



City Research Online

City, University of London Institutional Repository

Citation: de Grave, R., Rust, N., Reynolds, C., Watson, A., Smeddinck, J. & de Souza Monteiro, D. (2020). A catalogue of UK household datasets to monitor transitions to sustainable diets. *Global Food Security*, 24(100344), doi: 10.1016/j.gfs.2019.100344

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/23923/>

Link to published version: <https://doi.org/10.1016/j.gfs.2019.100344>

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online:

<http://openaccess.city.ac.uk/>

publications@city.ac.uk

1 **A Catalogue of UK household datasets to monitor transitions to sustainable diets**

2 Remco Benthem de Grave, Institute of Neuroscience, Newcastle University, UK
3 r.benthemdegrave2@newcastle.ac.uk

4 Niki A. Rust, Newcastle University, UK niki.rust@ncl.ac.uk

5 Christian J Reynolds, Department of Geography, University of Sheffield, and the Barbara Hardy
6 Institute, University of South Australia c.reynolds@sheffield.ac.uk, ORCID 0000-0002-1073-7394
7 (Corresponding author).

8 Anthony W. Watson, Human Nutrition Research centre, School of Biomedical Sciences, Newcastle
9 University UK. anthony.watson@newcastle.ac.uk

10 Jan D Smeddinck, Open Lab, Newcastle University jan.smeddinck@ncl.ac.uk

11 Diogo M. de Souza Monteiro, Centre for Rural Economy, School of Natural and Environmental
12 Sciences, Newcastle University, diogo.souza-monteiro@ncl.ac.uk

13 **Draft unformatted post-print (the version of the paper after peer-review, with revisions having**
14 **been made), Accepted to Global Food Security, 9th of December 2019.**

15 **Please see Global Food Security website for final version:**
16 <https://www.sciencedirect.com/science/journal/22119124>

17 **Abstract:** There is growing international consensus that current patterns of food consumption are not
18 sustainable and global change is needed. Understanding the mechanisms for a transition towards
19 more sustainable diets requires systematic temporal monitoring at the individual or household level.
20 Whilst many countries collect panel data on food expenditure and dietary intake, these datasets are
21 often not designed to monitor progress towards dietary sustainability, therefore using them to
22 understand how or why diets are becoming more or less sustainable can prove challenging. What is
23 also lacking is a curated dataset catalogue or a library where all relevant data could be easily accessible
24 to enable such evaluation. Our aim was to identify, classify and describe existing food expenditure and
25 diet datasets available in the UK and to assess the extent to which they can be used to monitor
26 transitions to sustainable diets. We found that despite the large number of datasets tracking UK
27 individual or household food purchases and consumption over time, these datasets are not suited to
28 understand how and why individuals are transitioning to sustainable diets. With the exception of
29 proprietary datasets, most datasets only collect data annually, making it challenging to understand
30 fine-scale behavioural change over shorter timeframes. There is an opportunity to design and
31 implement an open-access UK sustainable diets data collection effort at the individual and household
32 level. These efforts can be complemented with recent innovations in data science methods and digital
33 technologies – such as dietary intake trackers – that, along with supporting individuals in their dietary
34 behaviour change, may enable the collection of high-quality datasets.

35 Keywords: Panel data; food consumption; sustainable diets; data science; digital technologies;
36 review.

37 **Highlights:**

- 38 • The current open-access UK datasets have limited effectiveness to monitor fine-scale
39 transitions to sustainable diets.
- 40 • No single dataset recorded purchased and consumed quantities, alongside
41 attitudes/perceptions of dietary sustainability and food consumption or purchase.
- 42 • Multiple UK datasets can be used to collectively conduct analyses of general trends and to
43 compare different cohorts regarding the changes toward sustainable dietary patterns.
- 44 • Not all UK datasets are linked to databases containing environmental impact information.

45

46 **1. Introduction**

47 Current food purchase and consumption patterns are leading to unhealthy diets (Kearney, 2010),
48 which in turn are linked to increased prevalence of non-communicable diseases, such as obesity, type
49 2 diabetes and cardiovascular disease (Aston et al., 2012; Blundell and Cooling, 2000). Moreover, there
50 is mounting evidence that the production, processing, transport and final preparation of food to
51 support current dietary patterns have increasing environmental and social costs, creating an
52 unsustainable food and agricultural system that leads to increasing eutrophication, greenhouse gas
53 emissions, biodiversity loss, and food insecurity (Aleksandrowicz et al., 2016; Green et al., 2015; Poore
54 and Nemecek, 2018; Tilman and Clark, 2014; Willett et al., 2019).

55 As the evidence of the contribution of food production and consumption to the deterioration of
56 planetary health becomes clear, so too does the need to help consumers choose more sustainable
57 diets (Willett et al, 2019). The Food and Agricultural Organisation of the United Nations (FAO)
58 (Burlingame and Dernini, 2012) and the first and second US National Academies of Sciences,
59 Engineering and Medicine Workshops on Sustainable Diets, Food and Nutrition (Institute of Medicine,
60 2014; National Academies of Sciences, Engineering, and Medicine et al., 2019) suggest that sustainable
61 diets must be affordable and acceptable, as well as being healthy and nutritionally balanced with low
62 environmental impact. Transitioning towards sustainable diets is directly related to all of the United
63 Nations Sustainable Development Goals¹. There is thus a clear need for rapid, international change in
64 how we produce and consume food. Importantly, changes in demand patterns will help lead to
65 changes in production (Horton, 2017); indeed, Ingram (2017) argues that we need to change the way
66 we look at food systems and, rather than emphasizing the need to increase production, we should
67 focus on managing demand.

