The nature and extent of the literature on linked reproductive health datasets in the UK: a scoping review

HALL, Jennifer <http://orcid.org/0000-0002-2084-9568>, HARVEY-PESCOTT, Lois <http://orcid.org/0009-0003-6361-0164>, KO, Sum Yue Jessica, STEVENS, Rose, PATHAK, Neha, ALI, Ifra, BARRETT, Geraldine <http://orcid.org/0000-0002-9738-1051>, SHAND, Jenny and DICKSON, Kelly

International Journal of Population Data Science

Journal Website: www.ijpds.org

The nature and extent of the literature on linked reproductive health datasets in the UK: a scoping review

Jennifer Hall[1,*], Lois Harvey-Pescott[1], Sum Yue (Jessica) Ko[1,2], Rose Stevens[1], Neha Pathak[1,3], Ifra Ali[1], Geraldine Barrett[1], Jenny Shand[1,4], and Kelly Dickson[1,2]

Submission History

Submitted:	27/03/2025
Accepted:	26/06/2025
Published:	08/09/2025

[1] NIHR Policy Research Unit in Reproductive Health (PRU-RH), EGA Institute for Women's Health, UCL, London, United Kingdom
[2] EPPI-Centre, Social Research Unit, Social Research Institute, Faculty of Education, UCL, London, United Kingdom
[3] Community Sexual and Reproductive Health, Guy's and St Thomas's NHS Trust; Institute for Global Health, UCL, London, United Kingdom
[4] Department of Clinical, Edu & Hlth Psychology, Faculty of Brain Sciences, UCL, London, United Kingdom

Abstract

Introduction
Data linkage methodologies are increasingly being utilised across research, but there is currently no evidence on the extent and nature of studies that have used linked reproductive health data. The objective of this scoping review is to identify UK studies that use reproductive health data linkage, to improve our understanding of how data linkage could be used for policy, practice, and research in reproductive health.

Methods
We conducted a scoping review using a systematic search in five databases: MEDLINE, EMBASE, CINAHL, MIDIRS, and PSYCINFO to identify literature published in English between January 2000 – April 2024. Following duplication removal, piloting, and screening of titles/abstracts, screening of full texts was conducted. Publications using reproductive health data linkage among UK participants of reproductive age were included. Data was extracted from included articles to capture details relating to study characteristics and what, how, and why data was linked.

Findings
Of the 7,291 identified studies, 272 studies were included in the review. Most studies using data linkage answered questions around reproductive cancer and maternal and child health, whilst only a few studies focused on abortion, contraception, menopause, and preconception health. Several nationally agreed reproductive health indicators did not appear in any included study. Information on sample sociodemographic characteristics, such as ethnicity and deprivation, was often unreported, limiting the identification of health inequalities. Many different datasets were linked (n = 155) with routine health data sources, such as hospital episode statistics (HES), being the most frequently linked.

Interpretation
There is a growing body of research using linked UK reproductive health data, with gaps in which reproductive health domains are covered and which sample characteristics are reported. Further efforts to create a comprehensive, linked reproductive health data resource with robust linkage methods would enable us to fill data gaps, examine inequalities, and explore reproductive health trajectories.

Funding
National Institute for Health and Care Research (NIHR) Policy Research Unit in Reproductive Health.

Keywords
data linkage; medical record linkage; reproductive health; sexual health; United Kingdom

*Corresponding Author:
Email Address: Jennifer.hall@ucl.ac.uk (Jennifer Hall)

https://doi.org/10.23889/ijpds.v101.2989

Introduction

There is increasing availability of large health-related datasets due to mass digitisation and the exponential growth of born-digital archives over the past two decades [1]. However, most datasets only contain limited information, for example are from a specific setting (primary/secondary care) or are on a specific topic (e.g. cancer). Data linkage involves combining data from different sources related to the same individual to create a new, enhanced data resource [2]. Data linkage methodologies are increasingly being utilised in healthcare research [3–5].

Reproductive health is a broad concept that spans physical, mental, and social wellbeing in all matters relating to the reproductive system [6]. Linking reproductive health data enables us to answer questions about reproductive events across the life course, across generations, and across different levels of healthcare provision. Linked reproductive health data is already used globally to answer questions about topics including maternal and child health [7–9] and predicting reproductive cancer risks using health insurance data linked with surveillance or registry data [10].

In the UK, reproductive health data are held in different databases across multiple organisations meaning linkage is required to explore many research, policy, and surveillance questions. In 2021, Public Health England recommended 23 indicators for measuring population reproductive health, yet as of 2024, over half remain unavailable in UK public health surveillance profiles (i.e. are not captured in routine data) (Table 1) [11]. There has been no systematic approach to assessing the extent and nature of studies using linked reproductive health data in the UK, and no mapping of which reproductive health domains are prioritised for linkage, which populations are missing from linked data resources, or on the robustness of linkage methodologies used. Without this we are unable to understand inequalities in either the topics or the populations studied.

The aim of this scoping review is to map the landscape of UK reproductive health data linkage studies to improve our understanding of how data linkage could be used for policy, practice, and research in reproductive health. This can be used to inform the development of a linked data resource that considers the use and linkage of routine health data as well as large epidemiological datasets. Such linked reproductive health data would enable us to fill data gaps, examine inequalities, explore reproductive health trajectories, and understand care experiences and patient journeys that span providers in primary, secondary, tertiary and independent care.

