Evaluation of the reported data linkage process and associated quality issues for linked routinely collected healthcare data in Multimorbidity research: a systematic review

Maria Elstad (maria.elstad@kcl.ac.uk)
King’s College London

Saiam Ahmed
UCL

Jo Røislien
University of Stavanger

Abdel Douiri
King’s College London

Research Article

Keywords: data linkage, routinely collected data, multimorbidity, systematic review, reporting guidelines

Posted Date: August 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1824087/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background:

Datasets from multi-sources that routinely collect healthcare information such as patient medical records, admissions and disease registries are increasingly used for medical research. In some cases, multiple sources are combined using data linkage techniques to create comprehensive datasets. The patient records are linked on an individual level using available person level identifiers. Errors in this process can introduce bias of unknown size and direction. The objective of this systematic review was to examine how the record linkage process was reported and to understand challenges related to accessing, linking, and analysing linked routinely collected data.

Methods:

A systematic search for relevant studies was conducted in three online databases (Medline, Web of Science and Embase) in May 2021 using predefined search terms, and inclusion and exclusion criteria. All published studies using linked routinely collected data for multimorbidity research were included. Information was extracted on how the linkage process was reported, which conditions were studied together, which data sources were used, as well as challenges encountered during the linkage process or with the linked dataset.

Results:

Twenty studies were included, of which seventeen investigated at the relationship between two specified long-term conditions. Fourteen studies received the linked dataset from a trusted third party. Hospital Episode Statistics was the most common source of data (n = 5). Eight studies reported variables used for the data linkage, while only two studies reported pre-linkage checks. The quality of the linkage was assessed only by three studies, of which two reported linkage rate and one reported raw linkage figures. Only one study checked for bias by comparing patient characteristics of linked and non-linked records.

Conclusions:

The linkage process was poorly reported in multimorbidity research, even though this might introduce bias and potentially lead to inaccurate inferences drawn from the results. There is therefore a need for increased awareness of linkage bias and transparency of the linkage processes, which could be achieved through better adherence to reporting guidelines.

Background

Routinely collected healthcare data are increasingly used for medical research(1). Such data sources include disease registries, primary and secondary care databases, administrative health data and public health reporting data (1). While these are healthcare data collected for purposes other than research(2), there are several benefits of using such routinely collected healthcare data for medical research, including the accessibility of the data, the
wide geographical coverage and their comprehensive capture of individuals who access the health system for a
defined population\(^{3}\). Routinely collected data are also an efficient use of resources as they avoid the need for
new data collection.

Linkage of routinely collected healthcare data is generally done through person-level linkage using various
available identifiers. The two main types of record linkage methods are deterministic and probabilistic linkage.
Deterministic record linkage uses a uniquely shared key, and records are defined as matched if the same key is
found in both datasets and unmatched if not. Unique identifiers, such as the National Health Service (NHS)
number in the UK are the gold-standard for deterministic linkage. When a unique identifier is not available,
alternative approaches are used\(^{4}\). In probabilistic record linkage, the linkage is done by using information from
multiple, possibly non-unique, keys. (5).

To reduce the risk of disclosure, the linkage can be done by a third party. This can help create separation
between identifiers and sensitive personal information. However it can also lead to loss of important information
about the linkage process, potentially influencing the reliability of the linked dataset (6).

A concern when linking multiple datasets is the occurrence of false record matches and missed record matches,
so-called linkage error. False record matches happen when different individuals are assumed to be the same
person in the dataset, e.g. a pair twins being assigned the same NHS number. Missed record matches occur
when a match exists but has not been discovered through the linkage process, e.g., due to recording errors such
as misspelt names, mistyped unique identifiers, or missing information.

As some degree of linkage error is unavoidable, assessing the data linkage quality is important. A particular
concern is if the records that are linked – and thus can be used in the subsequent statistical analysis – differ
significantly from those that are not linked, potentially introducing bias of unknown magnitude and direction (7).

In recent years the challenges of accessing, linking, and analysing linked routinely collected healthcare data has
been highlighted (6). Reporting guidelines for studies using data linkage were first published in 2011 (8). In 2015
came the “Reporting of studies conducted using observational routinely collected health data (RECORD)”
statement\(^{1}\), while the “Guidance for information about linking data sets (GUILD)” was published in 2018 (9).
These publications all emphasise the importance of transparency before, during and after the data linkage
process, so that the potential bias can be assessed. Several statistical methods have been proposed to adjust for
the bias due to linkage error (10).

