Published in 'Violence Against Women'. Accepted for publication: 14 October 2025

Title: The Meaning of Missing: The Hidden Power of Police Data Recording Practices in Rape Cases

Authors
Gordana Uzelac (London Metropolitan University)*, David Buil-Gil (University of Manchester), Katrin Hohl (City St. George's, University of London), and Jo Lovett (London Metropolitan University)
*Corresponding author: g.uzelac@londonmet.ac.uk

Abstract

Missing values represent a major weakness of police-recorded crime data. This article examines missing data in rape cases recorded by five police forces in England and Wales between January 2018 and December 2020. A thematic analysis of semi-structured interviews with police practitioners reveals factors that influence missing information in police datasets, and quantitative data analyses explore the volume and patterns of missing data across cases defined by different victim, suspect and offense characteristics. Finally, we investigate the impact of missing data on outcomes of police investigations. We find that missing data is partly explained by victim, suspect and case characteristics and is also associated with outcomes.

Key words: Missingness, Missing Data, Police Data, Case Outcomes

Introduction

The missingness of rape and other sexual offenses from police recorded crime data epitomizes the place sexual violence has long held within policing and society at large: much of it is invisible, or rather, invisibilised. The vast majority of sexual offenses are never reported to the police and, as a result, do not appear in official crime records (Allen, 2007; Carretta et al., 2015; Hohl & Stanko, 2024). Survivors who do report to the police are frequently disbelieved and denied access to criminal justice, reflected in the historical discounting and undercounting of sexual offenses in police recorded data (Jordan, 2004; Kelly et al., 2005; Yung, 2013). Countless police inspections have concluded that police forces routinely fail to record, or inaccurately record suspect, victim and case characteristics on data systems, even when these details are known. For example, victim ethnicity is missing in two-thirds of all police recorded crimes (HMICFRS, 2023). Errors and omissions in recording are particularly prevalent in sexual offenses (HMIC, 2014; Hall, 2022).

Inaccurate and incomplete official crime data can result in underestimates of the number of sexual offences disclosed to the police (Allen, 2007; Hohl & Stanko, 2024). It also limits and, potentially, biases our understanding of the victims, suspects and context of sexual offences as well as our ability to reliably assess whether police forces provide equal service and justice outcomes to all victims and suspects, regardless of their ethnicity, sex, or other characteristics (HMICFRS, 2023; Lovett et al., 2022). Reliable empirical evidence is particularly important in the area of sexual offenses because of the pervasiveness of assumptions and misconceptions surrounding them (Lonsway & Fitzgerald, 1994). Systematic missingness of data, particularly from marginalized victims whose reports may be dismissed due to stereotypes about 'ideal victims' (Jordan, 2004; Lonsway & Fitzgerald, 1994), or because they involve suspects or circumstances that challenge common assumptions and misconceptions about rape, sex and race (Hohl & Stanko, 2015, 2024), for example, may reinforce existing power structures that determine whose experiences are deemed credible or worthy of official documentation. Furthermore, today's policing is driven by performance targets. In a world in which only what is counted counts, complete and accurate statistics matter. Missing and inaccurate data can limit officers' ability to identify repeat suspects, establish the needs of particular groups of victims, or design targeted preventative interventions (Taylor & Gassner, 2010). As such, police data both exemplifies and contributes to our limited and potentially biased understanding of the sexual offending that comes to police attention.

Research on missing data is common across disciplines, but minimal research has been conducted on the effect of missingness on police investigations (Harper et al., 2023). There is little empirical examination of missingness in police recorded rape cases, specifically about: (1) the patterns of missing data; (2) the reasons for missing and incorrectly recorded sexual offenses data; and (3) how this may impact on case outcomes. In this article, we set out to address these questions. First, we examine current debates on the meaning of missingness in datasets in general and examine the established practices for handling missing values. Second, we identify and critically discuss how missing data has been addressed in recent studies that are based on police

data in rape cases in England and Wales. We then draw on semi-structured interviews with police officers and police analysts to understand police crime recording and analysis practices to illuminate the reasons for missing and incorrectly entered data. Next, with reference to large-scale data comprising 37,961 rape offenses recorded in five police force areas in England and Wales between January 2018 and December 2020, we estimate the extent of and patterns of missing data in rape cases and examine the relationship between missing data and case outcomes. We conclude by discussing the wider implications of our findings for understanding the nature of sexual offending and for police practice.

Literature Review

Missing Data in Analyses of Rape Case Outcomes

The poor quality of police recorded data in rape cases is well documented. The Mayor's Office for Policing and Crime (MOPAC) report on rape cases in London (2019, p. 12) states that 'there were variables, such as nationality, which were missing in a substantial number of cases.' Lovett et al (2022, p. 291) warn that 'police data are generally subject to important limitations, including data gaps due to key fields such as sex, age and ethnicity not being completed, as well as errors and inconsistencies.'

Researchers have developed different strategies to mitigate these issues. Some avoid using police administrative data altogether; however, rape estimates derived from victimization surveys are not free from limitations (Koss, 1992). Some researchers create 'parallel' datasets by *manually coding* a sample of police case files (Walker et al., 2021; Murphy et al., 2022), while others avoid missing values by *systematically excluding cases missing* crucial information (e.g., victims' sex) (Hohl & Stanko, 2015; Walfield, 2016). Some reports *include warnings* about the quality of the police data but do not clarify how the problem of missing data is treated in the analysis (Lovett et al., 2022). Finally, some reports fail to provide information about missing data (Hester & Lilley, 2017).

The treatment of missing data is particularly important when some variables with missing information – such as unknown ethnicity or suspect-victim relationship – are included in predictor sets used for modeling case outcomes (Data Analytics Lab, 2020, p. 11) or imply a relationship between missing data and case outcome (Lovett et al., 2022). The impact of these predictors is unclear due to the problem of how missing values are labeled. Missing data may be coded as 'not known' (Data Analytics Lab, 2020) or 'not recorded or unknown', for example, for the suspect-victim relationship (Lovett et al., 2022).

The lack of an agreed convention for labeling missing data raises questions of comparability of findings between different studies. Both labels can be understood as indicating missing values, however, the label 'unknown' is more ambiguous. It could indicate a missing value in the narrow sense of the officer having the information but having failed to record it. However, 'unknown' could also indicate that the variable value is unknown to the police or even the victim (e.g., suspect age), or that conflicting or contradictory information has been gathered, making it unclear. Further, there also appears to be no agreed standard for reporting missing

values for variables included in analysis. Some studies report frequency distribution analysis but do not specify the volume of missing data. For example, MOPAC (2019, p. 35) adds frequency analysis of case characteristics included in the coding framework. The sum of all victim ethnicities (White, Black, Asian and Other) amounts to 100 under the column '% of sample.' Yet, the sum of all frequencies for the four ethnicities accounts for 465 out of 501 cases included in the analysis. Hence, the table fails to indicate whether '% of sample' refers to the percentage of valid cases or of the total number of cases included in the analysis.

Finally, a further issue arising from analyses of police data on rape cases relates to the treatment of binary or dummy variables. Analyses that seek to identify predictors of rape case outcomes tend to use datasets either produced by police forces or created by manual coding of police case files. In both cases, these analyses report use of dichotomous dummy variables when indicating the presence of certain characteristics. For example, Hohl and Stanko's (2015, p. 330) analysis of manually coded police case files stated that all their 'explanatory variables used in the analyses are binary and were coded as 1 if the particular characteristic was present and coded as 0 if it was absent.' Case characteristics – such as mental health issues (MHI), domestic abuse (DA), presence of alcohol, drugs or weapons, commonly denoted in police data systems using flags – are traditionally coded in this manner. Whether data were extracted from police systems or case files were manually coded, a value of 1 was assigned if these case flags were mentioned/recorded and with 0 if not. However, as MOPAC states in its reports, 'many of the variables were coded only for their *presence* which means we cannot determine whether the *absence* of a variable is due to omission in data recording or it not being a factor in the case' (MOPAC, 2019, p. 12). This has implications when drawing inferences from the results of analyses which are often given insufficient consideration or neglected altogether.

Murphy et al. (2022, p. 16) state that 'only 7% [of victims] were recorded as having a mental health issue.' However, since the remaining 93% were all coded as 0, it is impossible to distinguish between those with no MHI and those whose mental health status was unknown; that is, where data were missing. The use of dummy variables in this case only allows us to infer whether certain characteristics, like MHI, are recorded or not. For example, Walker et al. (2021: 15) correctly report that 'older complainants were significantly more likely to be *recorded* as having MHI.' The lack of clarity regarding the use of the value '0' in coding case characteristics prevents examination of the impact of missing values in such cases. The problem starts with the process of recording information and could easily be embedded into the computer system in use, by giving the officer the option of recording a value of 'unknown.'

In sum, the treatment of missing values in the analysis of rape cases in England and Wales suggests a lack of standardized procedures for coding and reporting missing data. There is awareness of the problem of missing data and, while some note the likely impact of missing data on outcomes of rape investigations, existing studies do not statistically model the impact of missing data on their outcome variable(s) of interest and are thus unable to fully consider the impact on research findings and conclusions.