68 While there is increasing scientific consensus over the need to shift to more sustainable diets, there is
69 less clarity on how to implement that transition. Food choices are complex and have numerous
70 determinants. They are influenced by geographical, economic and social factors along with a mix of
71 local, regional and national government policies, as well as business strategies. Due to these multiple
72 influences, it is vital to systematically monitor the effectiveness of different interventions and assess
73 if, how, where and which dietary transitions are occurring. To understand these trends, it is important
74 to identify what datasets are currently in the public domain that monitor individual or household food
75 expenditure and consumption both at home and away from home with a regular frequency to
76 determine micro-level change.

77 In many developed countries there are both private and public data collection efforts collecting
78 information on food expenses, consumption patterns and nutrition². However, to the best of our
79 knowledge, these data sets are not curated or catalogued in a systematic way and then made available
80 to the policy or research community to conduct further analysis that can be used to inform policy and
81 practice.

82 Having a catalogue or list of datasets can be beneficial for undertaking future research in this area.
83 Examples of such research is De Keyzer et al. (2015), Perignon et al. (2017), as well as Bandy et al.

¹ <https://www.stockholmresilience.org/research/research-news/2016-06-14-how-food-connects-all-the-sdgs.html>

² For example the World Bank Global Consumption Database compiles food expenditures across food and drinks expenses from a nationally representative sample of developing countries households (<http://datatopics.worldbank.org/consumption/sector/Food-and-Beverages>). Similar datasets are available from international organizations like the OECD, the European Union and the national statistics of all high-income countries.

84 (2019), who conducted systematic reviews of food consumption datasets investigating evidence of
85 progress to health or sustainable diets. They found important gaps and limitations regarding the
86 applicability of datasets for monitoring transitions to sustainable food consumption behaviour. The
87 reviews indicate that there does not appear to be a systematic data collection effort capturing all the
88 dimensions of sustainability. However, databases currently exist that allow estimating nutritional
89 values, greenhouse gas emission (GHGE) and cost from purchased or consumed products (see for
90 instance www.ggdot.org, (Hobbs et al., 2015; Horgan et al., 2016; Monsivais et al., 2013)).

91 This paper aims to start filling this gap by:

- 92 1. providing an overview of existing private and public datasets on food purchases, as well as
93 consumption and dietary intake patterns in the UK and,
- 94 2. discussing their suitability to assess transitions and changes towards sustainable diets, to offer
95 a catalogue of existing data sources enabling a monitoring or assessment of transitions to
96 sustainable diets in the UK, as well as highlighting the limitations and opportunities of
97 available datasets.

98 The UK is an interesting starting point and case for observation with respect to this topic because the
99 sustainability of current UK diets has been questioned (Reynolds et al., 2019a,b; Reynolds et al., 2015;
100 Wrieden et al., 2017). The UK is committed to meeting the UN Sustainable Development Goals,
101 recently also declaring the goal of reaching 'net zero' carbon emissions by 2050 (Pye et al., 2017;
102 Walker et al., 2019), and has an actively engaged political and civil society in developing approaches
103 to improving the current dietary and environmental situation. Moreover, the UK has a strong tradition
104 and capacity to collect data on food purchase and consumption (Oddy, 2003; Orr, 1937) and is in the
105 process of developing a National Food Strategy³ that focuses on human and environmental health, to
106 which an understanding of current food consumption trends will be integral.

107 Along with providing a comprehensive overview and discussion of available datasets on UK food
108 expenditure and consumption patterns to support future data collection efforts, we also provide
109 suggestions for approaches to improving the completeness, quality, and linking of existing datasets,
110 and we discuss the potential for improved data collection and monitoring with digital technologies. As
111 such, next to informing further research, this work provides guidance and evidence on improving data
112 collection that can lead to better monitoring and understanding of transitions towards more
113 sustainable diets. The outcomes can therefore be helpful to policy makers, research an industry alike.

114 3. **Methods**

115 We aimed to identify datasets that researchers can use to assess trends in individual and household
116 dietary behaviour that can be used to track the level of sustainability in their food consumption,
117 expenditures, purchases and dietary intakes. We therefore searched for UK datasets that had
118 temporal information on individuals or household food purchases/expenditure –i.e. panel data– or
119 consumption –i.e. dietary surveys. From these initial criteria, we added a second layer, using a range
120 of sustainability dimensions as the second criterion for selection, i.e. we identified in the panel
121 datasets whether they contained measures of:

- 122 1. healthiness of diets /purchases (estimated using nutrition profile tables);
- 123 2. affordability (using price and income information);

³ <https://consult.defra.gov.uk/agri-food-chain-directorate/national-food-strategy-call-for-evidence/>

124 3. environmental sustainability (specifically whether the data contained information on carbon
125 footprints or water use)^{4,5}.

126 Using these criteria, a first list of datasets was created from authors combined knowledge of
127 (publications about) data collection efforts describing diets in the UK. Next we consulted the UK data
128 service (see <https://www.ukdataservice.ac.uk/get-data/themes/food.aspx>). To identify additional
129 datasets, we contacted researchers through personal networks who undertake empirical analysis of
130 food consumption. We also reached out to private companies that collect diet information (not
131 necessarily in the UK) and to expert groups such as the Food and Climate Research Network (FCRN)
132 Google group (<https://groups.google.com/forum/#!topic/fcrn-l/TRMs4BnUWYc>).⁶

133 To be able to reconstruct a complete diet, we defined that the data should cover at least one complete
134 consumption day (e.g. through a 24h Dietary Recall (24h-DR), a Diet Diary (DD), an extensive Food
135 Frequency Questionnaire (FFQ), or a purchase diary of at least a week). To focus on UK population
136 dietary change, we excluded datasets that focus exclusively on children or the very old, as well as
137 datasets that consist of secondary data collection efforts (i.e. merging data collection efforts done
138 elsewhere).

139 For each dataset identified, we collected characteristics and metadata from the description that
140 accompanied the dataset. In some cases, we referred to the original survey questionnaires, the raw
141 data, or to publications that use the dataset to find this information. Where we required further
142 information, institutions were contacted to verify entries and asked for missing information, though
143 not all data holders returned answers (see table 2).