However, it is not yet known whether reporting of linkage studies is adequate, despite the availability of these
guidelines.

A field where data linkage is often used to create richer datasets is multimorbidity (11). Multimorbidity is
commonly defined as patients with at least two long-term conditions \(^{12}\), and detailed information about
different diseases is often captured in separate, national, or regional, disease specific registers. In UK alone there
are more than 200 disease registers \(^{13}\). Linked data sources from disease registries combined with primary
and/or secondary care data are therefore useful sources for understanding the clustering of diseases and
management of multiple long-term conditions.
Using multimorbidity as a case, the objective of this systematic review was to examine how the record linkage process is commonly reported and to explore challenges related to accessing, linking, and analysing linked routinely collected healthcare data. Findings from this study will feed into further guidance to understand and minimise bias due to linkage error in medical research.

Methods

Databases, search strategy and screening

Literature search strategies were developed using medical subject headings (MeSH) and text words related to data linkage, routinely collected data, and multimorbidity. MEDLINE, EMBASE, and Web of Science were searched for studies published in the 10-year period from January 2010 through December 2020. Only studies related to multimorbidity research with at least two specified conditions, following the definition of multimorbidity proposed by Hafezparast et al. (14), were included. Studies not explicitly stating the conditions studied in the abstract were excluded. The studies had to use linked data from at least two datasets of which one of the datasets had to be routinely collected healthcare data. The search was limited to the English language and human adult subjects. Studies of participants < 18 years old were excluded.

Titles and abstracts were screened in random order against the eligibility criteria. Studies with any uncertainty regarding eligibility underwent full text screening. Additionally, 20% of the full text papers were reviewed by a second reviewer. Any disagreements were discussed among the reviewers and moderated within the supervisory group.

A comprehensive protocol was written following the PRISMA-P guidelines (15) and registered with PROSPERO (16).

Data extraction and analysis

A data extraction form was created in order to standardise data collection (Appendix 2). The form was piloted on the first 10 full text papers, refined, and then used for all full text papers. The information extraction focused on description of data sources and the data linkage process. Supplementary materials were accessed when referenced with regards to the linkage process in the full text. To validate the data extraction, an independent researcher extracted data from 10 randomly selected full text papers.

A narrative synthesis in accordance with the guidance by Popay et al. (17) was carried out to summarise the multi-morbidity conditions studied together, data sources used and comprehensively describe the reported evaluation of data linkage quality, metrics used, concerns raised by researchers regarding linkage bias and adjustments made to account for linkage error.

The quality of the reported linkage was assessed using a customised checklist created for this study, as no standardised quality assessment tools were available. Other researchers have followed a similar approach (18), (19). The customised checklist was based on the items related to data linkage in the RECORD statement (1) and the proposed checklist for reporting key elements of the linkage process by Pratt et al (20). The customised checklist has 6 domains; ‘Identified as linked routinely collected data’, ‘Data source’, ‘Linkage variables’, ‘Linkage methods’, ‘Linkage results’ and ‘Linkage evaluation’. All questions were assigned four possible answers ‘yes’, ‘no’,...
‘partially’ and ‘not applicable’. The answers were weighted following a 5-point system; ‘yes’= 5, ‘partially’=3, ‘no’=1. The ‘not applicable’ questions were not included in the denominator when calculating the overall mean score. The quality of linkage was considered good when a paper scored 4 or more points and acceptable with 3 points.

Results

Study characteristics

Initially, 1872 records were identified. Of these, 608 were duplicate records, leaving 1264 titles and abstracts for further screening. The main reasons for exclusion were violation of the multimorbidity inclusion criteria (n=834) and conference abstracts (n=261). After a full text assessment, six more studies were excluded. In total 20 reports were included in this review. These 20 studies utilised data from 10 different countries, most commonly from the UK (n=8, 40%), including two studies that used Welsh data only, followed by data from the US (n=4, 20%). The review inclusion process is shown in figure 1.

All studies were published after the first reporting guidelines paper for linkage studies in 2011. 65% of the studies were published after the RECORD statement from 2015, with eight 8 (40%) published after the GUILD guidelines paper from 2018.

