Secondary use of electronic medical records for early identification of raised condition likelihoods in individuals: a machine learning approach

Jonathan Eric Turner

Centre for Health Informatics

School of Mathematics, Computer Science & Engineering

City, University of London

This dissertation is submitted for the degree of Doctor of Philosophy

June 2019

It is more important to know what person the disease has than what disease the person has.

Hippocrates of Cos

When you have exhausted all the possibilities, remember this: You haven't.

Thomas Edison

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ACKNOWLEDGEMENTS

During the course of the work leading to this thesis, I have had six supervisors. I would like to thank three of them: Dr Peter Weller, for his conversations leading to the commencement of this work; and Dr Dympna O'Sullivan and Dr Jon Bird, for their input and good humour in the later stages of this work. I thank also my examiners, Dr Matthew Sperrin and Dr Tarek Besold, for their comments on my thesis which have led to what I hope are major improvements in this final document.

DECLARATION

This dissertation is the result of my own work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text.

It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

In accordance with the City, University of London, guidelines this thesis does not exceed 100,000 words.

Powers of discretion are granted to the University Librarian to allow the thesis to be copied in whole or in part without further reference to the author. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Signed:____

Date:____

Jonathan Eric Turner

ABSTRACT

With many symptoms being common to multiple diseases, there is a challenge in producing an initial diagnosis or recommendation for diagnostic tests from a set of symptoms that could have been produced by a number of diseases. Often the initial choice of diagnosis or testing is based on a clinician's impression of the likelihood of that condition in a general population; however the opportunity may exist for modification of these likelihoods based on individuals' recorded medical histories. This data-driven approach utilises existing data and is thus cheap and non-invasive. A method is proposed by which an individual's likelihoods of having specified medical conditions are modified by the similarity of that individual's medical history to the medical histories of other individuals, comparing the prevalence of conditions in those other individuals' records who are similar to the individual of interest versus the prevalence of the conditions in those individuals who are dissimilar.

In order to maximise the number of records available for analysis, a process was developed for the merging of data from disparate sources that used different clinical coding systems, including extensive development of a technique for semi-automatically mapping clinical events coded in ICD9-CM to Clinical Terms Version 3 (CTV3), for which no existing mapping table was found. Semantically similar fields in the source code sets were identified and retained in the combined data set. 'Codelists' comprising multiple CTV3 codes for a variety of conditions were built that defined the presence of those conditions within individual records. The hierarchical structure of the CTV3 code table was utilised as a method of identifying codes that differed in structure but had clinically similar or related meaning. The optimum degree of granularity of the coded data to use in identifying similar records was investigated and used in subsequent analysis.

Two methods were used for discovering groups of similar and dissimilar individuals: the 'nearest neighbours' method and the grouping of records using a clustering process. Altered likelihoods for a range of conditions were investigated and results for the nearest-neighbours approach compared to the clustering approach. Results for adjusted condition likelihoods for 18 conditions are reported, together with a discussion of possible reasons for a change, or otherwise, in the condition likelihood, and a discussion of the clinical significance and potential use of information about such a change. logistic regressions performed on a selection of conditions KNN performed better than logistic regression when judged by F-score (or sensitivity and specificity separately), however situation more nuanced when looking at likelihood ratios: Logistic regression produced higher (better) positive likelihood ratios, but KNN produced lower (better) negative likelihood ratios. Logistic regression produced higher odds ratios.

List of Abbreviations and Acronyms

CTV3	Clinical Terms Version 3
EHR	Electronic Health Record
EMR	Electronic Medical Record
ICD-9	International Classification of Diseases, version 9
ICD-9-CM	International Classification of Diseases, Clinical Modification, version 9
ICD-10	International Classification of Diseases, version 10
ICD-10-CM	International Classification of Diseases, Clinical Modification, version 10
ICD-11	International Classification of Diseases, version 10
KNIME	Konstanz Information Miner
KNN	K nearest neighbours
NHS TRUD	NHS Technology Reference Data Update Distribution
NIH	National Institutes of Health
PHR	Personal Health Record
SNOMED	Systematized Nomenclature of Medicine
SNOMED CT	Systematized Nomenclature of Medicine, Clinical Terms
TRUD	*see* NHS TRUD
UMLS	Unified Medical Language System

1 INTRODUCTION

1.1 Motivation

There is a growing understanding that there may be useful information stored in the large amount of clinical records data that are now being created and stored electronically: if these records can be accessed and their data released for research then perhaps we can gather useful information on diseases, their prevalence and prevention for the purposes of general medical research and for enhancing the care of an individual. However there is also a growing understanding of the challenges involved in accessing and processing these data.

Medical diagnosis is an inexact science, with many signs & symptoms, or combinations thereof, being explicable by more than one condition. Indeed the number of conditions far outweighs the number of symptoms. Given a set of symptoms that could be caused by more than one different condition, it is natural for a clinician first to consider the most common conditions as being the prime candidates as the cause of the conditions. However, it is possible that a patient exists in a population sub-group in which the likelihood of different conditions varies from that of the general population. A technique is proposed which utilises information from an individual's record and from others to modify the likelihood of various conditions.

1.2 Aims and objectives

The objectives for this research were:

(i) A review of the literature relating to secondary use of clinical records, including an overview of relevant legislation and guidance. A review of the literature relating to secondary use of clinical records is presented in the second chapter of this thesis.

(ii) To investigate potential sources for data to use in the work described in this thesis. A survey of data sources is given in chapter three.

(iii) Develop techniques for aggregating data from different sources and which use different coding systems, in order to maximise the potential number of data sources. The techniques developed are described in the fourth chapter of this thesis; the fifth chapter shows the validation of a composite data set from three data sources.

(iv) To build 'codelists' – lists of clinical event codes that each indicate the diagnosed presence of particular conditions.

(v) Determination of the optimum granularity of clinical event data and the development of a technique of matching records in order to produce modified condition likelihoods for individuals and for groups. In the sixth chapter, the various factors required to prepare the data set for analysis and testing are described and derived. Results from the techniques developed are given in the seventh chapter of this thesis.

(vi) To discusses the results of this work and its clinical significance, including brief discussions for each condition evaluated concerning the benefits of earlier discovery of those conditions in individuals.

1.3 Contributions

Several contributions were made during the course of this work:

1.3.1 Code mapping

No existing mapping of ICD-9-CM codes to CTV3 codes was found. A semi-automatic technique for indirect mapping of these codes via SNOMED CT was developed and a mapping table for 4342 ICD-9-CM codes to equivalent CTV3 codes was generated and verified by a domain expert. The technique developed is generalizable. A paper based on this work is under review at the Health Informatics Journal.

1.3.2 Codelists

In order to determine whether a particular record is 'positive' or 'negative' for a particular condition it was necessary to generate sets of codes, or 'codelists', that each indicated the presence of that condition. The codelists generated for use in this project are presented in the Appendix 2.

1.3.3 Methods for patient matching by event history

Methods were developed for preparing data from clinical event histories to make the histories suitable for processing. These methods included mapping event codes to a specified granularity, classification of event codes as administration, symptom or diagnosis codes, and the development of techniques to identify individual records with raised condition likelihood for a particular condition, based on other events in those records.

2 LITERATURE REVIEW

2.1 Introduction

Although the general use of electronic records systems in healthcare has become increasingly common over the last 30 years [1], the use of electronic computers to benefit medical care was proposed longer ago. Lipkin and Hardy [2, 3] described the use of a mechanical punched card system as an aid to the differential diagnosis of 26 haematological diseases, with the system recommending further tests if it had insufficient information to recommend a diagnosis. By 1960 their punched-card mechanical system had been implemented on an electronic computer [4]; one year later, Warner et al [5] wrote of a computerized mathematical system to aid with the diagnosis of congenital heart disease. Also around this time, Ledley and Lusted [6] wrote of the potential for the use of electronic computers as an aid to clinical decision making, investigating the theoretical foundation for such assistance and concluding that “The great significance and importance of such a health computer network cannot be overestimated as an aid to increasing individual good health” but were forced to conclude that 'no such project is under investigation at present. This is surprising since the advantages of such a system are well recognized and present technological capabilities are more than adequate.'

By 1965, Spencer and Vallbonna [7] were able to write of several applications of computers in clinical practice and included a list of problems in the use of computers as a clinical aid – including 'lack of clinical relevance of the data provided as computer input', 'the paucity of proper statistical and mathematical techniques for analysing the data collected', 'insufficiency of data reduction techniques', 'the difficulty in establishing adequate usage of the computer by the physician' and 'equipment failures' – at least some of which challenges may still apply today.

These examples of the early use of computers in medical care are notable in that they are all examples of the secondary use of electronic medical records data - use that goes beyond the original, primary, purpose for which the data were recorded. Indeed, the first published article discussing electronic medical records appears to be Larry Weed's 1964 paper 'Medical Records, Patient Care and Medical Education' [8], with his PROMIS system in operation by the early 1970s [9].

For the purposes of this thesis, the definition of secondary use of clinical data proposed by the American Medical Informatics Association in 2007 is used: "non-direct care use of personal health information (PHI) including but not limited to analysis, research, quality/safety measurement, public health, payment, provider certification or accreditation, and marketing and other business including strictly commercial activities" [10].

Aickin [11] has written that challenges still remain in the re-use of existing clinical records as a basis for drawing research conclusions. He wrote that there is a 'paradox that the most prevalent conditions are also the research orphans', since it is 'difficult to do randomized controlled trials for these conditions' and there is a 'lack of generalizability of such trials to clinical populations.' Prokosch and Ganslandt [12] wrote of the challenges inherent in reusing medical records data, noting that '[c]onsideration of regulatory requirements, data privacy issues, data standards as well as people/organizational issues are prerequisites in order to vanquish existing obstacles'. D'Avolio et al [13] further discussed some of these challenges, crucially noting that most EMRs were designed to support clinician-patient interaction and not 'analysis of aggregated data as required by many secondary uses.' One effect of this is a problem discussed by D'Avolio et al, namely, that much potentially useful information is stored in medical records as free text, requiring the development of processing techniques to retrieve information. Kukafka et al [14] support the view that the design of electronic health records systems does not support the aggregated reuse of the data held in them, arguing the case for redesigning Electronic Health Records systems to support a 'focus on preventive health and socio-behavioural factors'. Judd and Kim [15] discuss the feasibility of having one system that can function both as an Electronic Medical Records system supporting patient care and as a medical research database, concluding that it was possible to design a system that allowed two views into the data held within

– one view giving a clinician all the required information about an individual, another giving researchers access to the data in the system but only after the system had stripped out information that could be used to identify individuals. Kim et al [16], in a paper specifically looking at the benefits of clinical data re-use for the pharmaceutical industry, wrote that 'as the [USA] continues towards increasing utilization of electronic health records, the potential value of ancillary activities such as monitoring quality, assessing population health, and clinical research, is becoming possible'.

The American Medical Informatics Association published a white paper in 2007 [10], which suggests that re-use of clinical data had an important role, but '[l]ack of coherent policies and standard "good practices" for secondary use of health data impedes efforts to strengthen the U.S. health care system. The nation requires a framework for the secondary use of health data with a robust infrastructure' [17]. A further AMIA white paper [18] built on the 2007 paper, suggesting possible items for inclusion in a national framework for reuse of clinical data.

The UK Department of Health published a 'Summary of Responses to the Consultation on the Additional Uses of Patient Data' in 2008 [19], covering several topics, including the use of anonymised data, pseudonymised data and identifiable data.

Examples of re-use of clinical data by computerised analysis dating back at least to the 1950s can be found in the literature – the work of Lipkin and Hardy , Ledley and Lusted, and Warner have been discussed earlier in this report. More recent work on the reuse of clinical records stored in electronic information systems is discussed here and is divided into particular areas of research.

2.2 Advantages of secondary use of clinical records data

Aickin [11] wrote that it was worth pursuing the potential of using information held in clinical records since formal clinical trials were expensive and that it was 'not obvious that a therapy administered in the setting of a trial is the same as would be administered in usual care.' Others too have written of the potential for reusing existing health data for research: Dean et al [20] described 'the Electronic Medical Records' flexibility to examine large cohorts as well as identify patients with rare diseases"; Pearson et al [21] felt that there was potential to combine data from health records with real-time information from online social networking sites and mobile technologies, which would 'undoubtedly play a role in future research efforts by making available a veritable flood of information, such as real-time exercise monitoring, to health researchers'. Walton et al [22] noted that reuse of clinical record data 'can provide information that is inaccessible to randomised, controlled clinical trials, which require ethical approval and informed patient consent because they are prospective and experimental. These requirements greatly reduce the inclusion of young children, pregnant women, very old and very sick people, and those unable to give informed consent. However, medical practice includes a high proportion of such patients who are underrepresented or excluded from clinical trials. Furthermore, considerations of feasibility and cost often limit the numbers of patients exposed to a drug to a maximum of a few thousand for comparatively short times. Computerised databases in primary care can extend times to many years of continuous care and the numbers of patients to millions; this would be impossible to do in any other way'.

In a large review of 136 published clinical studies, Grove et al [23] showed that "in general, mechanical prediction matched or out-performed expert prediction" both in terms of accuracy of prediction and in cost-effectiveness, though it was noted that this was not the case in all studies included in the review.

2.3 Challenges in re-using clinical data

It is not a simple matter to obtain a set of clinical data and analyse it. There are several areas of difficulty in doing this, in particular the identification of sources of suitable data, ensuring appropriate privacy and data security protection for patients who are the source of the data used, quality and content of the data set(s) used and homogeneity of coding. Elkin et al [24] has a list of some of the barriers to secondary use of clinical data, including issues with data interoperability (including having a common coding system), data being stored as free text rather than coded, errors in data entry. Other challenges include coding accuracy, equality of meaning, completeness and precision.

2.3.1 Identifying and obtaining data

One of the challenges involved in re-using clinical data for research occurs at the outset: identifying sources of data and gaining access to the data.

Publications concerning research based on existing clinical records are naturally focussed on the outcomes of analysis of their data sets and are often written by authors who deal with the source clinical data as part of their normal daily responsibilities and it is understandable that they do not address the challenge of obtaining a set of data to work with. No papers were found that specifically addressed the issue of identifying clinical data sets suitable for research and obtaining access to these data.

Gaining access to identified sources of clinical data requires ethical approval to use the data (or an exemption from ethical approval) and the agreement of the organisation holding the data – which may entail some cost.

A further potential source of data is the Personal Health Record (PHR). The PHR has been defined by the Medical Library Association/National Library of Medicine as "a private, secure application through which an individual may access, manage, and share his or her health information. The PHR can include information that is entered by the consumer and/or data from other sources such as pharmacies, labs, and health care providers. The PHR may or may not include information from the electronic health record (EHR) that is maintained by the health care provider and is not synonymous with the EHR. PHR sponsors include vendors who may or may not charge a fee, health care organizations such as hospitals, health insurance companies, or employers" [25].

In 2008 it was been estimated that around 70 million US citizens have access to a PHR, through their employer, healthcare provider or health insurer 'though most patients would not be aware of it' [26]. However, the number of patients actively using a PHR is less than this figure suggests: a survey for the California Healthcare Foundation in 2010 [27] suggested that only 7% of the US population (about 22 million people) used a PHR; a similar survey in early 2011 [28] found that about 10 % of the US population (about 31 million people) reported using a PHR. Aside from the low take-up rate, Zulman et al [29] in a survey of active users of PHRs found that users of PHRs were not representative of general population (90% were aged over 50, 92% were male, 39% reported 'poor' or 'fair' health).

No studies on the accuracy of the information entered into PHRs by patients or other non-professional sources were found, whether compared to information held within EHRs or independently Pre-dating the onset of PHRs, Harlow and Linet [30] evaluated, by review of the available literature, the accuracy of patients' recall of their medical histories (when compared to their medical records) and found that there were some significant differences: some conditions were more likely to be recalled by patients than to appear in their records (e.g. hay fever), others were less likely to be recalled (e.g. thyroid conditions) while other, perhaps more serious conditions were as likely to be recalled as to appear in their records (e.g. heart disease, diabetes). Van Deursen et al

[31] propose a method of ranking reliability of data according to its source by means of a reputation engine, but do not assess any actual data themselves.

2.3.2 Privacy of individuals and the security of their data

When reusing clinical records data it is essential to be mindful of the privacy of the individuals whose data form those records, from both a legal and ethical aspect. A review of the literature and legislation pertaining to patient privacy and data security is included in section 2.6 of this literature review.

It is important to limit access to identifiable clinical records only to those who have a need to see the records at any particular moment. Boxwala et al [32] investigated accesses to clinical records in one institution, taking a set of manually-categorized (as 'suspicious' or 'appropriate') accesses as a gold standard for machine learning models. The authors note that their methods "may not generalize because of interinstitutional differences', reflecting the challenges involved in re-using clinical data from different sources discussed in the introduction to this section.

2.3.3 Data quality, coding and text-mining

Effective re-use of clinical records data depends on that data being semantically consistent, accurate and complete enough for the purpose of the re-use. Ignoring image data, data items within records are either single values for a defined field (i.e. coded fields) or free text.

Stein et al [33] queried a clinical data set of around 5,000 discharge summaries that contained both coded fields and free text fields, to see whether the two types of fields contained information that was conflicting, confirming or complementary. Both the coded fields and the free text fields were searched for answers to particular clinical questions and the degree of concordance between the free and coded fields was calculated. The researchers concluded that there could be significantly disparate results between coded and free text fields, and that to obtain the best information it was necessary to search both types of field (turning to human assessment for the most accurate information). A similar study by Turchin et al [34] looked at 18,000 medical records which contained both structured data and free text fields. They found that around one third of the records had events recorded in both structured and free text fields, suggesting that, with two-thirds of the events recorded in only one of the structured and free-text fields, both types of field should be considered to obtain the most complete set of information. It is possible automatically to map text from clinical documents to codes: Friedman et al [35] describe a system that performs such work to a level of performance claimed to be comparable to human experts, while Turchin [36] utilised regular expressions to achieve a similar level of performance, reporting that ""By some estimates free text physician notes contain over 50% of the data in the patient's medical record."

Following on from a review of studies of EHR quality in primary care by Thiru [37] et al in 2003, Chan et al [38] published a review of the literature concerning data quality in electronic health records, reviewing 25 studies, each of which investigated some or all of data accuracy, data completeness or data comparability. They concluded that 'Issues related to data accuracy, completeness, and comparability must be addressed before routine EHR-based quality of care measurement can be done with confidence'.

A similar review by Liaw et al [39] suggested evaluating data quality using four dimensions, of 'completeness, consistency, correctness and timeliness', drawing similar conclusions to the Chan et al study as to the need to improve data quality to allow for improved reuse of clinical data.

2.4 Examples of clinical data re-use

Examples of re-use of clinical data by computerised analysis dating back at least to the 1950s can be found in the literature – the work of Lipkin and Hardy [2,3], Ledley and Lusted [6], and Warner [5] have been discussed earlier in this report. More recent work on the reuse of clinical records stored in electronic information systems is discussed here and is divided into particular areas of research.

2.4.1 Syndromic surveillance

Detection of outbreaks of diseases in populations have relied on clinicians informing public health bodies of patients in their care who have a notifiable condition – typically infectious diseases that can pose a serious health threat to an individual. Examples of such notifiable diseases in the UK include Legionnaires' disease, rabies and cholera [40]. Early identification of such outbreaks is vital for the control of the spread of these diseases, but delays in notification can hinder the detection of outbreaks [41]. Systems based on automatic interrogation of clinical records, with the intention of improving both speed and accuracy of notification, have been described by several authors. Gesteland et al [41] report on an 'Automated Syndromic Surveillance' system installed for the 2002 Winter Olympics in Utah. This system took information on patient encounters from 28 clinics, taking free text information on the reason for patient presentation and coding this for analysis. Two warnings were flagged during the period of the Winter Olympics but these proved to be false alarms, and fortunately there were no genuine disease outbreaks during this period.

Klompas et al [42] at Harvard University discuss their system for interrogating existing EMR data and automatically messaging information about any new cases of notifiable disease to the appropriate authority. This system was further developed for the particular case of Hepatitis B, scanning electronic medical records data for laboratory test results that would indicate the presence of hepatitis B [43]. Also from Harvard, Calderwood et al [44] discuss the algorithm they used to predict the presence of TB in patients, based on coded data in the records, determining that 'Live, prospective [tuberculosis] surveillance using EHR data is feasible and promising'.

Hripcsak et al [45] compared the use of structured data with free text information for syndromic surveillance, concluding that structured data performed best but required knowledge of the structure of the health records system, whereas the free text information performed less well but had applicability to a broader range of systems.

Buckeridge et al [46] developed a model that they suggest can be used as the basis for comparing different detection algorithms for their performance. Dailey et al[47] compared sources of data that are used to detect influenza outbreaks, including sources outside hospital records, including over-the-counter pharmaceutical sales and work absenteeism.

2.4.2 Performance and care quality measurement

Records have long been used in hospitals to keep a record of the number of patients examined or treated, to measure the quantity of work performed for the purposes of reimbursement. Korner Units were commonly used in the UK National Health Service, which quantified the workload of each treatment episode [48].

Several researchers have investigated the potential to analyse the information held within clinical records for the purposes of performance measurement and measurement of the quality of care received by patients. Owen et al [49] performed a feasibility study to see if the data held in an Electronic Medical Record system could be used to measure the quality of treatment of schizophrenia patients, deciding that it was possible to perform such measurements but that "electronic recording of depot prescriptions was possible but usually incomplete ... providers and facilities should improve recording so that automated data could be used to more accurately monitor and improve ... medication management for schizophrenia." Voorham and Denig [50] performed a study evaluating the feasibility of using free text data to extract the information required to assess the quality of care of diabetes patients, concluding that this was a practical technique. Also in the area of diabetes care quality measurement by use of free text data was studied by Pakhomov et al [51], who looked at findings from foot examinations (which are part of the programme of diabetes care), and again concluding that such automated analysis was practical.

Another study that relied on the processing of free text to assess quality of care was that of Chiang et al [52] who took quality control measures from electronic discharge notes in order to estimate the standard of care, claiming "reasonable agreement with medical experts." Chan et al [53] developed a process to assess the quality of co-ordination during the patient's transition from primary care to specialist care, although their method required recording of data additional to that recorded as standard in the records systems. The problems of using information sourced from several records systems were addressed by Lee et al [54], who described a "virtual medical record", a single repository for key data extracted from different systems, from which quality indicators were calculated.

2.4.3 Outcomes research

Dean et al [20] published a literature review of research covering the use of electronic medical records for outcomes research, for papers published between 2000 and 2006, finding 126 studies. It was noted that the number of studies published increased with each year, reflecting perhaps the increased uptake of EMR systems over this time period and the increased interest in using EMR systems as a source of data for research. The authors concluded with a comment that reflects others' concerns: “It is essential that standardized terms and codes be incorporated into EMR data for EMR-based research to be translated into clinical best practices.”

2.4.4 Specific outcomes predictions for individual patients

Prediction of patient outcomes following admission or treatment has been studied by several researchers. Himes et al [55] used a Bayesian network model to identify the clinical factors which could be used to predict asthma patients' progression to chronic obstructive pulmonary disease, taking data from the clinical records of one organisation. Testing on a set of nearly 10,000 patients achieved an accuracy of 83%.

A tool to predict an individual's chance of developing type II diabetes within the next 10 years was developed by Hippisley-Cox et al [56], who used a Cox proportional hazards model to estimate the effects of various risk factors.

Sebastiani et al [57], having selected a cohort of patients which covered all common phenotypes of sickle cell disease, used Bayesian network modelling to estimate the risk of death within the next 5 years, taking this as a measure of sickle cell disease severity. Their technique identified new markers, in addition to previously known risk factors, that contributed to the calculation of the risk score.

2.4.5 Decision support

A clinical decision-support system (CDSS) is, in its broadest definition, "any computer program designed to help health professionals make clinical decisions" [58]. Output from these programs can be derived from existing knowledge, from individual cases using artificial intelligence (AI) methods or from a combination of both; a variety of AI methods can be used, including rule-based reasoning, Bayesian inference, artificial neural networks and case-based reasoning.

Case-based reasoning (CBR) aims to solve new problems by "finding, adapting and reusing solutions to previously encountered problems" [59]. Useful introductions to this technique are given in Kolodner [60], Schmidt et al [61] and Yusof and Buckingham [62]. Recent developments are described by Bichindaritz and Marling [63] and Bichindaritz and Montani [64], who describe the CBR process from a physician's viewpoint. They list key application areas for CBR as being diagnosis, treatment planning, image analysis, long-term follow-up, quality control, tutoring and research assistance. A useful introduction to the usefulness of CBR is given by Ting et al [65], although they do note that "Despite numerous researches showing CBR is effective in problem-solving in the medical domain, several researchers argued that the chance of reusing a case from CBR is not high in some areas, such as ... multiple medical disorder cases".

Applications in the medical literature have generally utilised this technique to assist clinicians by employing information on similar previous cases to those under current consideration. For example, Kahn and Anderson [66] used CBR in a system that suggested the most appropriate diagnostic imaging procedure (within the ultrasound and computed tomography domains only), based on text information within case histories to choose the imaging procedure. Their study used 200 cases as the training set. Marling and Whitehouse [67] developed a system to aid in the care of Alzheimer's Disease, using CBR to determine whether a patient would benefit from administration of neuroleptic drugs, but using a rule-based procedure to choose precisely which of the available drugs should be prescribed. This study used 28 cases in its training set. In a study utilising 166 patients in their training set, Chuang et al [68] investigated using CBR with several other classification methods to support liver disease diagnosis, concluding that a hybrid model of CBR with back-propagation neural network gave the most accurate diagnosis results.

These studies take data, including outcomes data, from similar prior cases in order to help with decisions on care for newly-presenting cases. A key component of a CBR system is the task of finding similar matching cases has been addressed by researchers including. O'Sullivan et al [69] and van den Branden et al [70].

2.4.6 Drug actions and reactions

An active area of research is the study of identifying adverse drug reactions from medical records. Honigman et al [71], performing a retrospective analysis of data from an electronic medical record system, were successful in identifying adverse drug reactions, finding that "free-text searches were especially useful." Working exclusively with free text, Wang et al [72] processed discharge summaries to identify medications and adverse reactions, relying on natural language processing to do this.

Nadkarni [73] describes the problems in detecting adverse drug reactions using existing medical records data and in particular identifying problems with using ICD-9 and SNOMED CT, suggesting that free text fields can help with better detection of adverse events. Savova et al [74] looked at drug treatment patterns, rather than adverse events, combining drug treatment events from free text fields in clinical records with data from a prescribing system.

2.4.7 Identifying patients suitable for trials or other analysis

Wilke et al [75] describe a system for identifying patients with diabetes from electronic medical records. When searching using solely diabetes diagnosis codes they found false positive rates of up to 44 %; much reduced after implementing an algorithm that also included laboratory data and medical history. Kho et al (2012) also searched electronic medical records for diabetes patients, developing an algorithm that used a combination of diagnoses, medications and laboratory results. Subjects were identified across different records systems, with "the use of standard terminologies to define data elements ... across five different institutions"" noted as being key to the success of the work,

Clark et al [76] looked at free text in clinical reports to determine whether patients were smokers or non-smokers (or "unknown" if no references to smoking were found), reporting an accuracy of above 90 % in their data sets.

Attempting a more general approach to identifying patients for research purposes, Yamamoto et al [77] developed a system to identify patients from a single hospital medical records system based on appropriate clinical research criteria. They noted that "Enabling medical records retrieval system use in and across multiple institutions is an important future task."

2.4.8 General health outcomes events prediction

For more general health event predictions (i.e. predictions of likely future conditions, based on records of previous health events), fewer publications were found. McCormick et al [78] give a primarily theoretical description of a Bayesian hierarchical model for the selection of association rules, testing their method on a sample of patients from a clinical trial, predicting future medical conditions on the basis of common clinical histories.

2.5 Methods of analysis

Aickin [11] compared the maturity of analysis techniques used in clinical trials data with those techniques employed on data acquired during clinical practice, stating that 'Methods of analysis that deal with the biases caused by lack of a research intervention have not been developed to the same degree as methods for intervention trials.' This is not to say, however, that work has not been done on clinical practice data.

Doddi et al [79] studied medical insurance claims records, which included information on medical procedures and diagnosis, to see if there was an association between procedures and diagnoses, looking for association rules using similar techniques to market basket analysis (a technique used in retail business management to identify which products are most often bought together, in order to make predictions or recommendations of other products that the customer may purchase — see, for example, Tan et al [80] for a description of this technique). Tsui et al [81] used Bayesian text classifiers to analyse messages from health systems in real time, for subsequent statistical analysis in order to detect disease outbreaks. Investigating models for disease outbreak detection more deeply, Jiang and Cooper [82] took retrospective data from one US hospital's emergency department for one year, injecting synthetic data to simulate disease outbreaks. They used a Bayesian network framework to identify the disease outbreaks by time and geographic location. Also working in the area of disease outbreak detection, Que and Tsui [83] introduced a 'rank-based spatial clustering algorithm' as an alternative method for identifying disease outbreaks, claiming improved computational efficiency over previous methods.

Creighton and Hanash [84] developed an algorithm for mining association rules from genome data, although they did not suggest that this technique could be applied to clinical events.

A primarily theoretical paper on prognostic Bayesian networks was published by Verduijn et al [85], describing the potential for using this technique to clinical data. This paper was accompanied by a second paper [86] describing an application of this technique for predicting mortality following cardiac surgery. Also in the same year, Reynolds et al [87] described the use of Bayesian belief networks with test results from a variety of sources, giving an example of the use of this method for classifying tumours. Other examples of the use of Bayesian networks for clinical prediction have been given in the work of van Gerven et al [88] and Sakai et al [89].

2.6 Data privacy and security

2.6.1 Introduction

A key consideration in the re-use of data obtained from real patients is the confidentiality of the patients who are the source of the data. This section of the report describes the current thinking around confidentiality issues - thinking that is both influenced and captured by legislation, codes of practice and recommendations from professional bodies concerned with patient care. Confidentiality of medical records that are made available to researchers not directly involved in a patient's care is primarily achieved by anonymising or pseudo-anonymising the medical records.

2.6.2 Review of the legal issues

A literature review was performed with the aim of achieving an understanding of the issues relating to confidentiality, in particular anonymisation and pseudo-anonymisation (hereafter referred to as 'pseudonymisation', as is common in the healthcare literature), including definitions of terminology; the need for anonymisation and/or pseudonymisation; legal requirements, including relevant permissions required; techniques for anonymisation and pseudonymisation; risks of re-identifying patients from anonymised or pseudonymised data; and any other issues with use of patient-derived health records.

The NHS Confidentiality Code of Practice defines the terms 'anonymised' and 'pseudonymised" and their definitions and spellings are those adopted within this report, unless an original publication is being quoted. Their definitions are given here:

“This is information which does not identify an individual directly, and which cannot reasonably be used to determine identity. Anonymisation requires the removal of name, address, full post code and any other detail or combination of details that might support identification.”

“Pseudonymised Information: This is like anonymised information in that in the possession of the holder it cannot reasonably be used by the holder to identify an individual. However it differs in that the original provider of the information may retain a means of identifying individuals. This will often be achieved by attaching codes or other unique references to information so that the data will only be identifiable to those who have access to the key or index. Pseudonymisation allows information about the same individual to be linked in a way that true anonymisation does not.”

2.6.3 Legal requirements and professional guidance

The Hippocratic Oath, dating from the 5th Century BC, includes the words 'Whatever I see or hear in the lives of my patients, whether in connection with my professional practice or not, which ought not to be spoken of outside, I will keep secret, as considering all such things to be private' [90] and this requirement to keep information about a patient private remains. The General Medical Council writes that [the] “duty of confidentiality continues after a patient has died.” [91]

Kalra et al, quoting the European Parliament, noted that “data protection, and therefore the need for consent, does not apply if the data have been anonymised and the individual cannot be identified through linking the information to other publicly available data” [92, 93]. This requirement was enacted in UK law in the Data Protection Act of 1998 [94] and subsequently in the Data Protection Act 2018 [95]. A decision by the English Court of Appeal in 1999 ruled that use of anonymised patient data for research did not breach confidentiality [96].

Bourke and Wessely [97], writing from the UK, in a review of confidentiality issues in various areas of medical practice and research state that "data may be fully anonymised so that individual patients cannot be identified, in which case the Data Protection Act does not apply." Legislation and other literature relevant to issues of patient confidentiality are listed at the end of this section.

In the UK, the Department of Health's 1997 'Report on the Review of Patient-Identifiable Information' [98], known colloquially as "The Caldicott Report', reviewed the use and transfer of patient-identifiable information from between NHS organisations and from NHS to non-NHS organisations, making recommendations that should be implemented to safeguard patient privacy. A second report by the Caldicott Committee, ("Caldicott 2') was published in early 2013 [99].

2.6.4 The General Data Protection Regulation (GDPR)

Data protection regulations in the UK are now led by the General Data Protection Regulation (GDPR) of the EU (Regulation (EU) 2016/679) of 27th April 2016 [100], enacted into UK law on 23rd May 2018 by the Data Protection Act 2018 [95], although since it was enacted as a Regulation rather than a Directive the GDPR did not require UK legislation to become legally enforceable in the UK. Cornock [101] notes that the changes over the last two decades in the amount of information available on individuals, how that data is collected and the uses to which it can be put mean that the original 1995 directive is ' no longer fit for purpose'. Article 5 of the GDPR contains the seven principles of the Regulation, these being noted by Chico [102]:

[The GDPR requires] "that personal data is processed:

(a) lawfully, fairly and in a transparent manner;

(b) collected for specified, explicit and legitimate purposes and not further processed in a manner; that is incompatible with those purposes (purpose limitation);

(c) adequate, relevant and limited to what is necessary; in relation to the purposes for which they are processed (data minimization);

(d) accurate and, where necessary, kept up to date (accuracy);

(e) kept in a form which permits identification of data subjects for no longer than is necessary (storage limitation);

(f) in a manner that ensures appropriate security (integrity and confidentiality);

(g) in a way which demonstrates compliance (accountability)."

Chico goes on to state that Article 5.1(b) of the GDPR says that 'There is an exception to the 'purpose limitation' principle for scientific research (see principle (b) above) which states: further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), 'not be considered to be incompatible' with the initial purposes' (Article 5 1. (b)' and is thus 'a significant relaxation of the restrictions on repurposing personal data for scientific research purposes'.

The GDPR also includes some definitions of terms, which it is useful to note here:

"Pseudonymisation" means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person (Article 4, Recital 30(5).

“Anonymous information” is defined as information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable (Article 4, Recital 26)
“Identifiers” are pieces of information which are closely connected with a particular individual which could be used to single him out (Recital 159).
“Personal data” means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (Article 4:1).

Mourby et al [103] quote advice from the UK Information Commissioner’s Office that ‘pseudonymised data may be personal data or may be considered to be anonymised, depending on how easy it is to obtain the pseudonym keys’. Olimid et al [104] state that the GDPR does not apply to anonymous data, according to Recital 26 of the GDPR; the GDPR ‘recognizes the difference between two main categories of data: personal data and anonymous data’. Furthermore, quoting Schaar [105]“complete anonymisation of data is no longer explicitly required” .

The current GDPR legislation in Article 22:1 includes a ‘right to explanation’ of individuals regarding how decisions about them have been made. The individual ‘shall have the right not to be the subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.’ Further, the GDPR in Article 22 Paragraph 4 states that decisions “which produces legal effects concerning him or her” or are of similar importance shall not be based on the following categories of personal data specified in Article 9 Paragraph 1:

…personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person’s sex life or sexual orientation.

There is currently some debate on how this right is to be interpreted and implemented in practice. Wachter et al [106] in a paper entitled ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’, take the position that it is sufficient to inform individuals that their data has been used by an algorithm(s) to make decisions about them, giving them the basic design of the algorithm, but would not require giving details of any algorithms used. Wachter et al note that GDPR Article 22 Paragraph 3 states that a data controller “shall implement suitable measures to safeguard…at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision”, otherwise a person has “the right not to be subject to a decision based solely on automated processing”; they go on to say that this does not appear to be a legally-binding ‘right to explanation’.

However, Goodman and Flaxman [107] quote Articles 13 to 15 of the GDPR as giving persons the right to be told the purpose of collecting data about them and the right to access that data, including the right to receive “meaningful information about the logic (algorithm) and possible impact.”

Selbst and Powles [108] discuss both the Goodman and Flaxman [107] and the Wachter et al [106] interpretations of the Regulation, concluding that there is a right to explanation but that this right should be interpreted ‘functionally [and] flexibly”.

In summary, it appears that the GDPR allows for research use of data for purposes beyond which those data were originally collected, provided that the data is protected by anonymization or, provided that the keys are strongly protected, pseudonymisation. Legislation, guidance and codes of practice. This suggests that the work described in this report remains compliant with UK and EU law, as it did at the commencement of the work.It is unclear as yet as to how strictly the ‘right to explanation’ will be interpreted, however the algorithms explored are explainable.

2.6.5 Legislation and other literature relevant to issues of patient confidentiality

There is a substantial body of legislation, guidance, codes of practice and recommendations relevant to the storage, use, transmission, protection and anonymisation of health records in the UK. The key documents are listed here:

- UK legislation
    - Data Protection Act 1998 [94]
    - Human Rights Act 1998 [109]
    - Access to Health Records Act, 1990 [110]
    - Computer Misuse Act, 1990 [111]
    - Freedom of Information Act, 2000 [112]
    - Regulation of Investigatory Powers Act, 2000 [113]
    - Common Law
    - Copyright, Designs & Patents Act, 1988 [114]
    - Data Protection Act 2018 [95]
- European Union legislation
    - Data Protection Directive, 1995 [93] *(note: now repealed and superseded by the GDPR of 2016)*
    - General Data Protection Regulation (GDPR), 2016 [100]
- Guidance papers, codes of practice
    - Caldicott report [98]
    - NHS Care Records Guarantee 2011 [115]
    - NHS Confidentiality Code of Practice [116]
    - CfH Pseudonymisation Implementation Project [117]
    - Data Protection & Medical Research [118]
    - Anonymisation: managing data protection risk code of practice [119]
    - Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule Recommendations [120]
    - Confidentiality as part of a bigger picture--a discussion paper from the BMA [121]

2.6.6 Why anonymise or pseudonymise the patient data?

The right to use data derived from patients' records must be considered. A statement from the British Medical Association that 'Legally and ethically health professionals are responsible to patients for the confidentiality of the health information they hold.... there should be no use or disclosure of any confidential patient identifiable information gained in the course of professional work for any purpose other than the clinical care of the patient to whom it relates' [121] which reflects existing UK and European legislation governing the use of personal information, and implies that disclosure of clinical data obtained from patients to individuals other than those responsible for their care requires the removal of patient identifiable information. The implications for secondary use of clinical data are that the anonymisation or pseudonymisation process must be robust and non-reversible (except for cases where pseudonymised data should be traced back to the original patient, requiring relevant permissions so to do). The process should be reversible only when there is a potential benefit to the patient in doing so and when permission has been agreed by the patient's carer(s) that this re-identification can be done.

2.6.7 Techniques for anonymisation and/or pseudonymisation

The basic technique for achieving patient privacy when their records have a secondary use (e.g. in health research) is to remove those parts of the record that can be used to identify the patient. Some identifiers are obvious – the patient's name, for example – but others are not so immediately obvious – for example, should the patient live in a small community and within that community be the only sufferer from a particular disease. Also, simply removing identifying information and leaving the field blank or entering randomised information may not be in the patient's best interest; Pommerening [122] notes that "it could be important for the patient ... to learn about the results of a research project, for example, a genetic disposition."

The Health Insurance Portability and Accountability Act (HIPAA) of 1996 [123] lists 18 identifiers that need to be removed from patient information for it to be considered anonymous. Although HIPAA is American legislation having no jurisdiction in the UK, these identifiers provide a useful checklist when deciding whether patient data has been appropriately anonymised. A summary of the 18 items in the list is given in Table 2.1. As an alternative to removing the 18 identifiers, HIPAA also allows for 'professional statistical analysis and opinion regarding de-identification', deeming information to be de-identified for HIPAA purposes if a person “with appropriate knowledge and experience” deems that the risk of re-identification of an individual or individuals is 'very small" (all quotes from US Department of Health and Human Services guidance [120]).

Item number	Item
1	Names
2	Geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes
3	All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death
4	Telephone numbers
5	Fax numbers
6	Electronic mail addresses
7	Social Security number
8	Medical record numbers
9	Health plan beneficiary numbers
10	Account numbers
11	Certificate/license numbers
12	Vehicle identifiers and serial numbers, including license plate numbers
13	Device identifiers and serial numbers
14	Web Universal Resource Locators (URLs)
15	Internet Protocol (IP) address numbers
16	Biometric identifiers, including finger and voice prints
17	Full face photographic images and any comparable images;
18	Any other unique identifying number, characteristic, or code (note this does not mean the unique code assigned by the investigator to code the data)

Table 2.1 HIPAA patient identifiers [123]

In the UK, the Information Commissioner's Office has produced guidance on anonymisation [119], though not specific to health records. The guidance states that

"The current Data Protection Directive, dating from 1995, says that the principles of data protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable. It also says that a code of practice can provide guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible", adding that "The DPA does not require anonymisation to be completely risk free – you must be able to mitigate the risk of identification until it is remote." The guidance also draws a distinction between a general release of data and a more limited release, to known individuals or organisations, to which conditions can be attached.

2.6.8 Re-identification risk for individuals

There are many papers in the literature discussing work carried out on the risks of re-identifying patients from their medical records, even after key identifiers have been removed. Several researchers have identified free-text fields that may appear in medical records as an area of potential weakness. For fields that contain well-defined data items that can be used to identify patients (e.g. name, date of birth), it is clear that the original data in these fields must be removed or disguised. Free text fields, however, present a greater challenge – they contain clinically important information but may also contain information that can identify the patient. Dorr et al [124] describe the difficulties in ensuring all such identifiers are removed from medical records prior to making them available to researchers, concluding that a significant time input is needed to manually remove identifiers, and that automated removal is difficult. Beckwith et al [125], however, describe an open source software tool that attempts to find identifiers in pathology reports and they claim a high degree of success in doing this. Neamatullah et al [126] also describe a software tool that attempts to de-identify free text in medical records to HIPAA requirements, again claiming a high degree of success.

2.6.9 Patient attitudes towards secondary use of de-identified clinical data

A survey for the NHS Information Authority in 2002 [127] found that "people felt that any information released outside of the NHS, or used inside the NHS for purposes other than treatment, should be anonymised - or patient permission sought to use identifiable data. Once information was anonymised, a majority ... were happy not to be asked for consent to share it."

In a survey from Ireland, Buckley found that "89.5 % [of survey respondents] said they would agree to ... allowing the sharing by GPs of anonymous personal health information with researchers without the need for consent" [128].

In New Zealand, Whiddett et al [129] surveyed 200 patients, finding that although they were generally unwilling to have their information shared with "other [non-health professional] stakeholders such as ... researchers", "they were more prepared to share anonymous information." Again in New Zealand, Parkin et al [130] report the results from a "citizen's jury" who unanimously concluded, after discussion, that "researchers contracted by a public body should be permitted to use medical information about identifiable people, without their consent" provided that "existing ethical guidelines and relevant laws" were followed.

Page and Mitchell in Canada [131] found that, of the 278 patients they surveyed, "the majority of subjects wanted to be asked for their consent unless anonymity was assured."

2.7 Predictive Analytics

Predictive analytics has been defined by Kelleher et al [132] as 'the art of building and using models that make predictions based on patterns extracted from historical data'. Cousins et al [133]have written "Predictive modeling tools incorporate mathematical formulas that allow users to interpret historical data and make predictions about the future. More specifically, these tools are used to create a predictive model by mapping associations and their statistical relationships among data elements to a specific target. The empirically derived model is then used to forecast future events based upon the identified relationships." Steyerberg [134] notes that 'prediction is primarily an estimation problem' and introduces specific areas in healthcare where such estimation can be of benefit to individuals: Screening, diagnosis, therapy impact. There are a variety of techniques available to build these predictive models and these will be discussed in this section. Particular reference will be made to work that has repurposed data from electronic health records. Goldstein et al write that "there are multiple advantages to EHR-based risk prediction ... allows one to observe more metrics, on more individuals, at more time points, and at a fraction of the cost of prospective cohort studies. One can use the same set of data to predict a wide range of clinical outcomes – something not possible in most cohort studies. As data are sometimes observed with greater frequency … it is also easier to predict near-term risk of events. Furthermore, patient populations derived from the EHR may be more reflective of the real-world than cohort studies that rely on volunteer participation.” However, Rose notes that “it is critical to remember that these data are not collected to answer specific research questions, which is a central difficulty in relying on them for these purposes.”

It has been a long established ambition to be able to prediction future health states and events from previous knowledge. Rahe et al in 1970 [135], for example, aimed to predict near-future health events in sailors based on information about their recent-years life events gathered by questionnaire. They concluded that there was a positive correlation in the rank-order of recent life events and illnesses experienced by the sailors during the 8 months of the study.

A number of techniques have been described in the literature. Steyerberg [134]writes that “Statistical models for medicine can be discerned in main classes: regression, classification, and neural networks”. Islam et al [136] identify several main areas of data mining techniques:

- Regression - Relationship estimation between variables
- Association - Finding relation between variables
- Classification - Mapping to predefined class based on shared characteristics
- Clustering - Identification of groups and categories in data
- Anomaly - detection Detection of out-of-pattern events or incidents
- Sequential pattern mining - Identification of statistically significant patterns in a sequence of data

Common techniques found in the literature are briefly introduced, together with examples from the literature of work done using these techniques.

2.7.1 Linear regression

Regression analyses aim to describe the relationship between dependent variable(s) and independent variables. Linear regression particularly analyses the linear relationship between a dependent variable, which must be continuous, and one or more independent variables, which may be continuous, binary or categorical. An introduction to linear regression is given by Schneider et al[137].

Work done using linear regression models includes that of Flemons et al [138], who used this method to predict the likelihood of sleep apnoea in 200 patients, of whom 82 were diagnosed with sleep apnoea, finding that their model was ‘superior to physician impression’. More recently, in 2018, another clinical prediction rule using linear regression was developed by Sanchez-Santos et al [139], whose model predicted the likelihood of patient-reported pain after total knee replacement. Combes et al [140]used linear regression techniques to predict hospital length of stay following admission to an emergency department, concluding that although there were limitations to their model, perhaps because of non-linearity in the data, the simplicity of their model meant that the medical staff using it could understand it.

2.7.2 Logistic regression

Logistic regression analyses the relationship between a binary dependent variable and a set of independent variables. Unlike linear regression, which aims to predict the value of a dependent variable from the set of independent variables, logistic regression aims to predict the category of the dependent variable from the set of independent variables, for example presence or absence of a disease.

This was a very common method found for creating predictive models in healthcare. Recent work employing logistic regression models includes that of Devin et al [141] who developed a model to predict the likelihood of return to work 3 months after cervical spine surgery; - A predictive model and nomogram for predicting return to work at 3 months after cervical spine surgery; Park et al [142], who used information held within Electronic Health Records to predict future incidence of Alzheimer’s Disease; and Kim et al [143], who used linear regression to predict osteonecrosis of the jaw following dental extraction. Kim et al compared their logistic regression model to other methods, concluding that it worked better than decision tree model but not as well as random forest, neural network or support vector machine, although their logistic regression model had the advantage of explainability. Other work using logistic regression includes D’Agostino’s work [144] on cardiovascular risk profiles for calculating patient risk of heart disease as part of the Framingham Heart Study; Chhatwal et al [145]who used logistic regression methods to aid breast cancer diagnosis; Singal et al [146], who used the technique to identify cirrhosis patients who were at raised risk of re-admission to hospital; and Jacobs et al [147] who used the method to combine three diagnostic methods in order to predict the risk of individual women having ovarian cancer.

In a recent systematic review, Christodoulou et al [148]have shown that they found no performance benefit of machine learning over logistic regression for clinical prediction models, although there was some suggestion that other techniques could perform better on data with a large (>100) number of variables.

2.7.3 K Nearest Neighbours

Another long-established technique is k nearest neighbours. "K-nearest neighbor classification involves retrieving the nearest neighboring entities to a new entity and assigning a category, or set of categories, to this new entity based on those already assigned to other entities in the space." [149] Another conceptually simple technique, it has long-standing popularity. Many researcher have used the method to leverage similarity in patient records or other patient-related data, for example in 2017 Tayeb et al [150] used the method to predict medical conditions in individuals by inspecting conditions in similar patients; similarly, Zhu et al [151]matched new patients to existing patients in a community care database in rural Canada to the predict rehabilitation potential of the new patient, concluding that the method was an improvement over the existing clinical assessment protocol.

A common application of the k nearest neighbours method is in predicting or detecting heart disease: Polat et al [152] used the technique for this purpose in 2007, as did Shouman et al [153] in 2012 and Enriko in 2016 [154], all claiming success for the technique.

2.7.4 Neural Networks

A Neural network is "a computer program that operates in a manner inspired by the natural neural network in the brain. The objective of such artificial neural networks is to perform such cognitive functions as problem solving and machine learning. The primary appeal of neural networks is their ability to emulate the brain's pattern-recognition skills" [155] Neural networks are another technique that is frequently applied to making predictions from medical records. Rajkomar et al [156] have applied the technique to “predicting multiplemedical events from multiple centers”, achieving high accuracy for in-hospital mortality, 30-day unplanned readmission, prolonged length of stay and final discharge diagnoses and concluding that the technique “outperformed traditional, clinically-used predictive models in all cases.” Pham et al used a similar approach to predict future health events from medical records, as did Chen et al [157]. Ma et al combined neural networks with medical knowledge for their risk prediction method, stating that their method “outperformed existing risk prediction models.”

2.7.5 Naïve Bayes

“The Naïve Bayes classifier is a family of simple probabilistic classifiers based on a common assumption that all features are independent of each other, given the category variable” [158. Much work has been done using this method for the prediction and identification of heart disease and breast cancer. Hollon et al {Hollon, 2018 #259]used this method to predict early outcomes after pituitary adenoma.

2.7.6 Decision trees

“Decision trees are sequential models, which logically combine a sequence of simple tests; each test compares a numeric attribute against a threshold value or a nominal attribute against a set of possible values. Such symbolic classifiers have an advantage over “black-box” models, such as neural nets, in terms of comprehensibility. The logical rules followed by a decision tree are much easier to interpret than the numeric weights of the connections between the nodes in a neural network” [159]
Decision trees have been successfully used in many predicitive and prognostic models, incuding in work by Lynch et al [160], who used the method to predict lung cancer patient survival; and Scheer [161], who developed a preoperative predictive model for pseudoarthritis.

2.7.7 Clustering

The clustering method involves grouping objects or records together in some way such that those objects in a cluster are more ‘similar’ to each other than to objects in other clusters. Hivert et al [162]used this method to identify primary care patients who were at risk of future diabetes or cardiovascular disease, based on information in their and others’ medical records.

2.7.8 Collaborative filtering

Collaborative filtering is a method used primarily in recommender systems, where information on an individual’s past behaviour, for example purchasing history, can be used to predict future behaviour by comparing them to others with similar histories. It “analyzes relationships between users and interdependencies among products, in order to identify new user-item associations. For example, some CF systems identify pairs of items that tend to be rated similarly or like-minded users with similar history of rating or purchasing to deduce unknown relationships between users and items. The only required information is the past behavior of users, which might be their previous transactions or the way they rate products.” This method can be extended into healthcare, substituting medical event history for purchasing history, and patients for users.

Less work utilising the collaborative filtering method was found in the healthcare domain, however one paper, by Davis et al [163], used the method. In their work, Davis et al used collaborative filtering techniques on a large (13,039,018) database of Medicare records of elderly people in the USA to predict their future health risks. They utilised the structure of ICD-9-CM in order to collapse disparate 5-digit codes to aggregate 4- or 3-digit codes, although they noted that this code collapse did not always improve their results. They conclude that their system performed “well” at capturing future disease risks.

2.8 Code mapping

Medical records since their inception have contained free text, with medical conditions and symptoms described by natural language terms that may be imprecise or ambiguous. Recent decades have seen a steady increase in the uptake of electronic health records (EMR) systems [164, 165]. There are now a large number and variety of terminologies used to code events recorded in these EMR systems, with it being estimated that there are over 100 terminologies currently in use [166]. For aggregation of data or analysis over time, a controlled, pre-defined vocabulary is required, with codes representing concepts that allow for descriptive synonyms [167]. A number of coding systems have been created over recent decades, including the International Classification of Diseases [168]; the Read Codes [169], the most common system in UK primary care, which in its latest iteration is Clinical Terms Version 3 (CTV3) but is most commonly used in version 2; and SNOMED CT [170], the largest coding system in terms of number of concepts. Modern electronic health records systems may use any of these existing systems, with different countries or regions favouring one system over another: in 2001, de Lusignan et al [171], in a survey of systems in use in Europe, found that the Read Codes were the most common system in use in primary care in the UK, ICD-10 the most common in primary care in Austria and Germany, and ICPC in a further 10 European countries. ICD-9, until its recent supersession by ICD-10, has been the dominant system in primary care in the USA.

There are several reasons why it may be necessary to move from one coding system to another: government mandate; the desire to use an up-to-date coding system; compatibility with other data repositories, e.g. in a newly-shared EHR system; combining data sets from disparate sources for research or audit purposes. Code mapping is an approach to enable codes from one system to be translated to their semantically equivalent codes in another system, a process that has been defined as “the process of associating concepts or terms from one coding system to concepts or terms in another coding system and defining their equivalence in accordance with a documented rationale and a given purpose” [172][9]. In order to combine data from diverse datasets coded using different coding systems it is necessary to converge the data sets onto a common coding system. At a minimum, translation of data items recorded in the coding system used in one of the source data sets to another coding system is required. However, the opportunity exists for all event codes from multiple sources to be mapped to a third coding system if that system has advantages over either of the coding systems in use in the existing data sets. When combining or comparing data sets from different sources using different coding systems, it is necessary to map clinical event codes to a single common coding system, which may be a system used by one or more of the source datasets or may be a new coding system. Bonney et al [173][10] write: “Mapping data elements in EHRs to a reference classification and/or terminology system not only facilitate reuse of primary care data for multiple purposes, but they also support data analysis, health information exchange and interoperability, and data comparison across the continuum of different healthcare providers [and] improves the quality of the research output derived from EHRs.”

An issue which can occur when combining datasets is that of semantic interoperability, in particular equivalence in the coding of clinical concepts. It is relatively straightforward to combine demographic information between systems since, for example, “there is general agreement as to what ‘age’ means in relation to a patient” [174][11] and there is similar agreement for names of individuals and dates, but it is less straightforward to map clinical concepts or their coded representation between different terminologies. One long-standing method is to match the text description of concepts (e.g. [175-179][12] [13] [14]; [15] [16]). The majority of work in automatic mapping has focused on the lexical approach, using techniques similar to those used for automatic mapping from free text clinical notes to concept codes (for example [180] [17] and [181][18]). However, Fung et al [182][19] found that ‘Semantic mapping performed better than lexical mapping’. Cimino and Barnett [183][20] proposed a method of semantic mapping by which each concept in a terminology was characterised by a set of properties, with concepts being mapped across terminologies according to the closest similarity in properties. This method requires each concept to be characterised manually in a process described as ‘tedious [but] not complicated’. A similar approach was proposed by Rocha et al [184].

Mappings exist between some of the major coding systems in current use, particularly between older and newer versions of coding systems, e.g. ICD-9-CM and ICD-10-CM; Read Codes Version 2 and CTV3, provided by several organisations and individuals, often those responsible for the maintenance of the coding systems. In the UK, the Department of Health Technology Reference-data Update Distribution service (TRUD) [185] provides mappings between a number of coding systems, in particular those systems in common use in the UK: SNOMED CT, Read 2, CTV3. In the US, mappings are provided between systems more common in that country by the Centers for

Medicare and Medicaid Services (CMS), International Health Terminology Standards Development Organisation (IHTSDO), Unified Medical Language System (UMLS), National Library of Medicine (NLM). Brouch [186] gives an introduction to the mapping process and contains a glossary of relevant terms. Nandigam and Topaz [187], describing their work in mapping SNOMED CT to ICD10-CM, note that the SNOMED CT to ICD10-CM mappings from NLM "may need to be modified on the basis of the clinical specialty and patient population and further validated." Previous work on creating mapping tables between coding systems has been primarily by human experts comparing text descriptions of codes in different coding systems, e.g. [188]; with the assistance of a text search tool, e.g. [189], [183]; or by automated text matching [184]. One problem described by Nadkarni & Darer [189] was that of missing mappings: in their work in investigating the completeness of mapping a data set from ICD-9-CM to SNOMED CT they found that 784 (of 2199; 35.8%) ICD-9-CM codes in their data set had no map to an equivalent SNOMED CT code in the UMLS ICD-9-CM to SNOMED CT cross-map, requiring them to create these mappings by hand.

2.9 Potential sources of data

A number of sources of data were identified for possible use in this project. They are listed here in Table 2.2. All are from clinical records or other professional records. It is not intended to use data from personal health records for the reasons of relatively low take-up and population bias noted in 2.3.1, although there may be scope for future use should analysis based on clinical data only prove promising. Note that the Harlow and Linet review [30] suggests that individuals may recall more conditions than they take time to report to clinicians and so using data volunteered by individual patients may help to increase the richness of the data set. However there are caveats to this approach: conditions may be inaccurately described or may be imaginery; dates may be misleading; bias can be introduced by those with better memory (hyperthymesia) recalling more conditions or conversely age-related memory impairment may lead to fewer conditions being recalled or being recalled inaccurately; older (or more recent) conditions may be preferentially recalled [190]; self-medication may lead to inappropriate treatment; ;

Data source	Country	Notes
Helseundersøkelsene i Nord-Trøndelag (HUNT)	Norway	Information on around 20,000 individuals from a geographically small region of mid-Norway. An application submitted to relevant Norwegian Regional Ethics Committee for use of the data, however following negotiations regarding fields to be supplied and the cost of doing so it was decided not to use data from this source
EPI-CT	Norway	The Norwegian part of this project aimed to acquire data on around 30,000 patients, all of whom had had CT scans. Clinical history for each individual was limited to information relevant to their CT scans and potentially to later-life cancers. Data coding was project-specific.
Clinical Practice Research Datalink (CPRD)	UK	Data from UK general practice patients, coded in Read Codes version 2, containing longitudinal medical event histories
The Health Improvement Network (THIN)	UK	Data from UK general practice patients, coded in Read Codes version 2, containing longitudinal medical event histories
Nottingham University Hospitals NHS Trust	UK	Data from patients attending specific clinics at a Nottingham hospital, access to the data to be arranged via the original PhD supervisor to this project, following on from previous work. Data was uncoded but some had been coded into SNOMED CT as part of the earlier work. Access to data later proved not possible due to supervisor absence
Practice Fusion	USA	Data from US general practice patients, coded in ICD-9-CM, containing longitudinal medical event histories
Informatics for Integrating Biology and The Bedside (i2b2)	USA	Contains uncoded discharge summaries for around 1500 patients
Mexican Health and Aging Study	Mexico	Data on medical and lifestyle histories of individuals acquired by interview of subjects and relatives, not captured contemporaneously. 11,000 households

		invited to participate. Responses are uncoded
Boston Medical Centre Clinical Data Warehouse	USA	Contains data from various information systems within the Boston Medical Center. Data is coded in ICD-9-CM and CPT.
Danish National Database of Reimbursed Prescriptions	Denmark	Contains prescriptions records only, not diagnoses or other event history and so not useful for this project.
Health Informatics Centre, Dundee	UK	Contains hospital laboratory data, prescriptions data, but no primary care records. Do data processing on-site and return aggregated results only. Unlikely to be useful for this project.
QResearch	UK	
The Hampshire Healthcare Record	UK	
PHARMO	The Netherlands	Based in The Netherlands; have general practice and other data. PHARMO do research in-house but may be able to make raw data available.
German Statutory Health Insurance Claims	Germany	
Veterans Administration Encounter Data	USA	
US Hospitals Databases: Premier, Cerner Health Facts	USA	
GE Healthcare EMR	USA	
Hospital Episode Statistics	USA	
Mediplus		
Disease Analyzer project		
Oncology Analyzer project		
Healthcare Cost and Utilization Project		

(HCUP)		
Centers for Disease Control and Prevention National Health and Nutrition Examination Survey (NHANES)	USA	
The Trauma Audit and Research Network	TARN	

Table 2.2 Potential data sources

2.10 Conclusions

There were a number of trends evident from the review of the literature. There was clear evidence of successful secondary use of clinical record data in many areas. Data from electronic medical records systems has been utilised both in near-real-time (e.g. for detection of disease outbreaks) and retrospectively (e.g. for selection of patients suitable for clinical trials). Successful secondary use of clinical records data used well-coded data or utilised natural language processing of free text fields; it is a challenge to use coded data from a typical electronic medical record system and a greater challenge to combine data from several systems.

Little work has been done on the detection of patients with similar clinical histories to that of a sample patient, although some work has been done for detecting patients with similar genomes. Likewise, little work seems to have been done on general predictions of future health events based on lifetime clinical histories, although there has been some work in specific areas. This gives rise to the key research area for this thesis: is it possible to modify individuals’ likelihoods of future health events simply by matching them with others who have experienced similar medical events to the individual of interest?

3 RESEARCH OBJECTIVES

3.1 Background

The research area identified in the literature review as being an opportunity for further research will be explored here. "Strong patterns, if found, will likely generalize to make accurate predictions on future data" [191]. However, "Mathematics works in Physics because purely physical processes can be idealized, and therefore simplified, to an extent that permits their handling by mathematical formulas. When it comes to biological phenomena, one finds that they are too complex to be represented by ideal cases without destroying their true nature, If, however, their complexity is kept intact, sufficiently powerful mathematical techniques will be lacking for their satisfactory handling" [192] quoting [193]. This is not the only challenge when re-using clinical data, and one relevant to the challenge of making risk predictions from medical records event histories. Drake and McHugo [194] note that data may well exist within electronic medical records, but since it was not collected for the purposes of research it may not be of a quality sufficient for it to be suitable for research use.

3.2 Research question

Clinical trials have brought much benefit and key knowledge. However there are areas where clinical trials are inappropriate. They may be ethically impermissible – encouraging a group to smoke; giving a 'placebo' CT scan; trialling a drug on pregnant women; or there may be insurmountable practical challenges. We therefore need to look at other ways to draw conclusions about health outcomes for patients.

In a related area, that of genome research, work has been done in associating genetic variants with phenotypes and further into associating genetic variants with the risk of individuals with particular variants being affected by a disease, their prognosis if they have a disease and their likely response to a particular treatment. For example, Kruppa et al [195] investigate a machine-learning approach to genome association for rheumatoid arthritis. However, it appears that little work has been done on performing similar analyses based on previous diagnoses and clinical events rather than the presence of particular genetic variants.

This leads to the key research question for the work presented here:

Is there potential to re-use data from multiple data sets, acquired for the primary purpose of the care of individuals, to enhance our knowledge of health events of populations, and to improve the future health of individuals based on this knowledge?

3.3 Practical uses

There are some practical applications for which this knowledge can be beneficial:

3.3.1 Screening: Pre-emptive care

Should it become apparent that an individual's clinical history suggests a probable future health path, then some appropriate pre-emptive care may be available. For example, should an individual demonstrate a raised risk of atherosclerosis, appropriate interventions can be put in place, such as encouraging increased exercise in the individual at risk [196]

3.3.2 Screening: Lifestyle adjustments

Future health states may be improved more by some lifestyle changes than others. It may be possible to establish this from analysis of individuals with similar clinical histories who have subsequently made different lifestyle choices – for example with diet, exercise or tobacco use changes.

3.3.3 Treatment options decisions

Some types of patient may respond better to one type of treatment; another type may respond better to another treatment; some patients may respond best with no treatment

3.3.4 Healthcare enterprise resource management

If a healthcare enterprise is better able to predict the likely care needs of its patients it may better be able to allocate resources

3.3.5 Other potential benefits

Decision support: An individual’s raised likelihood of a condition compared to prior assumptions of that likelihood can be presented to a clinician as an additional source of information to aid diagnosis.

3.4 Contributions

It is anticipated that new contributions to knowledge may be possible as a result of this work. In particular:

3.4.1 Combining datasets

Many of the examples found in the literature of re-use of clinical data have relied on data from a single data set or have reported challenges when using multiple data sets. This suggests that improving the ease with which data sets can be combined would be a useful area of work.

3.4.2 Selection of patients with similar clinical histories

Part of the core work for this project will be to attempt to predict the likelihood of future clinical outcomes or health states by comparing a single patient with others who have similar longitudinal clinical histories. The important elements of those clinical histories need to be established.

3.5 Testing predictions

Predictions made following analysis of the data sets must be tested. It is planned to retain a subset of the obtained data to be used as a test set against the techniques developed using the rest of the data.

3.6 Project outline and methods overview

3.6.1 Summary

It is intended to acquire clinical event history data from existing clinical records repositories and to combine these data sets into a single repository of longitudinal records data. From this data set, methods will be developed to group individual records by the similarity of their contained events. Using the grouped data set, predictions will be made regarding the likelihood of a record containing a condition off interest. A summary of the process is described in this chapter.

3.6.2 Data acquisition and consolidation

Acquire data from existing repositories of longitudinal records and combine into a single composite repository, as shown schematically in Figure 3.1. A similar system was suggested by Celi et al [197].

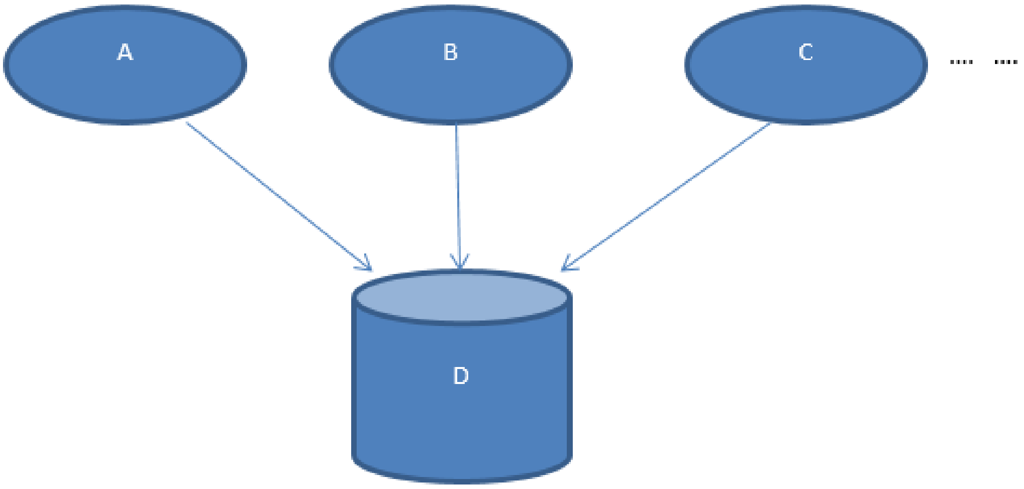


Figure 3.1 Schematic of data acquisition

Repositories A, B, C, … hold details of clinical events for individuals. These data may be held in different ways, with different fields and utilising different coding systems. Data from A, B, C ... are extracted and merged into a single repository D.

One issue that can arise with acquiring data from multiple sources is that of data consistency. In the traditional definition of data consistency, data across all systems reflect the same information and are synchronised with each other [198]. However, with the system proposed here, there may be subtleties with the consistency of the data caused by use of data from multiple clinical systems. It is possible that an individual patient is present in multiple systems but in the work here an assumption has been made that there is only a realistic possibility of that happening in the two UK-sourced data sets; it has been assumed that the chance of an individual having a primary care record in both the UK and the US is small. The two UK-sourced data sets have data sourced primarily from different system providers and so the chances are small (but still finite) that the same practices are used so only patients who move practice may be in multiple practices, but their records should move with them rather than be copied. Section 6.2.f tried to find duplicate records. No adult exact duplicates were found.

If an individual's records are split across multiple data sets (or , indeed, split across separate records within the same source data set) then they will not be recombined since all data used for this project were de-identified at source, and so they will be treated as separate records. The assumption is that for UK-sourced records the established system of maintaining the integrity of patient records works well. However, should an individual's records be split over multiple 'patient' records then this is likely to reduce the ability to make accurate predictions for that individual. However, since all records used in this work were sourced from primary care via reputable data aggregators an assumption was made that records were likely to be consistent,

Within single records there is a possibility that record events are internally inconsistent, with a later event occurring after an earlier event would have made the later event impossible, for example arthritis recorded in a foot that had previously been amputated. These events have not been tested for. Although it would be possible to do so, this would rely in many cases on specialist medical knowledge that was beyond the resources available for this work. It should also be noted that (i) should any 'impossible' event combination be detected, it would not necessarily be simple to determine whether the earlier or later event was incorrect, and (ii) any event determined to be incorrect should still remain in the medical record as part of the medical history, although labelled as incorrect. In this work it was accepted that there would be some inconsistency in the real-world data used.

3.6.3 Content of composite data repository

The repository will contain, for each record, a set of event codes. Codes that are significant for this work will be identified and retained. Figure 1.2 shows an illustrative timeline of events contained in one record.

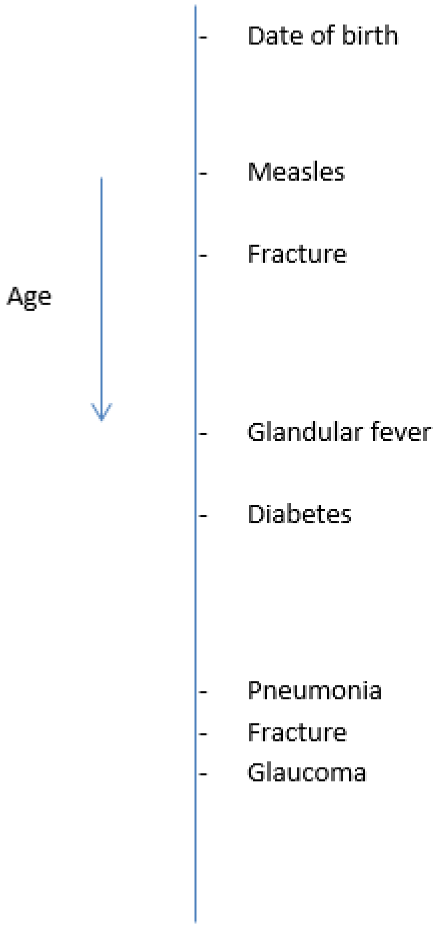


Figure 3.2 Illustrative timeline of events for one record

3.7 Considerations for data content

3.7.1 Level of detail

Consideration will be given to how detailed each data item should be. Recorded events may have a very fine level of detail, which may cause challenges in finding matches in other records, or may have a coarser level of detail, which may reduce the potential for differentiating between events and thus differentiating between records.

For example, a bone fracture can have information on the fracture site, in ascending order of granularity:

(a) Bone fracture (no site information)

(b) Fracture of the foot

(c) Fracture of a toe

(d) Fracture of the third toe on the left foot

A disease type ‘diabetes’ could be:

(a) Diabetes (with no particular information about the type of diabetes)

(b) Type 1 diabetes

(c) Type 2 diabetes

(d) Gestational diabetes

(e) One of several other less common forms

3.7.2 Finding matches and making predictions

It is intended to make predictions about future medical events (what, when, how bad) for newly-presenting individuals based on life histories of 'similar' individual(s) already in the composite repository. This breaks down in to two basic tasks:

(a) Finding matches

(b) Making predictions based on matches

As a simple example, if the patient illustrated in Figure 3.2 was in our database, and a new patient presented with a similar history but with one or more events not in the first patient’s history, we may be able to make a prediction of likely future health events (provided, of course, that there was not a greater weight of counter-predicting individuals also in our database). In practice it is expected that a number of matching individuals will be used to make predictions.

Possible differences in the ‘type’ of event have not been taken advantage of, although this remains possible for future work. A “likely externally induced fracture” could have had its likelihood increased by previous conditions or not (was it because the individual was a young person playing football? Or an elderly person with osteoporosis?); the fracture itself could make other conditions more likely, perhaps by reducing an individual’s exercise in both the short term and the long term. “Internal” conditions may have an internal cause (genetic) or an external cause (lung cancer, some diabetes). It is also difficult to say whether a particular event has a long “incubation” period (e.g. poor diet/poor exercise -> diabetes; smoking -> lung cancer) or a short one (food poisoning, for example). Without any a priori evidence for hard divisions between conditions it was felt inappropriate to make judgements about what conditions to include and exclude. Similarly, no different weighting was given to diagnoses over symptoms. This work focused on events in the record and made no medical judgement on the possibility of some conditions having a greater effect on other future conditions than others.

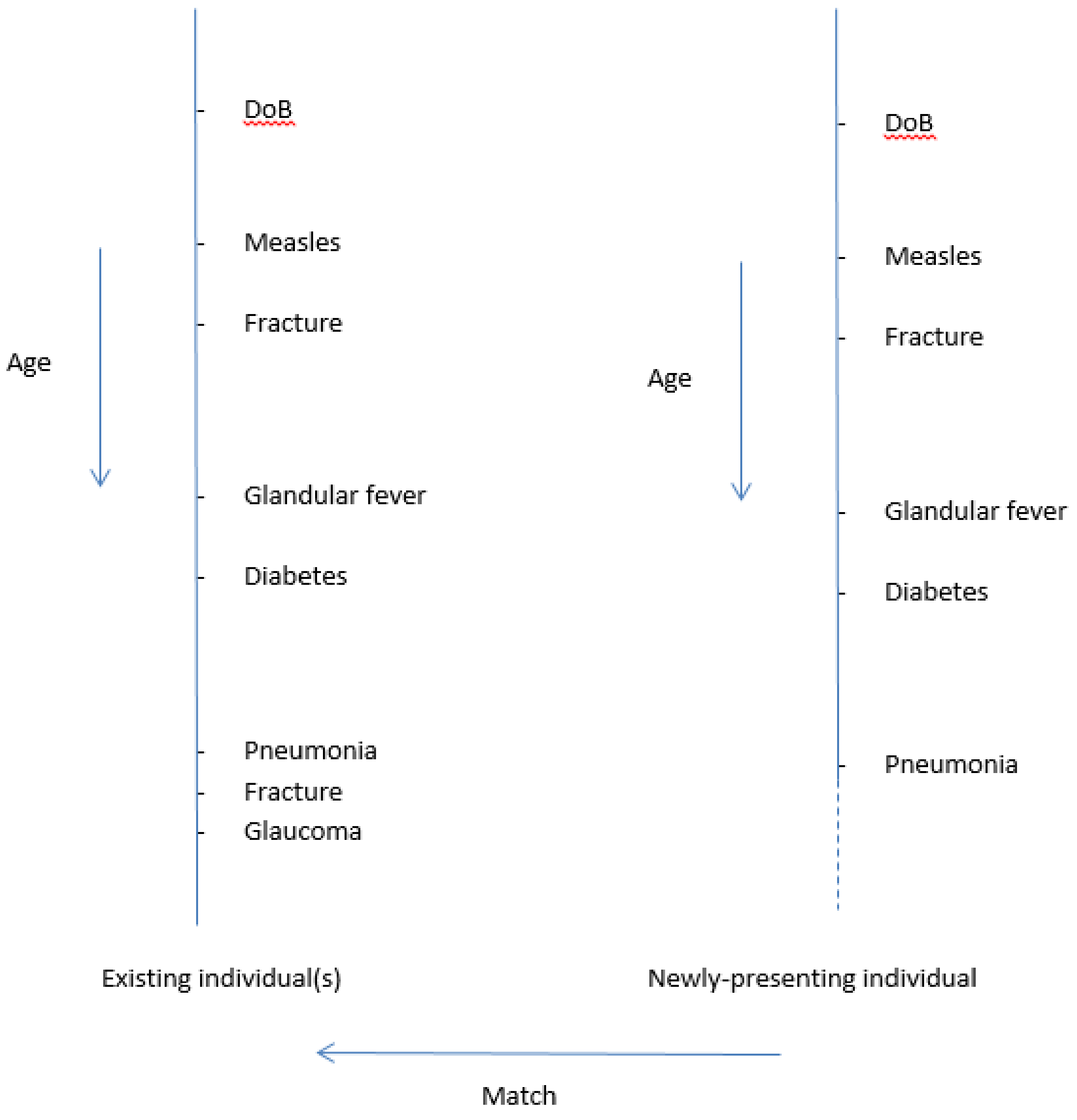


Figure 3.3 Timeline of lifetime events from a record existing in our database compared with timeline of newly-presenting record.

Thus, given a close match of the newly-presenting record to a record (or set of records) in our database, it would be possible to make a prediction of future events. In the example shown in Figure 3.3, there is a close match for most of the life history and so we could predict that future life events are likely to include fracture and glaucoma.

3.8 Finding matches

3.8.1 Choices for matching set

Several options exist for finding matches and acting on information gleaned from those matches to make predictions for the record of interest. These include:

(a) Find the single closest match and only use the information in that single match. See Kantardzic [199] for a discussion of the issues with this technique.

(b) Find a group of close matches, with the size of that group to be imposed in advance or determined from the characteristics of the data set under investigation, and use the information from that matching group. See, for example, Kelleher et al [[132]

(c) Find all the individuals that match to a defined degree (e.g. ’70 % or more of events in a record must match the target record to qualify as a match’). See Wu et al [200] for a discussion of this method.

(d) Where more than one record is included in the matching group, there may be potential to weight members of the matching set according to their individual degree of match to the target record. Dudani [201] has a description of this method.

3.8.2 Challenges

(a) 'Missing' events. Did a condition or event never happen for an individual? Or was the event just not recorded, or not reported by the individual?

(b) Incomplete longitudinal records. Is the complete set of events for an individual available, or are some records left-censored, right-censored or have gaps?

(c) Is it enough to look only at recorded conditions, or should such factors as age, tobacco use history, alcohol consumption history be used?

(d) Should different conditions be weighted relative to each other as well as with their own intra-condition severity? E.g. heavy cold vs mild pneumonia – which is more important?

3.9 Methods

This work falls into three main parts:

(i) Data aggregation;

(ii) Validation of the aggregated data set;

(iii) Calculation of modulated condition risks for a defined set of conditions.

Methods expected to be employed are briefly outlined here but will be described in more detail in the relevant sections of this report.

3.9.1 Data aggregation

Primary care records data will be taken from several sources. It is expected that the records will have some common fields (e.g. patient gender, age) but may have greater or lesser information on other demographic information, such as marital status or prescriptions. Events recorded in the records may be in different coding terminologies and so records from some sources may need to be mapped to a single coding terminology. It is expected that existing code mappings (e.g. from the UK NHS TRUD or NIH UMLS) will be sufficient for such code translations.

3.9.2 Validation of aggregated data set

Once the aggregated data set has been built, it will be examined to ensure it is representative of the general population. It is anticipated that this will be done by (i) analysing the data set for general demographic information and comparing this to demographic information available from population census information or similar and (ii) checking that prevalence of particular conditions in the data set are not significantly different from the prevalence of the same conditions found in the literature.

3.9.3 Calculation of modulated condition risks

Having prepared and validated the data set, its use in a system to use clinical records data to screen individuals for increased risk of particular conditions is investigated. Methods of calculating an individual’s risk modulated by that individual record’s similarity to other records will be developed and analysed. Any increase or decrease in risk will be presented in a clinically meaningful fashion, by likelihood ratio, prior and posterior probabilities, and odds ratio as appropriate, together with appropriate measures of uncertainty, and by comparison of change in absolute risks.

As a means of gaining familiarity with secondary use of data, a case study on the EPI-CT project is discussed in Chapter 4. A second case study looking at the acquisition of health histories directly from individuals and the feasibility of its use as a data source in this work is also discussed in Chapter 4.

4 CASE STUDIES

Two case studies are described: The EPI-CT project, which re-used clinical data as a key part of the project, and a project to obtain individuals' recollections of their medical histories. The projects illustrate the potential for secondary use of historical record data for research and the need for appropriate data management. Good and bad points from each use case will be discussed and used to inform later work.

4.1 Case Study: EPI-CT project

A presentation based on work in the EPI-CT project, an international epidemiological study to quantify risks for paediatric computerized tomography and to optimize doses, was made to the 'Data: storage, management, generation and legislation ' meeting in London in 2013: Turner J, Istad TS, Olerud HM, Flatabø S, Liland A, Ali W, Kjærheim K. "EPI-CT: International Epidemiological Paediatric CT Study. Data extraction and patient privacy protection: The approach in Norway." At IPEM Data: storage, management, generation and legislation, London, 16th April 2013. This case study report is based on that presentation.

Note: I worked with the EPI-CT project in Norway for six months, installing data extraction software in several hospitals, checking that data extraction was running as intended and with minimal impact on clinicians, and checking that the data extraction was appropriate and secure.

4.1.1 Introduction

A case study is described that illustrates some of the points discussed in Chapter 2 regarding re-use of clinical data for research purposes. In particular, the project shows some of the advantages and disadvantages of re-use of clinical data versus prospective clinical trials and also shows how the necessary patient privacy considerations have been addressed. The project is intended to investigate the effect of only one clinical event (exposure to diagnostic ionising radiation, which may be repeated) on the risk of another clinical event (cancer, although of several different types) in later life. However, although limited in the breadth of input and output events under investigation, the project illustrates the feasibility of re-using clinical data for research.

4.1.2 The EPI-CT project

The "Epidemiological study to quantify risks for paediatric computerized tomography and to optimise doses" (EPI-CT) investigated the relationship between exposure to ionizing radiation from diagnostic x-ray examinations (in particular, CT scans in childhood, adolescence and young adulthood) and increased health risks (specifically cancers). Eighteen centres from Belgium, Denmark, Germany, Finland, France, Luxembourg, the Netherlands, Norway, Spain, Sweden and the United Kingdom cooperated in this project, which aimed to enrol approximately one million patients over the 5-year duration of the study. Work was funded under programme FP7-EURATOM-FISSION, Grant agreement ID 269912. Table 4.1 gives information on each country’s contribution to the project, including projected cohort numbers [202].

The EPI-CT study was coordinated by the Section of Environment and Radiation at the International Agency for Research on Cancer (IARC). It received financial support from the Seventh Framework Program of the European Commission. The project completed in 2017, with results available on the CORDIS website [203]

This case study focuses on the contribution from Norway to this study. The Norwegian team comprises of staff from two centres: the Norwegian Radiation Protection Agency (NRPA) and Cancer Registry of Norway (CRN). Both centres are based in Oslo. Each centre had responsibility for different parts of the Norwegian contribution to the study.

	Cohort and exposure information				Outcomes	
Country	Cohort Age range	Start cohort accrual	Source of cohort information	Projected cohort size at outset	Childhood cancer incidence	Adult cancer incidence
Belgium	0-15	2002	PACS	30000	Yes	Yes
Denmark	0-18	2000	PACS	30000	Yes	Yes
France	0-5	2000	RIS/PACS	90000	Yes	Possible
Germany	0-15	1985	RIS/PACS	140000	Yes	No
Netherlands	0-18	1998	PACS	40000	Yes	Yes
Norway	0-20	2005	RIS/PACS	20000 (now 35 000+)	Yes	Yes
Spain	0-20	2005	RIS/PACS/other	200000	Yes	Since 2010
Sweden	0-18	1984	RIS/PACS/other	95000	Yes	Yes
UK	0-21	1985	RIS/PACS/other	400000	Yes	Yes
Total	0-21	1984-2002		1045000		

Table 4.1 Data Collection in Europe for the EPI-CT project

4.1.3 Project Design

At high levels of exposure, the effect of radiation on the human body is deterministic, i.e. above a particular threshold the severity of the effects of the radiation increases with increasing radiation dose (for example skin reddening, hair loss). Below the threshold effects are stochastic, i.e. effects are independent of the radiation dose, although their probability of occurring does depend on the radiation dose (for example cancer, genetic damage).

Most of our understanding of the long-term effects of ionising radiation exposure on the human body is derived from studies of survivors of the atomic bombs dropped on Hiroshima and Nagasaki in 1945 [204]. Although it is possible to find individuals who are estimated to have been exposed to similar levels of ionising radiation as those produced during diagnostic x-ray examinations, other factors may not be comparable, for example the length of time over which the radiation exposure took place or other factors that may pose a risk to health. There is also little information on early childhood cancers in Japan in the years immediately after 1945. However, studies on atomic bomb survivors and others have suggested that radiation at a level broadly similar to that used in diagnostic radiology can cause a small increase in the risk of induced cancer (Hall and Brenner, 2008).

In order to establish the existence of risks associated with diagnostic ionising radiation it is necessary to gather information on individuals who have received such diagnostic examinations. Since the effect of such relatively low levels of radiation is expected to be small, a large study cohort was required to achieve sufficient precision and statistical power to draw meaningful conclusions. In particular, the following design decisions were made [205]:

- To study only those individuals who have had one or more CT examinations when they were children, since CT examinations are responsible for approximately 80% of the population radiation dose due to medical examination and any effects of low-level radiation exposure are expected to take years, perhaps decades, to become apparent, so younger individuals would have more time for any ill effects to become apparent [206]. Prior to the relatively recent introduction of specific paediatric protocols, children were examined using the same protocols as were used for adults, thus causing them to receive higher effective radiation doses than adults during each CT examination [205]. For those individuals who had CT examinations as children, include also any CT examinations they may have had as adults
- To allow each country to run its own data collection programme, according to a common protocol. Record-keeping practices and definitions of "child" vary among the countries participating in the study; the protocol must allow for this.
- To incorporate existing studies underway in the United Kingdom, France and Germany.
- Use information on each examination to estimate radiation dose, producing a lifetime cumulative radiation dose for each individual due to CT examinations
- Follow patients over time to ascertain information regarding the incidence of leukaemia and other cancers.

Data collection commenced at any particular hospital by extraction of a list of examinations of patients who underwent at least one CT scan in the hospital when they were a child. This was done by querying the hospitals Radiology Information System

(RIS) for a list of all accession numbers (essentially serial numbers of examination requests) that match the stated criteria. It should be noted that in Norway, patients who were aged 20 or less at the time of their first CT examination were included; other countries vary as to the age at which they consider a patient to be a child.

Radiology Information Systems in Norwegian hospitals have been common since the early 1980s, approximately and fortuitously for this project coinciding with the widespread introduction of CT scanners in the country. Information held on the RIS included patient details (name, date of birth, ID number) and information on the examination (date of examination, body part scanned, scanner used, accession number). From the information on the body part scanned and the scanner used, a typical radiation dose for each type of CT scan (e.g. "head") was be assigned using the results of a CT radiation dose survey carried out in Norway in the early 1990s [207]. Other countries did not necessarily have such results of radiation dose surveys available and used other methods to estimate the radiation dose for each type of examination, typically specially developed questionnaires completed by staff working in the CT departments, results included in scientific publications or expert interviews with radiography staff still working in the relevant departments.

Beginning around the year 2000, Picture Archiving and Communications Systems (PACS) were installed in Norway and these provided a much richer set of data for each examination. In addition to the data provided by the RIS, the PACS gave information on the exposure parameters used, sufficient to enable calculation of radiation dose for each individual examination.

Data was extracted from PACS using PerMoS software [208]. This software ran on a PC attached to the PACS network and functioned like a normal PACS workstation. It used DICOM Query/Retrieve commands to retrieve individual CT examinations, based on the accession numbers obtained from the RIS. The metadata within each examination's data allows for calculation of exposure information. A feature developed during the course of the project was the use of the images within each examination's data to determine the physical start and end points of each scan on the patient's body. This was used to help determine more precise dose data for organs; for example, a "head" scan may or may not directly expose the thyroid, depending on where the start and end points of the scan were set, which will change the radiation dose received by the thyroid. This information will not be apparent from the metadata but can be determined from image data.

Installation of PerMoS was been technically straightforward but benefited from the cooperation of local staff; its use required some planning, since constant retrieval of CT examinations had the potential to overload some PACS networks. Image data for each CT examination can be large, in the order of tens or hundreds of megabytes. Continuous retrieval of CT examination data by PerMoS had the potential to slow down a hospital's PACS network and so PerMoS allowed for a configurable delay between each complete examination retrieval. Typically the setting was for a 10 minute delay between each retrieval but this was expected to reduce over time. PerMoS also allowed retrievals to run only at set times of the day, so that periods of high use of the network and PACS could be avoided. Retrieval times were set to avoid the busy morning reporting periods, the evening pre-fetch of examinations required for the next day's clinics and times (generally in the middle of each night) when system backups were set to run. This retrieval strategy ran successfully in Norway during the data collection period of the project and has caused no problems to local PACS.

The Norwegian cohort was originally projected to be around 20,000 patients but achieved more than 35,000 by project completion.

4.1.4 Patient privacy protection in the EPI-CT project Norway.

Following retrieval, each examination had its patient-identifying data removed and replaced with a pseudonym. Patient-identifying data and pseudonym only were sent to Cancer Registry of Norway for investigation of clinical histories; CT examination metadata with pseudonym were sent via the Norwegian Radiation Protection Authority to a central database for calculation of radiation doses. Calculated radiation doses were matched with clinical histories by their pseudonyms, then anonymised and made available for statistical analysis by the central EPI-CT study.

4.1.5 Advantages and disadvantages of secondary use of data in EPI-CT

The EPI-CT project illustrates some of the considerations and benefits when re-using clinical data for research purposes. These include:

4.1.5.1 Patient privacy protection

The minimum amount of patient-identifying information was acquired and was sent only to those groups needing it. In this case, the Cancer Registry of Norway required sufficient information to identify individuals' health records. Other organisations, those involved in estimating individuals' radiation doses, did not require knowledge of the identity of individuals and so received only pseudonyms with the data they were sent.

4.1.5.2 Economic access to large numbers of patients

The EPI-CT project utilised data which already existed in clinical information systems. This avoided the time and expense necessary to recruit individuals into a study. The effect of the intervention (the CT scan) on individuals was expected to be small and so large numbers of patients were required to demonstrate a significant effect. The effect of the intervention was expected to take many years to become apparent and so access to historical data on interventions was an advantage.

4.1.5.3 Ethical advantages

Some interventions do not lend themselves to clinical trials simply because it would be unethical to give the intervention if it not clinically indicated and likewise it would be unethical to withhold the intervention if it would normally clinically be requested. By using existing clinical records, this ethical issue is avoided.

4.1.5.4 Uncontrolled conditions

Initial criteria for inclusion of patients into the study were solely those of patient history of CT examination at a young age. The aim of the project was to establish whether there was a link between history of CT examination and incidence of cancer in later life. However, some patients may have had particular medical conditions that may have increased the incidence of cancer – and these conditions may also increase their likelihood of having CT scans. It may not prove as easy to allow for these conditions as it would be in a controlled trial.

Before data harvesting from PACS began, a RIS query is made for the patients to be included in the EPI-CT study, i.e. for the Norwegian cohort, all patients who have undergone a CT examination while being 0-20 years of age. From the RIS query a list of patient IDs was produced. These IDs were used to retrieve examinations from the PACS.

1. PerMoS Data Collector acted as a DICOM node on the PACS network, harvesting data from the PACS using standard DICOM query-retrieve. PerMoS Data Collector used the patient ID list from the RIS query to identify the relevant CT examinations from the PACS.

2. PerMoS Data Collector removed all image data, separated all patient-identifying information, generated pseudonyms and stored the harvested DICOM header data in local temporary files.

3. PerMoS Data Collector built a local database to keep track of which patients and which examinations had been harvested from the PACS. The database also contained the link between the pseudonyms and the identifying information for each patient.

4. The harvested DICOM header data was uploaded to the central PerMoS database. The data only contained CT scan parameters and pseudonyms, with no identifying information. This could be done either by automatic upload over secure Internet connection (HTTPS/SSL/TLS), or by manually moving the local temporary files on a hard disk (or other physical medium) from the hospital to the Norwegian Radiation Protection Authority for uploading from there. In both cases the hospital IT manager could inspect the files before uploading/copying. Figure 4.1 shows transfer of data out of the hospital using physical storage media.

5. The Norwegian Radiation Protection Authority checked the collected DICOM header data from all the Norwegian hospitals and used the collected CT scan parameters to calculate the radiation dose for each CT scan.

6. A linkage table consisting of pseudonyms + identifying info was transferred from the local PerMoS database at the hospital to the secure national database at the Cancer Registry of Norway. Only the database administrator and the personnel responsible for checking the harvested data at the Norwegian Cancer Registry had access to the identifying information.

7. The Cancer Registry used identifying information to collect health status (e.g. cancer disease status) and confounders from other Norwegian health registries and Statistics Norway.

8. Pseudonyms and calculated radiation doses were transferred from the central PerMoS database to the Cancer Registry. The pseudonyms were used for linking doses with patients in the national database at the Cancer Registry.

9. Dose, health status and confounders for each patient were transferred to the central epidemiological database at IARC (International Agency for Research on Cancer/WHO) for analysis. At this stage data was anonymous only, with all identifying information (including pseudonyms) having been removed.

All the relevant Norwegian data for the EPI-CT study was thus collected in a national database located at the Cancer Registry of Norway. This database contained the data from the RIS harvesting, including patient-identifying information, selected DICOM header parameters from the PACS, calculated radiation doses, the patient's health status, and confounders.

The database was secured according to the Cancer Registry's routines for handling personal information and according to the requirements of the Norwegian data protection act. Only anonymised data, which are impossible to link to individuals, were sent to the central EPI-CT research groups for analysis.

The number of workers handling personal information was kept at an absolute minimum. All those who had access to personal information were employees at the Cancer Registry of Norway, who signed a confidentiality agreement, and who were either database administrators or responsible for data cleaning.

The central PerMoS database in Luxembourg contained information on all the CT examinations included in the project, but no personal information. The purpose of this database was centralised, automated calculation of radiation doses from CT scan parameters.

No personally-identifiable information left Norway at any point.

In the central epidemiological database, the Norwegian anonymised data were combined with the data from the other European participant countries for epidemiological analysis.

4.1.6 Key points

- All data acquired for the project existed in various systems and databases, including hospital information and imaging systems and in cancer registries. No new data was required for the project.
- Impact on clinicians during data capture was the minimum necessary.
- Data acquired were kept secure, anonymised or pseudonymised as appropriate.

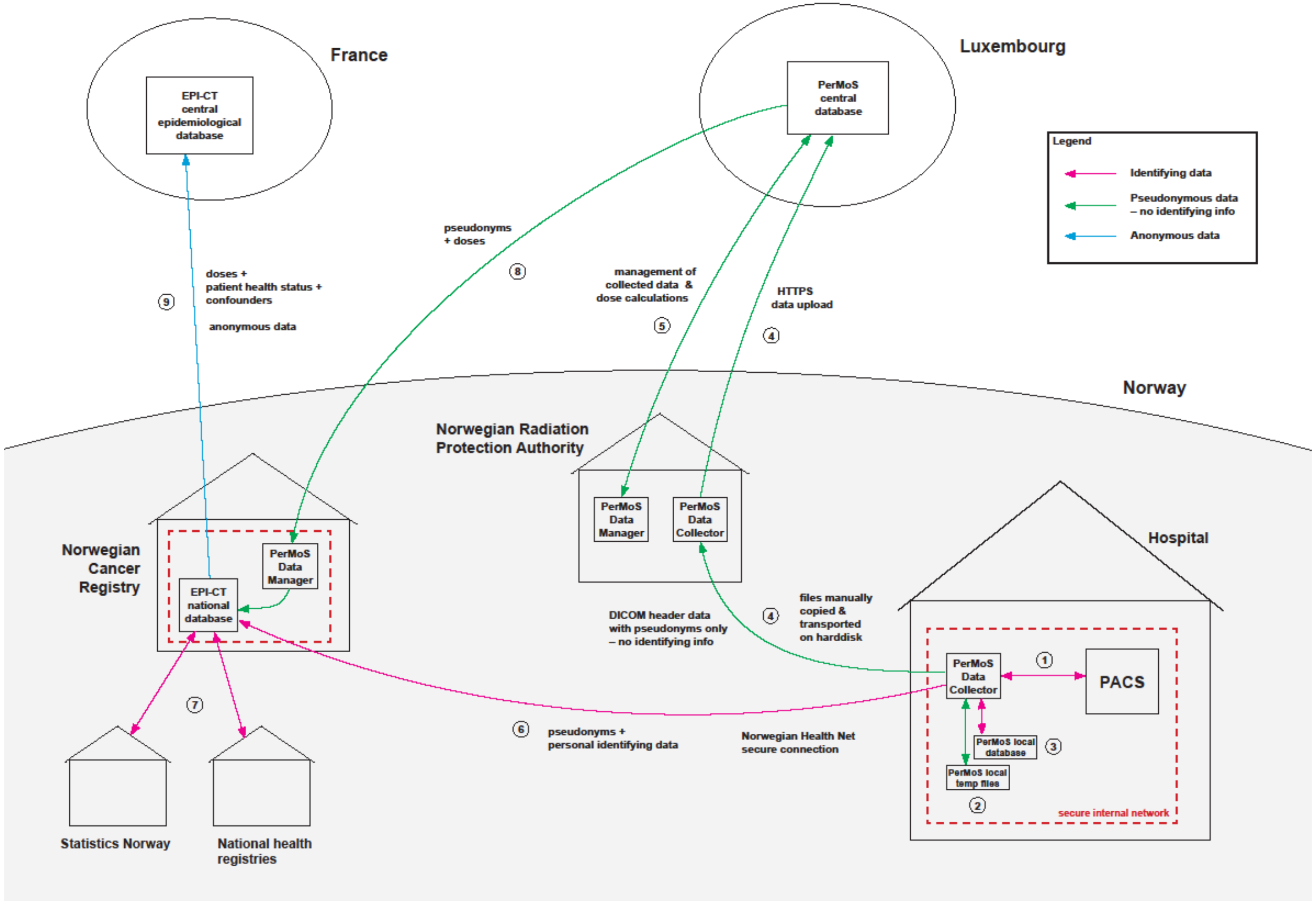

France
EPI-CT central epidemiological database
Luxembourg
PerMoS central database
Legend
Identifying data
Pseudonymous data – no identifying info
Anonymous data
pseudonyms + doses
8
doses + patient health status + confounders
9
anonymous data
management of collected data & dose calculations
5
HTTPS data upload
4
Norway
Norwegian Radiation Protection Authority
PerMoS Data Manager
PerMoS Data Collector
Hospital
Norwegian Cancer Registry
PerMoS Data Manager
EPI-CT national database
files manually copied & transported on harddisk
4
DICOM header data with pseudonyms only – no identifying info
PerMoS Data Collector
1
PACS
PerMoS local database
3
PerMoS local temp files
2
secure internal network
7
6
pseudonyms + personal identifying data
Norwegian Health Net secure connection
Statistics Norway
National health registries

Figure 4.1 Data flow using PerMoS for the EPI-CT project in Norway.

Diagram courtesy Norwegian Radiation Protection Authority

4.2 Case study 2: Self-reported medical histories

4.2.1 Background

This project was intended to collect anonymous information via the World Wide Web on individuals' recollections of their health events throughout their life, for the purpose of (i) comparing individuals' aggregated recollections of events to those stored in average health records and (ii) to see whether such individual-recalled events can be used to modulate predictions of future health events.

Note: I designed and implemented this project at City, University of London, Centre for Health Informatics, in order to investigate the challenges of acquired information from individuals about their medical histories, including quantity of responses, level of detail of individual events reported, and comparison with formal health records.

4.2.2 Methods

Survey respondents were invited to list, to the best of their recollection and without historical time limit, personal health events and ongoing conditions including, but not limited to, those events that were reported to or required the intervention of a general practitioner or other healthcare professional. Respondents were also invited to note their age group, weekly exercise habits, smoking status and alcohol consumption, and country of birth. In order to preserve anonymity, respondents were not asked for their name, exact age or current country of residence.

Once data had been collected, it was intended to use the data to:

i) To see how individuals' recollections of medical conditions compare, on average, to the quantity and detail of events typically stored in general practice records;

ii) As input to a health event prediction algorithm, in order to see whether patient-recalled data is of adequate quality to have practical use in such an algorithm.

It was expected that meaningful results would be obtained once 50 responses were received. This figure was calculated by use of the sample size calculator at http://www.raosoft.com/samplesize.html [209], using the default suggestions of 95% confidence and a 20,000 population size, but with a 10% margin of error rather than the suggested 5%, and a 10% response distribution (approximately the proportion of diabetes or hypertension in the general Western population) rather than the suggested

50%, which gave a sample size of 35, which was rounded up to 50 to allow for invalid or incomplete responses.

In order to maximise the survey completion rate, questions were designed to be simple and as few in number as possible, with only age and gender being compulsory. An open-ended question asking respondents to recall and list all medical conditions experienced through their life was deliberately asked before a closed list asking whether or not the respondent had suffered from particular named conditions.

The survey was implemented on a website using LimeSurvey survey software [210] installed on a server located in the UK using the URI http://www.predictivehealth.org.uk. The survey was advertised on www.callforparticipants.com, to the then current MSc Health Informatics students at City, University of London, and via Twitter.

The questions used in the survey are shown in Figure 4.2, taken directly from the survey web pages.

What is your age group?

Choose one of the following answers

- ○ Under 18
- ○ 18-25
- ○ 26-40
- ○ 41-50
- ○ 51-60
- ○ 61-70
- ○ 71-80
- ○ 81 or older

What is your gender?

Choose one of the following answers

- ○ Female
- ○ Male
- ○ Intersex
- ○ MtF Female
- ○ FtM Male
- ○ Other

What is your country of birth?

Answer

What ethnicity listed most closely matches yours?

Choose one of the following answers

- ○ Other
- ○ Black
- ○ Mixed/multiple
- ○ Asian
- ○ White
- ● No answer

What is your highest level of education attained? Please choose the closest match.

Choose one of the following answers

- ○ Compulsory school education only
- ○ Optional school education
- ○ Trade/technical/vocational qualifications
- ○ Bachelor's degree
- ○ Master's degree
- ○ Doctorate
- ● No answer

About how many cigarettes do you smoke in a typical day?

Choose one of the following answers

- ○ None - I have never smoked
- ○ None - I am an ex-smoker
- ○ Occasional (not a daily smoker)
- ○ 1-4 a day
- ○ 5-9 a day
- ○ 10-19 a day
- ○ 20-39 a day
- ○ 40 or more a day
- ● No answer

Do you regularly smoke a pipe and/or cigars?

Choose one of the following answers

- ○ Yes
- ○ No - have never regularly smoked a pipe or cigars
- ○ No - but used to
- ● No answer

Please enter all the conditions you can remember and please don't worry if you can only recall your approximate age at the time. We don't mind if you use the common name (e.g. "hay fever") or the formal medical term (e.g. "seasonal allergic rhinitis")

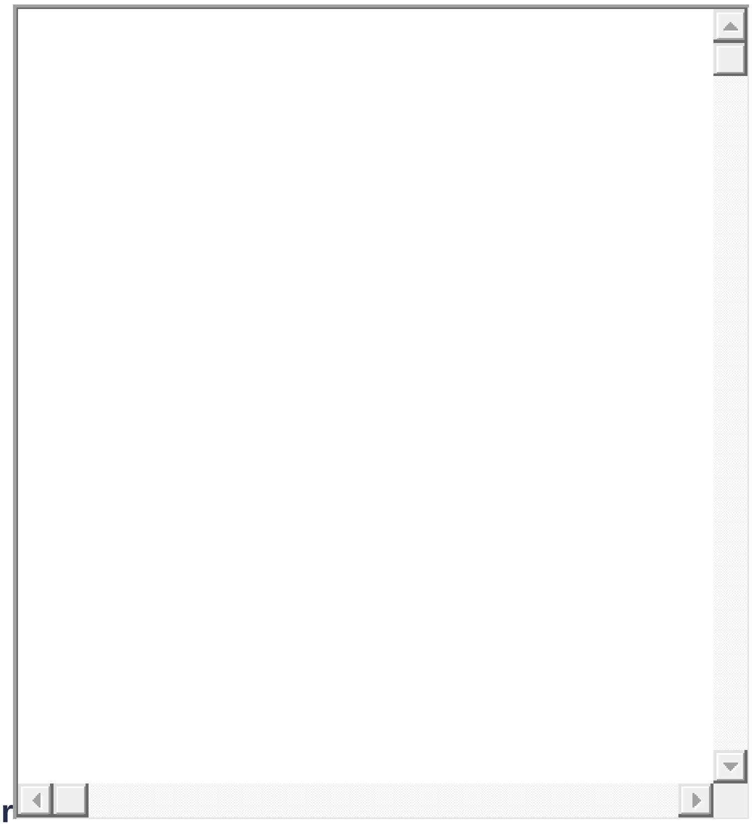
We are primarily interested in those conditions for which you attended a hospital or clinic, or consulted a GP. However, any conditions which you feel worth noting can be recorded.

Please enter each condition on a new line together with your age in years, as best you can recall, when the condition first occurred. The conditions do not need to be recorded in chronological order.