Missing Data, Bias and Measurement Error

Missing data can create bias and error. An analysis of data with missing values could systematically produce over- or underestimation of relevant parameters. Those biased parameters – e.g., a correlation or regression coefficient – would not be an accurate measure of the variable of interest in the population. The analysis of missing values in a dataset is important because the 'ultimate consequence of missing data is distortion from the truth; reducing the internal and external validity of study results' (Hardy et al., 2009, p. 2). Missing data can introduce bias into the analysis because cases with missing values could be systematically different from those where data is not missing (Rubin, 1987).

Of course, not all missing values would have the same effect on the analysis. An occasional missing value that occurs by accident in the process of collecting or inputting data need not affect the analysis, provided data are missing 'completely at random.' Missing data could be a result of an intentional decision to not provide requested information, or an unintentional act of forgetting to provide or input data, but it could also arise from technical errors embedded in the data system in use and the nature of data collection (Newman, 2014). Therefore, problems might occur if missing data is not completely at random and accidental, but systematic and patterned. For example, if the police data about rape cases fails to record an occasional ethnicity of a suspect, this may be inconsequential and the estimation of parameters within the analysis may not be biased. However, if the ethnicity of a particular group of suspects is systematically missing, then any analysis of such data would be biased. Even if data are not missing at random, provided the process by which data are missing is known, it may be possible to statistically account and correct for systematically missing data in the analysis.

Traditionally, studies based on police datasets handle missing data by excluding variables with a higher level of missing values from the analysis, by excluding cases with more missing values, or just by reporting but ignoring the number of missing values. More attention has been given to the problem of missing data especially after the publication of Little and Rubin (1987) and Rubin (1987). These works triggered a debate on the meaning of missingness and the development of methods for handling missing data, both generally and in application to particular datasets (Pina-Sánchez et al., 2023). A certain level of consensus on the stages of missing data analysis has been reached. For example, studies should include an analysis of *patterns* of missing data (what data are missing) and *mechanisms* of missing data (why data are missing) (Enders, 2010, p. 2). An analysis of missing data patterns identifies 'holes' in the data. Hence, the first stage of every analysis of missing data should include frequency analysis for every variable where variables with a substantial percentage of missing data are identified. According to Hardy et al. (2009, p. 6), the next stage of the analysis of missing data should include the 'characterization of missing data.' This process attempts to identify patterns of missingness where the analysis examines whether the presence of missing data is 'related to other known factors.'

Focusing on missing data in survey research, Newman (2014, p. 373) identifies three patterns of missing data: item-level, construct-level, and person-level. Item-level missingness

occurs when the respondent leaves a few items blank on a multi-item scale. Construct-level missingness occurs when the respondent omits an entire scale or an entire multi-item construct. Finally, person-level missingness 'involves failure by an individual to respond to any part of the survey' (Newman, 2014, p. 375). When applied to data collected by the police, *item-level missingness* occurs where an occasional value is missing within a specific crime record. Construct-level missingness would refer to the *case profile-level missingness,* where segment refers to one of the three main profiles of any case such as, victim and suspect profile, and offense and procedural characteristics. Person-level missingness applied to police administrative data would equate to *case-level missingness,* that is, failure to record any significant information about a rape disclosed to police. An analysis of missing values in a police dataset on rape cases should distinguish between cases where one profile of the case data is missing (e.g., all information on the victim or suspect) and cases where practically no information on any profile of the case is available. The level of missingness, Newman (2014) claims, can determine how the missing data is treated.

Once the patterns of missingness are examined, the analysis should seek to establish the data-generating process, or rather the process that produces missingness. Little and Rubin (1987) made a distinction between three categories of missingness mechanisms: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). 'If the likelihood of being missing is not related to either the value of the missing variable or to the values of any other variables in the data set' it would be a case of MCAR (Hardy et al., 2009, p. 14). In this case, missingness is not related to any aspect of case characteristics. Hardy et al. (2009) define MAR when the likelihood of missing data can be completely explained by other variables in the analysis, that is, when there is a systematic relationship between measured variables and the likelihood of missing data. For example, a MAR mechanism would be the case if missing values on victim age could be explained by third party reporting or suspect(s) not being identified. This will rarely be the case, as other variables might explain some missingness but rarely completely. Finally, Hardy et al. (2009) define MNAR occurring 'if missing values are not randomly distributed across participants, and the probability of being missing cannot be predicted from the other variables.' For example, missing values on victims' ethnicity may be related to the workload of the police officer or their reluctance to ask for that data. MNAR covers the majority of cases (Hardy et al., 2009). Graham (2012, p. 18) claims researchers cannot know 'whether missingness is MAR or MNAR in any particular case.' The hypothesis that MAR holds cannot be tested, except by collecting missing data. Researchers do not know why the data are missing, and therefore, it is not possible to describe the probability of missing data with any certainty (Enders, 2010). For practical reasons, Nakagawa and Freckleton (2008) suggest that all missing data should be treated under the assumptions of MAR. To what extent missing data is harmful for the analysis will depend on 'the amount of missing data, the pattern of missing data, and whether the data are missing in a strongly systematic […] fashion' (Newman, 2014, p. 372).

Against this backdrop, the aims of this study are threefold: (1) to examine patterns of missingness in the police datasets on rape in England and Wales; (2) to examine the effect of

missing data on outcomes of those cases, and (3) to assess possible reasons for missing and incorrectly entered data from the perspective of those who use and create the data.

Data and Methods

Data

Five police forces in England and Wales provided anonymized data on all rape cases recorded by them between January 2018 and December 2020. These cases amounted to a total of 47,213 recorded rape offences (for details, see Stanko, 2022). The five forces analysed in this study cover over 25 per cent of the population of England and Wales and represent diverse forces in terms of their rural-urban classification, size, and crime levels. Due to data confidentiality agreements, the names of police forces are anonymised in this study, and we refer to them as 'Police force A' to 'Police force E.' This was part of a larger action research project called Soteria (Stanko, 2022).

The research team provided a list of required variables to all forces included in the project, and the forces extracted this information about each case from their systems, collating it in an Excel spreadsheet in order to share it with us. The dataset included victim and suspect characteristics, offence and procedural characteristics, and case outcomes. For this study, all duplicate entries (including cases with multiple suspects, multiple victims or more than one crime classification) were removed. We also removed open cases (i.e., cases that were still ongoing and without a defined outcome), since the amount of missing data in the case may still change (typically decrease) as the investigation progresses. As a result, the dataset in this analysis consists of 37,961 cases.

While all police forces in England and Wales use a computerized data recording system, there is no unified system. Among the five forces participating in this study, there were four different systems in use. One force had changed to a new system a year before the project commenced. All systems include a combination of structured and unstructured data. Only some of these structured data are mandatory, such as crime classification and outcomes, as they must be reported periodically to the Home Office. However, due to the use of these different systems, only a small number of variables were available in comparable form across all five forces. These common variables include victim characteristics (age, sex and ethnicity), suspect characteristics (age, sex and ethnicity), offense characteristics (suspect-victim relationship, number of days from the offense to the police report, and number of days from the report until the investigation is closed), and case outcome. Ethnicity is recorded as police-defined ethnic appearance. Our analysis examines the impact of missing data on recorded rape case outcomes. The crime outcomes framework was introduced in 2013 and there are currently 22 possible outcomes, which were applied in the three-year dataset used in this analysis. For more detailed discussion on the Outcome Framework development, see Home Office (2023). We recoded this outcome framework into fewer categories to allow more meaningful quantitative analysis of the most

relevant outcomes (see Table 1; and detailed description of outcomes in Table A1 in the Appendix). The final crime outcome variable consisted of four categories as follows:

Table 1. Outcome groupings (details in Table A1 in the Appendix)

Outcome grouping	Home Office crime outcomes included
1 Charged	1, 2 and 3
2 Evidential difficulties: attributed to victim	14 and 16
3 Evidential difficulties: investigative	15 and 18
4 Prosecution prevented or not in the public interest	5, 9-13, 17, 21

Note: Outcomes 2 and 3 (cautions) were included with Outcome 1 because they are a form of sanction/acknowledgement that a crime has been committed. Outcome 19 did not feature in the dataset, as it relates to fraud cases only. All other outcomes (2, 3, 4, 6, 7, 8, 20 and 22) were excluded from the analysis, as they had marginal relevance and were only applied in a small number of cases, making groupings too fragmented for analysis.

The statistical analysis of the three-year dataset was complemented by qualitative analyses of semi-structured interviews with 32 police officers in operational and senior leadership roles, and data analysts (intelligence and performance) from across the five police forces. Interviews allowed us to capture the officers and analysts' experiences using the data recording systems and conducting data analysis as part of their routine tasks.

Methods

The analysis is structured in two parts: first, we present relevant sections of the thematic analysis that is directly related to the issue of missing data. The data consists of a set of 32 semi-structured interviews with police officers and police analysts to understand police crime recording and analysis practices and illuminate reasons for missingness and error in police data. Second, we draw on our large-scale dataset of all police recorded rapes across five police forces (n=37,961) to estimate the extent of and patterns in missing data and examine the impact of missing data on case outcomes.

For the purposes of the wider project, 32 qualitative semi-structured interviews were conducted with strategic and operational police leads, analysts, crime management staff, and force crime registrars in each of the five forces. Participants were selected within each force as those most relevant to the area of enquiry. Among others, the interviews explored their experiences of data use, performance monitoring, crime recording, and case progression processes within their force. The data that directly addresses these issues were thematically analysed using NVivo 12.