Conditions studied

Of the 20 studies, 17 (85%) studied the relationship between two specified conditions, while three studies (15%) investigated three conditions. Diabetes was the most common condition studied (n=7, 35%), with the combination of Diabetes and Chronic kidney disease being the most prevalent (n=4, 20%).

Data sources

Fourteen studies used data linked by a trusted third party. Among the studies using UK data (n=8), the most prevalent source was Hospital Episode statistics (HES) (n=5), linked to data from the Office for National Statistics (ONS) (n=4), Clinical Practice Research Datalink (CPRD) (n=2) and The Index of Multiple Deprivation (IMD) (n=1). Both Welsh studies used data from the Secure Anonymised Information Linkage (SAIL) Databank. Two of the studies from USA used data from large data providers: the Optum Clinformatics Data Mart (CDM) database and the Rochester Epidemiology Project (REP). The three studies from Asia – Japan, Korea and Taiwan – all used national insurance data in combination with clinical, and laboratory data from annual health screenings, national health survey data, and data from a disease specific register, respectively. Details about the data sources are provided in table 1.

Reported linkage process

Five studies provided a list of variables used for linkage without specifying the linkage method. These were all unique personal identifiers, such as the National Health Service number in the UK based studies. Only three (15%) studies explicitly mentioned the data linkage method. Notably, they were three somewhat different linkage strategies. These were:
i. Probabilistic matching using name, date of birth, gender, and address as the matching variables.

ii. Interactive deterministic approach using age, sex, postcode, centre ID, death date and treatment date as matching variables following an 8-rule system described in detail in the paper.

iii. Deterministic matching using a statistical linkage key devised from letters in the first name and surname, date of birth and gender.

Only two of the studies reported doing pre-linkage quality checks, of which one study reported doing a thorough cleaning of the date of birth variable – which was one of the key variables used for their data linkage – while the other group reported that they checked all the linkage variables. Details of the checks were not provided.

Quality measures of the linked dataset, checks for bias and statistical adjustment

Seventeen of the 20 studies (85%) did not report any measurements of the quality of the linked dataset. Two of the three studies that did report quality measurements only reported the percent linkage rate, which was 87% for one of the studies and 99.8% for the second study.

The third study reported the number of linked and non-linked records without any summary measures in the appendix. The expected linkage rate was not reported, it was therefore unknown if the non-linked records should have been linked or not.

Only one study performed checks for bias by comparing patient characteristics in the matched vs unmatched group. They concluded that there was an absence of any major selection bias. None of the 20 studies used statistical methods to adjust for potential linkage error.

Reported issues related to the linkage process

Five of the 20 studies reported issues related to the linkage process. There were six issues raised in total, details of the specific issues are reported below.

i. The linked data sources had different start dates, with at most a 9-year difference in the start dates between the electronic registers. The hospital admission data was available from 1991 to present, the data on death registrations from 1995 to present and the GP practice data were available from 2000 to present (21).

ii. The extent to which general practise (GP) data are retrospectively coded from paper records of early years of life into electronic health record varies among GP practices. Re-entering the data into electronic health records could lead to increased number of errors, which in turn can influence the linkage quality (21).

iii. Availability of datasets containing the variables needed to answer the research question. In the study that reported this issue, the team was looking for laboratory results to be linked with administrative claims data. The laboratory results were only available for a subset of patients, reducing the potential sample size by 70%, as only records with laboratory result were included in the final dataset (22).

iv. The lack of one unique identifier. The team that encountered this issue decided to use multiple variables that were available in both datasets. However, some of the overlapping variables were calculated in different
ways. For instance, age was calculated at different timepoints in the two datasets, resulting in potential discrepancies and thereby potentially an increased number of false and/or missed matches (23).

v. Time it took to access the data. The ethics approval took more than half of the time allocated to the project and was complicated by variations in parameters required for each site-specific study approval. The extraction of the data at local sites was made challenging by the outmoded hardware which struggled to handle the computational load (24).

vi. A subset of desired records was not linked. The study therefore decided to add non-linked patient records with the disease of interest to the linked dataset (25).

**Reported issues related to the datasets**

Eleven (55%) of the studies reported various issues related to the collected datasets. In total fifteen issues were reported, which can be split into two main categories: misclassification of disease status (n=7) and missing data (n=8).