For example:

Field's disease - 24
Progeria - 1 ongoing

We appreciate that it can be difficult to remember all your medical conditions but please try to remember as many as you can.



- **Answer**

Have you suffered or do you suffer from any of the following conditions? Please enter your age in years when the condition first occurred.

Hypothyroidism (underactive thyroid)	Age
Anxiety disorders	Age
Insomnia	Age
Chronic Pain (including back pain)	Age
Depression	Age
Diabetes (please note type)	Age
Chicken pox	Age
Chronic Obstructive Pulmonary Disease (COPD)	Age
Hypertension (high blood pressure)	Age
Attention Deficit Disorder	Age
A "rare disease" (please describe)	Age
Tooth decay / gum disease	Age
Neuropathic pain	Age
Gastrointestinal disorders	Age
Bipolar Disorder	Age
Crohn's Disease	Age
Otitis (ear infection)	Age

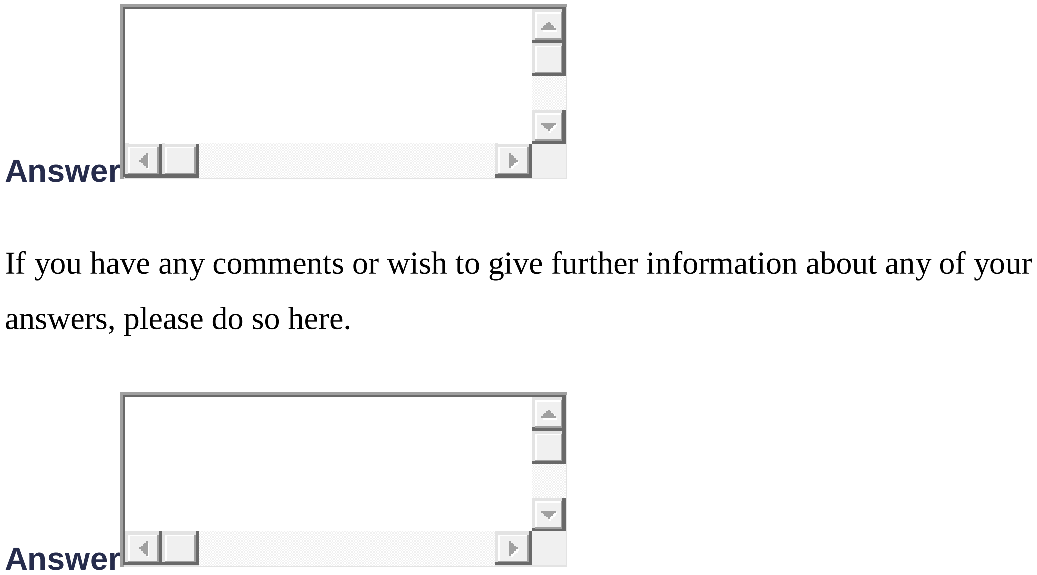
Asthma	Age
Allergies	Age
Sinusitis (sinus infection)	Age
Deep Vein Thrombosis (DVT)	Age
Chronic Dry Eye	Age
Arthritis/Osteoarthritis	Age
Rheumatoid Arthritis	Age
Anaemia	Age
Upper respiratory tract infection (coughs, colds, "flu")	Age
I have suffered from none of the above (write 'none' in the box)	

Do you have any allergies? Please list them here if you do, together with the age at which the allergy was first noticed (if you can recall this).

Answer

Please note any medications that you are currently taking.

These can be prescription medications or over-the-counter treatments.



Answer

If you have any comments or wish to give further information about any of your answers, please do so here.

Answer

Figure 4.2 Self-reported health histories web questionnaire.

4.2.3 Results

Unfortunately only a small number of responses were received: there were 17 responses of which 14 were completed and one partially complete. Of the completed or partially-completed responses, the gender of respondents was 10 female, 5 male; the country of birth of the respondents was UK 8 respondents, USA 2 respondents, with 1 respondent each from Canada, Australia and Poland. Two respondents declined to give their country of birth, although the rest of their survey responses were fully completed. 13 respondents described their ethnicity as ‘white’, with one describing their ethnicity as ‘Asian’. One respondent declined to give their ethnicity.

Further results are summarised in Table 4.2, Table 4.3 and Table 4.4.

Age range of respondent (years)	**Number of respondents**
Not given	1 (abandoned the survey at this point)
<18	1 (respondent not allowed to complete due to age restriction)
18-25	2
26-40	5
41-50	6
51-60	2
> 60	0

Table 4.2 Age of respondents to self-reported health histories study

Event category	Mean number of events in survey	Median number of events in survey
Unprompted recollection	7.3	8
Prompted recollection	4.9	6
In both categories	1.4	2
Total recalled conditions	9.2	10
Events recorded in formal records	25.8	11

Table 4.3 Number of recalled events entered to self-reported health histories study

The prevalence of common conditions from the 15 valid responses are shown in Table 4.4.

Condition	Number of respondents recalling condition	Prevalence in respondents	Prevalence in general records
Autism	1	7 %	0.2 %
Depression	6	43 %	-
Chicken pox	8	57 %	-
Asthma	2	14 %	12 %
Eczema	1	7 %	5 %
Hay fever/allergic rhinitis	6	43 %	20 %
Sinusitis	6	43 %	15 %
Anxiety	2	14 %	-
Bone fracture	3	21 %	-
Hypertension	2	14 %	-
Bronchitis	1	7 %	16 %

Table 4.4 Prevalence of common conditions entered to self-reported health histories study

Due to the low number of responses to the survey, no statistical comparison of the number of recalled events versus number of events recorded in formal records was attempted, although it can be seen that the median number of recalled events is close to the median from formal records. The rate of occurrence of conditions in the survey responses compared to the conditions prevalence is shown in Table 4.4, however the numbers for any one condition in the response set is small. Figures for formal records have been taken from the composite data set described in Chapter 6, with prevalences for some conditions calculated using the codelists described in Chapter 7.

4.2.4 Conclusions

The response rate to the survey was low and so no detailed analysis of the responses was attempted. It was clear that the responses were insufficient for analysis of associations between conditions for the purpose of calculating risks of unreported conditions. For this work, it was necessary to use record sets derived from formal medical records.

However, given the closeness of the median number of events reported by individuals when compared to the median number of events in formal records, it appeared that there would be potential for using information acquired in this way should it be possible to increase the response rate. It would also be possible for individual records to be compared to sets of formal medical records for the purpose of calculating condition risk, although this would require individuals’ responses to be coded into the same coding system as the formal records sets.

4.3 Discussion

Privacy and data security was important in both projects. The EPI-CT project allowed identifiable records to leave hospitals only under strict conditions, with use of pseudonyms (EPI-CT) elsewhere. The web survey of individuals’ recollections acquired data anonymously. The minimum useful data set was acquired in each project. The quantity of data acquired was important: EPI-CT was a very large project with good statistical power but required a large investment in time, money and staff. The web survey of individuals’ recollections project had a low response but gave some qualitative information and, once set up, was low maintenance, running unsupervised on a web server, and low cost.

5 CONSOLIDATION OF ELECTRONIC HEALTH RECORDS DATA FROM MULTIPLE SOURCES

5.1 Introduction and background

An Electronic Health Record (EHR) has been defined as ‘a system specifically designed to support users by providing accessibility to complete and accurate data …’ [211]. The ASTM E1384 Standard Guide on Content and Structure of Electronic Health Records [212] gives a comprehensive list of data items that an EHR system should be able to record. These data items include, but are not limited to, patient demographic or identifying items as date of birth, gender, occupation and address; clinically relevant information such as blood pressure, weight, height and allergy alerts; and outcomes of consultations and investigations such as diagnoses, prescriptions and referrals.

Three data sets of de-identified primary care records were obtained from three independent sources. The work required to combine them into a single data set is described, including translation of coded clinical events from the three source data sets to a common coding system, selection of fields common across the source data sets and the mapping of individual data items from the source data sets into a single composite data set.

The source data sets included data from both UK and US systems. The advantage of this was that this increased the quantity of data available. Possible disadvantages were that the two countries may have different definitions for some conditions and different population profiles. The potential problem of differing condition definitions was addressed to some extent by mapping codes to a single coding system and then using less granular codes to group together closely-related conditions.

5.2 The source data sets

5.2.1 THIN

As noted in section 2.9, THIN holds data from UK general practice patients, coded in Read Codes version 2, containing longitudinal medical event histories and sourced primarily from practices using Epic/Cegedim systems. Following discussion, THIN were able to supply a set of data that they had divided into 'train' and 'test' sets. The data were supplied in standard system agnostic .csv files.

Each set contained seven tables of patient, event and related data, plus 11 lookup tables which give the meanings behind the codes used to store information in the data tables. Figure 5.1 shows the THIN data table schema. All files are supplied as simple text files fields determined by their position within the text files. Five data tables from the sets supplied by THIN were used:[213]

- Two patient data tables, each of which includes patient pseudo-ID, date of birth, date of death (note that dates are displaced from the conventional calendar and need to be adjusted), gender, marital status, family number as well as a number of other items that are not used.
- Two clinical event tables, each of which includes patient pseudo-ID (allowing linkage to the patient data table), event pseudo-ID, event date, date of event data entry, 'medcode' (local THIN code for the recorded event), as well as a number of other items that are not used.
- THIN stores codes for medical events as 7-byte Read Codes version 2, where the last two bytes allow for synonyms but only the first 5 bytes of the code are clinically significant.

- A look-up table which converts a code for marital status to an English-language description of the status as a character string

[redacted]

5.2.2 CPRD

CPRD holds data from UK general practice patients, coded in Read Codes version 2, containing longitudinal medical event histories and sourced primarily from practices using EMIS systems. The data were supplied in standard system agnostic .csv files. The CPRD data structure is similar to that of the THIN data. Figure 5.2 shows the CPRD data table schema. CPRD supply nine tables of patient, event and related data, plus 99 lookup tables which give the meanings behind the codes used to store information in the data tables. All files are supplied as simple text files with tab-separated variables. Five data tables from the set supplied by CPRD are used:

- The patient data table, which includes patient pseudo-ID, date of birth, date of death (note that dates are displaced from the conventional calendar and need to be adjusted), gender, marital status, family number as well as a number of other items that are not used.
- The clinical event table, which includes patient pseudo-ID (allowing linkage to the patient data table), event pseudo-ID, event date, date of event data entry, ‘medcode’ (local CPRD code for the recorded event), as well as a number of other items that are not used.
- The immunisations table, which records immunisations for each patient and includes patient pseudo-ID, date of immunisation, date of immunisation data entry, ‘medcode’ as well as a number of other items that are not used.
- A look-up table which maps CPRD medcodes to their equivalent Read Code (Read version 2). The mapping is 1:1. The Read Codes mapped to are 7-byte codes, where the last two bytes allow for synonyms but only the first 5 bytes are clinically significant.
- A look-up table which converts a code for marital status to an English-language description of the status as a character string

[redacted]

5.2.3 Practice Fusion

Practice Fusion holds data from US general practice patients, coded using ICD-9-CM, containing longitudinal medical event histories and sourced exclusively from practices using Practice Fusion's systems. The data were supplied in standard system agnostic .csv files.The most complex of the three source data sets is that from Practice Fusion. The data set schema is shown in Figure 4.3. Practice Fusion supplied two sets of data, each containing 17 tables of patient data, event data and related data. Each table was stored as a text file with comma-separated variables. Five of these tables were used from each data set. Additionally, a bespoke lookup table was used to map clinical event codes recorded in the Practice Fusion data from ICD-9-CM to Clinical Terms version 3. The process of building this mapping table is described later in this chapter. A second lookup table was created to map information on patient smoking to CTV3 codes.

The 'Transcript' table records information pertaining to a clinic visit – date of visit, and any measurements made on the patient (e.g. weight, height, blood pressure). Also at these visits information about diagnoses will be recorded, in a separate 'Diagnosis' table. Linking these two tables is a third table, 'TranscriptDiagnosis'. In the Diagnosis table, conditions have a start date and end date recorded. Where diagnosis start date (i.e. the year in which the event was first recorded) is missing, the Transcript date is used.

The Practice Fusion data set was supplied as a set of 17 tables, shown in Figure 4.3. These tables are:

- Patient: Contains the basic demographic information for each patient: Gender, Year of Birth, State, Patient's practice identifier, and a unique identifier for each patient.
- Diagnoses: Contains patient identifiers, the ICD9-CM code for each diagnosis together with the description associated with the ICD-9CM code, the start year and end year for the diagnosis, whether the condition is acute or chronic, and an identidier for the provider who recorded the diagnosis.
- Condition: A table which contains valid patient conditions, as codes and descriptions

- Transcript: Records details of patient visits to providers. Contains patient ID, year of the visit, records of patient vital signs and indicators – height, weight, BMI, blood pressure, respiratory rate, heart rate and temperature.
- Transcript Diagnosis: an associative table that lists the diagnoses per transcript.
- Smoking Status: lists the valid values for smoking status together with the description of the status.
- PatientSmokingStatus: Lists smoking statuses for each patient together with the year of the recorded status.
- Allergy: The list of allergies recorded for each patient.
- TranscriptAllergy: an associative table recording the list of allergies recorded per transcript.
- Immunization
- LabResult: Contains patient lab test results. Lists the patient ID and the provider ID, the transcript ID for the visit that ordered the lab test; the identifiers for the patient's medical practice and for the lab test facility and the year of the test.
- LabPanel: Contains the lab test panels reported in the lab test result. An associative table linking LabResults with LabObservations.
- LabObservation: Contains laboratory test results: the HL7 code for the lab test observation and its name; the coding system used by the laboratory; the value of the observation, its units and its reference range; a flag indicating whether the result is abnormal; the status of the test; the year of the test.
- Medication: the list of medications (including NDC code, name, strength and schedule) for each patient and ID of the diagnosis linked to the medication.
- TranscriptMedication: an associative table linking medications and transcripts.
- Prescription: The prescription records for each patient, including year of prescription, quantity, number of repeats and whether the patient can order a repeat prescription.

[redacted]

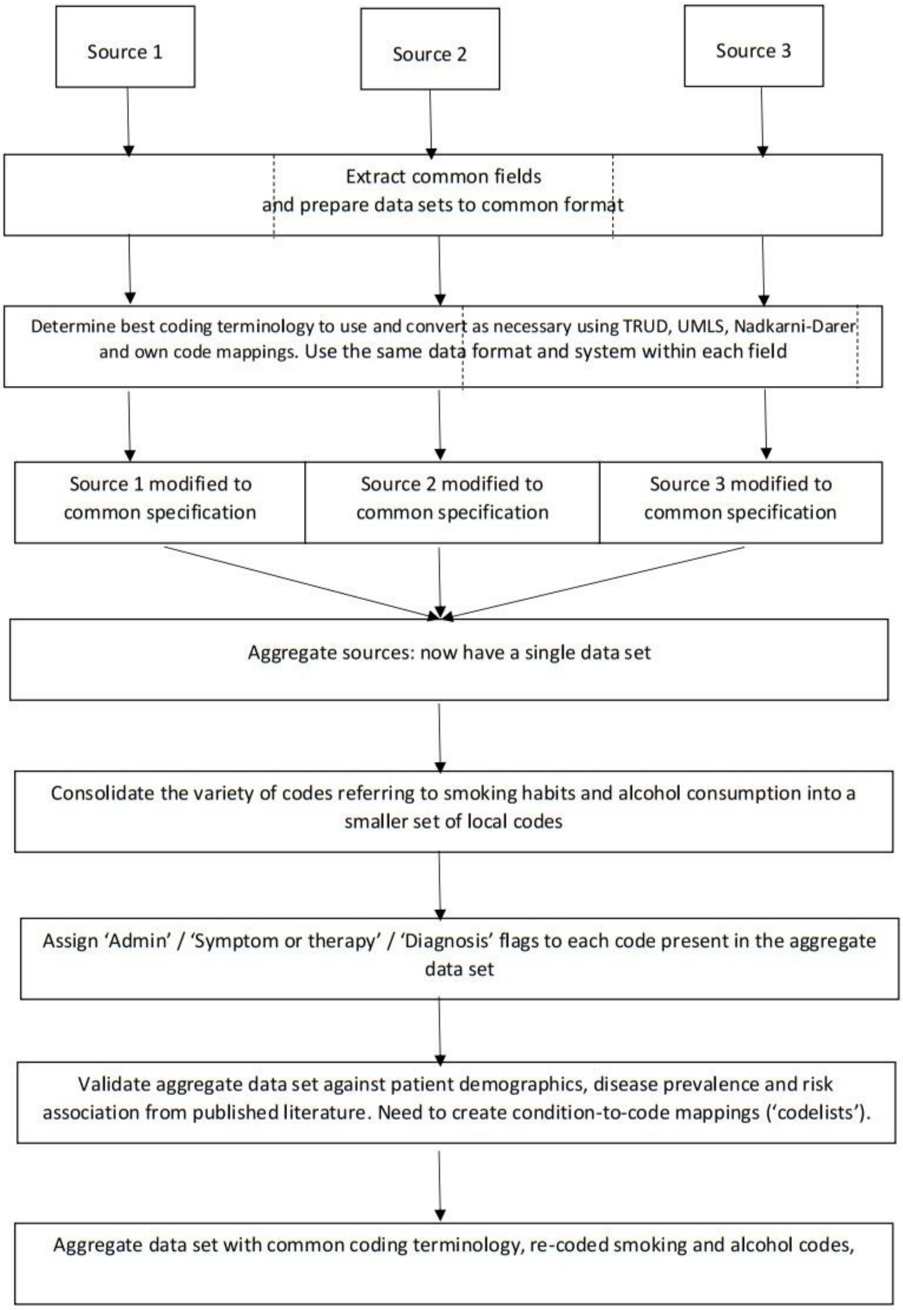
5.3 Merging of source data sets

The process for of combining data from the source data sets into the composite data set has several stages, illustrated in Figure 5.4. Part 1 of Figure 5.4 illustrates the process of creating an aggregated dataset whereas part 2 shows how aggregated data may be analysed. Each stage is of the aggregation process described in detail later in this chapter.

Additionally, for subsequent analysis it was advantageous to flag each code in the composite data set as being an ‘administrative’, ‘symptom or treatment’, or ‘diagnosis’ code. The methodology for automating the process of assigning such flags to CTV3 codes is described.

Once the composite data set has been produced, it must be validated (described in Chapter 6) and is then available for analysis (described in Chapter 7).

Part 1: Creation of aggregated data set


Source 1
Source 2
Source 3
Extract common fields
and prepare data sets to common format
Determine best coding terminology to use and convert as necessary using TRUD, UMLS, Nadkarni-Darer
and own code mappings. Use the same data format and system within each field
Source 1 modified to common specification
Source 2 modified to common specification
Source 3 modified to common specification
Aggregate sources: now have a single data set
Consolidate the variety of codes referring to smoking habits and alcohol consumption into a smaller set of local codes
Assign 'Admin' / 'Symptom or therapy' / 'Diagnosis' flags to each code present in the aggregate data set
Validate aggregate data set against patient demographics, disease prevalence and risk association from published literature. Need to create condition-to-code mappings ('codelists').
Aggregate data set with common coding terminology, re-coded smoking and alcohol codes,

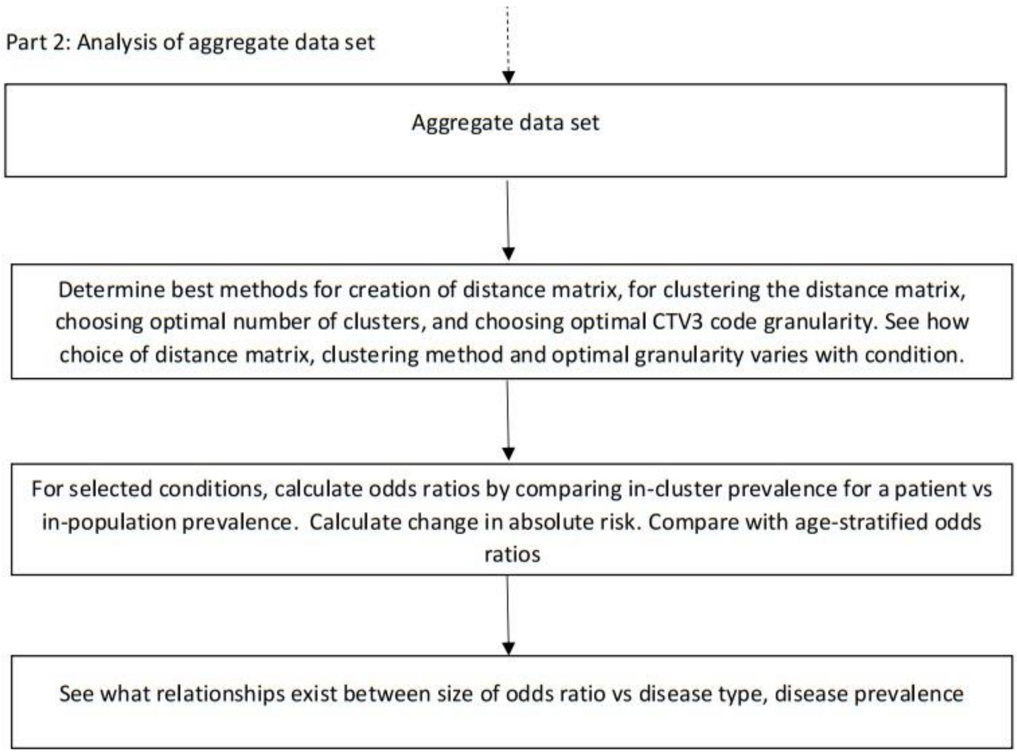

Part 2: Analysis of aggregate data set
Aggregate data set
Determine best methods for creation of distance matrix, for clustering the distance matrix, choosing optimal number of clusters, and choosing optimal CTV3 code granularity. See how choice of distance matrix, clustering method and optimal granularity varies with condition.
For selected conditions, calculate odds ratios by comparing in-cluster prevalence for a patient vs in-population prevalence. Calculate change in absolute risk. Compare with age-stratified odds ratios
See what relationships exist between size of odds ratio vs disease type, disease prevalence

Figure 5.4 Workflow for creation and analysis of composite data set

Data fields to be included in the composite data set are limited to those fields that are present in each of the three data sets to be merged or can be deduced from those or other fields. Inspection of the three source data sets reveals a number of fields that are common to all three data sets, and a number that are present in only one or two of the source data sets.

The set of common fields was inspected to ensure that each group of common fields were semantically interoperable. It was necessary to ensure data items in each field were converted to lowest common denominator. This was done by manual inspection of each field to determine which level of detail would ensure that data was captured at the broadest level of granularity across the source data sets. For example age and event dates in the composite data set were captured in years only: years for event occurrences (e.g. dates of birth) are available across all three data sets but in finer detail (for dates of birth, years and months) in only some of the data sets. The set of fields to be included in the composite data set is the intersection of the list of fields across all three data sets. Table 5.1 shows which fields are present in each data set, with a list of descriptions for each field that falls under that topic for each source data set.

Composite Data Set Field Name	Practice Fusion Field Name	CPRD Field Name	THIN Field Name
Patient ID	Pseudo-ID	Pseudo-ID	Pseudo-ID
Gender	M or F	Integer	integer
Date of birth	Year of birth	month and year of birth	year of birth (and month for children)
Date of death		date of death	date of death
Cause of death			cause of death
Address	State	GP practice region	urban or rural; ethnicity of ward; pollution in ward; whether a residential institute
marital status		current marital status	marital status
Family information		Family ID number	Family ID number
child health surveillance		whether registered with CHS, date of registration	
prescribing exemption		type of exemption	prescription exemption code
capitation supplement		type of supplement	
socio-economic status		[Included in CPRD data dictionary but not populated]	
First Registration with practice		date first registered with practice	registration date with practice
Current registration with practice		date current period of registration began	
registration status		registration status	registration status; whether they are a dispensing patient
registration gaps		count of days missing in registration status	
internal transfer outs		number of internal transfer out periods	
date transferred out		date the patient transferred out of the practice	date the patient transferred out of the practice
reason transferred out		reason patient transferred out of the practice	

other registration information, registration acceptance type			extended registration info
Practice ID	Practice identifier	encrypted identifier	
last collection date		date of last collection for the practice	
practice data quality		date at which the practice data is of research quality	
Diagnosis	ICD9 code	GPRD code	Read code
Diagnosis	description		
Diagnosis	start year		
Diagnosis	stop year		
Diagnosis	acute or chronic		
Diagnosis	provider recording info		
Diagnosis		Date of diagnosis event	Date of diagnosis event
Diagnosis		diagnosis type, e.g. diagnosis or symptom	
Diagnosis		ID of staff member entering info	ID of staff member entering info; source of record
Diagnosis		episode type	
Diagnosis			event end date
Allergy	type		
Allergy	start year		
Allergy	allergic reaction name		
Allergy	Severity		
Allergy	NDC code of medication taken for the allergy		
Allergy	name of medication taken for the allergy		
Allergy	provider recording info		
Conditions	Condition code		
Conditions	Condition name		
Conditions	Year		
Smoking status	status description		

Smoking status	NIST code		
Smoking status	year		
Immunizations	vaccine name		
Immunizations	year of administration	date of administration	
Immunizations	CVX code	compound administered	
Immunizations		individual components of compound administered	
Immunizations	immunization provider	staff ID, location of administration	
Immunizations		immunization type	
Immunizations		GPRD medcode	
Immunizations		stage of the immunization given	
Immunizations		immunisation status	
Immunizations		immunisation reason	
Immunizations		immunisation route	
Transcript	year		
Transcript	height		
Transcript	weight		
Transcript	BMI		
Transcript	blood pressure		
Transcript	respiratory rate		
Transcript	heart rate		
Transcript	temperature		
Transcript	physician specialty		
Transcript	Transcript provider		
Transcript	Diagnosis		
Consultation		Consultation date	
Consultation		type of consultation	
Consultation		consultation ID	
Consultation		Staff ID	
Consultation		Consultation duration	
Additional clinical details		dependent on entity type	
Referral		referral date	
Referral		referral category	
Referral		GPRD med code	
Referral		staff ID entering data	event recorded in practice; private or NHS

Referral		source of referral	
Referral		NHS classification of referral specialty	
Referral		FHSA classification of referral specialty	
Referral		Referral type	episode type
Referral		Attendance type	
Referral		Referral urgency	
Referral			Referral location
Referral			Cat of medical entry
Test		Test date	
Test		consultation type	
Test		GPRD medcode	
Test		staff ID	
Test		qualifier	
Test		normal range from, to, basis	
Test		various fields depending on test type	
Therapy	year of prescription	date of event	prescription date
Therapy	NDC code and medication name	GPRD product code	drug code
Therapy	User ID	Staff ID	staff ID
Therapy	strength	daily dose	dosage; calculated daily dosage
Therapy		BNF code	BNF chapter
Therapy	quantity	quantity, number of packs, pack size or type	quantity prescribed or number of packs; pack size
Therapy	start year, stop year, schedule	treatment days	duration
Therapy	number of refills	available for repeat prescription	acute or repeat prescription; sequence number for repeat prescriptions; max number of repeat issues
Therapy			private or NHS prescription
Therapy			source of drug
Therapy			event recorded in practice Y/N
Hospital event			Clinical specialty - code & description
Hospital event			Clinical sub-specialty - description

Table 5.1 Description of fields present in each data set

Table 5.2 shows, for each of the candidate fields for the composite data set whether that field is present in each of the source data sets or can be derived from other fields in the source data sets.

The fields for each source data set have been colour coded in Table 5.2 as follows:

🟪 Fields that match at a higher level of granularity

🟨 Fields that match at a lower level of granularity

⬜ Fields with no data in that particular source data set

Field number	Topic	Practice Fusion	example	CPRD	example	THIN	example
1	Patient ID	🟪	BC78C551	🟪		🟪	
2	Gender	🟪	F	🟪		🟪	
3	Date of birth	🟪	1981	🟪		🟪	
4	Date of death	🟧	NULL	🟪		🟪	
5	Cause of death					🟪	
6	Address	🟧	NY	🟧		🟧	
7	marital status			🟪		🟪	
8	family relationships			🟪		🟪	
9	child health surveillance			🟪			
10	prescribing exemption			🟪		🟪	
11	capitation supplement			🟪			
12	socio-economic status						
13	First Registration with practice			🟪			
14	Current registration with practice			🟪			
15	registration status			🟪		🟪	
16	registration gaps			🟪			

17	internal transfer outs			🟪			
18	date transferred out			🟪		🟪	
19	reason transferred out			🟪			
20	other registration information					🟪	
21	Practice ID	🟪	3E08ED81	🟪			
22	last collection date			🟪			
23	practice data quality			🟪			
24	Clinical event code	🟪		🟪		🟪	
25	Clinical event description	🟪		🟪		🟪	
26	Clinical event start date	🟪	2011	🟪		🟪	
27	Clinical event end date	🟪	NULL			🟪	
28	Acute or chronic	🟪	Acute				
29	Diagnoser provider information	🟪					
31	Diagnosis	🟪	272.2; 402.1; 715.16	🟪		🟪	
32	Diagnosis episode type			🟪			

33	Allergy type	🟪	Medication	🟨		🟪	
34	Allergy start date	🟪	2011	🟨		🟪	
35	Allergy name	🟪	Tongue swelling	🟨		🟪	
36	Allergy severity	🟪	Severe	🟨		🟨	
37	Allergy medication code	🟪	247224300	🟨		🟨	
38	Allergy medication name	🟪	Trilipix (fenofibric acid) oral delayed release capsule	🟨		🟨	
39	Allergy diagnoses provider info	🟪	F7998EB6	🟨		🟨	
40	Smoking status description	🟪	0 cigarettes per day (non-smoker or less than 100 in lifetime)	🟨		🟨	
41	Smoking status code	🟪	5ABBAB35	🟪		🟨	
42	Smoking status date	🟪	2010	🟪		🟨	
43	Immunization name	🟪		🟪		🟪	
44	Immunization date	🟪	2008 (Hepatitis B vaccine, adolescent (2 dose schedule), for intramuscular	🟪		🟪	

		🟪	use)	🟪		🟪	
45	Immunization code	🟪	43	🟪		🟪	
46	Immunization components			🟪			
47	Immunization provider	🟪		🟪		🟪	
48	Immunization type			🟪			
49	Immunization stage			🟪			
50	Immunization status			🟪			
51	Immunization reason			🟪			
52	Immunization route			🟪			
53	Physical characteristics date	🟪					
54	Physical characteristics height	🟪	63				
55	Physical characteristics weight	🟪	120				
56	Physical characteristics BMI	🟪	21.255				
57	Physical characteristics blood	🟪	120/80				

	pressure	🟪					
58	Physical characteristics respiratory rate	🟪	18				
59	Physical characteristics heart rate	🟪	NULL				
60	Physical characteristics temperature	🟪	95.2				
61	Physical characteristics physician speciality	🟪	Internal Medicine				
62	Physical characteristics provider	🟪					
63	Consultation date	🟪	2011	🟪			
64	Consultation type			🟪			
65	Consultation - staff ID		A75DB583	🟪			
66	Consultation duration			🟪			

67	Referral date			🟪			
68	Referral category			🟪		🟪	
69	Referral code			🟪			
70	Referral - staff ID			🟪			
71	Referral source			🟪			
72	Referral - classification of referral specialty			🟪			
73	Referral type			🟪			
74	Referral attendance type			🟪			
75	Referral urgency			🟪			
76	Test date			🟪			
77	Test type			🟪			
78	Test code			🟪			
79	Test - staff ID			🟪			
80	Test qualifier			🟪			
81	Test - normal range			🟪			
82	Test other			🟪			
83	Therapy date	🟪	2011	🟪		🟪	
84	Therapy - drug or	🟪	378710177 -	🟪		🟪	

	other product code	🟪	Fenofibrate oral tablet	🟪		🟪	
85	Therapy - staff ID	🟪	1E961D5D	🟪		🟪	
86	Therapy - dose	🟪	160 mg	🟪		🟪	
87	Therapy - BNF code			🟪		🟪	
88	Therapy quantity	🟪	14	🟪			
89	Therapy length	🟪	NULL	🟪		🟪	
90	Therapy repeat	🟪	0	🟪			
91	Therapy private or NHS prescription					🟪	
92	Therapy drug source					🟪	
93	Therapy event recorded in practice					🟪	
94	Hospital event - specialty code and description					🟪	
95	Hospital event - subspecialty description					🟪	

Table 5.2 Fields present in each of the source data sets

Combining all three data sets into a single data set, the following common fields are retained from the source data sets:

Patient ID [P]
Gender [P]
Year of birth [P]
Year of death [P]
Practice ID [P]
Clinical event code [M]
Clinical event start year [M]
Acute or chronic [M]
Allergy type [M]
Allergy start year [M]
Allergy name [M]
Smoking status code [P]
Smoking status date [P]
Immunisation code [M]
Immunisation date [M]
Therapy date [M]
Therapy code [M]
Therapy dose [M]
Therapy quantity [M]
Therapy length [M]

[P] patient-level fields – one per patient.
[M] event-level fields – many per patient.

5.4 Convergence of event codes onto a single coding system

5.4.1 Coding system convergence

There exist a number of different systems into which clinical events can be coded. for example the International Classification of Diseases (ICD) widely used in the US which in its latest mature iteration is version 10 but many healthcare enterprises are using earlier versions; SNOMED CT, the largest coding system in terms of number of concepts; Read Codes, the most common system in UK primary care, which in its latest iteration is Clinical Terms Version 3 (CTV3) but is most commonly used in version 2. It should be noted that the UK Department of Health has mandated the use of SNOMED CT from April 2020. When combining or comparing data sets from different sources using different coding systems, it is necessary to map clinical event codes to a single common coding system, which may be a system used by one or more of the source datasets or may be a new coding system. Bonney et al [214] write: "Mapping data elements in EHRs to a reference classification and/or terminology system not only facilitate reuse of primary care data for multiple purposes, but they also support data analysis, health information exchange and interoperability, and data comparison across the continuum of different healthcare providers [and] improves the quality of the research output derived from EHRs."

In this section, it is demonstrated how events recorded in one clinical coding system may be mapped to codes in a different system where no direct mapping between the two systems exists but where mapping is possible via a third coding system.

The work described in this section has been submitted as a paper to Methods of Information in Medicine and is under review at the time of writing. The paper can be found in Appendix 1.

Medical records since their inception have contained free text, with medical conditions and symptoms described by natural language terms that may be imprecise or ambiguous. Recent decades have seen a steady increase in the uptake of electronic health records (EMR) systems [1] [2]. There are now a large number and variety of terminologies used to code events recorded in these EMR systems, with it being estimated that there are over 100 terminologies currently in use [3]. For aggregation of data or analysis over time, a controlled, pre-defined vocabulary is required, with codes representing concepts that allow for descriptive synonyms [4]. A number of coding systems have been created over recent decades, including the International Classification of Diseases [5]; the Read Codes [6], the most common system in UK primary care, which in its latest iteration is Clinical Terms Version 3 (CTV3) but is most commonly used in version 2; and SNOMED CT [7], the largest coding system in terms of number of concepts. Modern electronic health records systems may use any of these existing systems, with different countries or regions favouring one system over another: in 2001, de Lusignan et al [8], in a survey of systems in use in Europe, found that the Read Codes were the most common system in use in primary care in the UK, ICD-10 the most common in primary care in Austria and Germany, and ICPC in a further 10 European countries. ICD-9, until its recent supersession by ICD-10, has been the dominant system in primary care in the USA.

There are several reasons why it may be necessary to move from one coding system to another: government mandate; the desire to use an up-to-date coding system; compatibility with other data repositories, e.g. in a newly-shared EHR system; combining data sets from disparate sources for research or audit purposes. Code mapping is an approach to enable codes from one system to be translated to their semantically equivalent codes in another system, a process that has been defined as "the process of associating concepts or terms from one coding system to concepts or terms in another coding system and defining their equivalence in accordance with a documented rationale and a given purpose" [9]. In order to combine data from diverse datasets coded using different coding systems it is necessary to converge the data sets onto a common coding system. At a minimum, translation of data items recorded in the coding system used in one of the source data sets to another coding system is required. However, the opportunity exists for all event codes from multiple sources to be mapped to a third coding system if that system has advantages over either of the coding systems in use in the existing data sets. When combining or comparing data sets from different sources using different coding systems, it is necessary to map clinical event codes to a single common coding system, which may be a system used by one or more of the source datasets or may be a new coding system. Bonney et al [10] write: "Mapping data elements in EHRs to a reference classification and/or terminology system not only facilitate reuse of primary care data for multiple purposes, but they also support data analysis, health information exchange and interoperability, and data comparison across the continuum of different healthcare providers [and] improves the quality of the research output derived from EHRs."

An issue which can occur when combining datasets is that of semantic interoperability, in particular equivalence in the coding of clinical concepts. It is relatively straightforward to combine demographic information between systems since, for example, "there is general agreement as to what 'age' means in relation to a patient" [11] and there is similar agreement for names of individuals and dates, but it is less straightforward to map clinical concepts or their coded representation between different terminologies. One long-standing method is to match the text description of concepts (e.g. [12] [13] [14]; [15] [16]). The majority of work in automatic mapping has focused on the lexical approach, using techniques similar to those used for automatic mapping from free text clinical notes to concept codes (for example [17] and [18]). However, Fung et al [19] found that 'Semantic mapping performed better than lexical mapping'. Cimino and Barnett [20] proposed a method of semantic mapping by which each concept in a terminology was characterised by a set of properties, with concepts being mapped across terminologies according to the closest similarity in properties. This method requires each concept to be characterised manually in a process described as 'tedious [but] not complicated'. A similar approach was proposed by Rocha et al [21]. Mappings exist between some of the major coding systems in current use, particularly between older and newer versions of coding systems, e.g. ICD-9-CM and ICD-10-CM; Read Codes Version 2 and CTV3, provided by several organisations and individuals, often those responsible for the maintenance of the coding systems. In the UK, the Department of Health Technology Reference-data Update Distribution service (TRUD) [22] provides mappings between a number of coding systems, in particular those systems in common use in the UK: SNOMED CT, Read 2, CTV3. In the US, mappings are provided between systems more common in that country by the Centers for Medicare and Medicaid Services (CMS), International Health Terminology Standards Development Organisation (IHTSDO), Unified Medical Language System (UMLS), National Library of Medicine (NLM). Brouch [23] gives an introduction to the mapping process and contains a glossary of relevant terms. Nandigam and Topaz [24], describing their work in mapping SNOMED CT to ICD10-CM, note that the SNOMED CT to ICD10-CM mappings from NLM "may need to be modified on the basis of the clinical specialty and patient population and further validated." Previous work on creating mapping tables between coding systems has been primarily by human experts comparing text descriptions of codes in different coding systems, e.g. [25]; with the assistance of a text search tool, e.g. [26], [20]; or by automated text matching [21]. One problem described by Nadkarni & Darer [26] was that of missing mappings: in their work in investigating the completeness of mapping a data set from ICD-9-CM to SNOMED CT they found that 784 (of 2199; 35.8%) ICD-9-CM codes in their data set had no map to an equivalent SNOMED CT code in the UMLS ICD-9-CM to SNOMED CT cross-map, requiring them to create these mappings by hand.

5.4.2 Choice of target coding system

UK data consisted of with clinical events coded in Read Codes V2, and US data was composed of clinical events coded in ICD-9-CM. Several coding systems were considered as candidates for the target common coding system. These coding systems were:

- Read Codes Version 2 (Read 2). The standard clinical terminology in use in UK general practice, introduced in 1990. The base Read Codes are 5 bytes, with an optional 2 byte extension to allow for synonyms. The UK dataset used in this work had clinical events coded using Read Codes version 2.

- Read Clinical Terms Version 3 (CTV3). Introduced in the late 1990s, with an increased number of codes compared to Read v2 and improvements to the code ontology and terminology. However, the majority of UK general practices continue to use Read v2.

- SNOMED CT. Created in 2001, a merger of CTV3 and SNOMED RT. Use of SNOMED CT rather than Read Codes (whether version 2 or 3) is mandated for UK General Practices by April 2018 and for UK NHS secondary care by April 2020.

- International Classification of Diseases 9th edition, Clinical Modification (ICD-9-CM) or 10th edition (ICD-10-CM). ICD codes are maintained by the World Health Organisation. Version 9 was introduced in 1978, with ICD-10 introduced from 1994 (ICD-9 is still maintained annually). The US dataset used in this work had clinical events coded using ICD-9-CM.

CTV3 was chosen as the target common coding system for the following reasons:

(i) This is the most recent development of the Read Codes, which allows for parent-child hierarchies to be represented by a separate table rather than by the structure of the codes themselves; Read Codes are optimised for secondary use [215];

(ii) A simple, complete and clinically validated mapping exists from Read 2 to CTV3. The mapping is freely available under licence from NHS TRUD;

(iii) No mappings exist from ICD-9-CM to either Read 2 or CTV3 and so there is no reduced effort required in mapping ICD-9-CM to Read 2 compared to CTV3;

(iv) CTV3 is closely aligned to SNOMED CT which is mandated for use in UK NHS primary care by April 2018 and in UK NHS secondary care by April 2020;

(v) CTV3 is a simpler coding system than SNOMED CT, with a single code per condition, which was an advantage for subsequent analysis of the merged data set;

(vi) Both Read 2 and ICD-9-CM include clinically obsolete terms (e.g. ICD-9-CM 318.0 'imbecile') or relationships (e.g. Read Codes version 2 code E220. 'Homosexuality' is categorised under code E22.. 'Sexual deviations or disorders' (whereas in CTV3 'Homosexuality' is categorised under code X766p 'Sexual orientation'), a situation which is addressed in CTV3;

(vii) CTV3 is well structured, with an existing, clinically-validated, table of parent-child relationships within the coding hierarchy;

(viii)There exists a simple mapping available from TRUD for mapping from Read 2 to CTV3 suitable for the UK data sets.

(ix) Mappings exist between some of the major coding systems in current use, provided by several organisations and individuals. In the UK, the Department of Health Technology Reference-data Update Distribution service (TRUD) provides mappings between a number of coding systems, in particular those systems in common use in the UK: SNOMED CT, Read 2, CTV3. In the US, mappings are provided between systems more common in that country by the Centers for Medicare and Medicaid Services (CMS), International Health Terminology Standards Development Organisation (IHTSDO), Unified Medical Language System (UMLS), National Library of

Medicine (NLM). Additionally, some mappings were found for ICD-9-CM to SNOMED CT in the work of Nadkarni and Darer [189], who had created some mappings for ICD-9-CM codes present in their data but not mapped in the NLM table.

(x) Existing mappings found are listed in Table 5.3. Note that mappings may not be bi-directional.

	Target coding				
Source coding	ICD-9-CM	ICD-10-CM	SNOMED CT	Read 2	CTV3
ICD-9-CM	-	CMS	NLM, ND		
ICD-10-CM	CMS	-			
SNOMED CT	TRUD, IHTSDO	UMLS	-	TRUD	TRUD
Read 2		TRUD	TRUD	-	TRUD
CTV3		TRUD	TRUD	TRUD	-

Table 5.3 Available inter-system code mappings and their sources

Sources for the available code mappings:

TRUD - Department of Health Technology Reference-data Update Distribution service

- CMS - Centers for Medicare and Medicaid Services
- IHTSDO - International Health Terminology Standards Development Organisation
- UMLS - Unified Medical Language System
- NLM – National Library of Medicine
- ND – Nadkarni & Darer

NLM provides mapping tables which map 6285 ICD-9-CM codes 1:1 to SNOMED CT codes and 3508 ICD-9-CM codes in 1:many maps to SNOMED CT codes, a total of 9793 unique ICD-9-CM codes. The data set which is required to be mapped to CTV3 contains 4342 unique ICD-9-CM codes, 44.3 % of the codes in the NLM mapping table.

NHS TRUD provides a SNOMED CT to CTV3 mapping table, which comprises of 747,717 unique SNOMED CT codes. The 4342 ICD-9-CM codes in the data set map to 2640 SNOMED CT codes, a mere 0.35 % of the codes in the NHS TRUD mapping table.

Both the Read Codes version 2 to CTV3 and the ICD-9-CM to CTV3 mapping processes were implemented using the Konstanz Information Miner (KNIME) [216]. KNIME is an open source data analytics and exploration modular environment providing a number of data manipulation and analysis modules. KNIME has several advantages over traditional programming which suggested it as a suitable tool for this work. These advantages include:

Rapid programming. Many of the required tasks, such as file reading and writing, SQL-type joining of tables, selection by field content, are available in pre-defined nodes that are quick to set up;

- Reduction in the programming required and thus reduction in the potential for programming error;
- It is possible to inspect the data after each step, helping find where errors have been made in the programming;
- Typographic errors are reduced by presenting a pull-down list of valid variables at each step;
- The program is open-source and well supported by the development team and community of users;
- Workflows created can be saved and are simple to share;
- Workflows created are easy to display and are a useful tool for describing the manipulation performed on data sets.

A simple KNIME workflow was written to enable the Read Codes version 2 mapping. This workflow is shown in Figure 1. Of the 14239 unique Read Codes Version 2 codes from the UK-sourced data, 100 % successfully mapped to a CTV3 code using tables from the UK Technology Reference Data Update Distribution (TRUD), mapping to 13947 unique CTV3 codes.

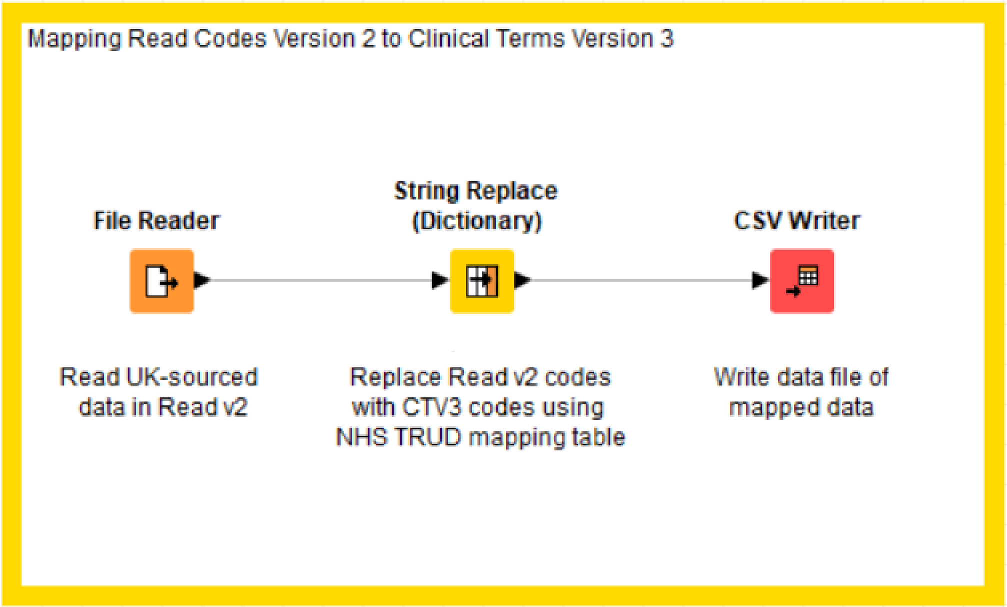

Mapping Read Codes Version 2 to Clinical Terms Version 3
File Reader
String Replace (Dictionary)
CSV Writer
Read UK-sourced data in Read v2
Replace Read v2 codes with CTV3 codes using NHS TRUD mapping table
Write data file of mapped data

Figure 5.5 Read Codes Version 2 to CTV3 mapping process using NHS TRUD mapping.

This simple workflow has three steps: (i) the source data file containing Read Codes Version 2 codes is read; (ii) using a simple look-up table derived from NHS TRUD mappings, Read 2 codes are paired with their equivalent CTV3 codes; (iii) the mapped file is saved.

No existing mapping was found from ICD-9-CM to CTV3 (see Table 5.3). However, mappings were available from ICD-9-CM to SNOMED CT from the US National Library of Medicine [217] ("NLM") and from SNOMED CT to CTV3 from the NHS Digital Technology Reference data Update Distribution [218] ("TRUD").

For mapping the US data to CTV3, a two stage process was proposed:

1. Map ICD-9-CM to SNOMED CT using NLM look up table;
2. Map SNOMED CT to Read CTV3 using TRUD look up table.

NLM supplies ICD-9-CM to SNOMED CT mapping tables in two tables: a 1:1 mapping table, where a single ICD-9-CM code maps to a single SNOMED CT code; and a 1:many mapping table, where a single ICD-9-CM code maps to many SNOMED CT codes, reflecting the increased nuance of description allowed by the larger number of SNOMED CT codes when compared to the less expressive ICD-9-CM codes. For example, the single ICD9-CM code 578.1 'Blood in stool' has 8 SNOMED CT codes: 405729008 'Hematochezia (finding)'; 2901004 'Melena (disorder)'; 59614000 'Occult blood in stools (finding)'; 300392005 'Stool flecked with blood (finding)'; 272045003 'Complaining of melena (finding)'; 269900004 'Feces: fresh blood present (finding)'; 249624003 'Blood in feces symptom (finding)'; 275782008 'Melena on examination of feces (disorder)'. Some ICD-9-CM codes have a very large number of matching SNOMED CT codes: ICD-9-CM code 995.29 'Unspecified adverse effect of other drug, medicinal and biological substance' maps to 1636 unique SNOMED CT codes, each specifying the particular adverse reaction, e.g. SNOMED CT code 293199005 'Glymidine adverse reaction (disorder)'. Given the potentially large number of possible matches in the 1:many table, a decision was made to map codes automatically using only the 1:1 mapping table, manually mapping any codes that were not mapped by the 1:1 mapping table.

Once the necessary mapping tables had been obtained from NLM and from TRUD, the ICD-9-CM codes present in the data set were mapped to SNOMED CT. These mapped codes were then further mapped from SNOMED CT to CTV3 and a single, direct ICD-9-CM to CTV3 mapping table was generated. The mapping process was then checked for completeness and exactness. Codes that failed the mapping process or were judged to have been incorrectly mapped were mapped manually. **Figure 5.6** illustrates the proposed mapping. Also included is a route for mapping directly from ICD-9-CM to CTV3 for codes which fail to map at either of the indirect mapping stages and which have to be mapped manually.

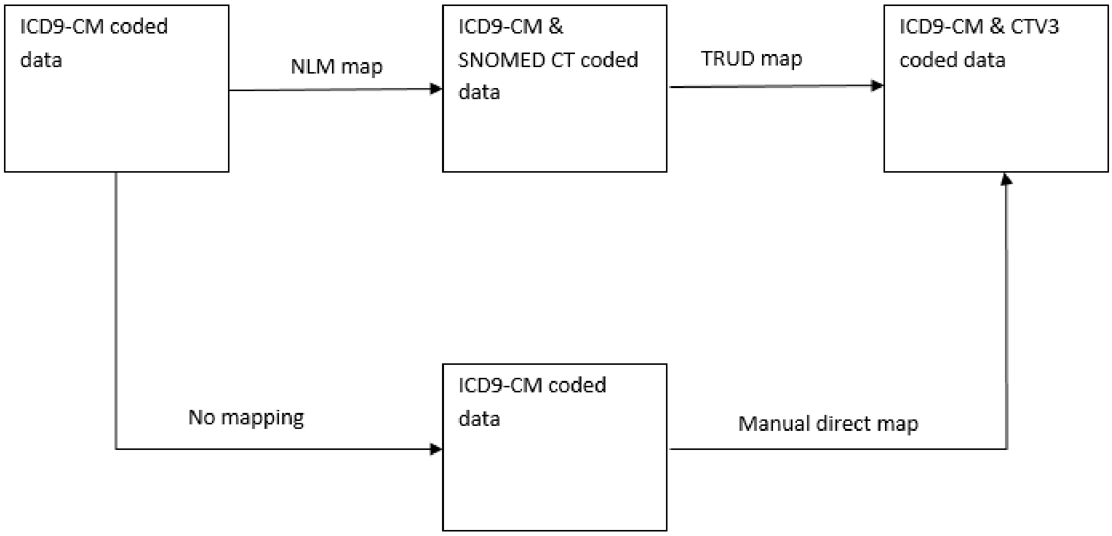


Figure 5.6 ICD-9-CM to CTV3 mapping process via SNOMED CT

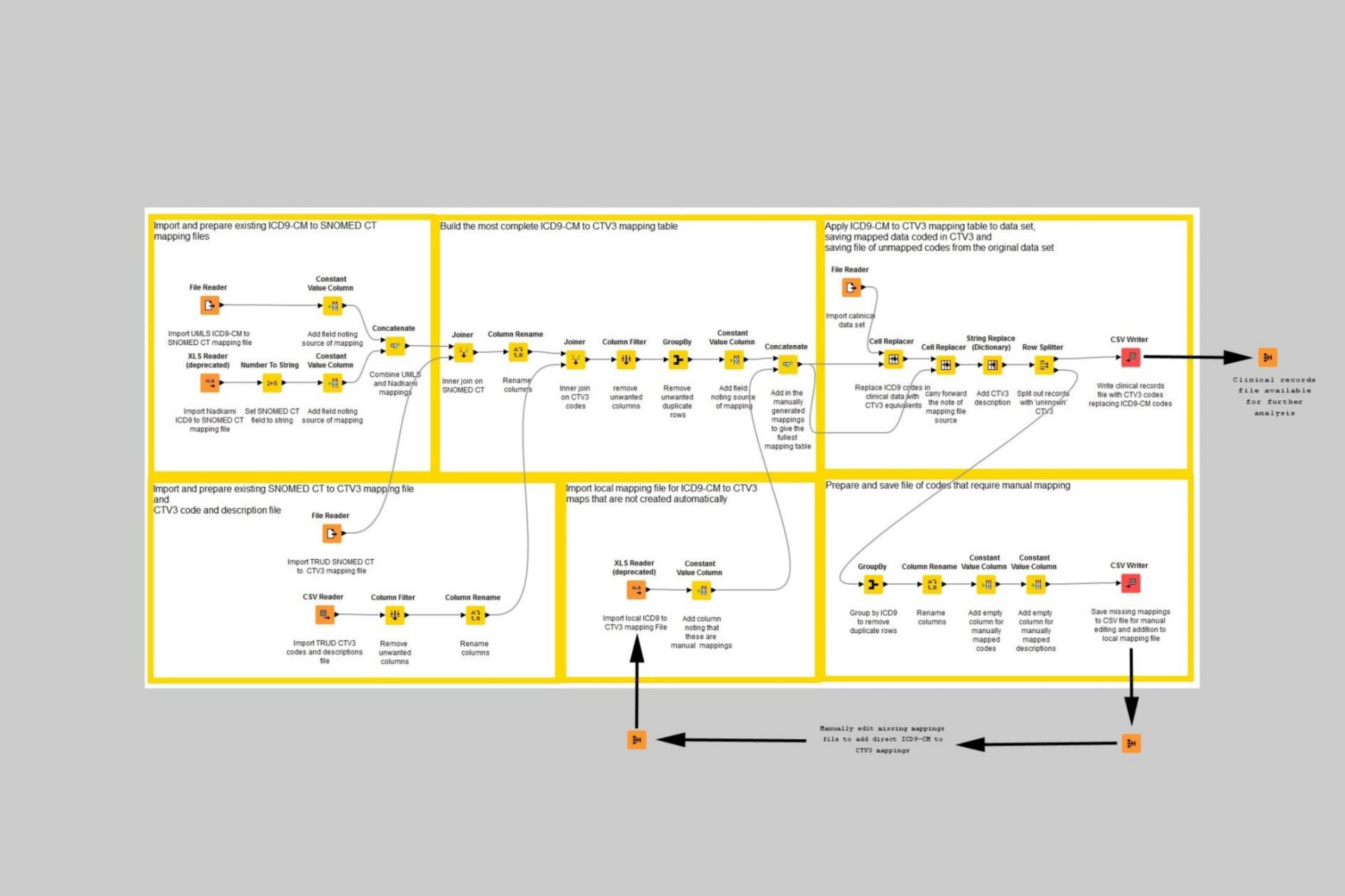

Import and prepare existing ICD9-CM to SNOMED CT mapping files
File Reader
Import UMLS ICD9-CM to SNOMED CT mapping file
Constant Value Column
Add field noting source of mapping
XLS Reader (deprecated)
Import Nadkarni ICD9 to SNOMED CT mapping file
Number To String
Set SNOMED CT field to string
Constant Value Column
Add field noting source of mapping
Concatenate
Combine UMLS and Nadkarni mappings
Build the most complete ICD9-CM to CTV3 mapping table
Joiner
Inner join on SNOMED CT
Column Rename
Rename columns
Joiner
Inner join on CTV3 codes
Column Filter
remove unwanted columns
GroupBy
Remove unwanted duplicate rows
Constant Value Column
Add field noting source of mapping
Concatenate
Add in the manually generated mappings to give the fullest mapping table
Apply ICD9-CM to CTV3 mapping table to data set,
saving mapped data coded in CTV3 and
saving file of unmapped codes from the original data set
File Reader
Import calinical data set
Cell Replacer
Replace ICD9 codes in clinical data with CTV3 equivalents
Cell Replacer
carry forward the note of mapping file source
String Replace (Dictionary)
Add CTV3 description
Row Splitter
Split out records with 'unknown' CTV3
CSV Writer
Write clinical records file with CTV3 codes replacing ICD9-CM codes
Clinical records file available for further analysis
Import and prepare existing SNOMED CT to CTV3 mapping file
and
CTV3 code and description file
File Reader
Import TRUD SNOMED CT to CTV3 mapping file
CSV Reader
Import TRUD CTV3 codes and descriptions file
Column Filter
Remove unwanted columns
Column Rename
Rename columns
Import local mapping file for ICD9-CM to CTV3 maps that are not created automatically
XLS Reader (deprecated)
Import local ICD9 to CTV3 mapping File
Constant Value Column
Add column noting that these are manual mappings
Prepare and save file of codes that require manual mapping
GroupBy
Group by ICD9 to remove duplicate rows
Column Rename
Rename columns
Constant Value Column
Add empty column for manually mapped codes
Constant Value Column
Add empty column for manually mapped descriptions
CSV Writer
Save missing mappings to CSV file for manual editing and addition to local mapping file
Manually edit missing mappings file to add direct ICD9-CM to CTV3 mappings

Figure 5.7 KNIME workflow implementing the ICD-9-CM to CTV3 mapping process

A KNIME workflow was written which combined the existing ICD-9-CM to SNOMED CT and SNOMED CT to ICD-9-CM mappings to create a single ICD-9-CM to CTV3 mapping table. Six steps are used within the KNIME workflow to produce the ICD-9-CM to CTV3 mapping table. These steps are combined into the single KNIME workflow shown in Figure 5.7.

. The workflow is broken down into six discrete sections:

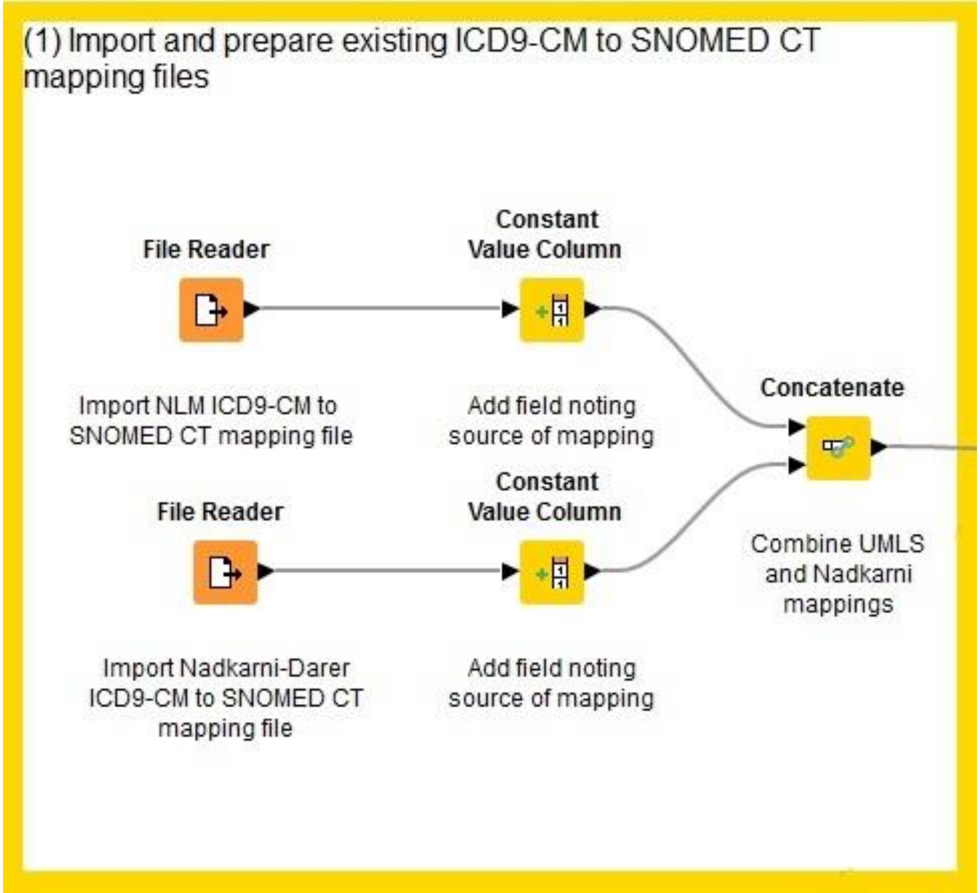

(1) Import and prepare existing ICD9-CM to SNOMED CT mapping files
File Reader
Constant Value Column
Import NLM ICD9-CM to SNOMED CT mapping file
Add field noting source of mapping
Concatenate
File Reader
Constant Value Column
Combine UMLS and Nadkarni mappings
Import Nadkarni-Darer ICD9-CM to SNOMED CT mapping file
Add field noting source of mapping

Figure 5.8 Import and prepare ICD-9-CM to SNOMED CT mapping files

The two external ICD-9-CM to SNOMED CT mapping files, from NLM and from the work of Nadkarni and Darer, are imported and combined into a single mapping table. An entry is made against each ICD-9-CM code to note the source of its mapping to SNOMED CT. **Figure 5.8** shows the subsection from the complete workflow that performs these tasks.

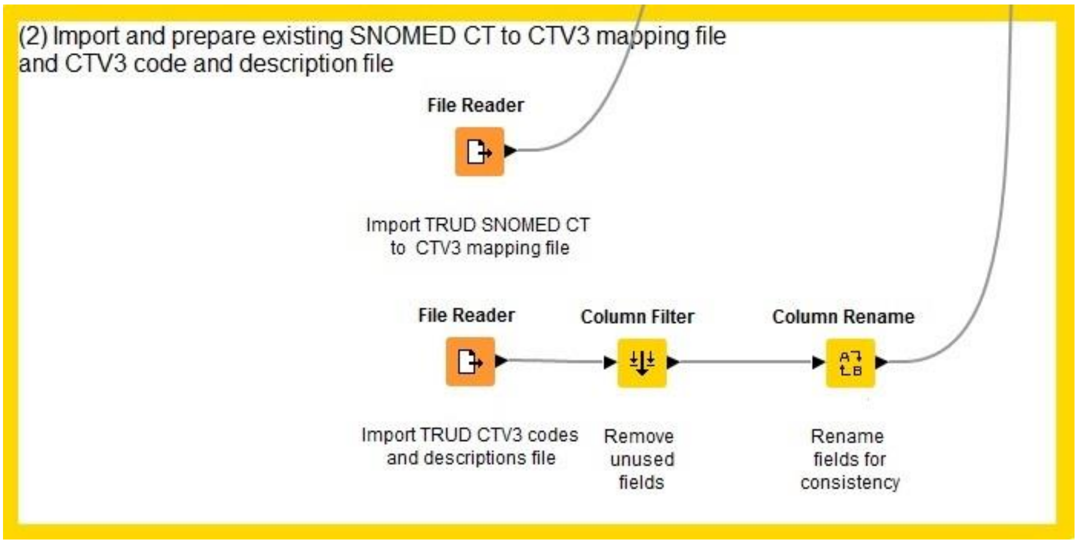


Figure 5.9 Import and preparation of the SNOMED CT to CTV3 mapping file and associated CTV3 codes descriptions file

The SNOMED CT to CTV3 mapping file obtained from TRUD is imported. The table of CTV3 codes and corresponding descriptions, also from TRUD, is imported. Unused fields in the CTV3 descriptions table are removed and the remaining fields renamed to ensure consistency across the complete workflow. This section of the workflow is shown in **Figure 5.9**.

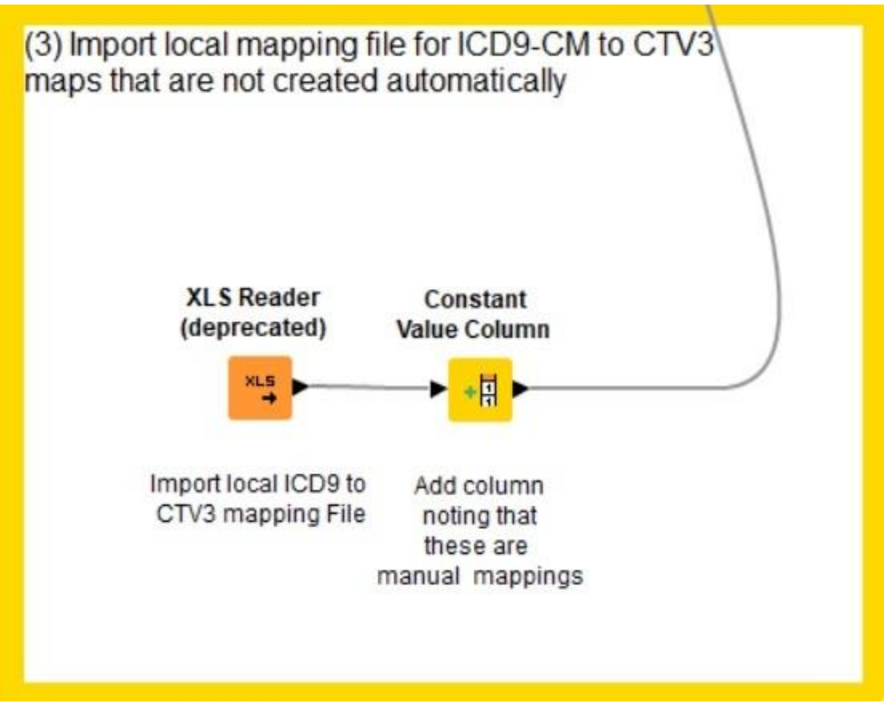


Figure 5.10 Import of local ICD-9-CM to CTV3 mapping file

For those ICD-9-CM codes that failed to map to SNOMED CT codes, or for which improved mappings have been found, a manually created local mappings file is imported. A field noting the source of these mappings is added. This stage of the workflow is shown in **Figure 5.10**.

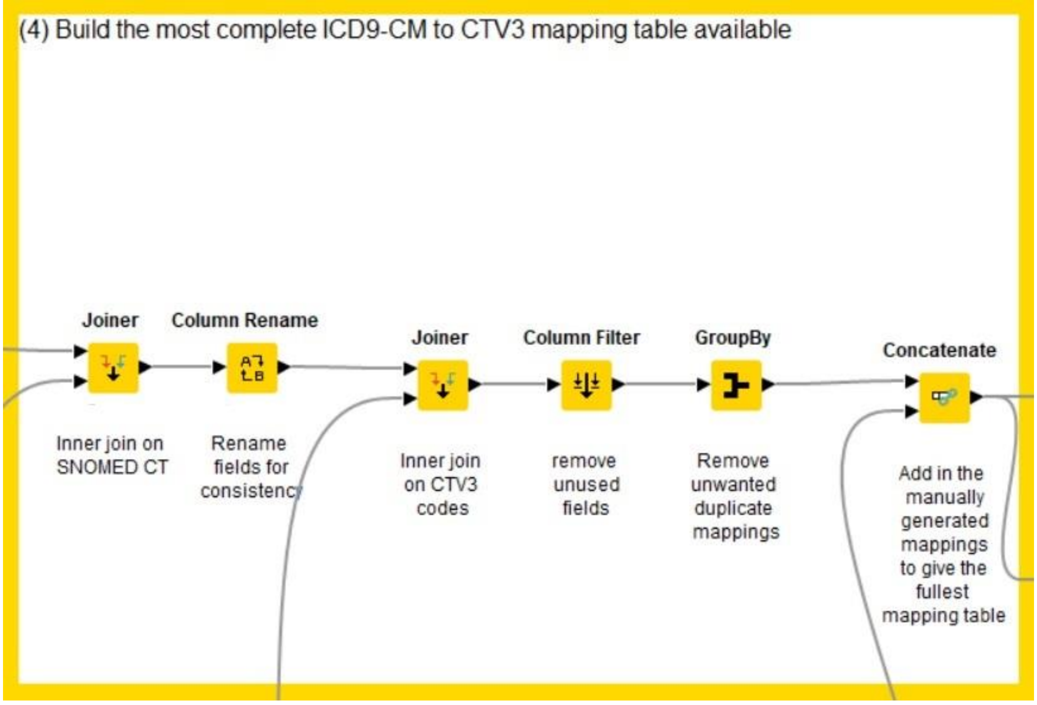
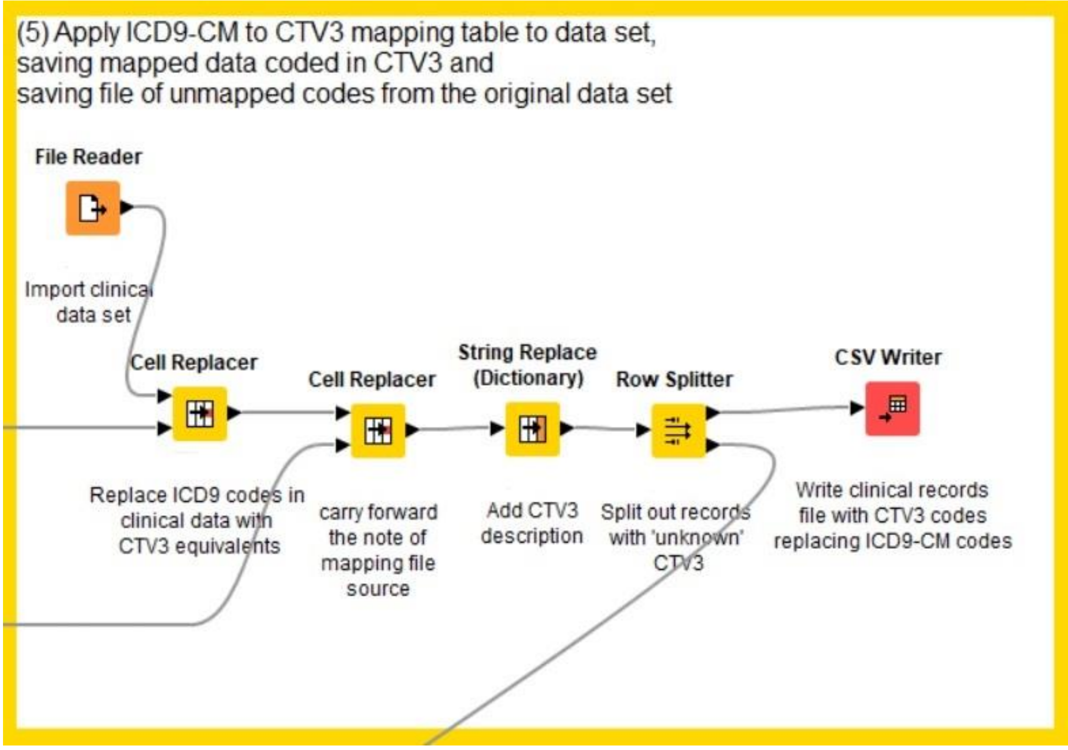

(4) Build the most complete ICD9-CM to CTV3 mapping table available
Joiner
Column Rename
Joiner
Column Filter
GroupBy
Concatenate
Inner join on SNOMED CT
Rename fields for consistency
Inner join on CTV3 codes
remove unused fields
Remove unwanted duplicate mappings
Add in the manually generated mappings to give the fullest mapping table

Figure 5.11 Building the complete ICD-9-CM to CTV3 mapping table

In this stage, the composite (NLM and Nadkarni-Darer) ICD-9-CM to SNOMED CT code mapping table and the SNOMED CT to CTV3 mapping table are joined by use of an inner join on SNOMED CT codes, i.e. all maps that have SNOMED CT values present in both the ICD-9-CM to SNOMED CT table and the SNOMED CT to CTV3 table are selected to produce a table of ICD-9-CM to SNOMED CT to CTV3 mappings. Examples of the mappings produced are shown in Table 2. The intermediate SNOMED CT fields are now dropped from the table, and duplicate mappings are removed by forcing unique values for ICD-9-CM codes. This automatically generated ICD-9-CM to CTV3 mapping table is then augmented by combining it with the manually-created local table to give the fullest mapping table. **Figure 5.11** shows the section of the workflow that builds the most complete ICD-9-CM to CTV3 mapping table.

ICD-9-CM code	ICD-9-CM description	SNOMED CT code	SNOMED CT description	CTV3 code	CTV3 description
427.31	Atrial fibrillation	49436004	Atrial fibrillation (disorder)	G5730	Atrial fibrillation
599.0	Urinary tract infection, site not specified	68566005	Urinary tract infectious disease (disorder)	XE0e0	Urinary tract infection
585.6	End stage renal disease	46177005	End stage renal disease (disorder)	X3030	End stage renal disease

Table 5.4 Example automatic ICD-9-CM - SNOMED CT - CTV3 mappings


(5) Apply ICD9-CM to CTV3 mapping table to data set,
saving mapped data coded in CTV3 and
saving file of unmapped codes from the original data set

File Reader
Import clinical data set
Cell Replacer
Replace ICD9 codes in clinical data with CTV3 equivalents
Cell Replacer
carry forward the note of mapping file source
String Replace (Dictionary)
Add CTV3 description
Row Splitter
Split out records with 'unknown' CTV3
CSV Writer
Write clinical records file with CTV3 codes replacing ICD9-CM codes

Figure 5.12 Save ICD-9-CM to CTV3 mapping table and file of unmapped codes

In this stage, shown in Figure 5.12, the data set coded in ICD-9-CM is imported. This is the data set containing codes that are required to be mapped to CTV3. To perform this mapping, each ICD-9-CM code in the data set is searched for in the ICD-9-CM to CTV3 mapping table and, should a mapping be found, the equivalent CTV3 code is added to the data set. If no mapping is found, the ICD-9-CM code is flagged as being an unmapped code. Those ICD-9-CM codes that mapped to a CTV3 code had that CTV3 code's description added. Items in the data set with ICD-9-CM codes that did not map to CTV3 codes were then split from the data set and passed to step 6; data set items which now had a CTV3 code were written to file for further analysis.

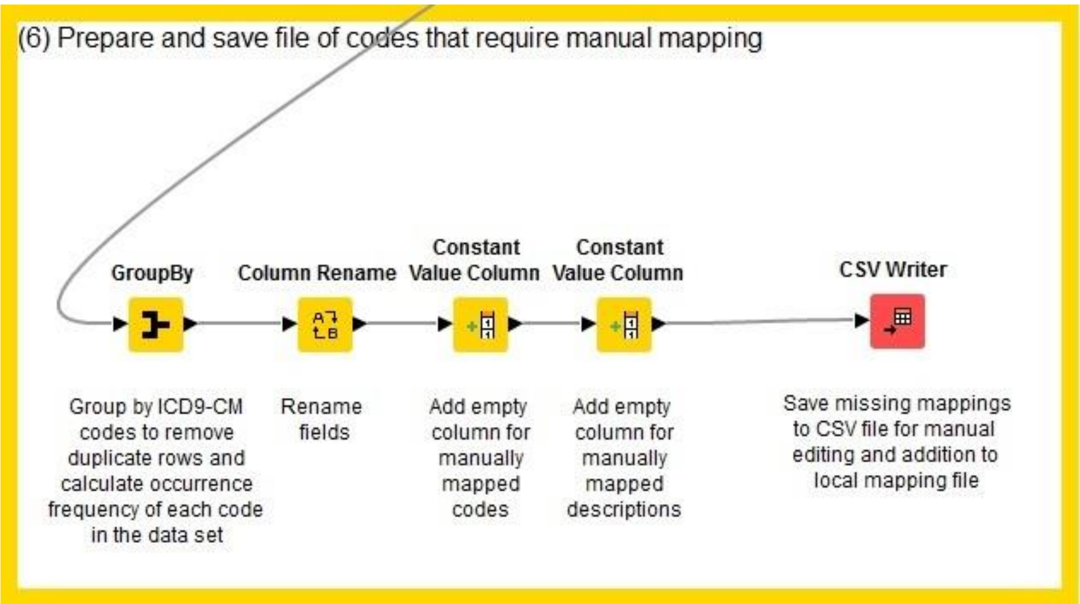

(6) Prepare and save file of codes that require manual mapping
GroupBy
Column Rename
Constant Value Column
Constant Value Column
CSV Writer
Group by ICD9-CM codes to remove duplicate rows and calculate occurrence frequency of each code in the data set
Rename fields
Add empty column for manually mapped codes
Add empty column for manually mapped descriptions
Save missing mappings to CSV file for manual editing and addition to local mapping file

Figure 5.13 Capture of ICD-9-CM codes that require to be manually mapped.

Data set items where the ICD-9-CM code did not have an equivalent CTV3 code in the mapping table were then operated on in this section of the workflow, shown in Figure 5.13. Firstly, unique values for the unmapped ICD-9-CM codes were extracted and the frequency of occurrence of each of these unmapped codes calculated. These unique codes and their frequencies were then placed in a table which was written to a file which was then available for manual inspection and mapping. Manually mapped codes in this stage were then appended to the existing local manual mapping table imported in stage 3, or if the local manual mapping table did not yet exist, this table was saved and used as the first iteration of the manual mapping table. The KNIME workflow was run each time that the manual mapping table was updated.

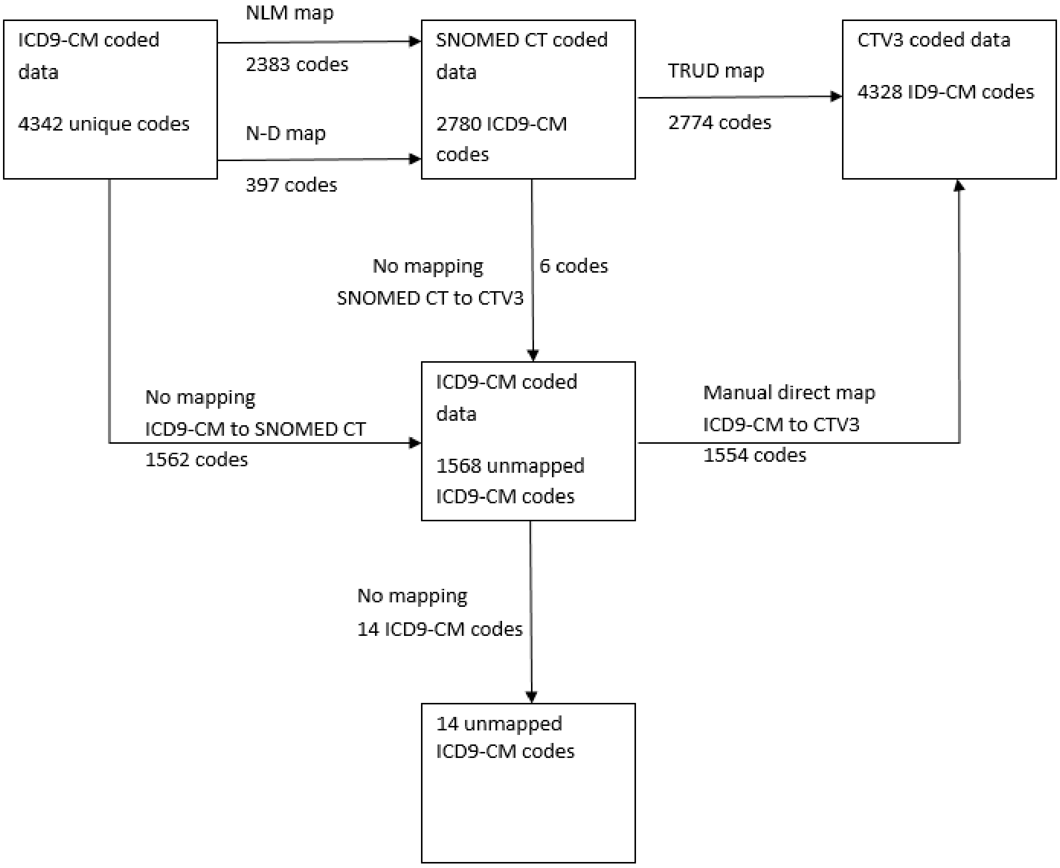


Figure 5.14 Results of the complete mapping processes.

When the complete workflow was run on the source dataset, it was observed that a significant number of codes did not map successfully, i.e. there was no ICD-9-CM to CTV3 mapping for these codes produced by the workflow, implying either that the ICD-9-CM to SNOMED CT mapping or the subsequent SNOMED CT to CTV3 mapping was incomplete. Additional mappings generated by Nadkarni and Darer [189], who had previously found this problem of missing mappings in their own work mapping ICD-9-CM codes to SNOMED CT codes, were added to the ICD-9-CM to SNOMED CT mapping table. Nadkarni and Darer had mapped 784 codes from ICD-9-CM to SNOMED CT, of which 399 were present in the data set used in this work (2 of which codes were also mapped in the NLM mapping table: ‘345.11 Generalized convulsive epilepsy, with intractable epilepsy’ and ‘V10.82 Personal history of malignant melanoma of skin’). This left a further 1562 codes in the data set that did not map from ICD-9-CM codes to SNOMED CT codes. For the subsequent SNOMED CT to CTV3 mapping, there were 6 SNOMED CT codes produced by the ICD-9-CM to

SNOMED CT mapping that failed the subsequent map to CTV3, giving a total of 1568 ICD-9-CM codes that did not map automatically from ICD-9-CM codes to CTV3. Figure 5.14 shows the results for each stage of the mapping process, both automatic and manual.

There were some areas of concern with the mappings. For the ICD-9-CM to SNOMED CT mapping, there were many one-to-one code mappings but there were also some one-to-many code mappings (for example, ICD-9-CM code 722.52 ‘Degeneration of lumbar or lumbosacral intervertebral disc’ maps to SNOMED CT code 26538006 ‘Degeneration of lumbar intervertebral disc (disorder)’ or 60937000 ‘Degeneration of lumbosacral intervertebral disc (disorder)’) and it was not always possible to map to more a more granular SNOMED CT code that would be a single parent of the ‘many’ target codes (see US National Library of Medicine, 2016 for a discussion of this issue). For those ICD-9-CM codes for which many SNOMED CT codes were suggested, the first suggested code was taken as the mapped code and this code was then used as the basis for the subsequent mapping to CTV3.

To further reduce the number of unmapped codes, a fourth mapping table was created, directly mapping ICD-9-CM to CTV3. This table was created by generating a list of unmapped codes, prioritised by the frequency with which these codes appeared in the source data set (in order to allow development work on analysis of this data set to proceed before this manual mapping table was complete by focussing mapping work on the most common unmapped codes) but also opportunistically mapping clinically related unmapped codes at the same time as mapping the unmapped more-frequent codes. Equivalent ICD-9-CM to CTV3 code mappings were deduced by inspection of code descriptions and by each code’s position in the CTV3 code hierarchy. Some codes in ICD-9-CM had very similar descriptions in CTV3, perhaps differing only in US vs UK spelling or in word ordering in the description and so the mapping was straightforward (e.g. ICD-9-CM code 782 “Disturbance of skin sensation” was mapped to CTV3 code XM07D “Skin sensation disturbance”), others had markedly different descriptions in the two coding systems and relied on an understanding of synonyms for the same conditions (e.g. ICD-9-CM code 734 “Flat foot” was mapped to CTV3 code N34.. “Pes planus”).

1944 ICD-9-CM to CTV3 manual mappings were created. There were 14 ICD-9-CM codes that were not mapped automatically and could not be mapped manually, since a search of ICD-9-CM and CTV3 concept descriptions (including synonyms) did not find a close match in the CTV3 descriptions to the ICD-9-CM concept description. It is believed that these ICD-9-CM codes remain unmapped because there is no equivalent CTV3 code to be found. These unmapped codes are listed in Table 4.

Examples of codes that required to be matched manually:

(1) Simple match

ICD-9-CM code 477 "Allergic rhinitis" mapped to CTV3 code XE0Y5 "Allergic rhinitis"

Failed to map at the ICD-9-CM to SNOMED CT stage: code 477 not found in UMLS 1:1 or 1:many mapping tables, nor in the Nadkarni-Darer mapping table.

(2) Match with minor US vs UK spelling variation:

ICD-9-CM code 599.7 "Hematuria" mapped to CTV3 code K197. "Haematuria"

Failed to map at the ICD-9-CM to SNOMED CT stage: code 599.7 not found in UMLS 1:1 or 1:many mapping tables, nor in the Nadkarni-Darer mapping table.

(3) More complex match: different word:

ICD-9-CM code 54.7 "Other Repair Of Abdominal Wall And Peritoneum" mapped to CTV3 code Xa9ZY "Repair of mesentery"

Failed to map at the ICD-9-CM to SNOMED CT stage: code 54.7 not found in UMLS 1:1 or 1:many mapping tables, nor in the Nadkarni-Darer mapping table.

(4) More complex match: different order of words:

ICD-9-CM code 715.04 ‘Osteoarthritis, generalized, involving hand’ mapped to CTV3 code XE1DW ‘Generalised osteoarthritis of the hand’

Failed to map uniquely at the ICD-9-CM to SNOMED CT stage: code 54.7 not found in UMLS 1:1 table, 6 options found in the 1:many mapping tables; not in the Nadkarni-Darer mapping table.

Mapping accuracy and efficiency was improved by experience and knowledge of the terms in the mapping tables, in particular an understanding of the differences between US and UK spellings and of English terms versus Latin terms. Searching for equivalent terms could also be expedited by search for word stems rather than complete words, for example if searching for a CTV3 equivalent to the ICD-9-CM code 601 ‘Inflam diseases of prostate’, a search for the stem ‘prostat’ would find the CTV3 code XE0e7 ‘Prostatic inflammatory disease’, which would have been missed by a search for the whole word ‘prostate’.

The complete workflow built the latest version of the mapping table using input from the US National Library of Medicine mapping table for ICD-9-CM to SNOMED CT, the Nadkarni-Darer mapping table for ICD-9-CM to SNOMED CT, the NHS Digital Technology Reference data Update Distribution mapping table for SNOMED CT to CTV3, and the manual direct mapping table for ICD-9-CM to CTV3. This allowed the latest version of the manual mapping table to be used and further allowed the latest versions of the NLM, Nadkarni-Darer and TRUD tables to be used should they be updated during the development period of this work.

Run time to build the complete mapping table and to remap the ICD-9-CM event codes in the source data set to CTV3 was primarily the time taken to read in the source data tables: the code mapping tables, the CTV3 code and description table, and the source data set. The complete workflow ran in approximately 45 seconds from a start position where no tables had been read (Asus, Windows 10 64-bit, intel Core i7-3610QM CPU @ 2.0 GHz, Nvidia GeForce GT 630M GPU, 8 GB RAM, 750GB hard, all inputs reset, all output tables re-written during the mapping table building process) to an end position where a mapped table of events coded in in ICD-9-CM and their code equivalents in CTV3 and a table of unmapped codes were both written.

There were 4342 unique ICD-9-CM codes present in this data set. Of these, 2780 (64.0 %) were mapped to SNOMED CT codes using mappings obtained from the NLM (2383 codes, 54.9 %) or from the work of Nadkarni and Darer (397 codes, 9.1 %), for a combined total of 2780 codes (64.0 %). Note that 2 codes further codes were mapped in the Nadkarni-Darer tables that were already in the NLM tables; for these codes the NLM table was given precedence. This overlap was likely due to these ICD-9-CM code mappings not being present in the NLM tables at the time (2007) that Nadkarni and Darer did their work but were present in the later (December 2016) version of the NLM table used in this work. It is recommended to always use the latest available versions of the mapping tables. These SNOMED CT codes were then mapped to CTV3 using mapping tables from NHS TRUD: 2774 codes mapped successfully, 63.9 % of the complete set of unique ICD-9-CM codes, 99.8 % of those codes that had been mapped to SNOMED CT. 6 codes failed to map from SNOMED CT to CTV3 codes using the TRUD mapping.

This left 1568 unique codes in the data set that did not map to CTV3 (36.0 % of the data set), failing to map at either the ICD-9-CM to SNOMED CT stage (1562 codes) or at the SNOMED CT to CTV3 stage (6 codes). It should be noted that 942 codes of these codes (21.7 % of the data set) were found in the NLM ICD-9-CM to SNOMED CT 1:many mapping table, however a decision had been made that it was not possible to select which of the "many" codes to select for the intermediate step towards CTV3 codes and so these codes were left as unmapped codes. These 1568 codes were then mapped manually. 1554 codes were mapped successfully, leaving 14 codes (0.3 %) that could not be mapped from ICD-9-CM to CTV3 due to no equivalent code being found. This gave a combined total of 4328 codes (99.7 %) that were mapped from ICD-9-CM to CTV3 using either the automatic or the manual process.

5.4.3 Summary of mapped codes.

Of the 4342 unique ICD-9-CM codes in the US-sourced data set, 2774 codes (64.0 %) were successfully mapped to a CTV3 code using an automatic approach, 96.3 % of these being mapped exactly or approximately as judged a domain expert. Of the 1568 remaining ICD-9-CM codes, 1554 were mapped manually, 95.6 % being mapped appropriately when judged by a domain expert. The success of the automatic mapping was compared to the success of the manual mapping, showing that automatic mapping was less successful than manual mapping in exact mapping ($p < 0.01$) but as successful when both exact and approximately successful mappings were compared ($p = 0.29$).

Number of unique ICD-9-CM codes present in the US data set: 4342
Number of unique ICD-9-CM codes that map to SNOMED CT using the NLM mapping table: 2383
Number of unique ICD-9-CM codes that map to SNOMED CT using the extra mappings from Nadkarni and Darer: 397*
Number of unique ICD-9-CM codes that map to SNOMED CT using NLM + Nadkarni & Darer: 2780
Number of unique ICD-9-CM codes that have no map from ICD-9-CM to SNOMED CT: 1562
* 2 further ICD-9-CM codes were mapped in the Nadkarni and Darer table that already had a map in the UMLS table.

Number of codes that map completely from ICD-9-CM to CTV3 (using the NLM and Nadkarni & Darer mappings to map from ICD-9-CM to SNOMED CT, and then the TRUD mapping to map from SNOMED CT to CTV3): 2774

A sample of the manual mapping table is shown in Table 3.

Source codes		**Mapped codes**		
ICD-9-CM code	**ICD-9-CM code description**	**CTV3 code**	**CTV3 code description**	**Search metho**d
250	type II diabetes mellitus [non-insulin dependent type] [NIDDM type] [adult-onset type] or unspecified type, not stated as uncontrolled, without mention of complication	X40J5	Type II diabetes mellitus	Search string matching
401	Essential hypertension	XE0Uc	Essential hypertension	Search string matching
311	Depressive disorder, NOS	E2B..	Depressive disorder NEC	Search string matching
305.1	Tobacco use disorder	Eu170	[X]Mental & behav dis due to use tobacco: acute intoxication	Key word matching
V70.0	Routine general medical examination at a health care facility	ZV700	[V]Routine health check-up	Concept matching
782	Disturbance of skin sensation	XM07D	Skin sensation disturbance	Key word matching
729.2	Neuralgia, neuritis, and radiculitis, unspecified	XE1Fn	Neuralgia, neuritis or radiculitis NOS	Search string matching
250.6	Diabetes with neurological manifestations	XE10H	Diabetes mellitus with neurological manifestation	Key word matching
54.7	Other Repair Of Abdominal Wall And Peritoneum	Xa9ZY	Repair of mesentery	Concept matching
519.11	Acute bronchospasm	Xa0Ns	Bronchospasm	Key word matching
250	type II diabetes mellitus [non-insulin dependent type] [NIDDM type] [adult-onset type] or unspecified type, not stated as uncontrolled, without mention of complication	X40J5	Type II diabetes mellitus	Search string matching
704.8	Other specified diseases of hair and hair follicles	M24..	Hair and hair follicle diseases	Search string matching
435.9	Unspecified transient cerebral ischemia	G65z.	Transient cerebral ischaemia NOS	Search string matching
333.1	Essential and other specified forms of tremor	F131.	Essential and other specified forms of tremor	Search string matching
272.4	Other and unspecified hyperlipidemia	Cyu8D	[X]Other hyperlipidaemia	Key word matching, spelling difference

Table 5.5 Example manual mappings from ICD-9-CM codes present in the US data set to CTV3 codes

ICD-9-CM code	ICD-9-CM code description	Frequency of code occurrence in data set
V45.8	Other postsurgical status	1
V45.86	Bariatric surgery status	6
V45.89	Other postsurgical status	20
V58.63	Encounter for long-term (current) use of antiplatelets/antithrombotics	2
V68.01	Disability examination	4
V68.8	Other specified administrative purpose	2
V68.89	Encounters for other specified administrative purpose	5
V68.9	Encounters for unspecified administrative purpose	1
V76.47	Screening for malignant neoplasms of the vagina	6
V78.8	Screening for other disorders of blood and blood-forming organs	6
V85.51	Body Mass Index, pediatric, less than 5th percentile for age	1
V85.52	Body Mass Index, pediatric, 5th percentile to less than 85th percentile for age	2
V85.53	Body Mass Index, pediatric, 85th percentile to less than 95th percentile for age	1
V87.2	Contact with and (suspected) exposure to other potentially hazardous chemicals	1

Table 5.6 ICD-9-CM codes present in the US data set for which no conceptual match in the CTV3 codes set was found

ICD-9-CM codes that did not map automatically to a CTV3 code and for which no equivalent CTV3 code could be found manually are shown in Table 5.6. These failed mappings were checked by a domain expert, who could find no suitable code match. It can be seen that none of the unmatched codes are codes for symptoms or conditions; should the data be used for analysis of symptoms and conditions, as is the case here, it may not be worth expending too much effort into finding maps for codes that will have no further use in any analysis.

5.4.4 Verification of the mapping process.

ICD-9-CM codes were mapped to CTV3 codes either by an automatic process via SNOMED CT or, for those codes that failed the automatic mapping process, by a manual process. Either route required verification and an assessment of the exactness of the mapping pairs.

Codes that failed to map automatically were mapped manually by the author of this report (JT), with these mappings verified by a domain expert (Dr Hugh O’Sullivan (HO’S), a general practice clinician based at Temple Street Children's University Hospital, Temple Street, Dublin 1, Ireland). To determine the degree of success of matching, code descriptions for ‘matched’ ICD-9-CM and CTV3 codes were inspected for equivalence and a judgement made using the success definitions defined by De [219]: exact matching: ‘both codes have the exact clinical meaning’; approximate matching: ‘the two codes have similar clinical meaning although the underlying clinical contexts are not the same’. When codes’ descriptive terms were identical or differed only in minor spelling variation the verification was simple. More complex differences required some knowledge of clinical terminology. Where matching was not exactly successful, consideration was given by HO’S to replacing the mapped code with a CTV3 code that achieved exact success status.

Similarly, the automatic mappings were inspected for equivalence using the same criteria, but this time the inspection was performed by JT. With the high degree of successful matching by JT (95.7%, as judged by HO’s) in the manual mapping, there was confidence in the judging of the success of the automatic mapping. Again, mapping success was judged using the definitions of De and for inexact or incorrect mappings an improved mapping was suggested.

ICD-9-CM code	ICD-9-CM description	CTV3 code	CTV3 description	Exact mapping	Approx. mapping	Improved code	Improved code description
493.21	Chronic obstructive asthma with status asthmaticus	H3121	Emphysematous bronchitis	N	N		
EP186	STATIN INTOLERANCE	Xa1pS	Drug allergy	N	Y	XaG2V	Statins contraindicated
385.24	Partial loss or necrosis of ear ossicles	F5523	Partial loss/necrosis,ossicles	Y	(Y)		
680.6	Carbuncle and furuncle of leg, except foot	M007.	Carbuncle of foot	N	N	M006.	Carbuncle of leg (excl. foot)
286.9	Other and unspecified coagulation defects	D30..	Bleeding diathesis	N	Y	XE14m	Coagulation defects
151.0	Malignant neoplasm of cardia	B110.	Malignant tumour of cardia	Y	(Y)		

Table 5.7 Examples of automatically-generated code mappings from ICD-9-CM to CTV3 with score of success of mapping and suggested improved mappings

(Y) indicates an implicit successful approximate mapping due to a successful exact matching

In Table 5.7, mapping for ICD-9-CM code 680.6 to CTV3 code M007. is scored as neither an exact mapping nor an approximate mapping. This is due to the ICD-9-CM code explicitly excluding the foot as the site of the carbuncle or furuncle, but the automatic mapping returning a CTV3 that explicitly includes the foot as the site.

Results from the verification of the success of the automatic code mappings are shown in Table 6.

	All auto	Nad	NLM	NAD %	NLM %	All %
Exact mapping	2219	235	1984	59.2	83.3	79.8
approximate, replacement code suggested	273	54	219	13.6	9.2	9.8
approximate, no replacement code suggested	186	73	113	18.4	4.7	6.7
incorrect, replacement code suggested	84	26	58	6.5	2.4	3.0
incorrect, no replacement code suggested	18	9	9	2.3	0.4	0.6
all correct (exact and approximate)	2682	366	2316	91.2	97.2	96.3
Total	2780	397	2383	100	100	100

Table 5.8 Results of verification of automatic code mapping.

79.8 % of the automatically-generated code mappings were judged to be exact mappings. A further 16.5 % of the automatically-generated code mappings were judged to be approximately successful code mappings, again using the success definitions of De [2012]. 3.6 % of the automatic code mappings were judged to have produced an incorrect mapping.

1568 ICD-9-CM codes failed to map, in that the automatic mapping process failed to produce a CTV3 code as an output. For each of these ICD-9-CM codes, JT proposed a best matching CTV3 code. All these manually mapped codes were scored for accuracy and exactness by a domain expert (HO'S). Results of this verification are shown in Table 7.

	Absolute	%
Exact mapping	1421	91.4
approximate, replacement code suggested	13	0.8
approximate, no replacement code suggested	53	3.4
incorrect, replacement code suggested	13	0.8
incorrect, no replacement code suggested	54	3.5
all correct (exact and approximate)	1487	95.7
Total	1554	100

Table 5.9 Results of verification of manual code mapping

Success of mapping	Manual	Automatic	Chi-square	P
Exact mapping	1421	2219	100.1	<0.01
Approximate mapping	66	459		
Correct mapping (exact + approximate)	1487	2678	1.1	0.29
Incorrect mapping	67	102	1.1	0.29
Total codes	1554	2780		

Table 5.10 Number of codes mapped by automatic and manual processes and their success

The relative success of the manual and automatic mapping processes was tested by comparison of the proportions of the codes that were successfully mapped by each method, using Pearson's chi-squared test as the test for significant difference between the groups. Table 5.10 shows the results from these tests.

Chi-square tests: manual vs automatic mapping.

Exact mapping: 1421 of 1554 codes (91.4 %) for manual mapping, 2219 of 2780 codes (79.8 %) for automatic mapping.

Chi-square: 100.1; $p < 0.01$. Manual mapping was significantly more successful for exact mapping than automatic mapping.

All ‘correct’ mapping (exact and approximate): 1487 of 1554 codes (95.7 %) for manual mapping, 2678 of 2780 codes (96.3 %) for automatic mapping.
Chi-square: 1.1; $p = 0.29$. No significant difference in the rate of successful mapping.

Incorrect mapping: 67 of 1554 codes (4.3 %) for manual mapping, 102 of 2780 (3.7 %) codes for automatic mapping.
Chi-square: 1.1; $p = 0.29$. No significant difference in the rate of incorrect mapping.

Automatic mapping performs as well as manual mapping when comparing the number of incorrect mappings or the number of successful (in the broadest definition, including both exactly successful and approximately successful) mappings. However manual mapping outperforms automatic mapping when the number of exactly successful mappings are considered.
Where it was determined that a code mapping, whether automatic or manual, could be improved, these improved code mappings were added to the manual mapping table. The manual mapping table was prioritised over the automatic mapping table in the final mapping table generation and used in subsequent data set analysis.

5.5 Merging process

Merging of the source data sets is performed as a four-stage process and illustrated in Figure 2. The stages in the process are:

5.5.1 Across all data sources:

Decide on a common set of fields and a common data standard for each field, for example ‘event date’ to be recorded as a year, since this is the coarsest granularity of date available across all source data sets.

5.5.2 For each source data set independently:

Merge the individual data set files supplied by each data source as appropriate: some sources supply data in more than one set; each set contains several tables. Source data sets may also be structured quite differently from each other, some as normalized tables, others as more flattened tables.

5.5.3 For each source data set independently:

Remove any fields that are not required in the final merged data set. Rename all fields to the common name set. The prepared data is then saved to disk, one file per source data set.

5.5.4 For each source data set independently:

Map the event codes from the coding systems used in the source to the target coding system (i.e. in our sets, from Read v2 to Clinical Terms Version 3 or from ICD-9-CM to Clinical Terms Version 3 as appropriate). This process is described separately later in this chapter in section 5.6.

5.5.5 Combine the source data sets to form a single composite data set:

This is a flattened data file with one line per patient, each line containing patient demographic information together with a complete set of event codes from the medical history.

5.5.6 Remove any events that are not required for the analysis, mostly administration-only events (e.g. "registered at practice").

These events are those that have no clinical significance, i.e. are neither a symptom, a treatment nor a diagnosis.

Figure 5.15 illustrates the data set merging process described above.

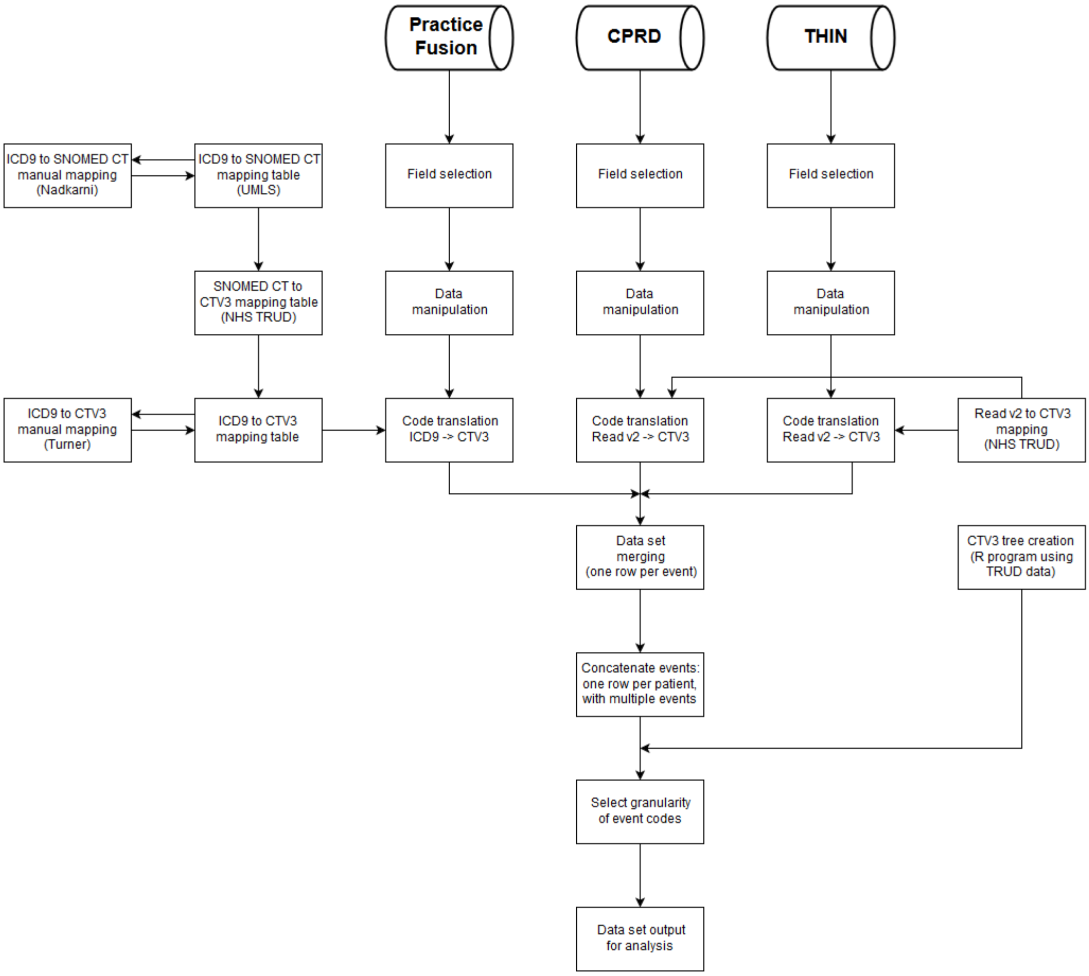

Practice Fusion
CPRD
THIN
ICD9 to SNOMED CT manual mapping (Nadkarni)
ICD9 to SNOMED CT mapping table (UMLS)
Field selection
Field selection
Field selection
SNOMED CT to CTV3 mapping table (NHS TRUD)
Data manipulation
Data manipulation
Data manipulation
ICD9 to CTV3 manual mapping (Turner)
ICD9 to CTV3 mapping table
Code translation ICD9 -> CTV3
Code translation Read v2 -> CTV3
Code translation Read v2 -> CTV3
Read v2 to CTV3 mapping (NHS TRUD)
Data set merging (one row per event)
CTV3 tree creation (R program using TRUD data)
Concatenate events: one row per patient, with multiple events
Select granularity of event codes
Data set output for analysis

Figure 5.15 Workflow for creation of single composite data set from several clinical data sources

Steps 1 to 3 in the merging process were run independently for each source data set using KNIME. One KNIME workflow was written for each source data set. These workflows vary slightly according to the demands of the structures of the source data sets and the format of the source data (e.g. date formats, some information in look-up tables, etc.). Each workflow is discussed in turn below. In step 4, event codes are mapped to a common coding system. For step 5, a further KNIME workflow is discussed. This workflow merges the separate data files and removes data items that are not required for future analysis. These are primarily records of administration events, which are not present in all data sources.