Methods

The scoping review was conducted using JBI methodology [12], the published protocol (https://osf.io/8s9c4) followed Lely *et al.'s* guidance [13], and we used the Preferred Reporting Items for Systematic Reviews and Meta-analyses extension for scoping reviews (Supplementary S1) in drafting the manuscript [14].

Search strategy

The search strategy was developed by the review team with a librarian. The databases searched were MEDLINE, EMBASE, CINAHL, MIDIRS, and PSCYINFO. Search terms were divided into three concepts: reproductive health, data linkage, and UK context. We took an inclusive approach to defining the scope of reproductive health, which at times blurs with maternal health, urology, andrology and sexual health. In keeping with this approach, we adapted the Guttmacher-Lancet Commission framework [15] by adding four additional domains[1] (see Figure 1). Publications were included if they reported health outcomes related to any of these domains.

Key terms within each area of reproductive health were identified through an initial limited search using Google Scholar and MEDLINE, inspecting the words contained in the titles and abstracts of relevant articles, and index terms. The search strategy for this concept, including identified text and index terms agreed by the team, was adapted for each of the five databases searched (see Supplementary S2). The search strategy for the data linkage concept was adapted for each database from a preexisting publication investigating data linkage in multimorbidity research [3]. For the UK concept, preexisting and validated filters for the retrieval of UK-based studies were used in MEDLINE and EMBASE [16–18]. An adapted, non-validated version of the MEDLINE UK filter was adapted for the search in PSYCINFO [19]. It was not possible to adapt the UK filters for the searches in CINAHL or MIDIRS, so any non-UK studies identified from these databases were removed during the screening phase. The search was limited by date, gathering studies published from 1st January 2000 to the date of the search, as few UK databases were digitised before 2000 [1]. The search strategy aimed to include peer-reviewed studies written in English. The databases were searched on the following dates: MEDLINE – 5 April 2024; EMBASE - 22 March 2024; PSYCINFO – 22 March 2024; CINAHL – 22 March 2024; and MIDIRS - 22 March 2024. All identified citations were collated and uploaded into EPPI Reviewer, where duplicates were removed.

Inclusion criteria

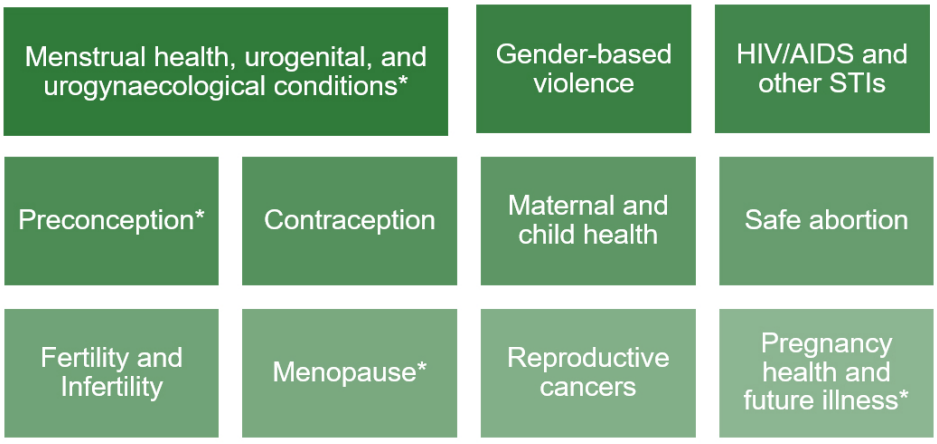
The eligibility criteria were developed collaboratively as a group through structured discussions to reach shared consensus and understanding. This was further reinforced through the pilot screening process, which familiarised the team with the criteria and provided opportunities to make any necessary refinements. The inclusion criteria specified that studies should contain primary data on a reproductive health topic and contain data linkage, be a UK study in English published from the year 2000 onwards, and report

[1]The four domains added in addition to the domains identified in the Guttmacher-Lancet Commission report [15] were: Menstrual health, urogenital, and urogynecological conditions; Preconception; Menopause; Pregnancy health and future illness. This was to ensure a broad and inclusive definition of reproductive health, fitting with the WHO definition of reproductive health: 'Reproductive health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity, in all matters relating to the reproductive system and to its functions and processes. Reproductive health implies that people are able to have a satisfying and safe sex life and that they have the capability to reproduce and the freedom to decide if, when and how often to do so' [6].

Table 1: Public Health England indicators for measuring population reproductive health

Public Health England Indicators for Measuring Population Reproductive Health (2021)	
Domain	**Indicator**
Psychosexual wellbeing	1. Perinatal mental health
Absence of violence	2. Domestic abuse-related incidents and crimes 3. Violent crimes – sexual offenses
Menstrual health	4. Women of reproductive age presenting with and having intrauterine system insertion primarily for heavy menstrual bleeding 5. Proportion of women who have self-reported an impact of reproductive health issues
Menopause health	6. Prevalence of women over 44 years old referred for menopause assessment
Contraception	7. Contraception prevalence rate 8. Women receiving emergency contraception 9. Average wait time to first available long-acting reversible contraception (LARC) fitting appointment
Unplanned pregnancy	10. Removal of babies at birth into care 11. Under age conception (i. under 16; ii. under 18) 12. London Measure of Unplanned Pregnancy (LMUP)
Abortion	13. Abortions under 10 weeks that are medical 14. Total abortion rate per 1,000 15. Under 18 conceptions leading to abortions
Preconception care	16. Low birth weight of term babies 17. Neonatal mortality and stillbirth rate
Infertility and fertility service	18. Prevalence of infertility in women 19. Live birth from assisted conception
Prevention of reproductive ill health	20. Females attending cervical screening within target period 21. Human papillomavirus (HPV) vaccination coverage for 2 doses (females 13 to 14 years old) 22. Total overall (LA and NHS) spend on sexual health provision per capita 23. Number of people who have received high quality sex and relationships education in their lifetime