The seven issues related to misclassification of disease status included the following:

- Four studies expressed concerns about the coding systems (21, 26-28). One study pointed out recording differences between versions 9 and 10 of the International Statistical Classification of Diseases and Related Health Problems (ICD) (27).
- One study pointed out that claims data carry a potential for misclassification of patients’ diagnoses, since the presence of a diagnosis code on a claim may not indicate the presence of a disease, but a rule-out code (22). To address this limitation, the study reportedly used a validated algorithm, yet details for this were not provided.
- A study noticed a 9.3% discrepancy in the recorded diabetes status between the Système National des Données de Santé database (SNDS) and the French Epidemiology and Information Network registry (REIN) (23). The study acknowledged that these records could be false-positive matches. As an alternative, they commented that some patients recorded as having type 2 diabetes in REIN might not have needed medication, and therefore were not recorded as diabetic in the SNDS database as that database is based on reimbursement of ambulatory healthcare procedures and hospital activity.
- A study mentioned a possible misclassification bias from the case definitions of epilepsy, dementia, and subtypes of dementia (28). The study noted that dementia and subtypes of dementia in general are challenging to classify.

The eight issues related to missing data included the following:

- Three studies mentioned that the project was confined by the recorded information, and that the researchers were unable to examine the records to ascertain accuracy (28-30).
- One study mentioned using missing data for disease specific variables as a proxy for a person not having the condition, e.g., individuals with no information on stroke status were classified as not having a stroke. Absence of evidence does however not equal evidence of absence, and the study acknowledged that this approach could lead to misclassification of the disease status (21).
- Four studies pointed out that key variables for the studies were not routinely recorded, not available or only recorded in a small subgroup (25, 29, 31, 32).
Reported linkage grading

All studies underwent detailed linkage grading (Table 2). The overall mean score was 2.5.

The first two domains, ‘Identified as linked routinely collected data’ and ‘Data source’ were well recorded. Fifteen (75%) of the studies were identified as studies using linked routinely collected data in the title or abstract. The data sources were either clearly or partially described in all twenty papers. Within the data source domain, the type of data was clearly described in all studies, while the origin of the data was clearly described in 17 (85%) and partially described in three (15%). Population coverage for each data source was clearly mentioned by seven (35%), partially mentioned by six (30%) and not mentioned by seven (35%) of the studies. None of the studies mentioned whether the selected data sources were representative for the study population.

The mean score for the linkage variables domain was 1.5. A total of eight (40%) of the studies provided the list of variables used for the linkage. Of these eight one (12.5%) described the quality of the linkage variables in terms of missingness, completeness and precision.

The linkage methods domain had a mean score of 1.9, with only three (15%) studies reporting the method of data linkage.

The fifth domain, linkage result, had only four (20%) studies. Two (10%) of these were clearly reported and two (10%) were partially reported.

The linkage evaluation domain had a median grade of 1 (IQR=1.2). The linkage verification was clearly reported by one (5%) study and partially reported by two (10%) studies. Linkage validation through providing discrete measures of true and false matches and describing the origin of the reference standard dataset was partially done by five (25%) of the studies.