Note that if data becomes available from other sources, this new data set will be required to go through steps 1 to 4 in the above process before running step 5 on all data sets - the previously-acquired data sets will not need to go through steps 1 to 3 again unless there is a change in the desired set of fields to be included in the merged data set. However if no new fields are to be included then field removal can be performed in the data set merging step. Note also that once the merged data set has been created, it is saved to disk, and is then read by the data analysis system; it is a simple matter to drop unrequired fields at the point of reading the merged data file.

5.6 Preparation of the individual source files

5.6.1 THIN data preparation using KNIME

Figure 5.16 shows the KNIME workflow used to prepare the data supplied in the THIN data tables. The process implemented in the KNIME workflow is as follows:

a. The ‘patient information’ section in the KNIME workflow prepares data from the patient information file:

   (i) The patient information files are read. They are plain text files - note that each row in the input file is a single text string, with no separator between fields – fields are split by position. For each source file, a field is added to the data table to note which file is the source of the data. The two tables are then concatenated to form a single data table.

(ii) The input character strings are now split into separate fields, according to the position in the file as described in the THIN data dictionary.

(iii) Unknown dates are recorded in the THIN data as the string ‘00000000’. These unknown date strings are replaced with an empty character string.

(iv) Years of birth and death are extracted from the date of birth and date of death fields, saved as new fields and converted to integers.

(v) Fields unwanted for later processing are removed from the data table.

b. The ‘Event information’ section in the KNIME workflow prepares data from the event information files.

(i) The event information files are read. They are plain text files - note that each row in the input file is a single text string, with no separator between fields – fields are split by position. The two data tables are then concatenated to form a single data table.

(ii) The input character strings are now split into separate fields, according to the position in the file as described in the THIN data dictionary.

(iii) Unknown dates are recorded in the THIN data as the string ‘00000000’. These unknown date strings are replaced with an empty character string.

(iv) Unknown event dates are replaced by the system dates, i.e. the date when the data item was entered into the GP system.

(v) Events are sorted by date and an “order number” assigned to each event in order to preserve event ordering.

(vi) The year of each event is extracted and placed in a new field in the data table.

(vii) Unwanted columns are removed from the data table.

(viii) The THIN table of Read Codes and descriptions is read into its own data table; like other THIN tables each row is a single character string. These strings are split by position into their constituent data items.

(ix) Read Code descriptions are added to the event table by an inner join on ‘medcode’ between the event table and the Read Code table. Note that at this point, Read Codes are still 7-byte, i.e. they retain the final 2 bytes that allow for synonyms.

c. The ‘Event information’ section in the KNIME workflow prepares data from the event information files. Patient information table and event table are combined by an inner join on Patient ID.

d. The combined patient and event information table is then tidied:

   (i) Numeric codes for marital status are converted to ‘M’ or ‘F’

   (ii) Read Codes for events are trimmed from 7 bytes to the base 5 bytes

   (iii) Patient age at each event was calculated and stored

   (iv) Numeric codes for marital status are converted to descriptive strings

   (v) Year of end of event is extracted (or missing value recorded if there is no end date)

   (vi) Patient age at date of data collection, or age at death if applicable, is calculated and saved.

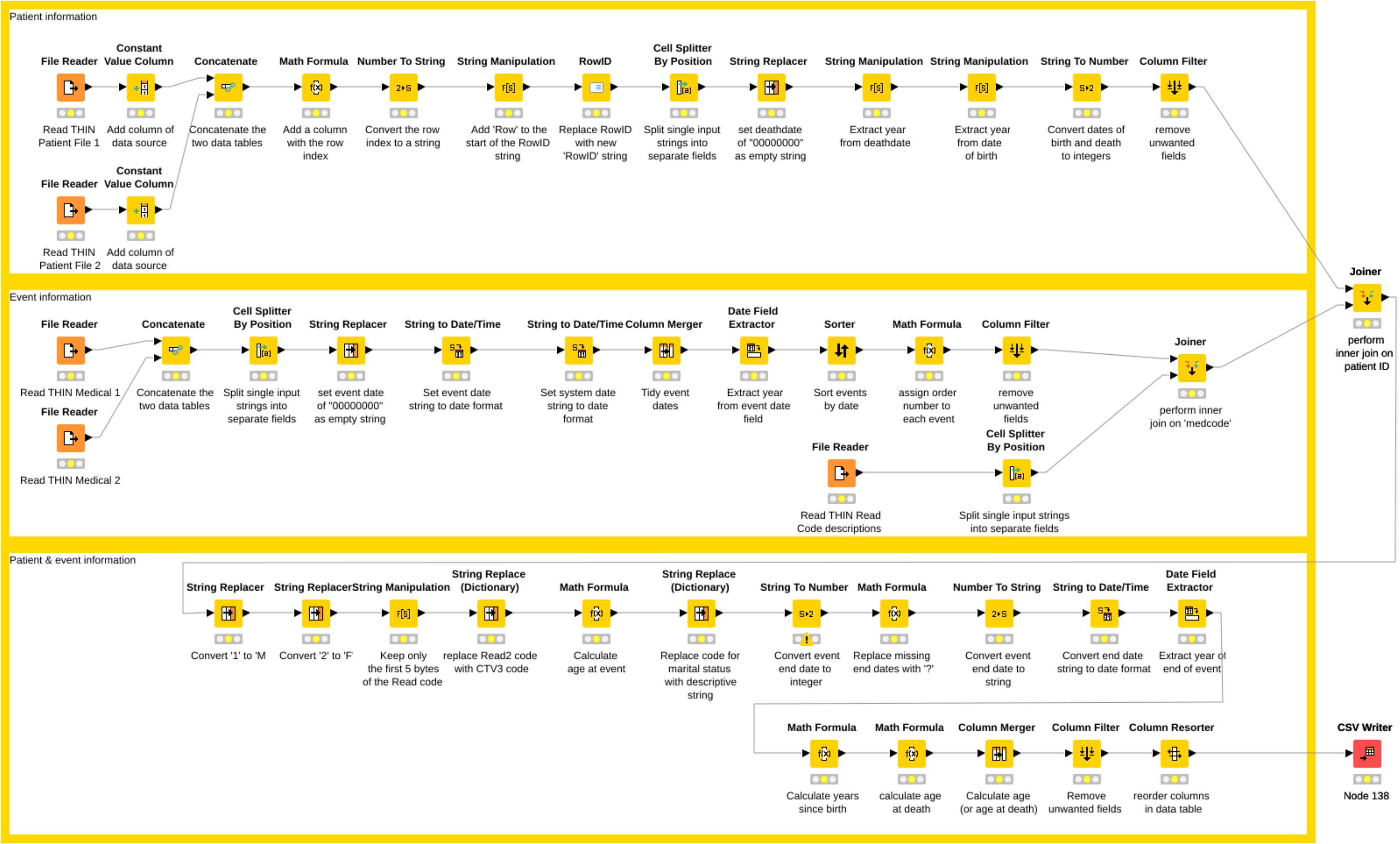
   (vii) Unwanted fields are removed

   (viii) Data table columns are reordered.

   (ix) The data table is written to csv file.

There were 732 events with an event start year of “2” after import into KNIME and conversion to year-only. The mode event year was 2009. The oldest event year was 1911, excluding years recorded as “2”. “2” in fact is an artefact of date conversion in KNIME, from a “missing” date, which is coded with a value of “00000000” and which gets converted to a year of 2 and month of November - but note that months get removed in this data tidying process. Investigating this problem highlighted another problem: dates that are stored in the THIN data set as a years-only string (e.g. “19270000”) are translated by KNIME to a date format which has the previous year – in this case 19270000 gets converted to 30-Nov-1926. The solution used in the data tidying process was to replace the substring “0000” in the date field with “0702” (i.e. 2nd July, the middle of the calendar year – note that we are not expecting to use months and days later in the analysis but this would allow for that). This was done AFTER converting “00000000” years to a missing value by searching for the complete strings “00000000” and replacing with “?”, the representation used for missing values in KNIME. Then strings “nnnn0000” were searched for with simple regex search for “0000$” and replaced with “nnnn0702”.

For events without a date, the date that the event was recorded on the practice system was assigned to the event date. For those events where both the system date and the event date were earlier than the patient’s date of birth, the year of birth was taken as the event year. Following this process, no events were without an event year.



Patient information
File Reader
Read THIN Patient File 1
Constant Value Column
Add column of data source
File Reader
Read THIN Patient File 2
Constant Value Column
Add column of data source
Concatenate
Concatenate the two data tables
Math Formula
Add a column with the row index
Number To String
Convert the row index to a string
String Manipulation
Add 'Row' to the start of the RowID string
RowID
Replace RowID with new 'RowID' string
Cell Splitter By Position
Split single input strings into separate fields
String Replacer
set deathdate of "00000000" as empty string
String Manipulation
Extract year from deathdate
String Manipulation
Extract year from date of birth
String To Number
Convert dates of birth and death to integers
Column Filter
remove unwanted fields
Joiner
perform inner join on patient ID
Event information
File Reader
Read THIN Medical 1
File Reader
Read THIN Medical 2
Concatenate
Concatenate the two data tables
Cell Splitter By Position
Split single input strings into separate fields
String Replacer
set event date of "00000000" as empty string
String to Date/Time
Set event date string to date format
String to Date/Time
Set system date string to date format
Column Merger
Tidy event dates
Date Field Extractor
Extract year from event date field
Sorter
Sort events by date
Math Formula
assign order number to each event
Column Filter
remove unwanted fields
File Reader
Read THIN Read Code descriptions
Cell Splitter By Position
Split single input strings into separate fields
Joiner
perform inner join on 'medcode'
Patient & event information
String Replacer
Convert '1' to 'M
String Replacer
Convert '2' to 'F
String Manipulation
Keep only the first 5 bytes of the Read code
String Replace (Dictionary)
replace Read2 code with CTV3 code
Math Formula
Calculate age at event
String Replace (Dictionary)
Replace code for marital status with descriptive string
String To Number
Convert event end date to integer
Math Formula
Replace missing end dates with '?'
Number To String
Convert event end date to string
String to Date/Time
Convert end date string to date format
Date Field Extractor
Extract year of end of event
Math Formula
Calculate years since birth
Math Formula
calculate age at death
Column Merger
Calculate age (or age at death)
Column Filter
Remove unwanted fields
Column Resorter
reorder columns in data table
CSV Writer
Node 138

Figure 5.16 KNIME workflow for preparation of THIN data set

5.6.2 CPRD data preparation using KNIME

The KNIME workflow shown in Figure 5.17 illustrates the CPRD data preparation process. The process implemented in this KNIME workflow is as follows:

a. The ‘patient information’ section prepares data from the patient information file.
   (i) The file is read
   (ii) Patient gender in the file is coded as ‘1’ or ‘2’. These are converted to ‘M’ or ‘F’
   (iii) Marital status is converted from a code (1 to 6) to a English language character string (e.g. ‘divorced’, ‘engaged’) by means of a look-up table supplied as part of the CPRD data set.
   (iv) Year of birth is stored in the CPRD data set as (actual year of birth – 1800) and so 1800 has to be added back to the stored value to obtain the real year of birth.
   (v) Date of death is stored as a fully day-specific date. The year of death is extracted from this.
   (vi) Fields that will not be used are removed.

b. The ‘Event information’ section prepares date from the clinical event information file and from the immunisations data file.
   (i) File ‘event information’ is read
   (ii) File ‘immunisation data’ is read
   (iii) The two tables are merged by simple concatenation – they have the same fields.
   (iv) Missing event dates are replaced by the system date, i.e. the date that the event was entered into the GP system. This give the latest date by which the event will have occurred, although it is possible that the event occurred earlier than the system date.
   (v) A column is added to the table containing the source of the event data, in all cases here this will be ‘C13’ for the CPRD 2013 data set.
   (vi) Event dates are extracted from the event date field.
   (vii) The event tables is sorted by event date
   (viii) An ‘order’ number is assigned to each event.

(ix) The CPRD lookup table to map the CPRD ‘medcode’ to Read Version 2 codes is read

(x) The Read codes in the look-up table are shortened to retain only the first 5 bytes (bytes 6 and 7 are to allow for synonyms but have no additional clinical significance).

(xi) An inner join is performed between the events table and the medcode-to-Read-code table so that the CPRD medcodes in the events table can be replaced by Read Codes (Read version 2).

(xii) Fields not needed are now dropped.

c. An inner join is now performed between the patient information table and the events table (modified to contain Read Codes rather than ‘medcodes’). This produces a flattened table, each row containing information on one event together with the patient information for that event.

d. The flattened patient-event table is now prepared for output:

  (i) Dates are converted from strings to integers

  (ii) Patient age at date of data set preparation is calculated (or date of death if the patient died before data set preparation)

  (iii) The patient’s age at the date of each event is calculated

  (iv) An ‘event end date’ field is added for compatibility with other data sets, although CPRD data does not contain this information.

  (v) Fields not needed are now dropped.

  (vi) Columns in the data table are sorted into the desired order, for compatibility with tables output from other data sets.

  (vii) The data table of data from the CPRD data set is saved.

This data set has 14,323 unique codes represented. This is a much higher figure than the other two data sets. Inspection of the data set suggests that a large proportion of the codes are administration codes that have little or no use for this analysis (e.g. "patient attended clinic"). These administration codes are kept in the data set at this stage but will be removed, along with any administration codes present in data sourced from the other data sets, prior to final analysis performed on the composite data set.

December is a very much more frequent month of event than other months recorded in this data set. It is not known why this is the case but it is conjectured that when the month of event is unknown, December is used as a default value.

One individual’s year of death was recorded as 1963. It is not clear why or how are they were in the data set. It is possible that 1963 was entered in error, or that they had a rare and/or hereditary condition and that their data was entered retrospectively for this reason. On inspection of this record, it was seen that the individual’s year of birth was also recorded as 1963. Looking further, this individual had tens of events recorded, the most recent being in 2005. There was no record of death in the patient's recorded events. It was therefore assumed that 1963 was entered erroneously. It is possible to check all data for event dates for an individual that were recorded as taking place before their birth or after their death.

Some further tidying of the data was also performed. For example, some event dates were recorded as occurring before the patient was born: these events were assigned a new date of the date on which the event was entered into the GP system (this date was available in the data set). Some other events were recorded as being the same year as the year of birth, which may thought to be unreasonable (for example, a patient recorded in 2000 as having “notes summary on computer” in 1912, their year of birth) or reasonable (a patient recorded in 2001 as having “normal birth” in 1941, their year of birth). These event dates were left unchanged.

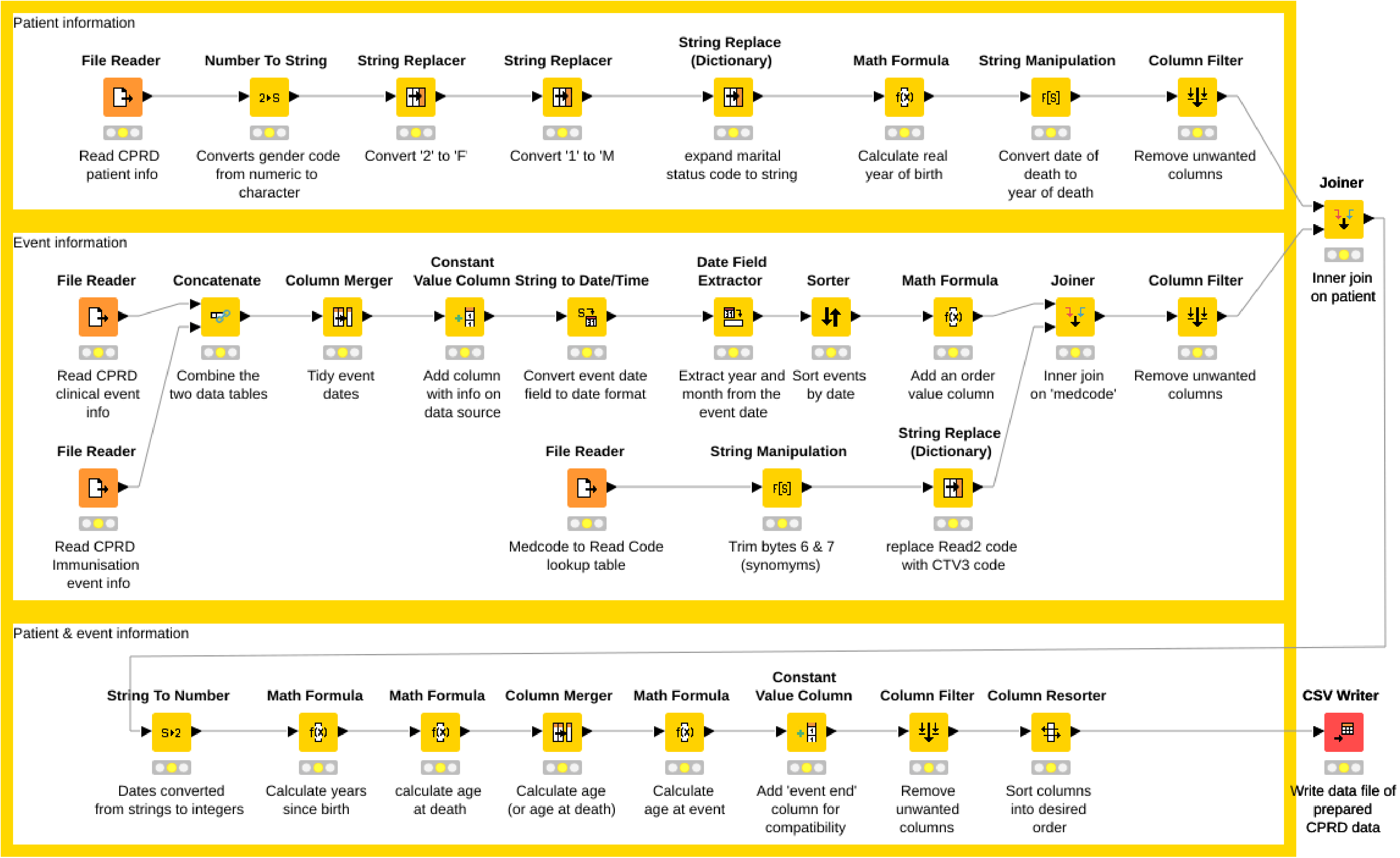

Patient information
File Reader
Read CPRD patient info
Number To String
Converts gender code from numeric to character
String Replacer
Convert '2' to 'F'
String Replacer
Convert '1' to 'M
String Replace (Dictionary)
expand marital status code to string
Math Formula
Calculate real year of birth
String Manipulation
Convert date of death to year of death
Column Filter
Remove unwanted columns
Joiner
Inner join on patient
Event information
File Reader
Read CPRD clinical event info
Concatenate
Combine the two data tables
Column Merger
Tidy event dates
Constant Value Column
Add column with info on data source
String to Date/Time
Convert event date field to date format
Date Field Extractor
Extract year and month from the event date
Sorter
Sort events by date
Math Formula
Add an order value column
Joiner
Inner join on 'medcode'
Column Filter
Remove unwanted columns
File Reader
Read CPRD Immunisation event info
File Reader
Medcode to Read Code lookup table
String Manipulation
Trim bytes 6 &amp; 7 (synomyms)
String Replace (Dictionary)
replace Read2 code with CTV3 code
Patient &amp; event information
String To Number
Dates converted from strings to integers
Math Formula
Calculate years since birth
Math Formula
calculate age at death
Column Merger
Calculate age (or age at death)
Math Formula
Calculate age at event
Constant Value Column
Add 'event end' column for compatibility
Column Filter
Remove unwanted columns
Column Resorter
Sort columns into desired order
CSV Writer
Write data file of prepared CPRD data

Figure 5.17 KNIME workflow for preparation of CPRD data set

5.6.3 Practice Fusion data preparation using KNIME

Data tables containing the various data items in the Practice Fusion data set were read in to the KNIME workflow, for both the supplied “train” and “test” file sets. Data sets were merged and flattened. Figure 5.18 shows the KNIME workflow for the preparation of this data set.

a. The ‘patient and event information’ section prepares and combines data from the ‘Patient’, Transcript’, ‘TranscriptDiagnosis’ and ‘Diagnosis’ files and the local ICD9-CM to CTV3 event codes look-up table..

    (i) The ‘Transcript’ files are read and concatenated.

    (ii) The ‘TranscriptDiagnosis’ files are read and concatenated.

    (iii) The Transcript and TranscriptDiagnosis tables are joined

    (iv) This table is then joined with the complete Diagnosis records table

    (v) This joined table is then joined with the combined Patient information table

    (vi) ICD9-CM codes in the Patient/Transcript/TranscriptDiagnosis/Diagnosis table are replaced with CTV3 codes from the local look-up table

b. The ‘Perform tidying of data and matching fields to other data sets’ section of the KNIME workflow calculates fields not natively in the Practice Fusion data set, remove fields not required in the final saved data set and renames fields for consistency with the final composite data set:

    (vii) The event date is checked for validity and corrected if necessary

    (viii) Events are sorted to be in chronological order and assigned a sequence number

    (ix) Fields no longer required are removed

    (x) Fields are renamed to conform to the fields required in the composite data set

    (xi) Missing value character for event year of ‘0’ replaced with ‘?’ for consistency with other fields and data sets

    (xii) Fields for family number and marital status added for consistency with other data sets, although information on these items not available in the Practice Fusion data sets

(xiii) Age in years at events calculated and stored

(xiv) Columns in the data table sorted to match the composite data set order

c. The ‘smoking information’ was read and prepared for joining to the patient event information table

(xv) Smoking information files were read and concatenated

(xvi) The ICD9-CM codes in the smoking information data were replaced with CTV3 codes from the local look-up table

(xvii) Fields no longer required were removed

d. The ‘Join smoking information to patient information’ section of the workflow adds tobacco use history to the patient information

(xviii) Smoking information joined to the patient event table

(xix) Fields are renamed to conform to the fields required in the composite data set

(xx) Age in years at events calculated and stored

e. The final section of the workflow, ‘Final internal data set merging’, combines the tobacco use information with the full table of patient information and event history and saves the prepared Practice Fusion data set in the composite data set format

(xxi) The tobacco use history information is concatenated to the patient event history table

(xxii) The patient age at date of data capture is calculated

(xxiii) The file is saved for further analysis

Events in the Practice Fusion data set were recorded in ICD9-CM and required translation to CTV3. This was done by use of a look-up table. No direct mapping of ICD9-CM codes to CTV3 (or ICD9-CM to Read v2) codes was found and so a mapping table required to be created. Creation of this look-up table is described in the paper included at the end of this chapter.

This data set has 4246 unique event codes represented. This is a smaller number than for either the CPRD or THIN data sets, each of which have fewer individuals’ records than the Practice Fusion data set. This suggests that ICD9-CM codes are less granular than Read Codes or that administration codes were not included in this data set. These 4246 ICD9 codes were mapped, in a process later described, to 3465 CTV3 codes. There was no explicit "date of death" for patients in the Practice Fusion data set. Individuals whose records were used in the Practice Fusion data set were aged 18 to 90 years inclusive only. This was most likely due to restrictions in the HIPAA regulations [123] designed to prevent re-identification of individuals from their medical records. Events have an “event start” date. For some patients this value was missing; in these cases the date of clinic visit was assigned to the “event start” date. Some transcripts did not have a recorded visit date. The final data set had 457,232 events with a recorded year and 240 events for which the year was unknown.

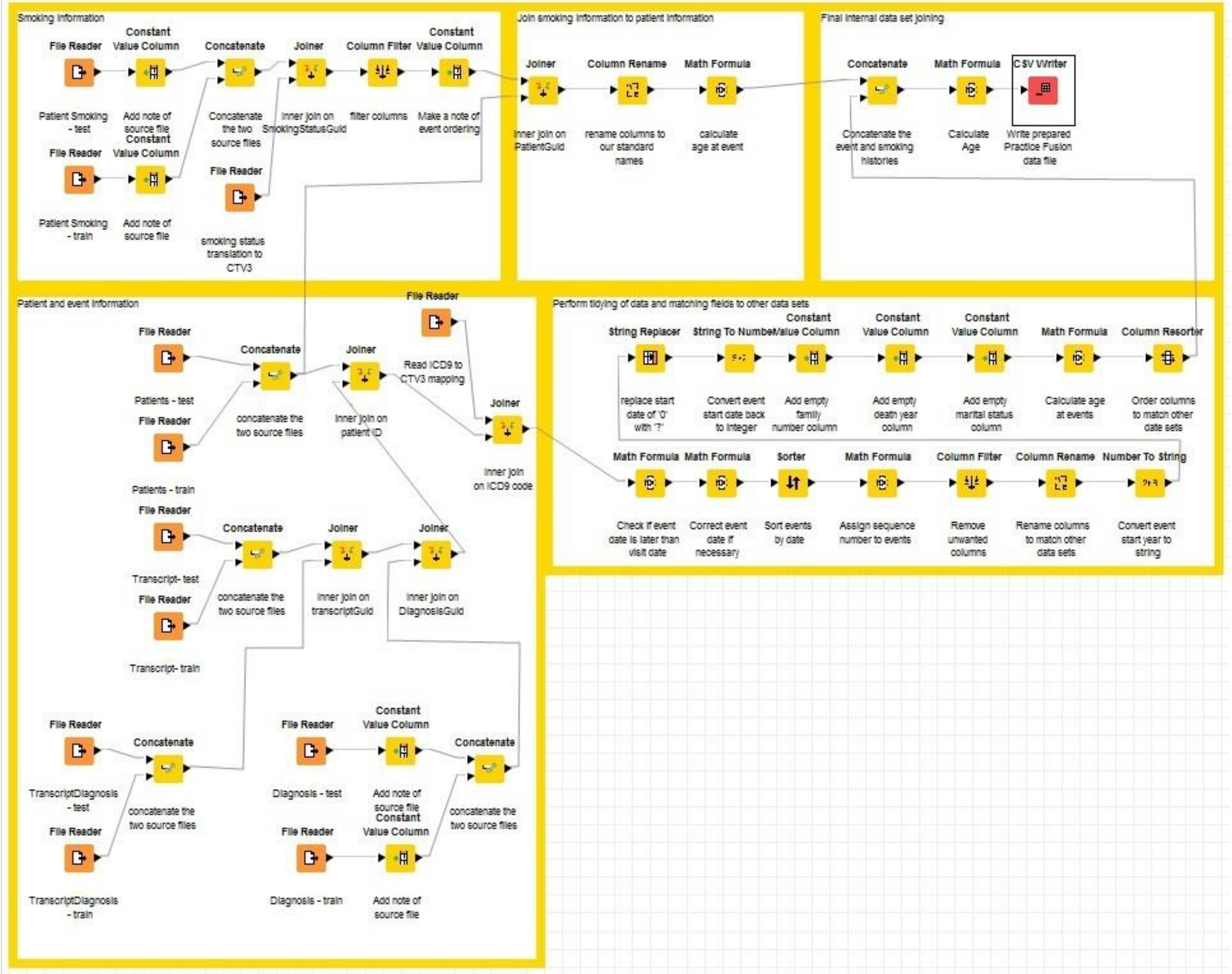

Smoking information
File Reader
Patient Smoking - test
Constant Value Column
Add note of source file
File Reader
Patient Smoking - train
Constant Value Column
Add note of source file
Concatenate
Concatenate the two source files
File Reader
smoking status translation to CTV3
Joiner
Inner join on SmokingStatusGuid
Column Filter
filter columns
Constant Value Column
Make a note of event ordering
Join smoking information to patient information
Joiner
Inner join on PatientGuid
Column Rename
rename columns to our standard names
Math Formula
calculate age at event
Final internal data set joining
Concatenate
Concatenate the event and smoking histories
Math Formula
Calculate Age
CSV Writer
Write prepared Practice Fusion data file
Patient and event information
File Reader
Patients - test
File Reader
Patients - train
Concatenate
concatenate the two source files
Joiner
Inner join on patient ID
File Reader
Read ICD9 to CTV3 mapping
Joiner
Inner join on ICD9 code
File Reader
Transcript- test
File Reader
Transcript- train
Concatenate
concatenate the two source files
Joiner
Inner join on transcriptGuid
Joiner
Inner join on DiagnosisGuid
File Reader
TranscriptDiagnosis - test
File Reader
TranscriptDiagnosis - train
Concatenate
concatenate the two source files
File Reader
Diagnosis - test
File Reader
Diagnosis - train
Constant Value Column
Add note of source file
Constant Value Column
Add note of source file
Concatenate
concatenate the two source files
Perform tidying of data and matching fields to other data sets
String Replacer
replace start date of '0' with '?'
String To Number
Convert event start date back to integer
Constant Value Column
Add empty family number column
Constant Value Column
Add empty death year column
Constant Value Column
Add empty marital status column
Math Formula
Calculate age at events
Column Resorter
Order columns to match other date sets
Math Formula
Check if event date is later than visit date
Math Formula
Correct event date if necessary
Sorter
Sort events by date
Math Formula
Assign sequence number to events
Column Filter
Remove unwanted columns
Column Rename
Rename columns to match other data sets
Number To String
Convert event start year to string

Figure 5.18 Workflow for preparation of Practice Fusion data set

5.7 Merging of source data sets into single composite data set

The source data sets having been prepared as described to consistent format, they can now be combined into a single composite data set. This process was carried out in a simple KNIME workflow and is shown in Figure 5.19. The single output data file is in comma-separated variable format and contains the fields chosen in the analysis of common fields described earlier in this chapter.

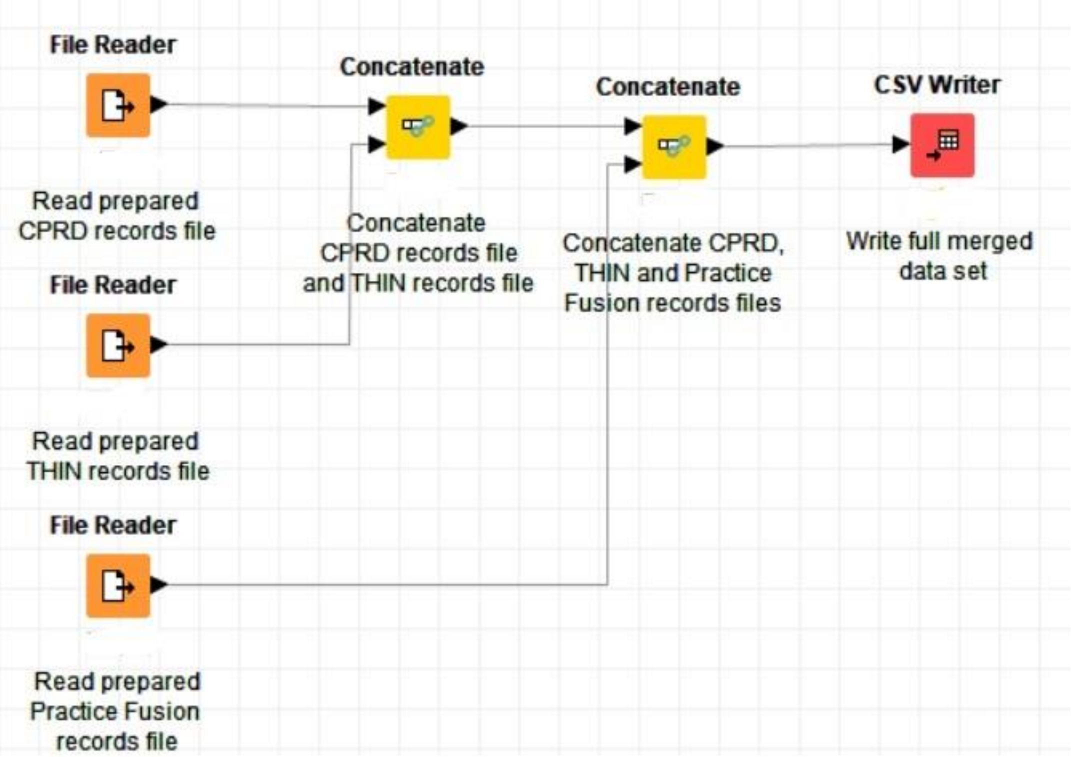

File Reader
Read prepared CPRD records file
File Reader
Read prepared THIN records file
File Reader
Read prepared Practice Fusion records file
Concatenate
Concatenate CPRD records file and THIN records file
Concatenate
Concatenate CPRD, THIN and Practice Fusion records files
CSV Writer
Write full merged data set

Figure 5.19 Aggregation of source data sets into single composite data set

5.8 Assignation of CTV3 code significance values

Clinical Terms Version 3 provides a coded thesaurus of clinical terms, structured in a hierarchical classification [220]. Classifications start from a single root node at the first level of the hierarchy, '…..' 'Read thesaurus', which has 18 child nodes, each corresponding to a broad subject heading. Each of these nodes, in turn, has its own child nodes, with the detail of the concept captured by each node increasing as the tree depth increases. New nodes are added to the hierarchy at each new level until the finest concept detail is reached, at the 19th level of the hierarchy. For convenience, the hierarchy levels will be referred to relative to their distance from the root node, i.e. the root node is 'level 1', all the child nodes of the root node is 'level 2', and so on.
An example of how codes are placed in the hierarchy is shown in figure n. At the top of the hierarchy, at 'level 1', is a placeholder code, '…..' Read thesaurus; all codes below this are members of the CTV3 hierarchy. ….. has 18 child codes, each at 'level 2'. For clarity, only one code is shown: XaBVJ 'Clinical findings', still a very broad subject heading which has no useful clinical significance. XaBVJ in turn has 7 child nodes at level 3, one of which is shown: A…. 'Infective disorder'. Again a broad subject heading but the granularity of the detail in this code is beginning to demonstrate some clinical significance. Below this code, at level 4, are 54 new, more granular, codes, one of which is shown: X70Gv 'Bacterial disease', which also has 54 child codes. Two of these child codes are shown. Following the tree down further, there are three terminal nodes: A3By1, X100G and H0608. Each of these is a precise concept describing a particular condition. None have child codes. A3By1 and H0608 are at 'level 7' of the hierarchy; X100G is at 'level 6'.
A history of the development of the Read Codes is given by Benson [220]. Robinson et al give detailed description of the structure of the various iterations of the Read Codes [221].

CTV3 code granularity

As noted in the section on assigning significance flags to individual CTV3 codes, the CTV3 codes are in a tree structure with 19 levels: level 1, the root node '…..', is merely a place-holder code indicating that these codes are CTV3 codes. Below this root node, at level 2, are 17 high-level codes indicating the broad category of codes that their descendant codes fit into, e.g. '9….' Administration, '0….' Occupations or 'XaBVJ' Clinical findings. Each code can have 0, 1 or many child codes:

CTV3 Level	Number of new codes introduced at this level	Mean number of child codes for codes at this level	Ratio of significant (flag value 1 or 2):not significant (flag value 0) codes
1	1	17	0/1: 0
2	17	25.4	0/17: 0
3	432	7.1	15/417: 0.04
4	3082	4.4	351/2731 : 0.13
5	13468	3.7	3572/3041 : 1.2
6	49825	1.1	10226/39599 : 0.3
7	56788	0.8	19880/36908 : 0.5
8	46332	0.7	21413/24919 : 0.9
9	34340	0.6	15896/18444 : 0.9
10	22092	0.7	10162/11930 : 0.9
11	15632	0.6	5787/9845 : 0.6
12	9165	0.6	2645/6520 : 0.4
13	5686	0.3	854/4832 : 0.2
14	1580	0.2	246/1334 : 0.2
15	345	0.2	70/275 : 0.3
16	56	0.1	0/56 : 0
17	7	0.4	0/7 : 0
18	3	0.7	0/3 : 0
19	2	0	0/2 : 0
20	0	-	-

Table 5.11 Number of CTV3 new codes at each level of the CTV3 hierarchy.

It can be seen from Table 4.7 that level 15 of the CTV3 hierarchy is the deepest level at which new clinically-significant codes are introduced; below this level no new significant codes are added. Levels 8, 9 and 10 have the highest ratio of significant to non-significant codes. Level 8 has the highest absolute number of new significant codes.

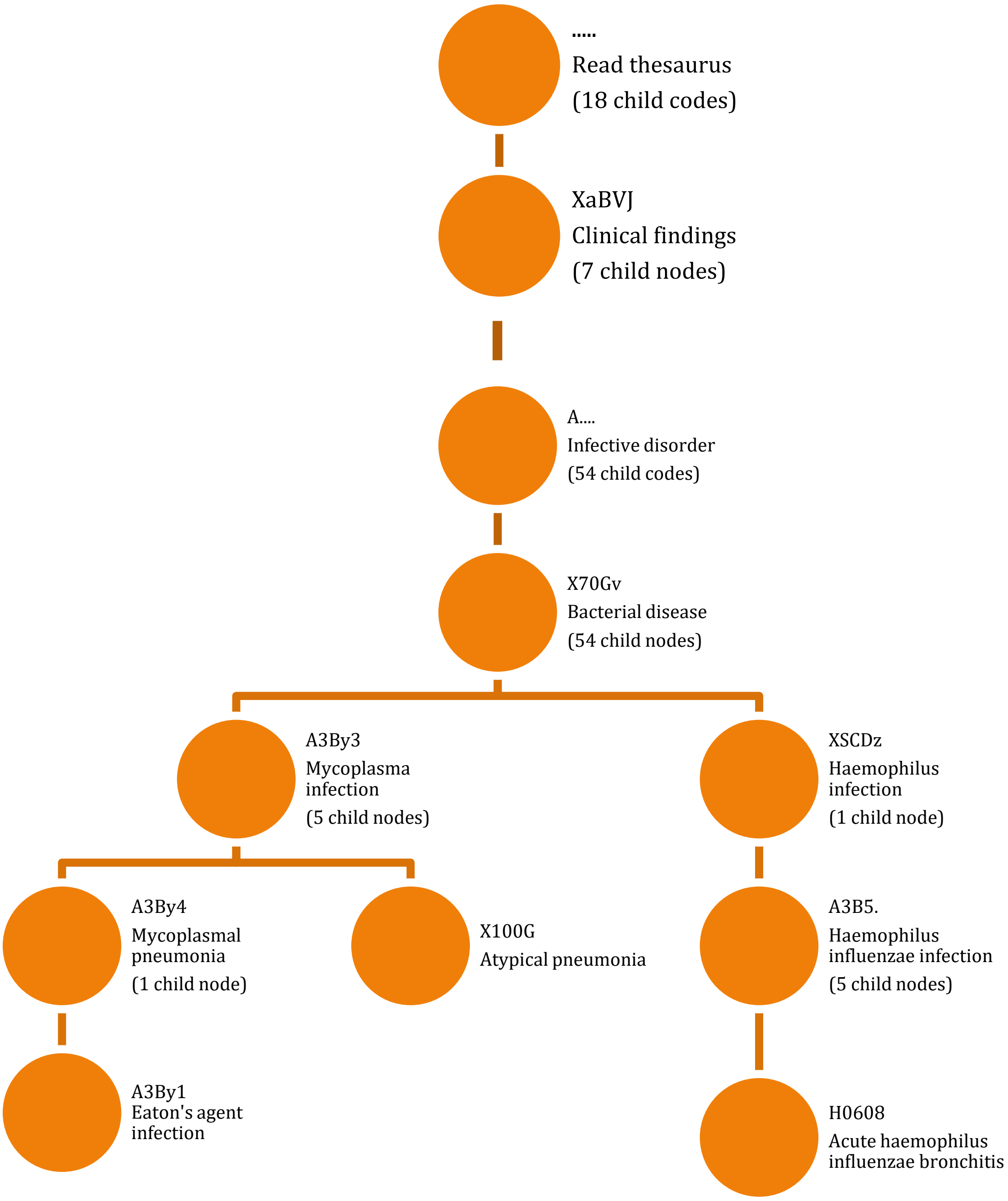


Figure 5.20 An example of codes in the CTV3 hierarchy

Figure 5.20 shows a small extract from the CTV3 hierarchy, illustrating the relationship between codes relating to bacterial chest infections. This Figure shows that Acute haemophilus influenza bronchitis is a Haemophilus influenzae infection; there are four other codes that are child codes of Haemophilus influenzae infection. Haemophilis influenzae infection itself is a Haemophilus infection, which in turn is a Bacterial disease. Bacterial disease is an Infective disorder, a child code of Clinical findings, a category coming under the root node of the CTV3 tree.

The composite data set holds all events that have been recorded for individuals. Each of these codes has a defined meaning – its “concept” - and significance, but this significance may only be for the duration of a particular circumstance, for example recording an episode of illness, for recording symptoms or for recording administrative events. The work described in this report focusses on events of clinical significance, in particular symptoms and diagnoses of conditions that form part of an individual’s clinical history. Events that are of no significance for this work need to be flagged as such and perhaps removed from the data set prior to further analysis of the composite data set.

Many codes have no or trivial clinical information, would add to processing time and/or memory requirements, would add noise to the information contained in clinically-significant codes. Dividing the clinically significant codes into two groups gives the potential for future work in which ‘most likely next condition’ can be predicted, in which case need to know what are the codes that could be predicted. However, in this work, the two groups of significant codes are not distinguished between.
Examples of codes that have little useful clinical information are code XaE42 ‘Medical records review’, or 932... ‘A4 records folder’. Other codes contain more obviously relevant information, such as A796. ‘Parvovirus infection’ or B2211 ‘Malignant neoplasm of hilus of lung’.

Codes were initially to be divided into two groups, ‘significant’ and ‘not significant’, with those codes determined to contain no useful clinical information assigned to the ‘not significant’ group. However, following early exploratory work on the data and consideration for future work it was decided to categorise codes into one of three groups: ‘Administrative’, ‘Symptom’ or ‘Condition’. These were broad definitions with Administrative codes including such events as the performing of screening tests, which do not of themselves indicate the presence or absence of a condition, or information topic headings where the information was not available or otherwise recorded, e.g. XaKTj ‘Abstinence history’ or XaBVJ ‘Clinical findings’. Codes were assigned a flag value according to the expected significance of the code: 0 for administration codes, 1 for symptom codes, 2 for condition codes.

Decision making process for assigning flags:
Three categories: ‘administration’, ‘symptom’ and ‘condition’. More fully, ‘administration and management’, ‘symptom, sign, complaint or procedure or physical attribute’ (i.e. the manifestation of an underlying condition) and ‘diagnosis, disorder or condition’ (i.e. the underlying cause of an illness or complaint).

Some codes were easy to assign:
A70z0 ‘Hepatitis C’ - an infectious disease caused by a viral infection, assigned a flag value of 2.
N094P ‘Ankle joint pain’ could have a number of causes, assigned a flag value of 1.
932.. ‘A4 record folder’ – a purely administrative piece of information containing no useful clinical information, assigned a flag value of 0.
Other decisions were more difficult:
X77BL ‘Tachycardia’ – can be caused by specific conditions (e.g. hyperthyroidism, anxiety) or be a condition of itself. A decision was made to assign a flag value of 2 but there are arguments that it should be assigned a value of 1. Note that for in this work it was ultimately decided to assign equal predictive weight to flag 1 codes and flag 2 codes, so in this case choosing to assign the alternative flag value would have had no influence on the outcome of the prediction.
Nodes that are indicative of the presence of a disorder but of themselves contain little further information, for example X0003 ‘Disorders’ – does suggest that a condition is present but that the code at this level is insufficiently granular to provide any meaningful information about the condition. For nodes with this minimum level of information, it was decided to assign a flag value of 0.

With 16,258 CTV3 codes in the composite data set (and 258,854 codes in the complete CTV3 code set) it was a large task to inspect each code and assign a significance flag value to each code. Advantage was therefore taken of the CTV3 table structure, with each code (bar the root node, '…..') having a parent code and 0:many child nodes. As a starting point for significance code assignation, each code passed its significance value to its child nodes, starting at the root node at level 1 and propagating down through each level in turn.

The highest level (level 1) of the CTV3 tree, the root node '…..', was assigned level 0, since it was a code with no clinical significance. All codes with the root node as their parent were assigned the same significance as the root node. There were 17 codes with the root node as their parent, i.e. at 'level 2'. Each of these codes was inspected to ensure that the significance level assigned was appropriate, with any changes being made manually as appropriate. With no specific clinical information being present in any of the 17 codes at this level, each code retained its significance value of 0.
Codes at level 3 were assigned the significance level of their parent codes. There were 432 codes introduced to the hierarchy at this level; all 432 codes were automatically assigned a significance value of 0, (inheriting the value of their parent code). On inspection of these codes, 15 codes were determined to have a clinical significance value of 1 and so had their significance flag value manually set to 1. The remained 417 level 3 codes retained their significance value of 0 inherited from their parent level 2 codes.

Similarly, level 4 codes inherited the significance values of their parent codes, 0 or 1 as appropriate. There were 3082 codes introduced at this level, 41 of which inherited a flag value of 1 from their parent codes, 3041 inheriting a value of 0. Again, each was inspected and the significance level adjusted as appropriate: there were now 65 codes assigned a significance of 2, 290 a level of 1 and 2727 a level of 0.
At level 5, there were 13,468 CTV3 codes making their first appearance. This quantity of codes was of a similar order to the total number of unique codes in the merged data set (22,764 codes, including codes in the data set plus ancestor codes) and so at this point it was decided only to inspect those codes that were present in the merged data set. Prior to doing this, after manual inspection and correction of level 4 codes, the

automatic propagation was continued until all CTV3 codes had been assigned a flag value.
Once all codes in the CTV3 tree had had significance codes assigned, all codes that were present in the composite data set were inspected and the significance assigned to each of those codes accepted or changed as appropriate. The inspection of the significance code level was expedited by this process: all codes from each significance level group (0, 1 or 2) could be inspected as a batch, with codes assigned an incorrect significance level being corrected.

Flag values for codes that were present in the merged data set were then inspected and corrected as necessary. Note that the presence of the automatically-assigned flags increased the efficiency with which this could be done: codes with one flag value were grouped together and quickly inspected to find any codes that belonged to a different group.

The corrected flag values for the codes in the composite data set were then fed back into the automatic flag value assignation program, allowing for improved automatic flag value assignation for codes that were not present in the merged data set. However these codes were superfluous to the analysis of the merged data set and so were not subject to further inspection and correction.

5.8.1 Results

(1) Automatic flag assignation followed by manual correction for levels 2 to 4; levels 5 to 18 automatic flag assignation only.

Code level	Flag 0 code count	Flag 1 code count	Flag 2 code count	Total codes at this level	Cumulative code count
1	1	0	0	1	1
2	17	0	0	17	18
3	432	0	0	432	450
4	3041	41	0	3082	3532
5	9896	2121	1451	13468	17000
6	39599	5045	5181	49825	66825
7	36908	8572	11308	56788	123613
8	24919	8456	12957	46332	169945
9	18444	5739	10157	34340	204285
10	11930	2664	7498	22092	226377
11	9845	1104	4683	15632	242009
12	6520	378	2267	9165	251174
13	4832	104	750	5686	256860
14	1334	62	184	1580	258440
15	275	20	50	345	258785
16	56	0	0	56	258841
17	7	0	0	7	258848
18	3	0	0	3	258851
19	2	0	0	2	258853
20	0	0	0	0	258853

Table 5.12 Number of new codes introduced at each level of the CTV3 hierarchy

After manually inspecting and correcting the flags assigned to the codes present in the merged data set, the automatic assignation program was run again, with the following results:

Inspect discrepancies between in-program corrections and manual look-up table, removing codes changes from the program or making changes to the look-up table as appropriate. Then need to re-run the program, again checking numbers before and after in-program correction

Once these flag values had been assigned to the CTV3 codes (codes in the data set and their ancestor codes were verified by manual inspection and correction, CTV3 codes not in the composite data set had flag values automatically assigned but not verified they do not appear in our data set) the flag values assigned to CTV3 codes were frozen and available for later use.

CTV3 codes not in our data set were also assigned significance flag values, but these were not manually verified. However, by automatically assigning these values it was possible to see how the code set divided up into administration, symptom and diagnosis codes.

The R code used to automatically assign the significance flags is given in Appendix 4. The significance flags assigned for a sample of CTV3 code are shown in Appendix 5.

5.9 Recommendations for good practice for merging data sets

The work required to create a single composite data set from three source data sets, combining fields with common meanings across the data sets and syntactically and semantically consistent data within those fields, has been described. For efficient implementation of this process, some recommendations are given:

1. Utilise existing mappings as much as possible – TRUD, NIH, independents
2. Do not spend time trying to get the automatic mappings as complete as possible: missing mappings can be defined manually
3. Be concerned only with codes that are present in the source data set and will be used in the final analysis
4. Manual mappings: Most can be mapped by searching for key words in the descriptions. But be aware of minor differences in spellings (UK English vs US English, singular vs plural) and completely different words to describe the same conditions (Common English terminology vs Latin-based terminology; disease-centred vs body part-centred ;).

5. Take care not to introduce artefacts into the data (e.g. KNIME converts a string of “00000000” to a date with a year of “2” so convert source data codes for missing values to appropriate values in your system.

6. Related coding systems may appear similar but have subtle differences and must be treated as separate systems. For example, Read Codes Version 2 is an ancestor system to Clinical Terms Version 3 and the two systems have many codes in common, but some codes are different and, unlike in Read Codes Version 2, it cannot be assumed that the first 4 characters of a CTV3 code is the parent code of the full five character code.

7. Coding systems that allow for synonyms within single concepts may be amenable to simple editing of codes to revert all synonyms to the base concept. For example for CTV3, the first 5 characters only are significant. Characters 6 and 7, if present, allow for synonyms for the same concept. The base concept as defined by the first 5 bytes is not changed by any characters in characters 6 and 7 and so these least significant characters can be dropped from the mapping (however, the synonyms may assist in manual mapping if a match between the first table (e.g. ICD9-CM codes) and the CTV3 table base concept description cannot be found).

8. Care should be taken with coding systems that use the complete ASCII character set and allow for character strings that comprise only numerals or numerals with a final decimal point. These strings may get converted automatically by some programs (e.g. Microsoft Excel) into numbers. For example, a trailing decimal point (e.g. E251.) can get dropped if the code is otherwise all-numeric (e.g. 1972. -> 1972) and the code is treated as a numeric value. It is therefore recommended to treat all codes as character strings at all times.

5.10 Results from this chapter

This chapter has described the process of creating a single composite data set from disparate source data sets, and the challenges involved in that process. Three data sets were merged and, where necessary, mapped from one coding terminology to the target terminology, to form a single aggregate data set. In this process, the fields common to each of the source data sets were determined and used for the minimum aggregate data set. The fields common across all three data sources are:

Patient ID; Gender; Year of birth; Year of death; Clinical event code; Clinical event start year; Acute or chronic; Allergy type; Allergy start year; Allergy name; Smoking status code; Smoking status date; Immunisation code; Immunisation date; Therapy date; Therapy code; Therapy dose; Therapy quantity; Therapy length.

Event codes were mapped to CTV3 from Read Codes Version 2 (for the UK sourced data) or from ICD9 (from the US sourced data). An existing mapping table obtained from NHS TRUD was used to map Read Codes Version 2 codes to CTV3 codes. This mapping was complete. For mapping event codes in the US data set, coded in ICD9, no existing mapping table was found and so a more complex process was required, a two-stage mapping process via an intermediate coding system which mapped a high proportion of the codes required to be mapped, provided that the mapping tables required for each of the intermediate mapping steps (into and out from SNOMED CT) existed. The automatic mapping process produce a list of failed mappings and the frequency of occurrence in the source data set of each code that failed to map, which was used as the basis for manual completion of the local ICD-9-CM to CTV3 mapping table.

6 VALIDATION AND VERIFICATION OF DATA SET

6.1 Introduction

Chapter 4 outlined how the three source data sets were combined into a single composite data set, with semantically equivalent data fields, a common coding system, a simplified record of smoking and alcohol consumption, and a field indicating the source data set for each record. This chapter describes how data items in the composite data were cleaned and checked for validity. The information in the source data sets and in the composite data set were verified against each other and against information on the general population to determine whether the source data were representative of the general population or whether the data needed to be weighted to compensate for being unrepresentative.

Areas for validation and verification in this chapter include:

(i) Data cleaning and error checking;

(ii) Demographic comparison of the source data sets by age and by gender;

(iii) Comparison of condition prevalences in the composite data set versus condition prevalences in the literature;

(iv) Calculation of risk factors from the composite data and comparison with known risk factors for particular conditions;

6.2 Data cleaning

Maletic and Marcus [222] give a useful introduction to data cleansing principles, though they do not deal with problems with health data specifically. They give some techniques for data cleaning, and refer to studies by Orr [223] and Redman [224] who states that "error rates in the data acquisition phase are typically around 5% or more." There is no reason to assume that the source data used in this work is 'clean': validity of data in records is not guaranteed to have been imposed at the time of data acquisition or entry, and so should be checked before use for implausible or outlying values. Further, Maletic and Marcus note that in any composite data set created from merging several sources the issue of duplicate records must be addressed and duplicates removed where possible. Data should be cleaned as early as possible and certainly before use. Note that some data cleansing had been performed as part of the data set merging described in Chapter 3, including removal of fields that were not common to all three source data sets and correction of obviously incorrect dates, such as date of death preceding other event dates.

The basic checklist used for cleaning the data in the composite data set was:

a. Data types should be valid for each field;
b. Data values must be within an allowed or plausible range for numerical values (e.g. age to be between 0 and 130 years inclusive) or within a set of allowed values (e.g. male/female/unknown);
c. Mandatory field values must be present;
d. Cross-field validation to be performed (e.g. the date of birth in a record must be earlier than all other recorded events);
e. Data must have consistent and appropriate units (e.g. all weights to be in kg) where units are recorded with data values;
f. Duplicate records to be identified and removed.

a. Data types should be valid for each field

Data type consistency was checked. All fields were either numeric or text fields as appropriate. Data type consistency had been enforced by default by the KNIME workflow for the data set merging stage in Chapter 4.

b. Data values to be within an allowed or plausible range.

Values for age were checked and all found to be in the range 0 to 130 years. All records had a gender of either male or female.

c. Mandatory field values must be present;

(i) Demographic information. Each record was checked for presence of year of birth and gender information. All records were found to be complete for gender and year of birth information. No other identifying information, such as name, address, telephone number, zip or postal code was included in any of the source data sets, apart from the inclusion of home state information in the US-sourced data.

(ii) Clinical events information. Each record contained a wide range of number of events, from 1 event to 1481 events. Irrespective of the number of events in each record, there is no guarantee that all events for any individual are recorded, either because a clinician neglected to record the condition or felt it was not worthy of recording, or the individual did not feel that the severity of the condition warranted a consultation, or for reasons of data loss or non-entry. Given the difficulty in deciding between an event that never happened versus an event that happened but was not recorded, it was decided to use the events records as they were with no attempt made to impute any ‘missing’ events records, which would be theoretically possible from a history of prescriptions or other treatments that may be condition-specific, though still likely to be incomplete.

d. Cross-field validation

Records were checked to ensure that event dates were the same year or later than the year of birth.

e. Data must have consistent and appropriate units

Unit checking and conversion had been carried out in the data set merging stage described in Chapter 4. All dates had been converted to the lowest common unit, years.

Event codes had all been mapped to a single common coding system, Clinical Terms Version 3.

f. Duplicate records to be identified and removed

It is possible that multiple records derived from the same individual are present in the data. As many duplicate records as possible should be removed in order to avoid giving undue weight to those duplicate records in later analysis. Given the limited demographic information available for each individual, identifying the duplicate records is no simple task. A simple set of rules was established in order to identify potential duplicates. These rules were:

(i) records should be from the same country source (i.e. it is possible for a record(s) derived from the same individual to be present in each of the UK data sources but not in both a UK source and the US source;

(ii) year of birth and gender in each record must match;

(iii) the first twenty events in the records must match.

Testing for duplicates was carried out after data set merging to allow for cross-source duplicate testing.

Analysis of the records using the above criteria showed that there were no duplicates within the US data set, no duplicates within the UK THIN data set, no duplicates across the UK THIN and CPRD data sets but six potential duplicates within the UK CPRD data set, in two sets of three records. These were all for minors – three aged five years at the time of data capture (three males), and three aged six years at the time of data capture (two females, one male). The duplicate records with the fewer number of events were removed, on the assumption that the records with more recorded events presented a more recent and fuller list of the events for the individual.

Following inspection of the data it was concluded that after the removal of potential duplicate records, the composite data set was clean and suitable for verification of its contents against known attributes of the general population.

6.3 Demographic comparison

There were two obvious pieces of demographic information on which to compare the source data sets: age and gender. The two UK source data sets were compared to each other and the combined UK data compared to the US data set.

6.3.1 Age

Age is recorded in all source data sets in years. The distribution of age in each of the source data sets is shown in Figure 5.1. As can be seen from the histograms, the US data set was both left-censored and right-censored for age. This finding had not been revealed by the earlier simple testing for presence of age (or, equivalently, date of birth) information in the data sets. This censoring is understood to be due to the HIPAA regulations [123] governing the release of data derived from individuals, even if the data are de-identified. The HIPAA regulations include a number of rules governing the release of data. Section 164.514(b) of the HIPAA regulations defines the ‘Safe Harbor’ method for de-identification by giving 18 rules for removal or de-granularisation of data items. Rule 3 states that ‘All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older’. Practice Fusion also chose to exclude records derived from those under 18 years of age. It can be seen from Rule 3 that HIPAA also recommend releasing only the year of birth, reflected in the detail of the data released by Practice Fusion It was necessary to filter the UK-sourced records to mimic the distribution of the US-sourced data; this was done by removing records with an age of 18 years or less or greater than 90 years. The number of records removed is shown in Table 6.1.

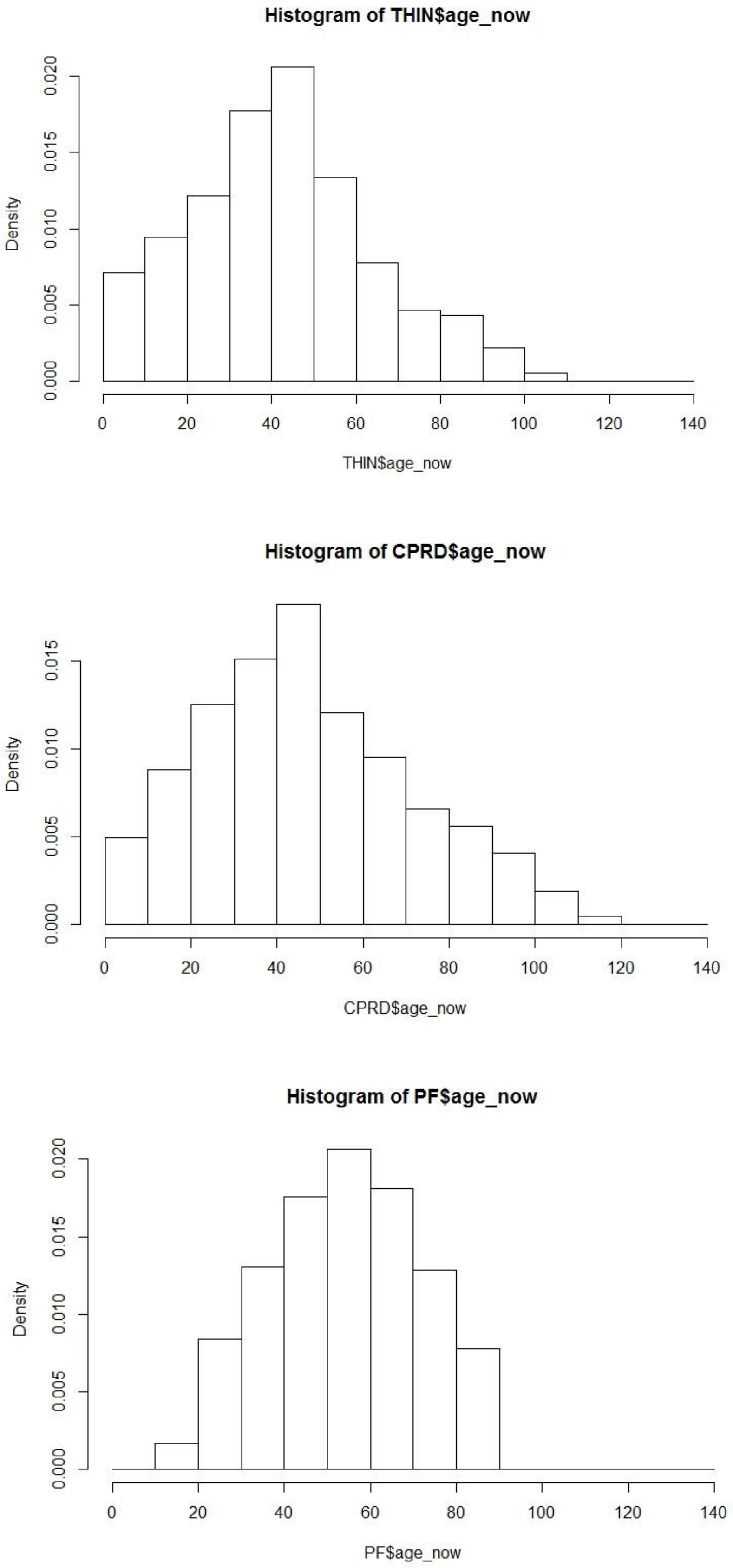


Figure 6.1 Age distribution from the three source data sets

Source data set	Number of records in source data set	Number of records after age filtering	Number of records removed
CPRD	4711	3908	803
THIN	3674	3070	604
All UK	8385	6978	1407
Practice Fusion	14740	14740	0
All UK + US	23125	21718	1407

Table 6.1 Number of records in the source data sets, before and after filtering by age

The filtered data set, with all records from individuals aged 17 years or younger or 90 years or older removed, was used for all subsequent analysis

6.3.2 Age distribution comparison

The age distributions in the age-filtered data sets were compared using a two-sample Kolgomorov-Smirnov test (see e.g. [225]), producing the Kolgomorov-Smirnov 'D' statistic, where $D_n = \sup_x |F_n(x) - F(x)|$

The 'ks.test' function from the R 'stats' package [226] is used to compare cumulative age frequencies in the CPRD data to cumulative age frequencies in the THIN data.

D = 0.429, p = 0.5752

Conclusion: the two UK data sets do not have significantly different age distributions ($p > 0.05$) and so can be merged.

Comparing the merged UK data set to the Practice Fusion data set:

D = 0.71429, p-value = 0.05303

Conclusion: the UK data set and the US data set do not have significantly different age distributions ($p > 0.05$) and so were merged without weighting.

6.3.3 Gender distribution

The gender balance between data sets was compared, using Pearson's chi-squared test (Pearson, 1900)

For CPRD vs THIN:

CPRD: 1980 Male, 1928 Female

THIN 1501 male, 1569 female

The chi-squared test was performed using the ‘chisq.test’ function from the base R ‘stats’ package [226]:

```
chisq.test (as.table(rbind(c(1980, 1928), c(1501,1569))))
            Pearson's Chi-squared test with Yates' continuity correction
data:  as.table(rbind(c(1980, 1928), c(1501, 1569)))
X-squared = 2.0911, df = 1, p-value = 0.1482
```

Conclusion: the two UK data sets do not have significantly different gender distribution and so can be merged

For combined UK vs US:

UK: 3481 male, 3497 female

US: 7225 male, 7505 female

```
chisq.test (as.table (rbind(c(3481, 3497), c(7225,7505))))
            Pearson's Chi-squared test with Yates' continuity correction
data:  as.table(rbind(c(3481, 3497), c(7225, 7505)))
X-squared = 1.2901, df = 1, p-value = 0.256
```

Conclusion: the UK data set and the US data set do not have significantly different gender distribution and so can be merged.

Overall conclusion: there is no significant difference between the source data sets that would preclude the merging of the source data sets into a single composite data set for further analysis.

6.4 Condition prevalences: composite data set vs population

In order to validate the combined data set as a source of data for further work, the combined data set was assessed to check that (1) that the most common conditions in the data set matched the most common conditions in the general population; (2) various conditions present in patients in the data set had similar prevalence to the same conditions present in the general population, as found in published studies; (3) that risk factors associated with particular conditions could be established and compared with risk factors published in the literature.

6.5 Most frequent conditions.

The first exercise was to discover the most frequent conditions in the composite data set and compare this list to an external study listing the most prevalent conditions in the general population.

A Practice Fusion study on most prevalent conditions in 2016 was used as the basis for this exercise (https://www.practicefusion.com/blog/25-most-common-diagnoses/)

Note that this is prevalence of condition – no account is taken of relative burden of each condition, or contribution to mortality. It is merely prevalence of condition as recorded in their patient records system. Note also that this data is derived from patients in the USA only.

The top 10 most frequent conditions in 2016 listed in the study were:

1. Hypertension
2. Hyperlipidaemia
3. Diabetes
4. Back pain
5. Anxiety
6. Obesity
7. Allergic rhinitis
8. Reflux esophagitis
9. Respiratory problems
10. Hypothyroidism

6.6 Semantic mapping.

The composite data set with events coded in a single coding system and data cleaned, with clear duplicates removed, was now ready for validation and analysis. However, from inspection of the event codes, it was clear that a single medical condition could be indicated by multiple codes. For example, the presence of type 2 Diabetes could be indicated by X40J5 ‘Type II diabetes mellitus’, XaIfI ‘Type II diabetes on diet only’, XaIfG ‘Type II diabetes on insulin’, Xa2hA ‘Dietary advice for type II diabetes’, and/or C1011 ‘Type 2 diabetes mellitus with ketoacidosis’ and a number of other codes. Other sets of codes, or ‘codelists’, exist for other conditions.

At the time of commencement of this work no standard sets of codelists were found to be available. It was therefore necessary to generate lists of CTV3 codes for each condition of interest, each list containing the set of codes that indicated the presence of the condition. Some recent work has been published that indicates the desirability of standardising such codelists in publically available repositories and a project is underway at the University of Manchester Institute of Population Health to acquire and disseminate such codelists (see codelists.org) and to which repository the codelists used in this work will be offered.

However, this project has yet to reach the point where they are able to make codelists publically available [227]. Other recent work on codelists has been described by Watson et al [228], who although working towards automated codelist generation, still rely in large part on manual processes.

Codelists for use in this project were created by a manual process of searching code descriptions and making a judgment as to whether the description of the code was indicative of the presence of the condition of interest. Skills acquired in the manual mapping of ICD9-CM codes to CTV3 codes were beneficial, in particular knowledge of synonyms for particular conditions and knowledge of the CTV3 hierarchy which enabled rapid searching for related codes.

In the list of example given above for Type 2 Diabetes, code Xa2hA ‘Dietary advice for type II diabetes’ was considered to be a strong enough indicator of the presence of the condition for it to be included in the code list, although it was not a formal record of a diagnosis of the condition. It should be noted that CTV3 codes exist in a hierarchy and so any ambiguity in a code description could be resolved by inspection of the code's parent code or child code(s).

With the unknown number of event codes that could indicate the presence of a condition, the variability in event code descriptions, the possibility that event codes could appear in disparate regions of the CTV3 hierarchy, and the large number of unique codes in the CTV3 hierarchy, it was challenging to produce codelists that were complete. A technique was developed to check that no significant event codes had been omitted from the codelist in which they should be present.

To check that no significant codes had been omitted from codelists, an R program was developed that implemented a decision tree algorithm on the composite set of records, following the process shown in

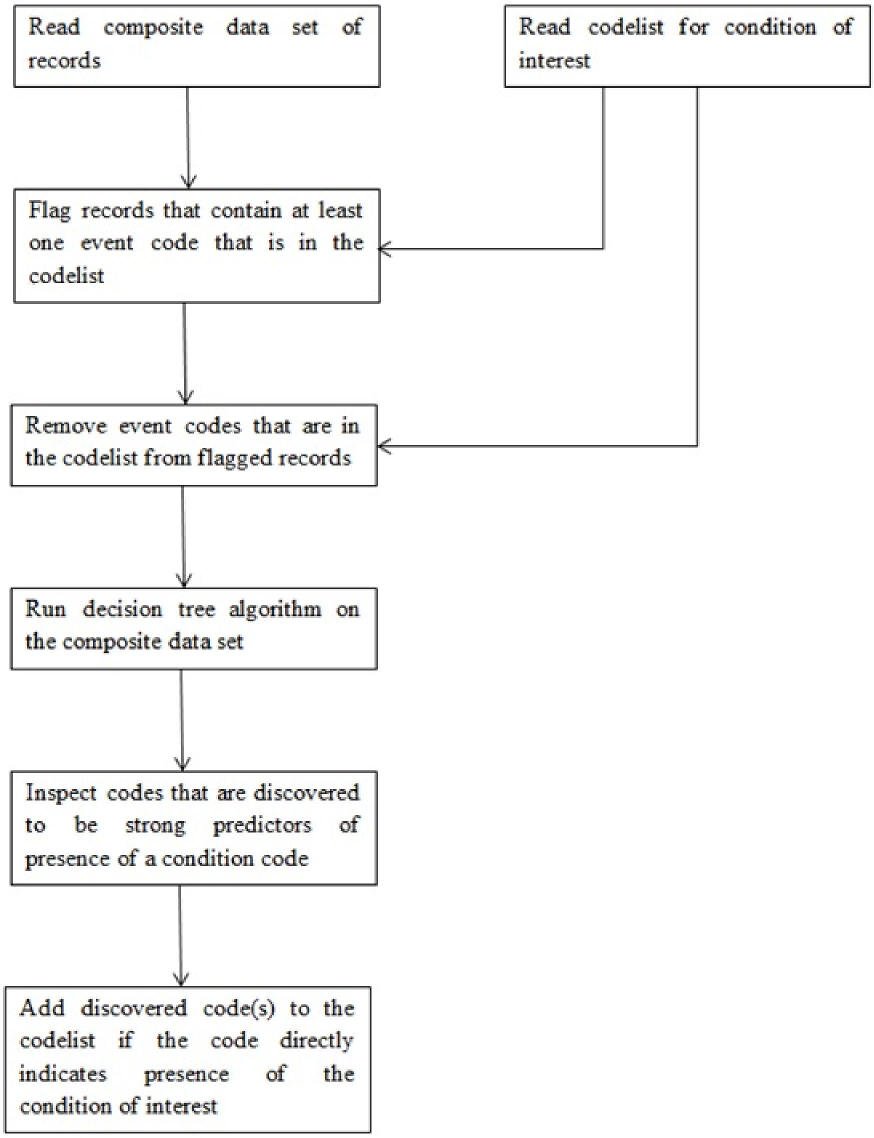

Read composite data set of records
Read codelist for condition of interest
Flag records that contain at least one event code that is in the codelist
Remove event codes that are in the codelist from flagged records
Run decision tree algorithm on the composite data set
Inspect codes that are discovered to be strong predictors of presence of a condition code
Add discovered code(s) to the codelist if the code directly indicates presence of the condition of interest

Figure 6.2.

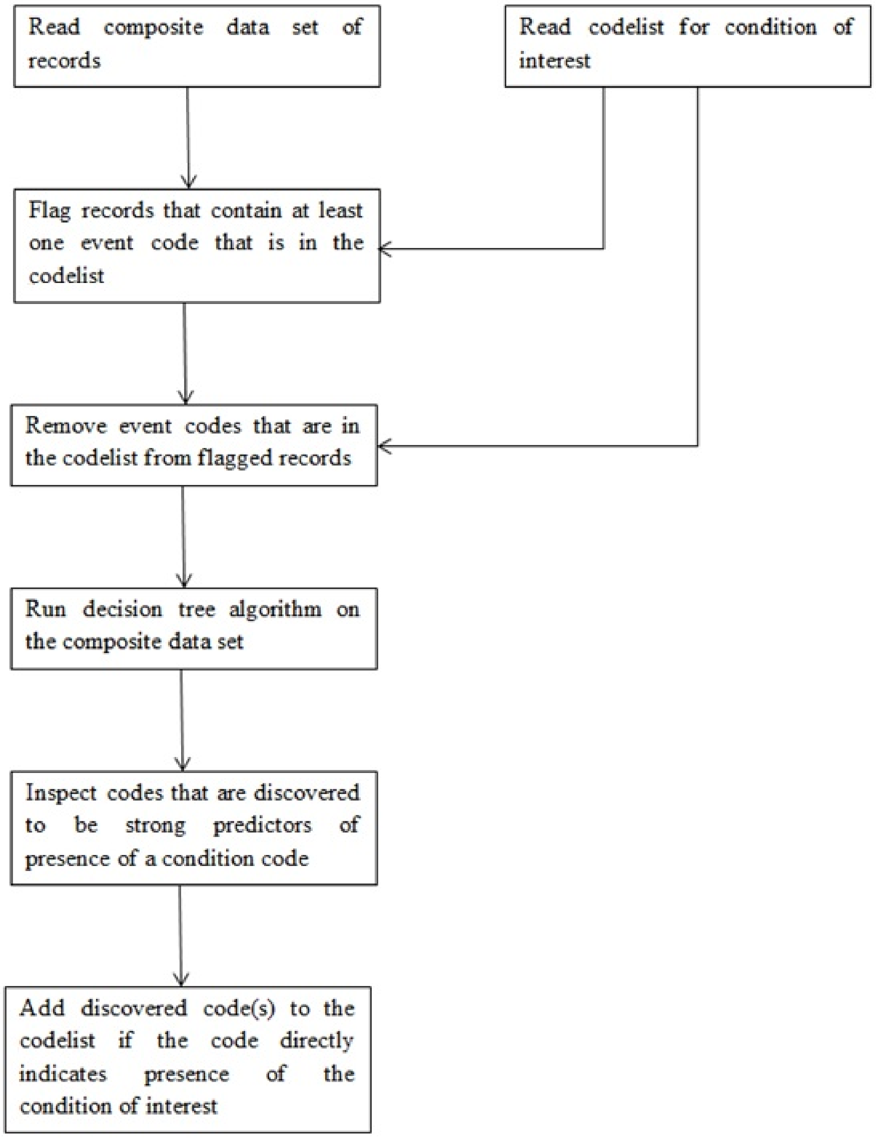

Read composite data set of records
Read codelist for condition of interest
Flag records that contain at least one event code that is in the codelist
Remove event codes that are in the codelist from flagged records
Run decision tree algorithm on the composite data set
Inspect codes that are discovered to be strong predictors of presence of a condition code
Add discovered code(s) to the codelist if the code directly indicates presence of the condition of interest

Figure 6.2 Process for discovering candidate codes for addition to codelists

For each condition, codes in the draft codelist were removed from the set of records, with those records having a ‘condition’ code present, i.e. a code that was in the codelist being flagged as a ‘condition positive’ record. The decision tree algorithm was then trained to use the remaining codes in all records to predict the ‘condition positive’ records. Codes that were strong predictors for condition positive records were inspected and those codes that themselves recorded the presence of the condition were added to the relevant codelist.

As an example, the codelist for type 2 diabetes is used. All codelists can be found in Appendix 2. This codelist has 32 event codes, each of which, if found in a record, is taken to be indicative of the presence of the condition ‘type 2 diabetes’. Removing two codes from the codelist and running the decision tree program on the composite data set produced the following results:

Codes temporarily removed from the data set:

XaOPu ‘Latent autoimmune diabetes mellitus in adult’, X40J5 ‘Type II diabetes mellitus’ and C1092 ‘Type II diabetes mellitus with neurological complications’

Running the partition tree algorithm produces a set of codes as being used to split the tree and should be inspected to see if they should be included in the codelist set for the condition under consideration. These codes produced in the example for type 2 diabetes with some event codes omitted are shown in Table 5.2.

CTV3 code	CTV3 code description
14L..	H/O: drug allergy
29H1.	O/E - vibration sense normal
66AE.	Feet examination (& diabetic)
X40J5	Type II diabetes mellitus
C100.	Diabetes mellitus with no mention of complication
XE10G	Diabetes mellitus with renal manifestation
Cyu8D	[X]Other hyperlipidaemia
XE2QC	Impacted wax
G2101	Malignant hypertensive heart disease with CCF
1361.	Teetotaller
Ub1na	Ex-smoker
XE15k	Diabetic polyneuropathy
Xa8Hh	Thrombocytopenic disorder
XE10G	Diabetes mellitus with renal manifestation
C1082	Type I diabetes mellitus with neurological complications
XE11U	Mixed hyperlipidaemia
66AP.	Diabetes: practice programme
XM06e	Dizziness and giddiness
G200.	Malignant essential hypertension
XE1EZ	Shoulder joint pain
24E1.	O/E -R.-leg pulses all present
24F1.	O/E - L.leg pulses all present
XaJvF	O/E - Right dorsalis pedis normal
XaJvH	O/E - left dorsalis pedis normal
22K4.	Body mass index index 25-29 - overweight
XE1hO	O/E - peripheral pulses R.-leg
XE1hP	O/E - peripheral pulses L.leg
ZV700	[V]Routine health check-up

Table 6.2 Codes suggested as being missing from a codelist

Note that not all the ‘missing’ codes were discovered, This is because the codelists were built on the complete CTV3 code set, whereas the partition algorithm is run only on the composite code set which contains only a subset of the complete CTV3 code set. Since later analysis could only be performed on the existing composite data set, it was important only to have codelists that were complete for event codes within the composite data set and not also those in the complete CTV3 code set.
Alternatively, the tree can be plotted, as shown in Figure 6.3.

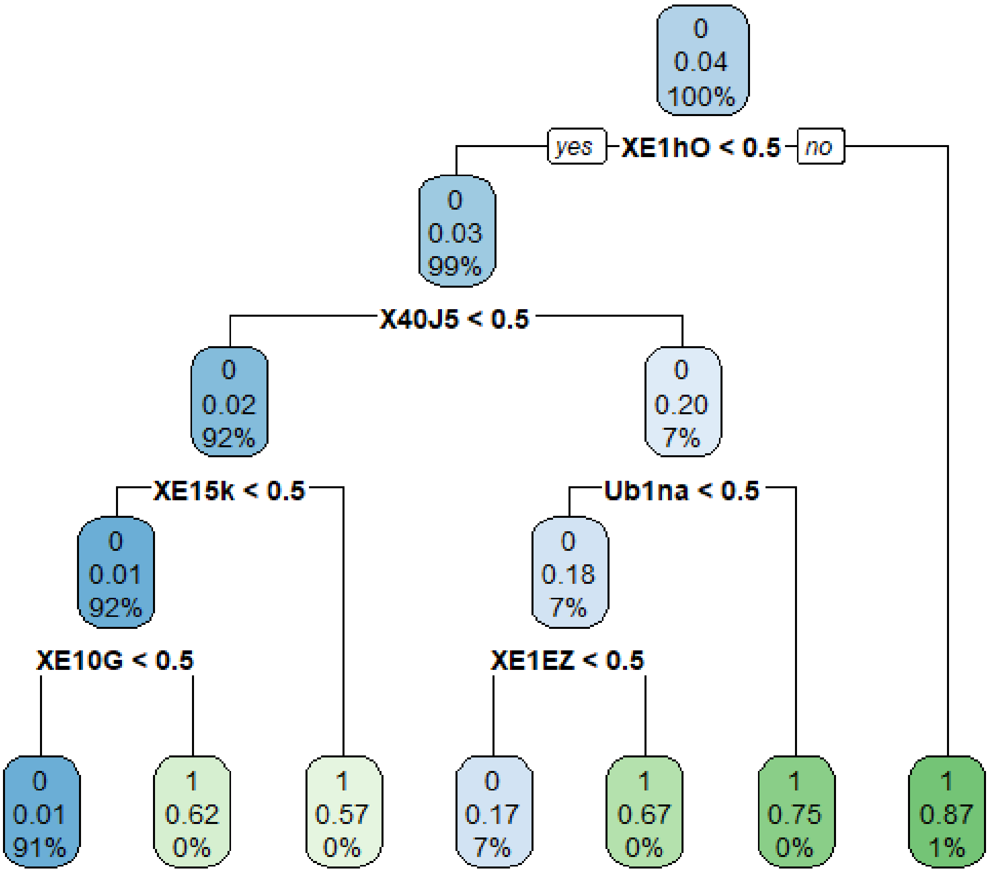


Figure 6.3 Plot of decision tree used to discover candidate codes for addition to codelists

Codelists were built for each of the 'top 10' conditions, with some conditions requiring a large number of codes to identify all occurrences of the condition. For example, the most common condition, hypertension, required 134 codes in its codelist. As a shorter example, codes making up the codelist for hyperlipidaemia are shown in Figure 6.3. Codelists for all the most frequent conditions can be found in Appendix 2.

CTV3 code	CTV3 code description
U60C6	[X]Antihyperlipidaem/antiarterioscl drg caus adv ef ther use
XaJYh	Hyperlipidaemia clinical management plan
Xa2hC	Dietary advice for hyperlipidaemia
Cyu8D	[X]Other hyperlipidaemia
XE13A	Disord lipid metab (& [Fredrick types] or [hyperlipidaemia])
X40Wy	Hyperlipidaemia
X40Vm	Familial combined hyperlipidaemia
XE11U	Mixed hyperlipidaemia
C324.	Hyperlipidaemia NOS
X40XI	Primary combined hyperlipidaemia
X40XO	Secondary combined hyperlipidaemia
C3202	Hyperlipidaemia, group A
C322.	(Mix hyperlipid) or (Fredr lip: [IIb][III]) or (xanthom tub)

Table 6.3 Codes in the codelist for hyperlipidaemia

Once codelists had been created and tested, the full record set could be validated for condition prevalences against condition prevalences in the literature.

Table 5.4 shows these top 10 conditions, their 'top 10' rank and their frequency rank in the composite data set following discovery using the codes in the codelists. Note that the composite data set prevalences shown in this table are raw prevalences, unadjusted for age.

Practice Fusion Prevalence Rank	Condition	Composite data set prevalence (per 100,000)	Composite Table rank	No. of CTV3 codes
1	Hypertension	30169	1	134
2	Hyperlipidaemia	21342	2	13
3	Diabetes	11749	6	73
4	Back pain	16552	3	41
5	Anxiety	11663	7	38
6	Obesity	13310	5	77
7	Allergic rhinitis	14754	4	18
8	Reflux oesophagitis	8832	8	14
9	Respiratory problems	3333	10	329
10	Hypothyroidism	7412	9	56

Table 6.4 Prevalence of the 'Top 10' most common conditions in the population in the composite data set, their rank order of prevalence in the general US population and in the composite data set, and the number of CTV3 codes in the codelists that indicate the presence of the condition

The degree of association between the two ranked lists was tested using Kendall's tau coefficient [229]. Calculation were performed using the on-line Kendall's tau rank correlation calculator at wessa.net [230]

This calculator produced a value for tau of 0.689, with a 2-sided p-value of 0.007. Given this result, the null hypothesis of independence between the two rank orders is rejected.

Investigation is now made of each of these conditions to compare the prevalence found in the composite data set to the prevalence in the general Western population as found in the literature. This serves both to validate the prevalence of the conditions in the composite data set and to give confidence in the codelists that have been built. A summary of the results is shown in Table 6.5. Each condition is then briefly discussed.

	Composite data set prevalence			**Literature prevalence**		**Literature source**
Condition	**Unadjusted Prevalence**	**US Age adjusted prevalence**	**UK Age adjusted prevalence**	**Unadjusted Prevalence**	**Age adjusted prevalence (if given)**	
Hypertension	29467	16786	20431	28000-44000	36000 (Europe), 31000 (“high income” countries)	Wolf-Maier et al [231], Mills [232]
Hyperlipidaemia	22542	13224	15914	39000		WHO [233]
Type 2 diabetes	8614	4748	5933		8700	Centers for Disease Control and Prevention [234]
All diabetes	10733	6201	7552	6000		Diabetes.co.uk [235]
Back pain	16433	12098	12992	12000 - 20000		Meucci et al [236]
Anxiety disorder	13118	11859	12064	2000-13000		Martin [237]
Obesity	7623	5663	6071	23000 (UK), 33800 (USA)		OECD [238]
Allergic rhinitis	15423	13417	13748	~23000		Bauchau and Durham [239]
Gastro-oesaphageal reflux disease	9369	6566	7309	6600-28000		Dent et al [240]
Respiratory problems	3333	2644	2351	4100		Eurostat [241]
Hypothyroidism	7753	5570	4815	600 - 12000		Vanderpump [242]

Table 6.5 Table of crude and age-adjusted prevalence (per 100,000) in composite data set

6.7 Condition prevalences in the data set versus in the literature

Age-adjusted prevalences were calculated for the US and UK derived records in the expectation that some of the data sources in the literature would report age-adjusted prevalence. Age-adjusted prevalence is useful when comparing populations that may have different age distributions. See Naing [243] for an introduction to age adjustment. The method used here is taken from the NIH National Cancer Institute [244] implemented locally in an R program.

Condition prevalences in the literature

This is a more challenging area for validation for several reasons:

a. Getting condition prevalence for general populations from the literature is not straightforward; definition of what is the condition is may not be well-established; the presence of a condition in a record can be indicated by multiple different CTV3 codes. It is therefore necessary, for each condition, to build lists of CTV3 codes that indicate the presence of each condition of interest, as has been previously described.
b. Condition prevalences in the literature are sometime vague, giving a wide range for prevalence value. Prevalences are often for a particular sub-population rather than for a general population.
c. Condition prevalences can be reported as an unadjusted prevalence or as an age-adjusted prevalence (prevalence may be different in different age groups; contribution of different age groups in the overall result is different – e.g. age 31to 40 years has more people than age 91-100 years). In the composite data set, both unadjusted and age-adjusted prevalence is calculated and the appropriate prevalence is compared to the prevalence noted in the literature. However it is not always clear which prevalence is being presented in published work.
d. Condition prevalences may be calculated following testing of a sample group, which may give a different (higher) prevalence than calculating prevalence in a set of records – undiagnosed conditions will not be recorded, diagnosed conditions may be mis-recorded or recorded elsewhere than the primary care record.

Each of the top 10 conditions is now briefly discussed, focussing on the condition prevalence discovered in the composite data set, using the generated codelists, and condition prevalences found in the literature.

6.7.1 Hypertension

The most common condition in the composite data set and in the figures found for the general population, hypertension is estimated to cause around 13 % of deaths globally. Information on the prevalence of hypertension was taken from the work of Wolf-Maier et al (2003) and a meta-analysis by Mills et al (2016).

Figures taken for comparison:

Wolf-Maier et al: USA: 28 % prevalence rate; Europe: 42 % prevalence rate.

Mills et al: globally 31 % prevalence rate; Europe 36 % prevalence rate; ‘high income economies’ 31 % prevalence rate.

Prevalence rates quoted in both studies were age-adjusted.

The unadjusted prevalence of 29.5 % found in the composite data set is similar to the prevalence rates presented by Wolf-Maier and by Mills. However, the age-adjusted prevalence rates in the composite data set of 16.8 % (USA) and 20.4 % (UK) are lower than the age-adjusted figures of Wolf-Maier and of Mills.

These lower figures may be due to under-reporting of hypertension in clinical records, with many cases not being reported to clinicians, or may be due to the codelist for hypertension being incomplete. However, given the composite data set figures are of similar magnitude to the estimates of the general population and they were not the subject of further investigation.

6.7.2 Hyperlipidaemia

Hyperlipidaemia, or raised cholesterol levels, is the leading risk factor in death from cardiovascular disease, the leading cause of death in the United States. Estimates vary as to the prevalence of hyperlipidaemia but reported prevalences are generally higher than that found in the composite data set (22.5 %). The Centers for Disease Control and

Prevention reported (2011) that an estimated 33.5 % of US adults had hyperlipidaemia, a figure not indicated as being age-adjusted.

Given the composite data set figures are of similar magnitude to the estimates of the general population, they were not the subject of further investigation

6.7.3 Type 2 diabetes

There are two main types of diabetes – type 1 and type 2. In the UK, around 90 % of individuals with diabetes have type 2. It is not always clear from the event codes in the records in the composite data set which type of diabetes is being recorded. The prevalence of specifically type 2 diabetes of 8.6 % in the composite data set is close to that described in CDC 2017; the prevalence of both types of diabetes is close to that shown on diabetes.co.uk. It may be that the prevalence discovered from the composite data set is closer to that in the literature because diabetes is less likely to remain unrecorded than some other conditions, or it may be that the event code that indicate diabetes are clearer and so the diabetes codelists are more complete.

Holman et al in 2014 [245] reported a recorded prevalence for type 2 diabetes of 5.7 % for adults in the UK, with an estimated actual prevalence (diagnosed and undiagnosed) of 8.9 %. The figures for prevalence given by Holman et al were not indicated in their paper to be age-adjusted (see Naing [243] for more information on age-adjusted prevalences and their calculation) and so the assumption was made that they were not age-adjusted.

6.7.4 Back pain

Back pain is a condition where it is challenging to find a definitive estimate of population prevalence, perhaps because it is a condition that an individual may not feel worthy of investigation or it may be recorded clinically under another name. The prevalence in the composite data set of 16.4 % is no different to the estimates found in the literature. Meucci et al, in a review of studies into back pain prevalence [236], find that prevalence rates in these studies varied from 4.2 % to 25.4 % depending on the population studied.

6.7.5 Anxiety

Anxiety was a condition with a wide range of estimates for its prevalence, perhaps reflecting a difficulty in precisely defining clinical anxiety. The prevalence in the composite data set of 13.1 % is no different to the estimates found in the literature. A review by Martin [237] quotes a prevalence of 2 % to 30 % over a lifetime, with a one-month prevalence of 7.3 %.

6.7.6 Obesity

A condition that appears to be under-recorded in the composite data set, with a prevalence of 7.6 %. This could be due to under-presenting or under-recording of individuals with the condition or the condition being recorded as other conditions. It should be noted that event codes indicating high BMI have been included in the obesity prevalence calculation

The US NHANES survey of 2013-2014 suggests an (age-adjusted) obesity prevalence of 70.2 % for US adults [246], while a UN Food and Agriculture report of 2013 [247] quotes an obesity prevalence in the UK of 26.9 %.

6.7.7 Allergic rhinitis

Allergic rhinitis, or hay fever, has a higher prevalence recorded in the literature than it does in the composite data set, which records a prevalence of 15.4 %. This may be because the condition has a range of severity, from mild to severe, and individuals with mild allergic rhinitis may not present themselves for medical care, preferring to self-medicate with over the counter medication. Bauchau and Durham [239] quote a prevalence of allergic rhinitis Italy of 17 % and in Belgium of 29 %; the World Allergy Organisation [248] quote a global prevalence of between 10 % - 30 %.

6.7.8 Reflux oesophagitis

Reflux oesophagitis, or simply “reflux disease”, has a wide range of prevalence in the literature. The prevalence recorded in the composite data set of 9.4 % is within the range of prevalence in Dent et al [240] of 6.6 % to 28 %.

6.7.9 Respiratory problems

"Respiratory problems" covers a number of complaints, including COPD, asthma, bronchitis, lung cancer, pneumonia. The prevalence found in the composite data set excludes bronchitis and asthma as these conditions are subject to their own codelists and are recorded separately in the table of most common conditions (at positions 20 and 21 respectively).

The prevalence of respiratory problems found in the composite data set of 3.3 % is broadly in line with that reported in the literature. Eurostat reported that in 2014, 4.1 % of the population of the EU self-reported a non-asthma respiratory problem [241].

6.7.10 Hypothyroidism

Estimates for the prevalence of hypothyroidism vary widely (see Vanderpump [242] for an analysis of work in this area, reporting estimates of prevalence ranging from 0.6 % to 12 %). The prevalence of 7.7 % calculated from the composite data set is roughly in the middle of the range of estimates in the literature.

6.8 Validation of risk factors

We know from previous work that some conditions have known risk factors, for example the risk of an individual suffering from type II diabetes is increased if they are overweight or have high blood pressure, among other risk factors.

It is possible to investigate whether the risk factors established in previous research can be detected for any particular condition in the composite data set. The top 10 conditions previously investigated were used to test for discovery of risk factors.

For each condition, the top risk factors can be determined. This was done by, for each record, determining the presence or absence of the condition of interest, placing the record into the set of records with the condition or the set of records without the condition, and then comparing the positive and negative condition groups for prevalence of other conditions and symptoms, determining whether any increase in presence of a potential risk factor was significant or not. Note that additional codelists for potential symptoms were required to be generated.

A program was written in R to group records, for each condition under investigation, by "contains this condition" and "does not contain this condition". For records which contain the condition, event codes that occurred prior to the diagnosis of the condition are checked for the presence of potential risk factors, i.e. they are checked against the risk factor codelists. For those records which did not contain the condition, all event codes were checked.

The group of records with the condition under investigation was then split into two groups: those with the potential risk factor and those without. Similarly, the group of records without the condition was split into a 'with risk factor' and 'without risk factor' groups. The ratio of with-risk to without-risk for with-condition against without-condition was then checked using a chi-squared test, using the chisq.test function from the base R stats package [226].

6.8.1 Bias between cases and controls

Index dates are different for cases and controls: for cases, those records that contain the condition of interest, the index date is the date on which the presence of the condition of interest was first recorded, with subsequent events discarded from the analysis, whereas for the controls, those records that do not contain the condition of interest, the index date is the date of the most recent event. These different ways of selecting index date can cause bias. In order to mitigate the effect of this bias, each 'case' record should have been matched with a 'control' case where later events after the index date were also discarded.

Grimes and Schulz [249] describe three main areas of bias: selection bias, information bias and confounding. The possibility of each of these areas of bias affecting the work described in this report is briefly described:

Selection bias: the data set is comparable to the general population in terms of age distribution, gender balance, and prevalence of conditions (see section 6.3).

Information bias: It cannot be guaranteed that information on each record has been gathered in the same way (for example, some patients may attend practices that record information more assiduously, or code data more accurately, than others; the threshold for recording an event may differ between practices and between nations). However, the training set and the test set are randomly sampled from a single data set and so records with a greater degree of precision or accuracy than others are as likely to be in the training set as the test set; they are as likely to have the condition of interest as not.

Confounding: Positive cases for each condition of interest are determined by identifying the presence of a condition code indicating that condition of interest in a record. Events occurring subsequent to the condition of interest are removed, since they cannot be used as predictors for the condition of interest for that record. However, negative cases - records that do not contain a code indicating the presence of the condition of interest - remain unchanged, and so the record is not curtailed. This could introduce bias by allowing for the negative cases to have records covering a longer time span and this potentially a greater number of events. The mean number of events for positive cases (up to the occurrence of the condition of interest) was compared to the mean number of events for negative cases; results of this are show in Table 6.6.

Table 6.6 shows that in general, the positive records had a greater or equal number of events before the occurrence of the condition of interest, with two exceptions: thyrotoxicosis and colon cancer, each of which had fewer recorded events before the occurrence of the condition of interest. There is no obvious explanation for this difference.

Condition	No of patient with condition	No of patients without condition	mean no of events for patients without the condition	mean no of events for patients with the condition	median no of events for patients without the condition	median no of events for patients without the condition	t-score	df	p
acuteSinusitis	1325	7936	19.6	20.1	11	12	-0.70	1668.6	0.49
Bronchitis	**1439**	**7822**	**19.4**	**21.1**	**12**	**12**	**-2.61**	**1825.2**	**0.01**
colonCancer	34	9227	20.2	27.4	15.5	12	-1.68	33.18	0.10
osteoarthritis	**1031**	**8230**	**19.5**	**22.2**	**12**	**12**	**-3.26**	**1184.25**	**0.001**
allergicRhinitis	1619	7642	20.0	20.2	13	12	-0.24	2222.81	0.81
anyCancer	**410**	**8851**	**19.6**	**32.5**	**17**	**12**	**-7.72**	**422.93**	**8.74E-14**
Asthma	**1100**	**8161**	**18.8**	**29.3**	**17**	**12**	**-11.18**	**1220.35**	**1.06E-27**
Autism	13	9248	20.2	30.7	28	12	-1.44	12.02	0.18
breastCancer	**105**	**9156**	**20.0**	**39.9**	**31**	**12**	**-5.89**	**104.87**	**4.73E-08**
Eczema	**464**	**8797**	**19.1**	**40.5**	**35**	**12**	**-14.69**	**483.33**	**1.17E-40**
Gastroparesis	19	9242	20.3	16.2	16	12	1.93	18.40	0.07
Obesity	**1446**	**7815**	**17.7**	**33.1**	**20**	**11**	**-17.83**	**1600.9**	**4.89E-65**
praderWilli	2	9259	20.3	27.0	27	12	-0.37	1.00	0.77
prostateCancer	96	9165	20.2	23.3	9.5	12	-0.92	95.84	0.36
refluxDisease	**987**	**8274**	**19.9**	**22.2**	**13**	**12**	**-2.54**	**1118.95**	**0.01**
Stress	**255**	**9006**	**19.7**	**40.0**	**28**	**12**	**-8.77**	**258.36**	**2.51E-16**
T2Diabetes	907	8354	20.2	19.8	11	12	0.44	1027.84	0.66
thyrotoxicosis	**39**	**9222**	**20.3**	**12.9**	**11**	**12**	**6.89**	**41.51**	**2.18E-08**

Table 6.6 Comparison of number of events for positive cases vs number of events for negative cases

Conditions that show a significant difference in mean no of events for patients with condition vs patients without condition are highlighted in bold. Generally, those individuals with the conditions have a higher number of reported events than those without, excepting for thyrotoxicosis. Also note that the data set has a representative sample of codes at all levels (see section 7.3.1)

A table of conditions vs risk factors is shown in Table 6.7. Risk factors that are found to be strongly associated with conditions ($p < 0.01$) are marked ■ in the table. Each condition and their associated 'risk factors' is then briefly discussed.

		Condition								
Risk factor	Hypertension	Hyperlipidaemia	Type 2 diabetes	Back pain	Stress	Obesity	Allergic rhinitis	Reflux disease	Respiratory problems	Hypothyroidism
Hypertension	-	■	■	■	■			■	■	■
Hyperlipidaemia	■	-		■				■	■	■
Type 2 diabetes	■		-			■		■	■	■
Back pain	■			-	■			■	■	■
Stress	■		■		-		■	■	■	■
Obesity	■			■		-	■	■	■	■
Allergic rhinitis				■			-			
Reflux disease								-		
Resp. problems	■								-	■
Hypothyroidism	■							■		-
Age > 45	■	■	■		■			■	■	■
History of heart attack							■	■		
Tobacco use	■	■	■	■	■	■	■	■	■	■

Alcohol use	■	■								
Passive smoking										
Low vitamin D in diet	■	■	■	■	■	■	■	■	■	■
Sleep apnoea	■							■		
Illegal drug use	■	■			■			■		
Eczema	■		■	■		■	■		■	■
Hiatus hernia										
Gastroparesis		■		■				■		
Physical inactivity	■							■	■	■

Table 6.7 'Risk factors' associated with conditions found in the composite data set

The analysis of risk factors for particular conditions shown in Table 6.7 shows associations between conditions but cannot establish causality. However it was considered a useful exercise to check whether known risk factors for conditions are discovered to have associations with those conditions.

6.8.2 Hyperlipidaemia

The Centers for Disease Control and Prevention (CDC) list a number of factors that increase the risk of hyperlipidaemia: having type 2 diabetes, eating a diet high in saturated fats, being physically inactive, being obese, and being older [250]. The composite data set picks out type 2 diabetes, hypertension, tobacco consumption, poor diet ('low vitamin D in diet' is taken as a marker for poor diet in general for this condition and for other conditions examined) and illegal drug use and age as risk factors for hyperlipidaemia.

6.8.3 Hypertension

The CDC [251] lists type 2 diabetes, an unhealthy diet, physical inactivity, obesity, high alcohol consumption and tobacco use and being older as risk factors for hypertension. The composite data set picks out type 2 diabetes and being older as having a significant association with hypertension.

6.8.4 Type 2 diabetes

The CDC lists being overweight, being older, being physically inactive as risk factors for developing type 2 diabetes, as well as other factors not tested such as having prediabetes, having a close relative with type 2 diabetes or being from certain ethnic groups. The composite data set picks out age > 45, tobacco use and poor diet as having a significant association with hypertension.

6.8.5 Anxiety and stress disorders

The Mayo Clinic [252] notes that there are a number of possible causes for anxiety disorders or stress, including medical problems such as heart disease, diabetes, thyroid problems, respiratory disorders, drug use, withdrawal from alcohol or some medications, chronic pain or irritable bowel syndrome, some rare cancers, some medications. Risk factors include trauma, stress due to illness, build-up of stress from multiple causes, some personality types, other mental health disorders, having close relatives with an anxiety disorder or drug or alcohol use or withdrawal.

6.8.6 Back pain

The Mayo Clinic [253] give being older, being inactive, obesity, having arthritis or some cancers, being depressed, tobacco use and improper lifting as risk factors for back pain.

6.8.7 Obesity

The Mayo Clinic [254] give genetics, lifestyle, physical inactivity, poor diet, certain medical conditions and medications, social and economic issues, being older, pregnancy, stopping smoking, and poor sleep patterns as risk factors for obesity. The composite data set picks out age > 45, tobacco use and poor diet as having a significant association with hypertension.

6.8.8 Allergic rhinitis

The Mayo Clinic [255] give eczema, having a close relative with allergies or asthma, having other allergies or asthma, being constantly exposed to allergens or having a mother who smoked while you were an infant as risk factors for allergic rhinitis. The composite data set picks out eczema, tobacco use, obesity and poor diet as having a significant association with allergic rhinitis.

6.8.9 Gastro-oesophageal reflux disease

The NHS Choices website [256] gives consuming certain foods, obesity, tobacco use, pregnancy, stress and anxiety, some medications and hiatus hernia as risk factors for gastro-oesophageal reflux disease (‘acid reflux’).

6.8.10 Respiratory problems

The NHS Choices website [257] lists tobacco use, regularly breathing in fumes or dust, air pollution and genetic factors as risk factors for respiratory problems. Hypertension, hyperlipidaemia, type 2 diabetes, back pain, stress, obesity, age over 45 years, tobacco use, poor diet, eczema and physical inactivity are associated with respiratory problems in the composite data set.

6.8.11 Hypothyroidism

The NHS Choices website [258] gives immune system problems and previous thyroid treatment as the main causes of underactive thyroid, with less common risk factors being type 1 diabetes, iodine deficiency or congenital factors. Hypertension, hyperlipidaemia, type 2 diabetes, back pain, stress, obesity, respiratory problems, being older, tobacco use, poor diet, eczema and physical inactivity are associated with hypothyroidism in the composite data set.

6.9 Summary

The composite data set has been verified for completeness and for plausible data values. Records that strongly appear to be duplicates have been removed. The source data sets have been validated against each other to ensure that there are no significant differences in age distribution or gender distribution.

Condition prevalences in the composite data set have been validated against condition prevalences described in the literature for the 10 most common conditions in the USA. The ordering of the most common conditions in the composite data set had some differences to the top 10 conditions in the US study, but this difference in ordering was not found to be significant. The prevalence of each these 10 most common conditions were calculated for the composite data set and compared to the prevalence reported in the literature. Some differences were found although all prevalences were within the same order of magnitude.

Risk factors for the most common conditions were investigated, comparing significant associations of risk factors to conditions in the composite data set against accepted risk factors for these conditions. In general, the accepted risk factors were identified in the analysis of the composite data set.

Contribution: In order to identify conditions and risk factors from the composite data set, it was necessary to build codelists comprising all the CTV3 codes that indicated the presence of a particular condition or risk factor. These codelists are detailed in Appendix 2. On completion of this work the codelists will be uploaded to the University of Manchester Institute of Population Health codelist repository.

7 ESTIMATION OF CONDITION RISK: METHODS

7.1 Introduction

At the population level, identification of groups of individuals at risk of particular conditions can help with health care resource planning, particularly for conditions that have a high individual cost of treatment and/or affect a high proportion of the population, and promotion of lifestyle choices that will reduce the numbers of individuals developing such conditions.

