



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

Maria Elstad ¹, Saïam Ahmed,² Jo Røislien ³, Abdel Douiri¹

To cite: Elstad M, Ahmed S, Røislien J, *et al.* 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. doi:10.1136/bmjopen-2022-069212

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2022-069212>).

Received 06 December 2022
Accepted 19 April 2023



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Faculty of Life Sciences and Medicine, King's College London, London, UK

²Institute of Clinical Trials and Methodology, UCL, London, UK

³Faculty of Health Sciences, University of Stavanger, Stavanger, Norway

Correspondence to

Maria Elstad;
maria.elstad@kcl.ac.uk

ABSTRACT

Objective The objective of this systematic review was to examine how the record linkage process is reported in multimorbidity research.

Methods A systematic search was conducted in Medline, Web of Science and Embase using predefined search terms, and inclusion and exclusion criteria. Published studies from 2010 to 2020 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. Fourteen studies received the linked dataset from a trusted third party. Eight studies reported variables used for the data linkage, while only two studies reported conducting prelinkage checks. The quality of the linkage was only reported by three studies, where two reported linkage rate and one 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.

PROSPERO registration number CRD42021243188.

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

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ This is the first systematic methodology review providing insight into how the data linkage process is reported in multimorbidity research.
- ⇒ Thorough literature search and reporting following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.
- ⇒ Small group of studies that met the inclusion criteria.
- ⇒ Publications included were restricted to English language only.

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, for example, 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, for example, 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 have 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. 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 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 (online supplemental materials 1 and 2). 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. The age criteria was set because while age in principle should not impact the linkage process, in practice children appear in datasets nested within families or schools, leading to a more advanced linkage process; governance regarding access to data on children is stricter in many countries adding potential challenges; and multimorbidity tends to increase with age.

The literature search took place in May 2021.

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 Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols guidelines¹⁵ and registered with PROSPERO.¹⁶

Data extraction and analysis

A data extraction form was created in order to standardise data collection (online supplemental material 3). 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. Online supplemental 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 multimorbidity 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. No subgroup analysis was performed.

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 statement¹ and the proposed checklist for reporting key elements of the linkage process by Pratt *et al*.²⁰ The customised checklist has six 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.

Patient and public involvement

No patients involved

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. About 65% of the studies were published after the RECORD statement from 2015, with 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 3 (15%) studies 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%).