Table 1: Study characteristics
<table>
<thead>
<tr>
<th>Authors</th>
<th>Ref</th>
<th>Year</th>
<th>Country</th>
<th>Conditions studied</th>
<th>Data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chou et al.</td>
<td>(32)</td>
<td>2020</td>
<td>Taiwan</td>
<td>Thyroid diseases and Myasthenia gravis</td>
<td>Taiwan National Health Insurance Database and Registry of Catastrophic Illness database</td>
</tr>
<tr>
<td>Folkerts et al.</td>
<td>(22)</td>
<td>2020</td>
<td>USA</td>
<td>Chronic kidney disease and Diabetes</td>
<td>Optum Clinformatics Data Mart database</td>
</tr>
<tr>
<td>Meier et al.</td>
<td>(26)</td>
<td>2020</td>
<td>UK</td>
<td>Schizophrenia, Bipolar disorder and Multiple sclerosis</td>
<td>HES and ONS</td>
</tr>
<tr>
<td>Raffray et al.</td>
<td>(23)</td>
<td>2020</td>
<td>France</td>
<td>Chronic kidney disease and Diabetes</td>
<td>French Epidemiology and Information Network and Système National des Données de Santé</td>
</tr>
<tr>
<td>Schnier et al.</td>
<td>(21)</td>
<td>2020</td>
<td>Wales</td>
<td>Epilepsy and Dementia</td>
<td>SAIL Databank</td>
</tr>
<tr>
<td>Choi et al.</td>
<td>(33)</td>
<td>2019</td>
<td>Korea</td>
<td>Metabolic Syndrome and Chronic Obstructive Pulmonary Disease</td>
<td>Korean National Health and Nutrition Examination Survey and National Health Insurance</td>
</tr>
<tr>
<td>Lawson et al.</td>
<td>(25)</td>
<td>2019</td>
<td>UK</td>
<td>Type 2 Diabetes and Heart Failure</td>
<td>CPRD, HES, ONS and IMD</td>
</tr>
<tr>
<td>Okosieme et al.</td>
<td>(31)</td>
<td>2019</td>
<td>Wales</td>
<td>Graves' disease and Cardiovascular morbidity</td>
<td>SAIL Databank</td>
</tr>
<tr>
<td>Shiels et al.</td>
<td>(34)</td>
<td>2018</td>
<td>USA</td>
<td>Cancer and HIV</td>
<td>HIV and Cancer registries</td>
</tr>
<tr>
<td>Cooper et al.</td>
<td>(29)</td>
<td>2017</td>
<td>USA</td>
<td>Heart Failure, Diabetes and Chronic Kidney Disease</td>
<td>American Heart Association's Get with the Guidelines-Heart Failure registry and Medicare claims</td>
</tr>
<tr>
<td>Ooba et al.</td>
<td>(30)</td>
<td>2017</td>
<td>Japan</td>
<td>Dyslipidaemia and Diabetes</td>
<td>Japanese health insurance claims data and Clinical and laboratory data for annual health screenings</td>
</tr>
<tr>
<td>Pakpoor et al.</td>
<td>(35)</td>
<td>2017</td>
<td>UK</td>
<td>Testicular hypofunction and Systemic lupus erythematosus</td>
<td>HES and ONS</td>
</tr>
<tr>
<td>Wotton et al.</td>
<td>(28)</td>
<td>2017</td>
<td>UK</td>
<td>Autoimmune diseases and Dementia</td>
<td>HES and ONS</td>
</tr>
<tr>
<td>Woodhead et al.</td>
<td>(36)</td>
<td>2016</td>
<td>UK</td>
<td>Cardiovascular disease and severe mental illness</td>
<td>Lambeth Data Net and South London and Maudsley</td>
</tr>
</tbody>
</table>

Table 1 continued
<table>
<thead>
<tr>
<th>Authors</th>
<th>Ref</th>
<th>Year</th>
<th>Country</th>
<th>Conditions studied</th>
<th>Data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>McDonald et al.</td>
<td>(37)</td>
<td>2015</td>
<td>UK</td>
<td>Chronic kidney disease and Diabetes</td>
<td>CPRD, HES and ONS</td>
</tr>
<tr>
<td>Howlett et al.</td>
<td>(24)</td>
<td>2014</td>
<td>Australia</td>
<td>Mental health and intellectual disability</td>
<td>New South Wales Disability Services Minimum Data Set and Community mental health services dataset</td>
</tr>
<tr>
<td>Pelucchi et al.</td>
<td>(38)</td>
<td>2014</td>
<td>Italy</td>
<td>Pancreatic cancer, Obesity and Diabetes</td>
<td>Regional health system databases and data from 2 case-control studies</td>
</tr>
<tr>
<td>Singh et al.</td>
<td>(39)</td>
<td>2014</td>
<td>USA</td>
<td>Chronic obstructive pulmonary disease and Mild cognitive impairment</td>
<td>Rochester Epidemiology Project</td>
</tr>
<tr>
<td>Bello et al.</td>
<td>(40)</td>
<td>2013</td>
<td>Canada</td>
<td>Obesity and Chronic kidney disease</td>
<td>Alberta Kidney Disease Network database</td>
</tr>
<tr>
<td>Nedkoff et al.</td>
<td>(27)</td>
<td>2013</td>
<td>Australia</td>
<td>Diabetes and Coronary Heart disease</td>
<td>Hospital Morbidity Data Collection and the Mortality register</td>
</tr>
</tbody>
</table>