144 One of the main challenges we faced was inconsistency across the way diets and nutrition have been
145 measured and reported. For example, some datasets record food consumption, others food
146 purchases. These cannot easily be combined, but they do complement each other, as there is high
147 correlation between what is purchased and what is consumed – with food waste data then used to
148 further ‘triangulate (Reynolds et al. 2019a)’. Some datasets maybe be combined using matching
149 methods⁷ (such as propensity score matching) which enable a construction of a comparison group
150 that is similar to the group of interest. However, there are caveats to these methods, namely that the
151 variables on which the matching is being made may have been collect in different ways and may not
152 be capturing the same characteristics used for the matching process.

⁴ We acknowledge that one of the limitations of this methods is that we have not referred datasets that are used to construct these panels. For instance, there are several food composition tables available in the UK that are the basis for the nutrition information provided in commercial panel data. Undoubtedly there is a need to curate those data sources, but that is beyond our goal on this paper.

⁵ Please note, as very few datasets have the capacity to immediately calculate these aspects, that for aspects of (1) healthiness and (3) environmental sustainability, we assessed whether each dataset contained ‘sufficient’ purchase/consumption data to calculate these dimensions by combining this data with tables of nutrition profiles and GHGE impact of the diet. We acknowledge that there are multiple additional factors that can be used to measure sustainability, such as water use, land use, biodiversity loss etc. however as stated in the main text, GHGE has multiple linked datasets already in wide use.

⁶ Final searches (and citation mining) of Google Scholar were carried out using search terms combinations of “Diet”, “Food”, “UK”, “Recall”, “Cohort”, “Questionnaire”, “Diary”, as well as the dataset names to find any additional datasets.

⁷ Matching methods are statistical and econometric methods were developed to combine datasets collecting similar information on different units observation (see Rosenbaum and Rubin (1983){ P.R. Rosenbaum, D.B. Rubin **The central role of the propensity score in observational studies for causal effects** *Biometrika*, 70 (1983), pp. 41-55}) to enable causal analysis by constructing a counterfactual.

153

154 **4. Results**

155 This section presents the datasets identified and provides further details on those that met
156 our main and secondary criteria. In table 1, we list all datasets that were identified in our search,
157 highlighting in bold those meeting our inclusion criteria. In table 2, we describe, in detail, nine datasets
158 that provided a complete overview of at least one full day of consumption or purchase data that can
159 be accessed for research purposes (noting which data is proprietary).

160

161 **Table 1:** List of UK household panel datasets gathering data on food expenditure and
 162 consumption. Public datasets are those collected by governmental agencies or funded by public
 163 research funds. Private sources are those collected by commercial companies, generally through
 164 home or retail scanners, surveys or through apps. Public datasets are divided into *open* or *restricted*,
 165 with *restricted* meaning that further access permissions where institutional associations need to be
 166 verified and sometimes special permission requests need to be provided. Private datasets are divided
 167 into those that are available for a fee and those that are generally not shared outside the company
 168 (restricted private datasets).

Dataset or survey name	Public		Private	
	Open	Restricted	Fee	Restricted
EPIC Norfolk (Day et al., 1999)		√		
EPIC Oxford (Davey et al., 2003)		√		
Family Food module of Living Cost and Food Survey (LCFS) (Department For Environment and Office For National Statistics, 2017) (Office For National Statistics, 2019)		√		
Fenland study ("Fenland Technical Summary - MRC Epidemiology Unit," n.d.)		√		
Kantar consumption panel			√	
Kantar purchase panel			√	
National Diet and Nutrition Survey (NDNS) (Laboratory and Research, 2019)		√		
UK Women Cohort Survey (UKWCS) (Cade et al., 2015)		√		
UKBiobank (Sudlow et al., 2015)	√			
Health Survey for England		√		
1000 family study		√		
85+ study		√		
ASH30		√		
ALSPAC		√		
FAO statistics	√	√ ²		
Food and Drink in Scotland		√		
Gateshead Millennium Cohort		√		
GfK (company)			? ¹	
Global Dietary Database (GGD)		√		
Loyalty card data collections (e.g. Dunhumby, Tesco, Sainsbury, Waitrose)			√	√ ³
Million Women Study		√		
MyFitnessPal (company)				√
Nielsen (company)			? ¹	
Scottish Health Survey		√		
Slimming world (company)				√
Weightwatchers (company)				√

169 ¹ Data for the UK for these companies may not be available, but this was not conclusively verified (the
 170 companies did not respond to an information request). ² Greater detail available via application for
 171 restricted data for some areas. ³ Some Loyalty card data available through UKDS and the CDRC.

172 Table 1 can be considered a tentative index, where we categorise the datasets identified
 173 according to their ownership and accessibility.

174

175 Table 2 presents and characterizes the nine datasets that met our main criteria. Next, we briefly
176 explain the characteristics of these data in three dimensions: sampling and recruitment, data
177 collection methods and economic information therein.

178 <<<<Table 2 here>>>

179 **Study design, recruitment and sample characteristics**

180 Three types of designs can be recognized in the overview. First, two of the nine studies
181 concerned non-cohort studies (National Diet and Nutrition Survey, and the Family Food Module of
182 Living Cost and Food Survey, FFM-LCFS). Both of these cross-sectional studies targeted UK households
183 using a multistage stratified sampling strategy in which households were identified from Postcode
184 Address Files (PAF) and recognized as small users, and clustered in Primary Sampling Units (PSUs).
185 Households were then drawn from a number of PSUs. Samples sizes ranged from about 1000
186 participants annually in the NDNS to 6000 households annually.