Thematic analysis was chosen for the flexibility of its approach and ability to identify manifest and latent content (Braun & Clarke, 2021). The coding process was applied on data that directly addressed the description of data inputting and analysis practices and evaluation of the data, including issues with missing values. Interview transcripts were analysed first separately for each police force and then analytically compared. After conducting initial coding of relevant text, themes were generated through the identification of recurring patterns. A level of rigour of the analysis was achieved through review and feedback from academics, police officers, specialist support service providers from the third sector and other subject matter experts as part of the project's quality assurance process.

Turning to the quantitative part of the study, the police dataset analysis aims to illuminate (a) types of missingness in the data; (b) the extent of missingness across different victim groups, suspect groups, and offence types; and (c) the impact of missing data on case outcomes. To do so, we first identify 'missing' data in our dataset by coding all cases in which variables contain empty records (i.e., no data), as well as information coded as 'not recorded', 'indeterminate', 'not specified' and other similar data entries. This allows for a comprehensive overview of the range of presentations of missingness in our data as they are used within and between police forces.

Secondly, we descriptively present the proportion of missing values across variables and police forces and utilise t-tests and ANOVA to explore whether differences in missingness are statistically significant across victim groups (e.g., males versus females), suspect groups (e.g., by suspect ethnicity), offence characteristics (e.g., by suspect-victim relationship), and across police forces. Then, we use regression models to estimate the combined influences of victim, suspect and case characteristics on the probability that cases have at least one missing value (binary: '0' for no missing data and '1' for at least one missing value), as well as on the total number of missing variables per case (numeric: 0 to 9). Logistic regression is used to analyse the binary measure of at least one missing value, and Poisson models to estimate the numeric variable of 'number of missing variables' (Osgood, 2017). In both cases, fixed effects are applied to police force areas to capture the effect of police force characteristics, including their size and workload, whether they cover mainly rural or urban areas, and any other effect directly associated with police forces rather than cases. Regression estimates are standardised to aid comparison and interpretation. Finally, we aim to gain a deeper understanding of the impact of data missingness on case outcomes. In other words, we test whether cases with missing data in key variables, and a larger amount of missing data overall, tend to be less likely to result in a suspect being charged (binary measure: '0' no charge and 1 'charge'). We once again make use of logistic regression models with fixed effects to control for the effect associated with police forces. We present standardised coefficients in tables. All data analysis has been conducted in R Software (R Core Team, 2024).

The analysis script used in this study is available on GitHub (https://github.com/davidbuilgil/meaning_missing) for reproducibility purposes, although the

underlying data cannot be shared due to confidentiality and data-sharing agreements with the participating police forces.

Findings

The findings section begins with the results of the thematic analysis of qualitative data to identify issues with data in general and missing data in particular from the standpoint of practitioners. We then analyse the quantitative database to explore patterns of missing data in records of rape investigations, including the labels used to capture missing information and the proportion of missing values across variables and cases. Finally, we investigate the impact of missing data on outcomes of police investigations.

Evaluation of Data Systems

The following section presents the findings from qualitative semi-structured interviews with participants across five police forces, analysed using thematic analysis. The analysis identified three main themes that reflect shared perspectives, recurring challenges, and differences in experiences regarding the use, management, and quality of police crime data systems: participants' assessment of system functionality, factors contributing to system limitations, and adaptive strategies for addressing system limitations.

Participants' Assessment of System Functionality

Across the five police forces, interviewees recognised certain **strengths in the computer systems** used to record, manage, and analyse crime records. Many described them as comprehensive and 'very good for larger amounts of data' [Detective Inspector], especially for monitoring workloads. Analysts noted the potential for 'data validity exercises' [Performance Analyst], suggesting that, in principle, the systems could be valuable tools for improving data quality and operational efficiency. However, several participants stressed that this potential was not always realised in practice.

The most frequently reported drawbacks were **the inflexibility of the systems and navigational challenges**. Because most systems are designed by external software companies, force-specific adjustments were often seen as overly bureaucratic and slow. As one Intelligence Analyst explained: 'If you want to make a change… some things can be done in force, [but] some… need to go to [an external country] to get done.' Navigational difficulties also emerged as a recurring concern. One system was described as a 'monster' [Intelligence Analyst], while a newly introduced platform was labelled 'clunky' by a Detective Chief Inspector: 'What took two clicks now takes 19.'

The participants consistently emphasized **inefficient data entry** and problems with system functionality. Data input was described as 'time consuming' [Intelligence Analyst], with systems designed primarily for record-keeping rather than analysis. Analysts reported that there was 'a lot of clicking' and 'no quick way' to input information, nor were there exportable fields

[Intelligence Analyst]. One Performance Analyst noted the extra step of having to access data via a separate system 'that plugs into the back,' further complicating workflow.

Factors Contributing to System Limitations

Many interviewees questioned the reliability and quality of the data produced by these systems, with one Detective Chief Inspector calling the data 'anecdotal' and a Chief Superintendent stating: 'I've got no confidence in that dataset.' Missing data was identified as the most common cause of this mistrust. Frequently missing fields included incident location, suspects' date of birth, ethnicity of victims and suspects, suspect-victim relationships, reporting party, and contextual factors such as domestic abuse, alcohol involvement, or mental ill health.

The interviewees identify several factors that increase the volume of missing data and affect reliability, which can be categorized into structural and subjective categories.

Identified **Structural or technical factors** include, first of all, the setup of the data system or, more specifically, the requirement for detailed information. As an Intelligence Analyst phrased: 'The more you add on for people to fill in, the greater the data quality issue becomes.' Second, it is noted that important data fields are not made mandatory to complete, such as victim or suspect ethnicity. The third problem relates to the form of the data. on the one hand, the interviewees stress the importance of introducing binary data fields (e.g. 'Yes/No' variables) to make data analysis and comparison easier. On the other hand, they highlight a lack of qualitative data. Further on, they stress how overly detailed variable entry options affect the reliability of analyses: 'there's about 40 different classifications that you have to go through and almost then make a judgement call through using filters' [Intelligence Analyst. Finally, another Intelligence Analyst points at the high level of demand and limited capacity: '[Call handlers are] logging lots of calls and I can imagine they're just trying to do them as quickly as possible, but the knock-on effect on analysis and the misinterpretation of data because those fields are not being entered properly, is huge - it totally changes our profiles.'

Among the **subjective or human factors** that link to poor data quality, the most common issues mentioned are low data literacy and insufficient training. This includes problems with some users who are unfamiliar with the system due to recent changes or the use of online-only training, which was perceived as less effective. Another issue stressed relates to the low attention to detail. For example, one interviewee emphasised how 'people are busy and they're out and about, or they're filling it in on their tablets. They're not really thinking about two months down the line, some intelligence analyst is going to need that data' [Intelligence Analyst]. Finally, respondents identified a lack of awareness of the importance of data quality among police officers as another of the subjective factors that affected the quality of data.

Adaptive Strategies for Addressing System Limitations

In response to these challenges, some forces developed **parallel datasets** maintained by analysts for analytical purposes. These datasets, described as 'a lot cleaner and more carefully inputted' [Intelligence Analyst], were often compiled manually from incident logs and free-text reports. While the process was labour-intensive – described as a 'time-consuming manual trawl' [Analyst] – it produced more reliable data for analysis. However, analysts typically lacked permission to update or correct the official police data systems directly, limiting the potential for system-wide data quality improvements.

Participants' evaluations of the crime data systems make clear that their reliability is inseparable from the quality and completeness of the data they produce. While the systems were recognised as potentially valuable for managing workloads and supporting analysis, this value was often compromised by inflexible design, cumbersome data entry processes, and limited training. These issues directly contributed to the high volume of missing data reported across forces – particularly in key fields such as incident location, suspect and victim characteristics, and contextual factors – undermining confidence in the accuracy of the datasets.

Patterns of Missing Data in Five Police Forces

The dataset on rape cases in the period 2018-20, collected from the five police forces, has information on 37,961 unique rape cases where nine variables (those listed in Table 2 plus the outcome variable) are deemed comparable across police forces. As indicated above, the starting point of our analysis of missingness focuses on 'counting missing values' in the data. Different data recording systems used in each of the five forces use various labels to categorise missing information. The collected Excel dataset that combined the extracted data from each force contained 17 different labels that indicate some form of missingness (see Table 2). Police forces used eleven different labels of missingness for the variable 'sex of suspect' and ten for 'sex of victim.' Data from Force D contains ten different labels for missing values over eight variables listed, while the data from Forces B and E reveal six. Differences between labels are not always easily distinguished, such as in the case of 'not recorded' or 'indeterminate' and 'not specified.' Some codes clearly indicate that the field was left empty (such as, ' ', '#VALUE!', 'NULL' or '-'). Labels such as 'not identified', 'not/unspecified' and 'indeterminate' might imply that the information was looked for but not obtained. The label 'not/none recorded' suggests that missingness might be a consequence of the data input procedure. While we can acknowledge that some labels of missingness try to capture the causes of missingness, it is doubtful whether the nuances between the 17 labels are justifiable.