Figure 1: Reproductive health domains used to define the scope of the review, based on the Guttmacher-Lancet Commission Report [15]



on reproductive health outcomes for people of reproductive age, defined by the study team as 12 years and up. An upper age limit was not specified to ensure reproductive health data was captured across the life-course, including data on menopause and reproductive cancers. Publications that reported reproductive health outcomes from participants of reproductive age and their offspring were included to capture any intergenerational effects. Publications that did not report

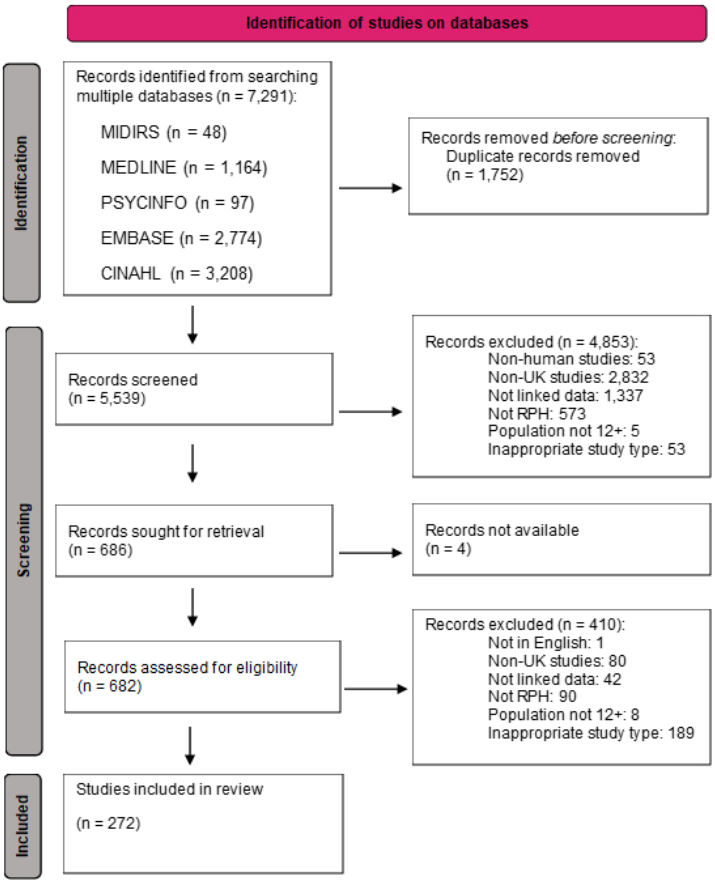
health outcomes from participants of reproductive age and only reported findings from children were excluded (e.g. studies examining educational outcomes of the children conceived via assisted reproductive technologies without other information on parental or child health).

Study selection

The above inclusion criteria were used for the screening stages of the scoping review, each stage of which was conducted in EPPI-Reviewer. One round of pilot title and abstract screening was conducted. For this, the review team independently screened the titles and abstracts of 50 records. A hierarchical pilot screening tool was used, which incorporated the eligibility criteria and facilitated the exclusion of articles based on the following criteria (0. Non-human studies 1. English Language, 2. UK Context, 3. Data linkage methodology, 4. Reproductive health outcomes, 5. Population, 6. Study type - contains analysis of empirical data). To pass the pilot stage, a cut-off percentage agreement of 90% was required. After the initial pilot screening, all titles and abstracts were single-screened against the eligibility criteria. Records that did not contain an abstract but mentioned a reproductive health concept in the title were automatically included for full-text screening.

After two initial rounds of piloting for full text screening to ensure a shared understanding, the full texts were single-screened against the same eligibility criteria, with any discrepancies settled through discussions. Throughout screening regular discussions helped to maintain consistency, address queries and resolve differences. Reasons for exclusion of records at full-text are reported in Figure 2.

Figure 2: PRISMA flow diagram



Adapted from Page et al 2020 [20].

Data extraction

Data was extracted using a bespoke coding tool in EPPI Reviewer software (see Supplementary S2). The data items extracted include study descriptives (e.g., aim, population sample, geography) as well as characteristics relating to reproductive health (e.g., reproductive health domain) and data linkage methods (e.g., what datasets have been linked, why said data have been linked, and how the data have been linked). Inductive and deductive approaches were used in developing the tool. At the end of the coding process, all codes were checked by a second reviewer to ensure consistency. Results from the data coding process were then descriptively summarised and presented through a narrative summary with accompanying figures or tables.