187 Second, six datasets concerned cohort studies (EPIC Norfolk, EPIC Oxford, the Fenland Study,
188 the UKBiobank and the UK Women’s Cohort Survey, UKWCS, Million Women Study). Targeted
189 populations varied considerably. Some studies targeted specific diets (non-red-meat-eating,
190 vegetarian), some geographical regions (Norfolk, Cambridgeshire) and two study targeted women
191 only. All studies targeted a middle-age range with participant ages ranging 20-79. NHS registers and
192 membership lists were used to recruit people. Cohort sizes of ranged from roughly 12,500 (Fenland
193 Study) up to approximately 211,000 (UKBiobank) or 688,000 (Million Women), although sample sizes
194 at the level of individual recordings range 1,600-100,000.

195 The datasets collected by Kantar are the only commercial datasets and the only ones that
196 monitor participants’ diets over an unrestricted time frame (4x per year with 10,000 people in the
197 consumption panel and 30,000 people in the purchase panel). Advertisements on social media were
198 used to recruit people, although more targeted methods were also used to obtain a representative
199 sample size.

200 **Dietary assessment methods, administration method and method of portion size estimation**

201 A variety of methods to assess dietary consumption or purchases can be found between and
202 within the databases. These include Food Frequency Questionnaires (FFQ), 24-Hour Dietary Recalls
203 (24h-DR), Diet Diaries (DDs), and purchase diaries. There are well known completion biases with all
204 food intake questionnaires when assessing the dietary intake of a free living population; with a linear
205 association between participant burden and accuracy. None of the datasets assessed used the “gold
206 standard” duplicate diaries to assess food intake and most used standard portions to assess food
207 quantities. It must also be noted that datasets which convert food intake to nutrient and energy intake
208 use food composition tables which are limited by the small number of foods they include and the age
209 of the data within the dataset. Therefore, only a “best fit” approach was used to crudely estimate
210 intake.

211 Food Frequency Questionnaires (FFQs) were used in four studies. These questionnaires asked
212 about habitual consumption frequency in the past 12 months on a range of food items (28 to 217 food
213 items). Participants were requested to rate their consumption frequency from never to six per day on
214 nine frequency choices. Some exceptions to this are that one study (UKWCS) used a 10-point
215 frequency scale and two smaller FFQs in EPIC Oxford used a 6-frequency scale. Portion sizes were
216 generally estimated by framing the question such that it asked for the consumption of standard
217 portion sizes. The standard portion size was then described with the item or category, for example

218 one sausage or one portion of carrots. Some questionnaires omitted portion size and only asked for a
219 frequency – this was to determine ‘general’ diet over a period of time rather than what was eaten on
220 a specific day. We note that some of the smaller FFQs do not include a full range of foods, only
221 categories of interest to the study. However, other assessments in the same study do. The mini FFQ’s
222 were included for completeness.

223 The 24h-DR was used in three studies. These asked about the consumption of the previous
224 day. Methods used varied from pen and paper recordings, accompanied with suggestions on standard
225 portion sizes, to online forms that required to rate their portion sizes in standard measures. The 24h-
226 DRs were all self-administered, either at the test centre or at home.

227 Diet Diaries (DDs) were used in five studies. These asked the participants to track their
228 consumption for several days (ranging between studies from 4 to 7 days). In both EPIC studies and the
229 NDNS paper, DDs were used in combination with suggestions for standard portion sizes, supported by
230 pictures of various portion sizes that participants could refer to. In the UKWCS, participants were
231 asked to list weight or volume of consumed products which had to be measured or read from
232 packaging (standard measures were allowed on some occasions). The DD in Kantar was performed on
233 a computer. Participants selected per meal the products that they had used, but did not specify
234 amounts consumed.

235 Purchase diaries were used in two studies. The FFM-LCFS used pen and paper entries or
236 allowed participant to attach their receipts. In the Kantar purchase panel, participants were asked to
237 scan each purchase receipt using a digital clicker. Both purchase diaries are self-administered and
238 completed at home. One of the limitations of purchased data is that they are only proxies for
239 consumption, as they don’t factor in wastage (though it can be examined with further inquiries or
240 complementary studies) or the delayed consumption. However, this type of data has information of
241 food prices and collects data on disposable income and permits estimation of expenditure by category.

242 **Economic information**

243 Income is recorded for five out of the nine studies we describe (the NDNS, the FFM-LCFS, the Fenland
244 Study and both Kantar datasets), while prices and/or expenditure are also recorded in the purchase
245 panels (FFM-LCFS and Kantar datasets). One of the problems with recording economic and income
246 information in the datasets we identified is that it is not consistent. For example, the Kantar data
247 enables a verifiable estimation of weekly expenditure as it is based on actual shopping receipts, but
248 this does not necessarily provide accurate information on what is actually consumed, nor does it
249 distinguish who in the household consumes what. Another issue is that some datasets collect
250 information on individuals, while others do so across households which prevents a combination of
251 different datasets. Still, insofar as these datasets capture information on disposable income and
252 purchases, they enable an assessment of affordability. Moreover, it may enable comparisons across
253 segments of the population and identify opportunities to improve the sustainability of diets within the
254 budget limits of household. When geographical information on location of households is available it
255 may be possible to understand how the food retail and service environment may determine the food
256 choices.

257 A full economic assessment of transitions to sustainable diets would need to include other
258 variables that are not currently collected, for example time spent planning, shopping and preparing
259 meals. There are datasets that provide information on time use (for example the Gershuny and
260 Sullivan (2017) survey on how much time different groups of the population spend their time), which
261 can be used to estimate more accurately the costs of sustainable diets.

262 **Environmental information**

263 There was no environmental impact information found within the datasets surveyed. However, GHGE
264 emission datasets have been linked to multiple datasets presented in table 1. This includes the NDNS
265 (Bates et al. 2019) and LCFS (Wriden et al 2019). We also found that the USDA’s FoodAPS (Boehm et
266 al. 2018) and the European Food Standard Agencies FoodEx2 (Reynolds et al. 2019c) have also been
267 matched to GHGE databases.