It is therefore proposed that identification of individuals at raised risk of particular conditions and an indication of the degree to which that risk is raised will have value to their healthcare. In order to estimate the raised risk for those individuals at increased risk of particular conditions, a system has been developed to interrogate clinical histories for previous diagnoses and symptoms and using these data estimate the changed risk. Opportunities for use of this information are two-fold: as a pre-screening tool to identify groups of individuals at raised risk of particular conditions, and as an opportunistic calculation of modified risk for an individual when attending a clinical practice, for presentation of modified risk to a clinician.

7.1.1 Background

One of the challenges of healthcare is to achieve timely diagnoses for conditions, i.e. a diagnosis that is early enough to improve outcomes for individuals with those conditions, whether that improved outcome is an improved chance of a cure, no cure but alleviation of ill-effects, or a combination of the two. Early diagnosis can be problematic since symptoms that are indicative of a particular condition may not be fully present, or symptoms may be present that could be indicative of a number of different condition. A measure of the challenge of diagnosis can be seen from the estimate of the number of diseases of more than 10,000 [259] versus the number of symptoms of less than 400 [158], and so it is clear that any one symptom could have many different underlying causes. When making a diagnosis it is natural to base that diagnosis on the symptoms, combinations of symptoms, history of the symptoms and of previously diagnosed conditions, and on the prevalence of candidate.

There are around 7,000 diseases considered to be rare diseases (diseases that affect less than 5 persons in 10,000 according to the EU definition [260]). Unless there are clear symptoms indicating that one of these conditions is present, it is natural to consider more common conditions as an explanation for symptoms presented. For conditions that are not rare but less prevalent than others, this preferential diagnosis may also be the case. For example, thyrotoxicosis (prevalence = 0.06 % in the composite data set) shares many symptoms with diabetes (prevalence = 6.0 %), depression (3.3 %) [261] and viral infection (varies by infection type) and is formally diagnosed by blood test, which has a financial cost.

A simple method of identifying individuals who are at increased risk of particular conditions can be the first step towards an earlier diagnosis than may otherwise be achieved. Should an early diagnosis be obtained, this can lead to earlier treatment with the possibility of improved outcome, although this may not be true for all conditions, particularly those that are not treatable or are slow-developing conditions identified in later life. Individuals who are identified as being in an ‘at-risk’ group for a particular condition but who have yet to develop the condition can have the opportunity to modify their behaviour to reduce the risk of developing that condition, for example improving diet to reduce the risk of developing type 2 diabetes or hyperlipidaemia. The NHS Health Check for over-40-year-olds in the UK aims to identify patients at-risk of several conditions, including diabetes and heart disease, diagnosing those who have developed conditions undiagnosed prior to their health check and offering lifestyle advice to those at risk of some conditions. However this health check is offered only to a sub-group of the population (those aged over 40), relies on individuals actively responding to the offer of a health check, and costs money to UK health providers and has some level of inconvenience, invasiveness and possibly cost to the individuals who are checked.

It is proposed that knowledge of the likelihood of a condition is useful information to a clinician when making a diagnosis. For example, around 10 % of the population has diagnosed diabetes, but this base figure varies between age groups and ethnicities. Given knowledge of the prevalence of conditions, clinicians can preferentially consider candidate diagnoses for a set of symptoms by considering and eliminating diagnoses in order of their statistical likelihood.

A second usage is as a pre-screening tool, selecting those patients at raised risk for a particular condition for screening (discuss screening only of benefit if there is a treatment for the condition that is commensurate with the condition’s risks, i.e. treatment for the common cold should be cheap and with few side effects; treatment for cancer can be expensive and have side effects, provided that these side effects are less harmful than the condition).

Described here is a technique for estimating risks of defined conditions that is based on an individual's history of diagnoses of clinical events, combined with two major lifestyle factors, history of tobacco use and alcohol consumption, and age. Based on these data, individuals' records are matched against other records to discover those records which share a similar set of event codes, lifestyle factors and age. The set of similar records can then be examined to determine whether the prevalence of the condition of interest is raised when compared to the prevalence within the complete composite data set.

7.1.2 Resources produced in previous chapters

This work will utilise the resources described and validated in the previous chapters. These are:

(i) A composite data set of records of clinical events derived from individuals, coded in a single coding system, Clinical Terms Version 3 (CTV3) and stored in a single file with each record comprising a list of clinical events, age and gender information;

(ii) A table of CTV3 event codes that contains, for each code, a ‘significance value’ that indicates whether an event code is an administration code, a symptom code or a diagnosis code, and a list of all parent codes up to the root node of the Clinical Terms Version 3 hierarchy;

(iii) A set of codelists that list the CTV3 codes that indicate the presence of a particular condition or symptom;

(iv) A list of the Top 10 most common conditions in the USA.

7.2 Methods

The risk of a particular individual developing a particular condition is affected by several factors: genetic predisposition and hereditary factors; environmental, dietary, employment and lifestyle choices; exposure to infectious diseases; chance. Some conditions are determined wholly by one of these factors (e.g. Methods for calculating an individual's risk of acquiring a particular condition can be calculated in a number of different ways: clinical markers from e.g. blood tests; genetic testing; environmental factors; lifestyle information.

The composite data set was randomly split into two sets using the base R ‘sample’ funcion: a training set for derivation of factors and a test set to score the predictions made on the basis of these factors. The training set and the test set each comprised 50 % of the composite data set, giving 11566 records in the training set and 11567 in the test set. No matching of the two sets for tobacco or alcohol use history or for age was performed.

A 50% training set/50 % test set split was used instead of the more common 90/10, 80/20, or 75/25 splits for two reasons: with a larger training set the computer was running out of memory and with a smaller test set the numbers of patients with rare conditions in that test set was small and sometimes, for very low prevalence conditions, zero.

Dobbin and Simon [262] in their analysis of optimal splits between training and test sets write that ‘the rule of thumb that assigns 2/3rds to the training set and1/3rdto the test set performs well’ but also note that ‘We discovered that the optimal proportion of cases for the training set tended to be in the range of 40% to 80% for the wide range of conditions studied.’

Choice of methods: Explainability is challenging for neural network methods and although work is being done in producing explainable AI models, currently this explainability remains a challenge, particularly if it is established that clear explainability is required by the GDPR (see section 2.6.4). With the motivation of having an explainable, patient-centric approach to condition prediction, methods such as clustering or k nearest neighbours were clear choices. Additionally, Wu [200]notes that knn methods are 'particularly well suited for multi-modal classes as well as applications in which an object can have many class labels'. Kim et al note that the k nearest neighbour method was "adaptive to relatively noisy training sets, simple to implement, and naturally handles multi-class cases. [It] also has a history of high success rates in the medical field" [143]. These methods have applicability to Case-Based Reasoning, in the retrieve part of the CBR process (see, for example, Sae-Hyun Ji et al [263] for a discussion of the use of nearest neighbour methods in Case-Based Reasoning.

Two techniques for discovering similar patients were investigated: clustering of similar records into groups, and simple nearest neighbours by means of minimisation of a distance metric. These methods are briefly introduced in Sections 7.2.1 and 7.2.2.

7.2.1 Clustering

In this method, the distance of each record from the current record of interest is again calculated. This distance information is then used to form the set of records into subgroups. The prevalence of the condition of interest in the subgroup containing the record of interest is used to make a prediction on the presence or absence of the condition of interest in the record of interest. The optimum number of subgroups into which the records set is to be divided must be determined in advance.

7.2.2 K nearest neighbours (KNN)

In this method, the distance of each record from the current record of interest is calculated. Records are then ordered by their calculated distance from the record of interest. The 'k' closest records, where the optimum value for k is to be determined, are used to make a prediction on the presence or absence of the condition of interest in the record of interest.

KNN is a non-parametric algorithm and so makes no assumptions about the distribution of the data. This is valuable with the data used here. However, as a lazy learning algorithm, it may run slowly, particularly as the number of data points increases. In the case of the composite data set, the number of data points depends on both the number of records and the level of the CTV3 hierarchy used for the predictions.

7.3 Choice of factors for record matching

Apart from the choice of the number of clusters in the clustering method or the value of ‘k’ in the nearest neighbours method, a number of other factors must be determined before predictions can be made using the record set. Indeed, these must be determined before the best values for the number of clusters or k can be determined. Factors to be determined in the technique described here are the level within the CTV3 hierarchy at which all event codes should be mapped; the minimum number of events that a record should have before it can be included in the analysis; the value of the number of clusters for the clustering method or the value of k (for the nearest neighbours method).

7.3.1 Event code level

Events are recorded in each record in the hierarchical clinical terminology Clinical Terms Version 3, as described in Chapter 4. Each CTV3 code, barring the root node, has a parent code and may have child codes. At the lowest level of the hierarchy, the granularity of the codes is finest and so can differentiate between records containing similar but not identical conditions. At higher levels of the hierarchy, information is less granular but related conditions are grouped together and so matching of records may be more successful. The total number of CTV3 codes reduces at higher levels of the hierarchy.

Table 7.1 shows the number of unique CTV3 codes present at each level of the CTV3 hierarchy, for both the complete CTV3 code set and the codes that are present in the composite data set. Figure 7.1 shows the cumulative count of unique codes present as the hierarchy is descended. Note that CTV3 codes that are introduced at higher levels of the hierarchy remain valid at lower levels of the hierarchy.

Depth level	unique codes count (complete CTV3 code set)	Unique codes count (composite data set only)
1	1	1
2	18	14
3	450	139
4	3532	702
5	17000	2359
6	66825	5804
7	123613	9787
8	169945	12708
9	204285	14591
10	226377	15492
11	242009	16009
12	251174	16195
13	256860	16246
14	258440	16252
15	258785	16253
16	258841	16253
17	258848	16253
18	258851	16253
19	258853	16253
20	258853	16253

Table 7.1 Unique CTV3 codes present at each level of the CTV3 hierarchy

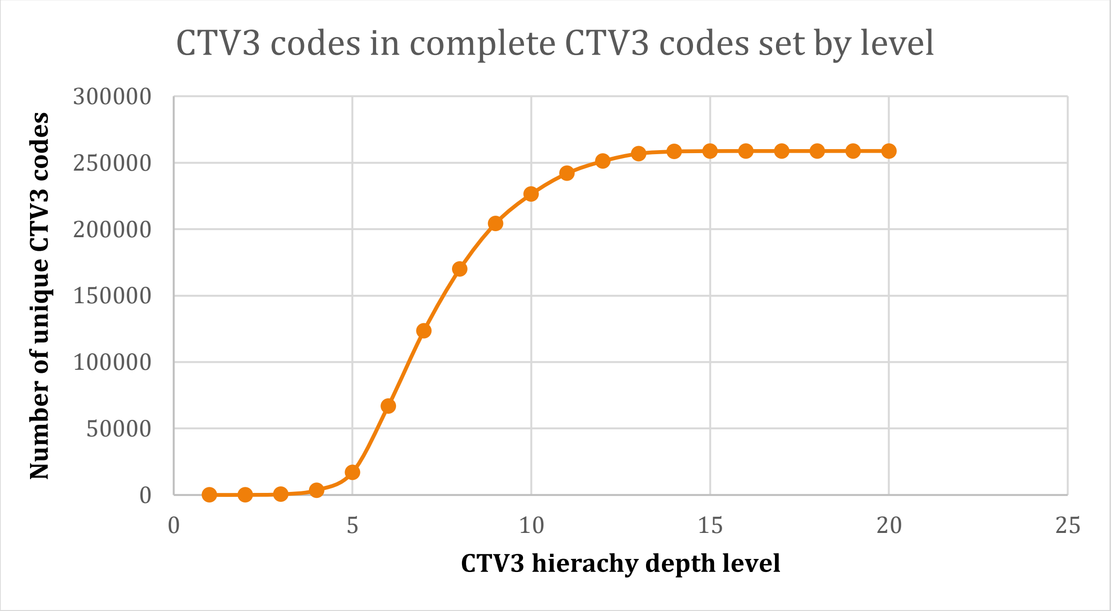


Figure 7.1 Unique CTV3 codes present in the complete CTV3 code set at each level of the CTV3 hierarchy

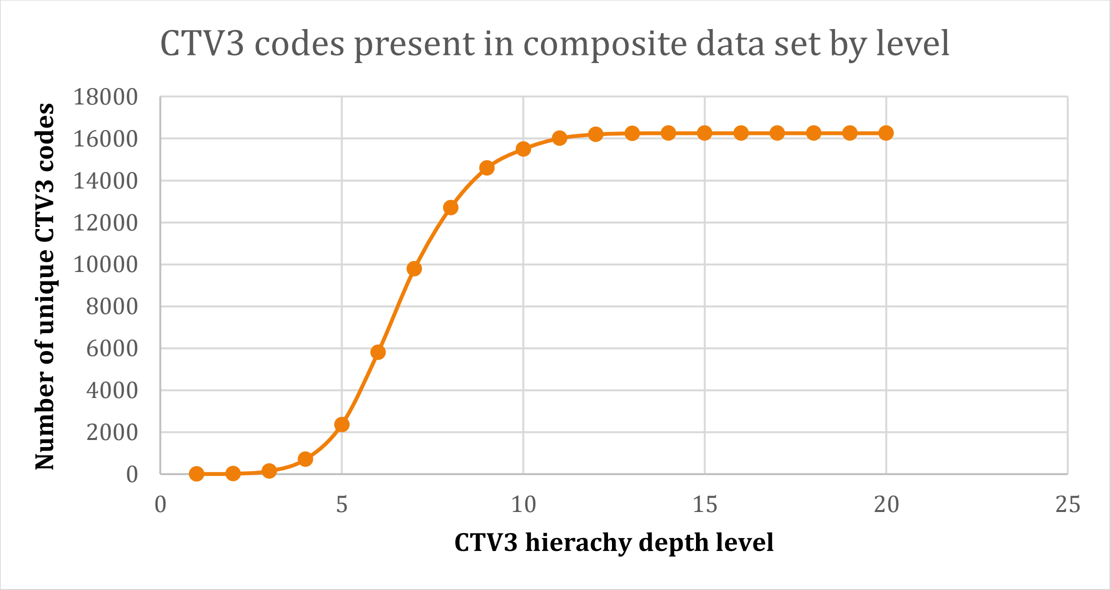


Figure 7.2 Unique CTV3 codes present in the composite data set at each level of the CTV3 hierarchy

Inspection of Figure 7.1 suggests that the rate of increase in the number of codes as the hierarchy is descended reduces at lower levels. Two methods for deriving the most useful level of the hierarchy to use (i.e. to which level of ancestor code to map CTV3 codes from lower levels) were used: (i) determination of the point of inflection of the curve shown in Figure 7.2 and (ii) the level of the hierarchy at which there was no significant increase in the number of unique CTV3 codes as the hierarchy was descended.

To investigate whether the subset of codes in the composite data set was a consistent sample of the complete code set at each level of the CTV3 code hierarchy, the number of CTV3 codes in the composite data set was plotted against the total number of CTV3 codes available, and a correlation coefficient calculated.

The value calculated for the correlation coefficient R^2 was 0.9877, suggesting a high degree of correlation.

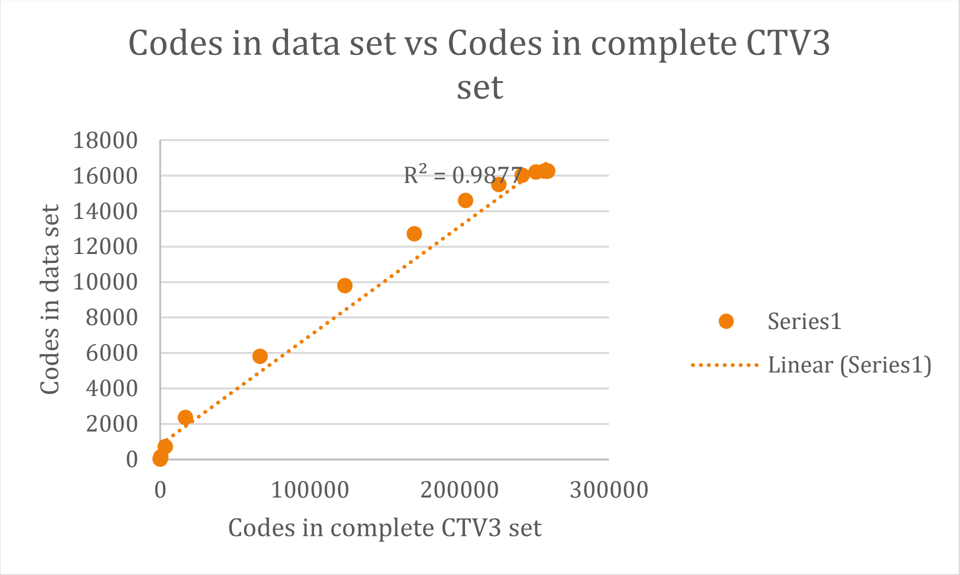


Figure 7.3 Comparison of number of codes in the composite data set vs number of codes in the complete CTV3 code set at each level

7.3.1.1 Determination of point of inflection

The point of inflection is the point on the curve at which a change in the direction of curvature occurs, in the case of Figure 7.1 from convex to concave. It is the point at which the rate of increase in the number of CTV3 codes starts to lessen as the hierarchy is descended. The point of inflection was calculated using the 'bede' function from the R package 'inflection' [264].

The point of inflection test was run on the complete set of CTV3 codes, returning a value of 7 for the level in the CTV3 hierarchy at which the point of inflection occurred. The test was repeated on only the CTV3 codes in the composite data set, again returning a value of 7.

7.3.1.2 By inspection

It can be seen from Figure 7.1 that the increase in the number of CTV3 codes as the CTV3 hierarchy is descended reduces at the lowest level of the hierarchy. As well as being the point of inflection (section 7.3.1.1), it was also noted that level 7 was the level at which half the CTV3 codes had been absorbed into higher level codes. Level 11 was also chosen since below this point the increase in the number of codes was trivial and not expected to add any value to the analysis.

With no further evidence to suggest which was the better level (7 or 11) for choosing the level of the CTV3 hierarchy at which to perform the nearest neighbour calculations, it was decided to use both values against the test set to see which performed better for each condition under consideration.

7.3.2 Minimum number of clinical events

The number of events recorded in the records varied from 1 to several hundred (minimum = 1; maximum = 332; mode = 7; standard deviation = 35.7). It was desired to include the maximum number of records in the analysis but also to ensure that there was sufficient information in each record for it to make a valid contribution to the analysis. No theoretical basis for determination of the minimum number of events per record was found and so the pragmatic decision was made to include 90 % of records, which excluded records comprising only three events or fewer. This decision was made on the basis that it would produce predictions for the majority of the records in the data set, excluding only those records with a small number of events. Increasing the minimum number of events to qualify for the analysis may increase the quality of the prediction but would reduce the number of records for which a prediction could be made. There was a need to balance being able to make a prediction for the highest number of individuals with the need for that prediction to have value. A high value for minimum events could have been chosen but this would restrict the individuals for whom predictions are being made and restrict the pool of their ‘nearest neighbours’

Unique events in record	Frequency	Cumulative percentage
1	979	4.2 %
2	770	7.6 %
3	954	11.7 %
4	1171	16.7 %
5	1205	22.0 %
6	1211	27.2 %
7	1290	32.8 %
8	1177	37.9 %
9	1086	42.5 %
10	962	46.7 %
11	845	50.4 %
12	736	53.5 %
13	677	56.5 %
14	641	59.2 %
15	590	61.8 %
16	429	63.6 %
17	467	65.7 %
18	328	67.1 %
19	304	68.4 %
20	287	69.6 %

Table 7.2 Frequency of number of event codes per record

7.3.6 Choice of distance metric

For both the clustering method and the k nearest neighbours method it is necessary to calculate the 'distance', 'similarity' or 'dissimilarity' between records. Note that 'distance' will be used loosely, to refer to any method that calculates a value indicative of (dis)similarity or distance between records.

There are many methods to calculate the distance between points in multidimensional space, as is the case with the records here which contain multiple events data, as well as age, gender, tobacco use and alcohol consumption data. A search of packages in the R programming language found 54 different methods implemented in functions available through various R packages. Table 7.3 lists the methods found and the R packages that implement them. The choice of methods investigated here was limited to those available in at least one R package. For completeness, empirical tests of all distance measures available in R were performed. No account was made of the theoretical basis of each distance measure, although it is acknowledged that some of the metrics are not appropriate for use on the data set used here, for example the Bray-Curtis dissimilarity measure [265], intended to give an indication of the dissimilarity of ecological sites based on the counts of species present at the sites.

7.3.6.1 Preliminary analysis of distance metrics.

Each distance calculation method was run against a small sub-sample, comprising 1000 event codes from 1000 records, randomly chosen from the complete data set. The same sub-sample was used for each method tested. Results were inspected for agreement and disagreement and validity for our particular data set. Note that some different methods give the same results when run against our binary data set.

Having run each method from each package against our test data set, the subset of methods that give meaningful results was selected, with the package that ran the fastest for that test being chosen as the package to use for that test.

Table 7.3 lists the distance calculation methods available in R, together with the libraries and functions that implement these methods, and the results from the first few cells of the distance matrix produced when run on the a sub-sample. Results were coded according to the results listed in Table 7.4. Note that some different methods give the same results when run against the binary sample data set.Although this does not imply that these methods were equivalent in their methods of calculation of distance metric, for the empirical, data-driven, purposes of the investigation in this section the ability of a method to differentiate between records was the important consideration.

Each distance method was then run against a fuller sample of 5000 records from the training data set, including age, gender, alcohol and smoking codes as well as the event codes.

In order to investigate the results returned by the different distance calculation methods, each method was run against a small test set and the results inspected. Some methods returned identical results, other methods returned results that were a simple multiple of other results, and yet other methods returned results that had no differentiation between different points in the sample.

R package:	vegan	amap	wordspace	fields	stats	parDist
Distance method	vegdist	Dist	dist.matrix	rdist	dist	parDist
euclidean	A	A	A	A	A	A
maximum		B	B		B	B
manhattan	C	C	C		C	C
canberra	B	D	C		D	E
binary		B			B	B
bray	B					B
kulczynski	B					
jaccard	B					
gower	F					
altGower	B					
morisita	Z					
horn	B					
mountford	B					A
raup	B					
binomial	G					
Chao	B					
Cao						
mahalanobis	H					
Pearson		B				
correlation		J				
spearman		K				
kendall		L				
minkowski			A		A	A
cosine			M			
simpson						Y
simple						F
russel						B
phi						N
ochiai						B

mozley						B
stiles						Z
tanimoto						P
yule						X
yule2						X
bhjattacharyya						A
chord						Q
divergence						C
dtw						R
fjaccard						B
geodesic						S
hellinger						Q
kullback						Z
podani						T
soergel						B
wave						Z
whittaker						B
braun-blanquet						B
dice						B
fager						U
hamman						V
kulczynski1						B
kulczynski2						B
michael						W
faith						Y

Table 7.3 Distance calculation methods and their packages available in R

Result set	Result set values
A	2, 3, 1.732051, 2.236068, 3.162278, 2.645751, 2.828427, 2, 3, 3.605551, 2.236068, 2.645751, 2.236068, …
B	1, 1, 1, 1, 1, 1, 1, …
C	[Results from set A, squared]
D	0.636537, 0.954805, 0.551257, 0.711670, 1.006454, 0.842059, 0.900199, 0.636537, 0.954806, 1.147533, 0.711670, 0.842059, 0.71167 …
E	[1 + results from set D]
F	0.03636364, 0.08181818, 0.02727273, 0.04545455, 0.09090909, 0.06363636, 0.07272727, …
G	2.772589, 6.238325, 2.079442, 3.465736, 6.931472, 4.85203, 5.545177, 2.772589, 6.238325, …
H	14.07125, 14.07125, 10.48701, 11.47232, 14.07125, 13.64563, 14.07125, 13.1542, 14.07125, …
J	1.016038, 1.040219, 0, 1.022786, 1.043651, 1.032527, 1.036539, 1.016038, 1.040219, 1.05295, …
K	27554, 52276, 14666, 23638, 48278, 51480, 53138, 26330, 42690, 73138, 26416, 39894, 26048, …
L	0.07773144, 0.1411176, 0.04553795, 0.08056714, 0.1421184, 0.1446205, 0.1497915, …
M	90, 90, NaN, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, …
N	0.9839618, 0.9597813, NaN, 0.9772138, 0.9563485, 0.9674729, 0.9634608, 0.9839618, …
P	0.07017544, 0.1512605, 0.05309735, 0.08695652, 0.1666667, 0.1196581, 0.1355932,
Q	1.414214 1.414214 NaN 1.414214 1.414214 1.414214 1.414214 1.414214 1.414214 …
R	2 5 3 1 6 3 4 2 5 8 1 3 1 3 2 2 3 3 1 7
S	1.570796 1.570796 NaN 1.570796 1.570796 1.570796 1.570796 1.570796 1.570796 …
T	0.1434529 0.3152627 0.1080901 0.1784821 0.3486239 0.2475396 0.281568 0.1434529 …
U	0.1339746 0.1339746 NaN 0.1339746 0.1339746 0.1339746 0.1339746 0.1339746 …
V	0.07272727 0.1636364 0.05454545 0.09090909 0.1818182 0.1272727 0.1454545 0.07272727 …
W	0.9989335 0.9929975 1 0.9978281 0.9916832 0.9954963 0.9942682 0.9989335 0.9929975 …
X	[All 0]
Y	0.5181818, 0.5409091, 0.5136364, 0.5227273, 0.5454545, 0.5318182, 0.5363636, 0.5181818, …
Z	[All NaN]

Table 7.4 Indicative results from distance matrix calculations on small data sample.

The Yule distance metric did not produce meaningful results. This is assumed to be because it is intended for use on Boolean data; the inclusion of age in the data has caused the function available in the parDist package to fail, returning a value of 0 for all. The Kullback function similarly did not produced meaningful results. The kullback function compares probability distributions rather than Boolean or continuous data and so the data presented to this function is inappropriate – the function expects the sum of the value presented to be less than or equal to 1.

Those results sets that did not produce results that would be useful in a distance matrix were excluded from consideration. Methods that produced the same result for the distance matrix as another method were also excluded from further analysis. This left a shorter list of methods for further investigation: Euclidean, Manhatten, Canberra (but not the version from the vegdist package), Gower, Morisita, binomial, Mahalanobis, correlation, Spearman, Kendall, Simpson.

Note that although the absolute values returned by the various distance matrix methods may vary between methods, when individual records are ordered by these values, the ordering of the records may be the same. This reflects some loss of information in the change from cardinal to ordinal numbering. Section 7.2.1 investigates a clustering method to make predictions, in which the cardinal distance information is retained.

For the k nearest neighbour method described in Section 7.2.2, a distance matrix method was sought that gave the most useful results on the set of records. For those methods that give similarly good results, the method that produces its results in the shortest time is preferred.

Determination of best distance method

(i) Brute force method

Run every method against a sub-sample of 5000 records from the training set. Use the level, minimum events, and a set of values for k as factors. Age and gender were not included: the test was intended to investigate the solely the ability to find similarities based on event histories. The test was run on one condition: type 2 diabetes, since this has a high prevalence and is explicitly coded in event codes.

Two CTV3 hierarchy levels were suggested, 7 and 11, and so both these will be used. Values for k were harder to establish a priori, and so a selection of arbitrary values was used: k = 100, k = square root of the number of valid observations, k = half of the number of valid observations. Only one value for minimum events will be used, 4 (see section 7.3.2 for choice of minimum events value). This gives six permutations for testing the distance matrix methods. The results from each run will be ranked and a final ranking produced.

For each run, each record will have a set of nearest neighbours. This set was tested for the prevalence of the condition under consideration, in this case type 2 diabetes. Should the prevalence of the condition in the nearest neighbours set be higher than the prevalence in the whole records set (i.e. a prevalence in the nearest neighbours set significantly higher than the prevalence in the complete test set ($p < 0.05$)), the record was predicted to have the condition, otherwise it was predicted not to have the condition. The prediction was than compared to the actual state for the record, which was then scored as a true positive, true negative, false positive or false negative as appropriate. The complete set of prediction results is then aggregated and scored. The score used in this instance is the Matthews Correlation Coefficient (MCC), chosen due to it being a balanced measure of the quality of all predictions in a binary classification, and because it is insensitive to class size.

Included in the table are the MCC scores resulting from a random allocation of positive and negative predictions to each record, allocations being made in proportion to the prevalence of the condition in the records sample.

CTV3 Level	7	7	7	7	11	11	11	11	Run time at level 7 (s)	Run time at level 11 (s)
K	10	100	Sqrt(records count)	Records count/2	10	100	Sqrt(records count)	Records count/2		
Scoring method	MCC									
random	0.102									
euclidean	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	27	72
maximum	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	83	308
manhattan	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	27	76
canberra	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	254	1200
binary	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	85	311
bray	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	212	629
kulczynski	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	80	287
jaccard	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	218	613
gower	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	219	648
altGower	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	221	643
morisita	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	228	679
horn	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	227	681
mountford	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	84	310
raup	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	227	602
binomial	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	247	663
chao	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	214	686
cao	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	214	634
Mahalanobis	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	211	653
Pearson	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	376	1152
correlation	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	373	1278
spearman	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	376	2434
kendall	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	317	1281
minkowski	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	445	1174
cosine	0.164	0.235	0.243	0.204	0.185	0.276	0.274	0.273	1	2

simpson	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	90	314
simple	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	84	291
russel	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	82	292
phi	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	81	294
ochiai	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	84	300
mozley	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	92	295
stiles	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	84	289
tanimoto	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	82	300
yule	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	82	299
yule2	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	87	294
bhjattacharyya	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	85	288
chord	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	80	301
divergence	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	78	274
dtw	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	79	314
fjaccard	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	81	309
geodesic	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	81	278
hellinger	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	80	280
kullback	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	78	297
podani	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	78	285
soergel	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	81	292
wave	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	80	291
whittaker	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	76	292
braun-blanquet	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	81	282
dice	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	81	280
fager	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	80	277
hamman	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	80	291
kulczynski1	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	78	284
kulczynski2	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	81	282
michael	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	80	280
faith	0.144	0.213	0.218	0.003	0.187	0.290	0.266	0.021	50	287

Table 7.5 MCC scores and run times for various distance calculation methods

Run times (in seconds) are for running the distance matrix function only and do not include other sections of the program such as file reading and preparing the data for the distance function. All runs were on a Dell PC running 64-bit Windows 7 Enterprise, Intel Core i5-6500 4-core CPU running at 3.2 GHz, 8 GB RAM, with minimal other programs running. The ‘parDist’ function used all 4 cores in parallel to generate the distance matrix; other functions were single-core. 5000 records were read to test the different functions, and were prepared by removing administration codes and by removing records that had only administration codes or the target condition.

With the exception of the cosine method, the results as ranked by F1-score fall into two distinct groups, with identical results within each group. All methods out-perform random prediction. The group with the better ranking F-scores contains the bray, jaccard, gower, morisita, horn, raup, binomial, chao, cao and mahalanobis methods, each of which had similar run times. Without a priori knowledge of correlations within the data and with no other reason to select one method over another other, the binomial method was chosen to use in the later analysis. This method was developed to investigate the effect of habitat on fish populations [266] and works with binary and continuous data, and so has no theoretical barrier to use in this work. However, the cosine method performs substantially faster than any other method with only an apparently small reduction in effectiveness of prediction and so the cosine method was also be tested in the final analysis on the test data set, in order to compare its predictive performance against the binomial method.

As described above, other factors that will be used are minimum number of events in a record to qualify for inclusion in the analysis, the level of the CTV3 hierarchy to which event codes should be relegated, and the size (or calculation method) of the nearest neighbours group on which to base prediction of presence or absence of a condition. A further consideration is the method by which the prevalence of the condition within a nearest neighbours group is deemed sufficiently high to warrant a positive prediction for the condition.

Those factors (level, k, calculation method) for which no single value can be decided from first principles were be optimised on the training set and the resultant optimal set of factors used to evaluate the performance against the test set of records.

7.3.7 Prevalence significance calculation method:

Once a set of nearest neighbours to a record has been produced, members of that nearest neighbours set are used to determine whether to predict the presence or otherwise of the condition of interest, i.e. condition positive or negative. There are several alternative methods suggested for determining whether the condition prediction should be 'positive' or 'negative':

(i) Simple majority vote: count the number of positive records in the nearest neighbours set and compare to the number of negative records. If there are more positive records, predict positive for the record of interest. Note that as the value of k increases (i.e. the size of the nearest neighbours set increases) the proportion of positive records in the nearest neighbours set will converge to the prevalence in the complete records set. For all conditions under consideration in this work, this prevalence is substantially less than 0.25 and so for large values of k (k > 50 % of the records set size), positive records cannot achieve a majority.

(ii) High prevalence: if the prevalence of the condition in the nearest neighbours set is greater than a multiple (to be determined) of the prevalence calculated from the complete records set, then predict positive for the record under consideration

(iii) Prevalence in the nearest neighbours set significantly greater than prevalence in the non-nearest neighbours set: compare the proportion of the records in the nearest neighbours set that are positive for the condition of interest to the proportion of records in the remainder of the records set that are condition positive. If the proportion in the nearest neighbours set is positive then predict positive for the record of interest.

Simple majority vote (i) was dismissed on the basis that the prevalence of any one condition would be very much less than 50 % and so it would be a challenge for any one group of records to have an intra-group prevalence high enough to trigger a positive vote. Conversely, simply setting the prevalence trigger level higher than an arbitrary threshold (ii) was dismissed because it takes no account of condition prevalence in the data set; the low (much less than 50%) prevelance of any one condition in the data set would make it difficult for any set to be scored as a 'positive'. The method chosen was (iii), comparing the prevalence within the group of interest against the prevalence in the rest of the record set, with a positive prediction for the condition made should the intra-group prevalence be significantly higher than the population prevalence, by setting a p-value threshold of 0.05.

7.3.8 Scoring prediction success

Once all records have had predictions made, the success of these predictions can be scored. Comparing the prediction for each record to the truth for each record has one of four possible outcomes: True Positive (i.e. the prediction was positive and the record actually contained the condition of interest); True Negative (i.e. the prediction was negative and the record did not contain the condition of interest); False Positive (i.e. the prediction for the record was positive but the record did not contain the condition); and False Negative (i.e. the prediction for the record was negative but the record did, in fact, contain the condition).

Over the complete set of records, the number of True Positives, True Negatives, False Positive and False Negatives can be summed. Froom these total values a variety of statistical scores can be calculated:

- Sensitivity, or Recall, or True Positive Rate: the proportion of records with the condition that were predicted to have the condition
- Specificity, or True Negative Rate: the proportion of records without the condition that were not predicted to have the condition
- Positive Predictive Value (PPV), or Precision: the proportion of those records predicted to be positive that actually were positive
- Negative Predictive Value (NPV): the proportion of those records predicted to be negative that actually were negative
- F-score: A measure of the accuracy of a test. It uses the Positive Predictive Value and the Sensitivity to calculate a score indicating the overall accuracy of a test. The F-score can be weighted to give more weight to PPV or to sensitivity according to the priorities for the test. The F1 score give equal weight to PPV and to sensitivity.
- Matthews Correlation Coefficient: a statistic which endeavours to give a balanced single value for the quality of classifications, giving equal value to True Positives, True negatives, False positive and False Negatives in its calculation.

- Accuracy: the proportion of correct results (i.e. True Positive + True Negatives) to the total set size.
- Likelihood Ratio: the ratio of the probability of the test result in diseased persons over the probability of the test result in non-diseased persons.

This work endeavours to utilise the information in clinical records as a test that suggests which records indicates an increased likelihood of a condition and thus which patients may benefit from a more formal screening test. As a non-invasive test, at this stage, the test can give greater value to increased sensitivity rather than to increased specificity, i.e. the test should aim to include most patients who are likely to have the condition rather than exclude those patients who do not have the condition. To this end, the F2 score was chosen as the primary statistic on which to optimise the input factors.

7.4 Determination of best factors for clustering approach

7.4.1 Method

In this technique, a distance matrix giving the calculated distance between records is created, as it was for the nearest neighbours method. Note that for records that contain the condition of interest, events occurring later than that condition of interest are not included in the distance calculation since these events would not yet have occurred in a real world situation. An index date in the control patients was not imposed. See section 6.8.1 for a discussion of the potential for bias here. Once the distance matrix has been formed, records are grouped together into groups or 'clusters', each cluster containing records that are more similar to each other than to those records placed in other clusters. Similarly to the nearest neighbours method, there are a number of input factors that can affect the output from the clustering algorithm:

(i) the level of the CTV3 hierarchy at which event codes are to be aggregated;
(ii) the minimum number of events per record to qualify for inclusion in the analysis;
(iii) the method by which the distance matrix is calculated;
(iv) the number of clusters into which the records are to be placed;
(v) the clustering method;
(vi) the method for calculating the predicted outcome for each record.

The choice of each of these factors can, in many cases, be informed by the preparation and analysis for the nearest neighbours method, and so the following factors are used:

Minimum events per record: 4 events;

Level of CTV3 hierarchy to which to group event codes: 7 or 11;

Distance matrix formation method: binomial;

The optimum number of clusters into which records are placed: to be determined;

Clustering algorithm: to be determined from the set of clustering methods implemented in base R or in R library functions.

Clustering method: agglomerative or divisive.

Method for calculating the predicted outcome for each record: predict positive for a record if condition prevalence in the rest of the record's cluster is significantly greater than population prevalence

To determine the optimum number of clusters and the clustering method, each clustering algorithm-clustering method pair was used in turn and at both candidate levels of the CTV3 hierarchy on the training data set in order to generate the clustering hierarchy. This clustering hierarchy was then used to produce a set of clusterings, with the number of clusters ranging from 1 (i.e. all records placed in a single cluster) to the number of records clustered (i.e. all records placed singly in their own cluster).

For each set of [CTV3 level – clustering algorithm – clustering method – number of clusters], the F2 score was calculated, using the same method as for the nearest neighbours technique.

The best performing combination of factors was selected for use on the test set of records and the same set of results produced for each condition.

In order to check that valid results were achieved, an exploratory analysis was performed using a selection of distance matrix method – clustering method pairs, on a subsample of the training set of records: CTV3 hierarchy level=7, 3000 records, 500 clusters, for a condition of type 2 diabetes.

Distance method	manhattan	euclidean	canberra	bray	kulczynski	jaccard	gower	morisita	horn	mountford	raup	binomial	chao
Clustering method													
ward.D	0.404	0.393	0.441	0.431	0.431	0.432	0.414	0.217	0.435	0.341	0.421	0.402	0.387
ward.D2	0.414	0.394	0.412	0.424	0.434	0.430	0.405	0.228	0.433	0.348	0.440	0.406	0.399
single	0.345	0.345	0.356	0.143	0.368	0.358	0.345	NA	0.358	0.371	0.378	0.345	0.357
complete	0.367	0.367	0.410	0.423	0.414	0.423	0.367	0.210	0.418	0.443	0.430	0.373	0.394
average	0.351	0.348	0.406	0.442	0.419	0.476	0.350	0.201	0.468	0.420	0.439	0.350	0.409
mcquitty	0.362	0.371	0.469	0.435	0.442	0.452	0.369	0.122	0.463	0.393	0.413	0.379	0.404
median	0.345	0.341	0.365	0.371	0.363	0.376	0.345	0.009	0.364	0.360	0.407	0.345	0.414
centroid	0.345	0.347	0.357	0.363	0.380	0.371	0.345	0.037	0.368	0.368	0.430	0.345	0.429

Table 7.6 F2 scores for distance method-clustering method combinations

In summary, the following factors will be used in optimisation:

Distance method: binomial, using the ‘vegdist’ function from the R ‘vegan’ library [267].

CTV3 hierarchy level: 7 or 11, results to determine which method to use

Prevalence significance calculation method: to be determined

K: to be determined

7.5 Nearest neighbours method

The clustering method has been described and investigated in 7.4 et seq. Some success for some investigated conditions was found, which are described in Chapter 8. A second method was also tested, the nearest neighbours method. This required determination of the best value for k, the number of neighbours to include in the nearest neighbours group.

7.5.1 Method for determination of best value for k

7.5.1.1 Read in a random selection of 5000 records from the training set
Exclude records with 3 or fewer clinical events (removes the data set size by about 10 %; further records filtering reduces the number of valid records by a further approximately 10 %).

7.5.1.2 Remove the admin flags

Retain the symptom and diagnosis flags.

7.5.1.3 Inspect the list of events in each record for the presence of a code indicating the presence of the condition of interest.

If there is a code or codes that indicates the presence of the condition of interest, remove those codes from the record and set a ‘has condition’ flag to be TRUE, otherwise retain all event codes and set the flag to be FALSE.

7.5.1.4 For each of the CTV3 code levels previously determined (i.e. 7, 11):

- Create the distance matrix using the distance method of choice (i.e. binomial)
- For each record, order its neighbours according to distance from the record (from nearest neighbour to furthest neighbour)
- For each record, cycle through the possible values of k for the k nearest neighbours, from 1 to the size of the complete record set. At each value for k, calculate the prediction for the presence of the condition according to each of the methods discussed (absolute majority vote; prevalence greater than a range of factors; significantly larger prevalence in the nearest neighbour group).

Results from these runs and calculations across k values and prediction calculation methods were evaluated in order to determine the best factors for each condition prediction. The F2 score was used to determine the best factors.

Figure 7.4 shows the process used to predict presence or absence of a condition for records in the data set.

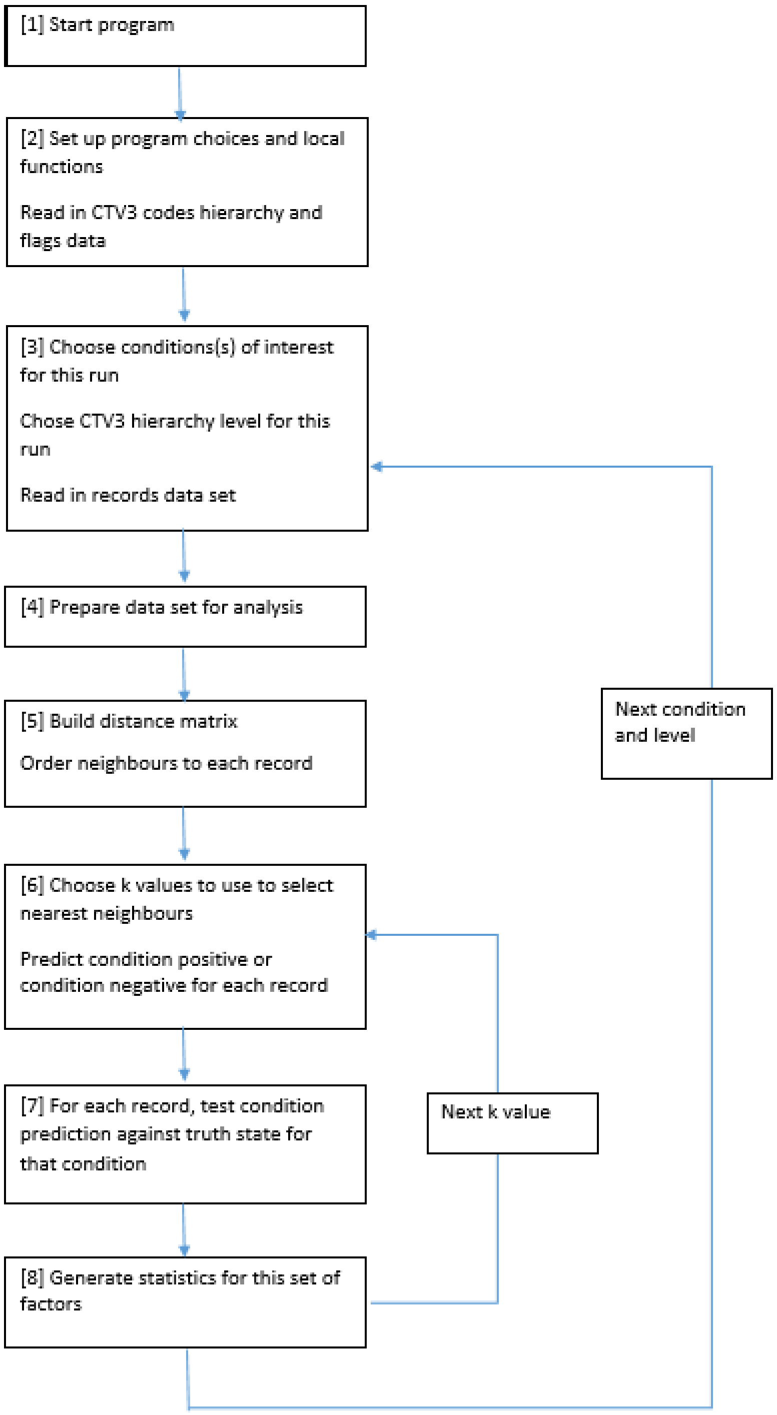

[1] Start program
[2] Set up program choices and local functions
Read in CTV3 codes hierarchy and flags data
[3] Choose conditions(s) of interest for this run
Chose CTV3 hierarchy level for this run
Read in records data set
[4] Prepare data set for analysis
[5] Build distance matrix
Order neighbours to each record
[6] Choose k values to use to select nearest neighbours
Predict condition positive or condition negative for each record
[7] For each record, test condition prediction against truth state for that condition
[8] Generate statistics for this set of factors
Next condition and level
Next k value

Figure 7.4 Process for predicting presence or absence of condition in the set of records

The steps within the program are described:

[1] The program is started.

[2] Initial set-up of program and CTV3 code information

(i) Various flags and factors controlling the program flow are set: whether or not to use the age and gender information in each record; how many records to read and whether the records should be read at random from the data file or sequentially from the start of the file (for reproducibility during program testing), the value of beta in the calculation of the Fbeta score (a value for beta of 2 has been used consistently in this work).

(ii) The R libraries required for particular functions in the program are loaded.

(iii) A set of local functions are defined:

a. calcOddsRatio(TN, FN, TP, FN, p)

This function calculates the odds ratio. It takes as its input the values for true positives, true negatives, false positives and false negatives, and a p-value, and returns an odds ratio together with upper and lower confidence limits calculated using the p-value passed to the function. A default value of 0.05 (giving a confidence interval of 95 %) is used if no p-value is passed to the function. This function has been adapted from the work of Ronald Pearson [268].

b. calcLikelihoodRatio(m, significance level)

   This function calculates the likelihood ratio. It takes as its input a 2x2 matrix of TP, FN, FP and TN values together with a significance level value. It returns a likelihood ratio together with upper and lower confidence limits calculated using the significance level passed to the function. This function has been adapted from the work of Tomas Karpati [269].

c. createBitVectors()

This function creates a binary vector indicating the presence or absence of particular event codes in each record. The vector is in the sequence of the set of event codes used as headings in the table of records read from the records file.

d. getHigherCode()

This function takes as its inputs a CTV3 code and a value for a level in the CTV3 hierarchy. It returns the ancestor code for the input CTV3 code at the requested hierarchy level.

e. includeFlagValue()

This function returns the significance value (i.e. whether it is an administration code, a symptom code or a diagnosis code) of a code from the table of CTV3 codes hierarchy and significance values. It performs a simple lookup of the significance value using the input CTV3 code.

f. TrueFalse()

Checks whether an event code or set of events codes is in the codelist for the target condition.

These functions are called as required in the program.

(iv) The table of CTV3 codes, their ancestor codes in the CTV3 hierarchy, and each codes significance value is read into the program.

[3] (i) The condition of interest is set for this run. The condition of interest must be a condition which has had a codelist of events previously created and stored. It is possible to select a set of conditions, each condition being tested sequentially.

(ii) The level in the CTV3 event code hierarchy at which the analysis is to be performed is selected. The value of the level must be between 1 (the root node) and 19 (the deepest level of the hierarchy). It is possible to select a set of levels, each level being tested sequentially.

(iii) The records set is read into the program. A number of options are available:

(a) The file to be read: the full data set has been split into a training set and a test set

(b) The number of records to be read from the file, or whether to read the complete file

(c) If limiting the number of records to be read, whether these records should be selected randomly from the complete set of records in the file or sequentially from the start of the file.

[4] The set of records read into the program is now prepared for analysis:

(i) Records with fewer than the minimum required number of events are dropped.

(ii) Records are examined for the presence of the condition of interest. Those with the condition (i.e. the record has one or more event codes in the codelist for the condition) are flagged as ‘positive’;

those without any event codes in the codelist are flagged as ‘negative’. Event codes that are in the codelist are dropped.

(iii) Records that now only have event codes that are ‘administration’ codes (i.e. have a significance value of 0) are dropped from the analysis.

(iv) Remaining records now have their event codes mapped to their ancestor codes at the chosen higher level of the CTV3 hierarchy.

(v) A table of all remaining records is built containing their mapped CTV3 codes, age, smoking status and alcohol status. Each code has a column in the table; the presence of the code in any record is indicated by a value of 1 in that record’s row, otherwise a value of 0 is stored.

(vi) If age is included as a dimension for input to the analysis, all age values are normalised, with the maximum value of 1 corresponding to the highest age value in the records set. If age is not to be included, all age values are set to 0.

(vii) If gender is not to be included in the analysis, all values for gender are set to 0.

(xiv) The prevalence of the condition of interest is calculated. This prevalence is used to randomly assign a prediction for presence of the condition to each record: the probability of a record being assigned a condition ‘positive’ value is equal to the condition prevalence in the record set. The random assignations are used to calculate F1, F2 and MCC scores as a baseline for comparison with the scores from the later analysis.

[5] The distance matrix is built and ordered

(i) A distance matrix (or, more precisely, a dissimilarity matrix) is calculated using the table of records and event codes from step 4(iv). The distance matrix calculation method is that suggested from step 6.2.1, the binomial method. This is implemented in the function ‘vegdist’ from the R package ‘vegan’ [267]. For a full discussion of the binomial method see [270].

(ii) The distance matrix is ordered by reverse order of dissimilarity for each record, i.e. each record has a list of all records, reverse ordered by dissimilarity score. So the first record in each record’s list is the record itself, since it will be the least dissimilar record.

[6] The prevalence of the condition in the least dissimilar neighbours is compared to the prevalence in the most dissimilar and appropriate predictions made

(i) A value (or set of values, in the optimisation phase) for ‘k’ is chosen or calculated. k is, by convention, the variable used to the size of the nearest neighbour set.

(ii) For each record, the prevalence of the condition of interest in the k nearest neighbours set is calculated and compared to the prevalence in the remainder of the records set, using Pearson's chi-squared test for significance. A significance value of 0.05 has been used in this work. For those records where the nearest neighbours set has a significantly raised prevalence of the condition of interest, the record is predicted 'positive', otherwise the record is predicted 'negative'. Predictions are recorded against each record.

[7] Test the predictions against the actual presence of the condition

(i) For each record, the prediction is compared to the actual presence of the condition. There are four possible results: a True Positive (TP), where both the prediction and the actual presence are positive; a False Positive (FP) where the prediction was positive but the actual presence was negative; a True Negative (TN), where both the prediction and actual presence were negative; and a False Negative (FN), where the prediction was negative but the original record did, in fact, contain an event code indicative of the condition.

(ii) The sums of each of the numbers of TP, TN, FP and FN results are calculated.

[8] A set of summary statistics is produced

(i) From the TP, FP, TN and FN scores, a number of summary statistical scores are produced. These scores include sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1, F2, MCC. In the optimisation phase, using the training set of records, the two candidate values for the CTV3 level and a range of values for k are used to calculate the predictions, with the combination of CTV3 level and k that produces the best F2 score being chosen as the combination to use in the testing phase.

(ii) In the testing phase, using the testing set of records, additional statistical scores are generated:

(a) The likelihood ratios, positive and negative (LR+ and LR-)

(b) The odds ratio

The likelihood ratios are used to illustrate the success, or otherwise, of this prediction method for each condition investigated.

For each condition, a set of training runs was performed, drawing 5000 records at random from the training set of records. For each run, a range of values of k, from only 1 record in the nearest neighbour group to all records bar one in the nearest neighbours group. For each value of k in each run the number of true positives, true negatives, false positives and false negatives was calculated and from these figures a value for F2 score at each value of k was calculated. The values for F2 from all training runs were plotted against the values for k and a curve fitted to the data using a linear method implemented in the ‘lm’ function from the R ‘stats’ package. The maximum F2 value of this fitted curve was used to select the optimum value of k, i.e. the optimum size of the nearest neighbours group used to predict the state for each record.

7.6 Training runs to determine best k

Each condition is presented separately, since it was not known in advance whether any or all conditions could be grouped together for analysis. For each condition, a chart is shown of F2 score versus value of k for several runs. Each run has one calculation of the distance matrix and subsequent analysis of predictions based on a range of k values from one nearest neighbour to one less than the size of the valid records set. For each condition, test runs are repeated at the derived candidate values for best level of the CTV3 hierarchy, level 7 and level 11.

Following a number of training runs, a curve is fitted to the data points using the linear method, for the level 7 set and the level 11 set. The maximum F2 value across the two curves is used to determine the optimum level and optimum value of k for each condition.

8 ESTIMATION OF CONDITION RISK: RESULTS

8.1 Best factors for clustering method

Results are presented for determination of the optimum level of the CTV3 hierarchy and the optimum number of clusters for the clustering method. The F2 score was used as the metric on which to optimise, which weights sensitivity greater than specificity. This was chosen in order to emphasise the benefit of this technique to select candidates for screening tests. By placing more weight on the calculated sensitivity than on the calculated specificity in choosing the optimum level of the CTV3 hierarchy and the optimum number of clusters, rather than giving them equal weight as would be the case for an F1 score, the method is intended to exclude fewer candidates for screening than would be the case than if the F1 score was used to determine the optimum factors.

8.1.1 Acute sinusitis

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for stress are shown in Figure 8.1 and Figure 8.2.

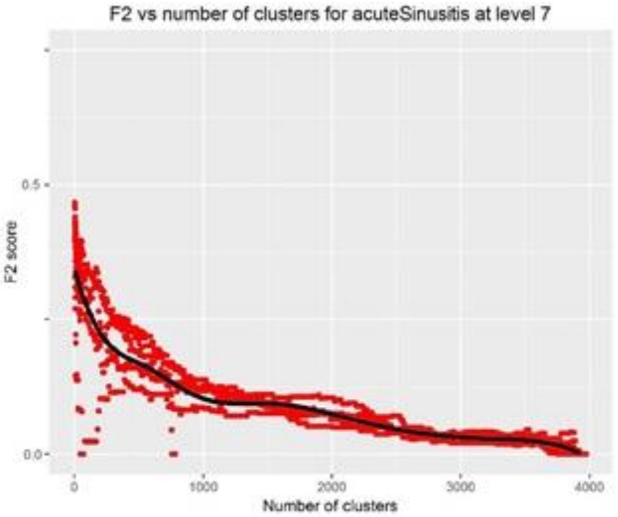


Figure 8.1 F2 score versus number of clusters for acute sinusitis at CTV3 hierarchy level 7

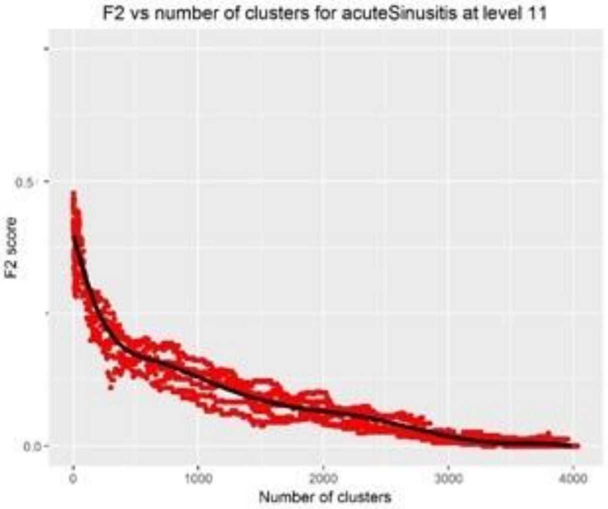


Figure 8.2 F2 score versus number of clusters for acute sinusitis at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.4587 with 2 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.4663 with 2 clusters.

Selected factors were therefore CTV3 hierarchy level 11 with 2 clusters.

8.1.2 Allergic rhinitis

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for allergic rhinitis are shown Figure 8.3 and Figure 8.4.

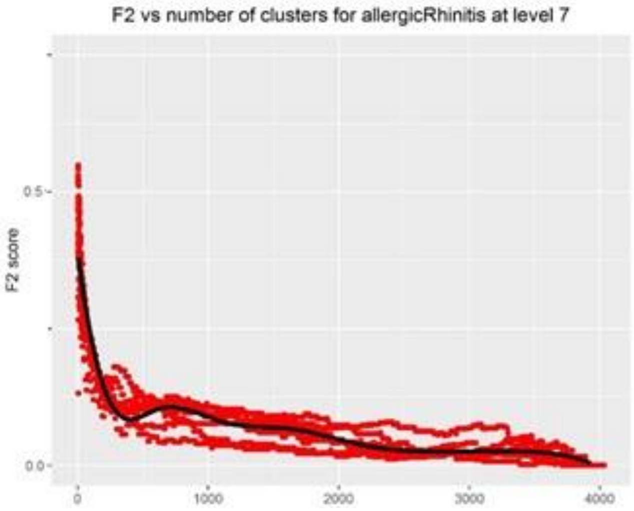


Figure 8.3 F2 score versus number of clusters for allergic rhinitis at CTV3 hierarchy level 11

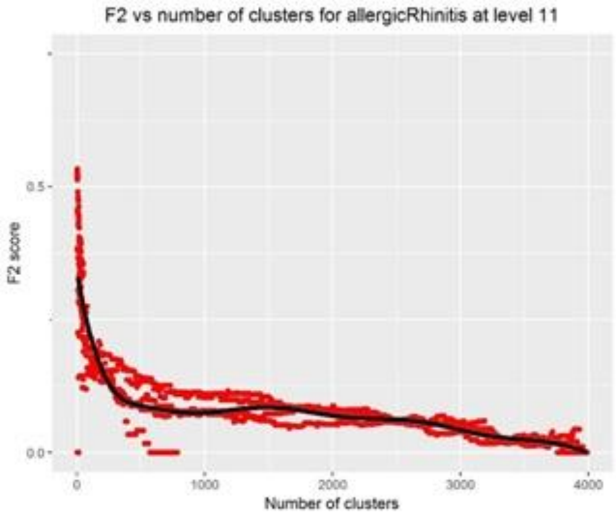


Figure 8.4 F2 score versus number of clusters for allergic rhinitis at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.5242 with 2 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.5289 with 3 clusters.

Selected factors were therefore CTV3 hierarchy level 11 with 3 clusters.

8.1.3 Any cancer

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for any cancer are shown Figure 8.5 and Figure 8.6.

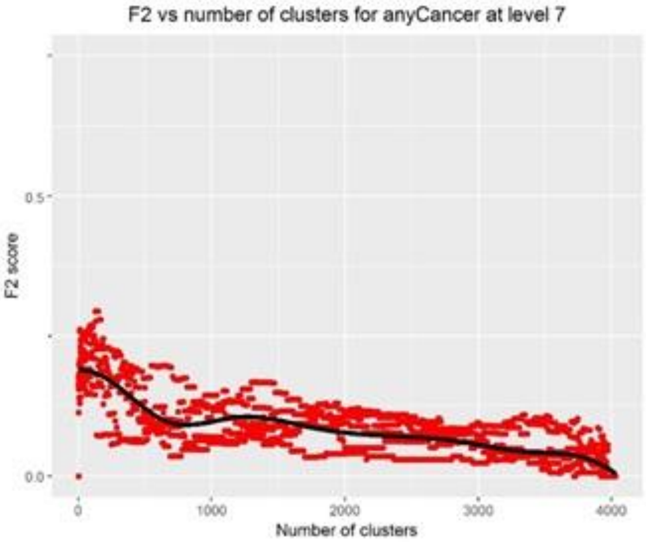


Figure 8.5 F2 score versus number of clusters for any cancer at CTV3 hierarchy level 7

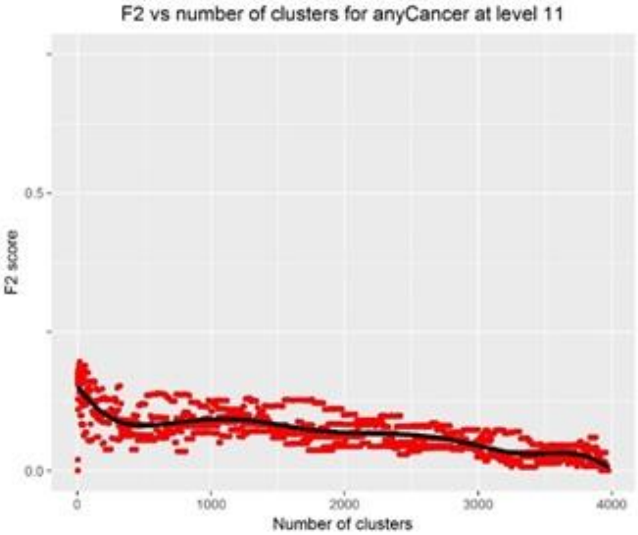


Figure 8.6 F2 score versus number of clusters for any cancer at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.2393 with 130 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.1759 with 14 clusters.

Selected factors were therefore CTV3 hierarchy level 7 with 130 clusters.

8.1.4 Asthma

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for asthma are shown in Figure 8.7 and Figure 8.8.

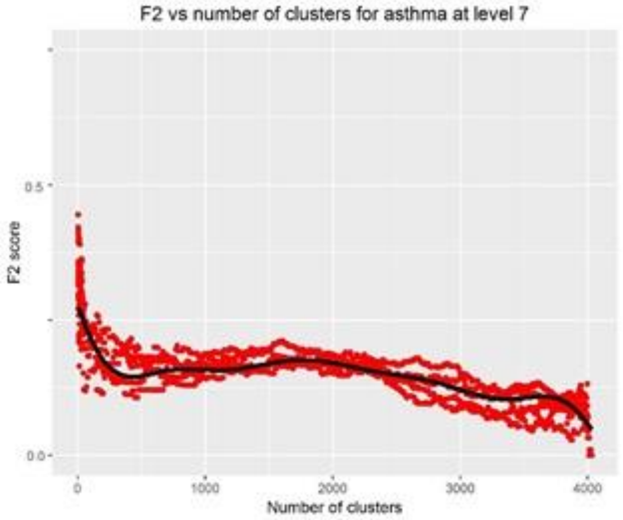


Figure 8.7 F2 score versus number of clusters for thyrotoxicosis at CTV3 hierarchy level 7

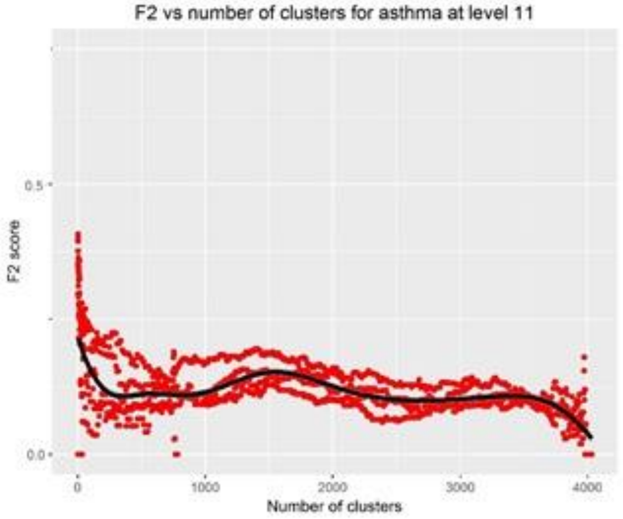


Figure 8.8 F2 score versus number of clusters for asthma at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.4023 with 2 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.3170 with 3 clusters.

Selected factors were therefore CTV3 hierarchy level 7 with 2 clusters.

8.1.5 Autism spectrum disorder

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for autism are shown Figure 8.9 and Figure 8.10.

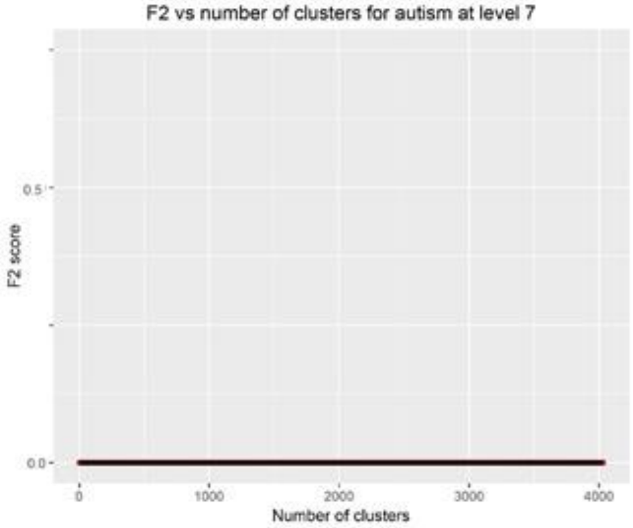


Figure 8.9 F2 score versus number of clusters for autism at CTV3 hierarchy level 7

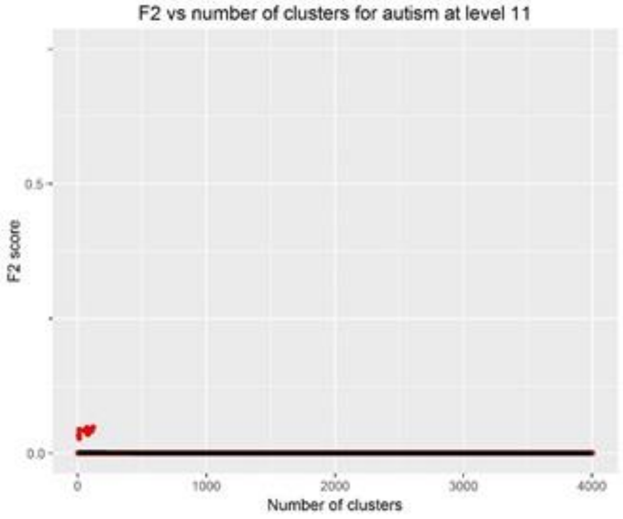


Figure 8.10 F2 score versus number of clusters for autism at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0 with NA clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.0170 with 90 clusters.

Selected factors were therefore CTV3 hierarchy level 11 with 90 clusters.

8.1.6 Breast cancer

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for breast cancer are shown in Figure 8.11 and Figure 8.12.

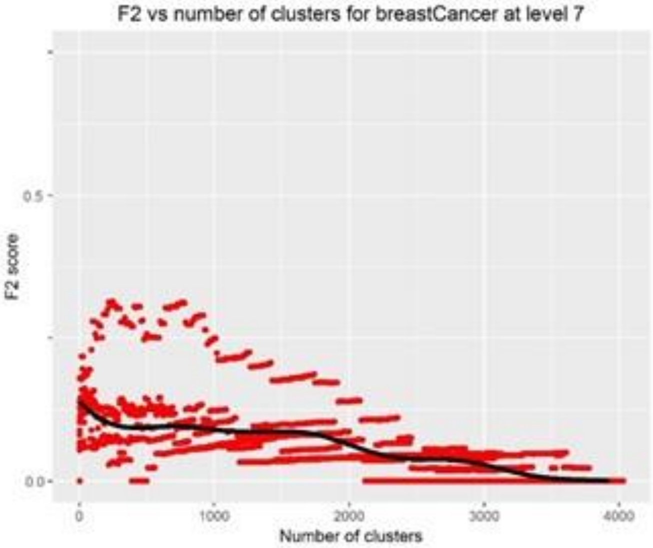


Figure 8.11 F2 score versus number of clusters for breast cancer at CTV3 hierarchy level 7

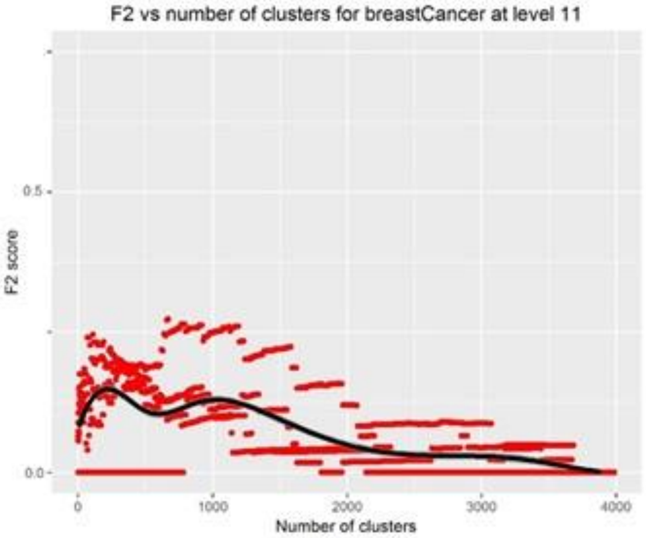


Figure 8.12 F2 score versus number of clusters for breast cancer at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.1469 with 26 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.1619 with 920 clusters.

Selected factors were therefore CTV3 hierarchy level 11 with 920 clusters.

8.1.7 Bronchitis

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for bronchitis are shown Figure 8.13 and Figure 8.14.

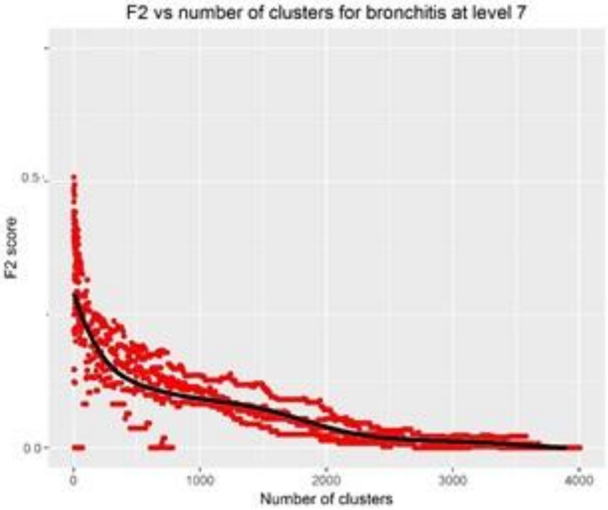


Figure 8.13 F2 score versus number of clusters for bronchitis at CTV3 hierarchy level 7

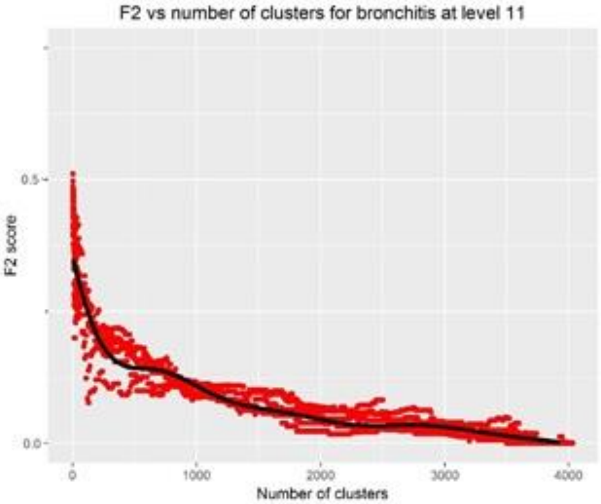


Figure 8.14 F2 score versus number of clusters for bronchitis at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.4807 with 2 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.4832 with 3 clusters.

Selected factors were therefore CTV3 hierarchy level 11 with 3 clusters.

8.1.8 Colon cancer

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for colon cancer are shown in Figure 8.15 and Figure 8.16.

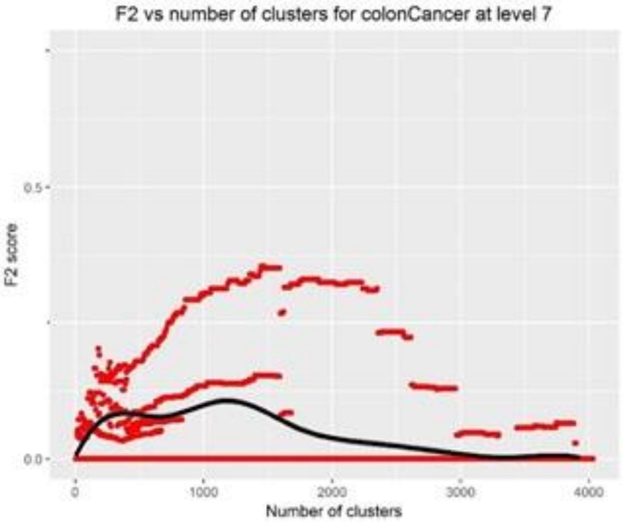


Figure 8.15 F2 score versus number of clusters for colon cancer at CTV3 hierarchy level 7

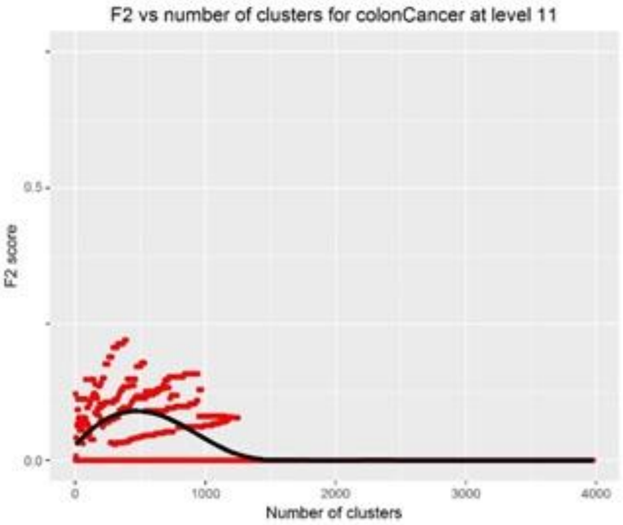


Figure 8.16 F2 score versus number of clusters for colon cancer at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.1077 with 1450 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.0972 with 380 clusters.

Selected factors were therefore CTV3 hierarchy level 7 with 380 clusters.

8.1.9 Eczema

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for eczema are shown in Figure 8.17 and Figure 8.18.

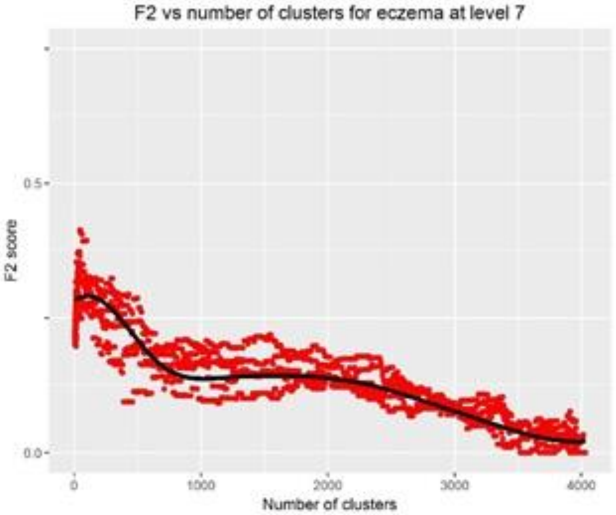


Figure 8.17 F2 score versus number of clusters for eczema at CTV3 hierarchy level 11

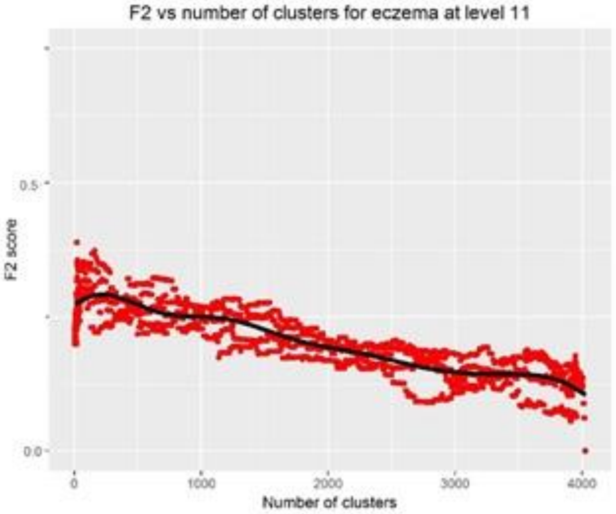


Figure 8.18 F2 score versus number of clusters for eczema at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.3301 with 47 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.3255 with 70 clusters.

Selected factors were therefore CTV3 hierarchy level 7 with 47 clusters.

8.1.10 Gastroparesis

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for gastroparesis are shown in Figure 8.19 and Figure 8.20.

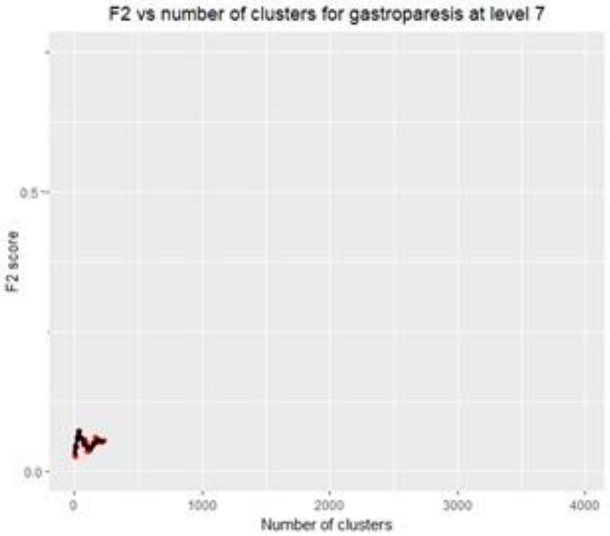


Figure 8.19 F2 score versus number of clusters for thyrotoxicosis at CTV3 hierarchy level 7

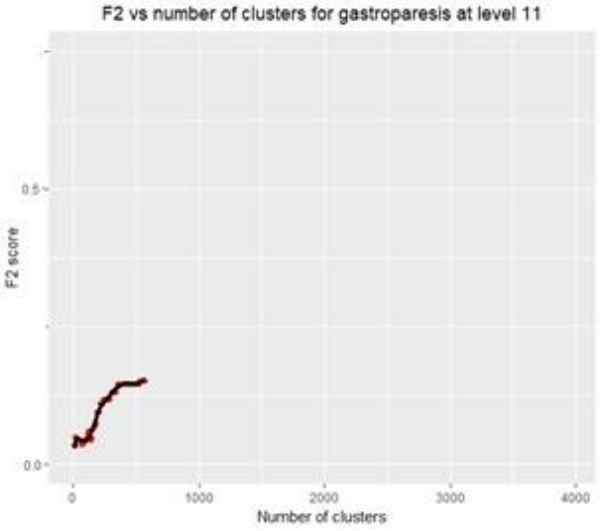


Figure 8.20 F2 score versus number of clusters for gastroparesis at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.0704 with 13 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.1515 with 530 clusters.

Selected factors were therefore CTV3 hierarchy level 11 with 530 clusters.

8.1.11 Gastro-oesophageal reflux disease

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for gastro-oesophageal reflux disease are shown in Figure 8.21 and Figure 8.22.

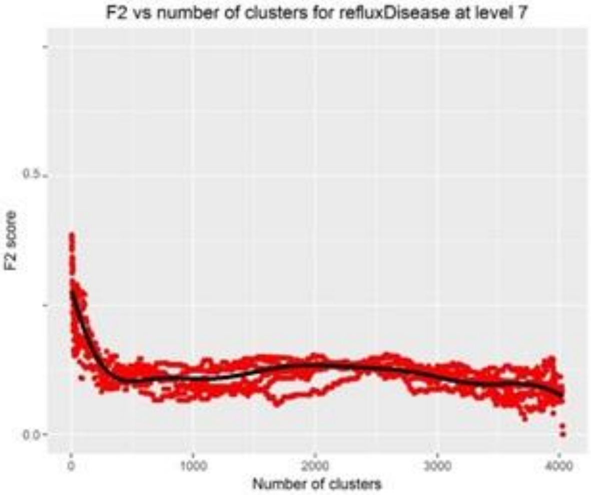


Figure 8.21 F2 score versus number of clusters for gastro-intestinal reflux disease at CTV3 hierarchy level 7

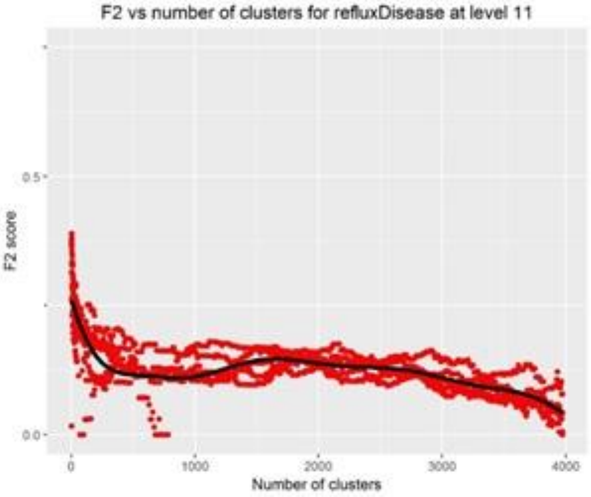


Figure 8.22 F2 score versus number of clusters for gastro-intestinal reflux disease at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.3742 with 2 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.3792 with 3 clusters.

Selected factors were therefore CTV3 hierarchy level 11 with 3 clusters.

8.1.12 Gout

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for gout are shown in Figure 8.23 and Figure 8.24.

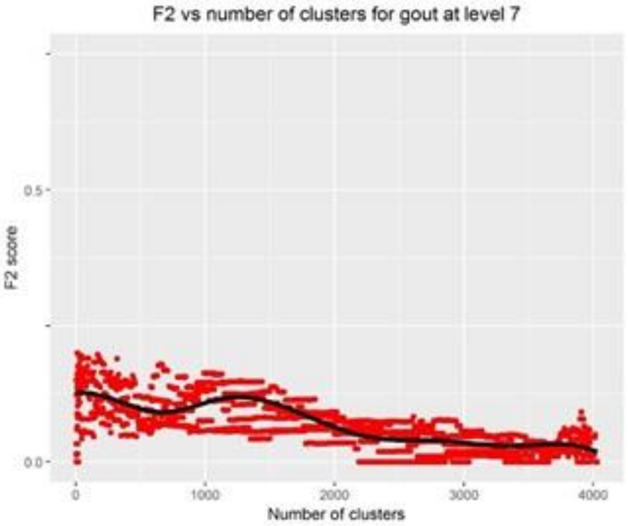


Figure 8.23 F2 score versus number of clusters for thyrotoxicosis at CTV3 hierarchy level 7

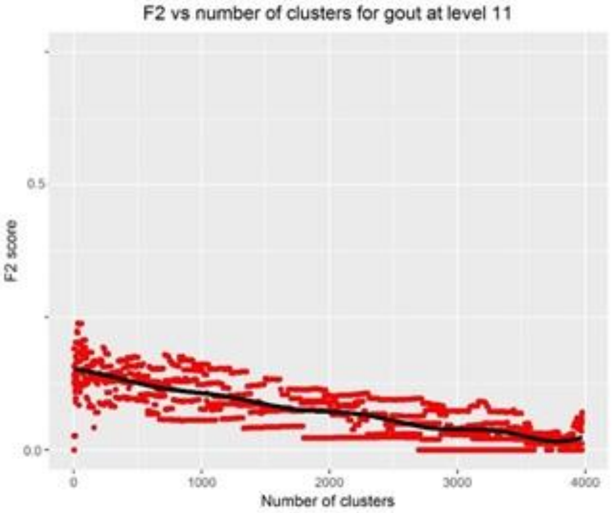


Figure 8.24 F2 score versus number of clusters for gout at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.1764 with 34 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.1864 with 37 clusters.

Selected factors were therefore CTV3 hierarchy level 11 with 37 clusters.

8.1.13 Obesity

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for obesity are shown in Figure 8.25 and Figure 8.26.

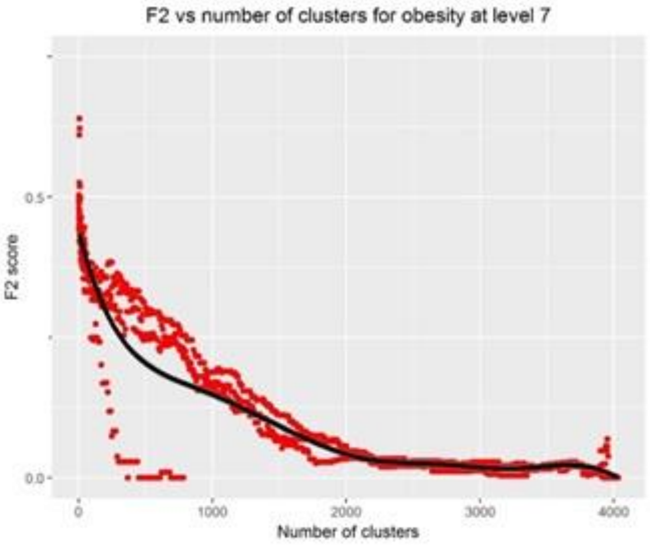


Figure 8.25 F2 score versus number of clusters for obesity at CTV3 hierarchy level 7

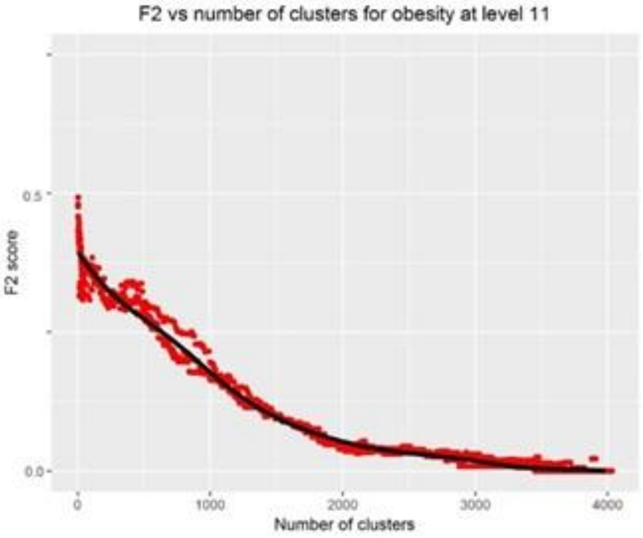


Figure 8.26 F2 score versus number of clusters for obesity at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.4820 with 2 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.4903 with 3 clusters.

Selected factors were therefore CTV3 hierarchy level 11 with 3 clusters.

8.1.14 Osteoarthritis

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for osteoarthritis are shown in Figure 8.27 and Figure 8.28.

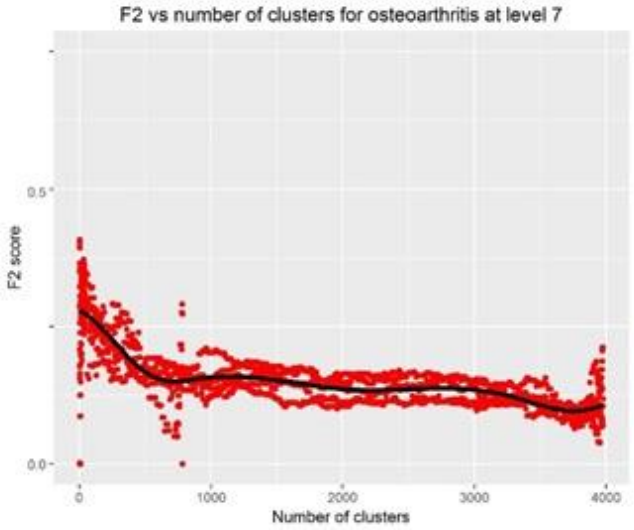


Figure 8.27 F2 score versus number of clusters for osteoarthritis at CTV3 hierarchy level 7

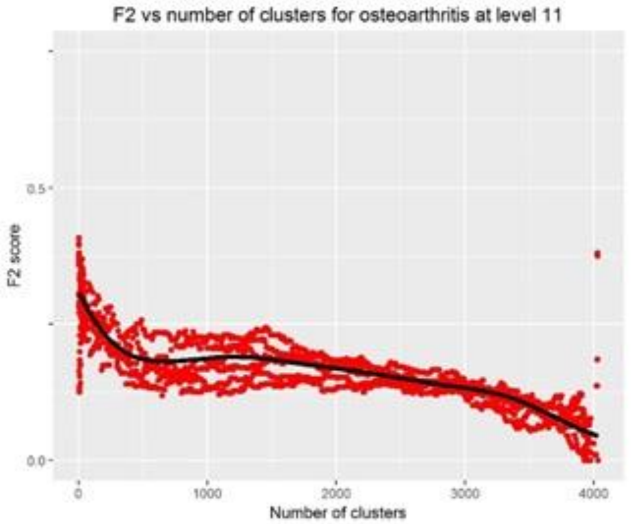


Figure 8.28 F2 score versus number of clusters for osteoarthritis at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.3978 with 2 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.3958 with 2 clusters.

Selected factors were therefore CTV3 hierarchy level 7 with 2 clusters.

8.1.15 Prader-Willi disease

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for Prader-Willi disease are shown in Figure 8.29 and Figure 8.30.

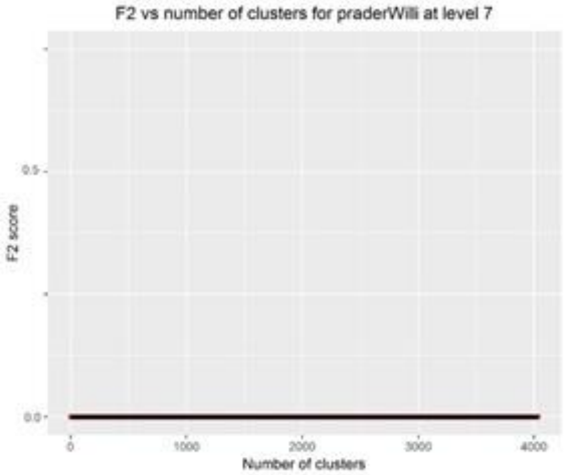


Figure 8.29 F2 score versus number of clusters for Prader-Willi disease at CTV3 hierarchy level 7

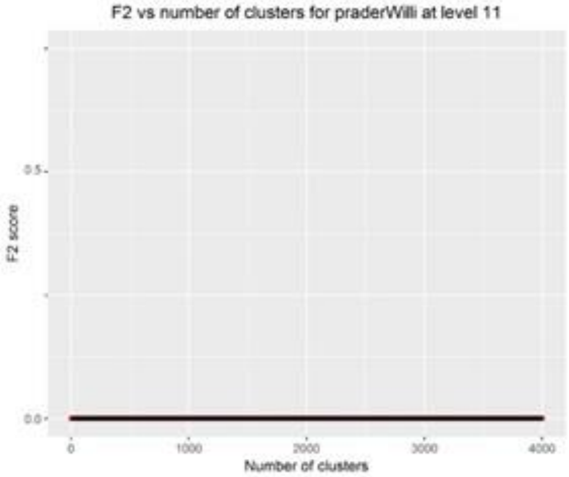


Figure 8.30 F2 score versus number of clusters for Prader-Willi disease at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0 with NA clusters.

At CTV3 hierarchy level 11, maximum F2 was 0 with NA clusters.

The low prevalence of Prader-Willi disease in the data set meant that it was not possible to deduce the optimum number of clusters.

8.1.16 Prostate cancer

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for prostate cancer are shown in Figure 8.31 and Figure 8.32.

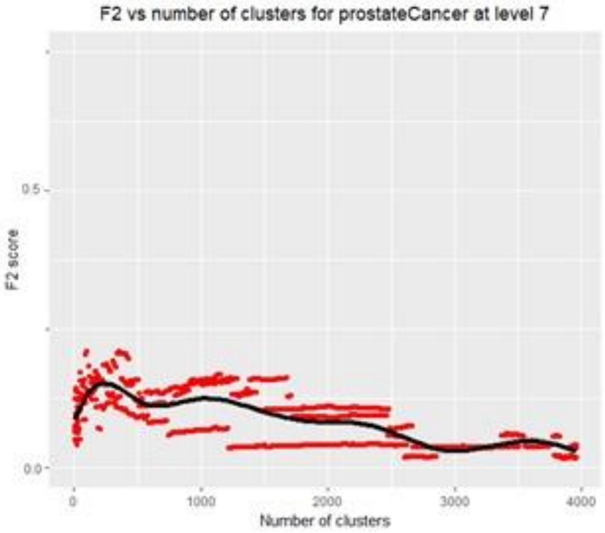


Figure 8.31 F2 score versus number of clusters for prostate cancer at CTV3 hierarchy level 11

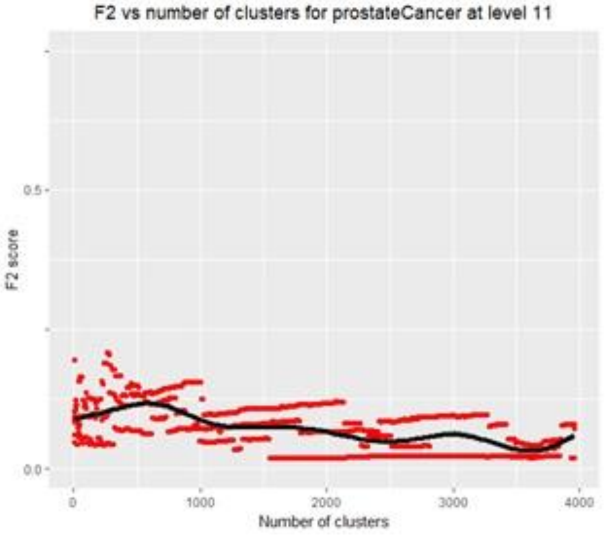


Figure 8.32 F2 score versus number of clusters for prostate cancer at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.1614 with 280 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.1278 with 514 clusters.

Selected factors were therefore CTV3 hierarchy level 7 with 280 clusters.

8.1.17 Stress

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for stress are shown in Figure 8.33 and Figure 8.34.

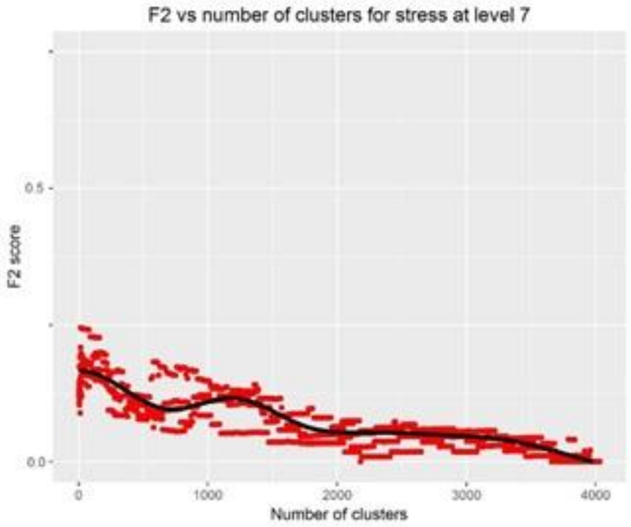


Figure 8.33 F2 score versus number of clusters for thyrotoxicosis at CTV3 hierarchy level 7

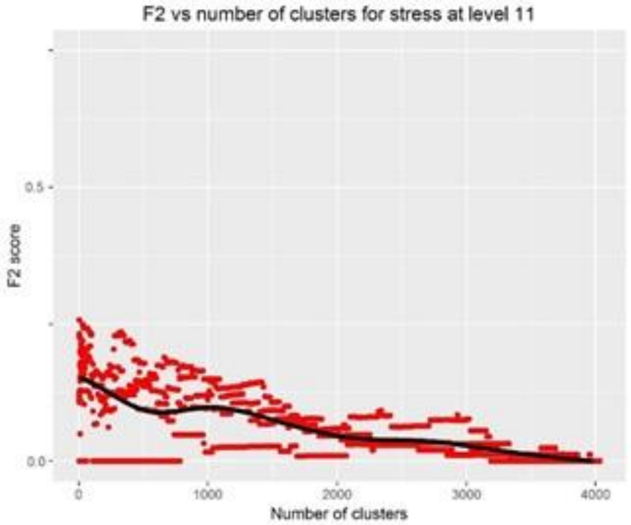


Figure 8.34 F2 score versus number of clusters for stress at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.1971 with 150 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.1902 with 70 clusters.

Selected factors were therefore CTV3 hierarchy level 7 with 150 clusters.

8.1.18 Thyrotoxicosis

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for thyrotoxicosis are shown in Figure 8.35 and Figure 8.36.

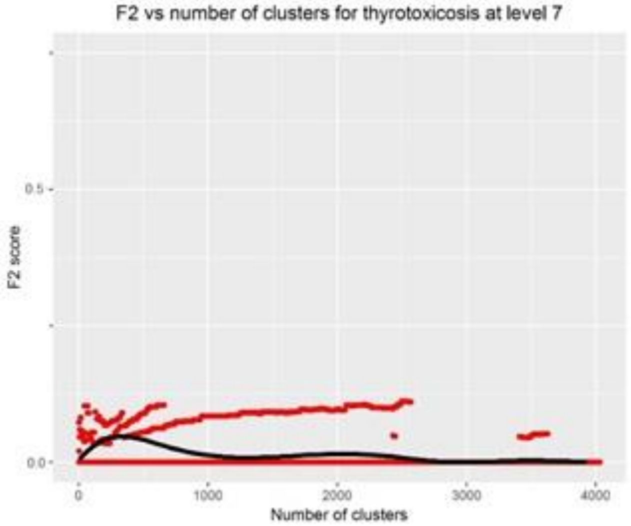


Figure 8.35 F2 score versus number of clusters for thyrotoxicosis at CTV3 hierarchy level 7

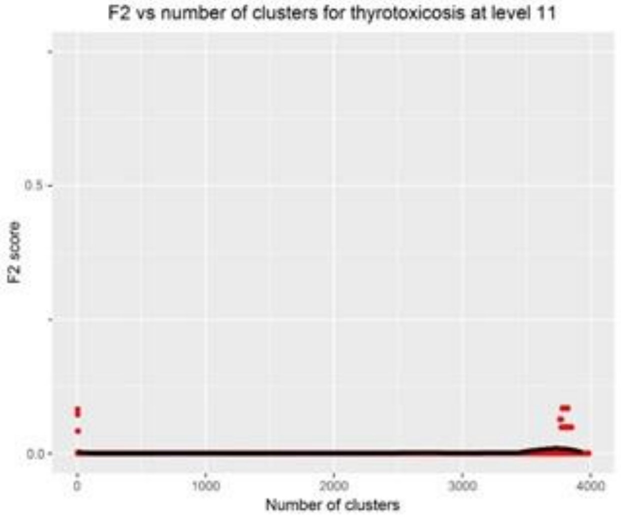


Figure 8.36 F2 score versus number of clusters for thyrotoxicosis at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.0599 with 660 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.0519 with 2 clusters.

Selected factors were therefore CTV3 hierarchy level 7 with 660 clusters.

8.1.19 Type 2 diabetes

Results of the analysis of F2 score versus the number of clusters at CTV3 hierarchy levels of 7 and 11 for type 2 diabetes are shown in Figure 8.37 and Figure 8.38.

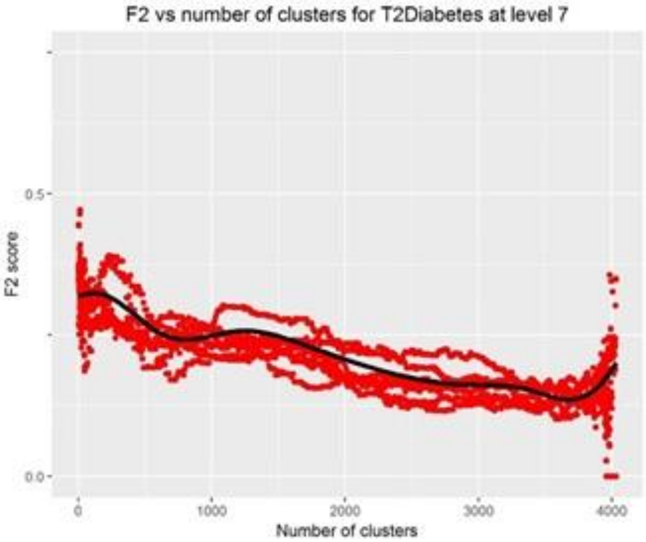


Figure 8.37 F2 score versus number of clusters for type 2 diabetes at CTV3 hierarchy level 7

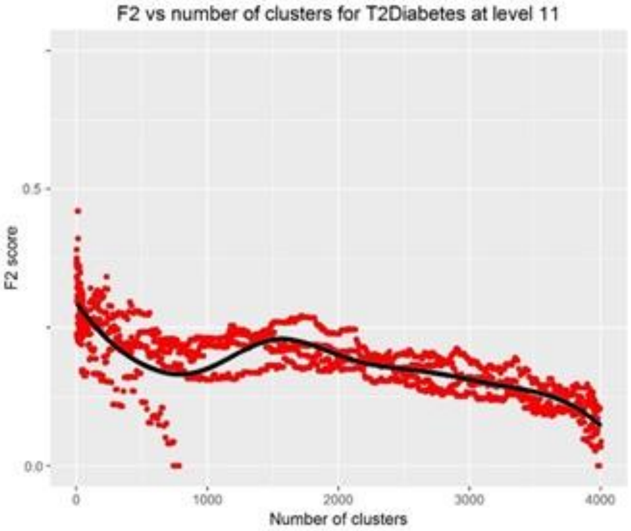


Figure 8.38 F2 score versus number of clusters for type 2 diabetes at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.3642 with 2 clusters.

At CTV3 hierarchy level 11, maximum F2 was 0.3730 with 3 clusters.

Selected factors were therefore CTV3 hierarchy level 11 with 2 clusters.

8.1.20 Summary of training runs for best number of clusters

The final set of best factors for each condition based on analysis of the training set has now been produced and is summarised in Table 6.25. Note that for two conditions, Prader-Willi Disease and autism (at level 7 only), the method failed to produce clusters. These conditions are the lowest prevalence conditions in the composite data set, and had no positive cases in the training set, thus the method was unable to determine the optimum number of clusters.

Condition	Best CTV3 level	Best no of clusters at CTV3 level	F2 at best level and k	Prevalence in training set
Acute sinusitis	11	2	0.466	0.154
Allergic rhinitis	11	3	0.529	0.198
Any cancer	7	130	0.239	0.042
Asthma	7	2	0.402	0.122
Autism	11	90	0.017	0.001
Breast cancer	11	920	0.162	0.011
Bronchitis	7	3	0.483	0.159
Colon cancer	7	359	0.108	0.005
Eczema	7	104	0.330	0.054
Gastroparesis	11	530	0..152	0.002
Gout	11	37	0.186	0.024
Obesity	11	3	0.490	0.160
Osteoarthritis	7	2	0.398	0.117
Prader-Willi	-	-	-	-
Prostate cancer	7	280	0.161	0.111
Reflux disease	11	3	0.379	0.109
Stress	7	150	0.197	0.026
Thyrotoxicosis	7	359	0.060	0.006
Type 2 diabetes	11	3	0.373	0.102

Table 8.1 Summary of results for optimum factors for the clustering method

8.2 Results from test runs for the clustering approach

As discussed in 7.3.8, there are several useful indicators to show the success or otherwise of a test in predicting the presence or absence of a condition. A key indicator of the success of a screening test is the likelihood ratio (LR). The likelihood ratio is the ratio of the probability of the test result in diseased persons over the probability of the test result in a non-diseased person (see e.g. Shortliffe and Perreault [271] for a description of likelihood ratios and their calculation, and Jacobs et al [147] for an illustration of their use in practice). More specifically, the positive likelihood ratio (LR+) and the negative likelihood ratio (LR-) are defined as

LR+ = probability that the test is positive in diseased persons
probability that the test is positive in non-diseased persons
= TPR/FPR

LR- = probability that the test is negative in diseased persons
probability that the test is negative in non-diseased persons
= FNR/TNR

A test that discriminates well between those persons with a disease and those without the disease will have an LR+ much greater than one; likewise an effective test will have an LR- much less than one. A likelihood ratio (+ or -) of 1 indicates that the test has no value in discriminating between those with the disease and those without.

Results are presented as positive and negative likelihood ratios, and as odds ratios. This is a standard way of presenting such results in the literature (see, for example, [147, 272, 273], with the likelihood ratios being calculated from the sensitivity and specificity values, and the odds ratio being calculated from the ratio of the positive and negative likelihood ratios (and so, by extension, also from the specificity and sensitivity). Clark [274] has a discussion of the calculation of likelihood ratios.

For each condition, the likelihood ratios are calculated. The positive likelihood ratio is then used to calculate a post-test prevalence of the condition for those records predicted to be positive for the condition; the negative likelihood ratio is used to calculate a post-test prevalence of the condition for those records predicted to be negative for the condition. Results of these tests are presented as Fagan nomograms. Note that the 95 % confidence intervals for likelihood ratios are calculated; should the range of any of these confidence intervals include 1 (indicating that the test has no discriminatory value), it is concluded that the likelihood ratio is not significantly different from 1 and so the test cannot be said to have a significant discriminatory value. Results are for five runs per condition, each run drawing a random sample of 5000 records from the training set of records. Note that in the Fagan nomograms shown here, the positive likelihood is shown in red and the negative likelihood is shown in green, with no change to likelihood shown by a single black line.