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Of the 7,291 studies identified from the search, 1,752 duplicates were removed. 5,539 studies were screened on title and abstract; 4,853 were excluded, most due to their context (i.e., non-UK studies) (n = 2,832) or because they were not about data linkage (n = 1,337). Four studies could not be retrieved. Of the 682 full-text studies screened, 410 were excluded. The most common reason was study design (n = 189) i.e. studies containing no empirical data, such as opinion papers and study protocols. Other common reasons for full text exclusion were geographical context (n = 80) and the lack of focus on reproductive health (n = 90). A total of 272 studies were included in the review (full list in Supplementary S2).

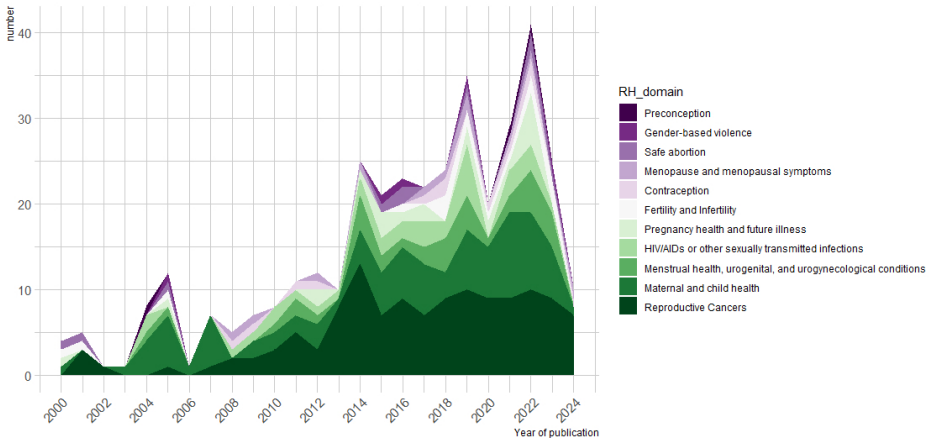
Trends over time and space

The majority of studies contained data based in England (n = 155); Scotland also has a notable presence (n = 75); Wales and Northern Ireland were less well represented (n = 26 and n = 11, respectively). Forty studies either covered the whole of the UK or did not specify which countries in the UK were included. National-level data dominated (n = 205); some studies combined both national and local level data (n = 43) and relatively few contained only local level data (n = 24). The number of published studies using linked reproductive health data showed a marked increase in 2014 (n = 20), doubling from the number of studies identified in 2013 (n = 10). As the searches were conducted in early 2024, only 8 studies were recorded for that year.

Reproductive health domains and PHE indicators

The identified studies were unequally distributed across reproductive health domains (Figure 3). The most frequent

Figure 3: Trends in the number of UK studies published across each reproductive health domain using linked reproductive health data each year since 2000. If studies covered more than one reproductive health domain, they were coded under each domain



domain included in these linked reproductive health data analyses was reproductive cancers (n = 129). Seventy-eight studies focus solely on reproductive cancers, the remainder included other cancers, such as bowel or lung. Maternal and child health was the second most common domain (n = 92). Other less common but relatively prominent reproductive health domains studied include "menstrual health" (n = 35) and "HIV/AIDS or other sexually transmitted infections" (n = 33). In contrast, there were minimal linked studies on "preconception" (n = 3), "gender-based violence" (n = 5), "safe abortion" (n = 10), "menopause and menopausal symptoms" (n = 11), and "contraception" (n = 12).

In terms of the 23 PHE population reproductive health indicators, studies including perinatal indicators, such as "Neonatal mortality and stillbirth rate" (n = 44) and "Low birth weight" (n = 30) were most common, and studies including indicators related to contraception and fertility were less common, with around 13-15 studies each. More than a quarter of the indicators were not addressed in any of the included studies, including sexual offenses, menopause assessment, wait-times for long-acting reversible contraception appointments, and high-quality sex and relationships education.

Populations sampled

The skew of reproductive health domains studied is reflected in the study populations. Studies had samples mostly consisting of individuals with reproductive cancer (n = 117), pregnant women (n = 83), and neonates (n = 73). Other specific groups, like mothers and hormone users, have n = 46 and n = 35, respectively. Populations including individuals with reproductive health conditions such as endometriosis (n = 6) or PCOS (n = 7) were rare, as were post-menopausal women (n = 9) and women with heavy menstrual bleeding (n = 9). Over half of the studies included both male and female participants (n = 147) but there were nearly four times more studies with only female participants (n = 99) than studies with only male participants (n = 26). The domains studied also impacted the stages of the life course most included in the literature. For instance, among studies that reported either the mean or median age of their participants, studies focusing on maternal and child health had average ages of around 20-30, but those focusing on reproductive cancers had average ages around 40-50. Studies mostly focused on linkage across health at one time point rather than across an individual's life course as they age. However, there were 22 studies examining relationships between health in pregnancy and health in later life, with 20 of these studies published since 2012 suggesting a trend of increasing research taking a life course approach.