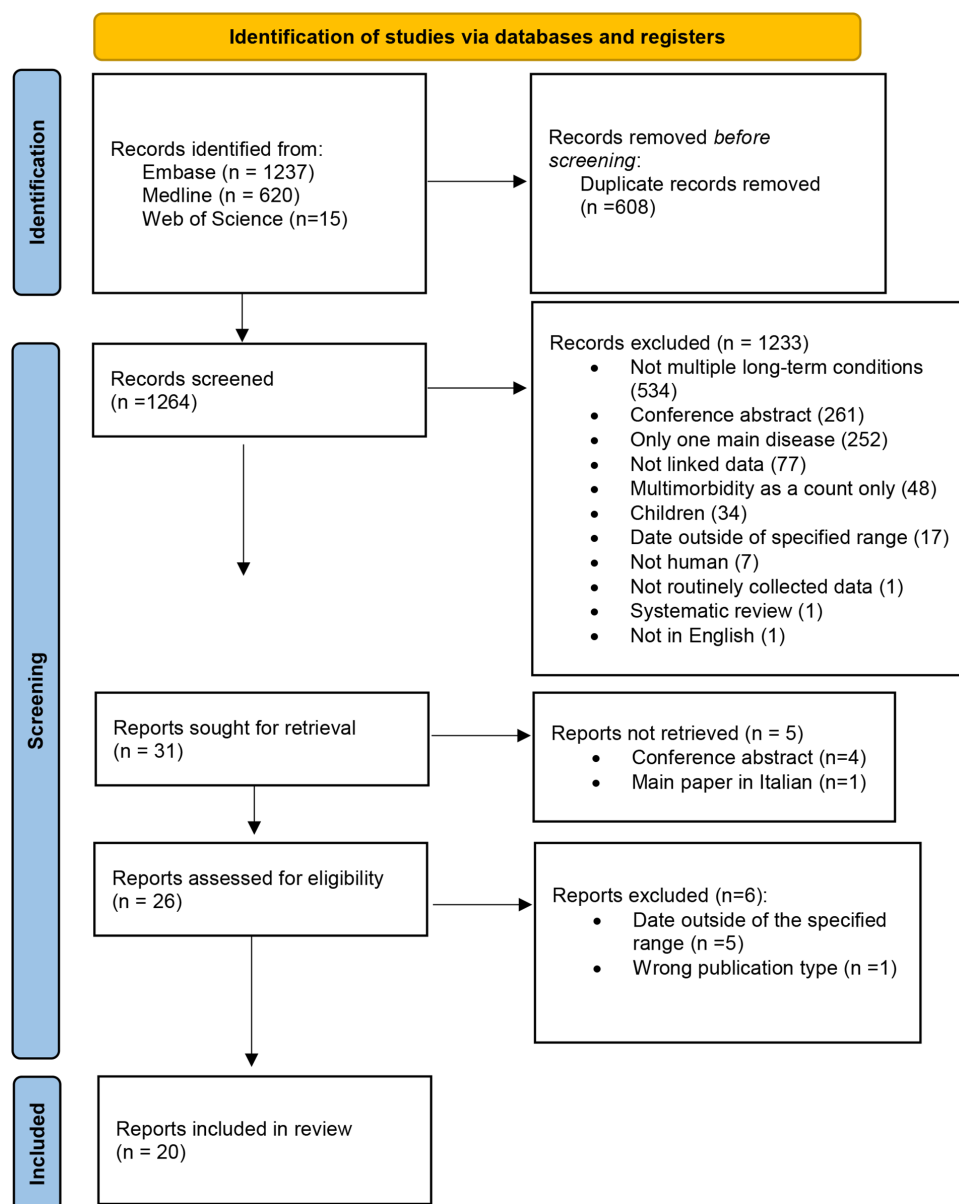


Figure 1 Flowchart of the paper selection process for studies into the review.



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](#).

Use of reporting guidelines

Only one study mentioned using data linkage reporting guidelines. Both the RECORD statement and the GUILD

Table 1 Study characteristics

Authors	Year	Country	Conditions studied	Data sources
Chou <i>et al</i> ³²	2020	Taiwan	Thyroid diseases and myasthenia gravis	Taiwan National Health Insurance Database and Registry of Catastrophic Illness database
Folkerts <i>et al</i> ²²	2020	USA	Chronic kidney disease and diabetes	Optum Clinformatics Data Mart database
Meier <i>et al</i> ²⁶	2020	UK	Schizophrenia, bipolar disorder and multiple sclerosis	HES and ONS
Raffray <i>et al</i> ²³	2020	France	Chronic kidney disease and diabetes	French Epidemiology and Information Network and Système National des Données de Santé
Schnier <i>et al</i> ²¹	2020	Wales	Epilepsy and dementia	SAIL Databank
Choi <i>et al</i> ³⁶	2019	Korea	Metabolic syndrome and chronic obstructive pulmonary disease	Korean National Health and Nutrition Examination Survey and National Health Insurance
Lawson <i>et al</i> ²⁵	2019	UK	Type 2 diabetes and heart failure	CPRD, HES, ONS and IMD
Okosieme <i>et al</i> ³¹	2019	Wales	Graves' disease and cardiovascular morbidity	SAIL Databank
Shiels <i>et al</i> ³⁷	2018	USA	Cancer and HIV	HIV and Cancer registries