8.2.1 Acute sinusitis

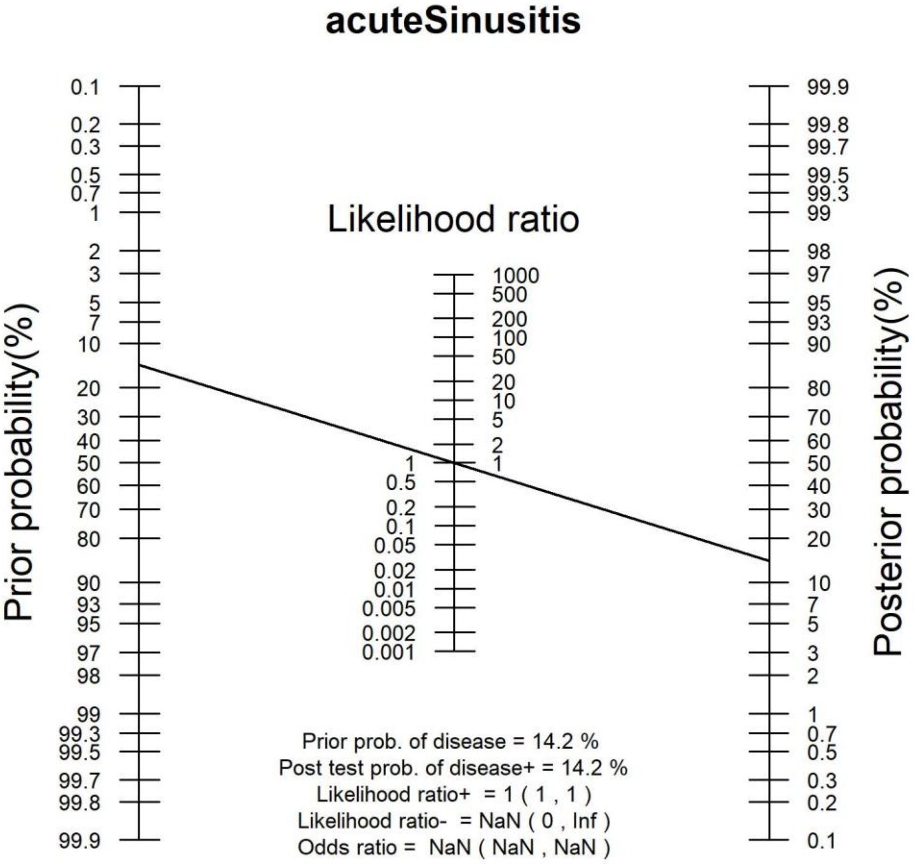


Figure 8.39 Results for acute sinusitis using clustering

True positives: 565; True negatives: 0; False positives: 3414; False negatives: 0

Sensitivity: 1; Specificity: 0

F1: 0.249; F2: 0.453; MCC: 0

Positive predictive value: 0.142; Negative predictive value: -

Positive likelihood ratio: 1 with 95 % CI: 1 to 1

Negative likelihood ratio: - with 95 % CI: 0 to >100

Odds ratio: - with 95 % CI: - to -

8.2.2 Allergic rhinitis

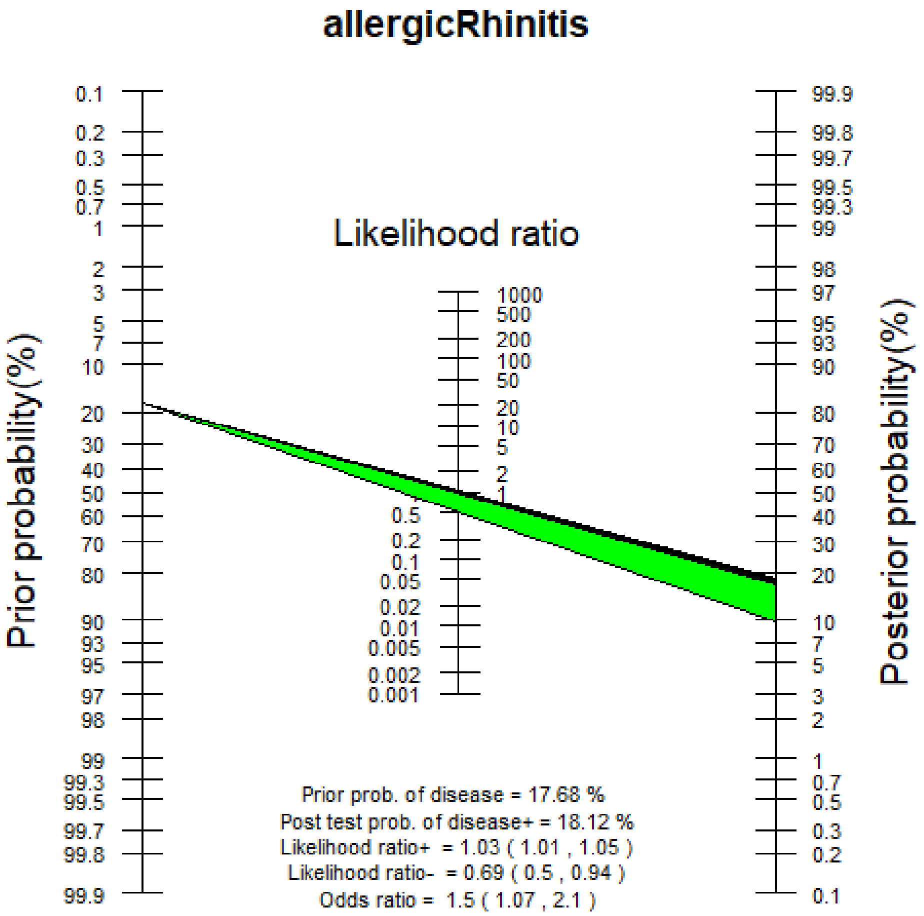


Figure 8.40 Results for allergic rhinitis using clustering

True positives: 667; True negatives: 278; False positives: 3018; False negatives: 41

Sensitivity: 0.942; Specificity: 0.084

F1: 0.304; F2: 0.5115; MCC: 0.037

Positive predictive value: 0.181; Negative predictive value: 0.871

Positive likelihood ratio: 1.03 with 95 % CI: 1.01 to 1.05

Negative likelihood ratio: 0.69 with 95 % CI: 0.5 to 0.94

Odds ratio: 1.5 with 95 % CI: 1.07 to 2.1

8.2.3 Any cancer

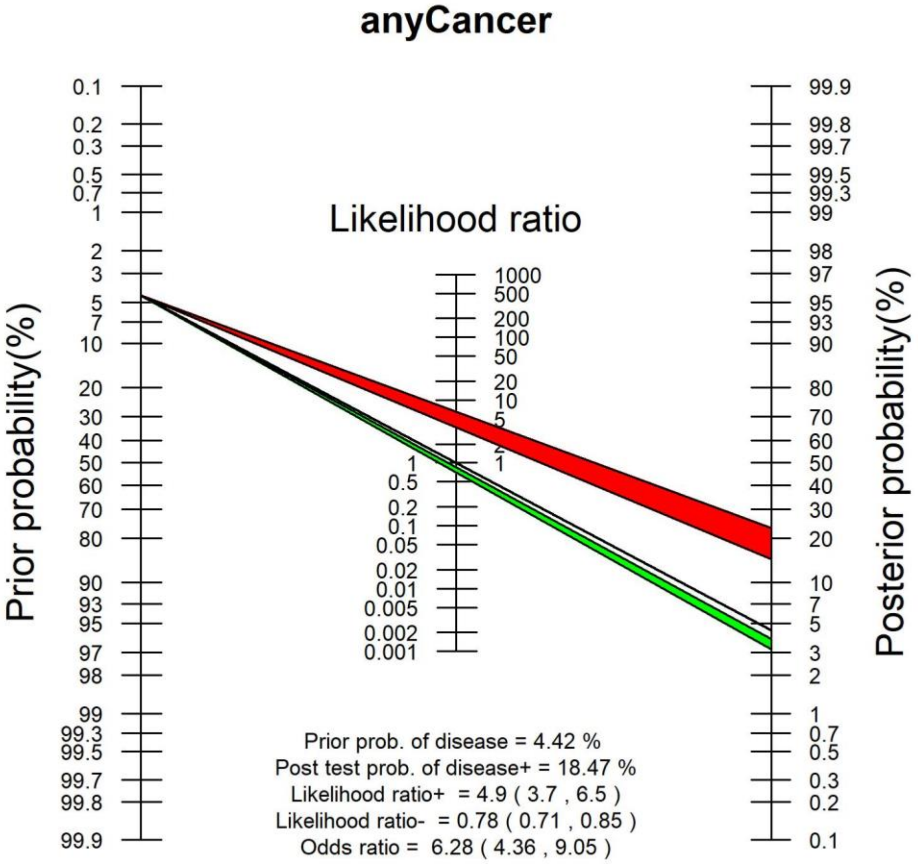


Figure 8.41 Results for any cancer using clustering

True positives: 46; True negatives: 3604; False positives: 232; False negatives: 130

Sensitivity: 0.261; Specificity: 0.947

F1: 0.216; F2: 0.241; MCC: 0.177

Positive predictive value: 0.185; Negative predictive value: 0.965

Positive likelihood ratio: 4.90 with 95 % CI: 3.70 to 6.50

Negative likelihood ratio: 0.78 with 95 % CI: 0.71 to 0.85

Odds ratio: 6.28 with 95 % CI: 4.36 to 9.05

8.2.4 Asthma

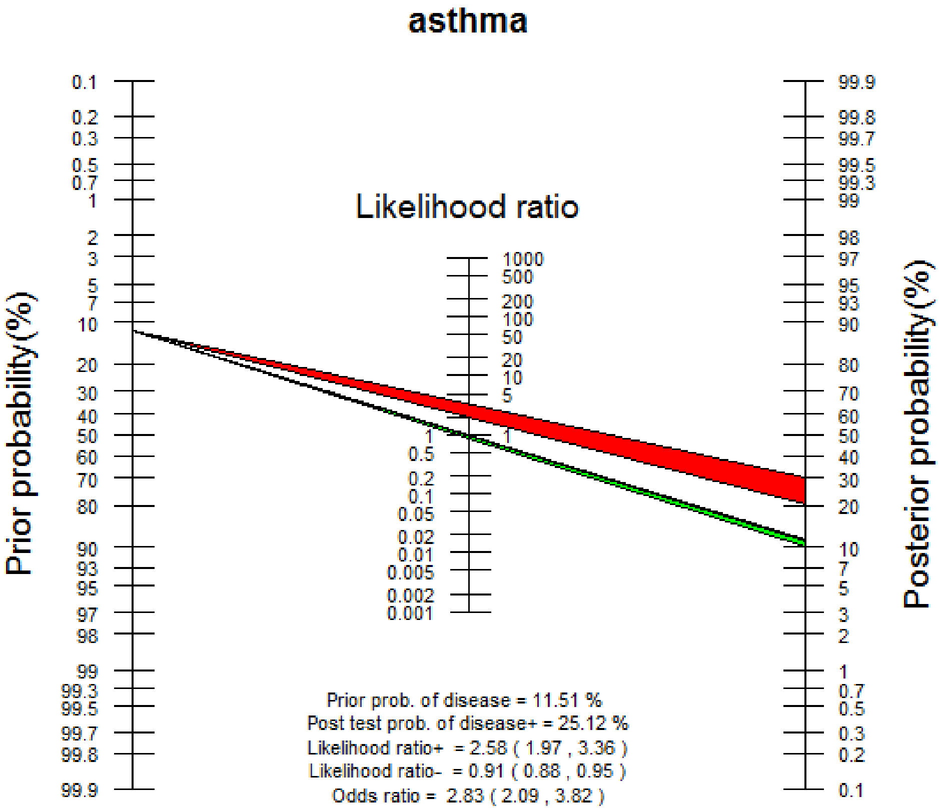


Figure 8.42 Results for asthma using clustering

True positives: 475; True negatives: 0; False positives: 3522; False negatives: 0

Sensitivity: 1; Specificity: 0

F1: 0.212; F2: 0.403; MCC: -

Positive predictive value: 0.119; Negative predictive value: -

Positive likelihood ratio: 2.58 with 95 % CI: 1.97 to 3.36

Negative likelihood ratio: 0.91 with 95 % CI: 0.88 to 0.95

Odds ratio: 2.83 with 95 % CI: 2.09 to 3.82

8.2.5 Autism spectrum disorder

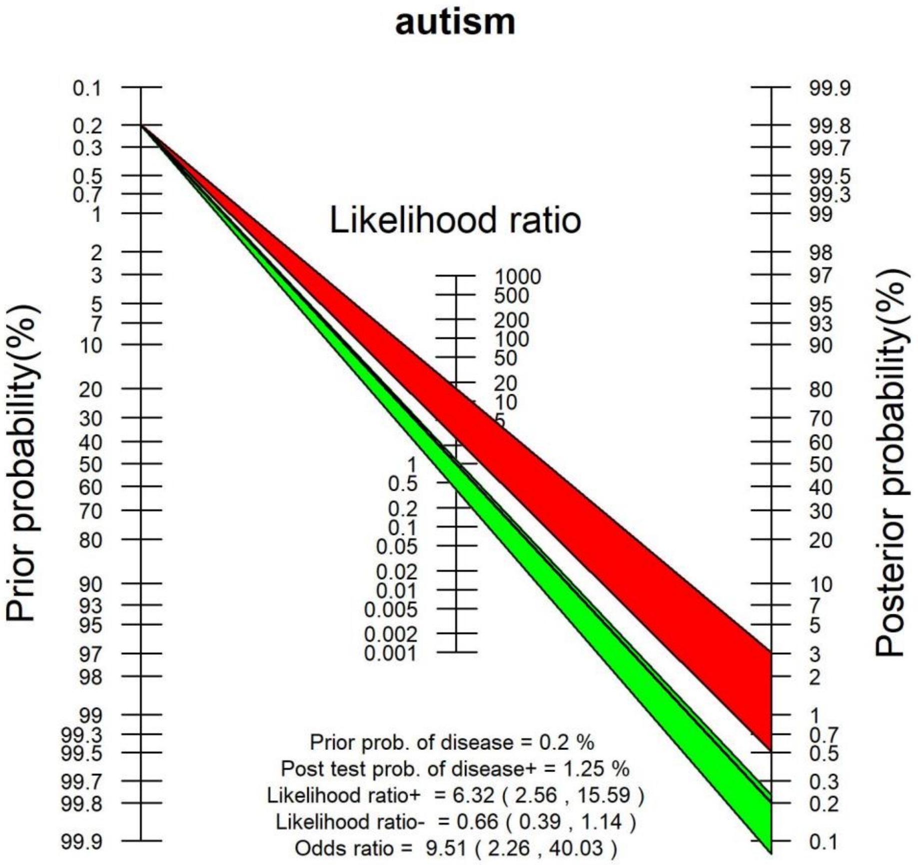


Figure 8.43 Results for autism using clustering

True positives: 3; True negatives: 3756; False positives: 237; False negatives: 5

Sensitivity: 0.375; Specificity: 0.941

F1: 0.024; F2: 0.05411; MCC: 0.059

Positive predictive value: 0.013; Negative predictive value: 0.999

Positive likelihood ratio: 6.32 with 95 % CI: 2.56 to 15.59

Negative likelihood ratio: 0.66 with 95 % CI: 0.39 to 1.14

Odds ratio: 9.51 with 95 % CI: 2.26 to 40.03

8.2.6 Breast cancer

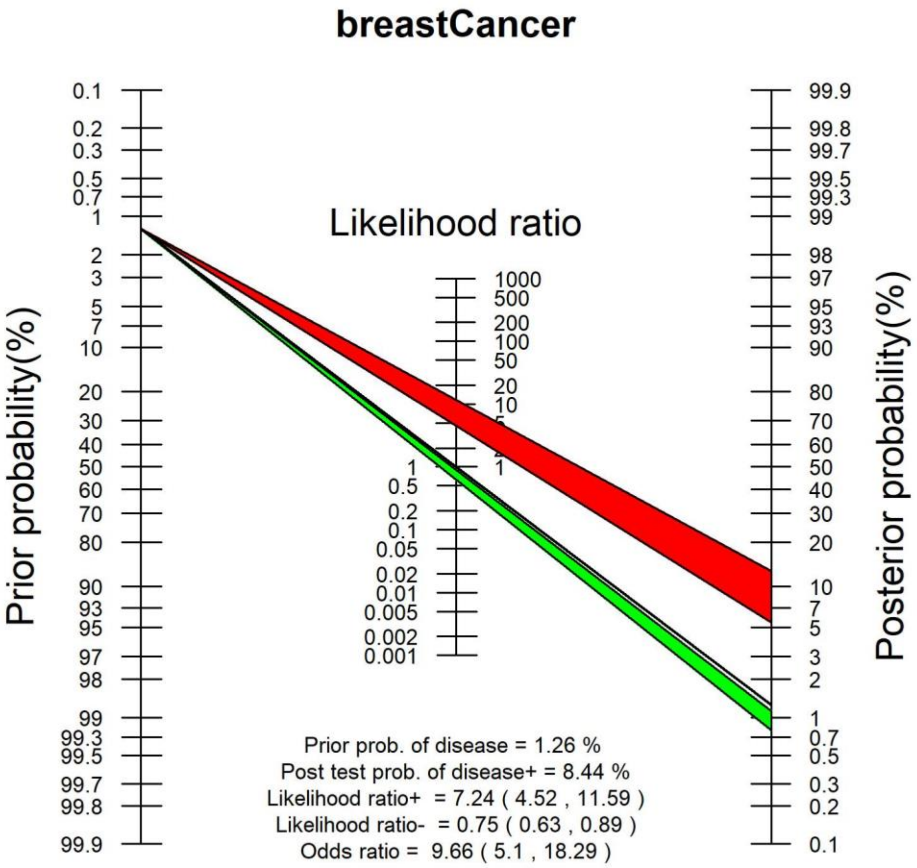


Figure 8.44 Results for breast cancer using clustering

True positives: 14; True negatives: 3776; False positives: 152; False negatives: 36

Sensitivity: 0.280; Specificity: 0.961

F1: 0.130; F2: 0.174; MCC: 0.134

Positive predictive value: 0.084; Negative predictive value: 0.991

Positive likelihood ratio: 7.24 with 95 % CI: 4.52 to 11.59

Negative likelihood ratio: 0.75 with 95 % CI: 0.63 to 0.89

Odds ratio: 9.66 with 95 % CI: 5.10 to 18.29

8.2.7 Bronchitis

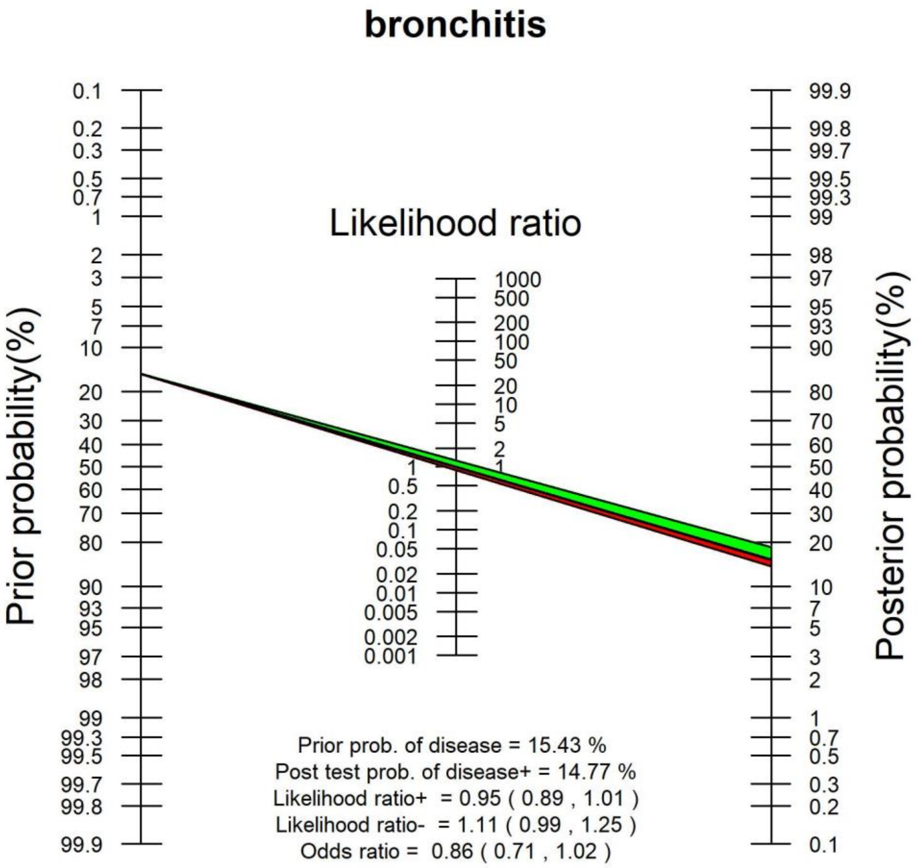


Figure 8.45 Results for bronchitis using clustering

True positives: 400; True negatives: 1082; False positives: 2311; False negatives: 219

Sensitivity: 0.646; Specificity: 0.319

F1: 0.240; F2: 0.386; MCC: -0.027

Positive predictive value: 0.269; Negative predictive value: 0.988

Positive likelihood ratio: 0.95 with 95 % CI: 0.89 to 1.01

Negative likelihood ratio: 1.11 with 95 % CI: 0.71 to 1.02

Odds ratio: 0.86 with 95 % CI: 0.71 to 1.02

8.2.8 Colon cancer

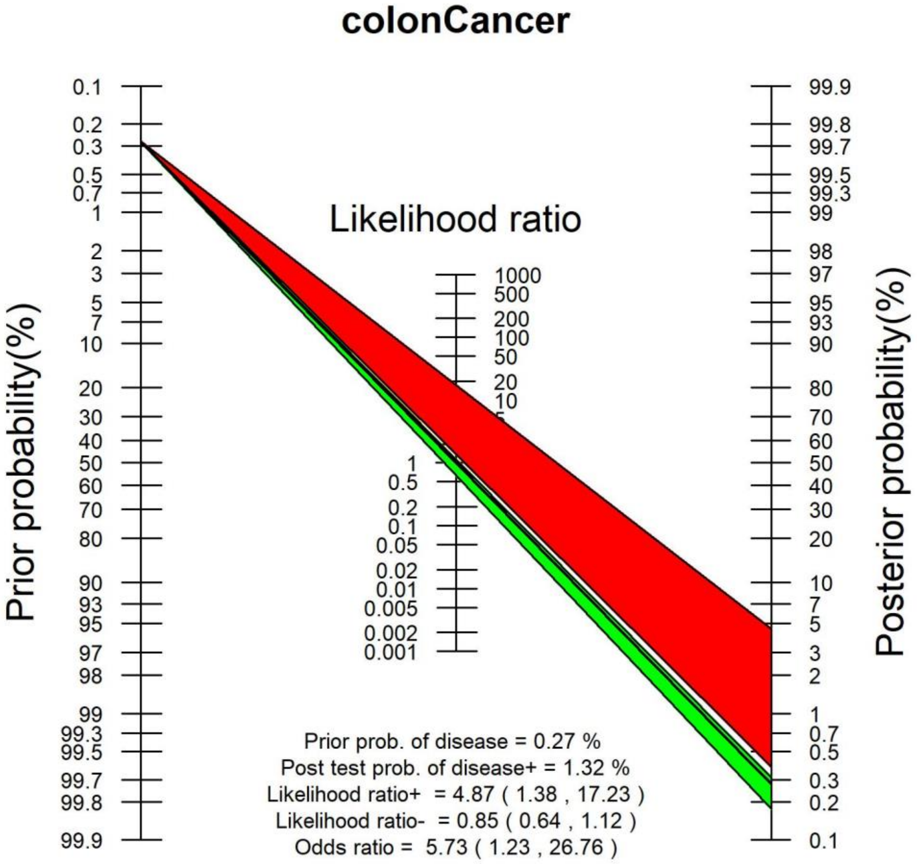


Figure 8.46 Results for colon cancer using clustering

True positives: 2; True negatives: 3843; False positives: 149; False negatives: 9

Sensitivity: 0; Specificity: 0.967

F1: 0; F2: 0; MCC: -0.010

Positive predictive value: 0; Negative predictive value: 0.997

Positive likelihood ratio: 4.87 with 95 % CI: 1.38 to 17.23

Negative likelihood ratio: 0.85 with 95 % CI: 0.64 to 1.12

Odds ratio: 5.73 with 95 % CI: 1.23 to 26.76

8.2.9 Eczema

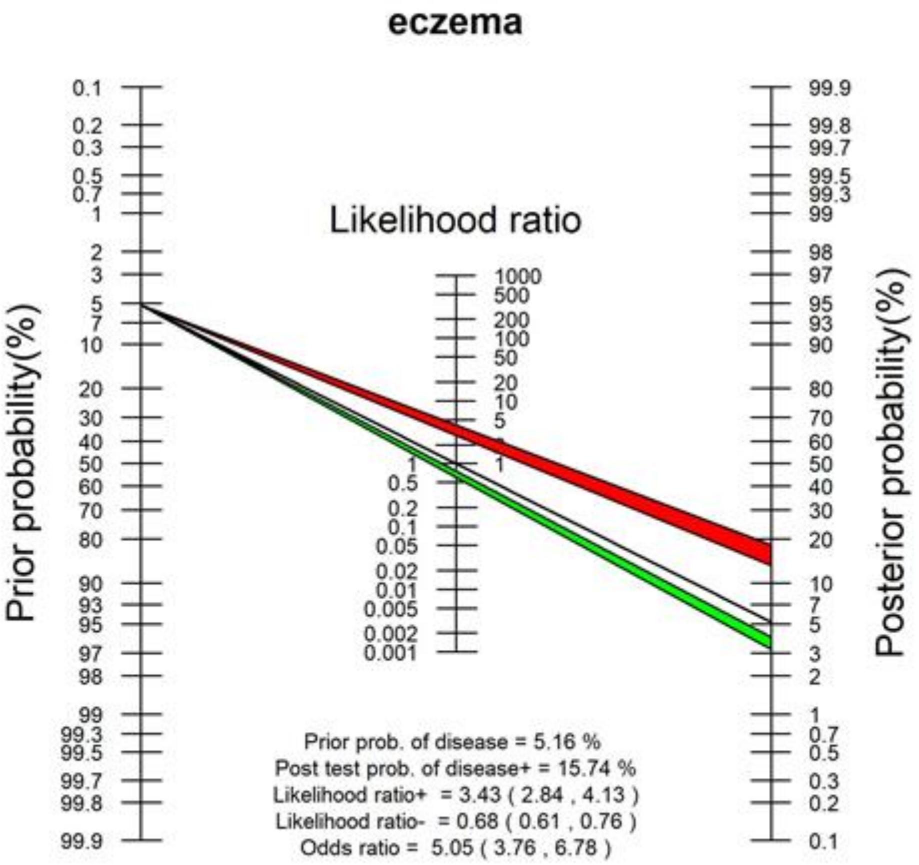


Figure 8.47 Results for eczema using clustering

True positives: 83; True negatives: 3357; False positives: 445; False negatives: 124

Sensitivity: 0.401; Specificity: 0.883

F1: 0.226; F2: 0.306; MCC: 0.186

Positive predictive value: 0.157; Negative predictive value: 0.964

Positive likelihood ratio: 3.43 with 95 % CI: 2.84 to 4.13

Negative likelihood ratio: 0.68 with 95 % CI: 0.61 to 0.76

Odds ratio: 5.05 with 95 % CI: 3.76 to 6.78

8.2.10 Gastroparesis

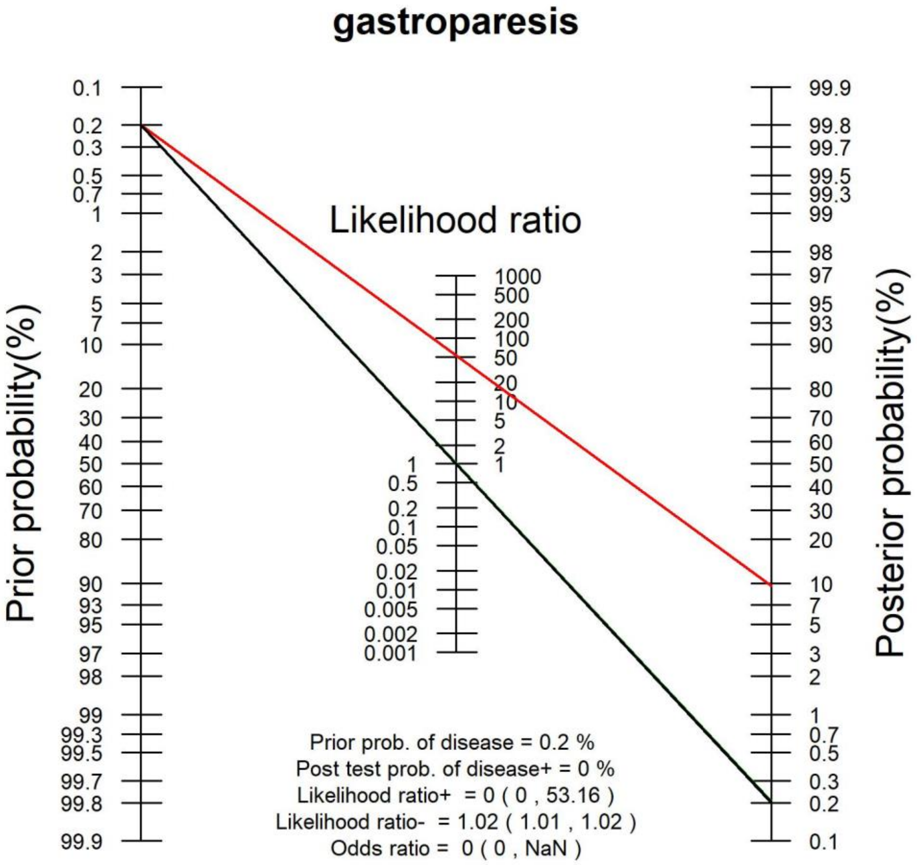


Figure 8.48 Results for gastroparesis using clustering

True positives: 0; True negatives: 3935; False positives: 70; False negatives: 8

Sensitivity: 0; Specificity: 0.983

F1: 0; F2: 0; MCC: 0.006

Positive predictive value: 0; Negative predictive value: 0.998

Positive likelihood ratio: - with 95 % CI: 0 to 53.16

Negative likelihood ratio: 1.02 with 95 % CI: 1.01 to 1.02

Odds ratio: 0 with 95 % CI: 0 to >100

8.2.11 Gout

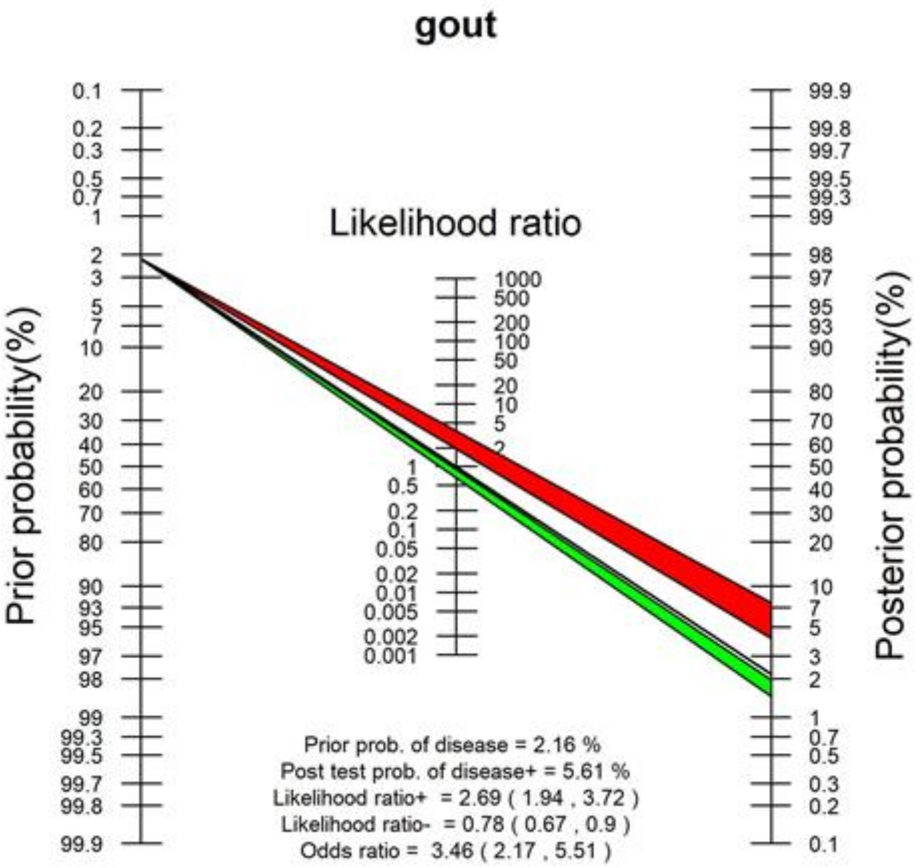


Figure 8.49 Results for gout using clustering

True positives: 27; True negatives: 3439; False positives: 445; False negatives: 59

Sensitivity: 0.314; Specificity: 0.883

F1: 0.095; F2: 0.164; MCC: 0.088

Positive predictive value: 0.056; Negative predictive value: 0.983

Positive likelihood ratio: 2.69 with 95 % CI: 1.94 to 3.72

Negative likelihood ratio: 0.78 with 95 % CI: 0.67 to 0.9

Odds ratio: 3.46 with 95 % CI: 2.17 to 5.51

8.2.12 Gastro-oesophageal reflux disease

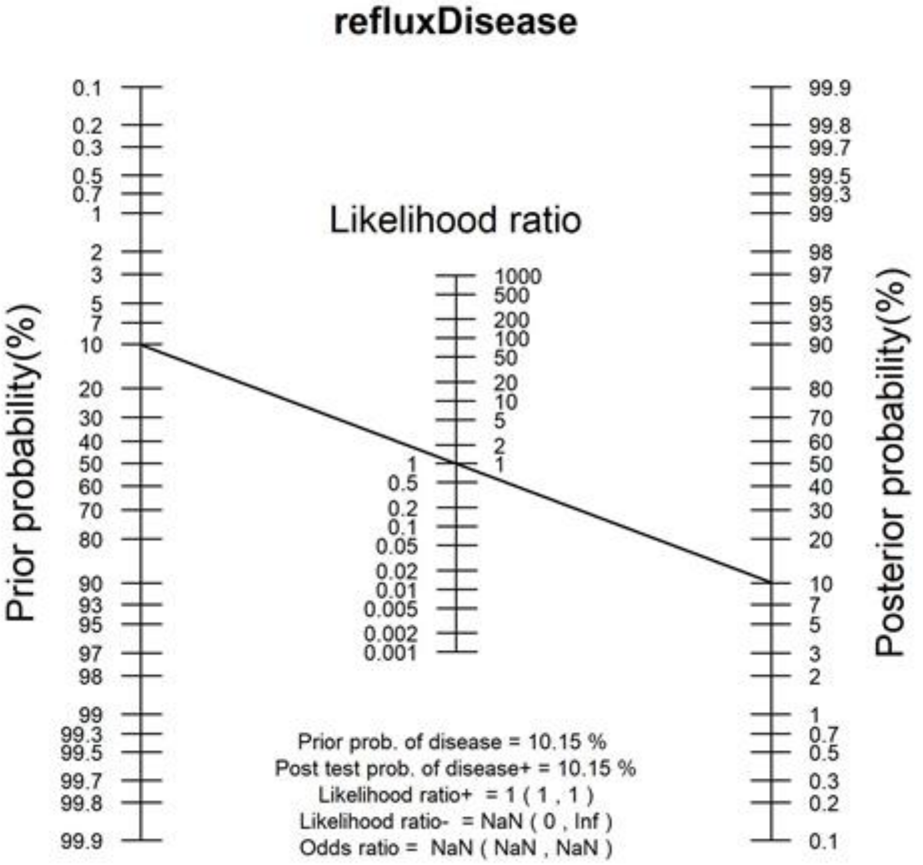


Figure 8.50 Results for gastro-intestinal reflux disease using clustering

True positives: 402; True negatives: 0; False positives: 3559; False negatives: 0

Sensitivity: 1; Specificity: 0

F1: 0.184; F2: 0.361; MCC: 0.-

Positive predictive value: 0.101; Negative predictive value: -

Positive likelihood ratio: 1 with 95 % CI: 1 to 1

Negative likelihood ratio: - with 95 % CI: 0 to >100

Odds ratio: - with 95 % CI: -

8.2.13 Obesity

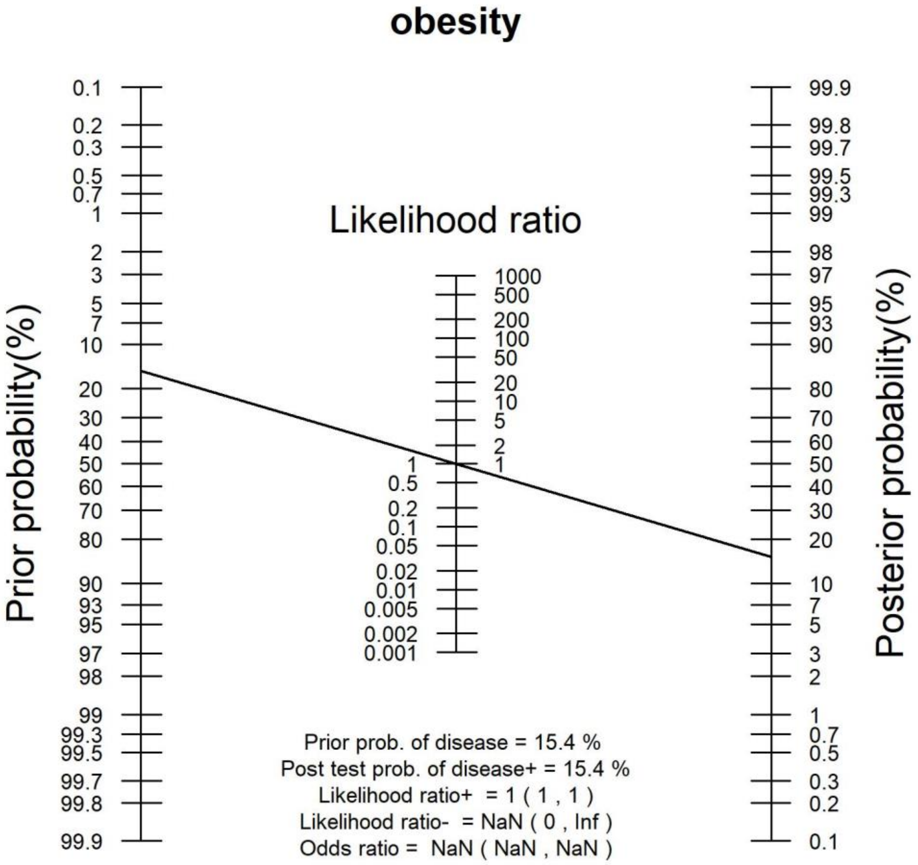


Figure 8.51 Results for obesity using clustering

True positives: 614; True negatives: 0; False positives: 3373; False negatives: 0

Sensitivity: 1; Specificity: 0

F1: 0.267; F2: 0.238; MCC: 0

Positive predictive value: 0.154; Negative predictive value: -

Positive likelihood ratio: 1 with 95 % CI: 1 to 1

Negative likelihood ratio: - with 95 % CI: 0 to >100

Odds ratio: - with 95 % CI: -

8.2.14 Osteoarthritis

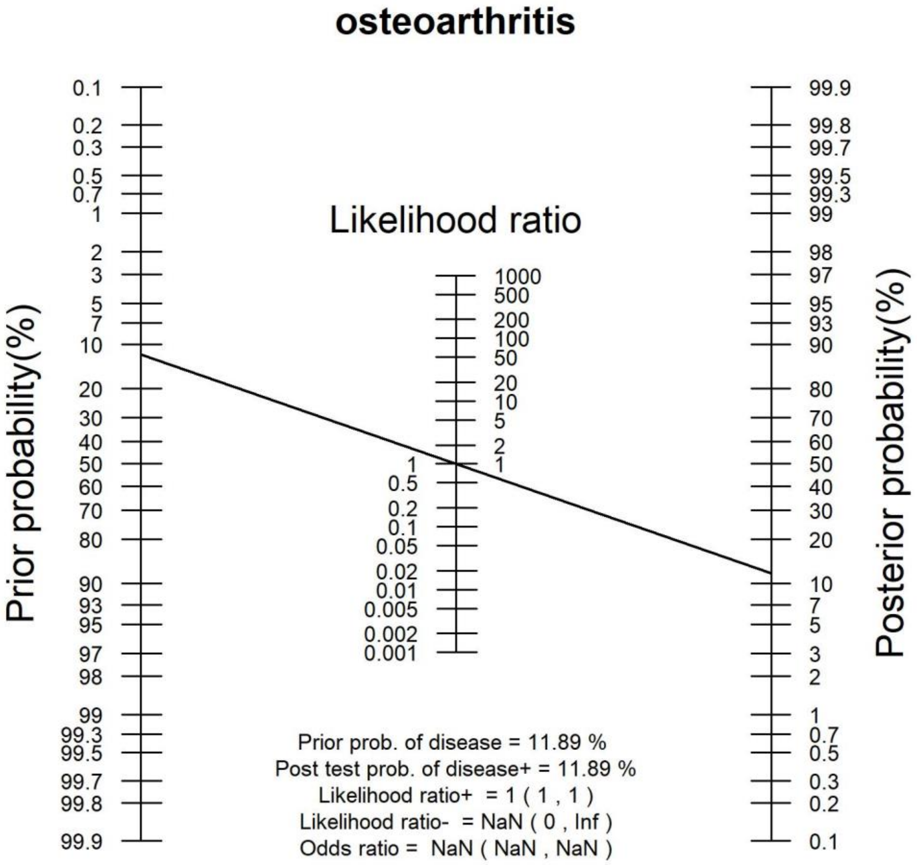


Figure 8.52 Results for osteoarthritis using clustering

True positives: 477; True negatives: 0; False positives: 3534; False negatives: 0

Sensitivity: 1; Specificity: 0

F1: 0.213; F2: 0.202; MCC: 0

Positive predictive value: 0.119; Negative predictive value: -

Positive likelihood ratio: - with 95 % CI: - to -

Negative likelihood ratio: - with 95 % CI: - to -

Odds ratio: - with 95 % CI: -

8.2.15 Prostate cancer

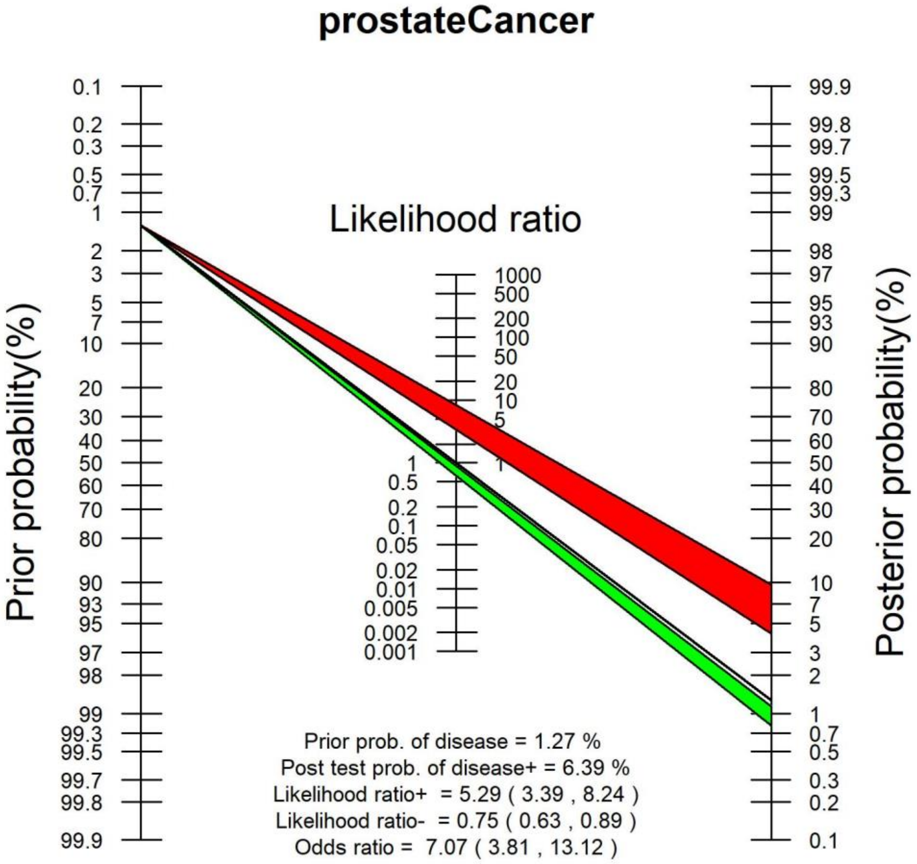


Figure 8.53 Results for prostate cancer using clustering

True positives: 15; True negatives: 3735; False positives: 220; False negatives: 36

Sensitivity: 0.294; Specificity: 0.944

F1: 0.105; F2: 0.086; MCC: 0.114

Positive predictive value: 0.064; Negative predictive value: 0.990

Positive likelihood ratio: 5.29 with 95 % CI: 3.39 to 8.24

Negative likelihood ratio: 0.75 with 95 % CI: 0.63 to 0.89

Odds ratio: 7.07 with 95 % CI: 3.81 to 13.12

8.2.16 Stress

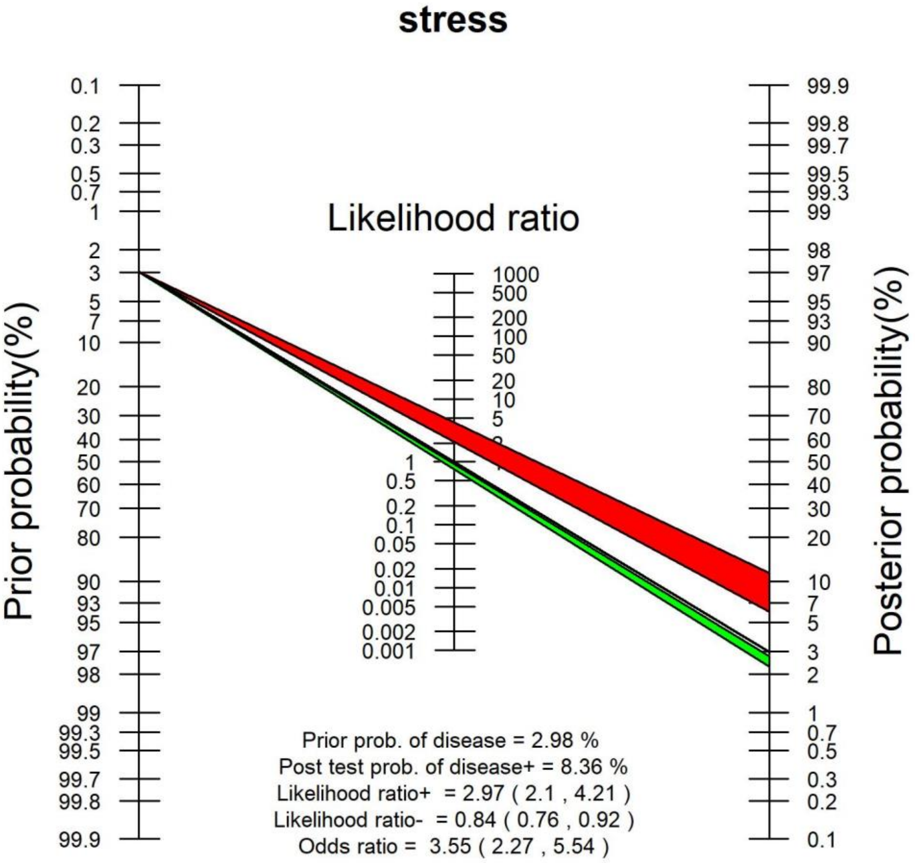


Figure 8.54 Results for stress using clustering

True positives: 27; True negatives: 3580; False positives: 296; False negatives: 92

Sensitivity: 0.227; Specificity: 0.924

F1: 0.122; F2: 0.169; MCC: 0.094

Positive predictive value: 0.084; Negative predictive value: 0.975

Positive likelihood ratio: 2.97 with 95 % CI: 2.1 to 4.21

Negative likelihood ratio: 0.84 with 95 % CI: 0.76 to 0.92

Odds ratio: 3.55 with 95 % CI: 2.27 to 5.54

8.2.17 Thyrotoxicosis

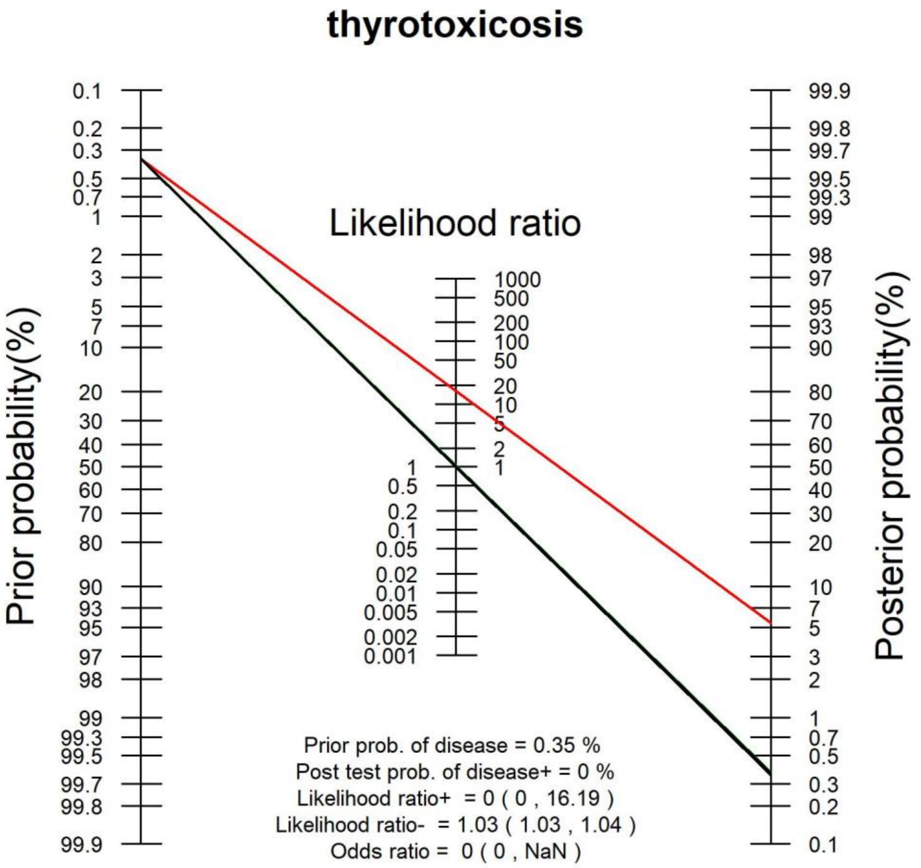


Figure 8.55 Results for thyrotoxicosis using clustering

True positives: 0; True negatives: 3833; False positives: 134; False negatives: 14

Sensitivity: 0; Specificity: 0.966

F1: 0; F2: 0; MCC: -0.011

Positive predictive value: 0; Negative predictive value: 0.996

Positive likelihood ratio: - with 95 % CI: -

Negative likelihood ratio: 1.03 with 95 % CI: 1.03 to 1.04

Odds ratio: 3.55 with 95 % CI: 2.27 to 5.54

8.2.18 Type 2 diabetes

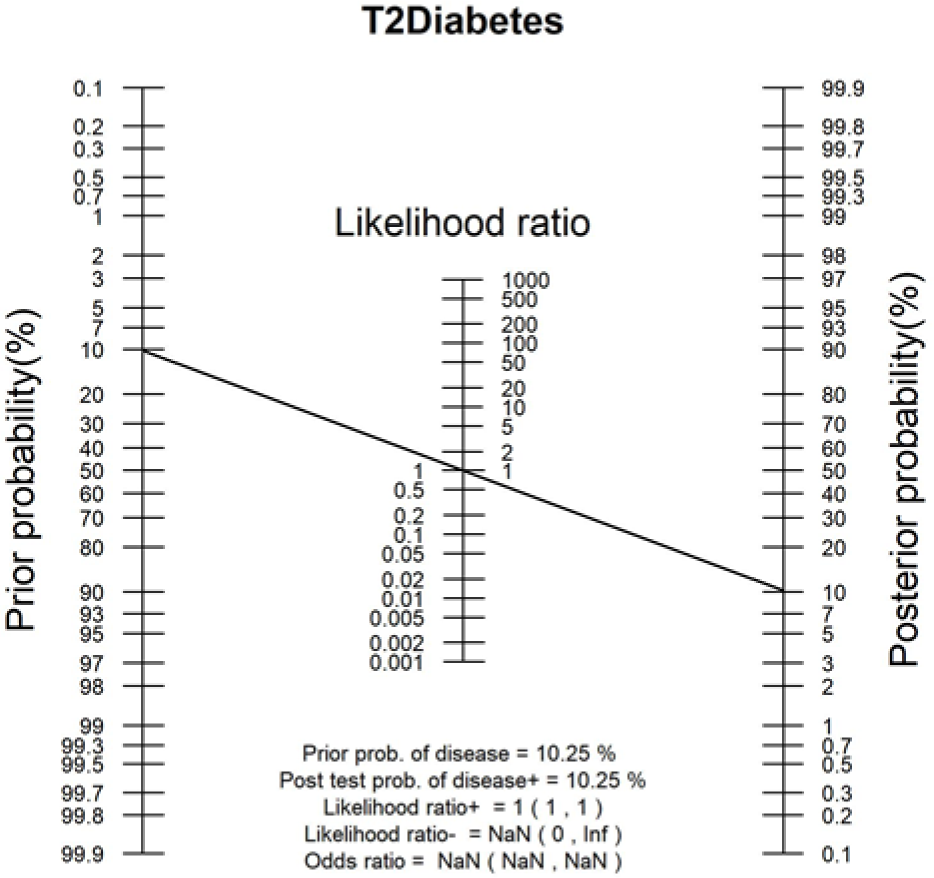


Figure 8.56 Results for type 2 diabetes using clustering

True positives: 409; True negatives: 0; False positives: 3581; False negatives: 0

Sensitivity: 1; Specificity: 0

F1: 0.186; F2: 0.364; MCC: -

Positive predictive value: 0.103; Negative predictive value: -

Positive likelihood ratio: 1 with 95 % CI: 1 to 1

Negative likelihood ratio: - with 95 % CI: -

Odds ratio: - with 95 % CI: -

8.2.19 Summary of results from clustering

A summary of the results from the test runs using the clustering method is shown in Table 8.2.

Condition	Positive Likelihood Ratio (95 % CI)	Negative Likelihood Ratio (95 % CI)	Condition Prevalence	Clusters
Acute sinusitis	1 (1 to 1)	-	14.2 %	2
Allergic rhinitis	1.03 (1.01 to 1.05)	0.69 (0.50 to 0.94)	17.7 %	3
Any cancer	4.90 (3.70 to 6.50)	0.78 (0.71 to 0.85)	4.4 %	130
Asthma	2.58 (1.97 to 3.36)	0.91 (0.88 to 0.95)	11.5 %	2
Autism	6.32 (2.56 to 15.59)	0.66 (0.39 to 1.14)	0.2 %	90
Breast cancer	7.24 (4.52 to 11.59)	0.75 (0.63 to 0.89)	1.3 %	920
Bronchitis	0.95 (0.89 to 1.01)	1.11 (0.99 to 1.25)	15.4 %	3
Colon cancer	4.87 (1.38 to 17.23)	0.85 (0.64 to 1.12)	0.3 %	359
Eczema	3.43 (2.84 to 4.13)	0.68 (0.61 to 0.76)	5.2 %	104
Gastro-intestinal reflux disease	-	-	10.15 %	3
Gastroparesis	-	1.02 (1.01 to 1.02)	0.2 %	530
Gout	2.69 (1.94 to 3.72)	0.78 (0.67 to 0.9)	2.2 %	37
Obesity	-	-	15.4 %	3
Osteoarthritis	-	-	11.9 %	2
Prostate cancer	5.29 (3.39 to 8.24)	0.75 (0.63 to 0.89)	1.3 %	280
Stress	2.97 (2.10 to 4.21)	0.84 (0.76 to 0.92)	3.0 %	150
Thyrotoxicosis	-	1.03 (1.03 to 1.04)	0.4 %	359
Type 2 diabetes	1 (1 to 1)	-	10.3 %	3

Table 8.2 Summary of results from test runs using clustering method

It can be seen from the summary of the results that the clustering method for predicting presence or absence of a condition in the set of records generally discriminates well between positive and negative-predicted groups when the optimum number of clusters derived from the training runs is high (> 3), even for conditions with low prevalence. For conditions with a low optimum number of clusters (2 or 3 clusters), the discrimination is generally poor. For these poor-discrimination conditions (acute sinusitis, asthma, gastro-intestinal reflux disease, obesity, osteoarthritis. Type 2 diabetes) ,the method predicted all records to be positive for the condition of interest, appearing to have placed all records in one of the clusters which it then scored as a 'positive' cluster. Results are more fully discussed in Chapter 9.

8.3 Best factors for nearest neighbours method

After testing using the clustering method, and inspection of the results, a second method was implemented, in order to see if results could be improved. Results are presented for determination of the optimum level of the CTV3 hierarchy and the optimum value for k, the number of nearest neighbours, using a nearest neighbours method. Again, the F2 score was used as the metric on which to optimise, which weights sensitivity greater than specificity.

8.3.1 Acute sinusitis

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for acute sinusitis are shown in Figure 8.57 and Figure 8.58.

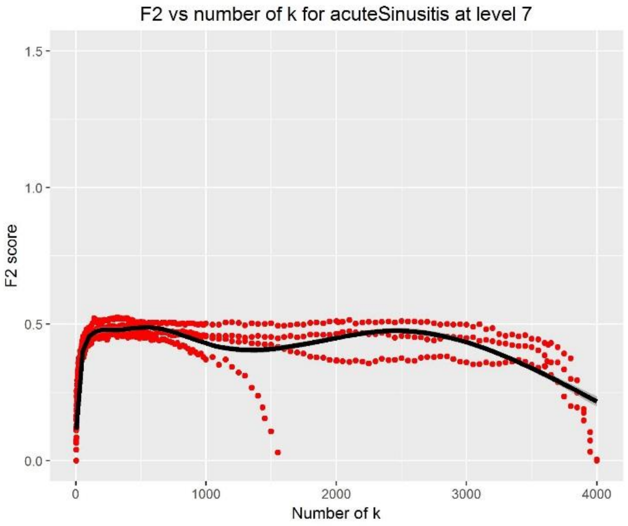


Figure 8.57 F2 score versus size of nearest neighbours group for acute sinusitis at CTV3 hierarchy level 7

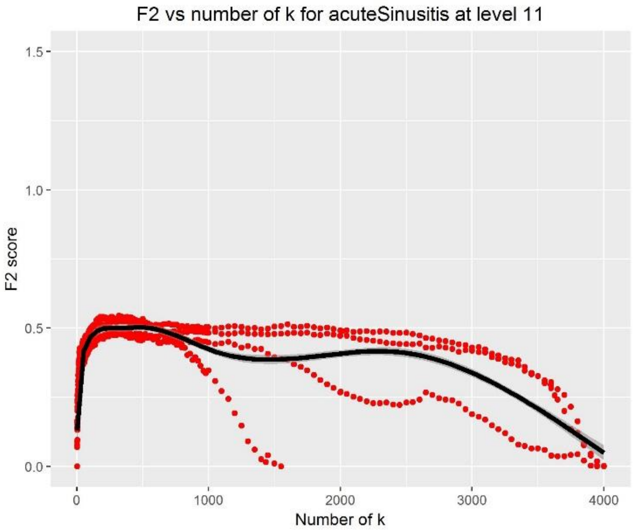


Figure 8.58 F2 score versus size of nearest neighbours group for acute sinusitis at CTV3 hierarchy level 11

At CTV3 hierarchy level 7, maximum F2 was 0.5143 at a value for k of 403.

At CTV3 hierarchy level 11, maximum F2 was 0.5364 at a value for k of 399.

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 399.

8.3.2 Allergic rhinitis

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for allergic rhinitis are shown in Figure 8.59 and Figure 8.60 .

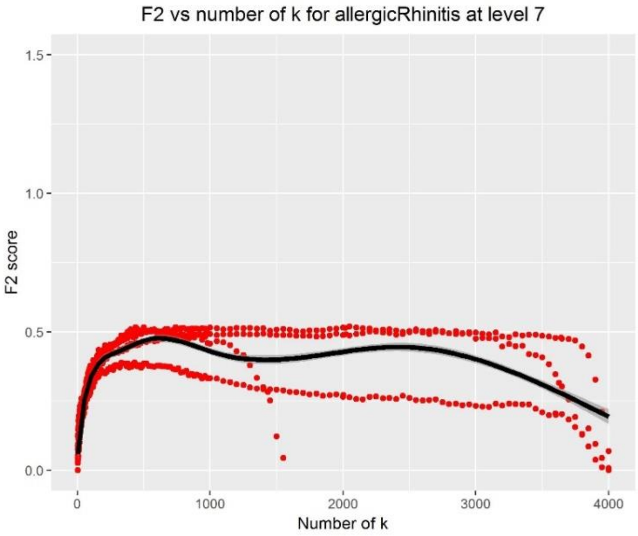


Figure 8.59 F2 score versus size of nearest neighbours group for allergic rhinitis at CTV3 hierarchy level 7

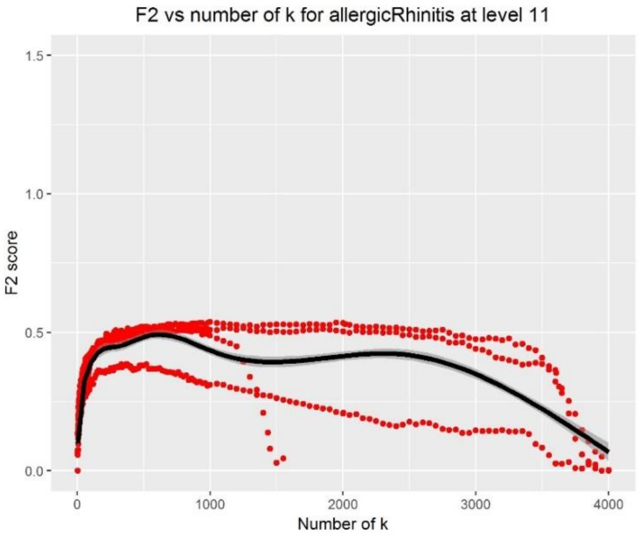


Figure 8.60 F2 score versus size of nearest neighbours group for allergic rhinitis at CTV3 hierarchy level 7

At CTV3 hierarchy level 7, maximum F2 was 0.5139 at a value of k of 2016.

At CTV3 hierarchy level 11, maximum F2 was 0.5203 at a value of k of 2016.

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 399 .

8.3.3 Any cancer

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for any cancer are shown in Figure 8.61 and Figure 8.62.

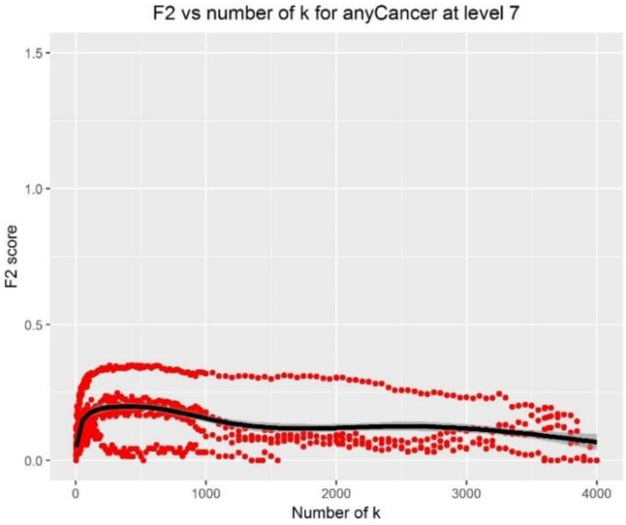


Figure 8.61 F2 score versus size of nearest neighbours group for any form of cancer

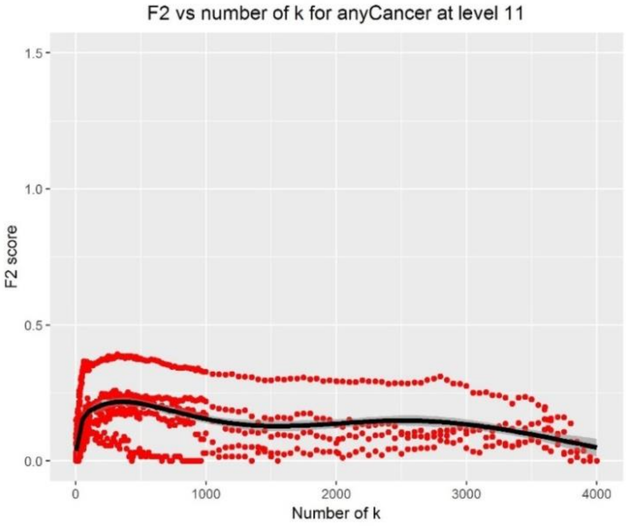


Figure 8.62 F2 score versus size of nearest neighbours group for any form of cancer

At CTV3 hierarchy level 7, maximum F2 was 0.3313 at a value for k of 286.

At CTV3 hierarchy level 11, maximum F2 was 0.3821 at a value for k of 286.

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 399 .

8.3.4 Asthma

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for autism are shown in Figure 8.63 and Figure 8.64.

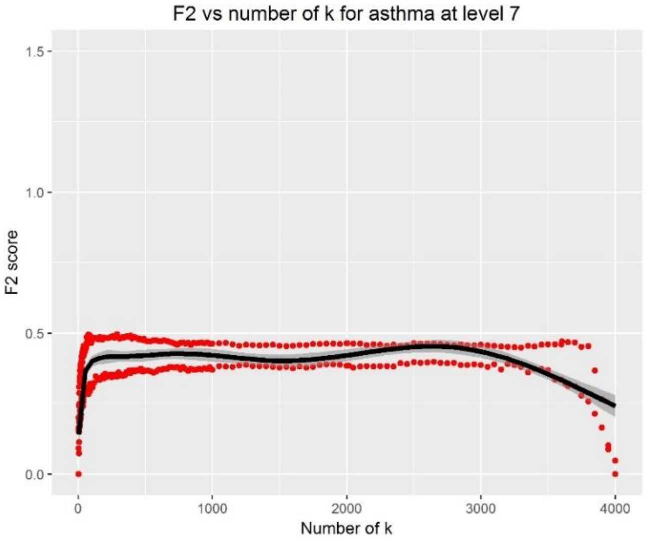


Figure 8.63 F2 score versus size of nearest neighbours group for asthma

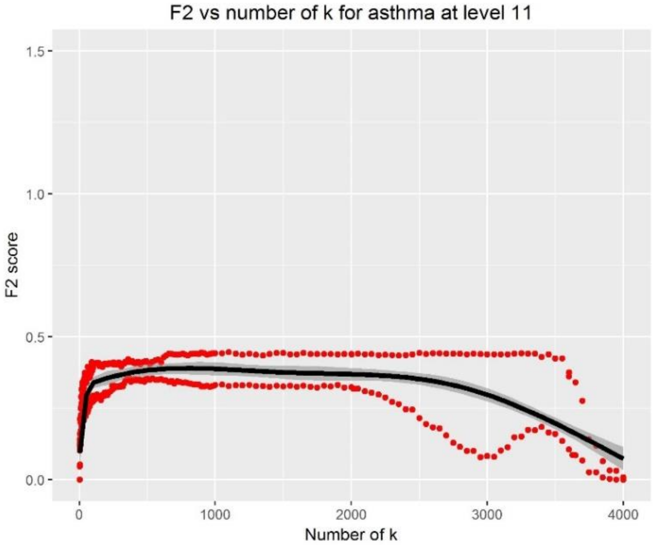


Figure 8.64 F2 score versus size of nearest neighbours group for asthma

At CTV3 hierarchy level 7, maximum F2 was 0.4806 at a value for k of 172.

At CTV3 hierarchy level 11, maximum F2 was 0.4378 at a value for k of 2002.

Selected factors were therefore CTV3 hierarchy level 7 with a value of k of 172 .

8.3.5 Autism

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for autism are shown in Figure 8.65 and Figure 8.66.

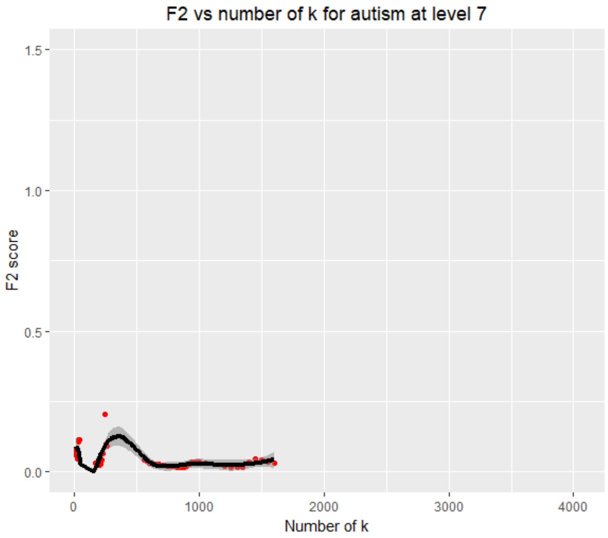


Figure 8.65 F2 score versus size of nearest neighbours group for autism

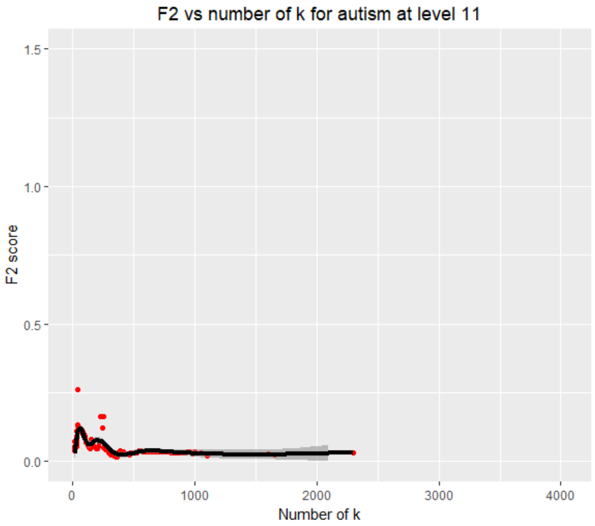


Figure 8.66 F2 score versus size of nearest neighbours group for autism

At CTV3 hierarchy level 7, maximum F2 was 0.2041 at a value for k of 240.

At CTV3 hierarchy level 11, maximum F2 was 0.2632 at a value for k of 37.

Selected factors were therefore CTV3 hierarchy level 7 with a value of k of 37 .

8.3.6 Breast cancer

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for breast cancer are shown in Figure 8.67 and Figure 8.68.

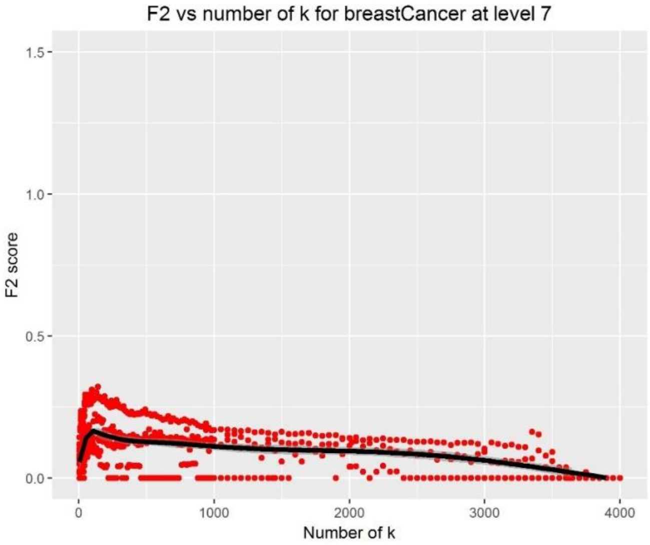


Figure 8.67 F2 score versus size of nearest neighbours group for breast cancer

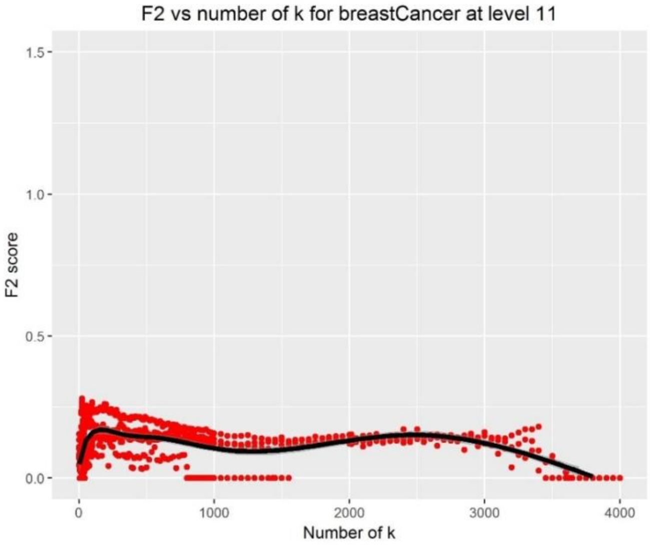


Figure 8.68 F2 score versus size of nearest neighbours group for breast cancer

At CTV3 hierarchy level 7, maximum F2 was 0.2255 at a value for k of 525.

At CTV3 hierarchy level 11, maximum F2 was 0.2030 at a value for k of 525.

Selected factors were therefore CTV3 hierarchy level 7 with a value of k of 525.

8.3.7 Bronchitis

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for bronchitis are shown in Figure 8.69 and Figure 8.70.

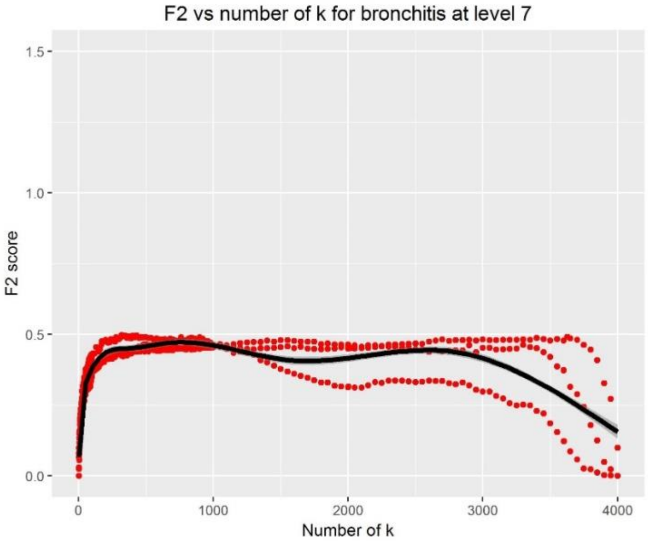


Figure 8.69 F2 score versus size of nearest neighbours group for bronchitis

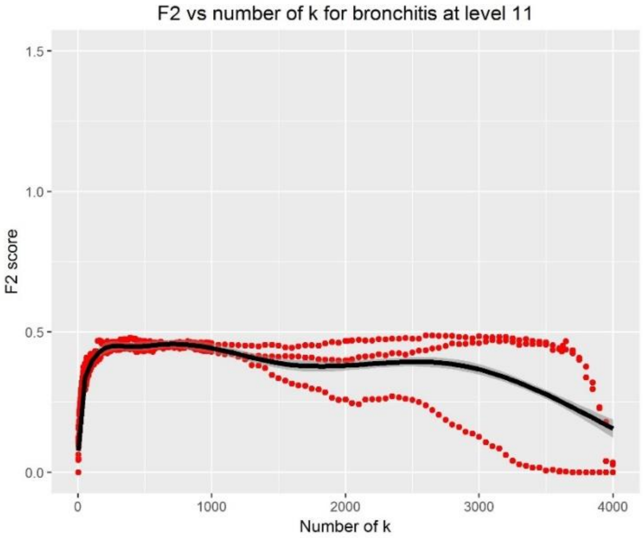


Figure 8.70 F2 score versus size of nearest neighbours group for bronchitis

At CTV3 hierarchy level 7, maximum F2 was 0.4904 at a value for k of 3628.

At CTV3 hierarchy level 11, maximum F2 was 0.4684 at a value for k of 2016.

Selected factors were therefore CTV3 hierarchy level 7 with a value of k of 3628 .

8.3.8 Colon cancer

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for colon cancer are shown in Figure 8.71 and Figure 8.72.

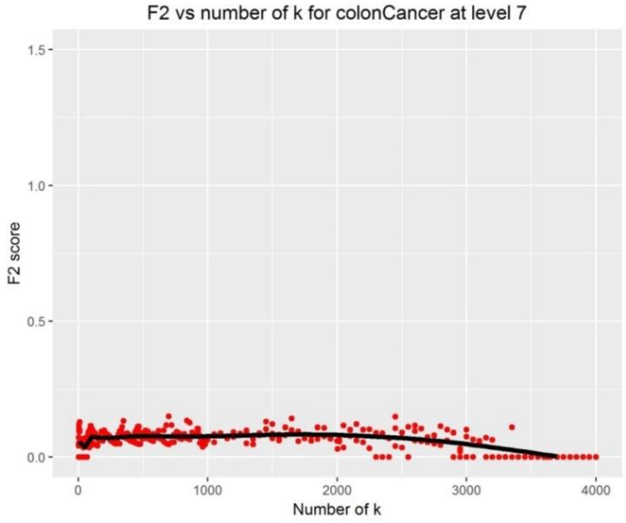


Figure 8.71 F2 score versus size of nearest neighbours group for colon cancer

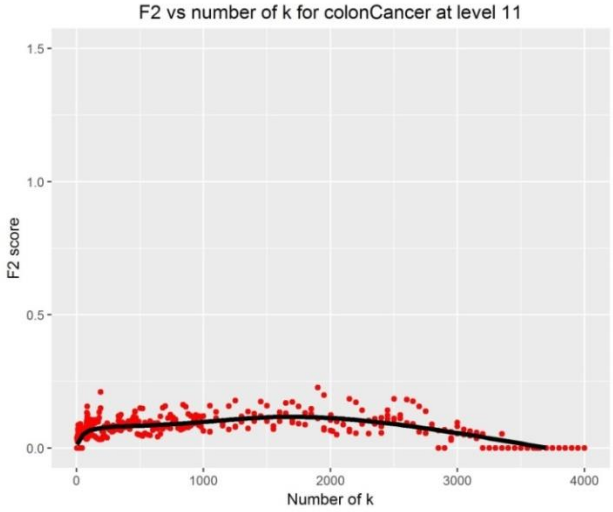


Figure 8.72 F2 score versus size of nearest neighbours group for colon cancer

At CTV3 hierarchy level 7, maximum F2 was 0.1198 at a value for k of 6.

At CTV3 hierarchy level 11, maximum F2 was 0.1497 at a value for k of 1900.

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 1900 .

8.3.9 Eczema

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for eczema are shown in Figure 8.73 and Figure 8.74.

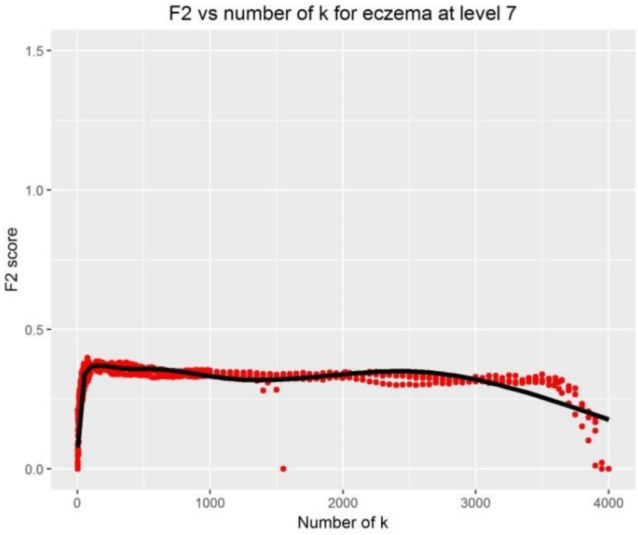


Figure 8.73 F2 score versus size of nearest neighbours group for eczema

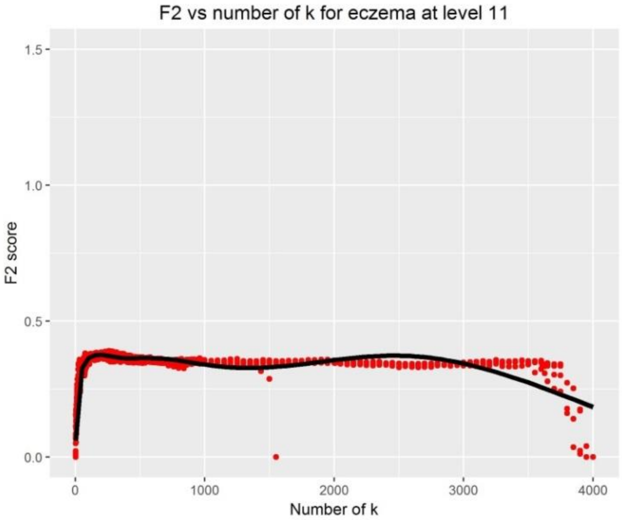


Figure 8.74 F2 score versus size of nearest neighbours group for eczema

At CTV3 hierarchy level 7, maximum F2 was 0.3736 at a value for k of 75.

At CTV3 hierarchy level 11, maximum F2 was 0.3827 at a value for k of 273.

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 273 .

8.3.10 Gastro-oesophageal reflux disease

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for gastro-intestinal reflux disease are shown in Figure 8.75 and Figure 8.76.

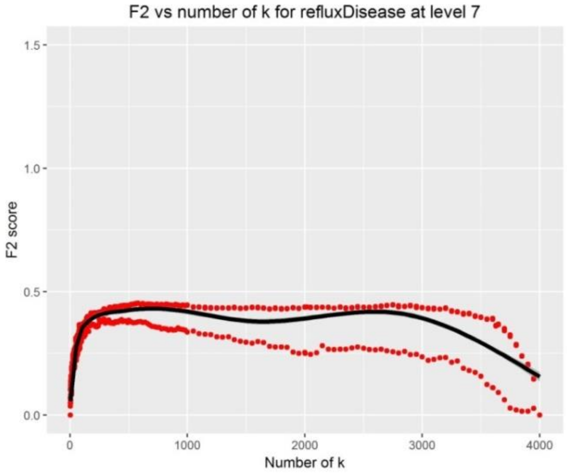


Figure 8.75 F2 score versus size of nearest neighbours group for gastro-intestinal reflux disease

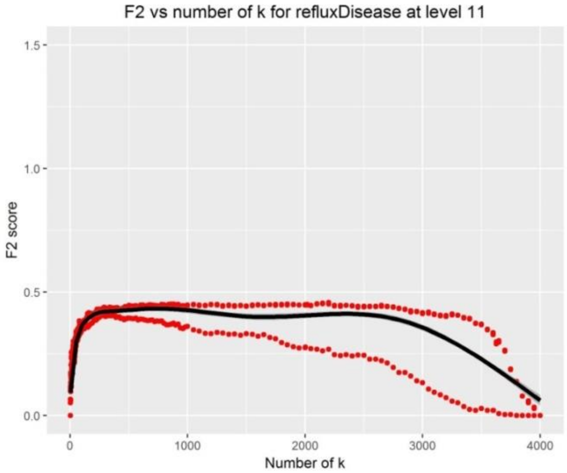


Figure 8.76 F2 score versus size of nearest neighbours group for gastro-intestinal reflux disease

At CTV3 hierarchy level 7, maximum F2 was 0.4388 at a value for k of 403.

At CTV3 hierarchy level 11, maximum F2 was 0.4485 at a value for k of 2016.

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 2016 .

8.3.11 Gastroparesis

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for gastroparesis are shown in Figure 8.77 and Figure 8.78.

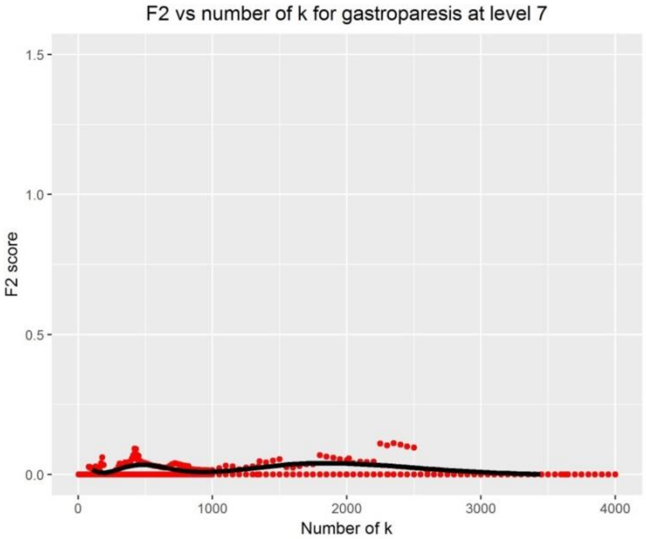


Figure 8.77 F2 score versus size of nearest neighbours group for gastroparesis

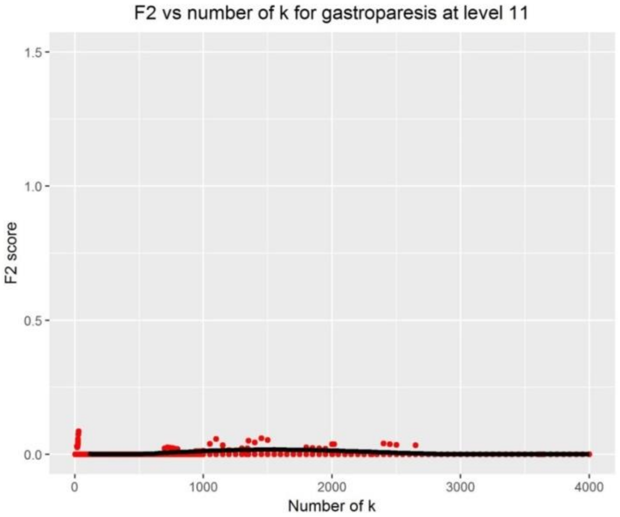


Figure 8.78 F2 score versus size of nearest neighbours group for gastroparesis

At CTV3 hierarchy level 7, maximum F2 was 0.1108 at a value for k of 2350.

At CTV3 hierarchy level 11, maximum F2 was 0.0847 at a value for k of 30.

Selected factors were therefore CTV3 hierarchy level 7 with a value of k of 2350 .

8.3.12 Gout

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for gout are shown in Figure 8.79 and Figure 8.80.

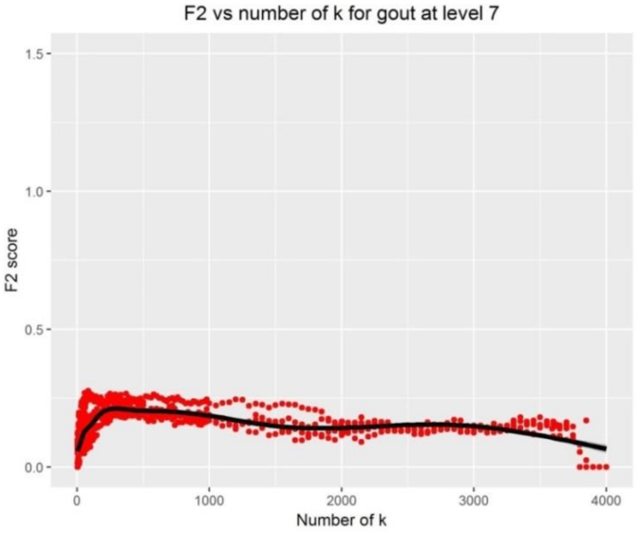


Figure 8.79 F2 score versus size of nearest neighbours group for gout

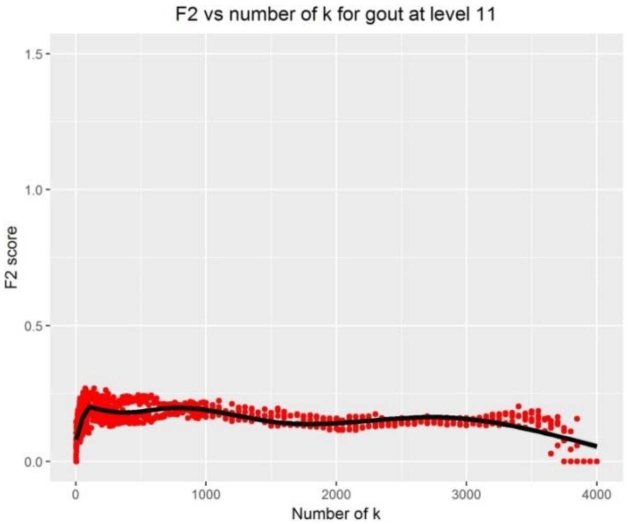


Figure 8.80 F2 score versus size of nearest neighbours group for gout

At CTV3 hierarchy level 7, maximum F2 was 0.2390 at a value for k of 406.

At CTV3 hierarchy level 11, maximum F2 was 0.2348 at a value for k of 399.

Selected factors were therefore CTV3 hierarchy level 7 with a value of k of 406.

8.3.13 Obesity

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for obesity are shown in Figure 8.81 and Figure 8.82.

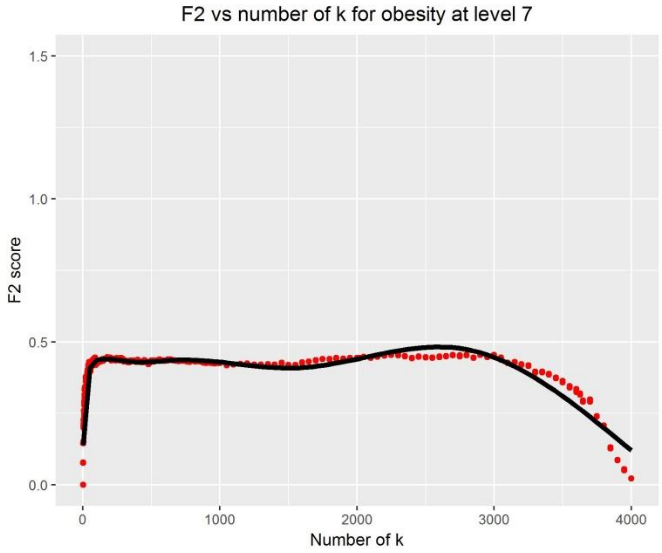


Figure 8.81 F2 score versus size of nearest neighbours group for obesity

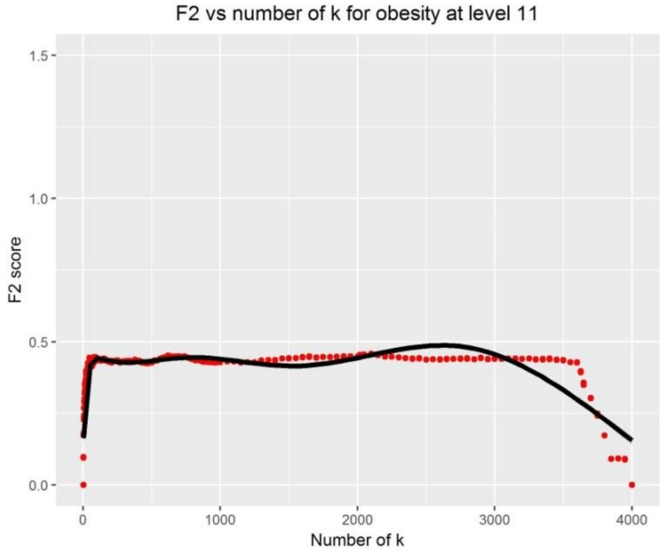


Figure 8.82 F2 score versus size of nearest neighbours group for obesity

At CTV3 hierarchy level 7, maximum F2 was 0.4547 at a value for k of 2900.

At CTV3 hierarchy level 11, maximum F2 was 0.4582 at a value for k of 2100.

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 2100 .

8.3.14 Osteoarthritis

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for osteoarthritis are shown in Figure 8.83 and Figure 8.84.

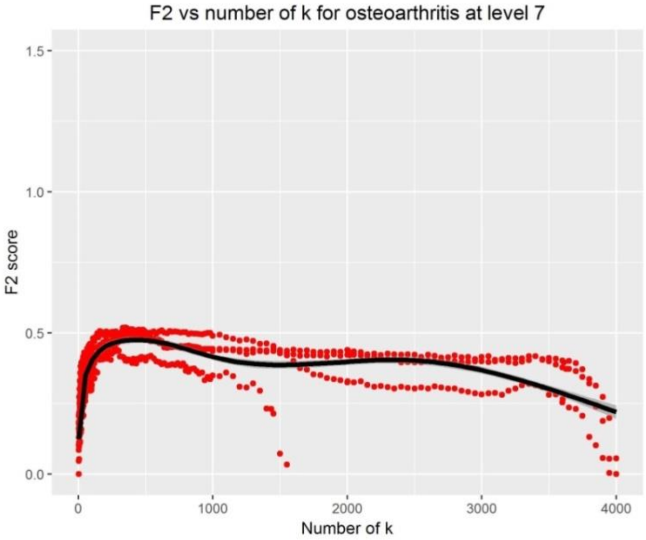


Figure 8.83 F2 score versus size of nearest neighbours group for osteoarthritis

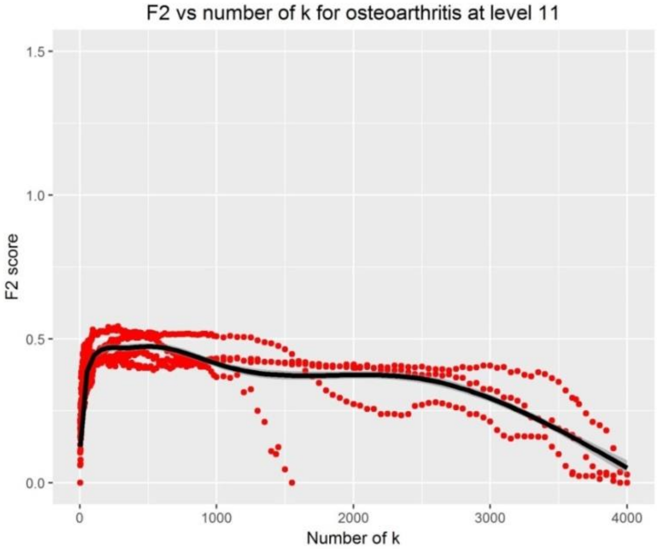


Figure 8.84 F2 score versus size of nearest neighbours group for osteoarthritis

At CTV3 hierarchy level 7, maximum F2 was 0.5058 at a value for k of 209.

At CTV3 hierarchy level 11, maximum F2 was 0.5410 at a value for k of 209.

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 209.

8.3.15 Prostate cancer

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for prostate cancer are shown in Figure 8.85 and Figure 8.86.

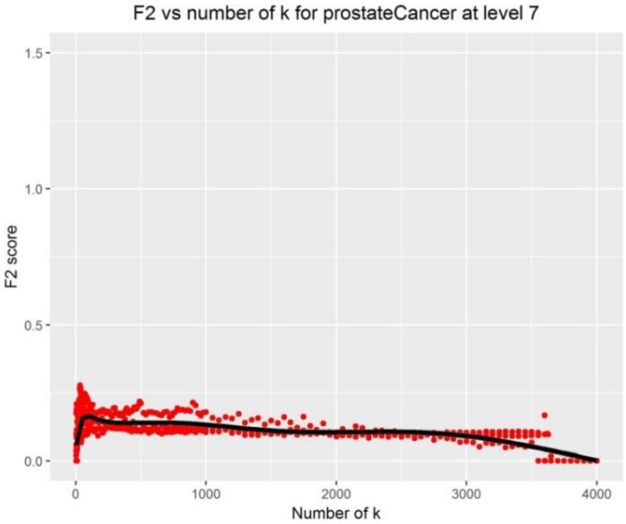


Figure 8.85 F2 score versus size of nearest neighbours group for prostate cancer

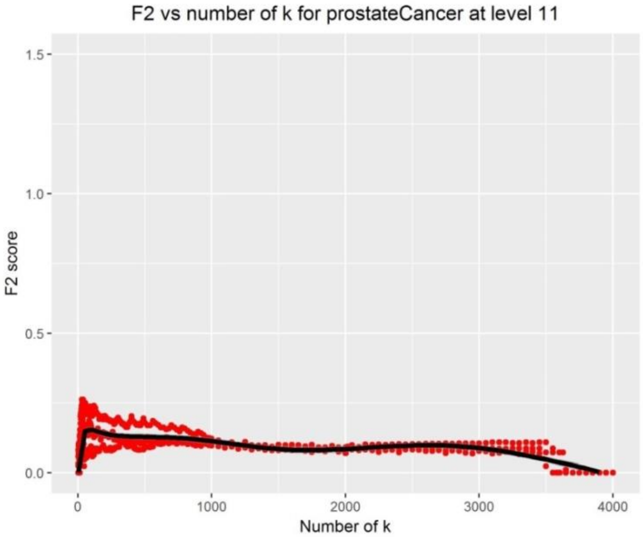


Figure 8.86 F2 score versus size of nearest neighbours group for prostate cancer

At CTV3 hierarchy level 7, maximum F2 was 0.1771 at a value for k of 138.

At CTV3 hierarchy level 11, maximum F2 was 0.2167 at a value for k of 138.

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 138 .

8.3.16 Stress

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for stress are shown in Figure 8.87 and Figure 8.88.

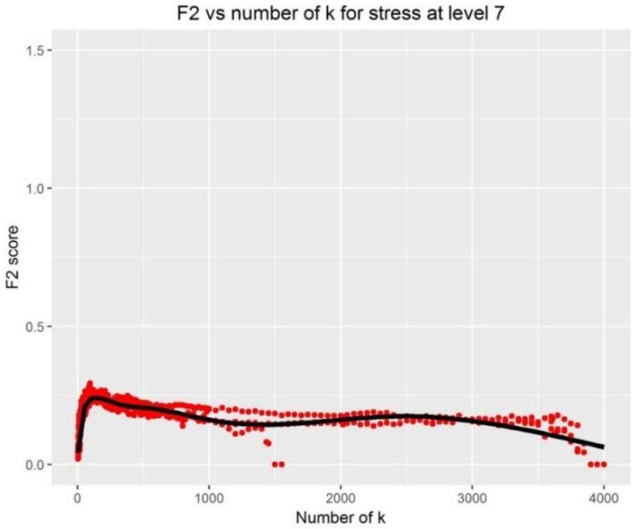


Figure 8.87 F2 score versus size of nearest neighbours group for stress

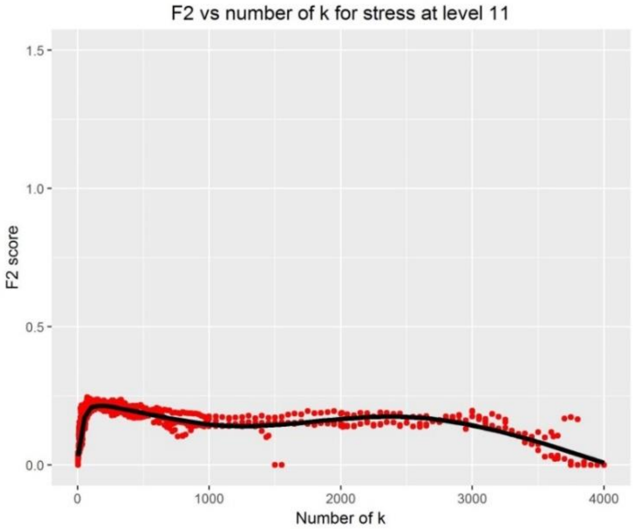


Figure 8.88 F2 score versus size of nearest neighbours group for stress

At CTV3 hierarchy level 7, maximum F2 was 0.2446 at a value for k of 180.

At CTV3 hierarchy level 11, maximum F2 was 0.2536 at a value for k of 180

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 180 .

8.3.17 Thyrotoxicosis

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for stress are shown in Figure 8.89 and Figure 8.90.

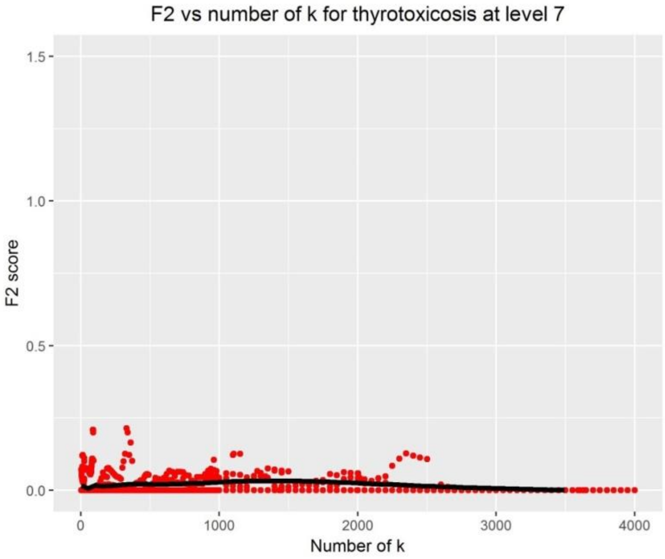


Figure 8.89 F2 score versus size of nearest neighbours group for thyrotoxicosis

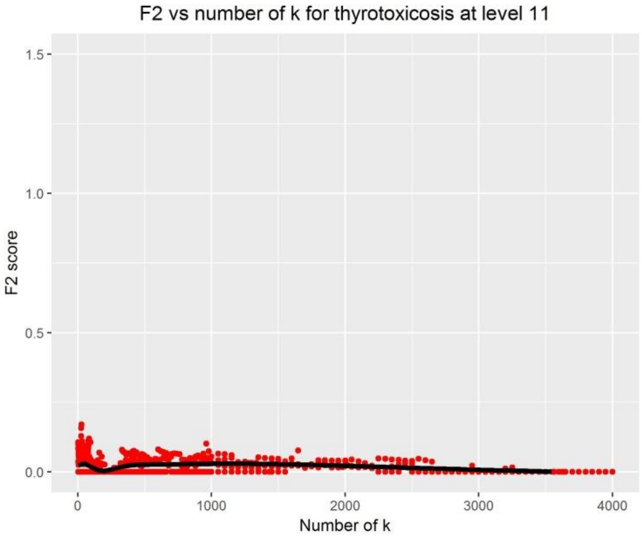


Figure 8.90 F2 score versus size of nearest neighbours group for thyrotoxicosis

At CTV3 hierarchy level 7, maximum F2 was 0.1263 at a value for k of 2350.

At CTV3 hierarchy level 11, maximum F2 was 0.1695 at a value for k of 28.

Selected factors were therefore CTV3 hierarchy level 11 with a value of k of 28 .

8.3.18 Type 2 diabetes.

Results of the analysis of F2 score versus size of the nearest neighbours group at CTV3 hierarchy levels of 7 and 11 for stress are shown in Figure 8.91 and Figure 8.92.

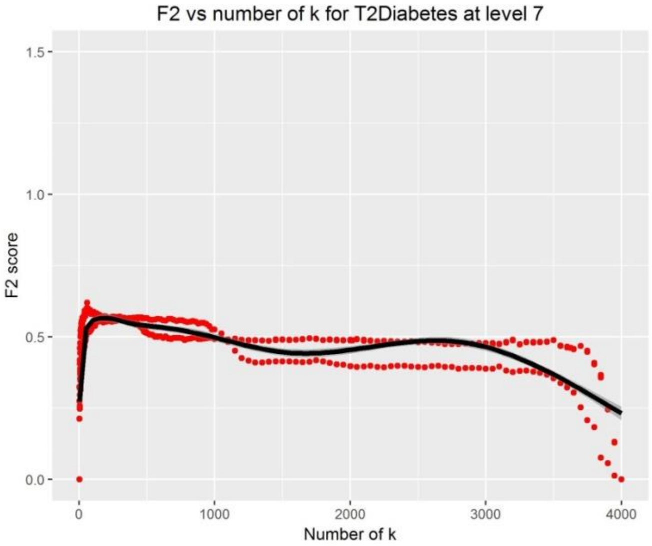


Figure 8.91 F2 score versus size of nearest neighbours group for type 2 diabetes

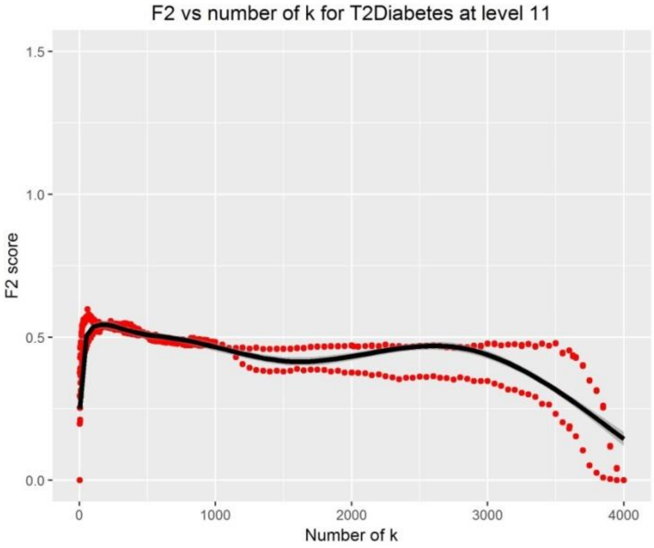


Figure 8.92 F2 score versus size of nearest neighbours group for thyrotoxicosis

At CTV3 hierarchy level 7, maximum F2 was 0.5685 at a value for k of 300.