268 **2. Discussion**

269 In this study, we have identified, classified and described nine datasets on diet, food
270 consumption, or expenditure that are available in the UK to the research community. Individually,
271 each dataset has limited effectiveness to monitor transitions to sustainable diets and for direct
272 comparisons between datasets. This is because they were not designed for either of these purposes.
273 The datasets use different units of observation, sampling sizes⁸, sampling rates, and study durations.
274 In addition, the datasets recorded either food purchased or consumed. In this regard, our findings are
275 consistent with the data limitations identified by Perignon et al. (2017), who found that there is a lack
276 of relevant and good-quality datasets for assessing the environmental, health and socio-economics
277 impact of current diets.

278 However, we propose that collectively these datasets have the potential to assess transitions and
279 changes towards sustainable diets in the UK. This is because they are complementary and can become
280 elements of multi-layered analysis combining food consumption or purchase with other information
281 affecting the households or individuals on which data is collected. For these purposes, the identified
282 datasets have to be matched with other existing databases containing geographical information,
283 further socio-economic information and environmental impact information of the foods consumed or
284 purchased. As we pointed out in the results, some of these datasets already existed, though
285 necessarily easily accessible. As already mentioned, matching methods (Rosenbaum and Rubin (1983);
286 Stuart (2010)) are increasingly used to combine datasets and construct counterfactuals that enable
287 causal analysis when, as is the case, it is challenging to design suitable experiments. While this
288 matching may not always be feasible and could be labour-intensive to varying degrees, due to the
289 different levels of food classification and dimensions for data-aggregation in each database, there are
290 already ways to automate the mapping and linking of dietary and environmental impact databases
291 (Eftimov et al., 2017). Even if they are not linked directly to environmental impacts, these databases
292 can still be used to collectively conduct analyses at the social-economic strata level to investigate
293 general trends and to compare different cohorts regarding the changes in dietary patterns.

294 In the best case, a collaborative and coordinated data collection effort - that takes account of possible
295 linkages, and upcoming data needs - must be part of any new food strategy for the UK. This strategy
296 could extend beyond the datasets identified in this paper to include linkages to food composition
297 tables, lifecycle and environmental impact studies from different food categories, along with data
298 from alternative production systems, and archived consumer survey instruments. It is beyond the
299 purpose of this study to provide those sources of information, but we acknowledge the need for such
300 a strategy and repository of complementary datasets that could be easily searched and used. For
301 example, a preliminary search for Composition tables in the UK identified a Governmental source of
302 data (the CoFID- Composition of Foods Integrated dataset) and the Carter et al (2016) new branded
303 UK composition database. Along with a list of databases, it may be informative to provide potential

⁸ For some datasets it is uncertain whether they present a representative sample of the British population.

304 users with a quality assessment of the data in repositories commenting on the methods used to collect
305 the data and its limitations.

306 The public datasets we identified are generally accessible, have a snapshot nature, and are
307 suitable to evaluate how different groups have changed diets and facilitate cross sectional analysis.
308 The value of the household food purchases panel data (such as Kantar) is that it enables researchers
309 to observe transitions with a much finer granularity. However, this analysis has the caveat that it does
310 not capture individual consumption, but rather expenditures. Still, it enables comparison on how
311 different households are changing consumption of a given food category and whether they are shifting
312 to healthier, more sustainable food categories, as well as across household types, and time periods
313 (52 weeks over a year in the case of Kantar, or weekly once a year e.g. for LCFS). In isolation, these
314 datasets do not necessarily gather information on the health status of the household they recruit. In
315 addition, there is a lack of detail in current panel data on the traceability and origin of food; this
316 additional information is needed to truly understand sustainability of different foodstuffs.

317 It should be highlighted that there is a certain degree of self-selection bias on the households
318 that are included in both public and private panels that were reviewed. Moreover, these datasets have
319 not inquired about households' attitudes to - or perceptions of - sustainable dimensions of food
320 consumption or purchase (this would be required to understand reasons why people make changes
321 in what they eat). Moreover, there is limited information about the home and neighbourhood context
322 as well as on the food preparation and consumption practices with which to explore more deeply what
323 may motivate or hinder transitions at the households or individual level. Indeed, the food availability
324 landscape is not necessarily captured in the datasets we have identified. However, those factors are
325 important determinants of consumption and purchase. Consequently, as the existing datasets do not
326 carry data on 1) attitudes and 2) the food environment there must be caution when interpreting this
327 data to assess and draw conclusions as to what may have changed dietary behaviour and
328 consumption/purchase patterns over time.

329 Still, the complementarity between the more frequent and rich information on products
330 gathered in panel data and the broad coverage of large cohort studies presents a clear opportunity
331 for assessing general transitions to sustainable diets. The household panel data could be employed to
332 identify trends and micro-responses to interventions, in turn the cohort studies can be used to confirm
333 how they are impacting broader aggregate measures. Another opportunity lies with matching both
334 private and public datasets to geographical information (which is recorded in differing detail in each
335 dataset) to further our understanding of how changes in regional or urban food policies may be
336 affecting consumption patterns, as well as environmental and health outcomes.

337 To overcome the aforementioned limitations of current datasets and to develop new datasets,
338 we suggest harnessing technological developments to better assess dietary transitions and changes
339 towards sustainable diets. We therefore briefly highlight the potential of digital wearable devices to
340 collect data on food choices, as well as the use of data science methods to provide new methods of
341 data harmonization and mapping.

342 In principle, data science methods (including frequentist statistics, probabilistic methods, data
343 matching as well as different techniques from machine learning and artificial intelligence) can be used
344 for two main purposes with respect to the existing datasets:

345 1) improving the data-quality and reducing sparsity (filling gaps, e.g. data imputation (Jerez et
346 al., 2010)),

- 347 2) linking datasets (e.g. through auto-correlation) (“Automated census record linking: a
348 machine learning approach,” n.d.),
- 349 3) clustering datasets or supersets, creating new sectioning or subsets (e.g. using
350 autoencoders (Baldi, 2012)),
- 351 4) optimizing future / ongoing data collection (Sra et al., 2011) and
- 352 5) prediction.