Table 2. Labels for missing data on eight variables in five forces

	Force A	**Force B**	**Force C**	**Force D**	**Force E**	**Number of codes**
Victim sex	'NULL', 'U', ' '	'none recorded', 'not specified', 'decoy'	'#N/A'	'#N/A', 'indeterminate',	' ', 'indeterminate', 'Unknown'	10

				‘not recorded’, ‘NULL’		
Victim age	‘#VALUE!’	‘none recorded’	‘#VALUE!’	‘#N/A’, ‘NULL’	‘ ’	5
Victim ethnicity	‘#N/A’, ‘ ’	‘decoy’, ‘none recorded’, ‘unknown’	‘Missing/not known’	‘#N/A’, ‘not identified’	‘unknown’, ‘ ’	7
Suspect sex	‘NULL’, ‘U’, ‘ ’	‘unknown’, ‘none recorded’, ‘not specified’, ‘decoy’, ‘ ’	‘-’, ‘#N/A’, ‘unknown’, ‘unspecified’	‘not recorded’, ‘NULL’, ‘unknown’, ‘indeterminate’	‘ ’, ‘unknown	11
Suspect age	‘#VALUE!’, ‘#N/A’, larger than 99	‘ ‘, ‘not recorded’	‘#N/A’, ‘-’	‘NULL’	‘ ‘, less than 0	8
Suspect ethnicity	‘#N/A’, ‘ ’	‘not recorded’, ‘none recorded’, ‘unknown’, ‘ ’	‘#N/A’, ‘missing/not known’	‘missing/not known’, ‘not identified’	‘unknown’	7
Suspect-victim relationship	‘Not recorded or unknown’	‘none recorded’, ‘unknown’, ‘ ’	‘Not recorded or unknown’	‘Not seen by victim’, ‘NULL’, ‘Victim refuses to identify’	‘Not recorded/ unknown’	7
Time incident-report	none	‘#VALUE!’	none	‘#VALUE!’	none	2
Number of codes	8	6	9	10	6	

This descriptive exercise does not allow inferences on the reasons for missingness (e.g., whether the victim/suspect/third party does not provide the information, the officer fails to record it, or the police are unable to interview the victim or suspect). Table 2, however, makes clear that differing computer systems used across forces add further complexities. While it would have been informative to further explore differences in the possible reasons for ‘missingness’ in our data, the inconsistency of labels within and across police forces makes this analysis impossible. We, therefore, coded all labels reflecting missingness as ‘unknown’ to make further analysis possible, thus potentially collapsing a variety of mechanisms of missingness – from victims/third parties failing to provide information, to police forces failing to include data in records, to cases in which the suspect was never identified.

Table 3 reports the frequency of missing records for each variable, grouped by ‘case profiles’ (i.e., victim, suspect and offence characteristics) and by police force. Across all forces, suspect ethnicity is the variable with the largest proportion of cases with missing information (50%), followed by suspect age (35%), suspect-victim relationship (34%) and victim ethnicity (30%). The variable with the fewest missing values is victim sex (1%), followed by the time between report and outcome (2%), the time between incident and report (3%), and victim age

(3%). Overall, missing data appears to be more common for suspect characteristics (on average, 1.11 out of 3 variables missing across all forces; median: 1) than case characteristics (on average, 0.39 out of 3; median: 0) and victim characteristics (on average, 0.34 out of 3; median: 0). This is to be expected; the police typically have knowledge of the victim in an incident and/or contact with them, while a suspect cannot always be identified. Suspect and victim ethnicity is less frequently recorded compared to age and sex. Overall, 75% of cases have at least one of these key variables missing. Only 4 cases in the whole dataset (0.01%) have all variables missing.

While missing data is common across all forces, it is more frequent in Forces D (on average, 2.42 variables missing out of 9; median: 2), C (average, 2.34; median: 2) and E (average, 2.15; median: 2) than in Forces B (average, 1.78; median: 1) and A (average, 1.45; median: 1). Force B has more missing values than all other forces for the variables capturing the time between the incident and the police report, and time between the report and case outcome, while police force A has more missing values than police forces D and E for the suspect ethnicity variable. We found no evidence that the presence of missing data across police forces was related to the amount of cases (caseload) in each police force, the police force size (workforce), or the population size (see Figure A1 in the Appendix).

Table 3. Frequency and percentage of missing values for each variable and police force

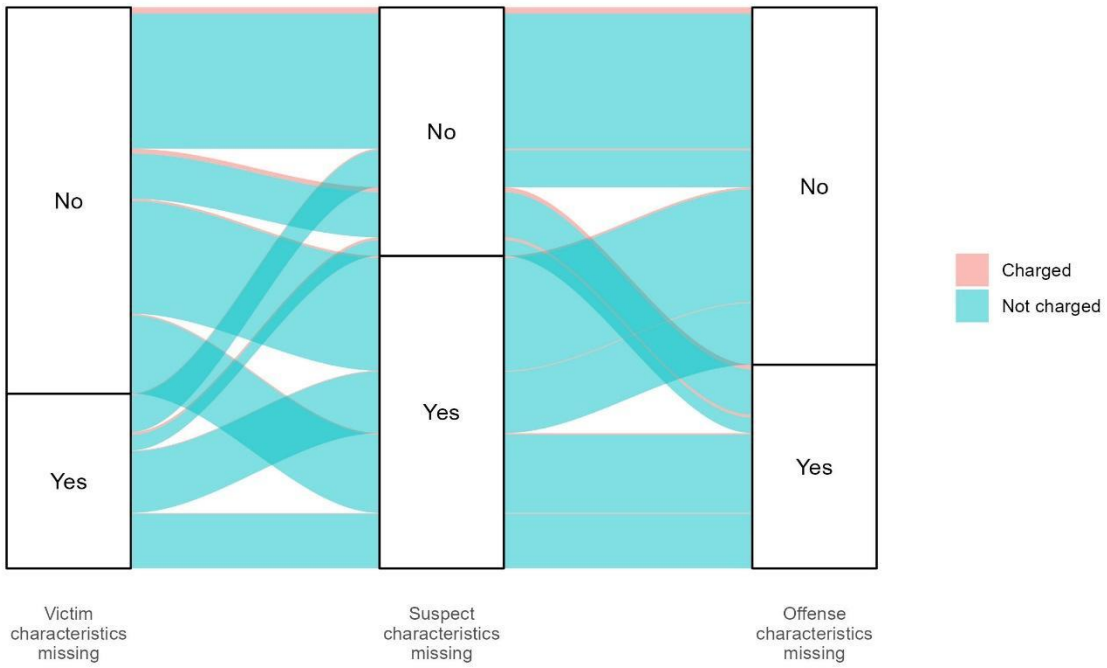
	Force A	Force B	Force C	Force D	Force E	All forces
Victim characteristics						
Victim sex	0.47%	1.32%	0.00%	1.19%	1.83%	241 (0.63%)
Victim ethnicity	18.02%	8.41%	54.84%	42.31%	34.39%	11,381 (29.98%)
Victim age	3.97%	0.35%	1.11%	5.41%	0.85%	1,103 (2.91%)
Suspect characteristics						
Suspect sex	12.63%	33.03%	49.54%	36.95%	32.48%	10,093 (26.59%)
Suspect ethnicity	49.33%	29.48%	57.06%	48.63%	49.18%	18,964 (49.96%)
Suspect age	34.32%	23.66%	37.50%	38.82%	34.54%	13,258 (34.93%)
Offense characteristics						
Suspect-victim relationship	24.39%	35.61%	33.79%	63.93%	52.84%	12,948 (34.11%)
Time incident-report	0.28%	17.53%	0.00%	4.87%	8.52%	952 (2.51%)
Time report-outcome	1.60%	28.93%	0.00%	0.00%	0.04%	884 (2.33%)
Overall summary						
At least one variable missing	68.49%	63.58%	82.22%	87.99%	82.81%	38,401 (74.82%)

Since missing data differs across variables and police forces, we explored whether cases with more missing information in one of our three 'case profile' sets of variables also displayed more missing information across the other two 'case profiles.' One might expect, for instance,

that cases in which there is missing victim data will also suffer from missing suspect and offence data. This could happen where the victim does not support the investigation and therefore chooses not to speak to the police or does not provide full information. We analyse the Pearson's bivariate correlations of the number of missing variables per 'case profile' and observe a moderate correlation between the amount of missing values in suspect and offence data ($r=0.27$, $p<0.001$). Such a correlation, although still statistically significant, is very weak in the case of victim and suspect data ($r=0.18$, $p<0.001$) and victim and offence data ($r=0.09$, $p<0.001$). This is apparent from Figure 1, which shows that the majority of cases with missing offence data also suffered from missing suspect or victim data. Nevertheless, a non-negligible number of cases with missing victim data did not have missing suspect data, which could be explained by third-party reporting not supported by the victim, and many of those with missing suspect data were not missing other offence data. Some cases have missing victim but not suspect data, and others have missing suspect but not offence data. The mechanisms that explain missing information across case profiles are, therefore, far more complex and nuanced than a linear causal process in which suspect variables are dependent upon successful completion of victim characteristics and offence variables are, in turn, dependent upon completion of victim and suspect characteristics.

Figure 1 further displays the proportion of cases with 'charged' and 'not charged' outcomes. While the proportion of charged cases remains remarkably small, it is comparatively greater in cases without missing data. We will revisit this observation below.