At CTV3 hierarchy level 11, maximum F2 was 0.5473 at a value for k of 230.

Selected factors were therefore CTV3 hierarchy level 7 with a value of k of 300 .

8.3.19 Summary of training runs for best k

The final set of best factors for each condition based on analysis of the training set has now been produced and is summarised in Table 6.25.

Condition	Best CTV3 level	Best k at CTV3 level	F2 at best level and k	Prevalence in training set
Acute sinusitis	11	399	0.536	0.154
Allergic rhinitis	11	2016	0.530	0.198
Any cancer	11	286	0.382	0.042
Asthma	7	172	0.481	0.122
Autism	11	37	0.263	0.001
Breast cancer	7	525	0.226	0.011
Bronchitis	7	3628	0.490	0.159
Colon cancer	11	1900	0.150	0.005
Eczema	11	273	0.383	0.054
Gastroparesis	7	2350	0.111	0.002
Gout	7	406	0.239	0.024
Obesity	11	2100	0.458	0.160
Osteoarthritis	11	209	0.541	0.117
Prader-Willi				
Prostate cancer	11	138	0.217	0.111
Reflux disease	11	2016	0.448	0.109
Stress	11	180	0.254	0.026
Thyrotoxicosis	11	28	0.169	0.006
Type 2 diabetes	7	300	0.569	0.102

Table 8.3 Summary of best factors for k nearest neighbours for optimum F2 score

It should be noted that these factors have been determined on a subset of the data set. In particular the value of k for the number of nearest neighbours and for the number of clusters was not scaled up for use on the full data set. There are a number of ‘rules of thumb’ for choosing values for k, for example taking the square root of the number of samples, as suggested by Duda et al [275], however ultimately the optimum value for k is likely to be chosen empirically, as is the case with the work described here. Nevertheless, not scaling up the value of k for larger data sets is doubtless a limitation.

These factors will be used as the input to the next stage of the process, seeing how well different techniques for predicting presence or absence of the condition from a set of near neighbours.

8.4 Results from test runs for nearest neighbours method

As described in Section 8.2 for testing of the clustering method, the positive and negative likelihood ratios were calculated for test runs on each condition using the optimum factors derived in Section 8.2 and shown in Table 8.4. Results of these tests are presented as Fagan nomograms.

Five test runs were carried out, with each run drawing a random sample of 5000 records from the training set of records. The mean of the True Positive, True Negative, False Positive and False Negative counts from each run were used to calculate the scores given and to produce the nomograms.

8.4.1 Acute sinusitis

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.93.

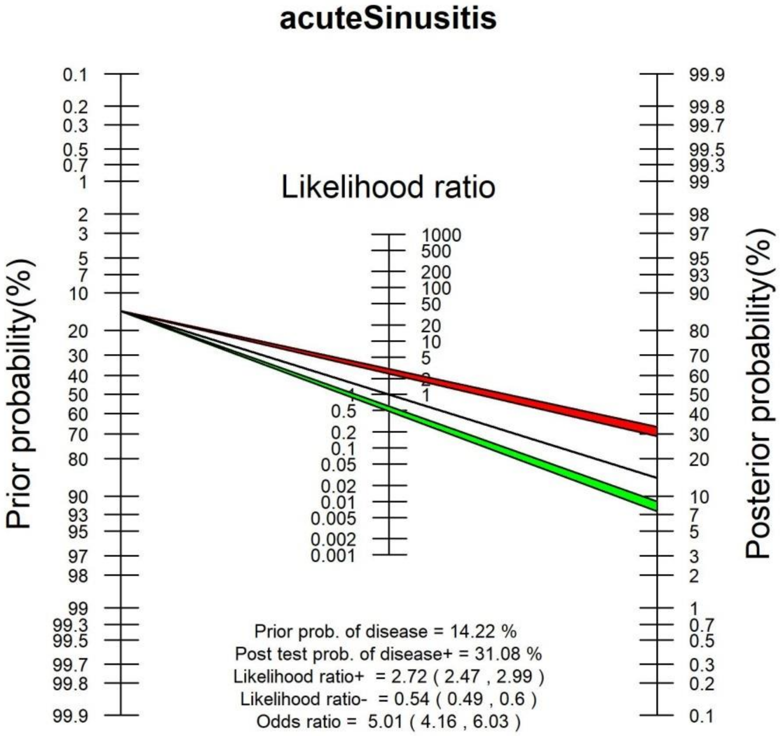


Figure 8.93 Results for acute sinusitis using nearest neighbours

True Positives: 325; True Negatives: 2711; False Positives: 721; False Negatives: 244

Sensitivity: 0.571; Specificity: 0.790

F1: 0.402; F2: 0.489; MCC: 0.287

Positive Predictive Value: 0.311; Negative Predictive Value: 0.917

Positive Likelihood Ratio = 2.719 (2.469, 2.994)

Negative Likelihood Ratio = 0.543 (0.493, 0.598)

Odds ratio = 5.008; 95% CI = (4.161, 6.027)

8.4.2 Allergic rhinitis

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.94.

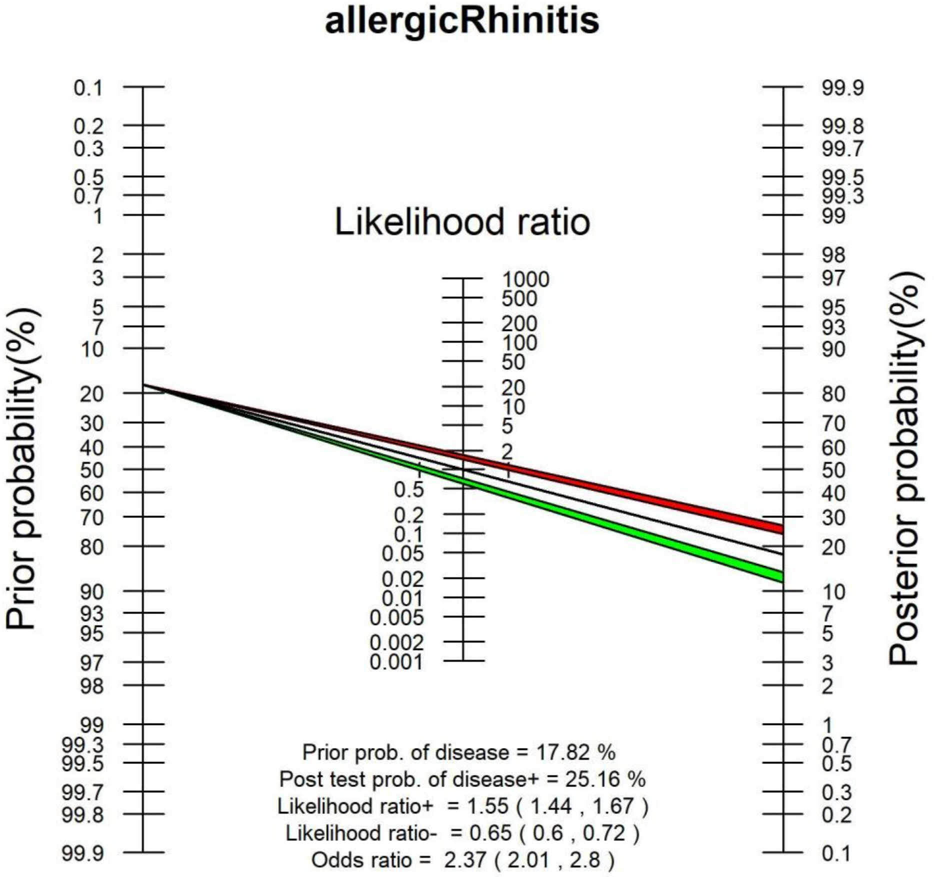


Figure 8.94 Results for allergic rhinitis using nearest neighbours

True Positives: 427; True Negatives: 2018; False Positives: 1270; False Negatives: 286

Sensitivity: 0.599; Specificity: 0.614

F1: 0.354; F2: 0.469; MCC: 0.165

Positive Predictive Value: 0.252; Negative Predictive Value: 0.876

Positive Likelihood Ratio = 1.55 (1.44, 1.669)

Negative Likelihood Ratio = 0.654 (0.595, 0.718)

Odds ratio = 2.372; 95% CI = (2.011, 2.799)

8.4.3 Any cancer

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.95.

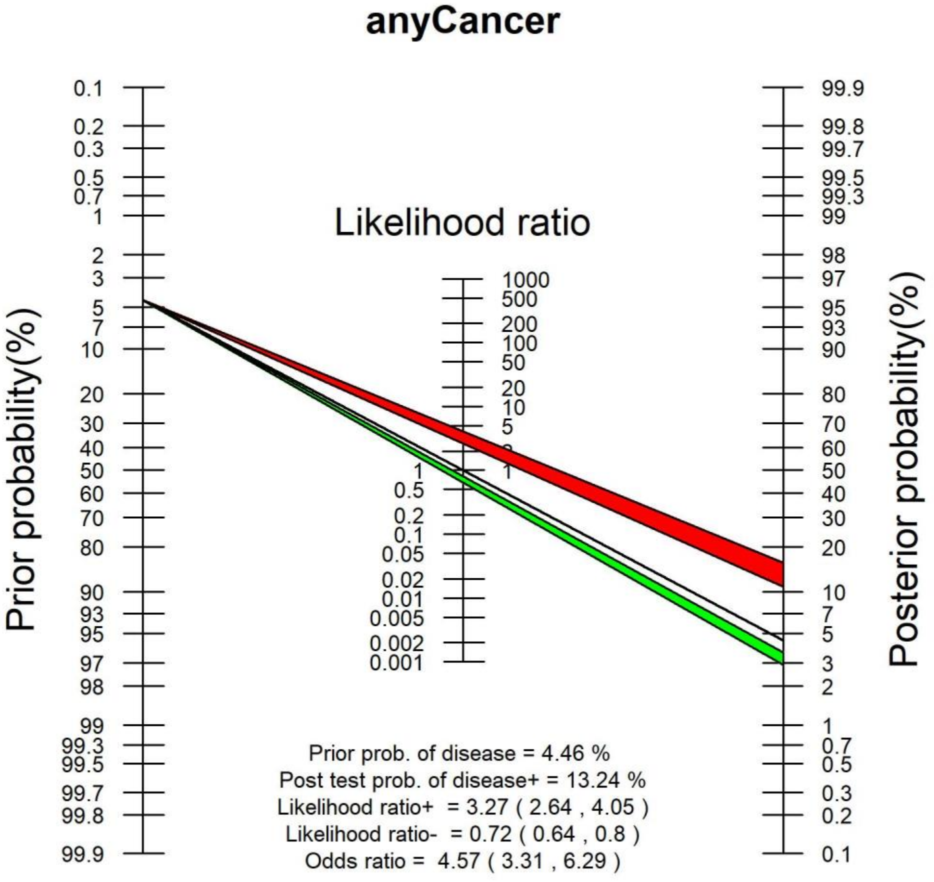


Figure 8.95 Results for any cancer using nearest neighbours

True Positives: 65; True Negatives: 3411; False Positives: 426; False Negatives: 114

Sensitivity: 0.363; Specificity: 0.889

F1: 0.194; F2: 0.269; MCC: 0.158

Positive Predictive Value: 0.132; Negative Predictive Value: 0.968

Positive Likelihood Ratio = 3.271 (2.641, 4.05)

Negative Likelihood Ratio = 0.716 (0.641, 0.801)

Odds ratio = 4.565; 95% CI = (3.312, 6.292)

8.4.4 Asthma

Results of the test run showing the prior prevalence of asthma and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.96.

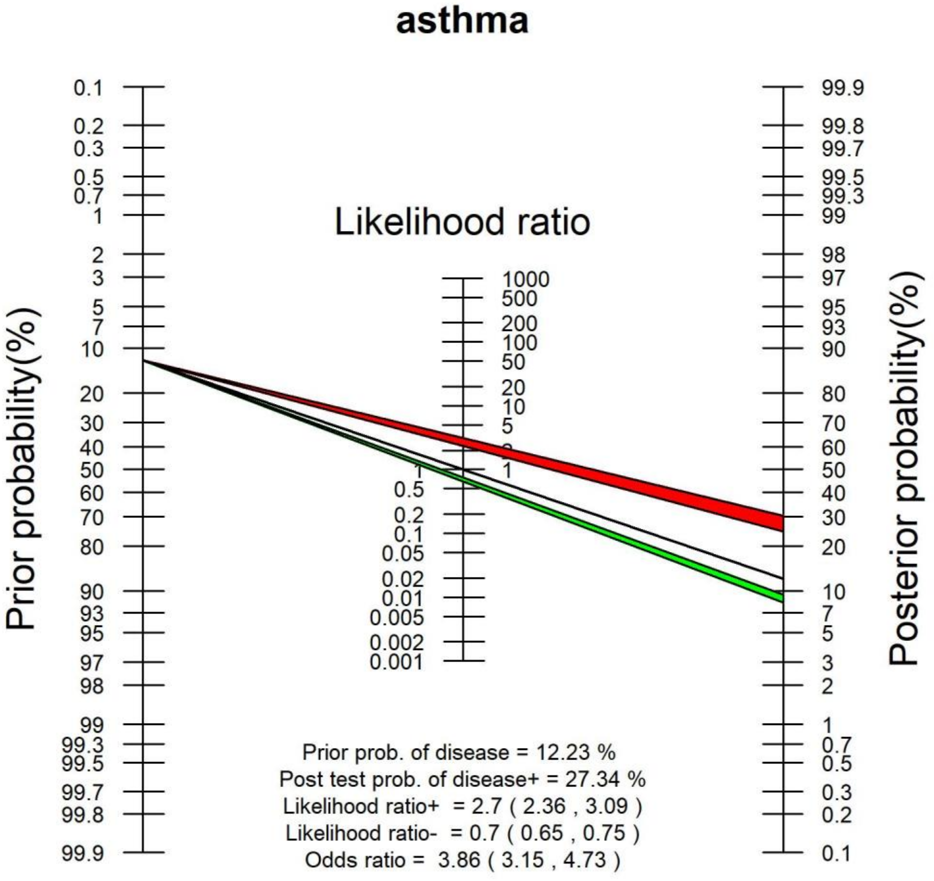


Figure 8.96 Results for asthma using nearest neighbours

True Positives: 198; True Negatives: 2974; False Positives: 526; False Negatives: 290

Sensitivity: 0.406; Specificity: 0.85

F1: 0.327; F2: 0.37; MCC: 0.218

Positive Predictive Value: 0.273; Negative Predictive Value: 0.911

Positive Likelihood Ratio = 2.701 (2.364, 3.085)

Negative Likelihood Ratio = 0.699 (0.649, 0.754)

Odds ratio = 3.862; 95% CI = (3.152, 4.731)

8.4.5 Autism

Results of the test run showing the prior prevalence of autism and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.97.

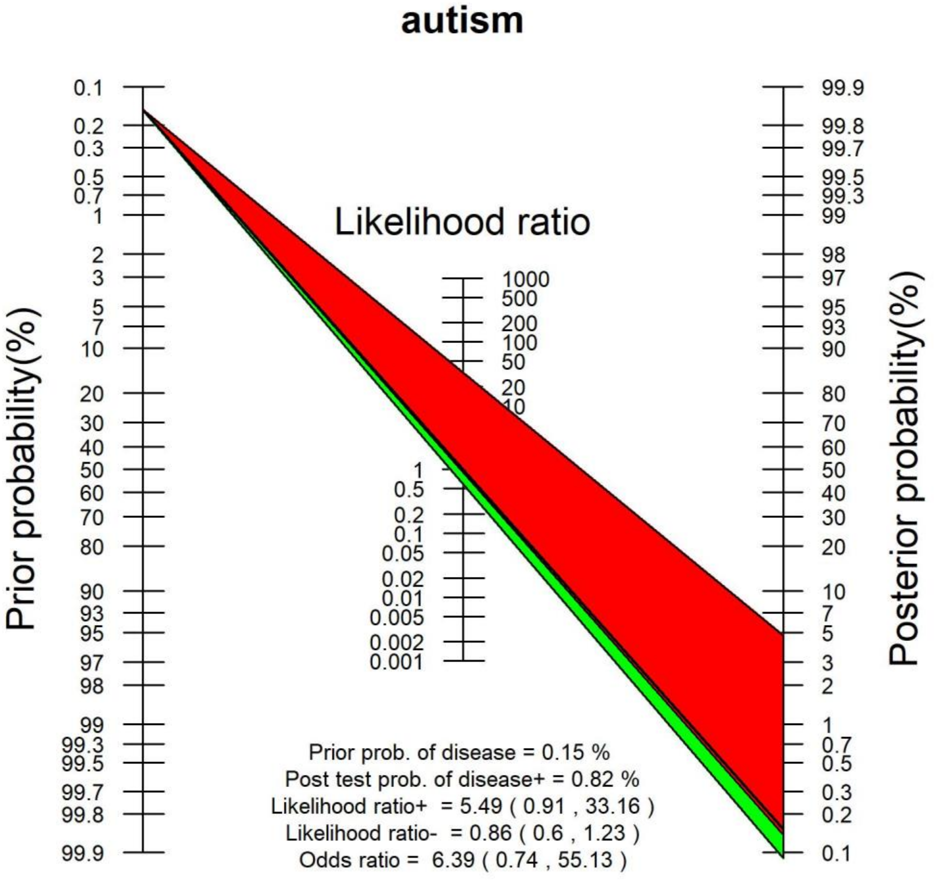


Figure 8.97 Results for autism using nearest neighbours

True Positives: 1; True Negatives: 3867; False Positives: 121; False Negatives: 5

Sensitivity: 0.167; Specificity: 0.970

F1: 0.016; F2: 0.034; MCC: 0.018

Positive Predictive Value: 0.008; Negative Predictive Value: 0.999

Positive Likelihood Ratio = 5.493 (0.910, 33.158)

Negative Likelihood Ratio = 0.854 (0.601, 1.229)

Odds ratio = 6.392; 95% CI = (0.741, 70.848)

8.4.6 Breast cancer

Results of the test run showing the prior prevalence of breast cancer and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.98.

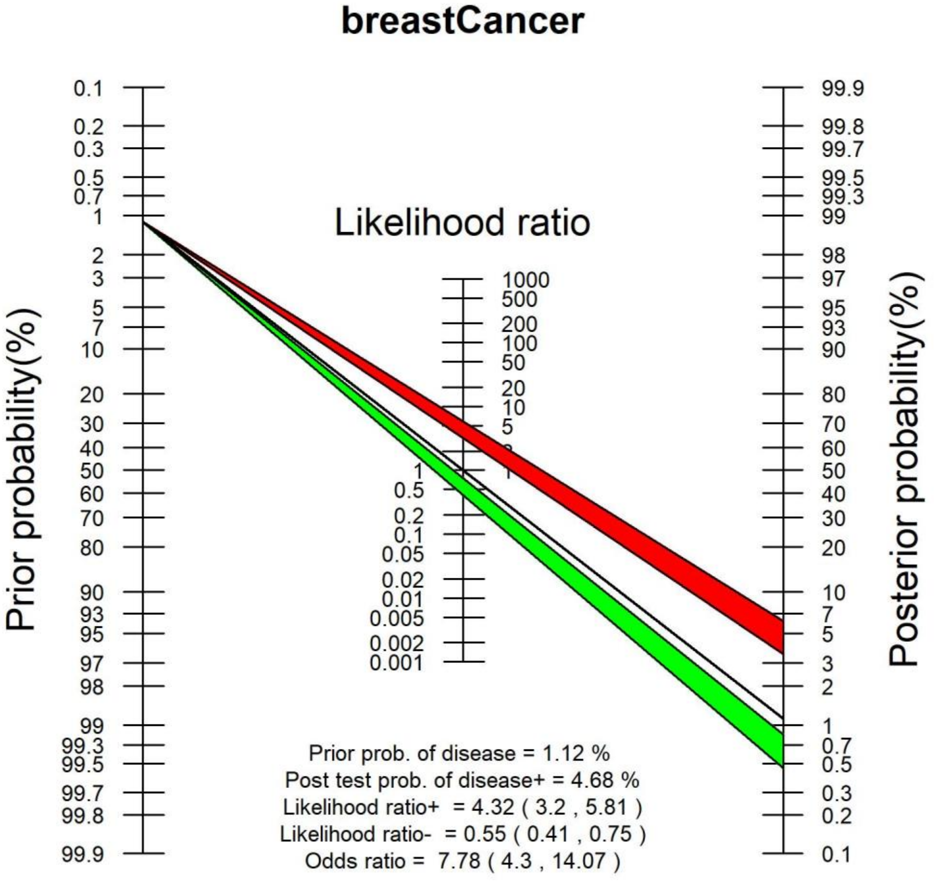


Figure 8.98 Results for breast cancer using nearest neighbours

True Positives: 23; True Negatives: 3491; False Positives: 469; False Negatives: 22

Sensitivity: 0.511; Specificity: 0.882

F1: 0.086; F2: 0.171; MCC: 0.125

Positive Predictive Value: 0.047; Negative Predictive Value: 0.994

Positive Likelihood Ratio = 4.316 (3.203, 5.814)

Negative Likelihood Ratio = 0.555 (0.411, 0.748)

Odds ratio = 7.782; 95% CI = (4.303, 14.072)

8.4.7 Bronchitis

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.99.

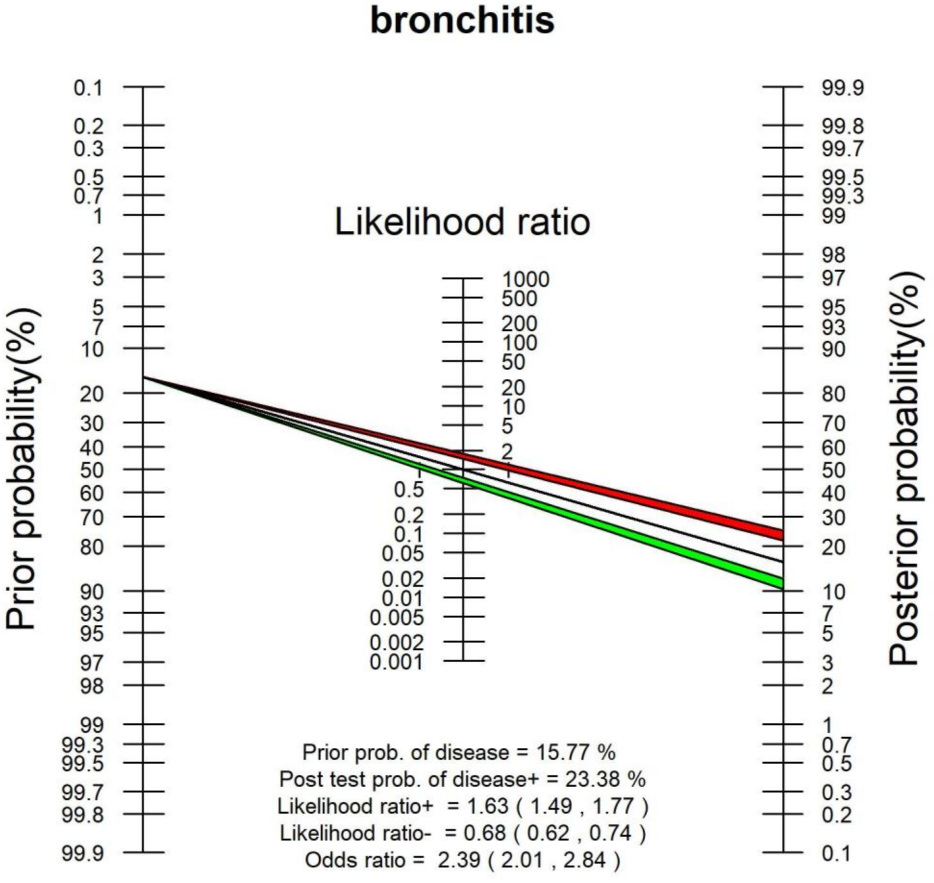


Figure 8.99 Results for bronchitis using nearest neighbours

True Positives: 348; True Negatives: 2227; False Positives: 1144; False Negatives: 283

Sensitivity: 0.552; Specificity: 0.661

F1: 0.328; F2: 0.433; MCC: 0.16

Positive Predictive Value: 0.233; Negative Predictive Value: 0.887

Positive Likelihood Ratio = 1.625 (1.493, 1.769)

Negative Likelihood Ratio = 0.679 (0.621, 0.743)

Odds ratio = 2.394; 95% CI = (2.015, 2.844)

8.4.8 Colon Cancer

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.100.

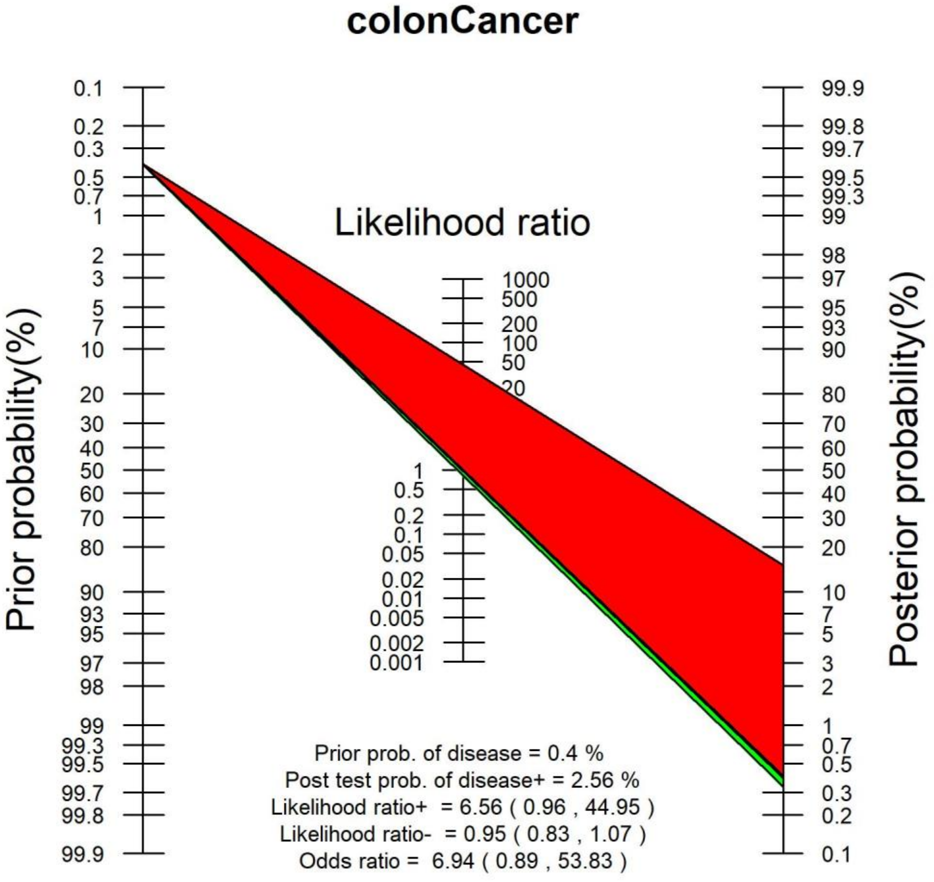


Figure 8.100 Results for colon cancer using nearest neighbours

True Positives: 1; True Negatives: 3953; False Positives: 38; False Negatives: 15

Sensitivity: 0.062; Specificity: 0.990

F1: 0.036; F2: 0.049; MCC: 0.019

Positive Predictive Value: 0.026; Negative Predictive Value: 0.996

Positive Likelihood Ratio = 6.564 (0.959, 44.95)

Negative Likelihood Ratio = 1.006 (0.834, 1.074)

Odds ratio = 6.935; 95% CI = (0.893, 53.836)

8.4.9 Eczema

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.101.

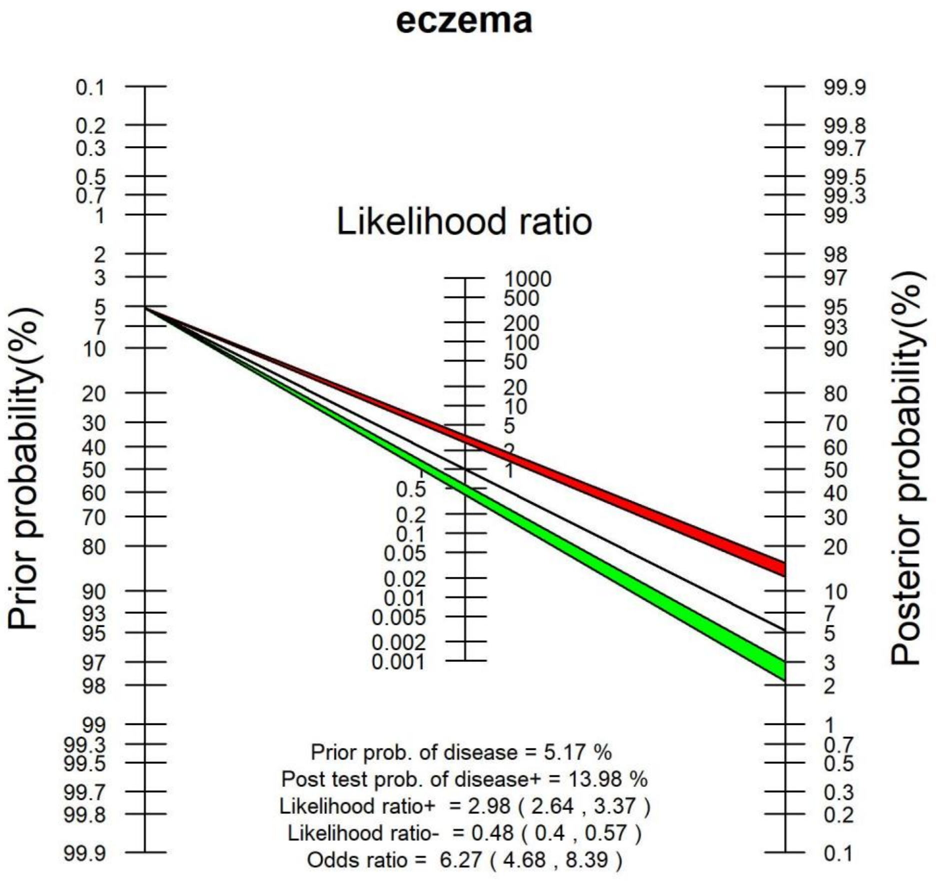


Figure 8.101 Results for eczema using nearest neighbours

True Positives: 129; True Negatives: 3004; False Positives: 793; False Negatives: 78

Sensitivity: 0.623; Specificity: 0.791

F1: 0.229; F2: 0.369; MCC: 0.217

Positive Predictive Value: 0.14; Negative Predictive Value: 0.975

Positive Likelihood Ratio = 2.984 (2.639, 3.373)

Negative Likelihood Ratio = 0.476 (0.399, 0.568)

Odds ratio = 6.265; 95% CI = (4.679, 8.388)

8.4.10 Gastro-oesophageal reflux disease

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.102.

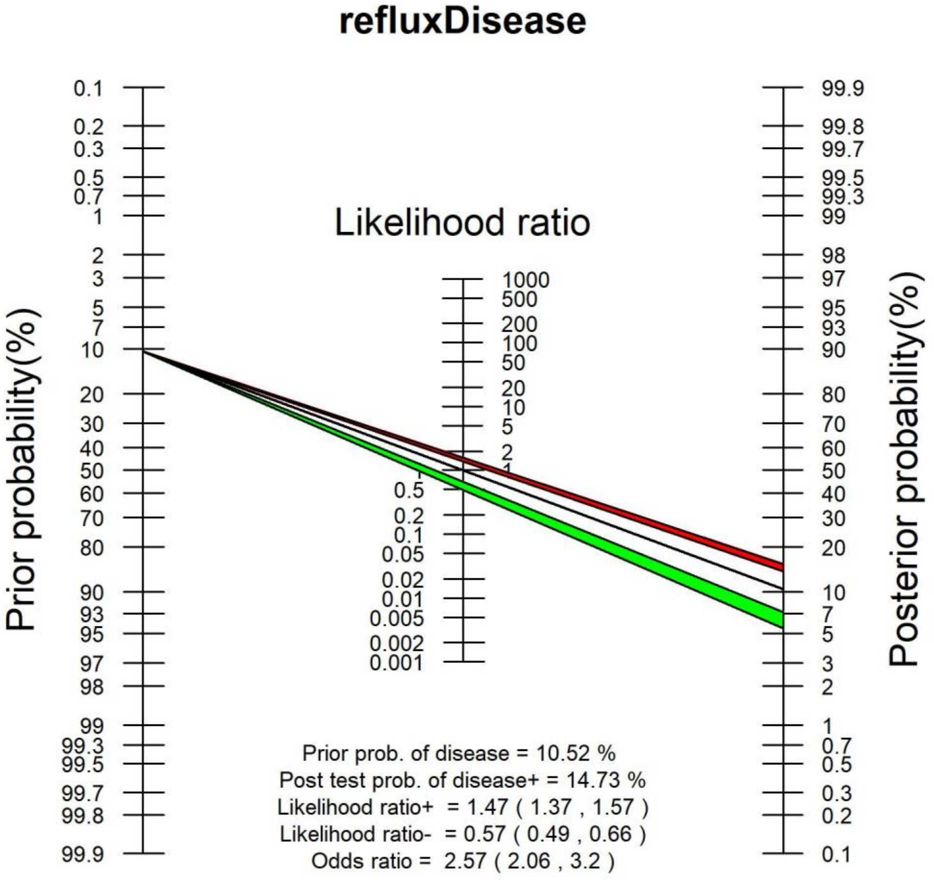


Figure 8.102 Results for gastro-oesophageal reflux disease using nearest neighbours

True Positives: 296; True Negatives: 1864; False Positives: 1718; False Negatives: 125

Sensitivity: 0.703; Specificity: 0.52

F1: 0.243 F2: 0.4; MCC: 0.137

Positive Predictive Value: 0.147; Negative Predictive Value: 0.937

Positive Likelihood Ratio = 1.466 (1.366, 1.574)

Negative Likelihood Ratio = 0.571 (0.491, 0.663)

Odds ratio = 2.569; 95% CI = (2.064, 3.199)

8.4.11 Gastroparesis

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.103.

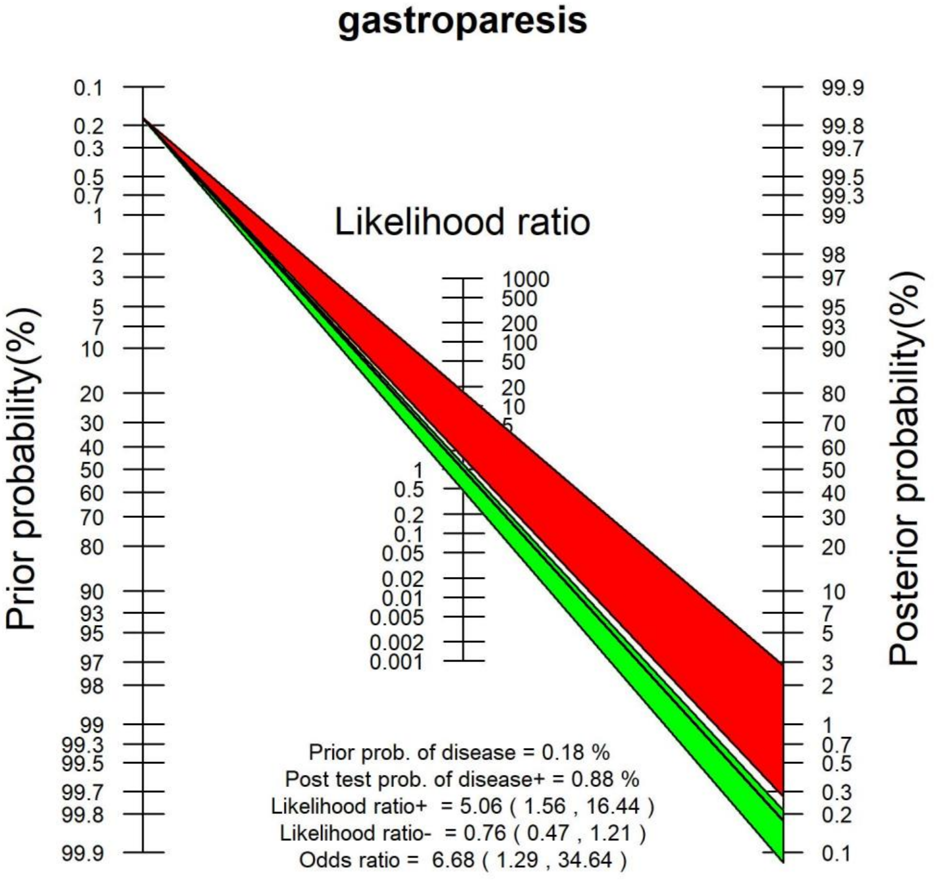


Figure 8.103 Results for gastroparesis using nearest neighbours

True Positives: 2; True Negatives: 3760; False Positives: 225; False Negatives: 5

Sensitivity: 0.286; Specificity: 0.944

F1: 0.017; F2: 0.039; MCC: 0.031

Positive Predictive Value: 0.009; Negative Predictive Value: 0.999

Positive Likelihood Ratio = 5.06 (1.558, 16.438)

Negative Likelihood Ratio = 0.757 (0.474, 1.21)

Odds ratio = 6.684; 95% CI = (1.29, 34.644)

8.4.12 Gout

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.104.

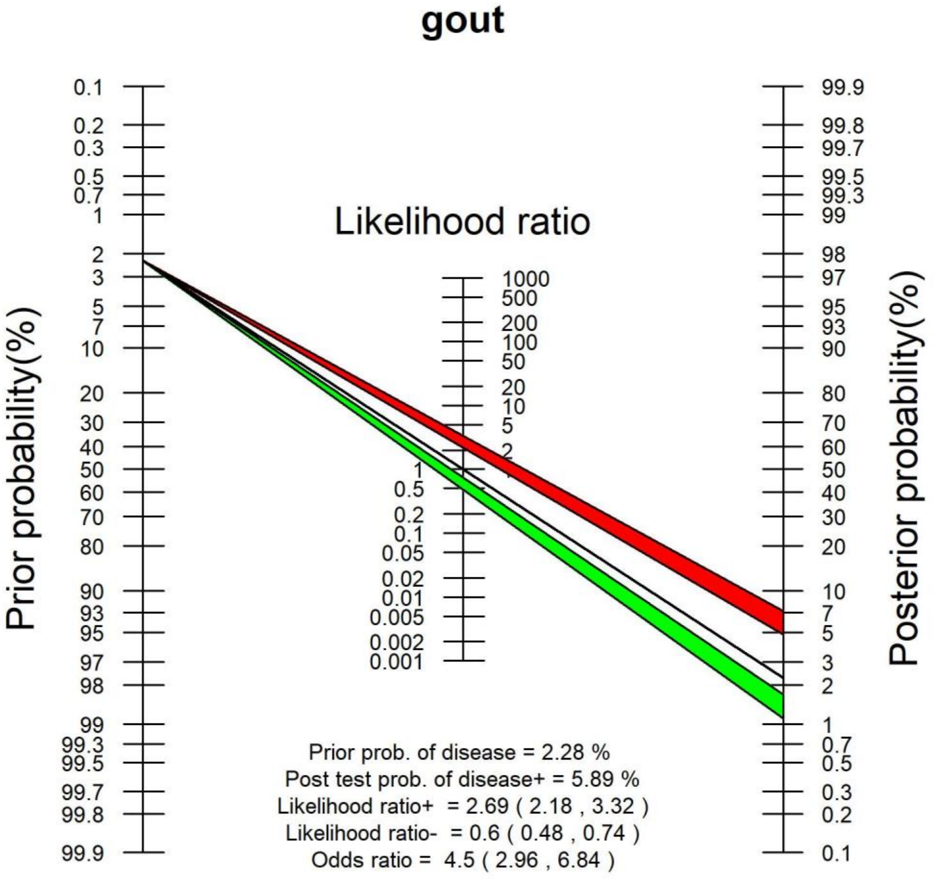


Figure 8.104 Results for gout using nearest neighbours

True Positives: 47; True Negatives: 3158; False Positives: 750; False Negatives: 44

Sensitivity: 0.516; Specificity: 0.808

F1: 0.106; F2: 0.202; MCC: 0.12

Positive Predictive Value: 0.059; Negative Predictive Value: 0.986

Positive Likelihood Ratio = 2.691 (2.184, 3.317)

Negative Likelihood Ratio = 0.598 (0.484, 0.74)

Odds ratio = 4.498; 95% CI = (2.959, 6.837)

8.4.13 Obesity

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.105.

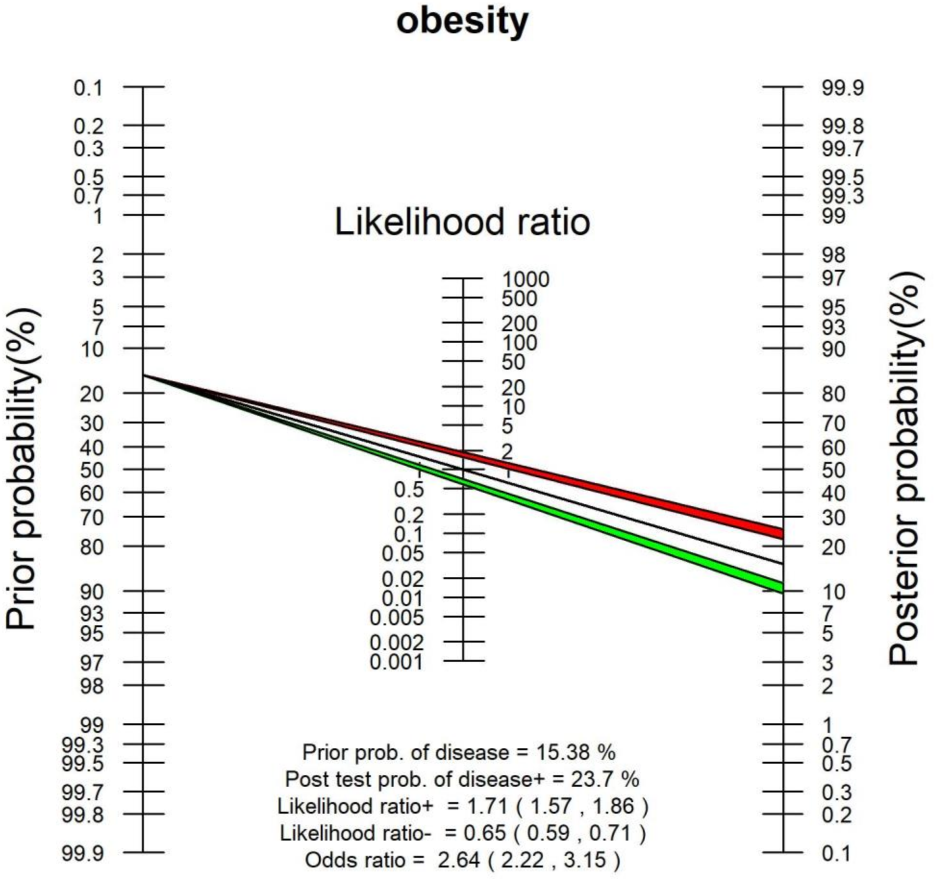


Figure 8.105 Results for obesity using nearest neighbours

True Positives: 348; True Negatives: 2267; False Positives: 1118; False Negatives: 267

Sensitivity: 0.566; Specificity: 0.67

F1: 0.334; F2: 0.443; MCC: 0.176

Positive Predictive Value: 0.237; Negative Predictive Value: 0.895

Positive Likelihood Ratio = 1.713 (1.575, 1.864)

Negative Likelihood Ratio = 0.648 (0.591, 0.712)

Odds ratio = 2.643; 95% CI = (2.219, 3.148)

8.4.14 Osteoarthritis

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.106.

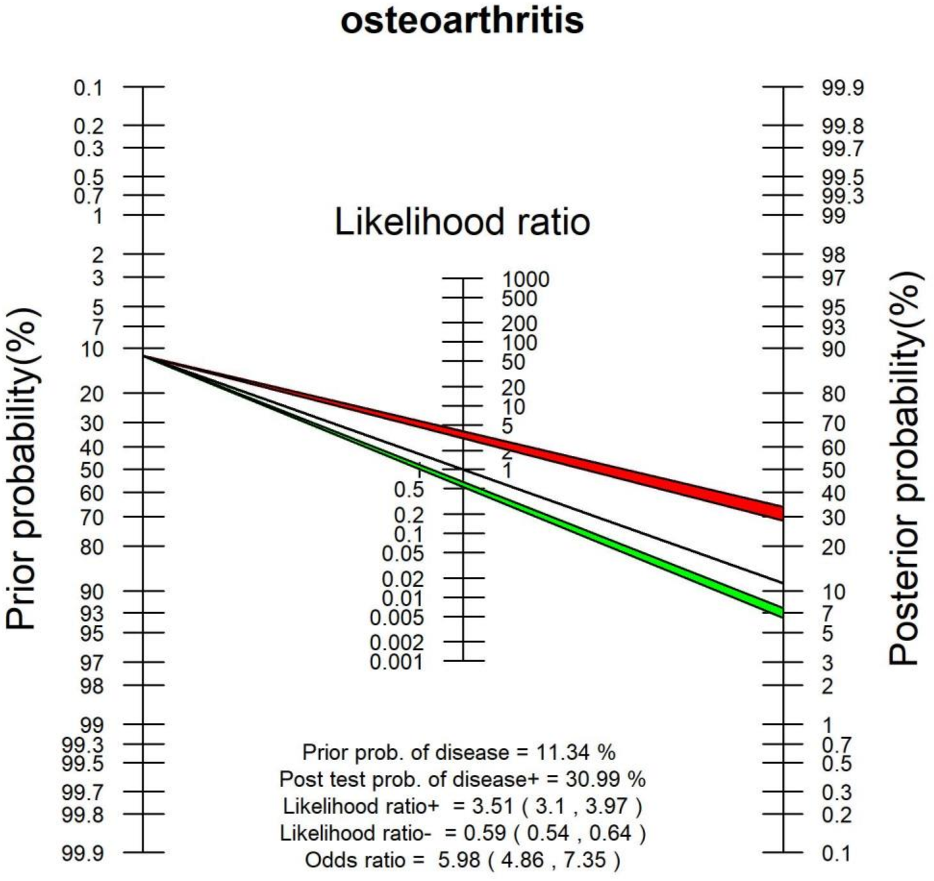


Figure 8.106 Results for osteoarthritis using nearest neighbours

True Positives: 225; True Negatives: 3047; False Positives: 501; False Negatives: 229

Sensitivity: 0.496; Specificity: 0.859

F1: 0.381; F2: 0.443; MCC: 0.291

Positive Predictive Value: 0.31; Negative Predictive Value: 0.93

Positive Likelihood Ratio = 3.51 (3.103, 3.97)

Negative Likelihood Ratio = 0.587 (0.536, 0.644)

Odds ratio = 5.976; 95% CI = (4.859, 7.349)

8.4.15 Prostate cancer

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.107.

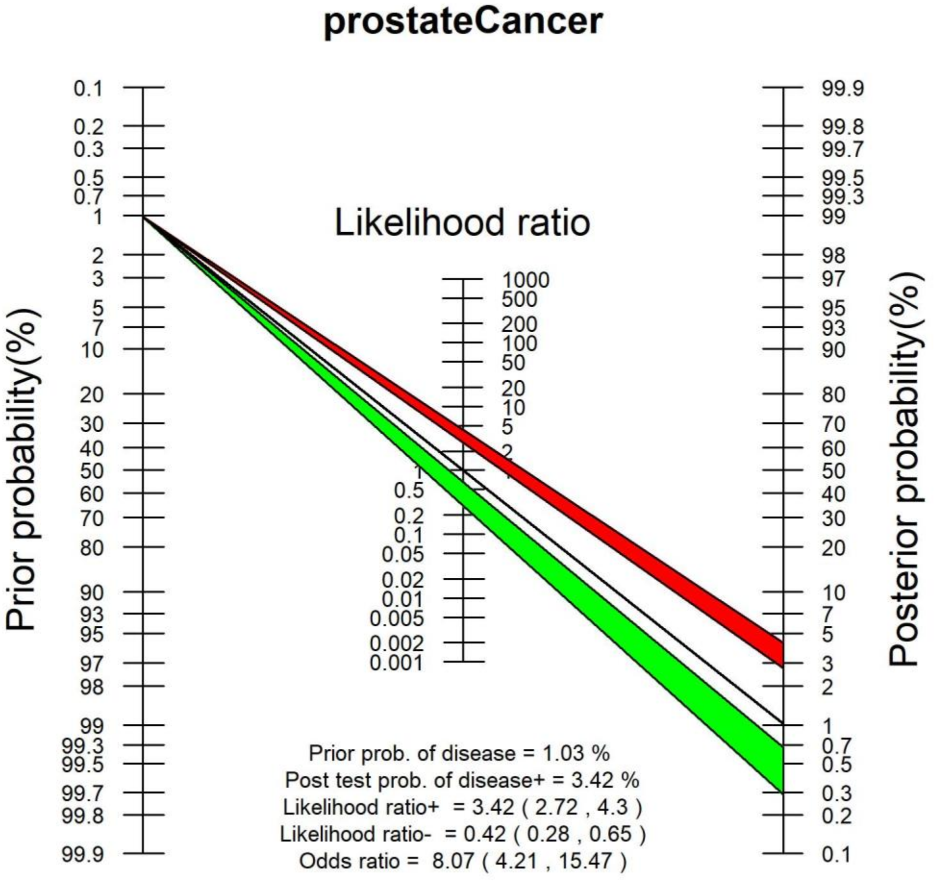


Figure 8.107 Results for prostate cancer using nearest neighbours

True Positives: 27; True Negatives: 3194; False Positives: 763; False Negatives: 14

Sensitivity: 0.659; Specificity: 0.807

F1: 0.065; F2: 0.142; MCC: 0.117

Positive Predictive Value: 0.034; Negative Predictive Value: 0.996

Positive Likelihood Ratio = 3.415 (2.715, 4.296)

Negative Likelihood Ratio = 0.423 (0.276, 0.647)

Odds ratio = 8.073; 95% CI = (4.213, 15.469)

8.4.16 Stress

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.108.

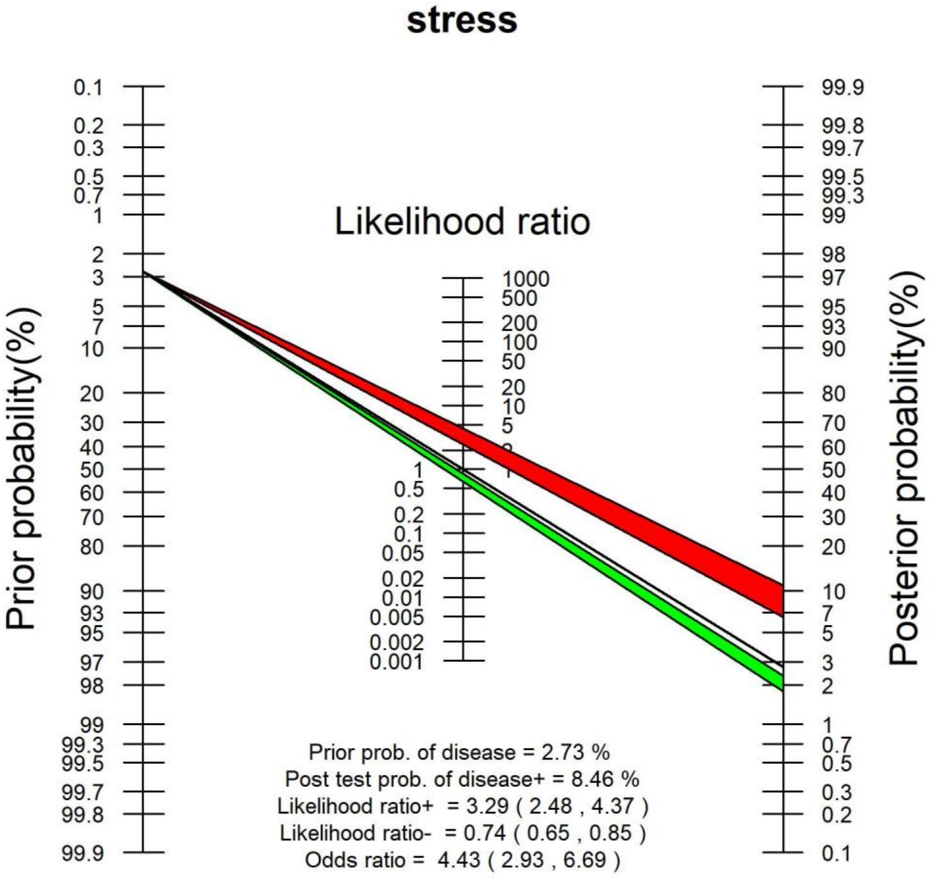


Figure 8.108 Results for stress using nearest neighbours

True Positives: 36; True Negatives: 3491; False Positives: 389; False Negatives: 73

Sensitivity: 0.33; Specificity: 0.9

F1: 0.135; F2: 0.209; MCC: 0.121

Positive Predictive Value: 0.085; Negative Predictive Value: 0.98

Positive Likelihood Ratio = 3.294 (2.481, 4.374)

Negative Likelihood Ratio = 0.744 (0.652, 0.85)

Odds ratio = 4.426; 95% CI = (2.929, 6.687)

8.4.17 Thyrotoxicosis

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.109.

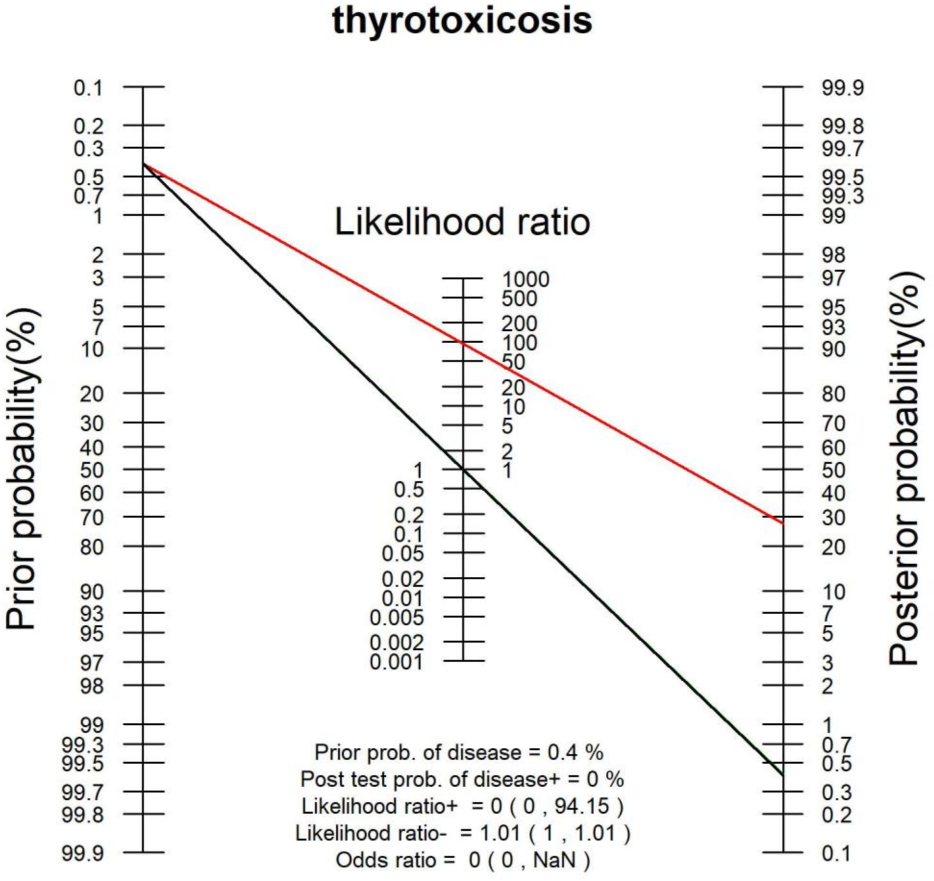


Figure 8.109 Results for thyrotoxicosis using nearest neighbours

True Positives: 1; True Negatives: 3973; False Positives: 21; False Negatives: 16

Sensitivity: 0; Specificity: 0.995

F1: 0; F2: 0; MCC: 0.017

Positive Predictive Value: 0; Negative Predictive Value: 0.996

Positive Likelihood Ratio = 0 (0, 94.152)

Negative Likelihood Ratio = 1.005 (1.003, 1.008)

Odds ratio = 0; 95% CI = (0, -)

8.4.18 Type 2 diabetes

Results of the test run showing the prior prevalence of acute sinusitis and the prevalences in the group predicted to have the condition and in the group predicted not to have the condition are shown in Figure 8.110.

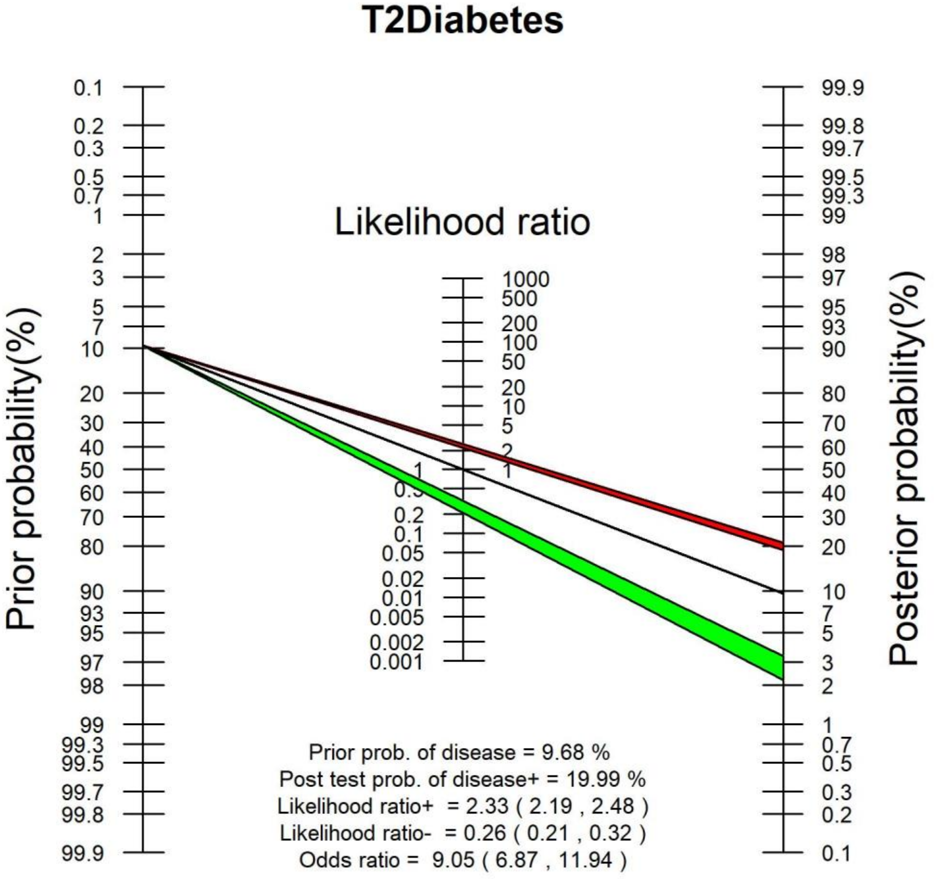


Figure 8.110 Results for type 2 diabetes using nearest neighbours

True Positives: 324; True Negatives: 2320; False Positives: 1298; False Negatives: 64

Sensitivity: 0.835; Specificity: 0.641

F1: 0.322; F2: 0.51; MCC: 0.287

Positive Predictive Value: 0.2; Negative Predictive Value: 0.973

Positive Likelihood Ratio = 2.328 (2.188, 2.477)

Negative Likelihood Ratio = 0.257 (0.205, 0.322)

Odds ratio = 9.052; 95% CI = (6.865, 11.937)

8.5 Summary of results for nearest neighbours and clustering methods.

A summary of the results from both the nearest neighbours method and the clustering method for predicting the presence or absence of conditions in a record is shown in Table 8.4.

		Clustering			**Nearest neighbours**		
Condition	Prevalence	Clusters	Positive Likelihood Ratio (95 % CI)	Negative Likelihood Ratio (95 % CI)	K (level)	Positive Likelihood Ratio (95 % CI)	Negative Likelihood Ratio (95 % CI)
Acute sinusitis	14.2 %	2	1 (1 , 1)	-	399 (11)	2.72 (2.479, 2.9)	0.54 (0.49, 0.60)
Allergic rhinitis	17.7 %	3	1.03 (1.01, 1.05)	0.69 (0.50, 0.94)	2016 (11)	1.55 (1.44, 1.67)	0.65 (0.60, 0.72)
Any cancer	4.4 %	130	4.90 (3.70, 6.50)	0.78 (0.71, 0.85)	286 (11)	3.27 (2.641, 4.05)	0.72 (0.64, 0.80)
Asthma	11.5 %	2	2.58 (1.97, 3.36)	0.91 (0.88, 0.95)	172 (7)	2.70 (2.364, 3.09)	0.70 (0.65, 0.75)
Autism	0.2 %	90	6.32 (2.56, 15.59)	0.66 (0.39, 1.14)	37 (11)	5.49 (0.91, 33.16)	0.86 (0.60, 1.23)
Breast cancer	1.3 %	920	7.24 (4.52, 11.59)	0.75 (0.63, 0.89)	525 (7)	4.32 (3.203, 5.81)	0.56 (0.41, 0.75)
Bronchitis	15.4 %	3	0.95 (0.89, 1.01)	1.11 (0.99, 1.25)	3628 (7)	1.63 (1.493, 1.77)	0.68 (0.62, 0.74)
Colon cancer	0.3 %	359	4.87 (1.38, 17.23)	0.85 (0.64, 1.12)	1900 (11)	6.56 (0.96, 44.95)	0.95 (0.83, 1.07)
Eczema	5.2 %	104	3.43 (2.84,	0.68 (0.61,	273 (11)	2.98 (2.639,	0.48 (0.40, 0.57)

			4.13)	0.76)		3.37)	
Gastro-intestinal reflux disease	10.15 %	3	-	-	2016 (11)	1.47 (1.366, 1.57)	0.57 (0.49, 0.66)
Gastroparesis	0.2 %	530	-	1.02 (1.01, 1.02)	2350 (7)	5.06 (1.558, 16.44)	0.76 (0.47, 1.21)
Gout	2.2 %	37	2.69 (1.94, 3.72)	0.78 (0.67, 0.9)	406 (7)	2.69 (2.184, 3.32)	0.60 (0.48, 0.74)
Obesity	15.4 %	3	-	-	2100 (11)	1.713 (1.575, 1.86)	0.648 (0.59, 0.71)
Osteoarthritis	11.9 %	2	-	-	209 (11)	3.51 (3.10, 3.97)	0.59 (0.54, 0.64)
Prostate cancer	1.3 %	280	5.29 (3.39, 8.24)	0.75 (0.63, 0.89)	138 (11)	3.00 (2.24, 4.00)	0.58 (0.42, 0.79)
Stress	3.0 %	150	2.97 (2.10, 4.21)	0.84 (0.76, 0.92)	180 (11)	3.29 (2.481, 4.37)	0.74 (0.65, 0.85)
Thyrotoxicosis	0.4 %	359	-	1.03 (1.03, 1.04)	28 (11)	0 (0, 94.15)	1.01 (1.00, 1.01)
Type 2 diabetes	10.3 %	3	1 (1, 1)	-	300 (7)	2.33 (2.188, 2.48)	0.26 (0.21, 0.32)

Table 8.4 Summary of results from nearest neighbours method and clustering method

8.6 Comparison to Logistic Regression

Conditions selected for comparison were two high prevalence conditions (Allergic rhinitis, Bronchitis, obesity), two low prevelance conditions (Gastroparesis and autism – Prader-Willi Disease, the lowest-prevalence condition in the data set, was omitted due to its extreme low prevalence) and one mid-prevalence condition (eczema).

Logistic regression performed using KNIME. The regression used stochastic average gradient solver since this was understood to work well with large tables and with tables where the number of columns is comparable to or greater than the number of rows[276], maximum epochs = 120, laplace priors (for high-dimensional data). Granularity levels of the CTV3 tree were chosen to be the same as previously established for the clustering and k nearest neighbours tests, i.e. level 7 and level 11. As for the clustering and k nearest neighbours tests, there was a 50% test / 50 % train split on the whole data set. Prior to performing the regression, input values were z-normalised.

Condition	Method	Level	F1 score	F2 score	Sensitivity	Specificity	Positive Likelihood Ratio	Negative Likelihood	Odds Ratio
Allergic Rhinitis	KNN	11	0.354	0.469	0.599	0.614	1.55	0.654	2.372
Allergic Rhinitis	LR	7	0.076	0.052	0.042	0.983	2.443	0.975	2.508
Allergic Rhinitis	LR	11	0.113	0.080	0.067	0.974	2.518	0.959	2.626
Bronchitis	KNN	7	0.328	0.433	0.552	0.661	1.625	0.679	2.394
Bronchitis	LR	7	0.139	0.981	0.223	0.164	7.240	0.881	5.146
Bronchitis	LR	11	0.213	0.155	0.131	0.982	7.439	0.885	8.408
Obesity	KNN	11	0.334	0.443	0.566	0.67	1.713	0.648	2.643
Obesity	LR	7	0.34	0.269	0.233	0.984	14.823	0.779	19.030
Obesity	LR	11	0.409	0.326	0.287	0.979	13.422	0.729	18.420

Gastroparesis	KNN	7	0.017	0.039	0.286	0.944	5.06	0.757	6.684
Gastroparesis	LR	7	0	0	0	0.999	0	0	0
Gastroparesis	LR	11							
Autism	KNN	11	0.016	0.034	0.167	0.970	5.493	0.854	6.392
Autism	LR	7	0	0	0	1	0	0	0
Autism	LR	11	0	0	0	1	0	0	0
Eczema	KNN	11	0.229	0.369	0.623	0.791	2.984	0.476	6.265
Eczema	LR	7	0.093	0.061	0.05	0.999	33.885	0.951	35.615
Eczema	LR	11	0.137	0.095	0.079	0.996	21.649	0.925	23.413

KNN generally performed better than Logistic Regression when judged by F-scores (both F1 and F2), sensitivity and specificity, for all conditions (note that the low prevalence conditions – autism and gastroparesis – performed poorly with logistic regression, predicting no positive cases). However, for mid- and high-prevalence conditions, logistic regression produced a higher positive likelihood ratio than did KNN. KNN produced a lower negative likelihood ratio than did logistic regression in all cases (noting that logistic regression was unable to make predictions in the low prevalence condition cases). As a result of the logistic regression's better performance in positive likelihood ratio, it also gave higher results for odds ratio than did KNN. This suggests that if the logistic regression predicts positive for a condition then there is greater confidence in the record containing that condition than if KNN predicts positive. However, if the logistic regression predicts negative for a record then there is less confidence that the record does truly not contain the condition than if the KNN predicted negative.

8.7 Conclusions

For some conditions, the methods developed were able to predict the presence of a condition of interest, and both the nearest neighbours method and the clustering method were able to discriminate records likely to contain the condition against those unlikely to contain the condition. These conditions were:

Acute sinusitis (knn method only), allergic rhinitis, any cancer, asthma, autism (clustering method only), breast cancer, bronchitis (knn method only), eczema, gastro-intestinal reflux disease (knn method only), gout, obesity (knn method only), osteoarthritis (knn method only), prostate cancer, stress and type 2 diabetes (knn method only).

For colon cancer, it was possible to identify a group of records with a raised likelihood of having the condition from the general records set, but not state that the remaining records had a lowered likelihood of containing the condition.

For a few conditions it was not possible to identify groups with raised or lowered likelihoods of having particular conditions. These conditions were thyrotoxicosis and gastroparesis, both conditions with a low prevalence in the records set. Work using a larger record set may demonstrate the potential for applying these techniques to these conditions, since currently the absolute presence count of the lowest prevalence conditions is extremely small, and having a larger quantity of records with these low prevalence conditions may help in identifying clusters or establishing whether near neighbours for records with these conditions do or do not have a prevalence significantly higher (or lower) than the population prevalence. Currently the low numbers for these conditions mean that it is difficult to establish significance.

The results show that for many conditions there may be value in utilising the event history contained in primary care records in order to select candidates for screening tests (e.g. diabetes) or for pre-emptive health advice (e.g. obesity or gout).

One set of distance calculation methods was clearly better performing on this data set than other methods (the bray, jaccard, gower, morisita, horn, raup, binomial, chao, cao and mahalanobis methods, see section 7.3.6.1). It is not clear why these methods performed best, although it can be noted that all these methods were developed for, or have been adopted for use for, ecological site similarity analysis and all are implemented in the ‘vegdist’ function in the R vegan package [267]. Other factors were less clear: for the majority of the conditions tests performance was better when event codes were aggregated at CTV3 hierarchy level 7, with others performing better at level 11.There was no clear discriminator between these groups. The value of k for the k nearest neighbours was also a factor that was difficult to optimise perfectly, with the heuristic method used for the selection of k value for each condition not wholly satisfactory.

9 DISCUSSION

9.1 Motivation and methods

9.1.1 Primary motivations

The general motivation for this work was to explore the secondary uses of medical records, looking at benefits for individuals, populations and for organisations. This was covered in the literature review of Chapter 2 and in the case studies of Chapter 4. A specific key motivation for this work was to investigate whether existing medical histories could be used to calculate revised likelihoods of individuals having particular conditions, and, if the likelihood was increased, to present that information at an individual level to clinicians, for example in a general practice surgery visit, or as a pre-screening tool to select individuals calculated to be at increased risk of the condition and suggest them as candidates for screening. As an example of the benefit of discovering increased likelihood in some individuals, thyrotoxicosis was considered to be a condition which would benefit from early detection but whose discovery could be delayed due to the similarity of its symptoms with other more prevalent conditions. Conditions producing symptoms similar to those produced by early-stage thyrotoxicosis include depression, diabetes and viral infection.

Rarely does a single factor guarantee that a disease will occur within an individual's lifetime - cigarette smoking, faulty genes, poor environment, workplace factors. But each can modulate the risk of having a particular condition. Similarly, a history of other conditions does not guarantee the occurrence of another condition nor guarantee that it won't occur.

9.1.2 Aggregation of data from disparate data sets

Sets of records derived from disparate clinical data sets were combined into a single consolidated data set suitable for further analysis. Event codes within the composite data set were mapped to a common coding system of Clinical Terms Version 3. In creating the composite data set, the fields common to the data sets were identified and retained; fields present in only some of the source data sets were ignored. This was to ensure that the data set was as balanced as possible, with predictions not based on data fields that were available for only a subset of the data set. Event codes were mapped to a single coding system, Clinical Terms Version 3, chosen because of its simple hierarchical structure, the ready availability of a mapping table from Read Codes version 2 to CTV3, its alignment with, but relative simplicity when compared to, SNOMED CT, and its more modern categorisation of some term relationships. Mapping to this coding system was performed using an existing mapping table for events coded in Read v2 (Read v2 to CTV3, NHS TRUD) and via a semi-automatic indirect mapping technique for events coded in ICD-9-CM (ICD-9_CM to SNOMED CT via UMLS/Nadkarni; SNOMED CT to CTV3 via NHS TRUD). This was to ensure that all codes were in a common coding system that was relatively simple and understood. Of the 4342 unique ICD-9-CM codes in the US data set, 4328 (99.7 %) were mapped to CTV3 codes by a combination of automatic and manual techniques. Of those codes that were mapped, 96.0 % were mapped appropriately when judged by a domain expert.

9.1.3 Generation of codelists

Sets of event codes, or ‘codelists’, were generated, by primarily manual methods of searching the term descriptions for the complete CTV3 code set. As described in section 5.4, successful searching relied on knowledge of differences between spellings and complete term names in US English versus British English. However it was not possible to be completely sure that all codes were included in the appropriate codelist: Some codes that were not included in the manual generation of codelists were discovered by use of decision trees, also described in section 5.4, although this method of checking was only possible for codes that were in the composite data set.

9.1.4 Development of techniques for record matching

Individual records from the composite record set, each record containing a set of event codes, were then grouped according to similarities between the records. The grouping process used standard machine learning techniques of clustering and k nearest neighbours, but with significant data preparation. Target conditions were defined using the assembled codelists. For each method, optimum factors were derived, by initial analysis of the data set and by training runs, in order to simplify subsequent testing and to ensure that programs ran in a feasible time. The grouping process utilised the hierarchical structure of the CTV3 coding system to group similar codes. For any one chosen condition, the condition prevalence within the group of similar records was calculated and compared to the prevalence of the condition in the group of non-similar records. These prevalence calculations were used to determine the positive and negative likelihood ratios, showing whether prediction of presence of a condition in the records was significantly raised or lowered when compared to the prevalence in the complete records set. Results from the two approaches for the selection of conditions were generated and compared to see which conditions performed best and which method performed best.

9.1.4.1 Clustering

Clustering distributes the records in the data set into a set of groups according to each record’s similarity to other record assigned to the same group. The number of clusters must be pre-defined, as are a number of other important factors, including the measure by which similarity is measured and the technique used for creating the clusters. Once a record was placed into a cluster, a prediction was made regarding the presence or absence of a condition by calculating the prevalence of that condition in other records in the cluster and testing to see if that prevalence was significantly higher than the population or not. Factors used in Chapter 6 in the clustering method were to use the binomial method for calculating the similarity between records and the Ward.D2 method for forming clusters, each chosen following testing on a small sample of the complete data set. Other factors were selected separately for each condition following test runs on multiple samples from the test portion of the data set: these were the choice of level of the CTV3 hierarchy (choices were limited to level 7 or level 11 following analysis of the codes in the complete data set) and the optimum number of clusters. This analysis is shown in Chapter 6, section 6.2, with the results from that analysis shown in Table 9.1.

Note that the prevalence of Prader-Willi disease is very small, meaning that there were insufficient number of records containing that condition for meaningful analysis.

In the testing to determine the optimum number of clusters, a small number of clusters was suggested for some conditions, but a much larger number of clusters for other conditions. It appeared that conditions with a higher prevalence had a small number of clusters suggested (e.g. type 2 diabetes had 3 clusters suggested), whereas conditions with a low prevalence had a high number of clusters suggested (e.g. thyrotoxicosis had 359 clusters suggested). It would seem that for low prevalence conditions, with few absolute numbers of cases in the data set, those cases are clustered with a small number of other records, whereas with high prevalence conditions, the cases are clustered with a high number of other records. Further investigation is required to determine what is causing this to happen, in particular whether the small number of cases in low prevalence conditions is insufficient to establish a strong pattern, or whether it is something intrinsic to the conditions that causes them to have a small or large number of clusters suggested.

Condition	Best CTV3 level	Best no of clusters at CTV3 level	F2 at best level and k	Prevalence in training set
Acute sinusitis	11	2	0.933	0.154
Allergic rhinitis	11	3	1.058	0.198
Any cancer	7	130	0.479	0.042
Asthma	7	2	0.805	0.122
Autism	11	90	0.034	0.001
Breast cancer	11	920	0.324	0.011
Bronchitis	7	3	0.966	0.159
Colon cancer	7	359	0.215	0.005
Eczema	7	104	0.660	0.054
Gastroparesis	11	530	0.303	0.002
Gout	11	37	0.373	0.024
Obesity	11	3	0.981	0.160
Osteoarthritis	7	2	0.796	0.117
Prader-Willi	-	-	-	0
Prostate cancer	7	280	0.323	0.111
Reflux disease	11	3	0.758	0.109
Stress	7	150	0.394	0.026
Thyrotoxicosis	7	359	0.120	0.006
Type 2 diabetes	11	3	0.746	0.102

Table 9.1 Summary of results for optimum factors for the clustering method

9.1.4.2 Nearest neighbours

The nearest neighbours method starts with a record of interest and seeks to find the most similar, by some measure, records in the rest of the data set. The technique employed is described in Chapter 6. In summary, a distance matrix was formed for a sample from the training set, using the binomial method as previously established. This distance matrix contained a distance measure for every record in the training sample to every other record in the training sample. For each record, all other records were ordered by this distance. Those other records closest to the record of interest were classed as 'near neighbours' the prevalence of a condition in the set of nearest neighbours was then used to calculate a prediction for the record of interest. The optimum number (by convention, 'k') of nearest neighbours, and the optimum level of the CTV3 hierarchy, was selected using multiple runs on the test set.

This analysis is shown in Chapter 6, section 6.2, with the results from that analysis shown in Table 9.2. Again, the low prevalence of Prader-Willi disease means that there was an insufficient number of records containing that condition for meaningful analysis. In a similar manner to that for clustering (section 9.1.4.1), some conditions had a high value of k recommended, others had a low value of k recommended. There was a less consistent effect than for clustering, however, for example autism and gastroparesis, both low prevalence conditions in the data set, had values of k suggested of 37 and 2350 respectively. It was noted, however, that for high prevalence conditions the plot of F2 score against k in the testing phase produced relatively flat and smooth curves, with the value of F2 dropping only at high values for k, suggesting that cases in the neighbours were relatively evenly distributed closer to record of interest, but the density of the cases was falling at large distance from the record of interest. For low prevalence conditions this was not the case, with curves not being smooth, suggesting that for low prevalence conditions randomness has a large impact. This can be seen, for example, when comparing the curves of F2 score against k for colon cancer and for eczema, a typical low prevalence condition and a typical high prevalence condition. For colon cancer, the F2 vs k plot (Figures 8.71 and 8.72) has a noisy curve, and its corresponding nomogram (Figure 8.100) shows no significant Likelihood Ratio. However for eczema, the F2 vs K plots (Figures 8.73 and 8.74) are smooth and the corresponding nomogram shows significantly raised positive likelihood ratio and significantly lowered negative likelihood ratio. Again, as for the cluster analysis, it cannot be stated absolutely whether this effect is due to the low or high prevalence of each condition, or something intrinsic to the condition - perhaps some conditions present a number of effects and related conditions, whereas other conditions exist in some isolation. Further work is needed to establish the reason for this behaviour.

Condition	Best CTV3 level	Best k at CTV3 level	F2 at best level and k	Prevalence in training set
Acute sinusitis	11	399	0.489	0.154
Allergic rhinitis	11	2016	0.469	0.198
Any cancer	11	286	0.269	0.042
Asthma	7	172	0.370	0.122
Autism	11	37	0.034	0.001
Breast cancer	7	525	0.171	0.011
Bronchitis	7	3628	0.433	0.159
Colon cancer	11	1900	0.049	0.005
Eczema	11	273	0.369	0.054
Gastroparesis	7	2350	0.039	0.002
Gout	7	406	0.202	0.024
Obesity	11	2100	0.443	0.160
Osteoarthritis	11	209	0.443	0.117
Prader-Willi	-	-	-	-
Prostate cancer	11	138	0.142	0.111
Reflux disease	11	2016	0.400	0.109
Stress	11	180	0.209	0.026
Thyrotoxicosis	11	28	0	0.006
Type 2 diabetes	7	300	0.510	0.102

Table 9.2 Summary of best factors for k nearest neighbours for optimum F2 score

9.1.5 What was found

For each method, a group of records that were ‘similar’ to a record of interest were found. These similar records were then checked for presence or absence of the condition of interest (the condition of interest having been removed prior to grouping) and the prevalence of the condition of interest within the similar group was counted. Should this prevalence be significantly higher than the prevalence within the wider records set (i.e.

the ‘population’ prevalence), the record of interest was predicted to be positive for the condition of interest. This analysis was repeated for all records in the test set, enabling a total count of true positives, true negatives, false positives and false negatives to be calculated and so subsequently calculations of sensitivity, specificity, positive and negative likelihood ratios and odds ratio. Results are shown in Table 8.4.

9.2 Results and implications for practice

A number of conditions were tested, the list of conditions being drawn primarily from the Practice Fusion list of the most prevalent conditions in the USA, together with some lower prevalence conditions to test performance with less common conditions.

It was found that for many conditions, both the clustering method and the nearest neighbours method could discriminate between records with a raised likelihood of containing the condition and those with a lesser likelihood of containing the condition. The nearest neighbours method was able to produce both significant positive likelihood ratios and significant negative likelihood ratios for more conditions. Both methods generally performed better with higher-prevalence conditions and less well with lower-prevalence conditions.

Table 9.3 (for clustering) and Table 9.4 (for nearest neighbours) show for which conditions the applied techniques were able to discriminate between raised and not raised/lowered likelihoods of the conditions. Those conditions in the lower right hand quadrant show a clear separation between raised and lowered likelihoods; those conditions in the upper left quadrant show no change in likelihoods after application of the grouping technique. It can be seen that the nearest neighbour technique produces more conditions (14 of 18) with both a significant positive likelihood ratio for those conditions predicted to be positive and a significant negative likelihood ratio for those conditions predicted to be negative than did the clustering method (7 of 18).

Results for each condition, reasons for those results, and the implications for clinical practice are discussed in section 9.3.

	Positive Likelihood Ratio not significant	**Positive Likelihood Ratio significant**
Negative Likelihood Ratio not significant	Acute sinusitis Bronchitis Gastroparesis Gastro-intestinal reflux disease Obesity Osteoarthritis Thyrotoxicosis Type 2 diabetes	Allergic rhinitis Autism Colon cancer
Negative Likelihood Ratio significant		Any cancer Asthma Breast cancer Eczema Gout Prostate cancer Stress

Table 9.3 Successes and failures of the clustering method

	Positive Likelihood Ratio not significant	**Positive Likelihood Ratio significant**
Negative Likelihood Ratio not significant	Autism Colon cancer Thyrotoxicosis	Gastroparesis
Negative Likelihood Ratio significant		Acute sinusitis Allergic rhinitis Any cancer Asthma Breast cancer Bronchitis Eczema Gastro-intestinal reflux disease Gout Obesity Osteoarthritis Prostate cancer Type 2 diabetes Stress

Table 9.4 Successes and failures of the nearest neighbours method

9.3 Implications for clinical practice

Early discovery of a condition can often improve the chances of successful treatment or of mitigation of the effects of the condition and many conditions can be detected through screening tests on asymptomatic individuals. Disadvantages of screening tests include risks of harm to the individual if the screening test is an invasive test; inconvenience and/or stress to the individual; and financial cost to the individual and/or the health care system. It is therefore advantageous to perform screening tests on subsets of the population selected to be at great risk of the condition being screened for. Increased detection of conditions via screening can be beneficial, but of most use if effective treatment is available once a condition has been detected. Similarly, identifying population sub-groups as being at raised risk for a condition is beneficial only if treatments are available to address that raised risk. For example, many patients are known to have cardiovascular risk already but are not effectively treated [277]. Results for likelihood ratios prediction, both positive (LR+) and negative (LR-) for each of the conditions investigated in this work are now presented, together with a brief discussion of the results and the implications for care.

9.3.1 Acute sinusitis

Clustering method: LR+: 1 (1, 1); LR-: -

Nearest neighbours method: LR+: 2.72 (2.47, 2.99); LR-: 0.55 (0.4, 0.60)

The test discriminates well between records that contain the condition and those that do not when using the nearest neighbours method: the positive likelihood ratio is significantly above 1, the negative likelihood ratio is significantly below 1. There is no discrimination between the groups when using the clustering method.

Advantage of early detection: avoids possible further complications, particularly in immunocompromised patients, e.g. patients with diabetes, HIV or other conditions.

Treatment available: Yes - generally nasal decongestants and/or antibiotics.

.

9.3.2 Allergic rhinitis

Clustering method: LR+: 2.16 (1.92, 2.43); LR-: 0.69 (0.5, 0.94)

Nearest neighbours method: LR+: 1.55 (1.44, 1.67); LR-: 0.65 (0.60, 0.72)

The test discriminates well between records that contain the condition and those that do not. The positive likelihood ratio is significantly above 1, the negative likelihood ratio is significantly below 1, for both the clustering method and the KNN method.

Advantage of early detection: There are advantages to early detection of allergic rhinitis, and allergic disease generally: “early diagnosis of allergic diseases makes specific immunotherapy more efficient. In this way, comorbidities can be avoided (e.g. bronchial asthma in patients with allergic rhinitis” [278].

Treatment available: Yes - antihistamines, decongestants, eye drops, nasal sprays, immunotherapy are all possible treatments for allergic rhinitis.

9.3.3 Any cancer:

Clustering: LR+: 4.90 (3.70, 6.50); LR-: 0.89 (0.71, 0.85)

Nearest neighbours method: LR+ 3.27 (2.64, 4.05); LR-: 0.72 (0.64, 0.80)

Advantage of early detection: Treating cancers while they are still small and/or before they have spread can improve the chances of successful treatment. The WHO note that "Early diagnosis is particularly relevant for cancers of the breast, cervix, mouth, larynx,

colon and rectum, and skin" [279]. As well as a greater chance of successful treatment, early treatment is likely to require a lower level of treatment and so fewer side effects.
Treatment available: Yes: radiotherapy, chemotherapy, surgery;

9.3.4 Asthma:

Clustering: LR+: 2.58 (1.97, 3.36); LR-: 0.91 (0.88, 0.95)
Nearest neighbours method: LR+: 2.70 (2.36, 3.09); LR-: 0.70 (0.65, 0.75)
Both the clustering method and the nearest neighbours method discriminate well between records that contain the condition and those that do not. The positive likelihood ratio is significantly above 1, the negative likelihood ratio is significantly below 1.
Advantage of early detection: There are advantages of earlier positive diagnosis, particularly in the young: “The diagnosis of asthma is of severely delayed, a fact which influences the prognosis and efficacy of therapeutic interventions… inhaled steroids seem to have a disease-modifying effect if started early enough” [280] . “Early detection and counselling is expected to reduce the prevalence of asthma symptoms and improve health-related quality of life at age 6 years” [281].
Treatment available: There is, as yet, no cure for asthma, but treatment is available and beneficial in helping control the condition. Treatments include use of inhalers or medication.

9.3.5 Autism spectrum disorder:

Clustering: LR+: 2.58 (6.32, 2.56, 15.59); LR-: 0.66 (0.39, 1.14)
Nearest neighbours method: LR+: 5.49 (0.91, 33.16); LR-: 0.86 (0.60, 1.23)
The clustering method was able to identify well a group of records with a higher likelihood of containing the condition autism spectrum disorder, although the group with reduced likelihood was not significantly lower in likelihood. The nearest neighbours method was not able to discriminate between the two groups.
Advantage of early detection: There are some advantages to early detection of autism spectrum disorder: “intervention for childhood autism based on applied behavior

analysis and delivered intensively … during the preschool period can bring about significant changes in children’s functioning” [282].

Treatment available: Although there is no ‘cure’ for autism spectrum disorder, specialist interventions can improve communication, educational and social development in those with the condition. Some conditions associated with autism spectrum disorder, such as epilepsy, ADHD, can be treated with appropriate medication.

9.3.6 Breast cancer:

Clustering: LR+: 7.24 (4.52, 11.59); LR-: 0.75 (0.63, 0.89)

Nearest neighbours method: LR+: 4.31 (3.20, 5.81); LR-: 0.56 (0.41, 0.75)

Both the clustering method and the nearest neighbours method discriminate well between records that contain the condition and those that do not. The positive likelihood ratio is significantly above 1, the negative likelihood ratio is significantly below 1.

Advantage of early detection: “tumour stage at diagnosis of breast cancer … greatly affects overall survival” [283]

Treatment available: Radiotherapy, chemotherapy, surgery, hormone therapy

9.3.7 Bronchitis:

Clustering: LR+: 0.95 (0.89, 1.01); LR-: 1.11 (0.71, 1.02)

Nearest neighbours method: LR+: 1.63 (1.49, 1.77); LR-: 0.68 (0.62, 0.74)

The test discriminates between records that contain the condition and those that do not when using the nearest neighbours method but is less successful when using the clustering method, where a group with significant positive likelihood ratio for the condition was identified, but a group with significant lowered likelihood ratio.

Advantage of early detection: There is little advantage to early treatment prior to condition becoming symptomatic, although there is a possible delay in condition onset or reduction in condition severity if early lifestyle changes are made – cigarette smoking is the most common cause of chronic bronchitis.

Treatment available: There is no direct treatment available for the cure of bronchitis, however a healthy diet, exercise and not smoking can all help to alleviate symptoms. Some medications are available to alleviate symptoms.

9.3.8 Colon cancer:

Clustering: LR+: 4.87 (1.38, 17.23); LR-: 0.85 (0.64, 1.12)

Nearest neighbours method: LR+: 6.56 (0.96, 44.95); LR-: 0.95 (0.83, 1.07)

The clustering method was able to identify groups with a raised likelihood of colon cancer, but not groups with a lowered likelihood. The nearest neighbours method was unable to identify groups with significantly raised or lowered likelihood ratio.

Advantage of early detection: 5-year survival if detected at Stage I is over 90 %; 5-year survival if not treated until Stage IV is only 11 % (National Cancer Institute's SEER database, quoted at https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html)

Treatment available: Early stage colon cancer may be treatable by surgery alone. Later stage colon cancer may require chemotherapy and/or radiotherapy in addition to surgery.

9.3.9 Eczema:

Clustering: LR+: 3.43 (2.84, 4.13); LR-: 0.68 (0.61, 0.76)

Nearest neighbours method: LR+: 2.98 (2.64, 3.37); LR-: 0.48 (0.40, 0.57)

Both the clustering method and the nearest neighbours method discriminate well between records that contain the condition and those that do not. The positive likelihood ratio is significantly above 1, the negative likelihood ratio is significantly below 1 for both methods.

Advantage of early detection: Early intervention in young people may help not only to alleviate the symptoms of eczema but also to prevent other allergic diseases such as asthma or allergic rhinitis [284].

Treatment available: "Atopic dermatitis is not curable… the treatment of atopic dermatitis aims to minimise the number of exacerbations of the disease, so-called flares [and] reduce the duration and degree of the flare" [285].

9.3.10 Gastro-oesophageal reflux disease:

Clustering: LR+: 1 (1, 1); LR-: -

Nearest neighbours method: LR+: 1.47 (1.37, 1.57); LR-: 0.57 (0.49, 0.67)

The clustering method could not discriminate between groups with raised likelihood and lowered likelihood of the condition. The nearest neighbours method did discriminate well between the two groups, with the positive likelihood ratio being significantly above 1 and the negative likelihood ratio significantly below 1.

Advantage of early detection: “Early detection and treatment of GERD in children may prevent, attenuate, or heal complications such as failure to thrive or feeding refusal as well as pulmonary, ear-nose-and-throat disorders, erosive esophagitis, and peptic stricture” [286]. “Early detection can help prevent minor heartburn from becoming a major health issue” [287].

Treatment available: Several treatment options are available, including medication, or surgery. Lifestyle changes may help to alleviate the condition, with weight loss, eating smaller but more frequent meals, decreasing stress and avoiding alcohol being beneficial.

9.3.11 Gastroparesis:

Clustering: LR+: 1 (0, 53.16); LR-: 1.02 (1.01, 1.02)

Nearest neighbours method: LR+: 5.06 (1.56, 16.44); LR-: 0.76 (0.47, 1.21)

Identification of groups with significant positive likelihood ratios or significant negative likelihood ratios was not demonstrated by the clustering method by was so with the nearest neighbours method.

Advantage of early detection: Gastroparesis can reduce the ability to control blood sugar in individuals with diabetes and so early identification of the presence of gastroparesis can help with blood sugar management.

Treatment available: Gastroparesis cannot usually be cured but some actions can help control the condition: eating smaller, more frequent meals, softer or liquid foods; chewing foods; drinking non-fizzy drinks with meals. Some medications may help alleviate the symptoms. Surgery may also be required in more serious cases.

9.3.12 Gout:

Clustering: LR+: 2.69 (1.94, 3.72); LR-: 0.78 (0.67, 0.90)

Nearest neighbours method: LR+: 2.69 (2.18, 3.32); LR-: 0.60 (0.48, 0.74)

Both the clustering method and the nearest neighbours method discriminate well between records that contain the condition and those that do not. The positive likelihood ratio is significantly above 1, the negative likelihood ratio is significantly below 1.

Advantage of early detection: reducing the level of uric acid prevents the formation of crystals in joints and tissues which trigger attacks of gout. Other benefits of reducing levels of uric acid include reducing the risk of liver disease.

Treatment available: anti-inflammatory drugs to alleviate the symptoms, with steroids for more severe cases. Keeping an affected joint cool can help reduce the symptoms. To prevent recurrence, medications are available. Lifestyle changes can help reduce the risk of recurrence.

9.3.13 Obesity:

Clustering: LR+: 1 (1, 1); LR-: -

Nearest neighbours method: LR+: 1.71 (1.58, 1.86); LR-: 0.65 (0.59, 0.71)

The clustering method could not discriminate between groups with raised likelihood and lowered likelihood of the condition. The nearest neighbours method did discriminate well between the two groups, with the positive likelihood ratio being significantly above 1 and the negative likelihood ratio significantly below 1.

Advantage of early detection: Can help to ensure that diet and exercise is sufficient for treatment, ensuring that more major treatment is not required and that risks of conditions associated with obesity are not raised.

Treatment available: diet and exercise. Medication is also available. In severe cases surgery may be appropriate.

9.3.14 Osteoarthritis

Clustering: LR+: - ; LR-: -

Nearest neighbours method: LR+: 3.51 (3.10, 3.97); LR-: 0.59 (0.54, 0.64)

The nearest neighbours method discriminates well between records that contain the condition and those that do not. The positive likelihood ratio is significantly above 1, the negative likelihood ratio is significantly below 1. The clustering method is ineffective here.

Advantage of early detection: There is an advantage to positive prediction: "osteoarthritis typically develop over a long period of time, offering a long window of time to potentially alter its course… it's etiology is multifactorial… with highly modifiable risk factors of mechanical overload, obesity and joint injury" and "Osteoarthritis ... currently lacks disease-modifying treatments" [288] So early detection of pre-osteoarthritis may lead to treatments and lifestyle changes that could delay or prevent full osteoarthritis developing.

Treatment available: There is, as yet, no cure for osteoarthritis. However, treatment once presence of condition is detected is available, including the use of medication to reduce pain and lifestyle changes such as weight loss and exercise. For more severe cases, surgery may be required.

.

9.3.15 Prostate cancer:

Clustering: LR+: 5.29 (3.39, 8.25); LR-: 0.75 (0.63, 0.89)

Nearest neighbours method: LR-: 3.42 (2.72, 4.30); LR-: 0.42 (0.28, 0.65)

Both the clustering method and the nearest neighbours method discriminate well between records that contain the condition and those that do not. The positive likelihood ratio is significantly above 1, the negative likelihood ratio is significantly below 1.

Advantage of early detection: it is unclear if the benefits of early detection outweigh the risks of detection and treatment

Treatment available: treatment depends on the stage of the cancer. Radiotherapy and/or chemotherapy treatments are available but slow-growing prostate cancers may simply be left under observation.

9.3.16 Stress:

Clustering: LR+: 2.97 (2.1, 4.21); LR-: 0.84 (0.76, 0.92)

Nearest neighbours method: LR+: 3.29 (2.48, 4.37); LR-: 0.74 (0.65, 0.85)

Both the clustering method and the nearest neighbours method discriminate well between records that contain the condition and those that do not. The positive likelihood ratio is significantly above 1, the negative likelihood ratio is significantly below 1.

Advantage of early detection: A number of conditions are related to stress, including cardiovascular disease, asthma and Alzheimer’s disease. Early intervention to reduce stress may reduce the risk of other conditions.

Treatment available: Treatments available for stress include medication, cognitive behaviour therapy and other talking therapies, alternative therapies

9.3.17 Thyrotoxicosis:

Clustering: LR+: - ; LR-: 1.03 (1.03, 1.04)

Nearest neighbours method: 1 (0, 94.15); LR-: 1.01 (1.00, 1.01)

Neither the clustering method nor the nearest neighbours method were able to identify groups with a significant positive likelihood for thyrotoxicosis or a significant negative likelihood. This is likely due to the low prevalence of the condition (0.6 % in the composite data set) but may be due to factors intrinsic to the condition.

Advantage of early detection: There are advantages to early diagnosis of thyrotoxicosis: “treatment of subclinical hyperthyroidism may decrease the risk of atrial fibrillation and may decrease the risk of low bone density in menopausal women” [289].

Treatment once presence of condition is detected: available. Medication can suppress thyroid function to normal levels. Radioiodine treatment can destroy some of the thyroid function. Surgery can remove some or all of the thyroid. Any shortfall in function following over-treatment can be compensated by medication.

9.3.18 Type 2 diabetes:

Clustering: LR+: 1 (1, 1); LR-: -

Nearest neighbours method: LR+: 2.33 (2.19, 2.48); LR-: 0.26 (0.21, 0.32)

The test discriminates well between records that contain the condition and those that do not. The positive likelihood ratio is significantly above 1, the negative likelihood ratio is significantly below 1.

Advantage of early detection: “the burden of the disease could be further reduced with early intervention to address prediabetes, as mounting evidence suggests that this approach could prevent, or at least delay, progression to overt diabetes. However, most people do not benefit from this, as prediabetes is largely underdiagnosed” [290].

Treatment available: Medication and dietary & exercise changes can control type 2 diabetes. Pre-diabetes can be control by lifestyle changes including diet, exercise and smoking cessation

9.4 Comparison to existing work

The produced composite data set had condition prevalences similar to those reported in the literature for the developed world and had age and gender demographics in accordance to those figures published by the UK and USA governments. The age and gender demographics in the composite data set were also no different to those in the source data sets. The objective for producing a composite data set and the intention for its use reflected the ambitions suggested by Celi et al [197] for the analysis of data sets from multiple sources. Other techniques for predicting likelihoods of future conditions are described in the literature, e.g. [56] for prediction of diabetes risk or [291] for prediction of cardiovascular disease risk, but in each case the risk calculation is based on prior knowledge of risk factors, whereas the techniques described here do not rely on any prior knowledge (save the code grouping by common ancestors in the CTV3 hierarchy) and have the potential to suggest further risk factors.

The code mappings generated for ICD-9-CM to CTV3 produced a success rate similar to that reported by Nadkarni and Darer [189] for their manual (though software-assisted) mapping of ICD-9-CM to SNOMED CT.

9.5 Implications for research

9.5.1 Research privacy

If it can be demonstrated that there is an optimum level or levels of the CTV3 hierarchy to perform work such as that described here, it can be hypothesised that useful data for research can be released with coding at a less granular level than in the originally recorded data but can still produce useful research, audit or predictive results. Decreasing granularity can be a way of increasing privacy – it may increase privacy by increasing the k-number in k-anonymity (see [292] for a discussion of k-anonymity) - but by decreasing the granularity it may decrease the research potential of the data should fine detail be required.

9.5.2 Computer processing requirements

Use of CTV3 codes at a higher level of the hierarchy may reduce RAM requirements for processing and reduce processing time, in essence by dimensionality reduction. The number of unique event codes on which the algorithms are run is reduced by aggregating similar events (as determined by common ancestors in the CTV3 hierarchy) into single parent codes.

9.5.3 Discovery of risk factors

Although the methods described cannot deduce the aetiology of medical conditions it can provide pointers towards risk factors by investigation of the most common event codes within clusters or near neighbour groups with raised condition prevalence.

9.6 Study limitations and future work

9.6.1 Small data set size

One limitation of the work was the relatively small data set size, requiring development of techniques to compensate for this. The primary technique employed was to work with less granular levels of the CTV3 hierarchy, trading a decrease in granularity for an increase in nearest neighbour matching.. Conditions with low prevalence had, of course, very few occurrences in the data set and so did not perform as well as conditions with greater prevalence. Although it was assumed that this was due to low prevalence, it is possible that the performance was due to the nature of the condition. Analysis using a large quantity of data, or data that is skewed to contain a quantity of the rarer conditions higher than their general prevalence would usually give, would allow for this to be investigated further.

9.6.2 Codelists not verified

Codelists may not be complete: any codes that should be in the list but aren’t may increase the likelihood ratios. This was mitigated by CART analysis which suggested codes that divided the data set into those records that were more likely to have the condition and those that were less likely, however this method was not foolproof and could only suggest codes that were in the composite data set and not in the complete CTV3 code set. Codelists have not been clinically validated for accuracy.

9.6.3 Time period of adjusted likelihoods

Records were predicted to have an increased or not increased likelihood of particular conditions, but there was no time period set on this prediction. So rather than predicting an increased likelihood within 1 year, 5 years or 10 years, a lifetime increased likelihood was presented. This may not be useful for some conditions.

9.6.4 Weighting of events

There was no weighting of event codes to allow for severity - a condition or symptom was either present or absent. Equally, no allowance was made for the age of an event, i.e. whether it was a recent event or occurred further into the past. However, without a priori knowledge of which events take have effects which increase with time and which events have effects which lessen with time it is difficult to decide how to allow for this.

9.6.5 Coding system used

The choice of coding system on which to converge was Clinical Terms Version 3. During the course of this work, this coding system was deprecated by the UK Department of Health, with the NHS mandated to use SNOMED CT in its place, the conversion to take place by April 2010. SNOMED CT is a more complex system and it is not clear how well the techniques applied here would work under SNOMED CT.

9.6.6 Coincidence of predictors and conditions

The technique has been tested by removing conditions of interest from records and making predictions based on the remaining events in the record. However it is not clear whether individuals who have yet to be diagnosed with a condition of interest will have yet had the events that are predictive of those conditions, i.e. the symptoms or diseases which match them with records that contain the condition of interest.

9.6.7 Future work

There are a number of areas for further research, many of which seek to address the study limitations. The analysis can be repeated using the existing methods but translating event codes to SNOMED CT and comparing results to those described here. The slow run time of the methods empoyed was disappointing and limited development of the techniques. This would have implications for use in clinical practice: currently the system takes around 30 minutes to run on a sample of 5,000 records. This is likely to be too long for opportunistic likelihood predictions for an individual arriving in a clinic, although acceptable for use in identification of individuals appropriate for screening or invitation to consultation.

Other machine learning techniques can be applied to the data set, with comparison of the results and the running times to the analysis performed here. These other techniques include neural networks, decision trees, correspondence analysis and others.

There may be potential for combining the results from the work described here with results from genomic analysis, combining nature with nurture: a person's genes can predict the risks of various conditions, but for some conditions these risks can be modified by events that occur later in life. For example, an individual may genomically have a raised risk of diabetes, but their lifestyle can also have an effect on this risk. Other factors can be brought in to the analysis: geographic factors such as index of deprivation, occupation history and ethnicity. Allowance must also be made for accidents and random factors.

9.7 Contributions to knowledge

9.7.1 Code mapping

A system was developed to generate code mappings between two coding systems, indirectly via a third coding system. This addressed a gap in the available code mappings, a mapping from ICD-9-CM to CTV3. Although developed to aid in the generation of that specific code mapping, the system is generalizable.

9.7.2 Codelists

Codelists are sets of codes from clinical coding systems that indicate the presence of particular conditions. Very few existing codelists were found and so a set of codelists defining each condition evaluated in this work was assembled. These codelists were of CTV3 codes and were intended to include the entire CTV3 code set, without restriction to the CTV3 codes in the data used **using data volunteered by** in this work. The codelists produced for this work are shown in the Appendix 2.

9.7.3 Analysis techniques

Using existing machine learning methods a number of techniques were developed in order to optimise the effectiveness of these methods. These included, in particular, analysing and utilising the hierarchical structure of the CTV3 coding system, and assignation of significance codes to each CTV3 event code. Analysis of the coding system hierarchy, and subsequent testing, derived an effective level of the CTV3 hierarchy at which to operate with the methods and data employed in this work, allowing for aggregation of similar codes to increase the chances of finding matching records but retaining sufficient granularity to distinguish between dissimilar records. It is believed that the values selected for the candidate optimum levels of the CTV3 hierarchy (i.e. levels 7 and 11) will remain generally true, since they are the result of inspection and analysis of the CTV3 hierarchy (see section 7.3.1); however the final choice of level must be determined by performing test runs on the condition being investigated – note that there are many more conditions than the relatively small number analysed in this work. Values for the optimum number of near neighbours or for the optimum number of clusters were produced independently for each condition.