353 At the same time, with the growing capabilities and affordability of sensors and increased
354 computational capacity easily available in the cloud, digital technology, including devices and software
355 applications opens interesting opportunities for improving data collection and research efforts. Digital
356 data streams can be very complex and have a high sampling rate – which can at times even emulate
357 real-time “natural fidelity” recording, compared to what is feasible with more traditional data
358 collection efforts. This area can be split into four main elements:

- 359 1) quantified self and community applications with a) self-reporting tools, such as consumption /
360 intake trackers (Bradley et al., 2016), or b) habit tracking / forming apps (Stawarz et al., 2015),
- 361 2) general dietary information tools (Boulos et al., 2015),
- 362 3) professional practice support (Simons et al., 2012) and
- 363 4) indirect information sources (such as product sales data, raw materials uptake / tracking, supply-
364 chain monitoring, distributed ledgers, as well as production and transport cost /energy expenditure
365 monitoring).

366 The ethical implications (and possibilities for additional bias) due to the use of such technologies,
367 sensors, wearables, and the internet of things are of considerable extent and beyond the scope for
368 this paper. Possible future research questions include: when and how should researchers be allowed
369 to gain access to data from wearables? How can researchers ensure that an individual’s data is used
370 with care? How can researchers ensure that we are not neglecting harder-to-access members of
371 society such as the poor? Who pays for these wearables? And how do we overcome the “big brother”
372 nature of these devices?

373 5. Conclusions and future work

374 We identified and classified existing data sources with the potential to be used in research on
375 monitoring transitions towards more sustainable diets in the UK. We present a catalogue of datasets
376 classified in key sustainability dimensions and discuss potential of these datasets for such analysis. We
377 conclude that neither of the datasets fulfils the requirements for reliable monitoring or prediction.
378 Most of the datasets are also limited to traditional data sources, such as survey responses. This clearly
379 suggests two pathways for future work: improving the quality and enable matching of the existing
380 data sets, as well as a broader effort to collect coherent data on transitions towards more sustainable
381 diets that combines - in a single data collection instrument - individual-level data, including
382 motivations, and objective behaviour and food consumption over regular time periods. This
383 instrument needs to be carefully designed tacking into account existing datasets with complementary
384 information, such as food composition tables, environmental impact assessment and economic
385 information. If designed and implemented ethically, digital technologies can play a key role and enable
386 novel approaches and insights. These technologies include software with supportive algorithms and
387 user interfaces, which can, for example, gauge shopping behaviour, shopping, and the engagement

388 with – and social communication about – diet information sources, as well as (sensing) hardware
389 devices that allow for objective measurements e.g. of eating behaviour.

390 We also acknowledge that we have not documented and critically examined other data
391 sources that complement the datasets we covered – this includes a) food composition tables; b)
392 datasets of environmental outcomes for different foods; c) food price datasets; d) survey data on
393 attitudes surrounding food purchasing behaviour. These were outside the scope of this particular
394 effort. We recognize that such information is valid and believe that future work should fill that gap.
395 Also, we hope our limitations inspire researchers interested in measuring sustainable diets to create
396 and curate a library of datasets facilitating further work in this area.

397

398 6. Bibliography

- 399 Aleksandrowicz, L., Green, R., Joy, E.J.M., Smith, P., Haines, A., 2016. The impacts of dietary change
400 on greenhouse gas emissions, land use, water use, and health: A systematic review. *PLoS*
401 *One* 11, e0165797. doi:10.1371/journal.pone.0165797
- 402 Aston, L.M., Smith, J.N., Powles, J.W., 2012. Impact of a reduced red and processed meat dietary
403 pattern on disease risks and greenhouse gas emissions in the UK: a modelling study. *BMJ*
404 *Open* 2. doi:10.1136/bmjopen-2012-001072
- 405 Automated census record linking: a machine learning approach [WWW Document], n.d. URL
406 <https://open.bu.edu/handle/2144/27526> (accessed 6.30.19).
- 407 Bates, R.L., Chambers, N.G. and Craig, L.C.A., 2019. Greenhouse gas emissions of UK diets.
408 *Proceedings of the Nutrition Society*, 78(OCE2). doi:10.1017/S0029665119000910
- 409 Baldi, P., 2012. Autoencoders, unsupervised learning, and deep architectures. *Proceedings of ICML*
410 *workshop on unsupervised and transfer learning* 37.
- 411 Bandy, L., Adhikari, V., Jebb, S., Rayner, M., 2019. The use of commercial food purchase data for
412 public health nutrition research: A systematic review. *PLoS One* 14, e0210192.
413 doi:10.1371/journal.pone.0210192
- 414 Blanquer, M., García-Alvarez, A., Ribas-Barba, L., Wijnhoven, T.M.A., Tabacchi, G., Gurinovic, M.,
415 Serra-Majem, L., 2009. How to find information on national food and nutrient consumption
416 surveys across Europe: systematic literature review and questionnaires to selected country
417 experts are both good strategies. *Br. J. Nutr.* 101 Suppl 2, S37–50.
418 doi:10.1017/S0007114509990572
- 419 Blundell, J.E., Cooling, J., 2000. Routes to obesity: phenotypes, food choices and activity. *Br. J. Nutr.*
420 83 Suppl 1, S33–8.
- 421 Boehm, Rebecca, Parke E. Wilde, Michele Ver Ploeg, Christine Costello, and Sean B. Cash. "A
422 comprehensive life cycle assessment of greenhouse gas emissions from US household food
423 choices." *Food policy* 79 (2018): 67-76.
- 424 Boulos, M., Yassine, A., Shirmohammadi, S., Namahoot, C., Brückner, M., 2015. Towards an “internet
425 of food”: food ontologies for the internet of things. *Future Internet* 7, 372–392.
426 doi:10.3390/fi7040372

427 Bradley, J., Simpson, E., Poliakov, I., Matthews, J.N.S., Olivier, P., Adamson, A.J., Foster, E., 2016.
428 Comparison of INTAKE24 (an Online 24-h Dietary Recall Tool) with Interviewer-Led 24-h
429 Recall in 11-24 Year-Old. *Nutrients* 8. doi:10.3390/nu8060358

430 Burlingame, B., Dernini, S., 2012. SUSTAINABLE DIETS AND BIODIVERSITY DIRECTIONS AND
431 SOLUTIONS FOR POLICY, RESEARCH AND ACTION.