Figure 1. Alluvial plot of missing values across three 'case profiles'



Next, we explore whether missing data is more prevalent for some types of victims and suspects than others. For instance, we investigate whether the total amount of missing data in a case (min: 0 and max: 9 variables) is more prevalent when the victim is male or female, and whether it varies across victim and suspect age and ethnicity. In other words, we aim to explore whether cases with a given victim, suspect or offense category (e.g., young victim) are typically defined by more missing information - in other variables - than others (e.g., older victim). Results are presented in Table 4. With the sole exception of suspect sex, all other variables in our data show statistically significant differences in the amount of missing data. With respect to victim characteristics, missing data is more prevalent when the victim is male, young, and White or Black as opposed to other ethnicities. Interestingly, similar patterns emerge when we look at suspect characteristics: missing data is more common in cases involving young, White and Black suspects than all others. As expected, missing data is more frequent when the suspect was a stranger to the victim than when they were family, acquaintance/friend or current/ex-partner (i.e., the victim knew the suspect characteristics even before the incident). Interestingly, cases reported to the police on the same day as the incident tended to suffer from more missing values than those reported 1 to 100 days after the incident, and the amount of missing data again increased in historical allegations reported over 100 days later than the incident. Cases closed after more than 100 days of investigation showed fewer missing values than those that reached an outcome earlier.

Table 4. T-tests and ANOVA analysis of missing data by categories of each variable

Variable	Test	Result	Mean missing variables per category
Victim characteristics			
Victim sex	Two Sample t-test	t = -19.31***	Male: 2.29; Female: 1.77
Victim age	ANOVA	F = 84.69***	Under 18: 1.91; 18 to 25: 1.71; 26 to 40: 1.61; Over 40: 1.90
Victim ethnicity	ANOVA	F = 96.52***	Asian: 1.04; Black: 1.28; White: 1.47; Another ethnic background: 1.17
Suspect characteristics			
Suspect sex	Two Sample t-test	t = -1.21	Male: 1.13; Female: 1.02
Suspect age	ANOVA	F = 76.07***	Under 18: 1.14; 18 to 25: 1.00; 26 to 40: 0.86; Over 40: 0.96
Suspect ethnicity	ANOVA	F = 63.39***	Asian: 0.54; Black: 0.57; White: 0.71; Another ethnic background: 0.47
Offense characteristics			

Suspect-victim relationship	ANOVA	F = 591.12***	Acquaintance/friend: 1.19; Family: 1.20; Current/ex-partner: 0.85; Stranger: 1.78; Other: 1.24
Time incident-report	ANOVA	F = 23.79***	Same day: 1.89; 1 to 2 days: 1.73; 3 to 10 days: 1.72; 11 to 100 days: 1.65; Over 100 days: 1.79
Time report-outcome	ANOVA	F = 570.40***	Same day: 2.13; 1 to 2 days: 2.75; 3 to 10 days: 2.70; 11 to 100 days: 2.05; Over 100 days: 1.50

***p<0.001; **p<0.01; *p<0.05

Some of these differences are further reflected in the results of the multivariate regression models shown in Table 5, including a logistic regression analysis of the binary outcome of at least one missing value (Model 1), and a Poisson regression exploring the number of missing variables per case (Model 2). Cases with male victims are more likely to have at least one missing value (Model 1), and more missing values overall (Model 2), than those with a female victim ($p<0.001$). Cases with older victims – 26 and over – are less likely to have missing data than those with younger victims (though this is not reflected in Model 2 for victims aged over 40). Additionally, cases in which the victim is Black, Asian or from another ethnic minority background are less likely to have missing data (Model 1), and have fewer missing variables (Model 2), than cases with White victims (all at $p<0.001$ level). With regard to suspect characteristics, cases involving younger suspects are more likely to have at least one missing value (Model 1) and have more missing data (Model 2) than cases with older suspects (all at $p<0.001$ level). Cases with suspects who are Black, Asian or from another ethnic background are also less likely to have at least one missing variable and suffer from less overall missingness than cases with White suspects (all at $p<0.001$ level). All other suspect-victim relationships are less likely to have at least one missing value (Model 1), and fewer missing values overall (Model 2). While the victim will typically know the characteristics of the suspect when they are or were in an intimate relationship, acquiring full data during the investigation appears more challenging in these cases than in situations where the suspect is an acquaintance or friend, family member, or even a stranger. Cases that take longer to reach an outcome are more likely to have missing values and have more missing data than cases closed on the day of the report. Cases reported after the day of the incident tend to be less likely to have missing data and have fewer missing values than cases reported on the day of the incident, though not all categories show statistically significant effects. Overall, we find evidence that the characteristics of the case, including the characteristics of the victim, suspect, and offence itself, explain a notable proportion of the variation of missing data.

Table 5. Fixed effects logistic regression (Model 1) and fixed effects Poisson model (Model 2) of missing data. Fixed effects considered for police force areas. Standardized coefficients

	Model 1: One missing variable		Model 2: Number of missing variables	
	Beta	CI	Beta	CI
Victim characteristics				
Victim sex (ref: female)				
Male	1.07***	1.03 – 1.10	1.02***	1.01 – 1.03
Victim (ref: under 18)				
18 to 25	1.02	0.98 – 1.06	1.01*	1.00 – 1.02
26 to 40	0.91***	0.87 – 0.94	0.99*	0.98 – 1.00
Over 40	0.92***	0.88 – 0.95	1.00	0.99 – 1.01
Victim ethnicity (ref: White)				
Asian	0.90***	0.88 – 0.93	0.95***	0.94 – 0.96
Black	0.95***	0.92 – 0.98	0.96***	0.95 – 0.96
Another ethnic background	0.94***	0.91 – 0.96	0.97***	0.96 – 0.98
Suspect characteristics				
Suspect sex (ref: male)				
Female	0.98*	0.95 – 1.00	0.99**	0.98 – 0.99
Suspect age (ref: under 18)				
18 to 25	0.48***	0.47 – 0.50	0.71***	0.71 – 0.72
26 to 40	0.38***	0.36 – 0.39	0.63***	0.63 – 0.64
Over 40	0.46***	0.44 – 0.47	0.69***	0.69 – 0.70
Suspect ethnicity (ref: White)				
Asian	0.71***	0.70 – 0.73	0.80***	0.79 – 0.81
Black	0.64***	0.62 – 0.65	0.76***	0.75 – 0.75
Another ethnic background	0.85***	0.83 – 0.87	0.89***	0.88 – 0.91
Offense characteristics				
Suspect-victim relationship (ref: current/ex-partner)				
Acquaintance or friend	0.71***	0.69 – 0.73	0.84***	0.84 – 0.85
Family	0.77***	0.75 – 0.79	0.88***	0.87 – 0.89
Stranger	0.93***	0.90 – 0.96	0.91***	0.90 – 0.92
Other	0.97**	0.94 – 0.99	0.97***	0.96 – 0.98
Time incident-report (ref: same day)				
1 to 2 days	0.95**	0.92 – 0.98	0.97***	0.96 – 0.98
3 to 10 days	0.98	0.95 – 1.01	0.98***	0.97 – 0.99
11 to 100 days	0.94***	0.91 – 0.97	0.96***	0.95 – 0.97
Over 100 days	1.02	0.98 – 1.06	0.98***	0.97 – 0.98
Time report-outcome (ref: same day)				
1 to 2 days	0.82***	0.79 – 0.86	0.97***	0.97 – 0.98
3 to 10 days	0.78***	0.75 – 0.82	0.97***	0.96 – 0.98

11 to 100 days	0.52***	0.48 – 0.56	0.89***	0.87 – 0.90
Over 100 days	0.46***	0.42 – 0.50	0.85***	0.83 – 0.86
Observations	37,961		37,961	
Pseudo R^2 Tjur	0.281		0.851	
Pseudo R^2 McFadden	0.260		0.235	
Pseudo R^2 Nagelkerke	0.375		0.591	