9.7.4 Other contributions

- Peer reviewer for: Leo Anthony Celi, Andrew J Zimolzak, David J Stone. Dynamic Clinical Data Mining: Search Engine-Based Decision Support. JMIR Med Inform 2014;2(1):e13) doi:10.2196/medinform.3110 [197]
- Presentation: Jonathan Turner, Peter Weller. "Modulation of Medical Condition Likelihood by Patient History Similarity." Medicine 2.0'14 Europe (Malaga, Spain, 10th October 2014) (& also chair of this session)
- Presentation: Turner J, Istad TS, Olerud HM, Flatabø S, Liland A, Ali W, Kjærheim K. "EPI-CT: International Epidemiological Paediatric CT Study. Data extraction and patient privacy protection: The approach in Norway." At IPEM Data: storage, management, generation and legislation, London, 16th April 2013

9.8 Summary

Examples from the literature have been given, showing a number of trends. Successful use of clinical records data for other purposes use well-coded data or utilise natural language processing of free text fields; it is a challenge to use coded data from a typical electronic medical record system and a greater challenge to combine data from several systems.

Data from electronic medical records systems has been utilised both in near-real-time (e.g. for detection of disease outbreaks) and retrospectively (e.g. for selection of patients suitable for clinical trials).

Some work by others has been done on general predictions of future health events based on lifetime clinical histories, using a variety of techniques, and development in this areas has continued in recent years (see section 2.7 for a review of some of the particular methods and areas where this work has been done). However, the work presented here is suggested as a useful starting point for a relatively explainable and understandable system of prediction of raised likelihood of conditions, which has performed well when benchmarked against a well–established method, logistic regression.

10 REFERENCES

References

1. Berner, E.S. and J. Moss, *Informatics Challenges for the Impending Patient Information Explosion.* Journal of the American Medical Informatics Association, 2005. **12**: p. 614-617.
2. Lipkin, M. and J. Hardy, *Differential Diagnosis of Hematologic Diseases Aided by Mechanical Correlation of Data.* Science, 1957. **1257**(3247): p. 551-552.
3. Lipkin, M. and J. Hardy, *Mechanical Correlation of Data in Differential Diagnosis of Hematological Diseases.* JAMA, 1958. **166**(2): p. 113-125.
4. Lipkin, M., *Correlation of data with a digital computer in the differential diagnosis of haematological diseases.* IRE Transactions on Medical Electronics, 1960. **ME-7**(4): p. 243-246.
5. Warner, H., et al., *A Mathematical Approach to Medical Diagnosis: Application to Congenital Heart Disease.* JAMA, 1961. **177**: p. 177-183.
6. Ledley, R. and L. Lusted, *The Use of Electronic Computers to Aid in Medical Diagnosis.* Proc IRE, 1959: p. 1970-1977.
7. Spencer, W. and C. Valbonna, *Application of Computers in Clinical Practice.* JAMA, 1965. **191**(11): p. 121-125.
8. Weed, L., *Medical Records, Patient Care and Medical Education.* International Journal of Medical Sciences, 1964: p. 271-282.
9. Schultz, J., *A History of the PROMIS Technology: An Effective Human Interface*, in *A history of personal workstations*. 1988, ACM: New York, NY. p. 439-488.
10. American Medical Informatics Association *A Taxonomy of Secondary Uses and Re-Uses of Healthcare Data*. 2007.
11. Aickin, M., *Patient-Centered Research from Electronic Medical Records.* The Permanente Journal, 2011. **15**(4): p. 89-91.

12. Prokosch, H. and T. Ganslandt, *Reusing the Electronic Medical Record for Clinical Research.* Methods of Information in Medicine, 2009. **48**(1): p. 38-44.

13. D'Avolio, L., et al., *Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). J Am Med Inform Assoc 17:375-382.* Journal of the American Medical Informatics Association, 2010. **17**: p. 375-382.

14. Kukafka, R., et al., *Redesigning electronic health record systems to support public health.* Journal of Biomedical Informatics, 2007. **40**(4): p. 398-409.

15. Judd, R. and R. Kim, *Electronic Medical Records and Medical Research Databases – Can They Be Synonymous?* Business Briefing: US Cardiology, 2006: p. 13-139.

16. Kim, D., S. Labkoff, and S. Holliday, *Opportunities for Electronic Health Record Data to Support Business Functions in the Pharmaceutical Industry – A Case Study from Pfizer. Inc.* Journal of the American Medical Informatics Association, 2008. **15**(5): p. 581-584.

17. Safran, C., et al., *Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper.* Journal of the American Medical Informatics Association, 2007. **14**(1): p. 1-19.

18. Bloomrosen, M. and D. Detmer, *Advancing the Framework: Use of Health Data – A Report of a Working Conference of the American Medical Informatics Association.* Journal of the American Medical Informatics Association, 2008. **15**(6): p. 715-722.

19. Department of Health, *Summary of Responses to the Consultation on the Additional Uses of Patient Data*. 2008: London.

20. Dean, B., et al., *Use of Electronic Medical Records for Health Outcomes Research: A Literature Review.* Medical Care Research and Review, 2009. **66**: p. 611-638.

21. Pearson, J., C. Brownstein, and J. Brownstein, *Potential for Electronic Health Records and Online Social Networking to Redefine Medical Research.* Clinical Chemistry, 2011. **57**(2): p. 196-204.

22. Walton, J., et al., *Consequences for research if use of anonymised patient data breaches confidentiality.* BMJ, 1999. **319**(7221): p. 1366.

23. Grove, W., et al., *Clinical versus mechanical prediction: a meta-analysis.* Psychological Assessment, 2000. **12**(1): p. 19-30.

24. Elkin, P., et al., *Secondary use of clinical data.* Studies in health technology and informatics, 2010. **155**: p. 14-29.

25. Jones, D., et al., *Characteristics of personal health records: findings of the Medical Library Association/National Library of Medicine Joint Electronic Personal Health Record Task Force.* Journal of the Medical Library Association, 2010. **98**(3): p. 243-249.

26. Kaelber, D., et al., *A Research Agenda for Personal Health Records (PHRs).* Journal of the American Medical Informatics Association, 2008. **15**: p. 729-736.

27. Undem, T., *Consumers and Health Information Technology: A National Survey*. 2010.

28. The Markle Foundation, *Markle Survey on Health in a Networked Life 2010*. 2011.

29. Zulman, D., et al., *Patient Interest in Sharing Personal Health Record Information: A Web-Based Survey. Annals of Internal Medicine.* Annals of Internal Medicine, 2011. **155**: p. 805-811.
30. Harlow, S. and M. Linet, *Agreement Between Questionnaire Data and Medical Records: The Evidence for Accuracy of Recall.* American Journal of Epidemiology, 1989. **129**: p. 233-248.
31. Van Deursen, T., P. Koster, and M. Petkovi. *Reliable Personal Health Records*. in *eHealth Beyond the Horizon - Get IT There - MIE2008 - The XXIst International Congress on the European Federation for Medical Informatics*. 2008. IOS Press.
32. Boxwala, A., et al., *Using statistical and machine learning to help institutions detect suspicious access to electronic health records.* Journal of the American Medical Informatics Association, 2012. **18**: p. 498-505.
33. Stein, H.D., et al., *Exploring the Degree of Concordance of Coded and Textual Data in Answering Clinical Queries from a Clinical Data Repository - JAMIA 2000.* Journal of the American Medical Informatics Association, 2000. **7**(1): p. 42-54.
34. Turchin, A., et al., *Comparison of Information Content of Structured and Narrative Text Data Sources on the Example of Medical Intensification.* Journal of the American Medical Informatics Association, 2009. **16**: p. 362-370.
35. Friedman, C., et al., *Automated Encoding of Clinical Documents Based on Natural Language Processing.* Journal of the American Medical Informatics Association, 2004. **11**: p. 392-402.
36. Turchin, A., et al., *Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information form the Text of Physician Notes. J Am Med Inform Assoc.* Journal of the American Medical Informatics Association, 2006. **13**: p. 691-695.
37. Thiru, K., A. Hassey, and F. Sullivan, *Systematic review of scope and quality of electronic patient record data in primary care.* BMJ, 2003. **17**(326).
38. Chan, K., J. Fowles, and J. Weiner, *Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature.* Medical Care Research and Review, 2010. **67**(5): p. 503-527.
39. Liaw, S., et al., *Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature.* International Journal of Medical Informatics, 2013. **82**(1): p. 10-24.
40. Health Protection Agency, *List of notifiable diseases*. 2012.
41. Gesteland, P., et al., *Automated Syndromic Surveillance for the 2002 Winter Olympics.* Journal of the American Medical Informatics Association, 2003. **10**(6): p. 547-554.
42. Klompas, M., et al., *Electronic Medical Record Support for Public Health (ESP): Automated Detection and Reporting of Statutory Notifiable Diseases to Public Health Authorities.* Advances in Disease Surveillance, 2007. **3**(3).
43. Klompas, M., et al., *Automated Identification of Acute Hepatities B Using Electronic Medical Record Data to Facilitate Public Health Surveillance.* PLoS ONE, 2008. **3**(7): p. e2626.

44. Calderwood, M., et al., *Real-Time Surveillance for Tuberculosis Using Electronic Health Record Data from an Ambulatory Practice in Eastern Massachusetts.* Public Health Reports, 2010. **125**: p. 843-850.
45. Hripcsak, G., et al., *Syndromic Surveillance Using Ambulatory Electronic Health Records. J Am Med Inform Assoc 16:354-361.* Journal of the American Medical Informatics Association, 2009. **16**: p. 354-361.
46. Buckeridge, D., et al., *Understanding Detection Performance in Public Health Surveillance: Modeling Aberrancy-Detection Algorithms.* Journal of the American Medical Informatics Association, 2008. **15**: p. 760-769.
47. Dailey, L., R. Watkins, and A. Plant, *Timeliness of Data Sources Used for Infuenza Surveillance. J Am Med Inform Assoc 14:626-631.* Journal of the American Medical Informatics Association, 2007. **14**: p. 626-631.
48. NHS/DHSS Steering Group on Health Services Information and E.E. Korner, *Steering Group on Health Services Information: A report on the Collection and Use of Information about Hospital Clinical Activity in the NHS (First Report)*, E. Korner, Editor. 1982: London.
49. Owen, R., et al., *Use of Electronic Medical Record Data for Quality Improvement in Schizophrenia Treatment.* Journal of the American Medical Informatics Association, 2004. **11**(5): p. 351-357.
50. Voorham, J. and P. Denig, *Computerized Extraction of Information on the Quality of Diabetes Care from Free Text in Electronic Patient Records of General Practitioners.* Journal of the American Medical Informatics Association, 2007. **14**(3): p. 349-353.
51. Pakhomov, S., et al., *Automatic Classification of Foot Examination Findings Using Clinical Notes and Machine Learning.* Journal of the American Medical Informatics Association, 2008. **15**: p. 198-202.
52. Chiang, J.-H., J.-W. Lin, and C.-W. Yang, *Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (medLEE).* Journal of the American Medical Informatics Association, 2010. **17**: p. 245-252.
53. Chan, K., et al., *EHR-Based Care Coordination Performance Measures in Ambulatory Care*. 2011, Commonwealth Fund.
54. Lee, W.-N., S. Tu, and A. Das. *Extracting Cancer Quality Indicators from Electronic Medical Records: Evaluation of an Ontology-Based Virtual Medical Record Approach*. in *AMIA Annual Symposium Proceedings*. 2009.
55. Himes, B., et al., *Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records.* Journal of the American Medical Informatics Association, 2009. **16**: p. 371-379.
56. Hippisley-Cox, J., et al., *Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore.* BMJ, 2009. **338**: p. b880.
57. Sebastiani, P., et al., *A Network Model to Predict The Risk of Death in Sickle Cell Disease.* Blood, 2007. **110**: p. 2727-2735.
58. Musen, M., Y. Shahar, and E. Shortliffe, *Clinical Decision Support Systems*, in *Medical Informatics: Computer Applications in Health Care and Bioinformatics*, E. Shortliffe and L. Perrault, Editors. 2001, Springer: New York.

59. Riesbeck, C. and R. Schank, *Inside Case-Based Reasoning*. 1989, Hillsdale NJ.: Erlbaum.

60. Kolodner, J., *Case-based reasoning*. 1993, San Francisco, CA: Morgan Kaufmann Publishers Inc. 668.

61. Schmidt, R., et al., *Cased-Based Reasoning for medical knowledge-based systems.* International Journal of Medical Informatics, 2001. **64**: p. 355-367.

62. Yusof, M. and C. Buckingham. *Medical Case-based Reasoning: A Review of Retrieving, Matching and Adaptation Processes in Recent Systems*. 2009. ACTA Press.

63. Bichindaritz, I., C. Marling, and C.-b.r.i.t.h.s.W.s.n.A.I.i.M. 36:127-135, *Case-based reasoning in the health sciences: What's next?* Artifical Intelligence in Medicine, 2006. **36**: p. 127-135.

64. Bichindaritz, I. and S. Montani, *Advances in case-based reasoning in the health sciences.* Artifical Intelligence in Medicine, 2011. **51**(2): p. 75-79.

65. Ting, S., et al., *RACER: Rule-Associated CasE-based Reasoning for supporting General Practitioners in prescription making.* Expert Systems with Applications, 2010. **37**: p. 8079-8089.

66. Kahn, C.J. and G. Anderson, *Case-Based Reasoning and Imaging Procedure Selection.* Investigative Radiology, 1994. **29**(6): p. 643-647.

67. Marling, C. and P. Whitehouse, *Case-Based Reasoning in the Care of Alzheimer's Disease Patients*, in *ICCBR 2001*, D. Aha and I. Watson, Editors. 2001. p. 702-715.

68. Chuang, C.-L., *Case-based reasoning support for liver disease diagnosis.* Artificial Intelligence in Medicine, 2011. **53**: p. 15-23.

69. O'Sullivan, D., et al., *Mobile case-based decision support for intelligent patient knowledge management.* Health Informatics Journal, 2007. **13**(179-193).

70. Van den Branden, M., et al., *Integrating case-based reasoning with an electronic patient record system.* Artificial Intelligence in Medicine, 2011. **51**: p. 117-123.

71. Honigman, B., et al., *Using Computerized Data to Identify Adverse Drug Events in Outpatients.* Journal of the American Medical Informatics Association, 2001. **8**: p. 254-266.

72. Wang, X., et al., *Active Computerized Pharmacovigilance Using Natural Language Processing Statistics and Electronic Health Records: A Feasibility Study.* Journal of the American Medical Informatics Association, 2009. **16**: p. 328-337.

73. Nadkarni, P., *Drug safety surveillance using de-identified EMR and claims data: issues and challenges.* Journal of the American Medical Informatics Association, 2010. **17**: p. 671-674.

74. Savova, G., et al., *Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record.* Journal of the American Medical Informatics Association, 2011.

75. Wilke, R., et al., *Use of an Electronic Medical Record for the Identification of Research Subjects with Diabetes Mellitus.* Clinical Medicine and Research, 2007. **5**(1): p. 1-7.

76. Clark, C., et al., *Identifying Smokers with a Medical Extraction System.* Journal of the American Medical Informatics Association, 2008. **15**: p. 36-39.

77. Yamamoto, K., et al., *A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research.* BMJ Open, 2012. **2**: p. e001622.

78. McCormick, T., C. Rudin, and D. Madigan, *Bayesian Hierarchical Rule Modeling For Predicting Medical Conditions.* Annals of Applied Statistics, 2012. **6**: p. 652-668.

79. Doddi, S., et al., *Discovery of association rules in medical data.* Medical Informatics and the Internet in Medicine, 2001. **26**(1): p. 25-33.

80. Tan, P., M. Steinbach, and V. Kumar, *Introduction to Data Mining*. 2005: Addison-Wesley.

81. Tsui, F., et al., *Technical Description of RODS: A Real-time Public Health Surveillance System.* Journal of the American Medical Informatics Association, 2003. **10**(5): p. 399-408.

82. Jiang, X. and G. Cooper, *A Bayesian spatio-temporal method for disease outbreak detection.* Journal of the American Medical Informatics Association. **17**: p. 462-471.

83. Que, J. and F.-C. Tsui, *Rank-based spatial clustering: an algorithm for rapid outbreak detection.* Journal of the American Medical Informatics Association, 2011. **18**: p. 218:224.

84. Creighton, C. and S. Hanash, *Mining gene expression databases for association rules.* Bioinformatics, 2003. **19**(1): p. 79-86.

85. Verduijn, M., et al., *Prognostic Bayesian networks I: Rationale, learning procedure and clinical use.* Journal of Biomedical Informatics, 2007. **40**: p. 609-618.

86. Verduijn, M., et al., *Prognostic Bayesian networks II: An application in the domain of cardiac surgery.* Journal of Biomedical Informatics, 2007. **2007**(40): p. 619-630.

87. Reynolds, G., A. Peet, and T. Arvanitis, *Generating prior probabilities for classifiers of brain tumours using belief networks. BMC Med Inform Decis Mak 7:27.* BMC Medical Informatics and Decision Making, 2007. **7**(27).

88. van Gerven, M., B. Taal, and P. Lucas, *Dynamic Bayesian networks as prognostic models for clinical patient management.* Journal of Biomedical Informatics, 2008. **41**: p. 515-529.

89. Sakai, S., et al., *Accuracy in the Diagnostic Prediction of Acute Appendicitis Based on the Bayesian Network Model.* Methods of Information in Medicine, 2007. **46**: p. 723-726.

90. National Institute of Health. *The Hippocratic Oath*. 2002 [cited 2013 13th January 2013]; Available from: http://www.nlm.nih.gov/hmd/greek/greek_oath.html.

91. General Medical Council, *Confidentiality*. 2009.

92. Kalra, D., et al., *Confidentiality of personal health information used for research.* BMJ, 2006. **333**: p. 196-198.

93. The European Parliament and The Council of The European Union, *Directive 95/46/EC of the European Parliament and of the Council of Europe of 24 October 1995, on the protection of individuals with regard to the processing of personal data and on the free movement of such data*. 1995, Official Journal of the European Communities L281. p. 0031-0050.

94. *Data Protection Act 1998*. 1998.
95. *Data Protection Act 2018*. 2018: United Kingdom.
96. Richard, T., *Court sanctions use of anonymised patient data.* BMJ, 2000. **320**: p. 77.
97. Bourke, J. and S. Wessely, *Confidentiality.* BMJ, 2008. **336**: p. 888-891.
98. The Caldicott Committee, *Report on the Review of Patient-Identifiable Information*. 1997.
99. Caldicott, F., *Information: To share or not to share? The Information Governance Review*. 2013.
100. *Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive)*. 2016, Official Journal of The European Union: EU. p. 1-88.
101. Cornock, M., *General Data Protection Regulation (GDPR) and implications for research.* Maturitas, 2018. **111**: p. A1-A2.
102. Chico, V., *The impact of the General Data Protection Regulation on health research.* Br Med Bull, 2018. **128**(1): p. 109-118.
103. Mourby M, M.E., Elliot M, Gowans H, Wallace SE, Bell J, Smith H, Aidinlis S, Kaye J, *Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK.* Computer Law & Security Review, 2018. **34**: p. 222-233.
104. Olimid, A.P., L.M. Rogozea, and D.A. Olimid, *Ethical approach to the genetic, biometric and health data protection and processing in the new EU General Data Protection Regulation (2018).* Rom J Morphol Embryol, 2018. **59**(2): p. 631-636.
105. Schaar, K., *What is important for data protection in science in the future? General and specific changes in data protection for scientific use resulting from the EU General Data Protection Regulation*. 2016, Working Paper of the German Data Forum (RatSWD). p. 1-13.
106. Wachter S, M.B., Floridi L, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.* International Data Privacy Law, 2017. **0**(0).
107. Goodman B, F.S., *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation".* AI Magazine, 2017. **38**: p. 50-57.
108. Selbst AD, P.J., *Meaningful information and the right to explanation.* International Data Privacy Law, 2017. **7**(4): p. 233-242.
109. *United Kingdom: Human Rights Act 1998 [United Kingdom of Great Britain and Northern Ireland]*. 1998 [cited 2013; Available from: http://www.refworld.org/docid/3ae6b5a7a.html.
110. *Access to Health Records Act 1990*. 1990: UK.
111. *Computer Misuse Act*. 1990.
112. *Freedom of Information Act*. 2000.
113. *Regulation of Investigatory Powers Act*. 2000.
114. *Copyright, Designs and Patents Act 1988*. 1988.
115. Care, N.I.G.B.f.H.a.S., *The Care Record Guarantee*. 2011.
116. DH/IPU/Patient Confidentiality, *NHS Confidentiality Code of Practice*. 2003.
117. Gowing, W., *Pseudonymisation Implementation Project (PIP)*. 2010.

118. Parliamentary Office for Science and Technology, *Data Protection & Medical Research*. 2005.
119. Information Commissioner's Office, *Anonymisation: managing data protection risk code of practice*. 2012.
120. Department of Health and Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. 2012.
121. British Medical Association, *Confidentiality as part of a bigger picture - a discussion paper from the BMA*. 2005, London: Nritish Medical Association.
122. Pommerening, K. and M. Reng, *Secondary Use of the EHR via Pseudonymisation.* Studies in health technology and informatics, 2004. **103**: p. 441-446.
123. *Health Insurance Portability and Accountability Act*. 1996.
124. Dorr, D., et al., *Assessing the Difficulty and Time Cost of De-identification in Clinical Narratives.* Methods of Information in Medicine, 2006. **45**: p. 246-252.
125. Beckwith, B., et al., *Development and evaluation of an open source software tool for deidentification of pathology reports.* BMC Medical Informatics and Decision Making, 2006. **6**(12).
126. Neamatullah, et al., *Automated de-identification of free-text medical records.* BMC Medical Informatics and Decision Making, 2008. **8**(32).
127. National Health Service Information Authority in conjunction with The Consumer's Association and Health Which?, *Share with care! Peoples' views on consent and confidentiality of patient information*. 2002.
128. Buckley, B., A. Murphy, and A. MacFarlane, *Public attitudes to the use in research of personal health information from general practitioners' records: a survey of the Irish general public.* Journal of Medical Ethics, 2011. **37**: p. 50-55.
129. Whiddett, R., et al., *Patients' attitudes towards sharing their health information.* International Journal of Medical Informatics, 2006. **75**: p. 530-541.
130. Parkin, L. and C. Paul, *Public good, personal privacy: a citizen's deliberation about using medical information for pharmacoepidemiological research.* Journal of Epidemiology & Community Health, 2011. **65**: p. 150-156.
131. Page, S. and I. Mitchell, *Patients' opinions on privacy, consent and the disclosure of health information for medical research.* Chronic Diseases in Canada, 2006. **27**(2): p. 60-67.
132. Kelleher, J.D., B. Mac Namee, and A. D'Arcy, *Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies*. 2015, Cambridge, Massachusetts: The MIT Press. xxii, 595 pages.
133. Cousins MS, S.L., Bander JA, *An Introduction to Predictive Modeling for Disease Management Risk Stratification.* Disease Management, 2002. **5**.
134. Steyerberg, E.W., *Clinical prediction models : a practical approach to development, validation, and updating*. Statistics for biology and health. 2009, New York, NY: Springer. xxviii, 497 p.
135. Rahe, R.H., J.L. Mahan, Jr., and R.J. Arthur, *Prediction of near-future health change from subjects' preceding life changes.* J Psychosom Res, 1970. **14**(4): p. 401-6.

136. Islam, M.S., et al., *A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining.* Healthcare (Basel), 2018. **6**(2).

137. Schneider, A., G. Hommel, and M. Blettner, *Linear regression analysis: part 14 of a series on evaluation of scientific publications.* Dtsch Arztebl Int, 2010. **107**(44): p. 776-82.

138. Flemons, W.W., et al., *Likelihood ratios for a sleep apnea clinical prediction rule.* Am J Respir Crit Care Med, 1994. **150**(5 Pt 1): p. 1279-85.

139. Sanchez-Santos, M.T., et al., *Development and validation of a clinical prediction model for patient-reported pain and function after primary total knee replacement surgery.* Sci Rep, 2018. **8**(1): p. 3381.

140. Combes C, K.F., Chaabane S, *Predicting Hospital Length of Stay Using Regression Models: Appliation to Emergency Department10ème Conférence* in *Francophone de Modélisation, Optimisation et Simulation- MOSIM'14*. 2014: Nancy, France.

141. Devin, C.J., et al., *A predictive model and nomogram for predicting return to work at 3 months after cervical spine surgery: an analysis from the Quality Outcomes Database.* Neurosurgical Focus, 2018. **45**(5): p. E9-E9.

142. Park JH, C.H., Kim JH, Wall M, Stern Y, Lim H, Yoo S, Kim H-S, Cha J, *Electronic Health Records Based Prediction of Future Incidence of Alzheimer's Disease Using Machine Learning*. 2019.

143. Kim KS, C.K., Moon CS, Mun CW, *Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions.* Current Applied Physics, 201. **11**: p. 740-745.

144. D'Agostino, R.B., Sr., et al., *General cardiovascular risk profile for use in primary care: the Framingham Heart Study.* Circulation, 2008. **117**(6): p. 743-53.

145. Chhatwal, J., et al., *A logistic regression model based on the national mammography database format to aid breast cancer diagnosis.* AJR Am J Roentgenol, 2009. **192**(4): p. 1117-27.

146. Singal, A.G., et al., *An automated model using electronic medical record data identifies patients with cirrhosis at high risk for readmission.* Clin Gastroenterol Hepatol, 2013. **11**(10): p. 1335-1341 e1.

147. Jacobs, I., et al., *A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer.* BJOG: An International Journal of Obstetrics & Gynaecology, 1990. **97**(10): p. 922-929.

148. Christodoulou, E., et al., *A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models.* J Clin Epidemiol, 2019. **110**: p. 12-22.

149. Cohen T, W.D., *Geometric Representations in Biomedical Informatics: Applications in Automated Text Analysi,* in *Methods in Biomedical Informatics,* S. N, Editor. 2014, Academic Press. p. 99-139.

150. Tayeb S, P.M., Sun J, Hall K, Chang A, Li J, Song C, *Toward predicting medical conditions using k-nearest neighbors* in *IEEE International Conference on Big Data*. 2017.

151. Zhu, M., et al., *The K-nearest neighbor algorithm predicted rehabilitation potential better than current Clinical Assessment Protocol.* J Clin Epidemiol, 2007. **60**(10): p. 1015-21.

152. Polat K, S.S., Gunes S, *Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing.* Expert Systems with Applications, 2007: p. 625–663.

153. Shouman M, T.T., Stocker R, *Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients.* International Journal of Information and Education Technology, 2012. **2**.

154. Enriko IKA, S.M., Gunawan D, *Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters.* Journal of Telecommunication, Electronic and Computer Engineering **8**.

155. https://www.britannica.com/technology/neural-network. *Neural network.* 2nd March 2019]; Available from: https://www.britannica.com/technology/neural-network.

156. Rajkomar A, O.E., Dean J, *Scalable and accurate deep learning with electronic health records.* npj Digital Medicine 2018. **1**.

157. Chen M, H.Y., Hwang K, Wang L, Wang LLU WANG1, AND LIN WANG, *Disease Prediction by Machine Learning Over Big Data From Healthcare Communities.* IEEE Access, 2017.

158. Zhou, X., et al., *Human symptoms–disease network.* Nature Communications, 2014. **5**.

159. SB, K., *Decision trees: a recent overview.* Artif Intell Rev 2013. **39**.

160. Lynch, C.M., et al., *Prediction of lung cancer patient survival via supervised machine learning classification techniques.* Int J Med Inform, 2017. **108**: p. 1-8.

161. Scheer, J.K., et al., *Development of a validated computer-based preoperative predictive model for pseudarthrosis with 91% accuracy in 336 adult spinal deformity patients.* Neurosurgical Focus, 2018. **45**(5): p. E11-E11.

162. Hivert, M.F., et al., *Identifying primary care patients at risk for future diabetes and cardiovascular disease using electronic health records.* BMC Health Serv Res, 2009. **9**: p. 170.

163. Darcy A. Davis, N.V.C., Nicholas A. Christakis, Albert-László Barabási, *Time to CARE: a collaborative engine for practical disease prediction.* Data Mining and Knowledge Discovery, 2010. **20**: p. 388-415.

164. Xierali, I.M., et al., *The rise of electronic health record adoption among family physicians.* Ann Fam Med, 2013. **11**(1): p. 14-9.

165. Simborg, D.W., D.E. Detmer, and E.S. Berner, *The wave has finally broken: now what?* J Am Med Inform Assoc, 2013. **20**(e1): p. e21-5.

166. Park HA, H.N., *Clinical Terminologies: A Solution for Semantic Interoperability.* J Kor Soc Med Informatics, 2009. **15**: p. 1-11.

167. Shortliffe EH, P.L.e., *Medical Informatics – computer applications in health care*. 1990, Reading, MA: Addison-Wesley.

168. Organisation, W.H. *Classification of Diseases*. 2nd January 2018]; Available from: http://www.who.int/classifications/icd/en/.

169. Digital, N. *UK Read Code*. 2016 2nd January 2018]; Available from: https://data.gov.uk/dataset/uk-read-code.

170. International, S. *SNOMED International: The Systematized Nomenclature of Medicine*. 2017 2nd January 2018]; Available from: http://www.snomed.org.

171. S. de Lusignan, C.M., J. Kennedy, M. Zeimet, J. Bommezijn and H. Bryant, *A survey to identify the clinical coding and classification systems currently in use across Europe.* Studies in Health Technology and Informatics, 2001. **84**: p. 86-89.

172. Standardization, I.O.f., *Health Informatics – Principles of mapping between terminological systems (ISO/TC 215).* 2014, ISO: Geneva.

173. Bonney, W., et al., *Mapping Local Codes to Read Codes.* Stud Health Technol Inform, 2017. **234**: p. 29-36.

174. M, C. *EHR Conversion Hurdles: Merging Data from Multiple Sources*. 2017 2nd January 2018]; Available from: https://www.healthdataarchiver.com/emr-conversion-hurdles-merging-data-from-multiple-sources/.

175. Miller, R., F.E. Masarie, and J.D. Myers, *Quick medical reference (QMR) for diagnostic assistance.* MD Comput, 1986. **3**(5): p. 34-48.

176. Barnett, G.O., et al., *DXplain. An evolving diagnostic decision-support system.* JAMA, 1987. **258**(1): p. 67-74.

177. D. Sherertz, M.T., M. Blois and M. Erlbaum. *Intervocabulary mapping within the UMLS: the role of lexical matching,*. in *Proceedingsof the Twelfth Annual Symposium on Computer Applications in Medical Care.* 1988. Washington DC: IEEE Computer Society Press.

178. Sun, J.Y. and Y. Sun, *A system for automated lexical mapping.* J Am Med Inform Assoc, 2006. **13**(3): p. 334-43.

179. Kim, T.Y., *Automating lexical cross-mapping of ICNP to SNOMED CT.* Inform Health Soc Care, 2016. **41**(1): p. 64-77.

180. Stenzhorn, H., et al., *Automatic mapping of clinical documentation to SNOMED CT.* Stud Health Technol Inform, 2009. **150**: p. 228-32.

181. Patrick J, W.Y., Budd P, *Automatic Mapping Clinical Notes to Medical Terminologies.* Australasian Language Technology Workshop, 2006: p. 75-82.

182. Fung, K.W., et al., *Combining lexical and semantic methods of inter-terminology mapping using the UMLS.* Stud Health Technol Inform, 2007. **129**(Pt 1): p. 605-9.

183. Cimino, J.J. and G.O. Barnett, *Automated translation between medical terminologies using semantic definitions.* MD Comput, 1990. **7**(2): p. 104-9.

184. Rocha, R.A., B.H. Rocha, and S.M. Huff, *Automated translation between medical vocabularies using a frame-based interlingua.* Proc Annu Symp Comput Appl Med Care, 1993: p. 690-4.

185. Digital, N. *NHS UK Read Codes Clinical Terms Version 3, Cross Maps*. 2017 2nd January 2018]; Available from: https://isd.digital.nhs.uk/trud3/user/guest/group/2/pack/9.

186. Brouch, K., *AHIMA project offers insights into SNOMED, ICD-9-CM mapping process.* J AHIMA, 2003. **74**(7): p. 52-5.

187. Nandigam H, T.M., *Mapping Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) to International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM): Lessons Learned from Applying the National Library of Medicine's Mappings.* Perspectives in Health Information Management, 2016. **Summer 2016**.

188. Goossen, W., *Cross-mapping between three terminologies with the international standard nursing reference terminology model.* Int J Nurs Terminol Classif, 2006. **17**(4): p. 153-64.

189. Nadkarni, P. and J. Darer, *Migrating existing clinical content from ICD-9 to SNOMED.* Journal of the American Medical Informatics Association : JAMIA, 2010. **17**(5): p. 602-607.

190. Kessels, R.P., *Patients' memory for medical information.* J R Soc Med, 2003. **96**(5): p. 219-22.

191. Witten, I. and E. Frank, *Data mining: practical machine learning tools and techniques*. 2005, San Francisco: Morgan Kaufmann Publishers.

192. Blois, M.S., *Information and Medicine: The Nature of Medical Descriptions*. 1984, London: University of California Press.

193. Jaki, S., *Brain, Mind and Computers*. 1969, South Bend, Ind.: Gateway Editions Ltd.

194. Drake, R. and G. McHugo, *Large data sets can be dangerous.* Psychiatric Services, 2003. **54**(2): p. 2:133.

195. Kruppa, J., A. Ziegler, and I. Konig, *Risk estimation and risk prediction using machine-learning methods. Hum Genet.* Human Genetics, 2012. **131**(10): p. 1639-54.

196. Thomas F Whayne, J., *Atherosclerosis: Current Status of Prevention and Treatment.* International Journal of Angiology, 2011. **20**(4): p. 213-222.

197. Celi, L.A., A.J. Zimolzak, and D.J. Stone, *Dynamic Clinical Data Mining: Search Engine-Based Decision Support.* Journal of Medical Internet Research, 2014. **2**(1): p. e13.

198. Thatipamula, S. *Data Done Right: 6 Dimensions of Data Quality*. 2013 12th November 2018]; Available from: https://smartbridge.com/data-done-right-6-dimensions-of-data-quality/.

199. Kantardzic, M., *Data mining : concepts, models, methods, and algorithms*. 2nd ed. 2011, Hoboken, N.J.: John Wiley : IEEE Press. xvii, 534 p.

200. Wu, Y.I., K; Govindaraju, V, *Improved k-nearest neighbor classification.* Pattern Recognition, 2002. **35**: p. 2311-2318.

201. Dudani, S., *The Distance-Weighted k-Nearest-Neighbor Rule.* IEEE Transactions on Systems, Man, and Cybernetics, 1976. **SMC-6**(4).

202. Cancer, I.A.f.R.o. *EPI-CT: International pediatric CT scan study. Epidemiology*. 2011 [cited 2014 2nd February 2014]; Available from: http://epi-ct.iarc.fr/epidemiology/index.php.

203. CORDIS. *Epidemiological study to quantify risks for paediatric computerized tomography and to optimise doses*. 2017 [cited 2017 11th September 2017]; Available from: https://www.cordis.europa.eu/project/rcn/97571_en.html.

204. Ozasa, K., et al., *Studies of the Mortality of Atomic Bomb Survivors, Report 14, 1950-2003: An Overview of Cancer and Noncancer Diseases.* Radiation research, 2012. **177**: p. 229-243.

205. Bosch de Basea, M., et al., *EPI-CT: design, challenges and epidemiological methods of an international study on cancer risk after paediatric and young adult CT.* Journal of Radiological Protection, 2015. **35**(3): p. 611-28.

206. Hall, E. and D. Brenner, *Cancer risks from diagnostic radiology.* British Journal of Radiology, 2008. **81**: p. 362-378.

207. Olerud, H., *Utsendelse av Rapport om CT Doser og Undersøkelsesteknik*. 1990, Statens Strålevern.
208. Jahnen, A., et al., *Automatic Computed Tomography Patient Dose Calculation Using DICOM Header Metadata.* Radiation protection dosimetry, 2011. **147**(1-2): p. 317-320.
209. Raosoft. *Sample size calculator*. 2004 4th May 2014]; Available from: http://www.raosoft.com/samplesize.html.
210. GmbH, L. *LimeSurvey*. [cited 2014 12th February 2014]; Available from: https://www.limesurvey.org/.
211. Dick, R. and E. Steen, *The Computer-Based Patient Record: An Essential Technology for Health Care*. 1991, Washington, DC: National Academy Press.
212. ASTM, *ASTM E1384 Standard Guide on Content and Structure of Electronic Health Records*. 2007.
213. Kho, A., et al., *Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc.* Journal of the American Medical Informatics Association, 2011.
214. Bonney, W., et al., *Mapping Local Codes to Read Codes*, in *Building Capacity for Health Informatics in the Future*, F.e.a. Lau, Editor. 2017, IOS Press.
215. Campbell, K. and K. Giannangelo, *Language barrier: Getting past the classifications and terminologies roadblock.* Journal of AHIMA. **78**(2): p. 44-8.
216. Berthold, M.R., et al., *KNIME - the Konstanz information miner: version 2.0 and beyond.* SIGKDD Explor. Newsl., 2009. **11**(1): p. 26-31.
217. US National Library of Medicine. *Unified Medical Language System ICD-9-CM Diagnostic Codes to SNOMED CT Map, Version 2016*. 2016 1st January 2018]; Available from: https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html.
218. NHS Digital, *NHS UK Read Codes Clinical Terms Version 3, Cross Maps.* 2017.
219. De, S., *8 Steps to Success in ICD-10-CM/PCS Mapping: Best Practices to Establish Precise Mapping Between Old and New ICD Code Sets.* Journal of AHIMA, 2012. **83**(6): p. 44-49.
220. Benson, T., *The history of the Read Codes: the inaugural James Read Memorial Lecture 2011.* Informatics in primary care, 2011. **19**: p. 173-82.
221. Robinson, D., et al., *Updating the Read Codes: User-interactive Maintenance of a Dynamic Clinical Vocabulary.* Journal of the American Medical Informatics Association, 1997. **4**(6): p. 465-472.
222. Maletic, J. and A. Marcus, *Data Cleansing: A Prelude to Knowledge Discovery*, in *Data Mining and Knowledge Discovery Handbook*, Maimon and L. Rokach, Editors. 2010, Springer.
223. Orr, K., *Data Quality and Systems Theory.* Communications of the ACM, 1998. **41**((2)): p. 66-71.
224. Redman, T., *The Impact of Poor Data Quality on the Typical Enterprise.* Communications of the ACM. **41**(2): p. 79-82.
225. Hays, W.L., *Statistics*. 4th ed. 1988, Fort Worth, TX: Holt, Rinehart and Winston Inc.

226. Team, R.C., *R: A language and environment for statistical computing*. 2016, R Foundation for Statistical Computing: Vienna.

227. Springate, D., et al., *ClinicalCodes: An Online Clinical Codes Repository to Improve the Validity and Reproducibility of Research Using Electronic Medical Records*. PLoS ONE, 2014. **9**(6): p. e99825.

228. Watson, J., et al., *Identifying clinical features in primary care electronic health record studies: methods for codelist development*. BMJ Open, 2017. **7**: p. e019637.

229. Kendall, M., *A New Measure of Rank Correlation*. Biometrika, 1938. **30**(1-2): p. 81-93.

230. Wessa. *Kendall tau Rank Correlation (v1.0.13) in Free Statistics Software (v1.2.1)*. 2017 20th January 2019]; Available from: https://www.wessa.net/rwasp_kendall.wasp/.

231. Wolf-Maier, K., et al., *Hypertension prevalence and blood pressure levels in 6 European countries, Canada, and the United States*. JAMA, 2003. **289**(18): p. 2363-9.

232. Mills, K.T., et al., *Global Disparities of Hypertension Prevalence and Control: A Systematic Analysis of Population-Based Studies From 90 Countries*. Circulation, 2016. **134**(6): p. 441-50.

233. Organization, W.H. *Global Health Observatory (GHO) data: Raised cholesterol*. [cited 2017 1st March 2017]; Available from: https://www.who.int/gho/ncd/risk_factors/cholesterol_text/en/.

234. Prevention, C.f.D.C.a., *National Diabetes Statistical Report 2017: Estimates of Diabetes and Its Burden in the United States*. 2017.

235. UK, D., *Diabetes Facts and Stats: 2015*. 2015.

236. Meucci, R.D., A.G. Fassa, and N.M. Faria, *Prevalence of chronic low back pain: systematic review*. Rev Saude Publica, 2015. **49**.

237. Martin, P., *The epidemiology of anxiety disorders: a review*. Dialogues Clin Neurosci, 2003. **5**(3): p. 281-98.

238. OECD, *Obesity Update 2012*. 2012.

239. Bauchau, V. and S.R. Durham, *Prevalence and rate of diagnosis of allergic rhinitis in Europe*. Eur Respir J, 2004. **24**(5): p. 758-64.

240. Dent, J., et al., *Epidemiology of gastro-oesophageal reflux disease: a systematic review*. Gut, 2005. **54**(5): p. 710-7.

241. Eurostat. *Persons reporting a chronic disease, by disease, sex, age and educational attainment level* 2014 [cited 2017 2nd February 2017]; Available from: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_ehis_cd1e&lang=en.

242. Vanderpump, M.P., *The epidemiology of thyroid disease*. Br Med Bull, 2011. **99**: p. 39-51.

243. Naing, N., *Easy Way to Learn Standardization : Direct and Indirect Methods*. Malaysian Journal of Medical Science, 2000. **7**(1): p. 10-15.

244. National Cancer Institute. *SEER*Stat Tutorials: Calculating Age-adjusted Rates*. [cited 2017; Available from: https://seer.cancer.gov/seerstat/tutorials/aarates/definition.html.

245. Holman, N., B. Young, and R. Gadsby, *What is the current prevalence of diagnosed and yet to be diagnosed diabetes in the UK.* Diabet Med, 2014. **31**(5): p. 510-1.
246. Diseases, N.I.o.D.a.D.a.K. *Overweight & Obesity Statistics*. 2015 [cited 2017 2nd February 2017]; Available from: https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity.
247. Europe, W.H.O.R.O.f., *Nutrition, Physical Activity and Obesity United Kingdon of Great Britain and Northern Ireland*. 2013.
248. Pawankar R, C.G., ST Holgate ST, Lockey RF, Blaiss M, *WAO White Book on Allergy*. 2013, World Allergy Organisation.
249. Grimes, D.A. and K.F. Schulz, *Bias and causal associations in observational research.* Lancet, 2002. **359**(9302): p. 248-52.
250. Prevention, C.f.D.C.a. *Knowing Your Risk for High Cholesterol*. 2019 [cited 2019 3rd March 2019]; Available from: https://www.cdc.gov/cholesterol/risk_factors.htm.
251. Prevention, C.f.D.C.a. *High Blood Pressure Risk Factors*. 2014 [cited 2018 3rd March 2018]; Available from: https://www.cdc.gov/bloodpressure/risk_factors.htm.
252. Mayo Clinic. *Anxiety disorders*. 2015 [cited 2017 3rd March 2017]; Available from: https://www.mayoclinic.org/diseases-conditions/anxiety/symptoms-causes/syc-20350961.
253. Mayo Clinic. *Back pain*. [cited 2017 3rd March 2017]; Available from: https://www.mayoclinic.org/diseases-conditions/back-pain/symptoms-causes/syc-20369906).
254. Mayo Clinic. *Obesity.*
255. Mayo Clinic. *Hay fever.*
256. Choices, N. *Heartburn and acid reflux*. 2014; Available from: https://www.nhs.uk/conditions/heartburn-and-acid-reflux/.
257. NHS Choices. *Chronic obstructive pulmonary disease (COPD)*. 2016 [cited 2017 3rd March 2017]; Available from: https://www.nhs.uk/conditions/chronic-obstructive-pulmonary-disease-copd/causes/.
258. NHS Choices. *Underactive thyroid (hypothyroidism)*. 2015 [cited 2017 3rd March 2017]; Available from: https://www.nhs.uk/conditions/underactive-thyroid-hypothyroidism/causes/.
259. World Health Organization. *Genes and human diseases*. [cited 2017 10th March 2017]; Available from: http://www.who.int/genomics/public/geneticdiseases/en/index2.html).
260. *Regulation (EC) N°141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products* 1999: European Union.
261. Mind. *How common are mental health problems?* 2013 [cited 2017 4th April 207]; Available from: https://www.mind.org.uk/information-support/types-of-mental-health-problems/statistics-and-facts-about-mental-health/how-common-are-mental-health-problems/#.
262. Dobbin, K.K. and R.M. Simon, *Optimally splitting cases for training and testing high dimensional classifiers.* BMC Med Genomics, 2011. **4**: p. 31.
263. Sae-Hyun Ji, M.P., Hyun-Soo Lee, You-Sang Yoon. *Sae-Hyun Ji, Moonseo Park, Hyun-Soo Lee, and You-Sang Yoon, Similarity measurement method of*

case-based reasoning for conceptual cost estimation. in *Proceedings of the International Conference on Computing in Civil and Building Engineering*. 2010. Nottingham: Nottingham University Press.

264. Christopoulos, D., *R package 'inflection' documentation.* 2014.
265. Bray J.R, C.J., *An ordination of upland forest communities of southern Wisconsin.* Ecological Monographs, 1957. **27**: p. 325-349.
266. Anderson MJ, M.R., *Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern New Zealand.* Journal of Experimental Marine Biology and Ecologyq, 2004. **305**: p. 191-221.
267. Oksanen, J., et al., *vegan: Community Ecology Package. R package*. 2017.
268. Pearson, R. *Computing Odds Ratios in R*. 2011 [cited 2017 6th June 2017]; Available from: https://www.r-bloggers.com/computing-odds-ratios-in-r/.
269. Karpati, T., *R/mechkar.R* 2017.
270. M.J. Anderson, R.B.M., *Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern New Zealand.* Journal of Experimental Marine Biology and Ecology, 2004. **305**: p. 191-221.
271. Shortliffe, E. and L. Perreault, *Medical Informatics: Computer Applications in Health Care*. 1990: Addison-Wesley.
272. Intragumtornchai, T., et al., *The role of serum ferritin in the diagnosis of iron deficiency anaemia in patients with liver cirrhosis.* J Intern Med, 1998. **243**(3): p. 233-41.
273. Kerlikowske, K., et al., *Likelihood ratios for modern screening mammography. Risk of breast cancer based on age and mammographic interpretation.* JAMA, 1996. **276**(1): p. 39-43.
274. Clark, M., *Prediction of clinical risks by analysis of preclinical and clinical adverse events.* J Biomed Inform, 2015. **54**: p. 167-73.
275. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern classification*. 2nd ed. 2001, New York: Wiley. xx, 654 p.
276. Schmidt M, L.R.N., Bach F, *Minimizing Finite Sums with the Stochastic Average Gradient.* Mathematical Programming, 2013. **162**(1-2).
277. Hobbs, F., *Cardiovascular disease: different strategies for primary and secondary prevention?* Heart, 2010. **90**(10): p. 1217-1223.
278. Klimek, L. and P. Schendzielorz, *Early detection of allergic diseases in otorhinolaryngology.* GMS Current Topics in Otorhinolaryngology, Head and Neck Surgery, 2008. **7**: p. Doc04.
279. World Health Organization. *Early detection of cancer*. [cited 2017 12th December 2017]; Available from: https://www.who.int/cancer/detection/en/.
280. Haahtela, T., *Early treatment of asthma.* Allergy, 1999. **54**(Suppl 49): p. 74-81.
281. Hafkamp-de Goren, E., *Early detection and counselling intervention of asthma symptoms in preschool children: study design of cluster randomised controlled trial.* BMC Public Health, 2010. **10**(555).
282. Remington, B., R. Hastings, and H. Kovshoff, *Early intensive behavioral intervention: outcomes for children with autism and their parents after two years.* American Journal on Mental Retardation, 2007. **112**(6): p. 418-438.
283. Saadatmand, S., et al., *Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients.* BMJ, 2015. **351**: p. h4901]8.

284. Hitokazu Esaki, P.M.B., Yael Renert-Yuval, Tali Czarnowicki, Thy Huynh, Gary Tran, Sarah Lyon, Giselle Rodriguez, Supriya Immaneni, Donald B. Johnson, Bruce Bauer, Judilyn Fuentes-Duculan, Xiuzhong Zheng, Xiangyu Peng, Yeriel D. Estrada, Hui Xu, Christina de Guzman Strong, Mayte Suárez-Fariñas, James G. Krueger, Amy S. Paller, Emma Guttman-Yassky, *Early-onset pediatric atopic dermatitis is TH2 but also TH17 polarized in skin.* Journal of Allergy and Clinical Immunology, 2016.

285. Thomsen, S., *Atopic Dermatitis: Natural History, Diagnosis, and Treatment.* ISRN Allergy, 2014. **2014**: p. 354250.

286. Hassall, E., *Early detection can help prevent minor heartburn from becoming a major health issue.* Jouranl of Pediatrics, 2005. **146**(3 Suppl): p. S3-12.

287. MedStar Washington Hospital Center. *Gastroesophageal Reflux Disease (GERD).* 23/03/2018]; Available from: https://www.medstarwashington.org/our-services/gastroenterology/conditions/gastroesophageal-reflux-disease-gerd/.

288. Chu, C.R., et al., *Early diagnosis to enable early treatment of pre-osteoarthritis.* Arthritis Research & Therapy, 2012. **14**(3): p. 212.

289. Pearce, E., *Diagnosis and management of thyrotoxicosis.* BMJ, 2006. **332**(7554): p. 1369-1373.

290. Pratley, R.E., *The Early Treatment of Type 2 Diabetes.* The American Journal of Medicine, 2013. **126**(9 Supplement 1): p. S2-S9.

291. Zomer, E., et al., *Effectiveness and cost-effectiveness of a cardiovascular risk prediction algorithm for people with severe mental illness (PRIMROSE).* BMJ Open, 2017. **7**(9).

292. Sweeney, L. *Achieving k-anonymity privacy protection using gerneralization and suppression.* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002. **10**: p. 571-588.

11 APPENDICES

The full text of this article has been removed for copyright reasons

Appendix 1: published article pages 382-421

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

APPENDIX 2: CODELISTS

Codelists were used in validation of the composite data set in chapter 5 and in determination of the presence of a condition in a record for risk prediction in chapter 6, since most conditions had many codes that indicated their presence, with different codes indicating a variation of the condition. Some rarer conditions had a codelist containing only one code.

The codes listed here are Read Codes Clinical Terms Version 3.

Codelist for acute sinusitis

Short name: acuteSinusitis

Number of codes: 18

CHILD	CTERM
H01yz	Other acute sinusitis NOS
XaNkV	Acute rhinosinusitis
XE0Xm	Acute sinusitis
H135.	Recurrent sinusitis
XM1QH	Sinusitis
H01..	Sinusitis (& acute)
XE0Yp	(Ac sinusitis: [NOS][ethmoidl][sphenoidl]) or ([pansinusit])
H010.	Acute maxillary sinusitis
H012.	Acute ethmoidal sinusitis
H011.	Acute frontal sinusitis
H013.	Acute sphenoidal sinusitis
X00m3	Suppurative sinusitis with complications
X00m4	Recurrent acute sinusitis
X00m5	Barotraumatic sinusitis
H01y0	Acute pansinusitis

H01y.	Other acute sinusitis
H01z.	Acute sinusitis NOS
Hyu00	[X]Other acute sinusitis

Codelist for allergic rhinitis

Short name: allergicRhinitis

Number of codes: 18

XE0Z5	**(Allerg rhinitis: [NOS][perenn][season]) or (vasomotor rhin)**
X00kx	Acute irritant rhinitis
X00lB	Perennial allergic rhinitis with seasonal variation
XE0Y5	Allergic rhinitis
Xa0lX	Seasonal allergic rhinitis
X00lA	Perennial allergic rhinitis
H17z.	Allergic rhinitis NOS
Hyu21	[X]Other allergic rhinitis
XE0Y6	Allergic rhinitis due to other allergens
XE0Y7	Allergic rhinitis due to unspecified allergen
H17..	Rhinitis: [perennial] or [allergic]
H172.	(Hay fever) or (allergic rhinitis) due to unspecif allergen
XE2QI	Allergic rhinitis due to pollens
Hyu20	[X]Other seasonal allergic rhinitis
X00l8	Hay fever - other allergen
X00l9	Hay fever - unspecified allergen
X1020	Hay fever with asthma
H170.	Pollinosis (& allergic rhinitis due to pollens)

Codelist for anxiety

Short name: anxiety

Number of codes: 38

CHILD	CTERM
1466	H/O: anxiety state
XaECG	Anxiety counselling
XE1aW	(Anxiety state (& [states][panic attack])) or (pseudocyesis)
E200.	Anxiety disorder
E2002	Generalised anxiety disorder
X00Sb	Mixed anxiety and depressive disorder
XE1YA	Phobic anxiety disorder
E2920	Separation anxiety disorder
X00RP	Organic anxiety disorder
E2000	Anxiety state unspecified
E2004	Chronic anxiety
E2005	Recurrent anxiety
E200z	Anxiety state NOS
X00Sc	Anxiety hysteria
Eu41y	[X] Anxiety disord: [other specified] or [anxiety hysteria]
E2924	Adjustment reaction with anxious mood
Eu930	[X]Separation anxiety disorder of childhood
E2D0z	Disturbance anxiety and fearfulness childhood/adolescent NOS
Eu40.	[X]Phobic anxiety disorders
Eu40y	[X]Other phobic anxiety disorders
Eu40z	[X]Phobic anxiety disorder, unspecified
Eu41.	[X]Other anxiety disorders
Eu46.	[X]Other neurotic disorders

Eu413	[X]Other mixed anxiety disorders
Eu41z	[X]Anxiety disorder, unspecified
XE1Zj	[X]Other specified anxiety disorders
Eu46z	[X]Neurotic disorder, unspecified
Ua2Dl	Alleviating anxiety
Eu412	[X]Mixed anxiety and depress disord (& mild anxiet depressn)
Eu410	[X]Panic disorder [episodic paroxysmal anxiety]
E202.	Phobic disorders (& [social] or [phobic anxiety])
Xa0XX	Anxiety about fainting
Xa0XY	Anxiety about having a heart attack
13WB.	Maternal: [concern] or [anxiety]
Ua1Fp	Acknowledging anxiety
XaABU	Castration anxiety complex
Ub0qs	Anxiety management training
XaL0q	Referral for guided self-help for anxiety

Codelist for any form of cancer

Short name: anyCancer

Number of codes: 200

CHILD	CTERM
XaPyg	Seen in fast track suspected cancer clinic
X78PC	Extrahepatic bile duct carcinoma
B162.	Malignant tumour of ampulla of Vater
X78Pf	Malignant tumour of endocrine pancreas
B340.	Malignant neoplasm of nipple and areola of female breast

B34y.	Malignant neoplasm of other site of female breast
B350.	Malignant neoplasm of nipple and areola of male breast
B35z.	Malignant neoplasm of other site of male breast
B430.	Malignant neoplasm of corpus uteri, excluding isthmus
B431.	Malignant neoplasm of isthmus of uterine body
X78Pn	Benign tumour of spleen
XE1w7	Benign tumour of breast
XE1wL	Neoplasm of uncertain behaviour of urinary organs OS/NOS
B932.	Neoplasm of uncertain behaviour of skin
Xa98f	Adnexal and skin appendage tumour
Xa98i	Cystic, mucinous and serous tumour
Xa98t	Ductal, lobular and medullary tumour
XM1FK	Paraganglioma and glomus tumour
XM1FS	Giant cell tumour
X77nC	Benign basal cell tumour
X77nA	Malignant basal cell tumour
XM1FG	Carcinoid tumour - argentaffin
X77oj	Malignant stromal tumour
Xa0Sr	Lymphoreticular tumour
Xa0KB	Tumour of external ear
Xa0KF	Tumour of lung
XaIpL	Cancer diagnosis discussed
Xa0os	Tumour of eyelid
X78Pz	Tumour of trachea
X78Q6	Tumour of bronchus
Xa0H4	Tumour of pericardium

Xa0De	Carcinoid tumour of intestine
X78PA	Tumour of extrahepatic bile duct
X78PF	Tumour of ampulla of Vater
X78PO	Tumour of body of pancreas
X78PR	Tumour of tail of pancreas
X78Pp	Tumour of peritoneum
B49..	Malignant tumour of urinary bladder
B46..	Malignant tumour of prostate
Xa1Ib	Tumour of testis
Xa1Id	Tumour of tunica vaginalis
B441.	Malignant tumour of fallopian tube
B43..	Malignant tumour of body of uterus
XE1wH	Carcinoma in situ of cervix
X77nX	Somatostatinoma
Xa98R	Tumour of adrenal cortex
Xa0ED	Tumour of adrenal medulla
Xa0GU	Odontogenic tumour of jaw
XaIlg	Procedure started
X77n8	Benign squamous cell tumour
X77n1	Squamous carcinoma in situ
X77mz	Malignant squamous tumour
Xa98A	Papillary carcinoma
Xa98B	Verrucous carcinoma
Xa98E	Squamous cell carcinoma
Xa98D	Squamous cell carcinoma in situ
X77n9	Papilloma
Xa98G	Basal cell carcinoma
Xa98L	Metastatic adenocarcinoma
X77nK	Adenocarcinoma in adenomatous polyp
Xa98Y	Solid carcinoma

Xa98e	Clear cell adenocarcinoma
Xa98P	Follicular adenocarcinoma
X77na	Chromophobe tumour
X77ng	Malignant endometrioid tumour
X77oM	In situ melanocytic morphology
X77oI	Malignant melanocytic lesion
X77pm	Nerve sheath tumour
Xa0a9	Neuroepithelial tumour morphology of uncertain origin
Xa0aA	Choroid plexus-derived tumours
Xa0aF	Embryonal neuroepithelial tumour
Xa0aI	Pineal tumour morphology
Xa0aL	Olfactory neuroepithelial-derived tumours
Xa0aQ	Meningeal-derived tumours
XaImo	Lymphoma staging system
XaIma	Gleason grading of prostate cancer
Xa0DI	Tumour of gastrointestinal tract
X78PD	Carcinoma in situ of extrahepatic bile duct
X78PZ	Neoplastic cyst of exocrine pancreas
X78Pq	Malignant tumour of peritoneum
Xa0WG	Primary malignant tumour of peritoneum
Xa0Bp	Melanocytic tumour of skin
Xa0D7	Malignant tumour of fibrous tissue
Xa0Pj	Tumour of soft tissue
B34..	Malignant neoplasm of female breast
B35..	Malignant neoplasm of male breast
B440.	Malignant tumour of ovary
XE1w9	Benign tumour of ovary
XE2vS	Malignant brain tumour
X77nb	Prolactinoma

X78id	Malignant tumour of male genital organ
X78iC	Malignant tumour of female genital organ
XE2vO	Malig neop of bone, connective tissue, skin and breast
XE2vP	Malignant neoplasm of genitourinary organ
XE2vR	Malignant neoplasm of other and unspecified sites
Xa0KC	Malignant tumour of external ear
X78gN	Malignant tumour of large intestine
B337.	Malignant neoplasm of skin of lower limb and hip
B33z.	Malignant neoplasm of skin NOS
Xa0CD	Malignant tumour of skin with pilar differentiation
Xa97s	Malignant tumour of soft tissue
B44..	Malignant neoplasm of ovary and other uterine adnexa
B45..	Malig neop of other and unspecified female genital organs
B47..	Malignant tumour of testis
XE2vQ	Malig neop of kidney and other unspecified urinary organs
BB...	Tumour morphology
X7A8T	Anatomical site notations for tumour staging
X7A6B	Generic tumour staging descriptors
X7A8A	Generic anatomical site tumour invasion status
X7A78	Specific tumour staging descriptors
Xa0LF	Tumour stages
Xa0LI	Metastasis stages
X7A6M	Tumour histopathological grade status values
X7A6W	Venous tumour invasion status values
X7A6b	Scleral tumour invasion status
X7A6l	Additional tumour staging descriptors

X7A6w	Tumour volume
X7A74	Generic tumour risk status stages
X7A6C	TNM tumour staging classifications
X7A6S	Lymphatic tumour invasion status stages
X7A6g	Residual tumour status stages
X7A70	Generic tumour extent
X7A79	Liver tumour staging descriptors
X7A7G	Lymphoma staging descriptors
X7A7q	Langerhans cell histiocytosis stages
X7A7S	Stannards retinoblastoma stages
X7A7A	Liver tumour size index
X7A7D	Timing of liver tumour staging
X7A7H	Lymphoma staging symptom status values
X7A7K	Lymphomatous extranodal involvement status values
X7A7T	Optic nerve tumour invasion status in retinoblastoma staging
X7A7Y	Choroidal tumour invasion status in retinoblastoma staging
X7A7e	Lymph nodal tumour invasion status in retinoblastoma staging
X7A7h	Brain tumour invasion status in retinoblastoma staging
X7A7r	Num of org systems involved Langerhans cell histiocytosis
X7A7u	Organ failure due to Langerhans cell histiocytosis
X7A80	Axillary lymph node level
X7A84	Abdominal lymph node tumour invasion status
X7A8B	Lung involvement stages
X7A8F	H+

X7A8G	Liver sectors
X7A8L	Markers for liver tumour staging
X90Tt	Tumour stage T1a
X90Tw	Tumour stage T1b
X90U2	Tumour stage T3b
Xa0IM	Tumour status
XM1FR	Blood vessel tumour
X78ef	Malignant tumour
Xa0KG	Malignant tumour of lung
B410.	Malignant neoplasm of endocervix
B411.	Malignant neoplasm of exocervix
B41y.	Malignant neoplasm of other site of cervix
B41z.	Malignant neoplasm of cervix uteri NOS
X77nE	Adenocarcinoma
X77nO	Endocrine tumour morphology
X77nf	Endometrioid tumour
XE1wF	Carcinoma in situ of digestive organ
Xa0IC	Size of occurrence
X7A6A	Cancer staging
Xa98g	Sweat gland tumour
Xa0Dg	Carcinoid tumour of large intestine
X90CP	Tumour stage T1
X90CX	Tumour stage T2
X90CZ	Tumour stage T3
Xa0ID	Tumour stage T4
X7A7V	Node stage N1
X7A7W	Node stage N2
X7A7X	Node stage N3
X7A7m	Metastasis stage M1
Xa0aS	Meningeal sarcoma

Xa7OT	Excision of basal cell carcinoma
Xa7Oh	Excision of skin carcinoma
Xa97r	Intrahepatic bile duct carcinoma
B1503	Hepatocellular carcinoma
XE2vT	Secondary malignant neoplasm of other specified sites
Xa988	Large cell carcinoma
Xa989	Small cell carcinoma
X70Ld	Intraepidermal squamous cell carcinoma - Bowen's type
X77n0	Metastatic squamous cell carcinoma
Xa98H	Basal cell carcinoma - sclerosing type
Xa98Q	Papillary carcinoma - follicular variant
Xa98k	Cystadenocarcinoma
Xa98m	Serous cystadenocarcinoma
Xa98o	Papillary cystadenocarcinoma
Xa98r	Mucinous cystadenocarcinoma
Xa98s	Papillary mucinous cystadenoma
X77nz	Non-infiltrating intraductal carcinoma
X77o6	Thecoma
X77oe	Malignant myomatous tumour
XM1FP	Brenner tumour
X77oz	Malignant haemangioma
XaIlC	Cancer dataset administrative items
XM1FQ	Plasma cell tumour
XaIls	Cancer treatment related morbidity
XaIlp	Reason for change in radiotherapy course
XaIlf	Reason for change in planned chemotherapy treatment
XaIlc	Presence of primary site synchronous tumours

XaIlD	Basis of cancer diagnosis
XaIlO	Reason for no specific anti-cancer treatment
XaL1F	Fast track cancer referral

Codelist for asthma

Short name: asthma

Number of codes: 131

CHILD	CTERM
14B4.	H/O: asthma
XaIer	Asthma follow-up
XaIfK	Asthma medication review
H33..	Asthma
X1024	Aspirin-sensitive asthma with nasal polyps
XaLPE	Nocturnal asthma
X101x	Allergic asthma
XE0YT	Non-allergic asthma
X1023	Drug-induced asthma
173A.	Exercise-induced asthma
X1025	Occupational asthma
XaKdk	Work aggravated asthma
H440.	Byssinosis
H441.	Cannabinosis
Xa0lZ	Asthmatic bronchitis
Xa9zf	Acute asthma
XE0YW	Asthma attack

Xa1hD	Exacerbation of asthma
Ua1AX	Brittle asthma
X101u	Late onset asthma
H332.	Mixed asthma
H33z.	Asthma unspecified
H44..	Pneumopathy due to inhalation of other dust
X101t	Childhood asthma
H330.	Asthma: [extrins - atop][allerg][pollen][childh][+ hay fev]
H3300	(Hay fever + asthma) or (extr asthma without status asthmat)
H331.	(Intrinsic asthma) or (late onset asthma)
H33z0	(Severe asthma attack) or (status asthmaticus NOS)
H33zz	(Asthma:[exerc ind][allerg NEC][NOS]) or (allerg bronch NEC)
XE0ZR	Asthma: [intrinsic] or [late onset]
XE0ZT	Asthma: [NOS] or [attack]
XE0YX	Asthma NOS
XaYZh	Number days absent from school due to asthma in past 6 month
X102C	Factitious asthma
XE0YQ	Allergic atopic asthma
X1021	Allergic non-atopic asthma
H330z	Extrinsic asthma NOS
X101y	Extrinsic asthma with asthma attack
X101z	Allergic asthma NEC
XE0YR	Extrinsic asthma without status asthmaticus
XE0YS	Extrinsic asthma with status asthmaticus
XaJFG	Aspirin-induced asthma
H47y0	Detergent asthma

X1026	Baker's asthma
X1027	Colophony asthma
X1028	Grain worker's asthma
X1029	Sulphite-induced asthma
XE0YV	Status asthmaticus NOS
XaBU3	Asthma monitoring status
663N.	Asthma disturbing sleep
663O.	Asthma not disturbing sleep
663P.	Asthma limiting activities
663Q.	Asthma not limiting activities
XaDvK	Asthma - currently active
XE0ZP	Extrinsic asthma - atopy (& pollen)
H3310	Intrinsic asthma without status asthmaticus
H331z	Intrinsic asthma NOS
X1022	Intrinsic asthma with asthma attack
XE0YU	Intrinsic asthma with status asthmaticus
H3311	Intrins asthma with: [asthma attack] or [status asthmaticus]
XM0s2	Asthma attack NOS
H3301	Extrins asthma with: [asthma attack] or [status asthmaticus]
H33z1	Asthma attack (& NOS)
663U.	Asthma management plan given
663V.	Asthma severity
663W.	Asthma prophylactic medication used
8791	Further asthma - drug prevention
XM1Xb	Asthma monitoring
XM1Xg	Chronic respiratory disease monitoring
XaBAQ	Recent asthma management
XaIQ4	Change in asthma management plan

XaIeq	Asthma annual review
XaIu6	Asthma monitoring by doctor
X1020	Hay fever with asthma
XaJ2A	Did not attend asthma clinic
8793	Asthma control step 0
8794	Asthma control step 1
8795	Asthma control step 2
8796	Asthma control step 3
8797	Asthma control step 4
8798	Asthma control step 5
X102D	Status asthmaticus
XaQij	Under care of asthma specialist nurse
663F.	Oral steroids started
663G.	Oral steroids stopped
663Y.	Steroid dose inhaled daily
663a.	Oral steroids used since last appointment
663c.	Nebulisation since last appointment
663d.	Emergency asthma admission since last appointment
XaDZF	Antiasthmatic agent
XaINb	Asthma causes daytime symptoms 1 to 2 times per month
XaINc	Asthma causes daytime symptoms 1 to 2 times per week
XaINd	Asthma causes daytime symptoms most days
XaIIW	Asthma accident and emergency attendance since last visit
XaIIX	Asthma treatment compliance satisfactory
XaIIY	Asthma treatment compliance unsatisfactory
XaIIZ	Asthma daytime symptoms
XaINa	Asthma never causes daytime symptoms

XaINf	Asthma limits walking up hills or stairs
XaINg	Asthma limits walking on the flat
XaINh	Number of asthma exacerbations in past year
XaINi	Number of times bronchodilator used in one week
XaINj	Number of times bronchodilator used in 24 hours
XaIoE	Asthma night-time symptoms
XaIww	Asthma trigger
XaY2V	Asthma never causes night symptoms
Xaa7Q	No asthma trigger identified by subject
XaIQD	Step up change in asthma management plan
XaIQE	Step down change in asthma management plan
XaINZ	Asthma causes night symptoms 1 to 2 times per month
XaXZm	Asthma causes night time symptoms 1 to 2 times per week
XaXZp	Asthma causes symptoms most nights
XaIuG	Asthma confirmed
XaLIm	Asthma trigger - respiratory infection
XaLIn	Asthma trigger - seasonal
XaLIr	Asthma trigger - animals
XaLJS	Asthma trigger - cold air
XaLJT	Asthma trigger - damp
XaLJU	Asthma trigger - emotion
XaObi	Asthma trigger - airborne dust
XaObj	Asthma trigger - exercise
XaObk	Asthma trigger - pollen
XaObl	Asthma trigger - tobacco smoke
XaObm	Asthma trigger - warm air
XaYja	Asthma trigger - wind
XaYpF	Asthma trigger - perfume

8H2P.	Emergency admission, asthma
XaYb8	Asthma self-management plan agreed
XaYZB	Asthma self-management plan review
21262	Asthma resolved
x02IG	Corticosteroids used in the treatment of asthma

Codelist for autism

Short name: autism

Number of codes: 5

CHILD	CTERM
X00TM	Autistic spectrum disorder
E14..	Psychoses with origin in childhood
XE2v2	Childhood autism
E1400	Active infantile autism
E1401	Residual infantile autism

Codelist for back pain

Short name: backPain

Number of codes: 41

CHILD	CTERM
8HTH.	Referral to back pain clinic
Xa7mE	Psychogenic back pain
E2780	Psychogenic pain unspecified
E2782	Psychogenic backache
XE1bM	Psychalgia: [tension backache] or [other]

XaZdZ	Low back pain clinical pathway
16C5.	C/O - low back pain
16C7.	C/O - upper back ache
Xa0ws	Thoracic back pain
Xa0wt	Low back pain
X75s1	Sacral back pain
XaIIv	Chronic back pain
16C..	Backache symptom
16C2.	Backache
16C3.	Backache with radiation
16C4.	Back pain worse on sneezing
16C6.	Back pain without radiation NOS
16CZ.	Backache symptom NOS
X75rz	Acute back pain with sciatica
XE1FE	Backache, unspecified
XaINe	Exacerbation of backache
N145.	(Backache unspecified) or (back pain unspecified & [acute])
XE1He	(Backache NOS) or (back pain [& low])
Xa0wp	Acute thoracic back pain
Xa0wq	Thoracic trigger point syndrome
Xa0wr	Thoracic segmental dysfunction
Xa0sM	Acute low back pain
Xa0sK	Chronic low back pain
Xa0wu	Mechanical low back pain
X75s3	Posterior compartment low back pain
Xa0wv	Lumbar trigger point syndrome
Xa0ww	Lumbar segmental dysfunction
N1420	Lumbago with sciatica
Xa7mB	Postural low back pain

XE0rW	Lumbar ache - renal
XE1FB	Pain in lumbar spine
N142.	(Back pain:[lumb sp][low][ac lum]) or (lumbalg) or (lumbago)
Xa0xt	Post-surgery back pain
XM1GI	Back pain
N12..	(Intervert disc: [disord][displ][slip]) or (acute back pain)
N143.	Acute back pain &/or sciatica

Codelist for bronchiectasis

Short name: bronchiectasis

Number of codes: 15

CHILD	CTERM
A115.	Tuberculous bronchiectasis
H34..	Bronchiectasis
P861.	Congenital bronchiectasis
X100m	Acquired bronchiectasis
H340.	Recurrent bronchiectasis
H34z.	Bronchiectasis NOS
XE1NO	(Congenital resp anomalies NOS) or (bronchiectasis - congen)
X100l	Congenital cystic bronchiectasis
H341.	Post-infective bronchiectasis
X100n	Idiopathic bronchiectasis
X100o	Obstructive bronchiectasis
X100p	Toxin-induced bronchiectasis

X100t	Post-lung transplantation bronchiectasis
X100q	Bronchiectasis due to toxic aspiration
X100r	Bronchiectasis due to toxic inhalation

Codelist for bronchitis

Short name: bronchitis

Number of codes: 90

CHILD	CTERM
XM1R4	H/O: bronchitis
14B3.	H/O: [COAD] or [bronchitis]
XE0tj	H/O: [obstructive airway disease(& chronic)] or [bronchitis]
H0608	Acute haemophilus influenzae bronchitis
H060B	Acute coxsackievirus bronchitis
X100A	Acute chlamydial bronchitis
XaREU	Aspergillus bronchitis
H301.	Laryngotracheobronchitis
H30..	Bronchitis: [unspecif (& chest infectn)] or [recurr wheezy]
XE0ZL	(Simple chron bronchitis)/(smok cough)/(sen tracheobronchit)
H06z.	Acute bronchitis or bronchiolitis NOS
XaDth	Acute infective tracheobronchitis
Xa0lW	Acute laryngotracheobronchitis
H300.	Tracheobronchitis NOS
H3122	Acute exacerbation of chronic obstructive airways disease

H312z	Obstructive chronic bronchitis NOS
H310.	Simple chronic bronchitis
XE0YM	Purulent chronic bronchitis
X101j	Occupational chronic bronchitis
H31y1	Chronic tracheobronchitis
H3121	Emphysematous bronchitis
H31y.	Other chronic bronchitis
H31z.	Chronic bronchitis NOS
XE0YL	Bronchitis unspecified
XE0ZN	Chronic: [bronchitis NOS] or [tracheobronchitis]
H310z	Simple chronic bronchitis NOS
Xa0lZ	Asthmatic bronchitis
H460.	Bronchitis and pneumonitis due to chemical fumes
H460z	Bronchitis and pneumonitis due to chemical fumes NOS
Xaa7C	Eosinophilic bronchitis
XE0Qw	Whooping cough
H0606	Acute pneumococcal bronchitis
H0609	Acute Neisseria catarrhalis bronchitis
H0607	Acute streptococcal bronchitis
H060x	Acute bacterial bronchitis unspecified
H060C	Acute parainfluenza virus bronchitis
H060D	Acute respiratory syncytial virus bronchitis
H060F	Acute echovirus bronchitis
H060E	Acute bronchitis due to rhinovirus
H060w	Acute viral bronchitis unspecified
XaYYt	Acute bronchiolitis due to human metapneumovirus
H0615	Acute bronchiolitis due to respiratory syncytial virus
X100D	Acute bronchiolitis due to adenovirus

XaDtg	Tracheobronchitis
XaDtP	Bronchitis
H4600	Acute bronchitis due to chemical fumes
XE0Zd	Acute chemical bronchitis
XaDtB	Acute infective bronchitis
H0600	Acute fibrinous bronchitis
H0601	Acute membranous bronchitis
H0602	Acute pseudomembranous bronchitis
H0604	Acute croupous bronchitis
H060v	Subacute bronchitis unspecified
H060z	Acute bronchitis NOS
XM1QX	Acute wheezy bronchitis
H060.	Acute bronchitis (& wheezy)
H06..	Acute bronchitis and bronchiolitis
XE0Yt	Acute: [bronchitis]/[chest infections]/[tracheobronchitis]
H30z.	Bronchitis NOS
H311.	Mucopurulent chronic bronchitis
H311z	Mucopurulent chronic bronchitis NOS
H313.	Mixed simple and mucopurulent chronic bronchitis
XM1QT	Acute fibrinous laryngotracheobronchitis
H3120	Chronic asthmatic bronchitis
X1007	Acute bacterial bronchitis
X1009	Acute mycoplasmal bronchitis
H0603	Acute purulent bronchitis
X100B	Acute viral bronchitis
Hyu10	[X]Acute bronchitis due to other specified organisms
X1006	Chest infection - unspecified bronchitis
XE0Xr	Acute bronchitis

H31..	Chronic bronchitis
H0605	Acute tracheobronchitis
X104u	Acute toxic tracheobronchitis

Codelist for bronchus cancer

Short name: bronchusCancer

Number of codes: 17

CHILD	CTERM
X78Q7	Malignant tumour of bronchus
X78QB	Benign tumour of bronchus
X78Q8	Squamous cell carcinoma of bronchus
X77nT	Carcinoid bronchial adenoma
X78kV	Metastasis to bronchus
B221.	Malignant neoplasm of main bronchus
B2210	Malignant neoplasm of carina of bronchus
B2220	Malignant neoplasm of upper lobe bronchus
B2230	Malignant neoplasm of middle lobe bronchus
B2240	Malignant neoplasm of lower lobe bronchus
XaEJe	Squamous cell carcinoma of bronchus in left lower lobe
XaEJf	Squamous cell carcinoma of bronchus in left upper lobe
XaEJg	Squamous cell carcinoma of bronchus in right lower lobe
XaEJh	Squamous cell carcinoma of bronchus in right middle lobe
XaEJi	Squamous cell carcinoma of bronchus in right upper lobe

Xa98a	Bronchial adenoma
X78QD	Papilloma of bronchus

Codelist for chronic pulmonary obstructive disease

Short name: COPD

Number of codes: 24

CHILD	CTERM
XaZd1	Acute non-infective exacerbation of COPD
Xa35l	Acute infective exacerbation chronic obstruct airway disease
H3y0.	Chronic obstruct pulmonary dis with acute lower resp infectn
X101i	Chron obstruct pulmonary dis wth acute exacerbation, unspec
XaX3c	Discussion about COPD exacerbation plan
XaPZH	COPD patient unsuitable for pulmonary rehabilitation
XaIUt	COPD self-management plan given
XaIet	Chronic obstructive pulmonary disease annual review
XaIu7	Chronic obstructive pulmonary disease monitoring by nurse
XaIu8	Chronic obstructive pulmonary disease monitoring by doctor
XaRCG	Step down change in COPD management plan
XaRCH	Step up change in COPD management plan
XaXCa	Chronic obstructive pulmonary disease 3 monthly review

XaXCb	Chronic obstructive pulmonary disease 6 monthly review
XaXnt	GP OOH service notified of COPD care plan
XaJFu	Admit COPD emergency
XaYbA	COPD self-management plan agreed
XaYZO	COPD self-management plan review
XaK8R	COPD accident and emergency attendance since last visit
XaK8S	Emergency COPD admission since last appointment
XaKzy	Multiple COPD emergency hospital admissions
XaXzy	Preferred place of care for next exacerbation of COPD
XaY0w	Referral to COPD community nursing team

Codelist for cystic fibrosis

Short name: cysticFibrosis

Number of codes: 14

CHILD	CTERM
XaZr7	Exacerbation of cystic fibrosis
XaREZ	Cystic fibrosis with distal intestinal obstruction syndrome
C3700	Cystic fibrosis with no meconium ileus
C3702	Cystic fibrosis with pulmonary manifestations
C3703	Cystic fibrosis with intestinal manifestations
C370z	Cystic fibrosis NOS
XaBDb	Cystic fibrosis with other manifestations
XaREa	Liver disease due to cystic fibrosis
XaXi9	Cystic fibrosis related cirrhosis
XaMzI	Cystic fibrosis related diabetes mellitus

XaQvc	Cystic fibrosis monitoring
XaVvv	Seen in cystic fibrosis clinic
XaQvd	Cystic fibrosis annual review

Codelist for eczema

Short name: eczema

Number of codes: 27

CHILD	CTERM
14F1.	H/O: eczema
XaQfn	Referral to eczema clinic
G831.	Varicose veins of the leg with eczema
G832.	Varicose veins of the leg with ulcer and eczema
XaEJY	Varicose veins of leg in long saph vein distribn with eczema
XaEJZ	Varicose veins of leg in short saph vein distrib with eczema
X505K	Eczema
XaY4o	Infected eczema
X505N	Atopic dermatitis of hands
M113.	Flexural atopic dermatitis
X505O	Inverse pattern atopic dermatitis
X505P	Discoid atopic dermatitis
X505Q	Erythrodermic atopic dermatitis
X505R	Follicular atopic dermatitis
X505S	Pruriginous atopic dermatitis
XaBsL	Chronic lichenified atopic dermatitis
X505T	Photosensitive atopic dermatitis

X505U	Photoaggravated atopic dermatitis
M112.	Infantile eczema
M115.	Besnier's prurigo
M117.	Atopic neurodermatitis
M11z.	Atopic dermatitis NOS
XE1C6	Atopic eczema/dermatitis NOS
M11..	Atopic dermatitis and related conditions
M111.	Atopic dermatitis

Codelist for emphysema

Short name: emphysema

Number of codes: 35

CHILD	CTERM
H32..	Emphysema
H3203	(Bullous emphysema with collapse) or (tension pneumatocoele)
H32y1	Emphysema: [acute interstitial] or [atrophic - senile]
X101n	Pulmonary emphysema
H32y2	MacLeods syndrome
H321.	Panlobular emphysema
H322.	Centrilobular emphysema
H32y.	Other emphysema
H32z.	Emphysema NOS
H3202	Giant bullous emphysema
H3200	Segmental bullous emphysema
H3201	Zonal bullous emphysema
H320z	Chronic bullous emphysema NOS
XE0YN	Bullous emphysema with collapse
H32y0	Acute vesicular emphysema

Hyu30	[X]Other emphysema
XE0YO	Atrophic (senile) emphysema
XE0YP	Other emphysema NOS
H32yz	(Sawyer-Jones syndrome) or (other emphysema NOS)
Q312y	Perinatal interstitial emphysema or related condition OS
Q312z	Perinatal interstitial emphysema or related condition NOS
Qyu34	[X]Oth conds relat/interstial emphysema orig perinatl period
X101o	Pulmonary emphysema in alpha-1 PI deficiency
X101p	Toxic emphysema
H320.	Chronic bullous emphysema
H582.	Compensatory emphysema
X101q	Congenital lobar emphysema
X101r	Scar emphysema
XaIQg	Interstitial pulmonary emphysema
H4640	Chronic emphysema due to chemical fumes
XaIQh	Mediastinal emphysema
H581.	(Emphysema [interstitial]/[mediastinal])/(pneumomediastinum)
XE1oC	Subcutaneous emphysema
Q3123	Perinatal pulmonary interstitial emphysema
Q312.	Perinatal interstitial emphysema and related conditions

Codelist for gastro-intestinal reflux disease or reflux esophagitis

Short name: refluxDisease

Number of codes: 14

CHILD	CTERM
X3003	Gastro-oesophageal reflux disease
XE0bv	Oesophagitis (& [reflux]) or oesophageal reflux
XE0aL	Gastro-oesophageal reflux disease with oesophagitis
XE0aO	Gastro-oesophageal reflux disease without oesophagitis
X3005	Peptic stricture of oesophagus
J1020	Gastro-oesophageal reflux disease with ulceration
Xa9Bz	Barrett's oesophagus
J1011	Acid reflux &/or oesophagitis
J1016	(Ulcerative oesophagitis) or (Barrett's oesophagus)
J10y4	Oesophageal reflux (& [without mention of oesophagitis])
X70jT	Radionuclide gastro-oesophageal reflux study
X70fi	Gastro-oesophageal reflux X-ray study

Codelist for gastroparesis

Short name: gastroparesis

Number of codes: 1

CHILD	CTERM
X301s	Gastroparesis

Codelist for hiatus hernia

Short name: hiatusHernia

Number of codes: 16

CHILD	CTERM
X30BB	Sliding hiatus hernia
X30BC	Rolling hiatus hernia
X30BE	Mixed hiatus hernia
PA6..	Congenital hiatus hernia
XaC18	Hiatus hernia - irreducible
XaC19	Hiatus hernia with gangrene
XaC1A	Hiatus hernia with obstruction
XaC1B	Simple hiatus hernia
XaC2M	Hiatus hernia NOS
X30BD	Rolling hiatus hernia with gastric volvulus
Xa9Ze	Hiatus hernia repair
760K4	Boerema repair of hiatus hernia
760K0	Transthoracic hiatus hernia repair (& [Allison] or [Mason])
X30BA	Hiatus hernia
XaJlJ	Laparoscopic repair of hiatus hernia

Codelist for hyperlipidaemia

Short name: hyperlipidaemia

Number of codes: 13

CHILD	CTERM
Cyu8D	[X]Other hyperlipidaemia
U60C6	[X]Antihyperlipidaem/antiarterioscl drg caus adv ef ther use
XE13A	Disord lipid metab (& [Fredrick types] or

	[hyperlipidaemia])
X40Wy	Hyperlipidaemia
X40Vm	Familial combined hyperlipidaemia
XE11U	Mixed hyperlipidaemia
C324.	Hyperlipidaemia NOS
X40XI	Primary combined hyperlipidaemia
X40XO	Secondary combined hyperlipidaemia
Xa2hC	Dietary advice for hyperlipidaemia
C3202	Hyperlipidaemia, group A
C322.	(Mix hyperlipid) or (Fredr lip: [IIb][III]) or (xanthom tub)
XaJYh	Hyperlipidaemia clinical management plan

Codelist for hypertension

Short name: hypertension

Number of codes: 134

CHILD	CTERM
X30BB	Sliding hiatus hernia
X30BC	Rolling hiatus hernia
X30BE	Mixed hiatus hernia
PA6..	Congenital hiatus hernia
XaC18	Hiatus hernia - irreducible
XaC19	Hiatus hernia with gangrene
XaC1A	Hiatus hernia with obstruction
XaC1B	Simple hiatus hernia
XaC2M	Hiatus hernia NOS
X30BD	Rolling hiatus hernia with gastric volvulus
Xa9Ze	Hiatus hernia repair

760K4	Boerema repair of hiatus hernia
760K0	Transthoracic hiatus hernia repair (& [Allison] or [Mason])
X30BA	Hiatus hernia
XaJlJ	Laparoscopic repair of hiatus hernia

Codelist for hypothyroidism

Short name: hypothyroidism

Number of codes: 55

CHILD	CTERM
XaLUg	Hypothyroidism review
Xa0l7	Congenital hypothyroidism with diffuse goitre
X40H8	Congenital hypothyroidism without goitre
C03y.	Other specified congenital hypothyroidism
XE107	Congenital hypothyroidism NOS
C03z.	Congenital hypothyroidism: [cretinism] or [NOS]
X40HN	Radioactive iodine-induced hypothyroidism
X40Hx	Hypothyroidism due to coupling defect
X40Hy	Hypothyroidism due to deiodase defect
F3814	Myasthenic syndrome due to hypothyroidism
X40HF	Hypothyroidism due to Hashimoto's thyroiditis
X40HG	Hypothyroidism due to TSH receptor blocking antibody
X40HM	Postablative hypothyroidism
C043.	Other iatrogenic hypothyroidism
C043z	Iatrogenic hypothyroidism NOS
C040.	Hypothyroidism: [postsurgical] or [post ablative]
XE109	Post-surgical hypothyroidism
C0410	Irradiation hypothyroidism

C041.	Other postablative hypothyroidism
C041z	Postablative hypothyroidism NOS
C0430	Hypothyroidism resulting from para-aminosalicylic acid
C0431	Hypothyroidism resulting from phenylbutazone
C0432	Hypothyroidism resulting from resorcinol
X40Hq	Euthyroid hypothyroxinaemia
X40IB	Subclinical iodine deficiency hypothyroidism
C042.	Iodine hypothyroidism
X40I7	Congenital iodine deficiency hypothyroidism
XaJ9F	Subclinical hypothyroidism
C03..	Congenital hypothyroidism
XE108	Acquired hypothyroidism
Cyu11	[X]Other specified hypothyroidism
Xa3ec	Hypothyroidism - congenital and acquired
C04..	Hypothyroidism: &/or (acquired)
C04z.	Hypothyroid (& [pretib myxoed][acq goitr][NOS][thyr insuf])
XE124	Hypothyroidism - congen and acquir (& [cretinism][myxoedem])
X40HE	Autoimmune hypothyroidism
Q4337	Neonatal jaundice with congenital hypothyroidism
X769C	Hypothyroid facies
X40IQ	Hypothyroidism
X40HH	Borderline hypothyroidism
X40HI	Compensated hypothyroidism
X40HL	Iatrogenic hypothyroidism
X40HO	Drug-induced hypothyroidism
X40HP	Post-infectious hypothyroidism
C04y.	Other acquired hypothyroidism

X40HD	Hypothyroid goitre, acquired
XE10A	Hypothyroidism NOS
Xa3ed	Acquired hypothyroidism NOS
C04z0	Premature puberty due to hypothyroidism
1432	H/O: hypothyroidism
Xa08g	Transient neonatal hypothyroidism
XaJYj	Hypothyroidism clinical management plan
XaOjl	Hypothyroidism annual review

Codelist for lung cancer

Short name: lungCancer

Number of codes: 68

CHILD	**CTERM**
XE1yR	Ca trachea/bronchus/lung NOS
B2231	Malignant neoplasm of middle lobe of lung
B223z	Malignant neoplasm of middle lobe, bronchus or lung NOS
XE1yN	Ca middle lobe bronchus/lung
B2241	Malignant neoplasm of lower lobe of lung
B224z	Malignant neoplasm of lower lobe, bronchus or lung NOS
XE1yP	Ca lower lobe bronchus/lung
Xa3A5	Metastasis to lung of unknown primary
X2032	Pulmonary tumour embolism
X78kX	Secondary lymphangitic carcinoma
X78QU	Carcinoma in situ of lung parenchyma
B812.	Carcinoma in situ of bronchus and lung
B81z.	Carcinoma in situ of respiratory organ NOS
B8122	Carcinoma in situ of upper lobe bronchus and lung

B8123	Carcinoma in situ of middle lobe bronchus and lung
B8124	Carcinoma in situ of lower lobe bronchus and lung
B812z	Carcinoma in situ of bronchus or lung NOS
B907.	Neoplasm of uncertain behaviour trachea, bronchus and lung
B90z.	Neoplasm of uncertain behaviour of respiratory organs OS/NOS
B9072	Neoplasm of uncertain behaviour of lung
B907z	Neop of uncertain behaviour of trachea, bronchus or lung NOS
X78QS	Non-small cell lung cancer
X78QF	Malignant tumour of lung parenchyma
X78QV	Benign tumour of lung parenchyma
X78QG	Adenocarcinoma of lung
X78QI	Carcinoid tumour of lung
X78QJ	Carcinoma of lung parenchyma
X78QQ	Epithelioid haemangioendothelioma of lung
X78QR	Lymphomatoid granulomatosis of lung
XaBAp	Bronchioloalveolar adenocarcinoma of lung
X78QK	Large cell carcinoma of lung
X78QN	Small cell carcinoma of lung
X78QP	Squamous cell carcinoma of lung
X78QL	Clear cell carcinoma of lung
X78QM	Giant cell carcinoma of lung
X78QO	Oat cell carcinoma of lung
X78QW	Histiocytoma of lung
X78QX	Adenoma of lung
Byu50	(Mesothelioma of lung) or ([X]mesothelioma of other sites)
B570.	Metastasis to lung

Xa3A4	Metastasis to bronchus of unknown primary
Xa0KG	Malignant tumour of lung
B22..	Malignant neoplasm of trachea, bronchus and lung
B225.	Malignant neoplasm of overlapping lesion of bronchus & lung
XE1yF	(Bronchus carc) or (lung carc) or (Ca trachea/bronchus/lung)
B222z	Malignant neoplasm of upper lobe, bronchus or lung NOS
XE1yL	Ca upper lobe bronchus/lung
B7230	Benign neoplasm of carina of bronchus
B7231	Benign neoplasm of main bronchus
B7232	Benign neoplasm of upper lobe bronchus and lung
B7233	Benign neoplasm of middle lobe bronchus and lung
B7234	Benign neoplasm of lower lobe bronchus and lung
B723z	Benign neoplasm of bronchus or lung NOS
B723.	Benign neoplasm of lung (& [bronchus])
X78QE	Tumour of lung parenchyma
XaFr7	Local recurrence of malignant tumour of lung
X78QT	Pancoast tumour
B2211	Malignant neoplasm of hilus of lung
B2221	Malignant neoplasm of upper lobe of lung
B223.	Malignant neoplasm of middle lobe, bronchus or lung
B224.	Malignant neoplasm of lower lobe, bronchus or lung
B22y.	Malignant neoplasm of other sites of bronchus or lung
Byu20	[X]Malignant neoplasm of bronchus or lung, unspecified
XE1vb	Malignant neoplasm of upper lobe, bronchus or lung
XE1vc	Malignant neoplasm of bronchus or lung NOS

B222.	Malig neopl of upper lobe/bronchus/lung: (& [Pancoast synd])
B22z.	Malig neopl lung: [of bronchus or lung NOS] or [lung cancer]

Codelist for obesity

Short name: obesity

Number of codes: 77

CHILD	CTERM
222A.	O/E - obese
XE1h3	O/E - weight 10-20% over ideal
XE1h4	O/E - weight greater than 20% over ideal
XM1YD	O/E - overweight
22A4.	O/E - overweight (& [weight 10-20% over ideal])
22A5.	O/E weight: [>20% over ideal] or [obese]
XaZ0S	Anti-obesity drug therapy
Xaa0k	Childhood obesity
X40YM	Android obesity
X40YN	Gynaecoid obesity
X40YO	Generalised obesity
XE2Q3	Localised obesity
X40YQ	Morbid obesity
X76BU	Central obesity
X40YT	Simple obesity
X76BV	Peripheral obesity
XSCIZ	Adult-onset obesity
XSCIX	Lifelong obesity
C3800	Obesity due to excess calories
C3801	Drug-induced obesity

Cyu70	[X]Other obesity
Xa0Cx	Pulmonary hypertension with extreme obesity
L161.	(Gest oedem/nonhypertens excess wt gain) or (mat obesit syn)
X40Lm	Hypothalamic obesity
X40YR	Pickwickian syndrome
C3802	Extreme obesity with alveolar hypoventilation
XaBM0	Simple obesity NOS
66C4.	Has seen dietitian - obesity
66C8.	Attends slimming clinic
66CC.	Wants to lose weight
66CD.	Difficulty maintaining weight loss
66CE.	Reason for obesity therapy - occupational
9OK1.	Attends obesity monitoring
9OK2.	Refuses obesity monitoring
9OK3.	Obesity monitoring default
9OK9.	Obesity monitoring deleted
9OKA.	Obesity monitoring check done
XaKiY	Weight management programme offered
XaKiZ	Weight management plan started
XaKia	Weight management plan completed
1444	H/O: obesity
C38z.	Obesity and other hyperalimentation NOS
66C5.	Treatment of obesity changed
66C6.	Treatment of obesity started
66C7.	Treatment of obesity stopped
66CA.	Ideal weight discussed
66CZ.	Obesity monitoring NOS
XE1T7	Target weight discussed
XaX5k	Inter risk health ass overwt ob gen adv hlthy wgt

	lifestyle
XaX5l	Int risk health ass overwt ob advice about diet physical act
XaX5m	Inter risk hlth overwght obesity adv diet phys act cons drug
XaX5n	Inter risk hlth owt ob adv diet phy ac cons dgs cons surgery
9OKZ.	Obesity monitoring admin.NOS
XM1U4	Obesity clinic administration
XaX5e	Risk health associa overweight obesity, at no increased risk
XaX5f	Risk health associ overweight and obesity, at increased risk
XaX5g	Risk health associated overweight and obesity, at high risk
XaX5h	Risk health associ overweight and obesity, at very high risk
22K2.	Body mass index high K/M2
22K4.	Body mass index index 25-29 - overweight
22K5.	Body mass index 30+ - obesity
XaJJH	Body mass index 40+ - severely obese
C380.	Obesity
C38..	Obesity and other hyperalimentation
Cyu7.	[X]Obesity and other hyperalimentation
Xa2hD	Dietary advice for obesity
XE1T6	Obesity monitoring
66C..	Weight monitoring (& obesity)
66C9.	Weight: [target discussed] or [loss advised]
XE13Y	(Hyperalimentation including obesity) or (adiposity)
XE2Nc	Obesity monitoring admin.