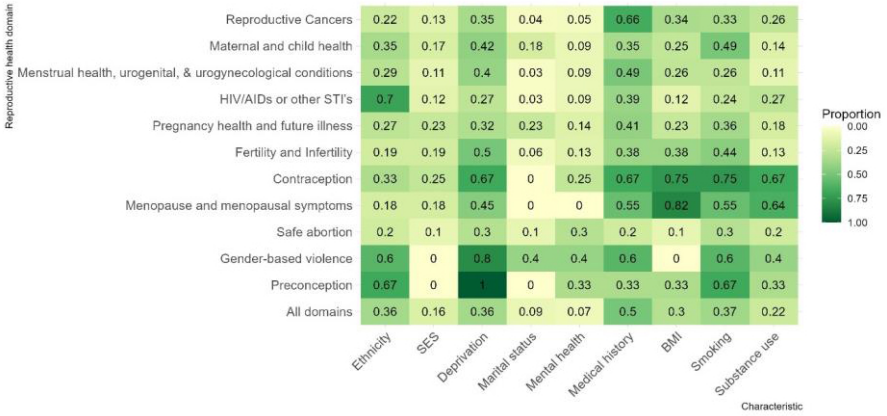
There was a wide range in other characteristics reported beyond the outcomes of interest in the study, limiting analysis of inequalities in the populations included in linked reproductive health data studies. For instance, while around half the studies included information on medical treatment history (n = 137), fewer contained information on participant ethnicity (n = 83, 31%) or deprivation (n = 98, 36%), and very few contained information on domestic abuse (n = 3), LGBTQIA+ identity (n = 6), or refugee status (n = 1).

There were trends in which characteristics were most reported across reproductive health domains (Figure 4). For instance, studies of reproductive cancers and maternal and child health commonly collected data on medical/treatment history and smoking status. Ethnicity was reported by fewer than half of all studies in most domains, but in 60% or more of studies on gender-based violence, preconception care, and HIV/AIDs or other sexually transmitted infections. Studies focusing on contraception were most likely to contain information on mental health, though still only a quarter of studies did so, and these studies were also more likely to report on deprivation and socioeconomic status.

Datasets linked

The datasets linked differed in terms of which setting of care they pertained to, ranging from demographic data (e.g., Office for National Statistics (ONS)) to hospital data (e.g., Hospital Episode Statistics (HES), Scottish Hospital In-Patient Statistics, Patient Episode Database for Wales) to primary care data (e.g., Clinical Practice Research Datalink (CPRD), QResearch, Royal College of General Practitioners data). The datasets linked also stemmed from a variety of study designs, including cohort study data (e.g., Born in Bradford, Million Women Study, UK Millennium Cohort

Figure 4: Percentage of studies reporting on additional sociodemographic and health characteristics across reproductive health domains



Study) and clinical trials (e.g., Pelvic Organ Prolapse PhysiotherapY Trial).

Within this evidence base, the most widely linked dataset was the HES (n = 104), followed by CPRD (n = 79), ONS data (n = 67), national and local cancer registries (n = 62) and the Scottish Morbidity Record (n = 42) (see Supplementary S3 for a full list of datasets included). It is noteworthy that the reporting of datasets linked could sometimes be unclear, with studies either not specifying the dataset (n = 4) or citing the linked dataset in generic terms such as "hospital data" (n = 18) or "hospital obstetric records" (n = 4)/ "obstetric records" (n = 1).

Whilst most studies linked different types of health data together, around a third of studies (n = 91) linked health data with wider social and environmental information. The vast majority of these linked to sociodemographic data from the census held by the Office for National Statistics (ONS) (n = 76). Four studies linked to education data e.g. the National Pupil Database (n = 2), one linked to police data (n = 1), and one to Meteorological Office data (n = 1).

Reason for linkage

More than 90% of the studies (n = 247) sought to utilise the linked datasets to analyse the relationships between two or more variables which were previously unlinked. For instance, studies linked data across different levels of healthcare (e.g., primary care and secondary care) (n = 89), data between individuals (e.g., mother and baby) (n = 63), and data across different research methodologies (e.g., cohort study data with hospital data) (n = 32). The latter, linking between different research methodologies (e.g. primary research data to routine health data), has shown a small increase compared to the first decade since 2000 but remains at low levels year on year (a mean of 1 study published per year from 2000-2009, raising to a mean of 1.5 studies published per year in 2010-2019.)

Other studies endeavoured to explore or understand the process of data linkage (n = 12), the level of agreement between data in different datasets (n = 13), the feasibility of using data linkage in measuring outcomes (n = 11) or assess public acceptability of the linked datasets (n = 1). The small number of studies seeking to understand the acceptability of linking reproductive health data is additionally reflected in the low proportion (<10%) of studies that report involving patients and/or the public in their research on linked datasets (n = 20).

Method and completeness of data linkage

Nearly half of the studies (n = 132) did not report the information used to link data between datasets. Among those that did, most used some form of unique identifier such as NHS number (England/Wales) or Community Health Index (CHI) number (Scotland). Other types of information used for linkage include date of birth (n = 54), postcode (n = 36), sex/gender (n = 29), and name (n = 20). Similarly, most studies (n = 144) did not report the methodology adopted for linkage. Of those that did, 102 used deterministic linkage and 34 used probabilistic linkage.[2] More than half of the studies did not provide information on the completeness of the data linkage (n = 187).

Discussion

We identified 272 studies published since 2000 which linked UK reproductive health data. Studies were unequally distributed across reproductive health domains and corresponding populations studied. Reproductive cancers and maternal and child health dominated, with fewer studies concerning preconception care, abortion, menopause, and contraception. Our review included studies that linked various datasets, the majority being routine health data and demographic datasets. Most studies aimed to use linked data to analyse relationships between previously un-linkable outcomes or explore relationships between variables across different individuals, time points in the life course, or healthcare levels. Few studies specifically set out to explore the process, public acceptability, or feasibility of data linkage [21].