Cooper <i>et al</i> ²⁹	2017	USA	Heart failure, diabetes and chronic kidney disease	American Heart Association's Get with the Guidelines-Heart Failure registry and Medicare claims
Ooba <i>et al</i> ³⁰	2017	Japan	Dyslipidaemia and diabetes	Japanese health insurance claims data and Clinical and laboratory data for annual health screenings
Pakpoor <i>et al</i> ³⁸	2017	UK	Testicular hypofunction and systemic lupus erythematosus	HES and ONS
Wotton <i>et al</i> ²⁸	2017	UK	Autoimmune diseases and dementia	HES and ONS
Woodhead <i>et al</i> ³⁹	2016	UK	Cardiovascular disease and severe mental illness	Lambeth Data Net and South London and Maudsley
McDonald <i>et al</i> ⁴⁰	2015	UK	Chronic kidney disease and diabetes	CPRD, HES and ONS
Howlett <i>et al</i> ²⁴	2014	Australia	Mental health and intellectual disability	New South Wales Disability Services Minimum Data Set and Community mental health services dataset
Pelucchi <i>et al</i> ⁴¹	2014	Italy	Pancreatic cancer, obesity and diabetes	Regional health system databases and data from two case-control studies
Singh <i>et al</i> ⁴²	2014	USA	Chronic obstructive pulmonary disease and mild cognitive impairment	Rochester Epidemiology Project
Bello <i>et al</i> ⁴³	2013	Canada	Obesity and chronic kidney disease	Alberta Kidney Disease Network database
Nedkoff <i>et al</i> ²⁷	2013	Australia	Diabetes and coronary heart disease	Hospital Morbidity Data Collection and the Mortality register