432 Cade, J.E., Burley, V.J., Alwan, N.A., Hutchinson, J., Hancock, N., Morris, M.A., Threapleton, D.E.,
433 Greenwood, D.C., 2015. Cohort profile: the UK women's cohort study (UKWCS). *Int. J.*
434 *Epidemiol.* doi:10.1093/ije/dyv173

435 Carter MC, Hancock N, Albar SA, et al. Development of a New Branded UK Food Composition
436 Database for an Online Dietary Assessment Tool. *Nutrients*. 2016;8(8):480. Published 2016
437 Aug 5. doi:10.3390/nu8080480

438 Davey, G.K., Spencer, E.A., Appleby, P.N., Allen, N.E., Knox, K.H., Key, T.J., 2003. EPIC-Oxford: lifestyle
439 characteristics and nutrient intakes in a cohort of 33 883 meat-eaters and 31 546 non meat-
440 eaters in the UK. *Public Health Nutr.* 6, 259–269. doi:10.1079/PHN2002430

441 Day, N., Oakes, S., Luben, R., Khaw, K.T., Bingham, S., Welch, A., Wareham, N., 1999. EPIC-Norfolk:
442 study design and characteristics of the cohort. *European Prospective Investigation of Cancer.*
443 *Br. J. Cancer* 80 Suppl 1, 95–103.

444 De Keyzer, W., Bracke, T., McNaughton, S.A., Parnell, W., Moshfegh, A.J., Pereira, R.A., Lee, H.-S.,
445 van't Veer, P., De Henauw, S., Huybrechts, I., 2015. Cross-continental comparison of national
446 food consumption survey methods--a narrative review. *Nutrients* 7, 3587–3620.
447 doi:10.3390/nu7053587

448 Department For Environment, F., Office For National Statistics, 2017. Living Costs and Food Survey,
449 2015-2016. UK Data Service. doi:10.5255/ukda-sn-8210-4

450 Eftimov, T., Korošec, P., Koroušić Seljak, B., 2017. StandFood: Standardization of Foods Using a Semi-
451 Automatic System for Classifying and Describing Foods According to FoodEx2. *Nutrients* 9.
452 doi:10.3390/nu9060542

453 Fenland Technical Summary - MRC Epidemiology Unit [WWW Document], n.d. URL
454 <http://doi.org/10.22025/2017.10.101.00001> (accessed 6.28.19).

455 Gershuny, J., Sullivan, O. (2017). United Kingdom Time Use Survey, 2014-2015. [data collection]. UK
456 Data Service. SN: 8128, <http://doi.org/10.5255/UKDA-SN-8128-1>

457 Green, R., Milner, J., Dangour, A.D., Haines, A., Chalabi, Z., Markandya, A., Spadaro, J., Wilkinson, P.,
458 2015. The potential to reduce greenhouse gas emissions in the UK through healthy and
459 realistic dietary change. *Clim. Change* 129, 253–265. doi:10.1007/s10584-015-1329-y

460 Hobbs, D.A., Lovegrove, J.A., Givens, D.I., 2015. The role of dairy products in sustainable diets:
461 modelling nutritional adequacy, financial and environmental impacts. *Proc. Nutr. Soc.* 74.
462 doi:10.1017/S0029665115003572

463 Horgan, G.W., Perrin, A., Whybrow, S., Macdiarmid, J.I., 2016. Achieving dietary recommendations
464 and reducing greenhouse gas emissions: modelling diets to minimise the change from
465 current intakes. *Int. J. Behav. Nutr. Phys. Act.* 13, 46. doi:10.1186/s12966-016-0370-1

466 Horton, P., 2017. We need radical change in how we produce and consume food. *Food Sec.* 9, 1323–
467 1327. doi:10.1007/s12571-017-0740-9

468 Ingram, J., 2017. Perspective: Look beyond production. *Nature* 544, S17. doi:10.1038/544S17a

469 Institute of Medicine, 2014. Sustainable Diets: Food for Healthy People and a Healthy Planet:
470 Workshop Summary, The National Academies Collection: Reports funded by National
471 Institutes of Health. National Academies Press (US), Washington (DC). doi:10.17226/18578

472 Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M., Franco, L., 2010.
473 Missing data imputation using statistical and machine learning methods in a real breast
474 cancer problem. *Artif Intell Med* 50, 105–115. doi:10.1016/j.artmed.2010.05.002

475 Kearney, J., 2010. Food consumption trends and drivers. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* 365,
476 2793–2807. doi:10.1098/rstb.2010.0149

477 Laboratory, M.E.W., Research, N.S., 2019. National Diet and Nutrition Survey Years 1-9, 2008/09-
478 2016/17. UK Data Service. doi:10.5255/ukda-sn-6533-13

479 Monsivais, P., Perrigue, M.M., Adams, S.L., Drewnowski, A., 2013. Measuring diet cost at the
480 individual level: a comparison of three methods. *Eur. J. Clin. Nutr.* 67, 1220–1225.
481 doi:10.1038/ejcn.2013.176

482 National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Food and
483 Nutrition Board, Food Forum, 2019. Sustainable diets, food, and nutrition: proceedings of a
484 workshop, The National Academies Collection: Reports funded by National Institutes of
485 Health. National Academies Press (US), Washington (DC). doi:10.17226/25192

486 Oddy, D.J., 2003. From plain fare to fusion food: British diet from the 1890s to the 1990s. Boydell
487 Press, Woodbridge, Suffolk.