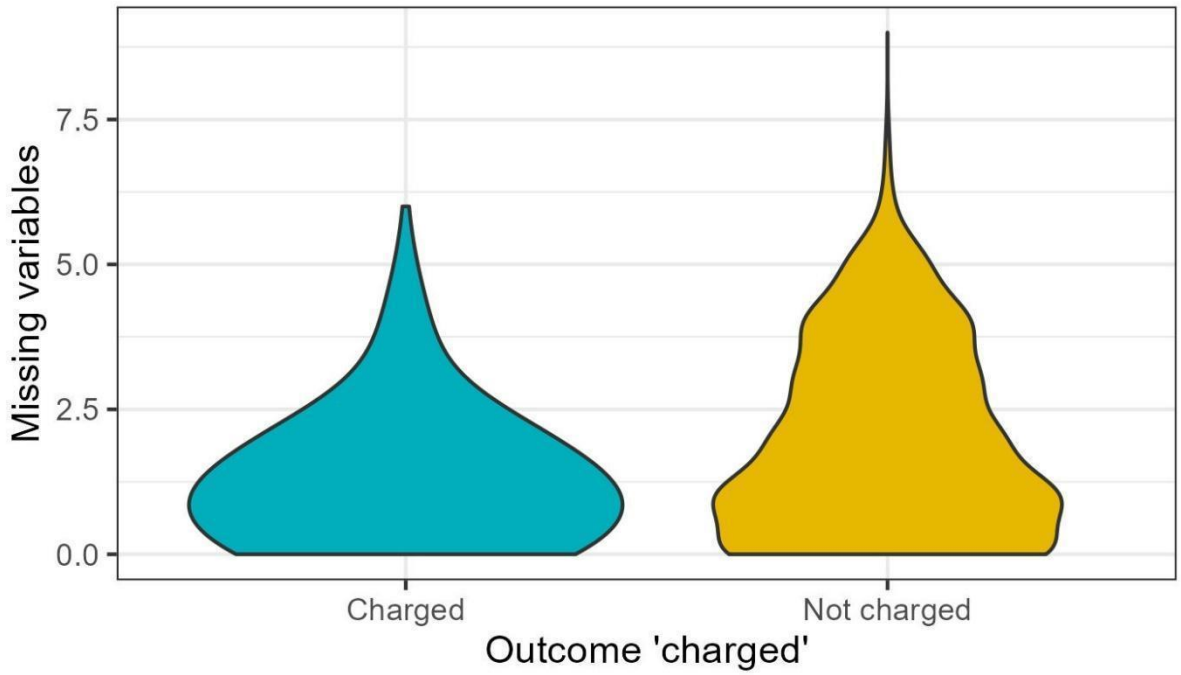
***p<0.001; **p<0.01; *p<0.05

Assessing the Effect of Missing Data on Rape Case Outcomes

We explore whether cases with more missing values tend to lead to a suspect being charged less often than those without missing data. Moreover, we analyse which missing values, for which variables, are more closely associated with the outcome of a police investigation.

Missing data appears to be associated with outcomes of police investigations. The average number of missing variables in police records significantly varies depending on the outcome of the case: 'Charged' (1.22), 'Evidential difficulties: attributed to victim' (1.70), 'Evidential difficulties: investigative' (2.02) and 'Prosecution prevented' (2.02) (ANOVA: $F = 247.17$; $p<0.001$). Cases that result in a suspect being charged have fewer missing variables (1.22) than cases where no suspect is charged (1.86) (Two Sample t-test: $t = 14.771$, p-value < 0.001). This is graphically visualised in Figure 2. Across non-charge outcomes, 'Evidential difficulties: investigative' and 'Prosecution prevented' tend to have more missing values than 'Evidential difficulties: attributed to victim.'

Figure 2. Violin plots of missing values in cases with a 'charged' and 'not charged' suspect



Furthermore, we explore which missing values, for which variables, are more closely associated with outcomes of police investigations. This is analyzed using logistic regression models predicting the binary outcome of 'charged' (1) as opposed to 'not charged' (0), with fixed effects added to control for the effect associated with each police force. Results are presented in Table 6, with Model 1 exploring the effect of each missing variable and Model 2 exploring the effect of the total number of missing values on the likelihood of a charge outcome. Firstly, looking at the results of Model 2, we note that the number of missing variables is indeed a statistically significant predictor of charge outcomes. Secondly, we observe that while missing victim data does not significantly predict the chances of a 'charge' outcome, missing suspect data - and particularly age - is strongly associated with lower likelihood of a charge. Missing data regarding the time between the incident and the report is negatively associated with charge outcomes, potentially indicating cases in which the date of the offence is unclear. Interestingly, when the type of suspect-victim relationship is unknown, there are higher chances of a 'charge', potentially masking the fact that those cases in which the relationship is well defined (e.g., current/ex-partner) are less likely to be supported by the victim. Missing data relating to the time between the report and the assignment of an outcome is also positively associated with a charge, though we do not have a clear explanation for why this may be the case.

If cases in which no suspect was identified are excluded (see Table A2 of the Appendix), most observed relationships remain robust, but the effect of the total number of missing variables on case outcomes (Model 2) is substantially weaker. This suggests that *part of* the observed association may be due to the fact that many variables cannot be recorded when the suspect's identity is unknown.

Table 6. Fixed effects logistic regression models of missing data on charge outcomes. Fixed effects considered for police force areas. Standardized coefficients

	Model 1		Model 2	
	Beta	CI	Beta	CI
Victim characteristics				
Sex unknown	0.90	0.78 – 0.99		
Age unknown	0.93	0.84 – 1.02		
Ethnicity unknown	1.04	0.98 – 1.10		
Suspect characteristics				
Sex unknown	0.79***	0.71 – 0.87		
Age unknown	0.53***	0.48 – 0.60		
Ethnicity unknown	0.77***	0.72 – 0.83		
Offense characteristics				
Suspect-victim relationship unknown	1.30***	1.23 – 1.38		
Time incident-report unknown	0.77***	0.72 – 0.82		
Time report-outcome unknown	1.45***	1.40 – 1.50		
Total missing variables			0.59***	0.56 – 0.63
Observations	37,961		37,961	
Pseudo R^2 Tjur	0.048		0.012	
Pseudo R^2 McFadden	0.104		0.038	
Pseudo R^2 Nagelkerke	0.120		0.044	

***p<0.001; **p<0.01; *p<0.05

Conclusions

Missing data in police recorded rapes is common. The analysis of three years of police data on rape cases from five police forces in England and Wales shows that missing data in police records is not randomly distributed but varies by population (victim and suspect) and offence characteristics. Police-recorded rapes cannot be assumed to be Missing Completely At Random (MCAR) and should be treated as Missing At Random (MAR) (Nakagawa & Freckleton, 2008). Item-level missingness, particularly regarding key suspect characteristics such as ethnicity and age, appears more problematic than overall case profile-level missingness. Crucially, the amount of missing data, the pattern of missing data, and its form correlate with outcomes of rape cases.

Missing data is an obstacle to grounding our understanding of rape and, consequently, to deriving evidence-driven practices to enhance police investigations. In police-recorded rape cases, this missingness bias potentially misinforms our understanding of victims, suspects, and the circumstances surrounding sexual offences. It also hampers our ability to assess accurately whether police forces are delivering justice outcomes to all victims and suspects, irrespective of their ethnicity, sex, or other characteristics. Missing data introduces biases into analyses, making use of such data critically increases the risk of false positives and false negatives in research (Rubin, 1987). On a practical level, missing data hinders police officers' ability to address the

needs of different victim groups and design targeted preventative measures (Taylor & Gassner, 2010).

Our review of the literature and interviews with police analysts show that those working with police recorded data either ignore or take varied approaches to addressing the issue of missing values. For example, some police analysts create 'parallel' datasets by manually coding a sample of police case files, leading to additional workload, while others avoid missing values by either systematically excluding cases where crucial information is missing, neglecting to report on how the problem of missing data is treated in the analysis or failing to indicate the volume of missing data. Because there is no consistent way of addressing missing data, findings from different studies and analyses are not directly comparable to one another. At a minimum, researchers and practitioners making use of such data should assess the patterns of missingness in the data as well as its mechanisms (Hardy et al., 2009; Enders, 2010). Furthermore, advances in research methods enable accounting for and correcting for the presence of missing data in descriptive and inferential research (Pina-Sánchez et al., 2023).

The reluctance to engage with the issue of missing data is not a surprise, even for those who are directly involved in creating, maintaining and analysing the data on rape cases. Our qualitative analysis shows that the volume and pattern of missingness are affected by both technical and human factors. Lessons for policing are that missing data could be reduced by standardising data systems used by police forces, for example, by reducing the complexity of the required information, the number of non-mandatory fields for relevant factors, or the complexity of variable entry options. Standardising data recording practices within and between police forces should also contribute to using standard labels to classify missing data. This would enable a better understanding of the mechanisms affecting the missing data in each case and potentially reduce variation in the presence of missing data across variables. Forces should consider tackling underlying issues such as low prioritisation of data recording, poor data literacy, lack of training, low attention to detail, and insufficient supervision and scrutiny around data completion.

Turning to the data systems in which rape reports are recorded (however incomplete), the four different data systems used by the five police forces observed here were described by police staff as inflexible, time-consuming, difficult to navigate, and challenging in extracting data for analysis. These systems produce data with low reliability and large amounts of missing data in crucial fields, such as suspect and victim ethnicity, suspect age, and suspect-victim relationship. Some forces use systems with mandatory fields for certain variables, while others do not; which may help explain variation in missing data across forces and variables. Further research with police officers about the barriers to ethnicity recording would be beneficial to understand the role of systems, institutional cultures, lack of training and guidance, and other factors.

In terms of the type of missingness we observed, the fact that the concentration of missing data was substantially greater in relation to suspects than victims was striking in our dataset. As our quantitative analysis revealed, these are also variables that more directly condition rape case outcomes. Missing suspect data – and in particular the age of the suspect –

appears associated with a lower likelihood of a suspect being charged, even accounting for cases where no suspect has been identified. As already explained, this may be partly connected to the fact that police are more likely to have contact with victims than suspects, but also that some victims may not be forthcoming with information about suspects, meaning little is known to police about them. However, it could also indicate police not expending time capturing or entering data in cases they deem to have no prospect of proceeding, a lack of prioritisation of pursuing rape suspects or a failure of intelligence-led policing, where investment in capturing relevant information about all suspects is not seen as a worthwhile strategy for assisting current and any future investigations that may be linked. Ultimately, this could lead to police failing to identify repeat offenders.