Abbreviations: CPRD = Clinical Practice Research Datalink, HES = Hospital Episode Statistics, HIV = Human immunodeficiency virus, IMD = Index of Multiple Deprivation, ONS = Office for National Statistics, SAIL = Secure Anonymised Information Linkage
Table 2: Reported data linkage summary by each domain

<table>
<thead>
<tr>
<th>Authors</th>
<th>Identified as linked routinely collected data</th>
<th>Data sources</th>
<th>Linkage variables</th>
<th>Linkage methods</th>
<th>Linkage results</th>
<th>Linkage evaluation</th>
<th>Reported linkage grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raffray et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Howlett et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Nedkoff et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Bello et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Woodhead et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Okosieme et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Cooper et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Ooba et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Choi et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Chou et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Meier et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>McDonald et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Lawson et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Wotton et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Folkerts et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Balwinder et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Schnier et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Pelucchi et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Shiels et al.</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Pakpoor et</td>
<td>●●●●●</td>
<td>●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
<td>●●●●●</td>
</tr>
</tbody>
</table>
Discussion

Main findings

The present literature review shows that in studies linking routinely collected healthcare data for use in multimorbidity research, the linkage process is rarely comprehensively reported. Although guidelines for reporting data linkage exist, the present study found that few studies adhere to the existing guidelines.

A possible explanation for the lack of data linkage reporting could be that the research teams do not have adequate information about the data linkage process of their dataset. Fourteen of the studies in this review used data that was linked by a trusted third party. From these studies it was unclear how much the authors knew about the linkage process for their dataset, including information about the origin of the datasets, linkage variables, linkage methods and evaluation of the linkage results. Insight into decisions made during the linkage is vital to understanding the dataset used for analysis, as insufficient linkage can lead to bias of unknown direction and magnitude. This information should thus be conveyed to the reader of the publication to give the reader the necessary context for interpreting the presented results.

Another explanation for the lack of reporting could be that most journals have a word limit for their publications, and detailed reporting of the linkage process might thus have been omitted. However, linkage information is important, and could at least have been included as supplementary material.

Multiple studies reported which variables were used for the data linkage but omitted to report the linkage method. A common theme for these studies were that they all used a form of unique person identifier. Access to a unique identifier is often highly valuable for linkage purposes and is sometimes seen as the gold standard of data linkage (4). They are commonly used in deterministic data linkage, and it is possible to assume that the information about the linkage method was omitted for this reason. Although the value of unique person identifiers is apparent, it is still important to consider the quality of the unique identifier in terms of completeness and accuracy (41). Unfortunately, only one study reported this information, highlighting the need for further knowledge about the impact of linkage bias and importance of clear reporting of the data linkage process.

The two main themes emerging from the reported issues regarding the dataset were misclassification and missingness. This finding is consistent with previous research using routinely collected healthcare data for research (42). A poorly or improperly recorded variable could lead to huge discrepancies between a person's actual disease status and the status they are assigned in the study. This is further emphasised as missing data for a disease specific variable often is used as a proxy for a person not having the condition. This could lead to misleading research results, and in turn can impact patient care.

This review demonstrates poor adherence to the currently available guidelines pointing to further need for clear reporting. A global initiative for enhancing the quality and transparency of health research (The EQUATOR network) highlights the importance of creating and using reporting guidelines as a tool to improve evidence-based decision making by clinicians, managers and other health professionals (43). All the included studies were published after the first reporting guidelines paper for linkage studies was published in 2011 (8). Over half were also published after the RECORD statement in 2015 and 40% were published after the GUILD guidelines paper in
2018. Although guidelines were available at the time of publication for all included papers in this review, many of their recommendations are still not being followed.

**Limitations of this review**

This review used a detailed literature strategy; however, it is possible that some studies using linked routinely collected data for multimorbidity research did not mention that they used linked data in the title, abstract or keywords and therefore were not included in this review.

The review was restricted to the field of multimorbidity, it is therefore possible that the reporting of data linkage is done differently in other medical fields.

Another limitation is that many of the studies were identified, screened, and extracted by only one reviewer, with a sample being checked by a second reviewer. Although the agreement between the reviewers were high, it is still possible that some selection and interpretation bias may exist.

**Generalisability**

The papers included in this review are international, which gives a broad overview of data linkage reporting worldwide. However, the review was limited to papers written in English language. Some key multimorbidity linkage papers might have been missed and some countries less represented due to this language criteria.