488 Office For National Statistics, 2019. Living Costs and Food Survey, 2017-2018. UK Data Service.
489 doi:10.5255/ukda-sn-8459-1

490 Orr, J.B., 1937. Food health and income: Report on a survey of adequacy of diet in relation to
491 income. Macmillan and Company Limited,, London.

492 Perignon, M., Vieux, F., Soler, L.-G., Masset, G., Darmon, N., 2017. Improving diet sustainability
493 through evolution of food choices: review of epidemiological studies on the environmental
494 impact of diets. *Nutr. Rev.* 75, 2–17. doi:10.1093/nutrit/nuw043

495 Poore, J., Nemecek, T., 2018. Reducing food’s environmental impacts through producers and
496 consumers. *Science* 360, 987–992. doi:10.1126/science.aaq0216

497 Pye, S., Li, F.G.N., Price, J., Fais, B., 2017. Achieving net-zero emissions through the reframing of UK
498 national targets in the post-Paris Agreement era. *Nat. Energy* 2, 17024.
499 doi:10.1038/nenergy.2017.24

500 Reynolds, C., Hodgson, H., Bajzelj, B., 2019b. An improved picture of the UK diet: Linking production,
501 consumption and waste data to provide a better dietary picture. WRAP.

502 Reynolds, C.J., Horgan, G.W., Whybrow, S., Macdiarmid, J.I., 2019a. Healthy and sustainable diets
503 that meet greenhouse gas emission reduction targets and are affordable for different

504 income groups in the UK. *Public Health Nutr.* 22, 1503–1517.
505 doi:10.1017/S1368980018003774

506 Reynolds, C.J. Schmidt Rivera, X. Frankowska, A. Kluczkowski, A. da Silva J. T., Bridle S. L. Levy, R.
507 Rauber, F. Quadros, V. P. Balcerzak, A. Sousa, R. F. Ferrari, M. Leclercq, C. Koroušić Seljak,
508 B. Eftimov, T. (2019c) A Pilot Method Linking Greenhouse Gas Emission Databases To The
509 Foodex2 Classification, Livestock, Environment and People (LEAP) Conference 2019. Saïd
510 Business School, Oxford 10th December 2019

511 Reynolds, C.J., Macdiarmid, J.I., Whybrow, S., Horgan, G., Kyle, J., 2015. Greenhouse gas emissions
512 associated with sustainable diets in relation to climate change and health. *Proc. Nutr. Soc.*
513 74. doi:10.1017/S0029665115003985

514 Simons, L.P., Hampe, J.F., Guldemond, N.A., 2012. Designing Healthy Consumption Support: Mobile
515 application use added to (e) Coach Solution. *Bled eConference* 34.

516 Sra, S., Nowozin, S., Wright, S.J. (Eds.), 2011. Optimization for machine learning. The MIT Press.
517 doi:10.7551/mitpress/8996.001.0001

518 Stawarz, K., Cox, A.L., Blandford, A., 2015. Beyond Self-Tracking and Reminders: Designing
519 Smartphone Apps That Support Habit Formation, in: *Proceedings of the 33rd Annual ACM*
520 *Conference on Human Factors in Computing Systems - CHI '15*. Presented at the the 33rd
521 Annual ACM Conference, ACM Press, New York, New York, USA, pp. 2653–2662.
522 doi:10.1145/2702123.2702230

523 Stuart E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical*
524 *science : a review journal of the Institute of Mathematical Statistics*, 25(1), 1–21.
525 doi:10.1214/09-STS313

526 Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J.,
527 Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T.,
528 Peakman, T., Collins, R., 2015. UK biobank: an open access resource for identifying the
529 causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12, e1001779.
530 doi:10.1371/journal.pmed.1001779

531 Tilman, D., Clark, M., 2014. Global diets link environmental sustainability and human health. *Nature*
532 515, 518–522. doi:10.1038/nature13959

533 Walker, P., Mason, R., Carrington, D., 2019. Theresa May commits to net zero UK carbon emissions
534 by 2050 | Environment | The Guardian. The Guardian.

535 Willett, W., Rockström, J., Loken, B., Springmann, M., Lang, T., Vermeulen, S., Garnett, T., Tilman, D.,
536 DeClerck, F., Wood, A., Jonell, M., Clark, M., Gordon, L.J., Fanzo, J., Hawkes, C., Zurayk, R.,
537 Rivera, J.A., De Vries, W., Majele Sibanda, L., Afshin, A., Chaudhary, A., Herrero, M.,
538 Agustina, R., Branca, F., Lartey, A., Fan, S., Crona, B., Fox, E., Bignet, V., Troell, M., Lindahl, T.,
539 Singh, S., Cornell, S.E., Srinath Reddy, K., Narain, S., Nishtar, S., Murray, C.J.L., 2019. Food in
540 the Anthropocene: the EAT-Lancet Commission on healthy diets from sustainable food
541 systems. *Lancet* 393, 447–492. doi:10.1016/S0140-6736(18)31788-4

542 Wrieden, W.L., Leinonen, I., Barton, K.L., Halligan, J., Goffe, L., 2017. Is the UK diet sustainable?
543 Assessing the environmental impact, cost and nutritional quality of household food
544 purchases. *Proc. Nutr. Soc.* 76. doi:10.1017/S0029665117001811

545 Wrieden, W., Halligan, J., Goffe, L., Barton, K., & Leinonen, I. 2019. Sustainable Diets in the UK—
546 Developing a Systematic Framework to Assess the Environmental Impact, Cost and
547 Nutritional Quality of Household Food Purchases. *Sustainability*, 11(18), 4974.
548 doi:10.3390/su11184974

549