A persistent area of missing population data concerns ethnicity. That this is the most frequently missing characteristic in both suspect and victim profiles suggests that there may be a specific problem relating to the gathering or recording of ethnicity data rather than simply being part of the broader overall tendency towards poorly completed data observed in relation to suspects. This also means we cannot fully rely on the accuracy of trends linked to cases where ethnicity is recorded because the large volume of cases with missing ethnicity data could harbour any number of different ethnic groups. This has far-reaching implications as ethnicity data linked to crime is subject to particular public scrutiny and commentary. The recent period in the UK has seen a re-emergence of the weaponisation and racialisation of sexual violence by far-right movements, particularly in narratives about sexual exploitation and 'grooming gangs' (Cockbain & Tufail, 2020). This has partly been enabled by historical failures to work through anxieties about police collecting and publishing data on the ethnicity of sexual offence suspects, leaving a perceived lack of transparency and a void where misinformation can circulate (see, for example, Gilroy 2002; Hall et al, 1982; Fatsis, 2021). In such a climate, there is little space, either for constructive public and academic debate about the necessities and potential dangers of how such data can be used, or for exploration of the complex factors that might contribute to explaining trends linked to ethnicity. Meanwhile, a continued lack of clear and accurate police data on the subject makes it very difficult to challenge inflammatory claims.

There also appear to be connections between missing data and victim sex and age. This could indicate greater police discomfort when talking to male victims, not recognising male or young victims as such due to assumptions about the context of the offence, or male/young victims feeling wary of providing the information because they do not feel at ease or believed (Javaid, 2018; Lonsway & Fitzgerald, 1994; Rumney, 2008), all of which could impact on police gathering key information.

Our analysis further suggests that there is more missing data in cases involving current/ex-partners compared to other relationship types. Some of these victims may choose not to identify their partner where they are ambivalent about the case proceeding or the suspect being criminalised. Although police recording of key data requires improvement, here we may also be seeing the impacts of wider issues within policing linked to recording practices, such as the growing proportion of cases not proceeding because victims are deemed not to support an

investigation or may not have originally chosen to make a report (see Lovett et al, 2024). These reporting contexts complicate findings and may be important in determining what information is available and known to the police, particularly in relation to certain types of suspect-victim relationships.

The higher level of missing data in cases recorded as being reported on the same day as the incident may reflect a recording artefact: when incident dates are unknown, officers may enter the report date to satisfy mandatory fields, which may also signal limited information on other aspects of the case. Alternatively, these cases may have been quickly closed with little effort spent on data entry. By contrast, cases investigated over longer periods may have involved more thorough enquiries, supportive victims, and identified suspects, resulting in more complete records.

It should be remembered that even if data systems and practices within policing are improved and this then leads to more reliable data containing fewer missing values, this will still only tell us about what is reported to and recorded by police (Kelly et al., 2005; Allen, 2007; Carretta et al., 2015). Therefore, further study should be devoted to deepening understanding of the population and offence-level characteristics associated with non-reporting through qualitative surveys with victim-survivors and/or support organisations. The extent to which sexual offences are being recorded as non-crime-related incidents rather than crimes, or are still being no-crimed, and whether this is related to particular item- or case profile-level characteristics should also be explored. Nevertheless, if, through better data quality enabling more robust analysis, we are able to better understand and make more visible who the victims and suspects of these offenses are, the nature of these offenses, and how they are concluded, then that is still a big step forward.

References

Allen, W.D. (2007). The reporting and underreporting of rape. *Southern Economic Journal*, 73(3), 623-641.

Braun, V., Clarke, V. (2021). *Thematic Analysis: A Practical Guide*. London: Sage

Carretta, C.M., Burgess, A.W., DeMarco, R. (2015). To Tell or Not to Tell. *Violence Against Women*, 21(9), 1145-1165.

Cockbain, E., Tufail, W. (2020) Failing victims, fuelling hate: challenging the harms of the 'Muslim grooming gangs' narrative. *Race & Class*, 1(3) 3–32.

Data Analytics Lab (2020). *Exploratory Analysis of Sexual Convictions*. Data Analytics Lab. Retrieved from https://www.westmidlands-pcc.gov.uk/wp-content/uploads/2020/02/rasso_findings_202001_s_no_mark.pdf?x57724.

Enders, C.K. (2010). *Applied Missing Data Analysis. Applied Missing Data Analysis.* New York, NY: Guilford Press.

Fatsis, L. (2021). Policing the Union's Black: The Racial Politics of Law and Order in Contemporary Britain. In: Gordon, F., Newman, D. (Eds.), *Leading Works in Law and Social Justice*, 137-150. Abingdon, UK: Routledge.

Gilroy, P. (2002). *There Ain't No Black in the Union Jack*. London: Routledge.

Graham, J.W. (2012). *Missing Data: Analysis and Design*. New York, NY: Springer.

Hall, M. (2022). Counting crime: Discounting victims? *International Review of Victimology*, 28(1), 3-32.

Hall, S., Critcher, C., Jefferson, T. et al. (1982). *Policing the Crisis: Mugging, the State and Law and Order*. London: Macmillan.

Hardy, S.E., Allore, H., Studenski, S.A. (2009). Missing Data: A Special Challenge in Aging Research. *Journal of the American Geriatrics Society*, 57(4), 722–729.

Harper, S.B., Davis, A., Shepp, V., O'Callaghan, E., Maskaly, J. (2023). What's Missing Matters: Examining Missing Data Problems in Sexual Assault Kit Data. *Deviant Behavior*, 45(1), 1-24.

Hester, M., Lilley, S.J. (2017). Rape Investigation and Attrition in Acquaintance, Domestic Violence and Historical Rape Cases. *Journal of Investigative Psychology and Offender Profiling*, 14(2), 175–88.

HM Inspectorate of Constabulary (2014). *Crime-recording: Making the victim count*. Retrieved from: https://assets-hmicfrs.justiceinspectorates.gov.uk/uploads/crime-recording-making-the-victim-count.pdf .

HM Inspectorate of Constabulary and Fire & Rescue Services (2023). *Police performance: Getting a grip*. Retrieved from: https://assets-hmicfrs.justiceinspectorates.gov.uk/uploads/peel-spotlight-report-police-performance.pdf.

Hohl, K., Stanko, E.A. (2015). Complaints of Rape and the Criminal Justice System: Fresh Evidence on the Attrition Problem in England and Wales. *European Journal of Criminology*, 12(3), 324–341.

Hohl, K., Stanko, E.A. (2024). *Policing rape. The way forward*. Routledge.

Home Office (2023). *Crime Outcomes in England and Wales: Technical Annex.* Home Office.

Javaid, A. (2018). The unheard victims: gender, policing and sexual violence. *Policing and Society*, 30(4), 412–428.

Jordan, J. (2004). *The word of a woman: Police, rape and belief*. Palgrave.

Kelly, L., Lovett, J., Regan, L. (2005). *A gap or a chasm? Attrition in reported rape cases.* Home Office Research Study 293. London: Home Office.

Koss, M.P. (1992). The under detection of rape: Methodological choices influence incidence estimates. *Journal of Social Issues*, 48(1), 61-75.

Little, P.J.A., Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley.

Lonsway, K.A., Fitzgerald, L.F. (1994). Rape Myths: In Review. *Psychology of Women Quarterly*, 18(2), 133-164.

Lovett, J., Hales, G., Kelly, L., Khan, A., Hardiman, M., Trott, L. (2022). What Can We Learn from Police Data About Timeliness in Rape and Serious Sexual Offence Investigations in England and Wales? *International Criminology*, 2(3), 286–298.

Lovett, J., Vera-Gray, F., Kelly, L. (2024) The unintended consequences of improving police recording of rape in England and Wales. *Policing: A Journal of Policy and Practice*, 18, paae086, https://doi.org/10.1093/police/paae086.

MOPAC (2019). *The London Rape Review: A Review of Cases from 2016.* MOPAC.

Murphy, A., Hine, B., Yesberg, J.A., Wunsch, D., Charleton, B. (2022). Lessons from London: A Contemporary Examination of the Factors Affecting Attrition among Rape Complaints. *Psychology, Crime & Law*, 28(1), 82–114.

Nakagawa, S., Freckleton, R.P. (2008). Missing Inaction: The Dangers of Ignoring Missing Data. *Trends in Ecology & Evolution*, 23(11), 592–596.

Newman, D.A. (2014). Missing Data: Five Practical Guidelines. *Organizational Research Methods*, 17(4), 372–411.

Osgood, D.W. (2017). Poisson-based regression analysis of aggregate crime rates. In S. Bushway, D. Weisburd (Eds.), *Quantitative methods in criminology* (pp. 577-599). Routledge.

Pina-Sánchez, J., Brunton-Smith, I., Buil-Gil, D., Cernat, A. (2023). Exploring the impact of measurement error in police recorded crime rates through sensitivity analysis. *Crime Science*, 12, 14.

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley.

Rumney, P.N.S. (2008). Policing male rape and sexual assault. *Journal of Criminal Law*, 72(1), 67–86.

Stanko, B. (2022). *Operation Soteria Bluestone Year One Report*. Home Office. Retrieved from: https://www.gov.uk/government/publications/operation-soteria-year-one-report/operation-soteria-bluestone-year-one-report-accessible-version#appendix-11-pillar-five---data-and-performance.

Taylor, S.C., Gassner, L. (2010). Stemming the flow: challenges for policing adult sexual assault with regard to attrition rates and under-reporting of sexual offences. *Police Practice and Research*, 11(3), 240–255.