There might be regional differences in data linkage procedures and reporting standards. Between country comparison was not possible due to the small sample of papers from each country. A more in-depth review on a national level is needed to uncover any systematic challenges related to the reporting of data linkage from specific national third-party data providers.

Both finding on issues related to the dataset and issues related the data linkage process are consistent with previously published literature.

**Conclusion**

Very little was found in the literature on the question of how researchers report the data linkage process, and which concerns they might have regarding linkage bias. Further awareness of the importance of clear reporting of the data linkage process is needed, as knowledge about the linkage process can influence the interpretation and understanding of the final research results.

**Abbreviations**
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKDN</td>
<td>Alberta Kidney Disease Network</td>
</tr>
<tr>
<td>CDM</td>
<td>Optum Clininformatics Data Mart</td>
</tr>
<tr>
<td>CPRD</td>
<td>Clinical Practice Research Datalink</td>
</tr>
<tr>
<td>DS-MDS</td>
<td>New South Wales Disability Services Minimum Data Set</td>
</tr>
<tr>
<td>EQUATOR</td>
<td>Enhancing the quality and transparency of health research</td>
</tr>
<tr>
<td>GP</td>
<td>General practise</td>
</tr>
<tr>
<td>GUILD</td>
<td>Guidance for information about linking data sets</td>
</tr>
<tr>
<td>HES</td>
<td>Hospital Episode Statistics</td>
</tr>
<tr>
<td>HIV</td>
<td>Human Immunodeficiency Virus</td>
</tr>
<tr>
<td>HMDC</td>
<td>Hospital Morbidity Data Collection</td>
</tr>
<tr>
<td>IMD</td>
<td>Index of Multiple Deprivation</td>
</tr>
<tr>
<td>KNHANES</td>
<td>Korean National Health and Nutrition Examination Survey</td>
</tr>
<tr>
<td>MeSH</td>
<td>Medical subject headings</td>
</tr>
<tr>
<td>MH-COM</td>
<td>Community mental health services dataset</td>
</tr>
<tr>
<td>NHI</td>
<td>National Health Insurance</td>
</tr>
<tr>
<td>NHS</td>
<td>National Health Service</td>
</tr>
<tr>
<td>ONS</td>
<td>Office for National Statistics</td>
</tr>
<tr>
<td>PRISMA-P</td>
<td>Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols</td>
</tr>
<tr>
<td>PROSPERO</td>
<td>International Prospective Register of Systematic Reviews</td>
</tr>
<tr>
<td>RECORD</td>
<td>Reporting of studies conducted using observational routinely collected health data</td>
</tr>
<tr>
<td>REIN</td>
<td>French Epidemiology and Information Network</td>
</tr>
<tr>
<td>REP</td>
<td>Rochester Epidemiology Project</td>
</tr>
<tr>
<td>SAIL</td>
<td>Secure Anonymised Information Linkage</td>
</tr>
<tr>
<td>SNDS</td>
<td>Système National des Données de Santé</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
</tbody>
</table>

**Declarations**

**Ethics approval and consent to participate**

Not applicable.
Consent for publication

Not applicable.

Availability of data and materials

Papers included in this systematic review are listed and referenced in table 1. The dataset used and analysed during the current study is available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

ME was funded by the Unit of Medical Statistics at Kings College London.

Authors' contributions

ME wrote the protocol, extracted and analysed the data and wrote the main manuscript. AS reviewed and extracted data from a subset of the included papers. AD and JR provided guidance and feedback to both the study protocol and the final systematic review paper. All authors reviewed the manuscript.

Acknowledgements

We would like to thank Dr Katie Harron from University College London, Dr James Doidge from Intensive Care National Audit & Research Centre (ICNARC), Dr Jessica Harris from the University of Bristol and Prof Martin Gulliford from King’s College London for continuing support and guidance. Additionally, we wish to thank Dr Mark Ashworth and Dr Patrick Redman both from King’s College London for clinical guidance.

Authors' information

1 School of Life Course & Population Sciences, Faculty of Life Sciences and Medicine, King’s College London, London, United Kingdom. 2MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, UCL, London, United Kingdom. 3Faculty of Health Sciences, University of Stavanger, Norway

References


**Figures**
Figure 1

Flowchart of the paper selection process for studies into the review

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- A1Searchstrategy.docx
- A2Dataextractionform.docx