[2]For the purposes of this review, deterministic linkage applies to the linkage of datasets using uniquely shared information. As such, records are linked only if the linkage fields agree, and any mismatch prevents a linkage [34]. Probabilistic linkage, on the other hand, attempts to establish linkage using multiple, possibly non-unique, pieces of information [35].

The studies identified predominantly link routinely collected data, reflecting the general focus of linkage in UK health data to date [21]. Thus, the topics represented in our review reflect the existence and prioritisation of high-quality routine data resources in these domains, such as comprehensive UK cancer registries. Data resources on contraceptive use, preconception care, and menopause experiences are not routinely available. Therefore, these domains are less represented in our review, potentially mirroring the lower focus on these topics in reproductive health research and policy to date. The distribution of topics covered in linkage studies may also reflect relative levels of funding within health research. While there are no data available on the allocation of research funds across reproductive health domains, cancer, which was the most common domain in our study, receives significantly more funding than reproductive health as a whole [22].

Several studies included linked to primary research data, including clinical trials and cohort studies, which may cover a greater range of topics than routinely collected data. However, further innovation is needed to understand and exploit the many potential data sources, both public and private, related to reproductive health and how they can be best utilised. For example, mHealth cycle tracking applications and wearable devices could represent significant sources of reproductive health data [23]. These data however do come with important considerations regarding data privacy and security, as a recent review found that, despite existing regulations, many popular women's mHealth applications had poor data privacy and security standards [24].

A benefit of data linkage is that it can increase our ability to study inequalities. For instance, routine health data records can be linked to datasets with environmental and social information to study both determinants of rare health outcomes and the experiences of hard-to-reach groups [2, 21]. However, our review found that more studies using linked data reported on additional medical/health-related sample characteristics, such as medical or treatment history and smoking status, rather than sociodemographic characteristics, such as ethnicity or deprivation. This limits our ability to identify population inequalities and to design services to meet the needs of vulnerable populations. Additionally, successful identification of inequalities requires reliably collected data. For instance, only six studies reported on LGBTQIA+ identity of their participants. This may be reflective of clinical information systems in the UK which inconsistently collect gender and sex data, and any patient who changes their gender marker must change their NHS number [25]. This risks the loss of vital health records, reducing the chance of accurately linking these individuals' experiences across the life course, and rendering invisible the health outcomes of this group.

Our review found that the quality and detail of reporting on methodologies used for linkage is highly variable across studies. This is consistent with a recent evaluation of the quality of reporting on linkage methodology in multimorbidity research [3] that linkage processes are generally poorly reported. Errors, changes over time, and missing data often hamper accurate linkage attempts [2]. In addition, the included studies did not investigate whether there were patterns in rates of consent to linkage of reproductive health data, as there are for the NHS national data opt-out [26], and thus we are unable to understand which populations may be systematically missed from reproductive health research. Therefore, increased transparency within the linkage process, reporting on consent to linkage, and better adherence to reporting guidelines [27–29] can improve our ability to estimate risk of introduced bias by linkage and any potentially inaccurate inferences drawn from results.

As well as low levels of reporting on how linkage was achieved, we found low engagement with the question of whether linkage is ethical or desirable among members of the public. Harron (2022) posits that the biggest barrier to realising the full potential of data linkage is gaining and maintaining public trust [2]. Fewer than 10% of studies mention undertaking any public involvement in their work and only one study was identified which explicitly set out to explore the acceptability of creating linked data. Whilst linkage of health data for research for public benefit, is generally acceptable to the wider public, as long as appropriate protections are in place [30], there has not been engagement specifically focused on understanding whether this is generalisable to reproductive health topics [27]. As more data are created, concern over how data are used and how identifiable an individual is within it has grown. Recent controversies, such as the feared use of UK Biobank data by race scientists [31] and the potential selling off of 15 million customers' genetic data as 23andMe faces bankruptcy [32], have damaged public trust and shown how data resources can cause harm. As data availability continues to increase, efforts to utilise data linkage to answer novel questions should first ensure that there is a strong ethical mandate driven by public consultation to do so [33].

Strengths and limitations

To our knowledge, this is the first scoping review of primary studies using linked reproductive data in the United Kingdom. Through a systematic search, we have explored the breadth of research activity in this field. Building on terms used in a preexisting publication on data linkage [2], we have generated sensitive search strings and identified many relevant studies. The search, however, was limited to electronic databases and did not include methods such as searching topic-related websites or checking references, which could have identified additional studies. Even without these additional searches, the number of studies linking reproductive health data, particularly over the last ten years, demonstrates growing interest in this area. We extracted and summarised key information about each primary study spanning both the scope and methodological approach used. Given the large volume of studies identified, we chose to utilise a closed category data extraction tool, rather than one which was comprised of open-ended questions, and instead included pre-defined categories to give a comprehensive and high-level picture of the identified studies' characteristics. Our analysis, like any other, is dependent on the accuracy with which review authors code their findings. Quality assurance procedures were in place to ensure consistency throughout the review team. However, despite these efforts, some details may still have been missed.