Walfield, S.M. (2016). When a Cleared Rape Is Not Cleared: A Multilevel Study of Arrest and Exceptional Clearance. *Journal of Interpersonal Violence*, 31(9), 1767-1792.

Walker, S.J.L., Hester, M., McPhee, D., Patsios, D., Williams, A., Bates, L., Rumney, P. (2021). Rape, Inequality and the Criminal Justice Response in England: The Importance of Age and Gender. *Criminology & Criminal Justice*, 21(3), 297–315.

Yung, C.R. (2013). How to lie with rape statistics: America's hidden rape crisis. *Iowa Law Review*, 99(3), 1197-1256.

Appendix

Table A1. Detailed description of outcomes and groupings

Grouping	Outcome	Description
Charged	1. Charge/Summons	
	1A. Charge and/or Summons – alternative offence	A person has been charged or summonsed for the crime, but […] the charge/summons relates to an alternative offence to that recorded (irrespective of any subsequent acquittal at court).
	2A. Youth Caution – alternative offence	A youth offender has been cautioned by the police but following the application of the CPS charging standards and the provisions of the HOCR, the caution relates to an alternative offence to that recorded.
	3A. Adult Caution – alternative offence	An adult offender has been cautioned by the police but following the application of the CPS charging standards and the provisions of the HOCR, the caution relates to an alternative offence to that recorded.
Evidential difficulties: attributed to victim	14. Evidential difficulties: suspect not identified; victim does not support further action (from April 2014)	Evidential difficulties victim based – named suspect not identified. The crime is confirmed but the victim declines or is unable to support further police action to identify the offender.
	16. Evidential difficulties: suspect identified; victim does not support further action (from April 2014)	Evidential difficulties victim based – named suspect identified. The victim does not support (or has withdrawn support from) police action.
Evidential difficulties: investigative	15. Evidential difficulties (suspect identified; victim supports action) (from April 2014)	Evidential difficulties named suspect identified – the crime is confirmed and the victim supports police action, but evidential difficulties prevent further action. This includes cases where the suspect has been identified, the victim supports action, the suspect has been circulated as wanted but cannot be traced and the crime is finalized pending further action.
	18. Investigation complete –no suspect identified (from April 2014)	The crime has been investigated as far as reasonably possible – case closed pending further investigative opportunities becoming available.
Prosecution prevented and not in public interest	5. The offender has died	
	9. Prosecution not in public interest (CPS) (all offences)	
	10. Formal action against the offender is not in the public interest (police decision)	
	11. Prosecution prevented - named suspect identified but is below the age of criminal responsibility	
	12. Prosecution prevented - named identified suspect identified but is too ill (physical or mental health) to prosecute	

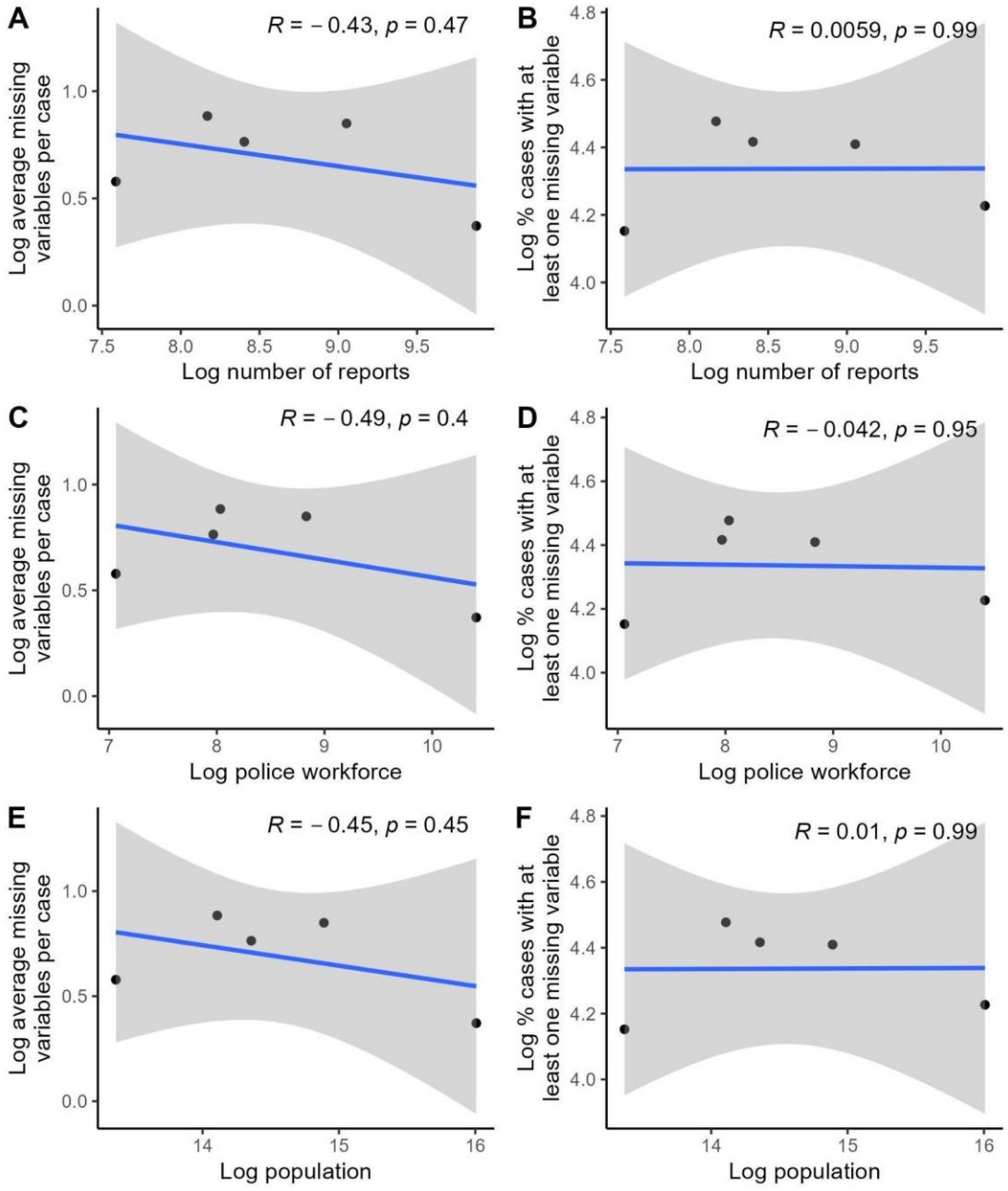
	13. Prosecution prevented - named suspect identified but victim or key witness is dead or too ill to give evidence 17. Prosecution time limit expired - suspect identified but the time limit for prosecution has expired 21. Further action, resulting from the crime report, which could provide evidence sufficient to support formal action being taken against the suspect is not in the public interest - police decision

Table A2. Fixed effects logistic regression models of missing data on charge outcomes (excluding cases with no suspect identified). Fixed effects considered for police force areas. Standardized coefficients

	Model 1		Model 2	
	Beta	CI	Beta	CI
Victim characteristics				
Sex unknown	0.90	0.78 – 0.98		
Age unknown	0.96	0.88 – 1.04		
Ethnicity unknown	1.03	0.96 – 1.09		
Suspect characteristics				
Sex unknown	0.85***	0.79 – 0.92		
Age unknown	0.86***	0.79 – 0.93		
Ethnicity unknown	0.86***	0.80 – 0.92		
Offense characteristics				
Suspect-victim relationship unknown	1.32***	1.25 – 1.40		
Time incident-report unknown	0.78***	0.73 – 0.84		
Time report-outcome unknown	1.48***	1.43 – 1.54		
Total missing variables			0.93*	0.88 – 0.99
Observations	27,030		27,030	
Pseudo R^2 Tjur	0.038		0.005	
Pseudo R^2 McFadden	0.060		0.012	
Pseudo R^2 Nagelkerke	0.073		0.014	

***p<0.001; **p<0.01; *p<0.05

Figure A1. Relationships between police force size, caseload and population and the extent of missing data



Author Bio Statement

Gordana Uzelac is a Reader in Sociology. She is a methodologist with a research focus on theories of nations and nationalism and topics in the area of sexual violence in collaboration with the Child and Women Abuse Studies Unit (CWASU), London Metropolitan University. UK.
David Buil-Gil is a Senior Lecturer in Quantitative Criminology and Research Director in the Department of Criminology at the University of Manchester. He is also Director of CrimRxiv, the global open access hub for criminology, and Associate Editor of Evidence Base: Criminal Justice Research, Policy & Action. His research interests include communities and crime, comparative criminology, adolescent offending, and quantitative research methods. He has published on measurement error in crime data, model-based estimation of crime in small areas, spare time and crime, cybercrime, and international comparisons of crime data quality.
Katrin Hohl is Professor of Criminology and Criminal Justice at the Department of Sociology and Criminology, City, St. George's, University of London, London, UK. Her research centers on police and wider criminal justice responses to sexual violence and domestic abuse.
Jo Lovett is a Senior Research Fellow at the Child and Woman Abuse Studies Unit, London Metropolitan University, UK. Her research interests are in sexual and domestic violence, including victim-survivor experiences, and policy, specialist service and criminal justice responses. She has published on topics such as sexual assault referral centres, attrition, alcohol and sexual violence, and discourses about child sexual abuse.