9OK4.	Obesity monitoring first letter
9OK5.	Obesity monitoring second letter
9OK6.	Obesity monitoring third letter
9OK7.	Obesity monitoring verbal invite
9OK8.	Obesity monitoring telephone invite
XaKko	Obesity resolved

Codelist for osteoarthritis

Short name: osteoarthritis

Number of codes: 151

CHILD	CTERM
N0507	Heberden's nodes with arthropathy
N0503	Bouchard's nodes with arthropathy
N0500	Generalised osteoarthritis of unspecified site
N0502	Generalised osteoarthritis of multiple sites
N050z	Generalised osteoarthritis NOS
XE1DW	Generalised osteoarthritis of the hand
N0501	(Heberden nodes) or (Bouchard nodes) or (gen osteoarth hand)
N0510	Localised, primary osteoarthritis of unspecified site
N0511	Localised, primary osteoarthritis of the shoulder region
N0512	Localised, primary osteoarthritis of the upper arm
N0513	Localised, primary osteoarthritis of the forearm
N0514	Localised, primary osteoarthritis of the hand
N0515	Localised, primary osteoarthritis of the pelvic region/thigh
N0516	Localised, primary osteoarthritis of the lower leg

N0517	Localised, primary osteoarthritis of the ankle and foot
N0518	Localised, primary osteoarthritis of other specified site
N051z	Localised, primary osteoarthritis NOS
N051A	Coxarthrosis resulting from dysplasia, bilateral
N0529	Post-traumatic coxarthrosis, bilateral
N0520	Localised, secondary osteoarthritis of unspecified site
N0521	Localised, secondary osteoarthritis of the shoulder region
N0522	Localised, secondary osteoarthritis of the upper arm
N0523	Localised, secondary osteoarthritis of the forearm
N0524	Localised, secondary osteoarthritis of the hand
N0526	Localised, secondary osteoarthritis of the lower leg
N0527	Localised, secondary osteoarthritis of the ankle and foot
N0528	Localised, secondary osteoarthritis of other specified site
N052z	Localised, secondary osteoarthritis NOS
XE1DX	Localised, secondary osteoarthritis of pelvic region/thigh
N0525	(Loc 2ndry osteoarth pelv regn/thigh) or (coxae malum senil)
N0530	Localised osteoarthritis, unspecified, of unspecified site
N0531	Localised osteoarthritis, unspecified, of shoulder region
N0532	Localised osteoarthritis, unspecified, of the upper arm

N0533	Localised osteoarthritis, unspecified, of the forearm
N0534	Localised osteoarthritis, unspecified, of the hand
N0537	Localised osteoarthritis, unspecified, of the ankle and foot
N0538	Localised osteoarthritis, unspecified, of other spec site
N0539	Arthrosis of first carpometacarpal joint, unspecified
N053z	Localised osteoarthritis, unspecified, NOS
XE1DY	Localised osteoarthritis, unspecified, pelvic region/thigh
XE1DZ	Localised osteoarthritis, unspecified, of the lower leg
N0540	Oligoarticular osteoarthritis, unspec, of unspecified sites
N0541	Oligoarticular osteoarthritis, unspecified, of shoulder
N0542	Oligoarticular osteoarthritis, unspecified, of upper arm
N0543	Oligoarticular osteoarthritis, unspecified, of forearm
N0544	Oligoarticular osteoarthritis, unspecified, of hand
N0545	Oligoarticular osteoarthritis, unspecified, of pelvis/thigh
N0546	Oligoarticular osteoarthritis, unspecified, of lower leg
N0547	Oligoarticular osteoarthritis, unspecified, of ankle/foot
N0548	Oligoarticular osteoarthritis, unspecified, other spec sites
N0549	Oligoarticular osteoarthritis, unspecified, multiple sites
N054z	Osteoarthritis of more than one site, unspecified, NOS
NyuC7	[X]Other hypertrophic osteoarthropathy
X702z	Toe osteoarthritis NOS

N051.	Localised, primary osteoarthritis
N05z9	Osteoarthritis NOS, of shoulder
N05zB	Osteoarthritis NOS, of acromioclavicular joint
N05zC	Osteoarthritis NOS, of elbow
XaEGf	Localised, primary osteoarthritis of elbow
N05zE	Osteoarthritis NOS, of wrist
XaEGd	Localised, primary osteoarthritis of the wrist
X703H	Osteoarthritis of distal interphalangeal joint
X703I	Osteoarthritis of proximal interphalangeal joint
X703J	Osteoarthritis of metacarpophalangeal joint of finger
X7035	Finger osteoarthritis NOS
N05zH	Osteoarthritis NOS, of DIP joint of finger
N05zG	Osteoarthritis NOS, of PIP joint of finger
N05zJ	Osteoarthritis NOS, of hip
XaYQD	Patellofemoral osteoarthritis
N05zL	Osteoarthritis NOS, of knee
N05zN	Osteoarthritis NOS, of ankle
N05zP	Osteoarthritis NOS, of subtalar joint
N352.	Hallux rigidus - acquired
N05zS	Osteoarthritis NOS, of 1st metatarsophalangeal joint
N05zU	Osteoarthritis NOS, of interphalangeal joint of toe
XaEGe	Localised, primary osteoarthritis of toe
N061.	Arthritis secondary to trauma
Xa1jD	Arthritis due to bleeding disorder
XE2sp	Neuropathic arthropathy
N0505	Secondary multiple arthrosis
N052.	Localised, secondary osteoarthritis
X7039	Spondylosis
X703A	Osteoarthritis of spinal facet joint
X703B	Osteoarthritis of shoulder joint

X703C	Osteoarthritis of acromioclavicular joint
X703D	Osteoarthritis of elbow
X703E	Osteoarthritis of wrist
X703F	Osteoarthritis of first carpometacarpal joint
X703G	Osteoarthritis of finger joint
X703K	Osteoarthritis of hip
X703L	Osteoarthritis of knee
X703M	Osteoarthritis of ankle
X703N	Osteoarthritis of subtalar joint
X703O	Osteoarthritis of first metatarsalphalangeal joint
X703P	Osteoarthritis of toe joint
XaBmY	Osteoarthritis of foot joint
N053.	Localised osteoarthritis, unspecified
XM1NQ	Osteoarthritis of metacarpophalangeal joint
Xa3gQ	Osteoarthritis - hand joint
Xa3gR	Osteoarthritis - ankle/foot
Xa3gS	Osteoarthritis - other joint
N0535	(Otto pel)(hip osteoart NOS)(loc osteoart uns pel reg/thigh)
N0536	Osteoarthritis: [localised low leg unsp] or [patellofemoral]
XE2Qb	Kaschin-Beck disease
X704S	Malemud disease
N060.	(Kaschin-Beck disease) or (endemic polyarthritis)
N312.	Hypertrophic osteoarthropathy
14G2.	H/O: osteoarthritis
XaIna	Exacerbation of osteoarthritis
X7041	Localised osteoarthritis
N050.	Generalised osteoarthritis
X703Q	Secondary osteoarthritis

X704R	Endemic osteoarthritis
X7038	Idiopathic osteoarthritis
N0506	Erosive osteoarthrosis
N054.	Oligoarticular osteoarthritis, unspecified
XE1Da	Osteoarthritis NOS
XE1Gm	Osteoarthritis -multiple joint
N05..	Osteoarthritis (& [allied disorders])
N05z.	[Joint degeneration] or [osteoarthritis NOS]
X7043	Coxae malum senilis
X7042	Otto's pelvis
N05z0	Osteoarthritis NOS, of unspecified site
N05z1	Osteoarthritis NOS, of shoulder region
N05z8	Osteoarthritis NOS, other specified site
N05zA	Osteoarthritis NOS, of sternoclavicular joint
N05zD	Osteoarthritis NOS, of distal radioulnar joint
N05zF	Osteoarthritis NOS, of metacarpophalangeal joint
N05zK	Osteoarthritis NOS, of sacroiliac joint
N05zM	Osteoarthritis NOS, of tibiofibular joint
N05zQ	Osteoarthritis NOS, of talonavicular joint
N05zR	Osteoarthritis NOS, of other tarsal joint
N05zT	Osteoarthritis NOS, of lesser metatarsophalangeal joint
X7030	Foot osteoarthritis NOS
X7031	Ankle osteoarthritis NOS
XE1Db	Osteoarthritis NOS, of the upper arm
XE1Dc	Osteoarthritis NOS, of the forearm
XE1Dd	Osteoarthritis NOS, of the hand
XE1De	Osteoarthritis NOS, pelvic region/thigh
XE1Df	Osteoarthritis NOS, of the lower leg
XE1Dg	Osteoarthritis NOS, of ankle and foot

N05z2	Osteoarthritis NOS: [of the upper arm] or [elbow]
N05z3	Osteoarthritis NOS: [of the forearm] or [wrist]
N05z4	Osteoarthritis NOS: [hand] or [finger] or [thumb]
N05z5	Osteoarthritis NOS: [pelvic region and/or thigh] or [hip]
N05z6	Osteoarthritis NOS: [lower leg] or [knee]
N05z7	Osteoarthritis NOS: [ankle &/or foot] or [toe]
X7034	Thumb osteoarthritis NOS
XaLsk	Delivery of rehabilitation for osteoarthritis
XE1DV	Osteoarthritis
N11..	(Spondyl & allied dis) or (arthr spine) or (osteoarth spine)

Codelist for pleural effusion

Short name: pleuralEffusion

Number of codes: 12

CHILD	CTERM
H51y0	Encysted pleurisy
H51yz	Other pleural effusion
Xa0lb	Pleural effusion
XE0Zl	(Pleural effusion NOS) or (haemothorax) or (hydrothorax)
XE2wM	Pleural empyema
Xa0IL	Malignant pleural effusion
X1012	Benign asbestos pleural effusion
X1013	Drug-induced pleural effusion
XaB1L	Haemorrhagic pleural effusion
H51y.	Other pleural effusion excluding mention of tuberculosis

H51z.	Pleural effusion NOS
Hyu70	[X]Pleural effusion in conditions classified elsewhere

Codelist for pneumonia

Short name: pneumonia

Number of codes: 143

CHILD	CTERM
14B2.	H/O: pneumonia
H24y6	Typhoid pneumonia
A116.	Tuberculous pneumonia
X100L	Meningococcal pneumonia
XE0YH	Pneumococcal lobar pneumonia
AyuK3	[X]Streptococ pneumon/cause/disease classified/oth chapters
A3By4	Mycoplasmal pneumonia
X100G	Atypical pneumonia
AyuK9	[X]Mycoplasma pneumoniae [PPLO]cause/dis classifd/oth chaptr
H242.	Ornithosis with pneumonia
H2470	Candidal pneumonia
H2471	Pneumonia with coccidioidomycosis
H2472	Pneumonia with histoplasmosis
AB4z5	Histoplasmosis with pneumonia
H24y5	Toxoplasma pneumonia
X100E	Pneumonia
H2701	Influenza with pneumonia, influenza virus identified
H2y..	Other specified pneumonia or influenza
H2z..	Pneumonia or influenza NOS

H22y.	Pneumonia due to other specified bacteria
H22z.	Bacterial pneumonia NOS
H2230	Group B streptococcal pneumonia
H22yz	Pneumonia due to bacteria NOS
XaDtl	Legionnaire's disease
H23z.	Pneumonia due to specified organism NOS
X100h	Giant cell pneumonia
H247z	Pneumonia with systemic mycosis NOS
H24yz	Pneumonia with other infectious diseases EC NOS
H260.	Lobar pneumonia due to unspecified organism
XE0YJ	Bronchopneumonia due to unspecified organism
XaBE9	Basal pneumonia due to unspecified organism
H270.	Influenzal pneumonia
H2700	Influenza with bronchopneumonia
H270z	Influenza with pneumonia NOS
SP131	Other aspiration pneumonia as a complication of care
H56y0	Endogenous lipoid pneumonia
H5303	Abscess of lung with pneumonia
H5830	Acute eosinophilic pneumonia
X100P	Fetal pneumonia
X100a	Neonatal pneumonia
Xa0B7	Congenital viral pneumonia
Xa0B8	Congenital bacterial pneumonia
Q310y	Other specified congenital pneumonia
Q310z	Congenital pneumonia NOS
Qyu32	[X]Congenital pneumonia due to other organisms
H2...	Pneumonia and influenza
H24..	(Pneumonia) or (chest infection) with infectious diseases EC
XaZ1l	Hospital acquired pneumonia

XaZ1k	Community acquired pneumonia
XaDsa	Infective pneumonia
X100M	Bronchopneumonia
X100X	Haemorrhagic pneumonia
H5400	Hypostatic pneumonia
Xa0lY	Lobar pneumonia
H56y1	Interstitial pneumonia
XaFrU	Postoperative pneumonia
H5302	Gangrenous pneumonia
Hyu0G	[X]Pneumonia in other diseases classified elsewhere
XM0rv	Pneumonia NOS
XaJEl	Bilateral pneumonia
XE0ZF	Pneumonia and influenza &/or pneumonia
XE0ZH	Pneumonia: [lobar] or [pneumococcal]
H24y0	Actinomycotic pneumonia
H222.	Haemophilus influenzae pneumonia
H22y2	Legionella pneumonia
X100J	Pneumococcal pneumonia
A0222	Salmonella pneumonia
H224.	Staphylococcal pneumonia
H220.	Pneumonia due to Klebsiella pneumoniae
H221.	Pseudomonal pneumonia
H22y0	Escherichia coli pneumonia
H22y1	Proteus pneumonia
H244.	Tularaemia pneumonia
H243.	Pertussis pneumonia
H245.	Anthrax pneumonia
H24y1	Nocardial pneumonia
H223.	Streptococcal pneumonia
X100Y	Mycobacterial pneumonia

H22..	Other bacterial pneumonia
H232.	Pneumonia due to pleuropneumonia-like organism
Hyu09	[X]Pneumonia due to other aerobic gram-negative bacteria
Hyu0A	[X]Other bacterial pneumonia
Hyu0C	[X]Pneumonia in bacterial diseases classified elsewhere
XaBfJ	Secondary bacterial pneumonia
H21..	Pneumococcal pneumonia (& lobar)
H25..	Bronchopneumonia: [unspec organism] or [chest infect - unsp]
X100O	Neonatal chlamydial pneumonia
X100S	Pulmonary mucormycosis
H24y2	Pneumocystis carinii pneumonia
H246.	Pneumonia with aspergillosis
AB405	Histoplasma capsulatum with pneumonia
Hyu0E	[X]Pneumonia in mycoses classified elsewhere
Xa0Y7	Non-tuberculous mycobacterial pneumonia
Q310.	Congenital pneumonia
Xa0B9	Acquired neonatal pneumonia
X70Ua	Capillaria aerophila chest infection
Hyu0F	[X]Pneumonia in parasitic diseases classified elsewhere
X100f	Mononuclear interstitial pneumonia
X102k	Seasonal crypt organising pneumonia, biochemical cholestasis
X1035	Cholesterol pneumonia
X1038	Lupus pneumonia
H571.	Rheumatic pneumonia
X1039	Traumatic pneumonia

A7893	HIV disease resulting in Pneumocystis carinii pneumonia
X100d	Rickettsial pneumonia
XaYYu	Pneumonia due to human metapneumovirus
X100e	Glandular fever pneumonia
H200.	Adenoviral pneumonia
H202.	Parainfluenzal pneumonia
H201.	Pneumonia due to respiratory syncytial virus
H20y.	Viral pneumonia NEC
H20z.	Viral pneumonia NOS
Hyu08	[X]Other viral pneumonia
Hyu0D	[X]Pneumonia in viral diseases classified elsewhere
H20..	Viral pneumonia (& chest infection)
H247.	Pneumonia with other systemic mycoses
H24y.	Pneumonia with other infectious diseases EC
H24z.	Pneumonia with infectious diseases EC NOS
Q3100	Congenital staphylococcal pneumonia
Q3101	Congenital group A haemolytic streptococcal pneumonia
Q3102	Congenital group B haemolytic streptococcal pneumonia
Q3103	Congenital Escherichia coli pneumonia
Q3104	Congenital pseudomonal pneumonia
Qyu31	[X]Congenital pneumonia due to other bacterial agents
Xa0BA	Meconium pneumonitis
Xa0BB	Neonatal aspiration pneumonia
X70Eg	Pneumonitis due to fetal aspiration
Xa7nL	Basal pneumonia
Xa7nU	Right upper zone pneumonia

Xa7nT	Right middle zone pneumonia
Xa7nP	Left upper zone pneumonia
Xa7nN	Right lower zone pneumonia
Xa7nM	Left lower zone pneumonia
X100H	Bacterial pneumonia
XE0YG	Viral pneumonia
X100R	Fungal pneumonia
X100b	Pneumonia due to parasitic infestation
XaDsb	Congenital infective pneumonia
H23..	Pneumonia due to other specified organisms
H26..	Pneumonia due to unspecified organism
Hyu0B	[X]Pneumonia due to other specified infectious organisms
Hyu0H	[X]Other pneumonia, organism unspecified
XE0YI	Pneumonia with infectious diseases EC

Codelist for Prader-Willi syndrome or Cushing’s disease

Short name: praderWillli

Number of codes: 14

CHILD	CTERM
C150.	Cushing's syndrome
X40MD	Adrenal Cushing's syndrome
X40MB	ACTH-dependent Cushing's syndrome
C1500	Idiopathic Cushing's syndrome
C1501	Iatrogenic Cushing's syndrome
X40ME	Cyclical Cushing's syndrome
C150z	Cushing's syndrome NOS
Cyu45	[X]Other Cushing's syndrome

PKy0.	(Multi syst cong anom NEC) or (Prader-Willi) or (Noonan syn)
PKy93	Prader-Willi syndrome
C1502	Pituitary-dependent Cushing's disease
C1503	Ectopic ACTH secretion causing Cushing's syndrome
F3951	Myopathy in Cushing's disease

Codelist for stress

Short name: stress

Number of codes: 45

CHILD	CTERM
XM0As	Stress and adjustment reaction
Eu4..	[X]Neurotic, stress-related and somatoform disorders
Ryu58	[X]State of emotional shock and stress, unspecified
X761N	Anxiety and fear
XM012	Mental distress
1B1J.	Emotional: [problem] or [upset]
XaX58	Delayed post-traumatic stress disorder follow military comb
XaX56	Chronic post-traumatic stress disorder follow military comb
XaX55	Acute post-traumatic stress disorder follow military combat
E2831	Acute post-trauma stress state
E29y1	Other post-traumatic stress disorder
XaEFB	Chronic post-traumatic stress disorder
XaI8j	Stress counselling

13H4.	Marital problems (& [stress])
XE0pM	Stress at home
X76AY	Work stress
XM1aI	Stress at work
XM1aJ	Work worries
Xa18j	Combat fatigue
Xa18v	Shell shock
E2830	Acute situational disturbance
E280.	Acute panic state due to acute stress reaction
E281.	Acute fugue state due to acute stress reaction
E282.	Acute stupor state due to acute stress reaction
E283.	Other acute stress reactions
E283z	Other acute stress reaction NOS
E284.	Stress reaction causing mixed disturbance of emotion/conduct
XE1Yn	Acute stress reaction NOS
E28..	Acute reaction to stress (& [combat fatigue])
Eu430	[X]Ac stress react (& [crisis][psych shock][combat fatigue])
Xa19c	Normal grief reaction
Ua18k	Abnormal grief reaction
E2900	Grief reaction (& [bereavement reaction])
XE1Ym	Acute stress reaction
X00Sf	Post-traumatic stress disorder
Xa7mz	Carer stress syndrome
E29..	Adjustment disorder
1B1L.	Stress-related problem
Eu43y	[X]Other reactions to severe stress
Eu43z	[X]Reaction to severe stress, unspecified
XE1bo	Acute reaction to stress (& [post-traum] or [shell-

	shock])
XM1Am	Undue concern and preoccupation with stressful events
Ua165	Feeling stressed
9ON1.	Attends stress monitoring
9ON2.	Refuses stress monitoring

Codelist for thyrotoxicosis

Short name: thyrotoxicosis

Number of codes: 49

CHILD	CTERM
C0240	Thyrotoxicosis from ectopic thyroid nodule with no crisis
C0241	Thyrotoxicosis from ectopic thyroid nodule with crisis
C024z	Thyrotoxicosis from ectopic thyroid nodule NOS
C02z0	Thyrotoxicosis without mention of goitre or cause no crisis
C02z1	Thyrotoxicosis without mention of goitre, cause with crisis
C02zz	Thyrotoxicosis NOS
F3816	Myasthenic syndrome due to thyrotoxicosis
XE1g7	(Perinatal endocr/metab NOS) or (thyrotoxicosis - perinatal)
X40Gx	HCG-induced thyrotoxicosis
X40H0	Thyrotoxicosis on thyroxine therapy
X40H1	Iodine-induced thyrotoxicosis
X40H2	Amiodarone-induced thyrotoxicosis
X40I4	Pituitary thyroid hormone resistance

X40H5	Thyrotoxicosis due to TSHoma
X40Hd	Chronic thyroiditis with transient thyrotoxicosis
X40Gk	Thyrotoxicosis due to Graves' disease
XE104	Thyrotoxicosis
C02..	([Thyrotoxicosis] or [hyperthyroidism]) or (toxic goitre)
C020.	Toxic diffuse goitre (& [Basedow disease] or [Graves dis])
X40Gl	Thyrotoxicosis due to Hashimoto's thyroiditis
X40Gn	Thyrotoxicosis due to acute thyroiditis
X40Go	Toxic nodular goitre
C024.	Thyrotoxicosis from ectopic thyroid nodule
X40Gs	T3 toxicosis
X40Gt	Borderline thyrotoxicosis
X40Gu	Autonomous thyroid function
X40Gv	Apathetic thyrotoxicosis
X40Gw	Thyrotoxicosis in pregnancy
Q443.	Neonatal thyrotoxicosis
X40Gy	Factitia thyrotoxicosis
X40Gz	Iatrogenic thyrotoxicosis
X40H4	Thyrotoxicosis due to inappropriate TSH secretion
X40H7	Thyrotoxicosis due to struma ovarii
C02z.	Thyrotoxicosis without mention of goitre or other cause
Cyu13	[X]Other thyrotoxicosis
XE106	Thyrotoxicosis of other specified origin
Xa3eb	Thyrotoxicosis with or without goitre
C02y.	Thyrotoxicosis: [other specified origin] or [factitia]

XE122	Thyrotoxicosis: [+/- goitr][tox goitr][Graves dis][thyr nod]
C02y0	Thyrotoxicosis of other specified origin with no crisis
C02y1	Thyrotoxicosis of other specified origin with crisis
C02yz	Thyrotoxicosis of other specified origin NOS

Codelist for type 2 diabetes

Short name: T2Diabetes

Number of codes: 32

CHILD	**CTERM**
66A..	Diabetic monitoring
XaX3o	Diabetic dietary review
XaOPu	Latent autoimmune diabetes mellitus in adult
XaOPt	Maternally inherited diabetes mellitus
X40J5	Type II diabetes mellitus
X40J6	Insulin treated Type 2 diabetes mellitus
C1011	Type 2 diabetes mellitus with ketoacidosis
C1031	Type II diabetes mellitus with ketoacidotic coma
XaF05	Type II diabetes mellitus with nephropathy
XaIzQ	Type II diabetes mellitus with persistent proteinuria
XaIzR	Type II diabetes mellitus with persistent microalbuminuria
C1096	Type II diabetes mellitus with retinopathy
XaFmA	Type II diabetes mellitus with diabetic cataract
XaJQp	Type II diabetes mellitus with exudative maculopathy
XaEnp	Type II diabetes mellitus with mononeuropathy

XaEnq	Type II diabetes mellitus with polyneuropathy
XaKyX	Type II diabetes mellitus with gastroparesis
XM19j	[EDTA] Diabetes Type II associated with renal failure
C1090	Type II diabetes mellitus with renal complications
C1091	Type II diabetes mellitus with ophthalmic complications
C1092	Type II diabetes mellitus with neurological complications
C1094	Type II diabetes mellitus with ulcer
C1095	Type II diabetes mellitus with gangrene
C1097	Type II diabetes mellitus - poor control
L1806	Pre-existing diabetes mellitus, non-insulin-dependent
XaELQ	Type II diabetes mellitus without complication
XaFWI	Type II diabetes mellitus with hypoglycaemic coma
XaFn7	Type II diabetes mellitus with peripheral angiopathy
XaFn8	Type II diabetes mellitus with arthropathy
Xa2hA	Dietary advice for type II diabetes
X405J	Postmortem caesarean section
XaFn9	Type II diabetes mellitus with neuropathic arthropathy

Appendix 3: Local R functions

A several R functions were written or adapted to use across all programs used in this work. They are shown here.

```
############################################################
# local functions
############################################################

calcOddsRatio <- function(n00, n01, n10, n11, alpha = 0.05){
  #from https://www.r-bloggers.com/computing-odds-ratios-in-r/
  #  Compute the odds ratio between two binary variables, x and y,
  # as defined by the four numbers nij:
  #    n00 = number of cases where x = 0 and y = 0
  #    n01 = number of cases where x = 0 and y = 1
  #    n10 = number of cases where x = 1 and y = 0
  #    n11 = number of cases where x = 1 and y = 1
  #
  OR <- (n00 * n11)/(n01 * n10)
  #
  #  Compute the Wald confidence intervals:
  #
  siglog <- sqrt((1/n00) + (1/n01) + (1/n10) + (1/n11))
  zalph <- qnorm(1 - alpha/2)
  logOR <- log(OR)
  loglo <- logOR - zalph * siglog
  loghi <- logOR + zalph * siglog
  ORlo <- exp(loglo)
  ORhi <- exp(loghi)
  oframe <- data.frame(LowerCI = ORlo, OR = OR, UpperCI = ORhi, alpha = alpha)
  oframe
}
```

```
calcLikelihoodRatio <- function(m, sig.level=0.95) {
  # from Tomas Karpati, https://rdrr.io/github/karpatit/mechkar/src/R/mechkar.R

  alpha <- 1 - sig.level
  a <- m[1, 1]
  b <- m[1, 2]
  c <- m[2, 1]
  d <- m[2, 2]
  spec <- d/(b+d)
  sens <- a/(a+c)
  lr.pos <- sens/(1 - spec)

  if (a != 0 & b != 0 ) {
    sigma2 <- (1/a) - (1/(a+c)) + (1/b) - (1/(b+d))
    lower.pos <- lr.pos * exp(-qnorm(1-(alpha/2))*sqrt(sigma2))
    upper.pos <- lr.pos * exp(qnorm(1-(alpha/2))*sqrt(sigma2))
  } else if ( a == 0 & b == 0 ) {
    lower.pos <- 0
    upper.pos <- Inf
  } else if ( a == 0 & b != 0 ) {
    a.temp <- (1/2)
    spec.temp <- d/(b+d)
    sens.temp <- a.temp/(a+c)
    lr.pos.temp <- sens.temp/(1 - spec.temp)
    lower.pos <- 0
    sigma2 <- (1/a.temp) - (1/(a.temp+c)) + (1/b) - (1/(b+d))
    upper.pos <- lr.pos.temp * exp(qnorm(1-(alpha/2))*sqrt(sigma2))
  } else if (a != 0 & b == 0) {
    b.temp <- (1/2)
    spec.temp <- d/(b.temp+d)
    sens.temp <- a/(a+c)
  lr.pos.temp <- sens.temp/(1 - spec.temp)
  sigma2 <- (1/a) - (1/(a+c)) + (1/b.temp) - (1/(b.temp+d))
  lower.pos <- lr.pos.temp * exp(-qnorm(1-(alpha/2))*sqrt(sigma2))
  upper.pos <- Inf
} else if ( (a == (a+c)) & (b == (b+d)) ) {
  a.temp <- a - (1/2)
  b.temp <- b - (1/2)
  spec.temp <- d/(b.temp+d)
  sens.temp <- a.temp/(a+c)
  lr.pos.temp <- sens.temp/(1 - spec.temp)
  sigma2 <- (1/a.temp) - (1/(a.temp+c)) + (1/b.temp) - (1/(b.temp+d))
  lower.pos <- lr.pos.temp * exp(-qnorm(1-(alpha/2))*sqrt(sigma2))
  upper.pos <- lr.pos.temp * exp(qnorm(1-(alpha/2))*sqrt(sigma2))
}

lr.neg <- (1 - sens)/spec
if (c != 0 & d != 0) {
  sigma2 <- (1/c) - (1/(a+c)) + (1/d) - (1/(b+d))
  lower.neg <- lr.neg * exp(-qnorm(1-(alpha/2))*sqrt(sigma2))
  upper.neg <- lr.neg * exp(qnorm(1-(alpha/2))*sqrt(sigma2))
} else if (c == 0 & d == 0) {
  lower.neg<- 0
  upper.neg <- Inf
} else if (c == 0 & d != 0) {
  c.temp <- (1/2)
  spec.temp <- d/(b+d)
  sens.temp <- a/(a+c.temp)
  lr.neg.temp <- (1 - sens.temp)/spec.temp
  lower.neg <- 0
  sigma2 <- (1/c.temp) - (1/(a+c)) + (1/d) - (1/(b+d))
  upper.neg <- lr.neg.temp * exp(qnorm(1-(alpha/2))*sqrt(sigma2))
} else if ( c != 0 & d == 0 ) {
    d.temp <- (1/2)
    spec.temp <- d.temp/(b+d)
    sens.temp <- a/(a+c)
    lr.neg.temp <- (1 - sens.temp)/spec.temp
    sigma2 <- (1/c) - (1/(a+c)) + (1/d.temp) - (1/(b+d))
    lower.neg <- lr.neg.temp * exp(-qnorm(1-(alpha/2))*sqrt(sigma2))
    upper.neg <- Inf
  } else if ( (c == (a+c)) & (d == (b+d)) ) {
    c.temp <- c - (1/2)
    d.temp <- d - (1/2)
    spec.temp <- d.temp/(b+d)
    sens.temp <- a/(a+c.temp)
    lr.neg.temp <- (1 - sens.temp)/spec.temp
    sigma2 <- (1/c.temp) - (1/(a+c)) + (1/d.temp) - (1/(b+d))
    lower.neg <- lr.neg.temp * exp(-qnorm(1-(alpha/2))*sqrt(sigma2))
    upper.neg <- lr.neg.temp * exp(qnorm(1-(alpha/2))*sqrt(sigma2))
  }
  list(
    lr.pos=lr.pos, lower.pos=lower.pos, upper.pos=upper.pos,
    lr.neg=lr.neg, lower.neg=lower.neg, upper.neg=upper.neg
  )
}

#function to create a binary vector showing whether or not event codes are present for a
patient
createBitVectors <- function(x){ # x is concatDF[row,], y is eventCodesGroup
  bitvector <- eventCodesGroup %in% x;
  return(bitvector)
}

#function to look up "include" flag for an event code
#getFlag <- function(x) {
```

```
#  event_flag <- include_flags[(match(x, include_flags$CHILD)),]$auto_flag;
#  return(event_flag)
#}

#function to look up ancestor code for an event code
getHigherCode <- function(x,y) {
  column_name <- paste("PARENT.",y,sep="");
  new_code <- include_flags[(match(x, include_flags$CHILD)),column_name];
  return(new_code)
}

#function to find position of most recent diagnosis code ('diagnosis' code as per flag
table)
most_recent_diagnosis_position <- function(x) {
  mrdp <- min(which(include_flags[match(x, include_flags$CHILD),]$auto_flag == 2));
#most recent diagnostic position
  if (mrdp == Inf) mrdp <- NA;
  return(mrdp);
}

#function to find the code of the most recent diagnosis code ('diagnosis' code as per flag
table)
mostRecentDiagnosisCode <- function(x) { # x is concatDF row, y is name of position
variable
  mrdc <- x[as.integer(tail(x,n=1))]
}

#function to look up the 'include flag' value for a code as stored in the flag table
includeFlagValue <- function(x) {
  flag_value <- include_flags[match(x, include_flags$CHILD),]$auto_flag
  return(as.integer(flag_value));
}
```

```
#function to generate true-false vector of any event in condition set occurring
TrueFalse <- function(x) {
  TFvalue <- x %in% targetCode;
  return(TFvalue);
}

#function to get predicted TRUE/FALSE value for presence of a condition for a patient
in a cluster
get_prediction <- function(x) {
  prediction <- as.logical(cc[match(x, cc$cluster),]$TrueFalsePred)
  return(prediction)
}

get_predictedRisk <- function(x) {
  predictedRisk <- as.numeric(cc[match(x, cc$cluster),]$prevalence)
  return(predictedRisk);
}

#########################################################
# end of local functions
#########################################################
```

APPENDIX 4: R CODE TO AUTOMATICALLY ASSIGN SIGNIFICANCE LEVELS TO CTV3 CODES

This program reads in CTV3 codes from the CTV3 table, assigns the code the significance of its parent code, and then corrects significance assignments as necessary. Corrections have only been made down to level 4 of the CTV3 hierarchy, beyond which level it was more practical to descend the CTV3 hierarchy, assign each code the significance of its parent, and once all codes had received a significance value, manually inspect the codes and their significances and correct as necessary. Note that the significances must be assigned and corrected in sequence of descending level in the CTV3 hierarchy since, initially, each code assumes the significance of its parent code.

```
cat("\nReading the full CTV3 codes flags and tree table and codes present in our data
set...\n");
#read in original CTV3 tree:
CTV3_flags <- read.csv("tree_rev_with_root_node.csv", header=TRUE,
stringsAsFactors=FALSE); # for ALL codes in TRUD CTV3 table
#read in revised CTV3 tree with corrected local codes:
CTV3_flags_local <- read.csv("finalIncludeFlagStatusForCTV3.csv", header = TRUE,
stringsAsFactors = FALSE); # list of all CTV3 codes with manually-assigned/corrected
flags for CTV3 codes in our data set
#CTV3_flags$flag <- CTV3_flags$INCLUDE
#colnames(CTV3_flags)[colnames(CTV3_flags) == "INCLUDE"] <- "flags"  # rename
the 'INCLUDE' flag column to 'flags' for consistency with code below

#read in the list of unique codes in our data set
unique_codes_list <- read.csv("unique_codes_in_merged_data_set.csv", header=TRUE,
stringsAsFactors=FALSE);

CTV3_flags$flag <- NA; # add column for semi-automatically generated flags
CTV3_flags$flag_auto <- NA; # add column for automatically generated flags
```

```
#####################################    Level 1
#####################################

#set "level 1" codes flag to 0 as our seed
CTV3_flags[(CTV3_flags$CHILD == "....."),]$flag <- 0

#work on level 1 codes - in fact only one code -
LEVEL.1.CODES <- unique(CTV3_flags$PARENT.1); # list of unique "level 1" codes
(these should all be ".....)
cat("\n", length(LEVEL.1.CODES), "Level 1 codes: ", LEVEL.1.CODES, "\n")
# set the level 1 codes to 0 (this should just be one CHILD code)
#CTV3_flags[(CTV3_flags$CHILD %in% LEVEL.1.CODES),]$flag <- 0; # set the
level 1 codes to 0 (this should just be one CHILD code)
#CTV3_flags[(CTV3_flags$PARENT == "....."),]$flag <-
CTV3_flags[(CTV3_flags$CHILD %in% LEVEL.1.CODES),]$flag
#####################################    Level 2
#####################################

#work on level 2 codes
LEVEL.2.CODES <- unique(CTV3_flags$PARENT.2); # list of unique "level 2" codes
(not many of these)
LEVEL.2.CODES <- LEVEL.2.CODES[!(LEVEL.2.CODES %in%
LEVEL.1.CODES)]
LEVEL.2.CODES <- LEVEL.2.CODES[(LEVEL.2.CODES %in%
unique_codes_list$CTV3)]
#LEVEL.2.CODES <- CTV3_flags[(CTV3_flags$CHILD %in%
LEVEL.2.CODES),]$CHILD # list of all codes at this levle
#codes <- CTV3_flags[(CTV3_flags$PARENT %in% LEVEL.1.CODES),]$flag;
cat("\n", length(LEVEL.2.CODES), "Level 2 codes: ", LEVEL.2.CODES, "\n")
for (j in LEVEL.2.CODES) { # go through the list of unique codes that are present at
this level
```

```
  parent <- CTV3_flags[(CTV3_flags$CHILD == j),]$PARENT # get the parent code
for the current code
  parent_flag <- CTV3_flags[(CTV3_flags$CHILD == parent),]$flag # get the flag
value for the parent code
  CTV3_flags[(CTV3_flags$CHILD == j),]$flag <- parent_flag  # assign the parent flag
value to the child code
  CTV3_flags[(CTV3_flags$CHILD == j),]$flag_auto <- parent_flag
  manualCode <- CTV3_flags_local[(CTV3_flags_local$CTV3 == j),]$flag
  cat("\n",CTV3_flags[(CTV3_flags$CHILD == j),]$CHILD,
CTV3_flags[(CTV3_flags$CHILD == j),]$CTERM,
CTV3_flags[(CTV3_flags$CHILD == j),]$flag_auto,
CTV3_flags[(CTV3_flags$CHILD == j),]$flag, manualCode);
}

codeAll <- CTV3_flags[!is.na(CTV3_flags$flag),];
codeAll2 <- codeAll[codeAll$CHILD %in% LEVEL.2.CODES,]
code0 <- codeAll2[(codeAll2$flag == 0),];
code1 <- codeAll2[(codeAll2$flag == 1),];
code2 <- codeAll2[(codeAll2$flag == 2),];

cat("\nCodes present at level2:", length(codeAll$CHILD), "; New codes introduced at
level2:", length(codeAll2$CHILD),"; flag0:", length(code0$CHILD),";
flag1:",length(code1$CHILD), "; flag2:",length(code2$CHILD), "sum: ",
(length(code0$CHILD)+length(code1$CHILD)+length(code2$CHILD)));

####################################    Level 3
####################################

#work on level 3 codes
LEVEL.3.CODES <- unique(CTV3_flags$PARENT.3); # list of unique "level 3" codes
(not many of these)
```

```
LEVEL.3.CODES <- LEVEL.3.CODES[!(LEVEL.3.CODES %in%
c(LEVEL.1.CODES, LEVEL.2.CODES))]
LEVEL.3.CODES <- LEVEL.3.CODES[(LEVEL.3.CODES %in%
unique_codes_list$CTV3)]
LEVEL.3.CODES <- sort(LEVEL.3.CODES);

#cat("\n", length(LEVEL.3.CODES), "Level 3 codes: ", LEVEL.3.CODES, "\n")
for (j in LEVEL.3.CODES) { # go through the list of unique codes that are present at
this level
  parent <- CTV3_flags[(CTV3_flags$CHILD == j),]$PARENT # get the parent code
for the current code
  parent_flag <- CTV3_flags[(CTV3_flags$CHILD == parent),]$flag # get the flag
value for the parent code
  CTV3_flags[(CTV3_flags$CHILD == j),]$flag <- parent_flag  # assign the parent flag
value to the child code
  CTV3_flags[(CTV3_flags$CHILD == j),]$flag_auto <- parent_flag
  manualCode <- CTV3_flags_local[(CTV3_flags_local$CTV3 == j),]$flag

  if (length(manualCode) > 0) {
    CTV3_flags[(CTV3_flags$CHILD == j),]$flag <- manualCode;
  }
  cat("\n",CTV3_flags[(CTV3_flags$CHILD == j),]$CHILD,
CTV3_flags[(CTV3_flags$CHILD == j),]$CTERM,
CTV3_flags[(CTV3_flags$CHILD == j),]$flag_auto,
CTV3_flags[(CTV3_flags$CHILD == j),]$flag, manualCode);

    #if (parent_flag != 0) {
  #  cat("\n",CTV3_flags[(CTV3_flags$CHILD == j),]$CHILD,
CTV3_flags[(CTV3_flags$CHILD == j),]$CTERM,
CTV3_flags[(CTV3_flags$CHILD == j),]$flag_auto,
CTV3_flags[(CTV3_flags$CHILD == j),]$flag);
```

```
  #cat("\n",CTV3_flags[(CTV3_flags$CHILD == j),]$CHILD,
CTV3_flags[(CTV3_flags$CHILD == j),]$CTERM,
CTV3_flags[(CTV3_flags$CHILD == j),]$PARENT,
CTV3_flags[(CTV3_flags$CHILD == j),]$PTERM, CTV3_flags[(CTV3_flags$CHILD
== j),]$flag_auto, CTV3_flags[(CTV3_flags$CHILD == j),]$flag);
  #}
}

codeAll <- CTV3_flags[!is.na(CTV3_flags$flag),];
codeAll3 <- codeAll[codeAll$CHILD %in% LEVEL.3.CODES,]
code0 <- codeAll3[(codeAll3$flag == 0),];
code1 <- codeAll3[(codeAll3$flag == 1),];
code2 <- codeAll3[(codeAll3$flag == 2),];

cat("\nAll codes present at level3:", length(codeAll$CHILD), "; New codes introduced
at level3:",length(codeAll3$CHILD),": flag0:", length(code0$CHILD),";
flag1:",length(code1$CHILD), "; flag2:",length(code2$CHILD), "sum: ",
(length(code0$CHILD)+length(code1$CHILD)+length(code2$CHILD)));

write.csv(code0, "level3code0.csv");write.csv(code1,
"level3code1.csv");write.csv(code2, "level3code2.csv");

####################################    Level 4
####################################

#work on level 4 codes

LEVEL.4.CODES <- unique(CTV3_flags$PARENT.4); # list of unique "level 4" codes
LEVEL.4.CODES <- LEVEL.4.CODES[!((LEVEL.4.CODES %in%
c(LEVEL.1.CODES, LEVEL.2.CODES, LEVEL.3.CODES)))];
LEVEL.4.CODES <- LEVEL.4.CODES[(LEVEL.4.CODES %in%
unique_codes_list$CTV3)]
```

```
LEVEL.4.CODES <- sort(LEVEL.4.CODES);

for (j in LEVEL.4.CODES) { # go through the list of unique codes that are present at
this level
  parent <- CTV3_flags[(CTV3_flags$CHILD == j),]$PARENT # get the parent code
for the current code
  parent_flag <- CTV3_flags[(CTV3_flags$CHILD == parent),]$flag # get the flag
value for the parent code
  CTV3_flags[(CTV3_flags$CHILD == j),]$flag <- parent_flag  # assign the parent flag
value to the child code
  CTV3_flags[(CTV3_flags$CHILD == j),]$flag_auto <- parent_flag
  manualCode <- CTV3_flags_local[(CTV3_flags_local$CTV3 == j),]$flag
  if (length(manualCode) > 0) {
    CTV3_flags[(CTV3_flags$CHILD == j),]$flag <- manualCode;
  }
  cat("\n",CTV3_flags[(CTV3_flags$CHILD == j),]$CHILD,
CTV3_flags[(CTV3_flags$CHILD == j),]$CTERM,
CTV3_flags[(CTV3_flags$CHILD == j),]$flag_auto,
CTV3_flags[(CTV3_flags$CHILD == j),]$flag, manualCode);
}



codeAll <- CTV3_flags[!is.na(CTV3_flags$flag),];
codeAll4 <- codeAll[codeAll$CHILD %in% LEVEL.4.CODES,]
code0 <- codeAll4[(codeAll4$flag == 0),];
code1 <- codeAll4[(codeAll4$flag == 1),];
code2 <- codeAll4[(codeAll4$flag == 2),];

cat("\nAll codes present at level4:", length(codeAll$CHILD), "; New codes introduced
at level4:",length(codeAll4$CHILD),": flag0:", length(code0$CHILD),";
flag1:",length(code1$CHILD), "; flag2:",length(code2$CHILD), "sum: ",
(length(code0$CHILD)+length(code1$CHILD)+length(code2$CHILD)));
```

```
write.csv(code0, "level4code0.csv");write.csv(code1,
"level4code1.csv");write.csv(code2, "level4code2.csv");

####################################    Level 5
####################################

#work on level 5 codes
LEVEL.5.CODES <- unique(CTV3_flags$PARENT.5); # list of unique "level 4" codes
LEVEL.5.CODES <- LEVEL.5.CODES[!((LEVEL.5.CODES %in%
c(LEVEL.1.CODES, LEVEL.2.CODES, LEVEL.3.CODES, LEVEL.4.CODES)))];
LEVEL.5.CODES <- LEVEL.5.CODES[(LEVEL.5.CODES %in%
unique_codes_list$CTV3)]

for (j in LEVEL.5.CODES) { # go through the list of unique codes that are present at
this level
  parent <- CTV3_flags[(CTV3_flags$CHILD == j),]$PARENT # get the parent code
for the current code
  parent_flag <- CTV3_flags[(CTV3_flags$CHILD == parent),]$flag # get the flag
value for the parent code
  CTV3_flags[(CTV3_flags$CHILD == j),]$flag <- parent_flag  # assign the parent flag
value to the child code
  CTV3_flags[(CTV3_flags$CHILD == j),]$flag_auto <- parent_flag
  manualCode <- CTV3_flags_local[(CTV3_flags_local$CTV3 == j),]$flag
  if (length(manualCode) > 0) {
    CTV3_flags[(CTV3_flags$CHILD == j),]$flag <- manualCode;
  }
  #cat("\n",CTV3_flags[(CTV3_flags$CHILD == j),]$CHILD,
CTV3_flags[(CTV3_flags$CHILD == j),]$CTERM,
CTV3_flags[(CTV3_flags$CHILD == j),]$flag_auto,
CTV3_flags[(CTV3_flags$CHILD == j),]$flag, manualCode);
}
```

```
codeAll <- CTV3_flags[!is.na(CTV3_flags$flag),];
codeAll5 <- codeAll[codeAll$CHILD %in% LEVEL.5.CODES,]
code0 <- codeAll5[(codeAll5$flag == 0),];
code1 <- codeAll5[(codeAll5$flag == 1),];
code2 <- codeAll5[(codeAll5$flag == 2),];

cat("\nAll codes present at level5:", length(codeAll$CHILD), "; New codes introduced
at level5:",length(codeAll5$CHILD),": flag0:", length(code0$CHILD),";
flag1:",length(code1$CHILD), "; flag2:",length(code2$CHILD), "sum: ",
(length(code0$CHILD)+length(code1$CHILD)+length(code2$CHILD)));

write.csv(code0, "level5code0.csv");
write.csv(code1, "level5code1.csv");
write.csv(code2, "level5code2.csv");
```

APPENDIX 5: CTV3 CODES AND SIGNIFICANCE FLAGS

A sample table of CTV3 codes together with their assigned significance flags is shown here. Code 0 implies an administration-type code that has little or no clinical significance and is not used for analysis. Code 1 implies a symptom-type code; code 2 implies a diagnosis of a condition. Both are used in the analysis.

CTV3 CODE	PARENT CODE	Significance flag	CTV3 CODE DESCRIPTION	PARENT CODE DESCRIPTION
XaBVJ	.....	0	Clinical findings	Read thesaurus
9....	.....	0	Administration	Read thesaurus
Xa22Y	.....	0	Operations, procedures and interventions	Read thesaurus
X78FG	.....	0	Staging and scales	Read thesaurus
0....	.....	0	Occupations	Read thesaurus
XE1Tr	22...	0	Adult health examination	General examination of patient
XM1Xs	22...	0	Screening - health check	General examination of patient
XE0qh	167..	1	Pale - symptom	Colour symptom
16AZ.	16A..	1	Stiff neck symptom NOS	Stiff neck symptom
16A..	16A2.	1	Stiff neck symptom	Stiff neck
17Z1.	17...	1	No respiratory symptoms	Respiratory symptoms
17ZZ.	17...	1	Respiratory symptom NOS	Respiratory symptoms
XE0qq	1738	2	Dyspnoea	Difficulty breathing
17Z..	17ZZ.	1	Respiratory	Respiratory

			symptoms NOS	symptom NOS
1822	182..	1	Central chest pain	Chest pain
X70Gv	A....	2	Bacterial disease	Infective disorder
X70Iu	A....	2	Viral disease	Infective disorder
X70OW	A....	2	Protozoal disease	Infective disorder
AB...	A....	2	Mycoses	Infective disorder
1829	182..	1	Retrosternal pain	Chest pain
1823	182..	1	Precordial pain	Chest pain
1828	182..	1	Atypical chest pain	Chest pain
PG52.	X78BM	2	Osteopetrosis	Dysplasia with increased bone density
Xa99T	X78Dz	2	Neurofibromatosis	Phakomatoses
PK5..	X78Dz	2	Tuberous sclerosis	Phakomatoses
X78EV	X78ES	2	Congenital malformation of thyroid gland	Congenital malformation of the endocrine glands
PJ63.	X78Ex	2	Turner's syndrome	Sex chromosome abnormality - female phenotype

APPENDIX 6: R CODE FOR CLUSTERING

Shown here is the body of the program used to determine the optium factors for clustering. The same basic code is used for the training and testing phases. In the training phase, the program loops over all options for the number of clusters and the options for CTV3 hierarchy level. In the testing phase, these values are fixed. Not shown are the header section of the program, which sets the working directory, reads in the local R functions, opens the required R libraries and reads in the conditions codelists.

```
for (condition in c("acuteSinusitis", "allergicRhinitis", "asthma", "anyCancer",
"prostateCancer", "stress", "T2Diabetes", "thyrotoxicosis","autism", "obesity",
"osteoarthritis", "praderWilli", "breastCancer",
"bronchitis","colonCancer","eczema","refluxDisease", "gastroparesis", "gout" )) {

  targetCode <- eval(parse(text = condition));

  for (minEventsCount in c(4)) {
    for (granularity_group in c(granularity_group_set)) {

      runNo <- runNo + 1;

      ########################################################
      # read in data files
      ########################################################

      cat("\n\nRun", runNo, "at", format(Sys.time(), date.format), "Condition:", condition,
"Maximum",nrows,"records; minimum events per record =", minEventsCount, ";
granularity level =", granularity_group, "\nreading data files ...");
      #read in the patient event histories. First build the file name to read. Use nrows to
limit the number of lines read
```

```
dataFileName <- "train50.csv";

#then read the first nrows rows from the record data file, or a random selection of nrows from the full records data file:
#concatDF <- read.csv(dataFileName, nrows=nrows, header=TRUE, fill=TRUE, stringsAsFactors = FALSE);
concatDF <- read.csv(dataFileName, header=TRUE, fill=TRUE, stringsAsFactors = FALSE);
concatDF <- concatDF[sample(nrow(concatDF), nrows), ];
concatDF[concatDF[,] == ""] <- NA; # convert empty string values (i.e. "no event") to NA

#drop columns that are all NA
#first get maximum number of codes at this level of granularity
max_events_count <- max(concatDF$unique.events.count);
#then drop the surplus columns and a couple of empty columns for safety:
concatDF <- concatDF[,-c((7+max_events_count):length(concatDF[1,]))];
concatDF[,(length(concatDF)+1):(length(concatDF)+2)] <- NA

#read in "include" flag look-up table
include_flags <- read.csv("include_flags_AHD.csv", header=TRUE,fill=TRUE,stringsAsFactors = FALSE);

############################################################
# end of read in data files
############################################################

############################################################
# do preparatory work on read data files
############################################################
```

```
    cat(" preparing data files ...");

    #get list of Read Codes present AT THIS LEVEL
    eventCodesIn <- unique(unlist(concatDF[,-c(1:6)])) # drop the first 6 columns of demographic information and leave only the event codes
    eventCodesIn <- eventCodesIn[!is.na(eventCodesIn)] #get final list of all event codes

    concatDF$mostRecentDiagnosisCode <- apply(concatDF,1,mostRecentDiagnosisCode) #find position in concatDF of most recent diagnosis code

    #drop the patients who have no diagnostic event - i.e. where the mostRecentDiagnosisCode is NA - since they are of no use to us
    concatDF <- subset(concatDF, !is.na(concatDF$mostRecentDiagnosisCode))
    #drop the patients who have a target diagnostic event but no other events - they are also of no use to us
    concatDF <- subset(concatDF, (concatDF$unique.events.count > minEventsCount))

    #get counts of number of event codes, number of patients, highest number of events for a patient
    number_of_eventCodesIn <- length(eventCodesIn) # how many event codes there in the data set at this level for this (sub)set of patients
    number_of_patientsIn <- length(concatDF$patid) # how many patients left after removing those with no events of interest
    number_of_patient_eventsIn <- max(concatDF$unique.events.count) #maximum number of events for a patient

    ################################################################
    # end of preparatory work on read data files
    ################################################################
```

```
#########################################################
# set up data frame ready for input into clustering
#########################################################

cat(" preparing data at higher Read Code level ...");

#need to flag up that the patient has the condition and remove the codes that mean they have the condition
#for each patient in concatDF, replace each event code with 1 (if in the condition list) or 0 (if not in the condition list)
concatDFisCondition <- apply(concatDF[,7:(7+number_of_patient_eventsIn)],2,TrueFalse)
#for each patient, get the total number of events that are in the condition list
concatDF$has.condition <- apply(concatDFisCondition,1,sum) >= 1;
#concatDF$has.condition <- concatDF$has.condition >= 1
#now remove the events that are in the condition list by replacing them with NA
# duplicate the event data frame and then replace event Codes with their ancestor codes
#concatDF_higher <- concatDF[,7:(length(concatDF[1,])-3)]; #probably don't need this now that the succeeding lines shuffle up the NA values...

for (i in 1:number_of_patientsIn) {
  concatDF[i,(which(concatDFisCondition[i,] == TRUE)+6)] <- NA
}

concatDF_higher <- concatDF[,7:(length(concatDF[1,])-3)]; #probably don't need this now that the succeeding lines shuffle up the NA values...
#concatDF_higher_old <- concatDF[,7:(length(concatDF[1,])-3)]; #probably don't need this now that the succeeding lines shuffle up the NA values...

concatDF_higher <- mapply(getHigherCode, concatDF[,7:(length(concatDF[1,])-3)],granularity_group);
```

```
for (i in 1:length(concatDF_higher[,1])) {
  concatDF_higher[i,duplicated(concatDF_higher[i,])] <- NA
}

#some of the codes in the event sequence are NA - remove these, shuffle up the existing event codes and fill with NA at the end of the event sequence
lentemp2 <- length(concatDF_higher[1,]);
for(i in 1:length(concatDF_higher[,1])) {
  temp1 <- concatDF_higher[i,];
  temp1 <- temp1[!is.na(temp1)] # try replacing these two lines with temp1 <- concatDF[!is.na(concatDF[i,7:(concatDF[i,]$unique.events.count+6)])]
  lentemp1 <- length(temp1);
  #concatDF_higher[i,] <- temp1;
  if (lentemp1 > 0) {
    concatDF_higher[i,1:lentemp1] <- temp1;
    concatDF_higher[i,(lentemp1+1):lentemp2] <- NA;
  }
}

#now put these higher level read codes back in to concatDF, replacing the lower level codes
concatDF[,7:(length(concatDF[1,])-3)] <- concatDF_higher



##########################################################
# do work on higher level data files
##########################################################

#get list of codes present AT THIS LEVEL
eventCodesGroup <- unique(unlist(concatDF[,7:(length(concatDF[1,])-3)])) # drop the first 6 columns of demographic information and leave only the event codes
eventCodesGroup <- eventCodesGroup[!is.na(eventCodesGroup)] #get final list of all event codes

#drop the patients who have no diagnostic event - i.e. where the mostRecentDiagnosisCode is NA
###### do we need to do this? we're no longer trying to predict events other than the condition of interest
concatDF <- subset(concatDF, !is.na(concatDF$mostRecentDiagnosisCode))
#drop the patients who have a target diagnostic event but no other events
concatDF <- subset(concatDF, (concatDF$unique.events.count > 1))  #### use this to control minimum number of events for a patient

#get counts of number of event codes, number of patients, highest number of events for a patient
number_of_eventCodesGroup <- length(eventCodesGroup) # how many event codes there in the data set at this level for this (sub)set of patients
number_of_patientsGroup <- length(concatDF$patid) # how many patients left after removing those with no events of interest
number_of_patient_eventsGroup <- max(concatDF$unique.events.count) #maximum number of events for a patient

if (nrow(concatDF) > nrows) {
  concatDF <- concatDF[sample(nrow(concatDF), nrows), ];
}
number_of_patientsGroup <- nrow(concatDF);

#########################################################
# end of work on higher level data files
#########################################################

#set up an empty data frame of patients(rows) by events (columns)
```

```
    eventTable <- setNames(data.frame(matrix(ncol =
(number_of_eventCodesGroup+4), nrow = number_of_patientsGroup)),
c("patid","gender","marital","age", eventCodesGroup))

    #populate the data frame
    eventTable[,1:4] <- concatDF[,1:4]; # assign patient IDs, gender, marital status, age,

    #now use the createBitVectors function to assign "present" or "absent" for each code
against each patient
    eventTable[,5:(number_of_eventCodesGroup+4)] <-
t(as.matrix(apply(concatDF[,7:(number_of_patient_eventsGroup+6)],1,createBitVector
s)))

    #drop the event codes which are only admin codes (i.e. the 'include flag' value is 0
for that event code)
    eventCodes_flags <- data.frame(includeFlagValue(eventCodesGroup),
stringsAsFactors = FALSE); #look up the 'include flag' value for each code
    eventCodes_and_flags <- cbind.data.frame(eventCodesGroup,eventCodes_flags,
stringsAsFactors = FALSE); # create local look-up table of event codes and their look-
up flags
    colnames(eventCodes_and_flags)[2] <- "flag";
    eventCodes_to_drop <- eventCodes_and_flags[eventCodes_and_flags[,2] == 0,1];
    eventTable <- eventTable[,!(names(eventTable) %in% eventCodes_to_drop)] #
remove the columns from the eventTable where column names (i.e. event codes) are in
'eventCodes_to_drop'
    number_of_eventCodesGroup <- number_of_eventCodesGroup-
length(eventCodes_to_drop) # need to adjust the number of event codes to show the
surviving codes

    eventTable$targetTrueFalse <- as.logical(concatDF$has.condition) # so we know
which patients have had the condition of interest and which haven't
    eventTable$marital <- NULL; # drop marital status
    if (useAge == TRUE) {
      ageMax <- max(eventTable$age); # set max age for age scaling
      eventTable$age <- eventTable$age/ageMax; # scale ages of patients 0 to 1
    } else eventTable$age <- 0; #
    eventTable[eventTable == TRUE] <- 1 #change value of 'TRUE' to '1'
    eventTable[eventTable == FALSE] <- 0 #change value of 'FALSE' to '0'

    if (useGender == TRUE) {
      eventTable[eventTable$gender == "F","gender"] <- 0; # set 0 for female
      eventTable[eventTable$gender == "M","gender"] <- 1; # set 1 for male
      eventTable$gender <- as.numeric(eventTable$gender);
    } else {
      eventTable$gender <- 0;
    }

    number_of_eventCodes <- length(eventTable[1,])-4

    ############################################################
    # end of set up data frame ready for input into clustering
    ############################################################

    ############################################################
    # get prevalence of condition
    ############################################################

    group_prevalence <-
sum(eventTable$targetTrueFalse)/length(eventTable$targetTrueFalse);

    cat("\nCondition:", condition,"; prevalence:", group_prevalence,"; number of
records with condition:",sum(eventTable$targetTrueFalse),"; total valid
records:",length(eventTable$targetTrueFalse));
```

```
    ################################################################
    # form distance matrix
    ################################################################


    n <- dim(eventTable[,(2:(number_of_eventCodesGroup+3))])[1]; # n is the number of rows in the event table
    nn_dup = matrix(0,n,n);
    nnTF_dup <- matrix(FALSE, n, n);

    cat("\n\nForming distance matrix using", dist.method)

    if (dist.method %in% c("euclidean", "manhattan", "cosine")) {
      cat(" from dist.matrix .... \nStart time:",format(Sys.time(), date.format));
      dist.mat <- dist.matrix(as.matrix(eventTable[,(2:(number_of_eventCodesGroup+3))]), method=dist.method, as.dist=TRUE); #get distance matrix
      #dist.mat <- dist.mat[2:(length(dist.mat)-1)]; # drop the "nearest" neighbour - it is our record of interest
      R_function <- "dist.matrix";
      cat("; End time: ",format(Sys.time(), date.format));
    } else if (dist.method %in% c("tanimoto", "dtw", "podani", "hamman", "michael", "faith", "mountford", "simpson", "fjaccard")) {
      cat(" from parDist .... \nStart time:",format(Sys.time(), date.format));
      dist.mat <- parDist(as.matrix(eventTable[,(2:(number_of_eventCodesGroup+3))],method=dist.method, threads=(no_cores)));
      R_function <- "parDist"
      cat("; End time: ",format(Sys.time(), date.format));
    } else if (dist.method %in% c("canberra")) {
      cat(" from dist .... \nStart time:",format(Sys.time(), date.format));
```

```
        dist.mat <-
dist(as.matrix(eventTable[,(2:(number_of_eventCodesGroup+3))],method=dist.method)
);
        R_function <- "parDist";
        cat("; End time: ",format(Sys.time(), date.format));
      } else if (dist.method %in% c("gower", "binomial", "mahalanobis", "cao", "chao"))
{
        cat(" from vegdist .... \nStart time:",format(Sys.time(), date.format));
        dist.mat <-
vegdist(as.matrix(eventTable[,(2:(number_of_eventCodesGroup+3))],method=dist.meth
od, binary=dist.binary, upper=FALSE));
        R_function <- "parDist"
        #cat("; End time: ",format(Sys.time(), date.format));
      } else if (dist.method %in% c("correlation", "spearman", "kendall")) {
        cat(" from Dist .... \nStart time:",format(Sys.time(), date.format));
        dist.mat <-
Dist(as.matrix(eventTable[,(2:(number_of_eventCodesGroup+3))],method=dist.method,
nbproc=no_cores, upper=FALSE));
        R_function <- "parDist"
        cat("; End time: ",format(Sys.time(), date.format));
      } else {
        cat("\nDistance method",dist.method,"not found: skipping")
      }

      dist.matrix <- as.matrix(dist.mat, nrow=n); # convert the distance matrix to a
standard matrix with n rows
```

```
   #########################################################
   # do clustering
   #########################################################

   for (hclust_method in cluster_method) {

    cat("\nCondition = ", condition, "; distance method = ", dist.method, "; Clustering using", hclust_method,"...");

    clust.res <- hclust(dist.mat,method=hclust_method)
    plot(clust.res);

    if (length(eventTable$patid) < clustersMax) clustersMax <- length(eventTable$patid)

    line <- data.frame(clusterNo=numeric(), level=numeric(), condition=character(), no_of_patients=numeric(), sensitivity=numeric(), specificity=numeric(), youden=numeric(), F1=numeric(), PPV=numeric(), FPV=numeric(), correct=numeric(), incorrect=numeric(), not_scored <- numeric(), stringsAsFactors=FALSE);
    count <- 0;
    cluster_set <- c(2:50, seq(60, (clustersMax-51), 10), (clustersMax-50):clustersMax);
    for (i in clusterNumbers) {
    for (i in clusters_set) {
     count <- count+1;
     temp_frame <- data.frame(temp=numeric());
     line <- data.frame(line, temp_frame);
     colnames(line)[(count+12)] <- paste("cluster",i,sep="");
    }
    temp_frame <- data.frame(total=numeric(),stringsAsFactors = FALSE);
    line <- data.frame(line, temp_frame);
    trackingTable <- eventTable[,c("patid", "targetTrueFalse")]
```

```
        trackingTableRisk <- eventTable[,c("patid", "targetTrueFalse")]

        ccF1 <- data.frame(computer_name=character(), date=character(), runNo =
integer(), dist_method=character(), dist.binary=logical(), hclust_method=character(),
condition=character(), nrows=integer(), validPatients=integer(),
minEVentsCount=integer(), level=integer(), clusters = integer(), prior=numeric(),
TP=integer(), TN=integer(), FP=integer(), FN=integer(), sensitivity=numeric(),
specificity=numeric(), F1=numeric(), adj_F1=numeric(), F2=numeric(),
adj_F2=numeric(), PPV=numeric(), NPV=numeric(), MCC=numeric(),
R_function=character(), stringsAsFactors=FALSE);

        cat("\nCycling through cluster numbers");

        for (clusterNo in clusters_set) {
          cat("\n\nStart cluster", clusterNo,"...")
          cluster.cut <- cutree(clust.res, k = clusterNo);   # k - an integer scalar or vector
with the desired number of groups;
          results <- data.frame(patid=eventTable$patid,
true.outcome=eventTable$targetTrueFalse, predicted.cluster=cluster.cut,
stringsAsFactors=FALSE);
          eventT <- table(results[,3], as.logical(results[,2]))

          #set up a data frame of patients(rows) by events (columns)
          cc <- data.frame(cluster = numeric(), ConditionNeg = numeric(), conditionPos =
numeric(), NoPatients = numeric(), prevalence = numeric())

          for (i in 1:clusterNo) {
            cc[i,1] <- rownames(eventT)[i];
            cc[i,2] <- eventT[i,1];
            cc[i,3] <- eventT[i,2];
            cc[i,4] <- sum(eventT[i,1:2]);
            cc[i,5] <- cc[i,3]/cc[i,4]; # this includes our patient of interest - shouldn't
```

```
}

cc$TrueFalsePred <- NA;

j <- i+1;
cc[j,1] <- "All";
cc[j,2] <- length(results[,2])-sum(results[,2]);
cc[j,3] <- sum(results[,2]);
cc[j,4] <- length(results[,2]);
cc[j,5] <- sum(results[,2])/length(results[,2]);

TP <- 0; TN <- 0; FP <- 0; FN <- 0; NPT <- 0; NPF <- 0;

for (patient in results[,1]) {
trueOutcome <- as.numeric(results[results$patid == patient,]$true.outcome);
   # did our patient have the condition?
predictedCluster <- results[results$patid == patient,]$predicted.cluster; # what cluster is our patient in?
rawPositive <- cc[predictedCluster,]$conditionPos;  # number of positive patients in the cluster - INCLUDING our patient of interest
rawNegative <- cc[predictedCluster,]$ConditionNeg;  # number of negative patients in the cluster - INCLUDING our patient of interest
rawNoPatients <- cc[predictedCluster,]$NoPatients;  # total number of patients in the cluster - INCLUDING our patient of interest
adjustedNoPatients <- rawNoPatients - 1; # number of patients in the cluster after removing our patient of interest
adjustedPositive <- rawPositive - trueOutcome; # number of positive patients in the cluster after removing our patient of interest
adjustedNegative <- rawNegative - (1 - trueOutcome); # number of negative patients in the cluster after removing our patient of interest
```

```
if (adjustedNoPatients > 0) { # if we have enough "other" patients in the cluster to calculate an in-cluster prevalence
 Ao <- adjustedPositive; Bo <- (sum(results[,2]) - rawPositive); Co <- adjustedNegative; Do <- length(results[,2]) - rawNoPatients - Bo;
 Ae <- (Ao+Bo)*(Ao+Co)/(Ao+Bo+Co+Do)
 Be <- (Ao+Bo)*(Bo+Do)/(Ao+Bo+Co+Do)
 Ce <- (Co+Do)*(Ao+Co)/(Ao+Bo+Co+Do)
 De <- (Co+Do)*(Bo+Do)/(Ao+Bo+Co+Do)
 CsqCalc <- (Ao-Ae)*(Ao-Ae)/Ae + (Bo-Be)*(Bo-Be)/Be + (Co-Ce)*(Co-Ce)/Ce + (Do-De)*(Do-De)/De;
 p <- 1;
 if (is.nan(CsqCalc)) {
  CsqCalc <- 0;
 }
 if (CsqCalc > 3.84) p <- 0.05;
 if (CsqCalc > 6.64) p <- 0.01;
 if (p < 0.05) {
  adjustedClusterPrediction <- 1;
 } else {
  adjustedClusterPrediction <- 0;
 }
}

cc[predictedCluster,]$TrueFalsePred <- adjustedClusterPrediction; # classifies a cluster's prediction as 'TRUE' if prevalence significantly > pop'n prevalence

if (trueOutcome == 1) { # if our patient of interest really did have the condition
 if (adjustedClusterPrediction == 1) { # and we predicted this
  TP <- TP + 1;
 } else if (adjustedClusterPrediction == 0) { # and we didn't predict it
  FN <- FN + 1;
 } else if (adjustedClusterPrediction == -1) {
       NPT <- NPT + 1; # no prediction made since our patient in a singleton cluster
      } else cat("adjustedClusterPredictionT = ", adjustedClusterPrediction);
     } else if (trueOutcome == 0) { # if our patient of interest didn't have the
condition
      if (adjustedClusterPrediction == 1) { # but we predicted that they did
       FP <- FP + 1;
      } else if (adjustedClusterPrediction == 0) { # or we correctly predicted that
they didn't
       TN <- TN + 1;
      } else if (adjustedClusterPrediction == -1) {
       NPF <- NPF + 1; # no prediction made since our patient in a singleton cluster
      } else cat("adjustedClusterPredictionF = ", adjustedClusterPrediction);
      #} # else cat("clusterPrediction = ", adjustedClusterPrediction); # safety check
     } # else cat("trueOutcome = ", trueOutcome); # safety check
    }
    results <- data.frame(results,
predicted.TrueFalse=get_prediction(results$predicted.cluster),
predicted.Risk=get_predictedRisk(results$predicted.cluster),
PosPatients=cc[results$predicted.cluster,]$conditionPos,
totalPatients=cc[results$predicted.cluster,]$NoPatients);

    cat("\nCondition:", condition, "; clustering method: ", hclust_method, ";
Clusters:", clusterNo, ";", nrows, "recs; level:", granularity_group);
    cat("\nTP:",TP, "; FP:", FP, "; TN: ", TN, "; FN:", FN,"; NP: ", (NPF+NPT),";");
    cat(dist.method,";",hclust_method);
    sensitivity <- TP / (TP + FN); # aka recall
    specificity <- TN / (TN + FP);
    cat("; sens:", format(round(sensitivity,3),nsmall=3), "; spec: ",
format(round(specificity,3),nsmall=3));

    PPV <- TP / (TP+ FP); # aka precision
    NPV <- TN / (TN + FN);
```

```
F1 <- 2 * (PPV * sensitivity) / (PPV + sensitivity)
cat("; F1 = ", F1);

#now add in the singleton clusters to the calculation
adj_TN <- TN + NPF;
adj_FN <- FN + NPT;
adj_PPV <- TP / (TP+ FP); # aka precision
adj_NPV <- adj_TN / (adj_TN + adj_FN);
adj_sensitivity <- TP / (TP + adj_FN); # aka recall
adj_specificity <- adj_TN / (adj_TN + FP);
# cat("; PPV = ", PPV, "; NPV = ", NPV);
adj_F1 <- 2 * (adj_PPV * adj_sensitivity) / (adj_PPV + adj_sensitivity)
if (is.nan(adj_F1)) adj_F1 <- 0;
cat("; adj_F = ", adj_F1);

ccF1[clusterNo,]$clusters <- clusterNo;
ccF1[clusterNo,]$TP <- TP;
ccF1[clusterNo,]$FP <- FP;
ccF1[clusterNo,]$TN <- TN;
ccF1[clusterNo,]$FN <- FN;
ccF1[clusterNo,]$sensitivity <- sensitivity;
ccF1[clusterNo,]$specificity <- specificity;
ccF1[clusterNo,]$F1 <- F1;
ccF1[clusterNo,]$computer_name <- computer_name;
ccF1[clusterNo,]$date <- date_today;
ccF1[clusterNo,]$nrows <- nrows;
ccF1[clusterNo,]$condition <- condition;
ccF1[clusterNo,]$dist.method <- dist.method;
ccF1[clusterNo,]$dist.binary <- dist.binary;
ccF1[clusterNo,]$hclust_method <- hclust_method;
ccF1[clusterNo,]$level <- granularity_group;
```

```
        beta <- 2;

        F2 <- 2 * ((1 +
beta*beta)*((ccF1[clusterNo,]$TP/(ccF1[clusterNo,]$TP+ccF1[clusterNo,]$FP)) *
ccF1[clusterNo,]$sensitivity)) /
(beta*beta*(ccF1[clusterNo,]$TP/(ccF1[clusterNo,]$TP+ccF1[clusterNo,]$FP)) +
ccF1[clusterNo,]$sensitivity);
        if (is.nan(F2)) F2 <- 0
        cat("; F2:",F2);
        adj_F2 <- 2 * ((1 + beta*beta)*((TP/(TP+FP)) * adj_sensitivity)) /
(beta*beta*(TP/(TP+FP)) + adj_sensitivity);
        cat("; adj_F2:",adj_F2);

        ccF1[clusterNo,]$adj_F1 <- adj_F1;
        ccF1[clusterNo,]$F2 <- F2;
        ccF1[clusterNo,]$adj_F2 <- adj_F2;
        ccF1[clusterNo,]$R_function <- "dist";

        MCC <- ((TP*TN)-(FP*FN)) / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN));
        write.table(t(c(computer_name, date_today, runNo, dist.method, dist.binary,
hclust_method, condition, nrows,length(eventTable$targetTrueFalse), minEventsCount,
granularity_group, clusterNo, group_prevalence, TP, TN, FP, FN, sensitivity,
specificity, F1, adj_F1, F2, adj_F2, MCC, PPV, NPV, MCC)), file=outputFileName,
append=TRUE,col.names=FALSE, row.names=FALSE, sep=",");

       }

       beta <- 2;
       ccF1$Fbeta <- 2 * ((1 + beta*beta)*((ccF1$TP/(ccF1$TP+ccF1$FP)) *
ccF1$sensitivity)) / (beta*beta*(ccF1$TP/(ccF1$TP+ccF1$FP)) + ccF1$sensitivity)

       ccF1$runNo <- runNo;
```

```
ccF1 <- ccF1[complete.cases(ccF1$computer_name),];

######################################################
# end of do clustering
#######################################################

sensitivity <- TP/(TP + FN); # aka recall
specificity <- TN/ (TN + FP);
PPV <- TP / (TP + FP); # aka precision
NPV <- TN / (TN + FN);
#TP[,j] <- TP; FP[,j] <- FP; TN[,j] <- TN; FN[,j] <- FN;
#prevPPV[,j] <- prevTPlocal / (prevTPlocal + prevFPlocal);
F1 <- 2 * (PPV * sensitivity) / (PPV + sensitivity);
#prevF2[,j] <- 2 * ((1 + beta*beta)*((TP/(TP+FP)) * TP / (TP + FN))) / (beta*beta*(TP/(TP+FP)) + TP / (TP + FN));
F2 <- ((1 + (beta*beta))*PPV * sensitivity) / ((beta*beta*PPV) + sensitivity);
MCC <- ((TP*TN)-(FP*FN)) / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

LRpos <- sensitivity/(1- specificity);
LRneg <- (1 - sensitivity)/specificity;
m <- matrix(c(TP, FN, FP, TN),2);
m_lr.ci <- calcLikelihoodRatio(m);
m_odds_ratio <- calcOddsRatio(TP, FP, FN, TN);

cat("\n", condition, ";", dist.method, "; level = ", granularity_group);
cat("; clusters:", clusters_set, "; TP:", TP, "; TN:", TN, "; FP:", FP, "; FN:", FN, "; total:", TP+FP+TN+FN);
cat("\nSensitivity:", sensitivity, "Specificity:", specificity, "F1:", F1, ": F2:", F2, "; MCC:", MCC, "; PPV:", PPV, "; NPV:", NPV);
cat("\nPos Likelihood Ratio =", m_lr.ci$lr.pos, "(",m_lr.ci$lower.pos, ",", m_lr.ci$upper.pos,")");
```

```
    cat("; Neg Likelihood Ratio =", m_lr.ci$lr.neg, "(",m_lr.ci$lower.neg, ",",
m_lr.ci$upper.neg,")")

    cat("\nOdds ratio = ", m_odds_ratio$OR, "; 95% CI = (", m_odds_ratio$LowerCI,
",", m_odds_ratio$UpperCI, ")");

    test.result <- "+";
    probs.pre.test <- group_prevalence;
    LR_pos <- round(m_lr.ci$lr.pos,2);
    if (is.nan(LR_pos)) LR_pos<-1;
    LR_neg <- round(m_lr.ci$lr.neg,2);
    if (is.nan(LR_neg)) LR_pos<-1;
    LR_pos_lower <- round(m_lr.ci$lower.pos,2);
    LR_pos_upper <- round(m_lr.ci$upper.pos,2);
    LR_neg_lower <- round(m_lr.ci$lower.neg,2);
    LR_neg_upper <- round(m_lr.ci$upper.neg,2);
    OR <- round(m_odds_ratio$OR,2);
    OR_lower <- round(m_odds_ratio$LowerCI,2);
    OR_upper <- round(m_odds_ratio$UpperCI,2);

   plotname <- paste("Fagan_test_clustering_test_random_", condition, "_",
clusters_set,"_",computer_name, ".jpg", sep="")
   jpeg(plotname, width = 6, height = 5, units = 'in', res = 300)

    opar <- par(no.readonly = T)
    on.exit(par(opar))
    par(mar = c(1.5, 6, 2, 6))
    stato <- ifelse(test.result == "+", "disease", "no disease")
    if (probs.pre.test > 1 | probs.pre.test < 0 | LR_pos < 0 | is.infinite(LR_pos) |
is.nan(LR_pos) | test.result %in% c("+", "-") == F) {
      cat("wrong values !!")
    } else {
     logits <- function(p) log(p/(1 - p))
    }
    logits.pre <- logits(probs.pre.test)
    logits.pos.post <- log(LR_pos) + logits.pre
    probs.pos.post.test <- exp(logits.pos.post)/(1 + exp(logits.pos.post))
    logits.pos.post_lower <- log(LR_pos_lower) + logits.pre
    probs.pos.post.test_lower <- exp(logits.pos.post_lower)/(1 +
exp(logits.pos.post_lower))
    logits.pos.post_upper <- log(LR_pos_upper) + logits.pre
    probs.pos.post.test_upper <- exp(logits.pos.post_upper)/(1 +
exp(logits.pos.post_upper))

    logits.neg.post <- log(LR_neg) + logits.pre
    probs.neg.post.test <- exp(logits.neg.post)/(1 + exp(logits.neg.post))
    logits.neg.post_lower <- log(LR_neg_lower) + logits.pre
    probs.neg.post.test_lower <- exp(logits.neg.post_lower)/(1 +
exp(logits.neg.post_lower))
    logits.neg.post_upper <- log(LR_neg_upper) + logits.pre
    probs.neg.post.test_upper <- exp(logits.neg.post_upper)/(1 +
exp(logits.neg.post_upper))

    compl.logit.pre <- logits(1 - probs.pre.test)
    compl.logit.post <- logits(1 - probs.pre.test)

    LR.vec <- c(0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2,
          0.5, 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000)
    prob.vec <- c(0.001, 0.002, 0.003, 0.005, 0.007, 0.01, 0.02,
           0.03, 0.05, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
           0.8, 0.9, 0.93, 0.95, 0.97, 0.98, 0.99, 0.993, 0.995,
           0.997, 0.998, 0.999)
    plot(0, 0, type = "n", ylim = range(logits(prob.vec)), axes = F,
      xlab = "", ylab = "")
```

```
axis(2, rev(logits(prob.vec)), 100 * prob.vec, pos = -1,
    las = 1, cex.axis = 0.7)
axis(2, rev(logits(prob.vec)), 100 * prob.vec, pos = -1,
    tck = 0.03, labels = F)
axis(4, logits(prob.vec), 100 * prob.vec, pos = 1, las = 1,
    cex.axis = 0.7)
axis(4, logits(prob.vec), 100 * prob.vec, pos = 1, tck = 0.03,
    labels = F)
axis(2, log(LR.vec[1:10])/2, LR.vec[1:10], pos = 0, las = 1,
    cex.axis = 0.7)
axis(2, log(LR.vec[1:10])/2, LR.vec[1:10], pos = 0, tck = 0.03,
    labels = F)
axis(4, log(LR.vec[10:19])/2, LR.vec[10:19], pos = 0, las = 1,
    cex.axis = 0.7)
axis(4, log(LR.vec[10:19])/2, LR.vec[10:19], pos = 0, tck = 0.03,
    labels = F)
text(0, 4.5, "Likelihood ratio", cex = 1.2)
segments(-1, compl.logit.pre, 1, logits.pos.post, lwd = 1.5,
       col = 2)
segments(-1, compl.logit.pre, 1, logits.pos.post_lower, lwd = 1.5,
       col = 2)
segments(-1, compl.logit.pre, 1, logits.pos.post_upper, lwd = 1.5,
       col = 2)


segments(-1, compl.logit.pre, 1, logits.neg.post, lwd = 1.5,
       col = 3)
segments(-1, compl.logit.pre, 1, logits.neg.post_lower, lwd = 1.5,
       col = 3)
segments(-1, compl.logit.pre, 1, logits.neg.post_upper, lwd = 1.5,
       col = 3)
```

```
        x_neg <- c(-1, -1, 1, 1);
        y_neg <- c(compl.logit.pre, compl.logit.pre, logits.neg.post_lower,
logits.neg.post_upper);
        polygon(x_neg,y_neg, col="green");

        x_pos <- c(-1, -1, 1, 1);
        y_pos <- c(compl.logit.pre, compl.logit.pre, logits.pos.post_lower,
logits.pos.post_upper);
        polygon(x_pos,y_pos, col="red");

        segments(-1, compl.logit.pre, 1, logits(probs.pre.test), lwd = 1.5, col = 1)

        mtext(side = 2, text = "Prior probability(%)", line = 2,cex = 1.2)
        mtext(side = 4, text = "Posterior probability(%)", line = 2, cex = 1.2, las = 3)

        title(main=condition);
        text(0, -6.3, paste("Prior prob. of disease =", round(100 * probs.pre.test, 2), "% \n",
                "Post test prob. of disease+ =", ifelse(test.result == "+", round(100 *
probs.pos.post.test,2), round(100 * (1 - probs.pos.post.test), 2)), "%", "\n",
                "Likelihood ratio+ ", "=", round(LR_pos,
2),"(",LR_pos_lower,",",LR_pos_upper,")", "\n",
                "Likelihood ratio- ", "=", round(LR_neg, 2),"(", LR_neg_lower,",",
LR_neg_upper,")", "\n",
               "Odds ratio = ", OR, "(", OR_lower, ",", OR_upper, ")","\n"), cex = 0.7)
        dev.off()
        if (writeColNames == TRUE) {
          write.table(t(c("condition","dist_method",
"sampleSize","validPatients","minEventsCount", "granularityLevel","k", "Prior",
"Posterior_pos", "Post_pos_lower", "Post_pos_upper",  "Posterior_neg",
"Post_neg_lower", "Post_neg_upper", "TP", "TN", "FP", "FN", "Sensitivity",
"Specificity", "F1","F2", "MCC", "PPV", "NPV", "LR_pos", "LR_pos_lower",
"LR_pos_upper", "LR_neg", "LR_neg_lower", "LR_neg_upper","OR", "OR_lower",
"OR_upper")),file=outputFileName,append=TRUE,col.names=FALSE,
row.names=FALSE, sep=",");
          writeColNames <- FALSE;
        }
      }
      }
    }
  }
```

APPENDIX 7: R CODE FOR NEAREST NEIGHBOURS

Using the nearest neighbours method used the same basic R code as for the clustering method, replacing the clustering section of the clustering program shown in Appendix5 with distance matrix and nearest neighbours code. Those sections only are shown here.

The same basic code is used for the training and testing phases. In the training phase, the program loops over all options for the value of k and the options for CTV3 hierarchy level. In the testing phase, these values are fixed.

```
  ##########################################################
   # nearest neighbours
   ##########################################################

    cat("\nCalculating nearest neighbours...");
    # for each record, order its nearest neighbours best on the distance matrix.        for
(i in 1:n) {
     nn_dup[i,] <- order(dist.matrix[i,]);
    }
    eventTable$generatedID <- 1:n;
    outcomes <- eventTable$targetTrueFalse;
    names(outcomes) <- eventTable$generatedID;
    # for each record, generate a table of "has condition" or "does not have the
condition" for each of a record's neighbours
    for (j in 1:n) {
     nnTF_dup[,j] <- outcomes[nn_dup[,j]];
    }

    ##########################################################
    # end do nearest neighbours
    ##########################################################
```

APPENDIX 8: APPROVALS

CITY UNIVERSITY
LONDON

Senate Research Ethics Committee

Application for Approval of Research Involving Human Participants

Please tick the box for which Committee you are submitting your application to

□	Senate Research Ethics Committee
□	Cass Business School
□	School of Arts & School of Social Sciences Research Ethics Committee
□	School of Health Sciences Research Ethics Committee
☒	School of Informatics
□	Learning Development Centre

For **Senate** applications: return one original and eight additional hardcopies of the completed form and any accompanying documents to Anna Ramberg, Secretary to Senate Research Ethics Committee, University Research Office, Northampton Square, London, EC1V 0HB. Please also email an electronic copy to [redacted] (indicating the names of those signing the hard copy).

For **School of Arts & School of Social Sciences** Research Ethics Committee submit a single copy of the application form and all supporting documentation to your Department's Research and Ethics Committee by email.

For **School of Health Sciences** applications: submit all forms (including the Research Registration form) electronically (in Word format in a single document) to

[redacted]

For **School of Informatics** applications: a single copy of the application form and all supporting documents should be emailed to Stephanie Wilson[redacted]

For **Learning Development Centre** a single copy of the application form and all the supporting documentations should be emailed to Pam Parker [redacted]

Refer to the separate guidelines while completing this form.

PLEASE NOTE

- Please determine whether an application is required by going through the checklist before filling out this form.
- Ethical approval **MUST** be obtained before any research involving human participants is undertaken. Failure to do so may result in disciplinary procedures being instigated, and you will not be covered by the University's indemnity if you do not have approval in place.
- You should have completed <u>every</u> section of the form
- The Signature Sections <u>must</u> be completed by the Principal Investigator (the supervisor and the student if it is a student project)

Project Title:
Self-reported health histories via anonymous web survey
Short Project Title (no more than 80 characters):
Self-reported health histories via anonymous web survey
Name of Principal Investigator(s) (*all* students are require to apply jointly with their supervisor and all correspondence will be with the supervisor):
Jonathan Turner Dr Peter Weller

Post Held (including staff/student number):

Jonathan Turner:[redacted]

Dr Peter Weller

Department(s)/School(s) involved at City University London:

Centre for Health Informatics

If this is part of a degree please specify type of degree and year

PhD Health Informatics year 3

Date of Submission of Application:

2/5/2014

1. Information for Non-Experts

Lay Title (no more than 80 characters)

Self-reported health histories via anonymous web survey

Lay Summary / Plain Language Statement (no more than 400 words)

This research project is intended to collect anonymous information on individuals' recollections of their health events throughout their life, for the purpose of (i) comparing individuals’ aggregated recollections of events to those stored in average health records and (ii) to see whether such individual-recalled event data can be used to modulate predictions of future health events.

Survey respondents will be invited to list, to the best of their recollection and without historical time limit, personal health events and ongoing conditions including, but not limited to, those events that were reported to or required the intervention of a general practitioner or other healthcare professional.

Respondents will also be invited to note their age group, weekly exercise habits, smoking status and alcohol consumption, and country of birth. In order to preserve anonymity, respondents will not be asked for their name, exact age or current country of residence.

Respondents who know the principal investigator will be discouraged from returning their information (they will be warned that by doing so they may inadvertently reveal medical conditions that they would prefer not to) but, given the anonymous nature of the process, cannot be prevented from participating. However, the investigators undertake not to attempt to re-identify any individuals from their submitted data.

Once data have been collected, they will be used:

i) To see how individuals' recollections of medical conditions compare, on average, to the quantity and detail of events typically stored in general practice records;

ii) As input to a health event prediction algorithm, in order to see whether patient-recalled data is of adequate quality to have practical use in such an algorithm.

2. Applicant Details

This project involves:

(tick as many as apply)

□	Staff Research	☒	Doctoral Student

□	Undergraduate	□	M-level Project
□	Externally funded	□	External investigators
□	Collaboration	□	Other
Provide details of collaboration and/or other			

Address for correspondence (including email address and telephone number) (Principal Investigator)

Jonathan Turner, Centre for Health Informatics, City University London, Northampton Square, London EC1V 0HB

[redacted]

(No City University telephone number)

Other staff members involved

Title, Name & Staff Number	Post	Dept & School	Phone	Email
Dr Peter Weller	Senior Lecturer	Computer Science / MCSE	[redacted]	[redacted]

All students involved in carrying out the investigation

Name & Student Number	Course / Year	Dept & School	Email

Jonathan Turner [redacted]	PhD Health Informatics year 3	Centre for Health Informatics, School of Informatics	[redacted]

External co-investigators

Title & Name	Post	Institution	Phone	Email
-				

Please describe the role(s) of all the investigators including all student(s)/external co-investigator(s) in the project, especially with regards to interaction with study participants.

Jonathan Turner has designed the study, will place the questionnaire form on the web and will be responsible for the promotion of the study, data collection and analysis and participant anonymity.

If external investigators are involved, please provide details of their indemnity cover.

No external investigators are involved

Application Details

2.1 Is this application being submitted to another ethics committee, or has it been previously submitted to an ethics committee? *This includes an NHS local Research Ethics Committee or a City University London School Research Ethics Committee or any other institutional committee or collaborating partners or research site.* (See the guidelines for more information on research involving NHS staff/patients/ premises.)

YES □ **NO** ☒

If yes, please provide details for the Secretary for the relevant authority/committee, as well as copies of any correspondence setting out conditions of approval.

-

2.2 If any part of the investigation will be carried out under the auspices of an outside organisation, e.g. a teaching hospital, please give details and address of organisation.

N/A

2.3 Other approvals required – has permission to conduct research in, at or through another institution or organisation been obtained? YES □ NO ☒

If yes, please provide details and include correspondence

2.4 Is any part of this research project being considered by another research ethics committee? YES □ NO ☒

If yes, please give details and justification for going to separate committees, and attach correspondence and outcome

2.5 Duration of Project

Start date: 1/6/2014 Estimated end date:1/10/2015

Funding Details

2.6 Please provide details of the source of financial support (if any) for the proposed investigation.

No expenses are expected to be incurred by this project

2.6a Total amount of funding being sought: 0

2.6b Has funding been approved? YES □ NO □

If no, please provide details of when the outcome can be expected

2.6c Does the funding body have any requirements regarding retention, access and storage of the data? YES □ NO □

If yes, please provide details

International Research

2.7 Is any part of the research taking place outside of England/Wales? (if not go to section 3) YES □ NO ☒

If yes, please provide details of where

Respondents completing the web form could be outside England/Wales – there are no geographic restrictions. Data will be collected, stored and processed in England.

2.7a Have you identified and complied with all local requirements concerning ethical approval & research governance*? YES □ NO □

2.7b Please provide details of the local requirements, including contact information.

2.7c Please give contact details of a local person identified to field initial complaints local so the participants can complain without having to write to or telephone the UK

All contact with participants will be via a web page which will contain information on how to contact the project team -likely to be the email address for Jonathan Turner at City University London[redacted]

*Please note many countries require local ethical approval or registration of research projects, further some require specific research visas. If you do not abide by the local rules of the host country you will invalidate your ethical approval from City University London, and may run the risk of legal action within the host country.

3. Project Details

3.1 Provide the background, aim and explanation for the proposed research.

This research project is intended to collect anonymous information on individuals' recollections of their health events throughout their life. Data collected will be used

i) To see how individuals' recollections of medical conditions compare, on average, to the quantity and detail of events typically stored in general practice records;

ii) As input to a health event prediction algorithm, in order to see whether patient-recalled data is of adequate quality to have practical use in such an algorithm.

Survey respondents will be invited to list, to the best of their recollection and without historical time limit, personal health events and ongoing conditions including, but not limited to, those events that were reported to or required the intervention of a general practitioner or other healthcare professional.

Respondents will also be invited to note their age group, weekly exercise habits, smoking status and alcohol consumption, and country of birth. In order to preserve anonymity, respondents will not be asked for their name, exact age or current country of residence.

3.2 Provide a summary and brief explanation of the design, methodology and plan for analysis that you propose to use.

Individuals will be asked to provide a few items of personal information - age group, smoking status, alcohol consumption, weekly exercise - and a list of health events that they can recall suffering from, together with the age at which they suffered these events.

Data will be collected anonymously by convenience sampling via a short web questionnaire.

Once data have been collected they will be analysed
i) by indication of smoking status, alcohol consumption, exercise quantity against conditions recorded by participants and known to be linked to these factors. This is to help understand how the sample compares to the general population.
ii) against those recorded in formal general practice records, to see how well the individuals' recollection of their health conditions compares with the formal records, for average number of conditions recorded per individual and distribution of recorded conditions (e.g. are more serious conditions more likely to be recalled?).
iii) as input to a health event predictor algorithm, to see how well predictions based on patient-supplied data perform compared to predictions based on formal medical records.

Comparisons will be at an aggregated level, i.e. , i.e. we will be looking to see how condition prevalences indicated by the survey responses compare to those indicated by published data, e.g. from the Quality and Outcomes Framework, Practice Fusion Insight or other published work (e.g. Blak et al, Generalisability of The Health Improvement Network (THIN) database, Informatics in Primary Care 2011).
Individuals providing information to the survey will not be identified in the study and it will not be possible to link them to particular GP records.

3.3 Please explain your plans for dissemination, including whether participants will be provided with any information on the findings or outcomes of the project.

This project forms part of Jonathan Turner's PhD work and results will be included there. Results will also be made available on the website used for data collection and a paper will be submitted to an appropriate journal or conference. It will not be possible to send results directly to participants due to the anonymous nature of data collection.

3.4 What do you consider are the ethical issues associated with conducting this research and how do you propose to address them?

Ethical issues are primarily around collection of personal, individual data. For this project, survey respondents are asked not to identify themselves; potential respondents are asked not to participate if they feel that they know the investigators in order to avoid the possibility of the investigators having the potential to identify participants by some unique combination of information supplied.

3.5 How is the research intended to benefit the participants, third parties and/or local community?

The research will not directly benefit the participants but may provide benefits to future patients, to the wider community and a general contribution to knowledge.

3.6a Will invasive procedures (for example medical or surgical) be used?

YES □ NO ☒

3.6b If yes, what precautions will you take to minimise any potential harm?

3.7a Will intrusive procedures (for example psychological or social) be used?

YES □ NO ☒

3.7b If yes, what precautions will you take to minimise any potential harm?

3.8a In the course of the investigation might pain, discomfort (including psychological discomfort), inconvenience or danger be caused? YES □ NO ☒

3.8b If yes, what precautions will you take to minimise any potential harm?

3.9 Please describe the nature, duration and frequency of the procedures?

4. Information on participants

4.1a How many participants will be involved?

The number of participants will be limited by time, dissemination speed of the survey invitation and response rate, but there is no absolute limit at which recruitment will halt. It is expected that meaningful results will be obtained once 50 responses are received.

This has been calculated by use of the the sample size calculator at http://www.raosoft.com/samplesize.html, using 10% margin of error, 95% confidence, 20,000 population size and 10% response distribution (very roughly the proportion of diabetes or hypertension in the general Western population), which gave a sample size of 35, which I rounded up to 50 to allow for invalid or incomplete questionnaires).

4.1b What is the age group and gender of the participants?

Participants are asked if they are 18 years old or older and excluded if they are younger than this. Not limited by gender.

4.1c Explain how you will determine your sample size and the selection criteria you will be using. Specify inclusion and exclusion criteria. If exclusion of participants is made on the basis of age, gender, ethnicity, race, disability, sexuality, religion or any other factor, please explain and justify why.

Inclusion criteria: all adults who have access to the world wide web.
Exclusion criteria: There are no exclusion criteria. Participants are not asked for their ethnicity, race, sexuality, religion or whether they have any disabilities.

4.2 How are the participants to be identified, approached and recruited, and by whom?

Participants are self-selecting, approached via general postings on Facebook, Twitter, LinkedIn and emails to general or professional discussion groups (specifically excluding those used by individuals to discuss medical conditions or care). Directly targeted invitations to named individuals will not be used.

4.3 Describe the procedure that will be used when seeking and obtaining consent, including when consent will obtained. Include details of who will obtain the consent, how are you intending to arrange for a copy of the signed consent form for the participants, when will they receive it and how long the participants have between receiving information about the study and giving consent.

Consent will be asked for at the time of participation in the survey. Due to the anonymous nature of the survey, participants will not be asked for a signature, merely to tick a box saying that they give their consent for their anonymous response to be used.

4.4 How will the participant's physical and mental suitability for participation be assessed? Are there any issues related to the ability of participants to give informed consent themselves or are you relying on gatekeepers on their behalf?

Due to the anonymous nature of the survey, it is not possible to assess participants' suitability for participation.

4.5 Are there any special pressures that might make it difficult to refuse to take part in the study? Are any of the potential participants in a dependent relationship with any of the investigators (for instance student, colleague or employee) particularly those involved in recruiting for or conducting the project?

Responses to the survey are anonymous and it will not be possible to ascertain whether any individual has or has not participated. There are no pressures on individuals to participate. Individuals known to the principal investigator are encouraged not to participate for the reasons outlined in 3.4

4.6 Will the participant's doctor be notified? YES □ NO ☒

(If so, provide a sample letter to the subject's GP.)

4.7 What procedures are in place for the appropriate referral of a study participant who discloses an emotional, psychological, health, education or other issue during the course of the research or is identified by the researcher to have such a need?

It will not be possible to identify any study participants who have need of referral due to the anonymous nature of the survey and so such procedures are not possible.

4.8 What steps will be taken to safeguard the participants from over-research? (I.e. to ensure that the participants are not being used in multiple research project.)

Participants are self-selecting from the general population and so it is assumed that they are comfortable with their participation.

4.9 Where will the research take place?

Data will be collected and analysed at City University London but participants will complete the survey over the world wide web from any location that is convenient for them.

4.10 What health and safety issues, if any, are there to consider?

There are no health and safety issues for participants or investigators beyond their usual use of networked computers within the City University London working space.

4.11 How have you addressed the health and safety concerns of the participants, researchers and any other people impacted by this study? (This includes research involving going into participants' homes.)

There are no health and safety issues for participants or investigators beyond their usual use of networked computers.

4.12 It is a University requirement that an at least an initial assessment of risk is undertaken for all research and if necessary a more detailed risk assessment be carried out. Has a risk assessment been undertaken?* YES ☒ NO □

4.13 Are you offering any incentives or rewards for participating? YES □ NO ☒

If yes please give details

*Note that it is the Committee’s prerogative to ask to view risk assessments.

5. Vulnerable groups

5.1 Will persons from any of the following groups be participating in the study? (if not go to section 6)

Adults without capacity to consent	□
Children under the age of 18	□
Those with learning disabilities	□
Prisoners	□
Vulnerable adults	□
Young offenders (16-21 years)	□
Those who would be considered to have a particular dependent relationship with the investigator (e.g. those in care homes, students, employees, colleagues)	□

5.2 Will you be recruiting or have direct contact with any children under the age of 18?

YES □ NO ☒

5.2a If yes, please give details of the child protection procedures you propose to adopt should there be any evidence of or suspicion of harm (physical, emotional or sexual) to a young person. Include a referral protocol identifying what to do and who should be contacted.

5.2b Please give details of how you propose to ensure the well-being of the young person, particularly with respect to ensuring that they do not feel pressured to take part in the research and that they are free to withdraw from the study without any prejudice to themselves at anytime.

5.3 Will you be recruiting or have direct contact with vulnerable adults? YES □ NO ☒

5.3a If yes, please give details of the protection procedures you propose to adopt should there be any evidence of or suspicion of harm (physical, emotional or sexual) to a vulnerable adult. Include a referral protocol identifying what to do and who should be contacted.

5.3b Please give details of how you propose to ensure the well-being of the vulnerable adult, particularly with respect to ensuring that they do not feel pressured to take part in the research and that they are free to withdraw from the study without any prejudice to themselves at anytime. You should indicate how you intend to ascertain that person's views and wishes.

Although vulnerable adults are not being targeted in this survey, a proportion of the general population being targeted may be classed as vulnerable. Information provided with the survey will make it clear that (a) the survey is anonymous, (b) that it does not form part of their medical care, (c) if they have concerns about their health they should consult their general practitioner or equivalent and (d) they can abandon participation in the survey at any time without and ill-effect on themselves. Any individual's survey results will only be made available to the researchers once the individual has completed the study by clicking the 'complete and send' button on their web browser; at this point it will be assumed that they are comfortable with sending the data they have entered.

5.3c Please give details of any City staff or students who will have contact with vulnerable adults and/or will have contact with young people (under the age of 18) and details of current (within the last 3 years) City University London Disclosure and Barring check.

Name	Dept & School	Student/Staff Number	Date of DBS	Type of disclosure

5.3d Please give details of any non-City staff or students who will have contact with vulnerable adults and/or will have contact with young people (under the age of 18) and details of current (within the last 3 years) Disclosure and Barring check.

Name	Institution	Address of	Date of DBS	Type of

		organisation that requested the disclosure		disclosure

5.4 Will you be recruiting any participants who fall under the Mental Capacity Act 2005? YES □ NO ☒

If so you MUST get approval from an NHS NRES approved committee (see separate guidelines for more information).

6. Data Collection

6.1a Please indicate which of the following you will be using to collect your data
Please tick all that apply

Questionnaire	☒
Interviews	□
Participant observation	□
Focus groups	□
Audio/digital-recording interviewees or events	□
Video recording	□
Physiological measurements	□
Quantitative research (please provide details)	□
Other	□
Please give details	

6.1b What steps, if any, will be taken to safeguard the confidentiality of the participants (including companies)?

Individuals will participate in the research by completing an on-line questionnaire. They are not asked for their name or location and are asked for their age group, not precise age. IP numbers are neither captured nor stored.

6.1c If you are using interviews or focus groups, please provide a topic guide

7. Confidentiality and Data Handling

7.1a Will the research involve:

complete anonymity of participants (i.e. researchers will not meet, or know the identity of participants, as participants, as participants are a part of a random sample and are required to return responses with no form of personal identification)**?**	☒
anonymised sample or data (i.e. an *irreversible* process whereby identifiers are removed from data and replaced by a code, with no record retained of how the code relates to the identifiers. It is then impossible to identify the individual to whom the sample of information relates)**?**	□
de-identified samples or data (i.e. a *reversible* process whereby identifiers are replaced by a code, to which the researcher retains the key, in a secure location)?	□
subjects being referred to by pseudonym in any publication arising from the research?	□
any other method of protecting the privacy of	□

participants? (e.g. use of direct quotes with specific permission only; use of real name with specific, written permission only)	

Please give details of 'any other method of protecting the privacy of participants' is used

7.1b Which of the following methods of assuring confidentiality of data will be implemented?

Please tick all that apply

data to be kept in a locked filing cabinet	□
data and identifiers to be kept in separate, locked filing cabinets	□
access to computer files to be available by password only	☒
storage at City University London	□
stored at other site	□

If stored at another site, please give details	

7.1c Who will have access to the data?

Access by named researcher(s) only YES □ NO ☒

Access by people other than named researcher(s) **YES** ☒ **NO** □

If people other than the named researcher(s), please explain by whom and for what purpose

In addition to the named researchers, data will be available on request to the 2nd supervisor and to PhD examiners.

7.2a Is the data intended for reuse or to be shared as part of longitudinal research?

YES □ NO ☒

7.2b Is the data intended for reuse or to be shared as part of a different/wider research project now, or in the future? YES □ NO ☒

7.2c Does the funding body (e.g. ESRC) require that the data be stored and made available for reuse/sharing? YES □ NO ☒

7.2d If you have responded yes to any of the questions above, explain how you are intending to obtain explicit consent for the reuse and/or sharing of the data.

7.3 Retention and Destruction of Data

7.3a Does the funding body or your professional organisation/affiliation place obligations or recommendations on the retention and destruction of research data?

YES □ NO ☒

If yes, what are your affiliations/funding and what are the requirements? (If no, please refer to University guidelines on retention.)

7.3bHow long are you intending to keep the data?

Twelve months beyond the end date of the PhD research (as noted in section 2.5)

7.3c How are you intending to destroy the data after this period?

Deletion of data files

8. Curriculum Vitae

CV OF APPLICANTS (Please duplicate this page for each applicant, including external persons and students involved.)

NAME:	Jonathan Turner
CURRENT POST (from)	October 2011
Title of Post:	Research Student
Department:	Centre for Health Informatics
Is your post funded for the duration of this proposal?	N/A
Funding source (if not City University London)	Self
Please give a summary of your training/experience that is relevant to this research project	
I am a current PhD research student in the Centre for Health Informatics at City University London. I have previously obtained a MSc in Health Informatics from City University London. I have worked for both PACSnet and ImPACT, medical device evaluation units based at St George's Hospital London who provided reports on medical devices for NICE and the MHRA. This work included working on clinial information systems containing real patient information. I have also worked as manager of PACS and RIS	

systems in hospitals, again working with real patient data. These posts have ensured an appreciation of the need and requirements for data confidentiality.

I am registered with the Health Professions Council and am a full member of both the Institute of Physics and Engineering in Medicine and the Institute of Physics, these memberships carrying with them an obligation of professional behaviour.

NAME:	Dr Peter Weller
CURRENT POST (from)	2003
Title of Post:	Senior Lecturer
Department:	Computer Science
Is your post funded for the duration of this proposal?	Yes
Funding source (if not City University London)	n/a

Please give a summary of your training/experience that is relevant to this research project

I have 16 years experience in the Health Informatics arena and in that time have carried out a large number of projects involving health data collection, including data from operating theatres and A&E departments. Recently I was awarded an NHS Innovation challenge award for the analysis of a web based questionnaire on Carpal tunnel syndrome.

8.1 Supervisor’s statement on the student’s skills and ability to carry out the proposed research, as well as the merits of the research topic (up to 500 words)

I fully support this application. Jonathan has strong experience in analysing patient records both from his MSc project work and practical experience in the NHS. The project is timely and has a great potential to provide clinicians (mainly GPs) with a novel tool for predicting possible illness for individual patients. The use of patient data is essential for this work so it is requested that this application be approved.

Supervisor’s Signature	[redacted]
Print Name	[redacted]

9. Participant Information Sheet and 10. Consent Form

Please use the templates provided below for the Participant Information Sheet and Consent Form. They should be used for all research projects and by both staff and students. Note that there are occasions when you will need to include additional information, or make slight changes to the standard text – more information can be found under the **application guidelines**.

11. Additional Information

12. Declarations by Investigator(s)

- I certify that to the best of my knowledge the information given above, together with any accompanying information, is complete and correct.

- I have read the University’s guidelines on human research ethics, and accept the responsibility for the conduct of the procedures set out in the attached application.
- I have attempted to identify all risks related to the research that may arise in conducting the project.
- I understand that **no** research work involving human participants or data can commence until **full** ethical approval has been given

	Print Name	Signature
Principal Investigator(s) (student and supervisor if student project)	Jonathan Turner Dr Peter Weller	[redacted]
Associate Dean for Research (or equivalent) or authorised signatory		
Date		

9. Template for Participant Information Sheet

This page will be adapted for presentation on the world wide web.

Branded web page – clear identification of the University as the responsible institution

Title of study Self-reported health histories via anonymous web survey

Standard text:
We would like to invite you to take part in a research study. Before you decide whether you would like to take part it is important that you understand why the research is being done and what it would involve for you. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information.

What is the purpose of the study?

This research project is intended to collect anonymous information on individuals' memories of their health events that have occurred in their life to date.

You will be invited to list, to the best of your recollection, personal health events and ongoing conditions including, but not limited to, those events that you reported to, or required the intervention of, a general practitioner or other healthcare professional.

You will also be invited to note your age group, weekly exercise habits, smoking status and alcohol consumption, and country of birth. In order to preserve your anonymity, you will not be asked for your name, exact age or current country of residence. The IP number of the computer you are using will not be collected.

If you know the principal investigators of this study you are discouraged from participating in the study, in case you inadvertently reveal information that you would prefer not to. However, in any case, the investigators undertake not to attempt to reidentify any individuals from their submitted data.

Once data have been collected, they will be used:

i) To see how individuals’ recollections of medical conditions compare, on average, to the quantity and detail of events typically stored in general practice records;

ii) As input to a health event prediction algorithm, in order to see whether patient-recalled data is of adequate quality to have practical use in such an algorithm.

Why have I been invited?

Any adult with access to the world wide web is able to participate in this study.

Do I have to take part?

Participation in this project is entirely voluntary. You do not have to take part and can stop completing the short survey at any point without penalty. All responses are anonymous and we will not know who has or who has not participated in the study.

What will happen if I take part?

If you decide to take part, you will be asked to complete a short survey on this website. We expect that you will take less than 10 minutes to complete the survey and you should complete the survey only once.

What are the possible disadvantages and risks of taking part?

There are no disadvantages or risks involved in taking part in this study.

What are the possible benefits of taking part?

There are no direct benefits to you in participating in this study. Indirect benefits include benefits to future patients, to the wider community and a contribution to knowledge.

Will my taking part in the study be kept confidential?

All data will be collected anonymously. At no point will we know the identity of any of the participants. In addition to the data anonymity, the data will be kept secure and accessed only by the research team.

What will happen to results of the research study?

The results of this project will be written up as part of a PhD thesis and may also be published in an appropriate journal or presented at a conference. No individual will be identifiable as part of this process.

Should you wish to be sent a copy of any report or publication that results from this study, please contact [redacted] to request a copy. You do not have to have participated in the study to request a copy of the report.

What will happen if I don’t want to carry on with the study?

You will be asked only to complete a short questionnaire as part of this study and will be asked to do this only once. However, even if you agree to complete the questionnaire, you are free to stop at any point before you complete it.

What if there is a problem?

If you have any problems, concerns or questions about this study, you should ask to speak to a member of the research team. If you remain unhappy and wish to complain formally, you can do this through the University complaints procedure. To complain about the study, you need to phone [redacted]. You can then ask to speak to the Secretary to Senate Research Ethics Committee and inform them that the name of the project is: 'Self-reported health histories via anonymous web survey'

You could also write to the Secretary at:
Anna Ramberg
Secretary to Senate Research Ethics Committee

Research Office, E214
City University London
Northampton Square
London
EC1V 0HB
[redacted]

City University London holds insurance policies which apply to this study. If you feel you have been harmed or injured by taking part in this study you may be eligible to claim compensation. This does not affect your legal rights to seek compensation. If you are harmed due to someone's negligence, then you may have grounds for legal action.

Who has reviewed the study?

This study has been approved by City University London *[insert which committee here]* Research Ethics Committee

Further information and contact details

For further information on this project, or to request a copy of reports and publications arising from it, please contact:

Jonathan Turner
Research student,
Centre for Health Informatics, City University London
[redacted]

or his supervisor:
Dr Peter Weller
Head of Centre
Centre for Health Informatics, City University London

Thank you for taking the time to read this information sheet.

- Researcher’s checklist for compliance with the Data Protection Act, 1998

This checklist is for use alongside the *Guidance notes on Research and the Data Protection Act 1998*. Please refer to the notes for a full explanation of the requirements.

You may choose to keep this form with your research project documentation so that you can prove that you have taken into account the requirements of the Data Protection Act.

	• REQUIREMENT	✓	
	Meeting the conditions for the research exemptions:	✓	
1	The information is being used *exclusively* for research purposes.	✓	Mandatory
2	You are not using the information to support measures or decisions relating to *any* identifiable living individual.	✓	Mandatory
3	You are not using the data in a way that will cause, or is likely to cause, substantial damage or substantial distress to any data subject.	✓	Mandatory
4	You will not make the result of your research, or any resulting statistics, available in a form that identifies the data subject.	✓	Mandatory
	Meeting the conditions of the First Data Protection Principle:		
1	You have fulfilled one of the conditions for using personal data, e.g. you have obtained consent from the data subject. Indicate which condition you have fulfilled here: Participants are asked to give consent before supplying any personal information. All data collected is collected anonymously.	✓	Mandatory
2	If you will be using sensitive personal data you have fulfilled one of the conditions for using sensitive personal data, e.g. you have obtained explicit consent from the data subject. Indicate which	✓	Mandatory if using sensitive

	condition you have fulfilled here: Participants are asked to give consent before supplying any personal information. All data collected is collected anonymously.		data
3	You have informed data subjects of: i. What you are doing with the data; ii. Who will hold the data, usually City University London; iii. Who will have access to or receive copies of the data.	✓	Mandatory unless B4 applies
4	You are excused from fulfilling B3 only if all of the following conditions apply: i. The data has been obtained from a third party; ii. Provision of the information would involve disproportionate effort; iii. You record the reasons for believing that disproportionate effort applies, please also give brief details here: ____________________ ____________________ ____________________ ____________________ N.B. Please see the guidelines above when assessing disproportionate effort.		Required only when claiming disproportionate effort
	Meeting the conditions of the Third Data Protection Principle:		
1	You have designed the project to collect as much information as you need for your research but not more information than you need.	✓	Mandatory
	Meeting the conditions of the Fourth Data Protection Principle:		
1	You will take reasonable measures to ensure that the information you collect is accurate.	✓	Mandatory
2	Where necessary you have put processes in place to keep the information up to date.		Mandatory
	Meeting the conditions of the Sixth Data Protection Principle:		

	• You have made arrangements to comply with the rights of the data subject. In particular you have made arrangements to: i. Inform the data subject that you are going to use their personal data. ii. Stop using an individual's data if it is likely to cause unwarranted substantial damage or substantial distress to the data subject or another. iii. Ensure that no decision, which significantly affects a data subject, is based solely on the automatic processing of their data. iv. Stop, rectify, erase or destroy the personal data of an individual, if necessary. Please give brief details of the measures you intend to take here: ____ ____ ____ ____ ____	✓	Mandatory