Conclusion

There is now a substantial body of research using linked UK reproductive health data. This has allowed researchers to answer novel questions across reproductive health domains, stages of the life course, service types, and different individuals. However, there are significant inequalities and gaps in which populations and reproductive health domains are represented in these studies and methods of linkage are often poorly reported, limiting the ability to assess the reliability of conclusions. Further efforts to create a comprehensive linked reproductive health data resource with robust linkage methods and public mandate, would enable us to fill data gaps, examine inequalities, explore reproductive health trajectories, and understand care experiences and patient journeys that span providers across primary, secondary, tertiary and independent care.

Contributors

Jennifer Hall: Conceptualisation; Methodology; Investigation; Writing – Original Draft; Supervision; Funding acquisition. Lois Harvey-Pescott: Methodology; Investigation; Writing – Original Draft; Visualisation; Project Administration; Sum Yue (Jessica) Ko: Investigation; Data Curation; Writing – Original Draft; Writing – Review & Editing; Rose Stevens: Writing – Original Draft; Writing – Review & Editing; Visualisation; Neha Pathak: Methodology; Investigation; Writing – Review & Editing; Ifra Ali: Investigation; Writing – Review & Editing, Geraldine Barrett: Methodology; Investigation; Writing – Review & Editing; Jenny Shand: Conceptualisation; Methodology; Investigation; Writing – Review & Editing; Funding acquisition; Kelly Dickson: Conceptualisation; Methodology; Validation; Investigation; Resources; Data Curation; Writing – Original Draft; Supervision, Funding acquisition.

Declaration of interests

We declare no competing interests.

Ethics statement

Ethics approval was not required for this scoping review as no new data was collected and all information included is available online.

Data sharing

As this study was a systematic scoping review of published studies, all data are publicly available in the specified references. The protocol for this study is available at https://osf.io/8s9c4.

Acknowledgements

This research is funded by the National Institute for Health and Care Research (NIHR) Policy Research Unit in Reproductive Health (Grant number: NIHR206129). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Jo Lysons contributed to defining the reproductive health concept, constructing the search strategy, and performing the searches for this scoping review. Patricia Lohr provided particular input on abortion search terms.

Supplementary material

S1: PRISMA checklist
S2: Example of search strategy for one data base (Medline) and any additional methods
S3: List of datasets in linked within included studies organised by name and by frequency of inclusion

References

1. Hawkins A. Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. Archival Science 2022;22:319–44. https://doi.org/10.1007/S10502-021-09381-0/FIGURES/3

2. Harron K. Data linkage in medical research. BMJ Medicine 2022;1:e000087. https://doi.org/10.1136/BMJMED-2021-000087

3. Elstad M, Ahmed S, Røislien J, Douiri A. Evaluation of the reported data linkage process and associated quality issues for linked routinely collected healthcare data in multimorbidity research: a systematic methodology review. BMJ Open 2023;13:e069212. https://doi.org/10.1136/BMJOPEN-2022-069212

4. Field E, Strathearn M, Boyd-Skinner C, Dyda A. Usefulness of linked data for infectious disease events: a systematic review. Epidemiol Infect 2023;151. https://doi.org/10.1017/S0950268823000316

5. Dinh NTT, Cox IA, de Graaff B, Campbell JA, Stokes B, Palmer AJ. A Comprehensive Systematic Review of Data Linkage Publications on Diabetes in Australia. Front Public Health 2022;10. https://doi.org/10.3389/FPUBH.2022.757987

6. World Health Organization. Reproductive Health. World Health Organization - Western Pacific 2024. https://www.who.int/westernpacific/health-topics/reproductive-health (accessed November 11, 2024).

7. Fleischer NL, Abshire C, Margerison CE, Nitcheva D, Smith MG. The South Carolina Multigenerational Linked Birth Dataset: Developing Social Mobility Measures Across Generations to Understand Racial/Ethnic Disparities in Adverse Birth Outcomes in the US South. Matern Child Health J 2019;23:787–801. https://doi.org/10.1007/S10995-018-02695-Z/TABLES/3

8. Abdullahi I, Wong K, Glasson E, Mutch R, De Klerk N, Downs J, Cherian S, Leonard H. Are preterm birth and intra-uterine growth restriction more common in Western Australian children of immigrant backgrounds? A population based data linkage study. BMC Pregnancy Childbirth 2019;19:1–16. https://doi.org/10.1186/S12884-019-2437-X/TABLES/7

9. Cheah SL, Scarf VL, Rossiter C, Thornton C, Homer CSE. Creating the first national linked dataset on perinatal and maternal outcomes in Australia: Methods and challenges. J Biomed Inform 2019;93:103152. https://doi.org/10.1016/J.JBI.2019.103152

10. Butler EN, Zhou CK, Curry M, McMenamin Ú, Cardwell C, Bradley MC, Graubard BI, Cook MB. Testosterone therapy and cancer risks among men in the SEER-Medicare linked database. British Journal of Cancer 2022 128:1 2022;128:48–56. https://doi.org/10.1038/s41416-022-02019-7

11. Public Health England. Measuring population reproductive health: Developing a new indicator set. 2021.

12. Peters MD, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Scoping reviews. JBI Manual for Evidence Synthesis, JBI; 2024. https://doi.org/10.46658/JBIMES-24-09

13. Lely J, Morris HC, Sasson N, Camarillo ND, Livinski AA, Butera G, Wickstrom J. How to write a scoping review protocol: Guidance and template. OSF Registries 2023. https://doi.org/10.17605/OSF.IO/YM65X

14. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L, Hempel S, Akl EA, Chang C, McGowan J, Stewart L, Hartling L, Aldcroft A, Wilson MG, Garritty C, Lewin S, Godfrey CM, MacDonald MT, Langlois E V., Soares-Weiser K, Moriarty J, Clifford T, Tunçalp Ö, Straus SE. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018;169:467–73. https://doi.org/10.7326/M18-0850

15. Starrs AM, Ezeh AC, Barker G, Basu A, Bertrand JT, Blum R, Coll-Seck AM, Grover A, Laski L, Roa M, Sathar ZA, Say L, Serour GI, Singh S, Stenberg K, Temmerman M, Biddlecom A, Popinchalk A, Summers C, Ashford LS. Accelerate progress—sexual and reproductive health and rights for all: report of the Guttmacher–Lancet Commission. The Lancet 2018;391:2642–92. https://doi.org/10.1016/S0140-6736(18)30293-9/ASSET/605C0074-A031-459D-B6A6-D5CDFBDEF9BA/MAIN.ASSETS/GR10.JPG

16. Ayiku L, Levay P, Hudson T, Finnegan A. The NICE UK geographic search filters for MEDLINE and Embase (Ovid): Post-development study to further evaluate precision and number-needed-to-read when retrieving UK evidence. Res Synth Methods 2020;11:669–77. https://doi.org/10.1002/JRSM.1431

17. Ayiku L, Levay P, Hudson T, Craven J, Finnegan A, Adams R, Barrett E. The Embase UK filter: validation of a geographic search filter to retrieve research about the UK from OVID Embase. Health Info Libr J 2019;36:121–33. https://doi.org/10.1111/HIR.12252

18. Ayiku L, Levay P, Hudson T, Craven J, Barrett E, Finnegan A, Adams R. The medline UK filter: development and validation of a geographic search filter to retrieve research about the UK from OVID medline. Health Info Libr J 2017;34:200–16. https://doi.org/10.1111/HIR.12187

19. Ayiku L, Craven J, Hudson T, Levay P. How to develop a validated geographic search filter: Five key steps. Evid Based Libr Inf Pract 2020;15:170–8. https://doi.org/10.18438/EBLIP29633

20. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372. https://doi.org/10.1136/BMJ.N71

21. Hughes T. An introduction to linked data research: what it means and how it can improve adult social care. The Insight Collective - Social Care Wales 2024. https://insightcollective.socialcare.wales/whats-on/news-and-blogs/an-introduction-to-linked-data-research-what-it-means-and-how-it-can-improve-adult-social-care (accessed November 11, 2024).

22. UK Clinical Research Collaboration. UK Health Research Analysis Report 2022. 2024.

23. Moglia ML, Castano PM. A Review of Smartphone Applications Designed for Tracking Women's Reproductive Health [111]. Obstetrics & Gynecology 2015;125:41S. https://doi.org/10.1097/01.AOG.0000463053.22473.AF

24. Alfawzan N, Christen M, Spitale G, Biller-Andorno N. Privacy, Data Sharing, and Data Security Policies of Women's mHealth Apps: Scoping Review and Content Analysis. vol. 10. JMIR Publications Inc.; 2022. https://doi.org/10.2196/33735

25. Kartik K. 'The computer won't do that' - Exploring the impact of clinical information systems in primary care on transgender and non-binary adults. The Ada Lovelace Institute 2024.

26. Tazare J, Henderson AD, Morley J, Blake HA, McDonald HI, Williamson EJ, Strongman H. NHS national data opt-outs: trends and potential consequences for health data research. BJGP Open 2024;8:BJGPO.2024.0020. https://doi.org/10.3399/BJGPO.2024.0020

27. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Peteresen I, Sørensen HT, von Elm E, Langan SM. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med 2015;12:e1001885. https://doi.org/10.1371/JOURNAL.PMED.1001885

28. Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. Eur J Epidemiol 2019;34:91–9. https://doi.org/10.1007/S10654-018-0442-4/TABLES/4

29. Harron K, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, van der Meulen JH. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol 2017;46:1699–710. https://doi.org/10.1093/IJE/DYX177

30. ADR UK. Trust, Security and Public Interest: Striking the Balance A review of previous literature on public attitudes towards the sharing and linking of administrative data for research. 2020.

31. Pegg D, Devlin H, Burgis T. Concerns raised over access to UK Biobank data after 'race scientists' claims — Genetics — The Guardian. The Guardian 2024. https://www.theguardian.com/science/2024/oct/25/concerns-raised-access-uk-biobank-data-race-scientists-claims (accessed November 9, 2024).

32. Brown K. Remember That DNA You Gave 23andMe? The Atlantic 2024. https://www.theatlantic.com/health/archive/2024/09/23andme-dna-data-privacy-sale/680057/ (accessed November 9, 2024).

33. Office for Statistics Regulation. Data Sharing and Linkage for the Public Good 2024. https://osr.statisticsauthority.gov.uk/publication/data-sharing-and-linkage-for-the-public-good/ (accessed November 8, 2024).

34. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. J Biomed Inform 2015;56:80–6. https://doi.org/10.1016/J.JBI.2015.05.012

35. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. Int J Epidemiol 2015;45:954. https://doi.org/10.1093/IJE/DYV322