CPRD, Clinical Practice Research Datalink; HES, Hospital Episode Statistics; IMD, Index of Multiple Deprivation; ONS, Office for National Statistics; SAIL, Secure Anonymised Information Linkage.

guidelines were referenced. The data linkage process was well reported for this study.

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 3 (15%) studies explicitly mentioned the data linkage method. Notably, they were three somewhat different linkage strategies. These were:

1. Probabilistic matching using name, date of birth, gender and address as the matching variables.
2. 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.
3. 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 prelinkage 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 (85%) studies 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 per cent 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 versus 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.

1. 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 were available from 1991 to present, the data on death registrations from 1995 to present and the general practice (GP) data were available from 2000 to present.²¹
2. The extent to which GP data are retrospectively coded from paper records of early years of life into electronic health record varies among GPs. Re-entering the data

into electronic health records could lead to increased number of errors, which in turn can influence the linkage quality.²¹

3. 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
4. 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 time-points in the two datasets, resulting in potential discrepancies and thereby potentially an increased number of false and/or missed matches.²³
5. 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
6. 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 15 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:

1. 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
2. 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.
3. 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,

**Table 2** Reported data linkage summary by each domain

Authors	Year	Identified as linked routinely collected data	Data sources	Linkage variables	Linkage methods	Linkage results	Linkage evaluation	Overall reported linkage score
Chou <i>et al</i> ³²	2020	●●●●●	●●●●○	●○○○○	●○○○○	●○○○○	●○○○○	●●○○○
Folkerts <i>et al</i> ²²	2020	●●●○○	●●●●○	●○○○○	●●○○○	●○○○○	●○○○○	●●○○○
Meier <i>et al</i> ²⁶	2020	●●●●●	●●●○○	●○○○○	●●○○○	●○○○○	●○○○○	●●○○○
Raffray <i>et al</i> ²³	2020	●●●●●	●●●●○	●●○○○	●●●●○	●●●●●	●●●●○	●●●●○
Schnier <i>et al</i> ²¹	2020	●●●●●	●●●○○	●○○○○	●●○○○	●○○○○	●●○○○	●●○○○
Choi <i>et al</i> ³⁶	2019	●●●○○	●●●○○	●○○○○	●●○○○	●○○○○	●○○○○	●●○○○
Lawson <i>et al</i> ²⁵	2019	●●●●●	●●●○○	●○○○○	●○○○○	●●●○○	●○○○○	●●○○○
Okosieme <i>et al</i> ³¹	2019	●●●●●	●●●●○	●●○○○	●●○○○	●○○○○	●●○○○	●●●○○
Shiels <i>et al</i> ³⁴	2018	●○○○○	●●●○○	●○○○○	●○○○○	●○○○○	●○○○○	●●○○○
Cooper <i>et al</i> ²⁹	2017	●●●●●	●●●○○	●●○○○	●●○○○	●●○○○	●○○○○	●●○○○
Ooba <i>et al</i> ³⁰	2017	●●●●●	●●●●○	●○○○○	●○○○○	●○○○○	●○○○○	●●○○○
Pakpoor <i>et al</i> ³⁸	2017	●●●●●	●●●○○	●○○○○	●○○○○	●○○○○	●○○○○	●○○○○
Wotton <i>et al</i> ²⁸	2017	●●●●●	●●●○○	●●○○○	●○○○○	●○○○○	●○○○○	●●○○○
Woodhead <i>et al</i> ³⁹	2016	●●●●●	●●●○○	●●○○○	●●○○○	●●●○○	●○○○○	●●●○○
McDonald <i>et al</i> ⁴⁰	2015	●●●●●	●●●○○	●○○○○	●○○○○	●○○○○	●○○○○	●●○○○
Howlett <i>et al</i> ²⁴	2014	●●●●●	●●●○○	●●●○○	●●●○○	●●●○○	●●●○○	●●●○○
Pelucchi <i>et al</i> ⁴¹	2014	●●●○○	●●●○○	●○○○○	●○○○○	●○○○○	●○○○○	●●○○○
Singh <i>et al</i> ⁴²	2014	●○○○○	●●●○○	●○○○○	●●○○○	●○○○○	●○○○○	●●○○○
Bello <i>et al</i> ⁴³	2013	●●●●●	●●●○○	●●○○○	●●○○○	●○○○○	●○○○○	●●●○○
Nedkoff <i>et al</i> ²⁷	2013	●●●●●	●●●●○	●●○○○	●●●○○	●○○○○	●●○○○	●●●○○

The black markers indicate the score for each item, out of 5. Where 5 is 'well reported' and 1 'not reported'.

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, for example, 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 assigned scores were between 5 'well reported' and 1 'not reported'. The overall mean score was 2.5, indicating that the data linkage process overall was only partially reported.

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 3 (15%). Population coverage for each data source was clearly mentioned by 7 (35%), partially mentioned by 6 (30%) and not mentioned by 7 (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 8 (40%) of the studies provided the list of variables used for the linkage. Of these 8, 1 (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 3 (15%) studies reporting the method of data linkage.

The fifth domain, linkage result, had only 4 (20%) studies. Two (10%) of these were clearly reported and two (10%) were partially reported.

The linkage evaluation domain had a median score of 1 (IQR=1,2). The linkage verification was clearly reported by one (5%) study and partially reported by 2 (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 5 (25%) of the studies.

There was no indication that the overall reported linkage score was associated with year of publication. The two best reported papers were published in 2020 and 2014.

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 were 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 online supplemental materials. Encouragement from the journal editors and reviewers to use available guidelines could also impact whether authors prioritise to use guidelines when writing the papers.

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.³³ 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.³⁴ 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, which 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.³⁵ 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.

Country policies on access, confidentiality and coverage could impact the availability of information and the reporting of the data linkage process. Although both the GUILD guidelines and the RECORD statement are created with an international audience in mind, the majority of the experts creating the guidelines were from western countries, such as UK, USA, Canada, Australia and Switzerland.

There was no clear indication of an improvement of data linkage reporting over time.

The research described in the included papers occurred before the COVID-19 pandemic. The impact of the pandemic on data linkage processes and quality of reporting was therefore not assessed in this review. Further research is required to access how the changes occurring during the COVID-19 pandemic have impacted current data linkage practise.

Strengths and limitations

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

Twitter Jo Røislien @joroislien

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.

Contributors ME wrote the protocol, extracted and analysed the data and wrote the main manuscript. SA 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. ME is the guarantor for this paper.

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

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. 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.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Maria Elstad <http://orcid.org/0000-0002-7882-7564>

Jo Røislien <http://orcid.org/0000-0002-7168-2833>

REFERENCES

- 1 Benchimol EI, Smeeth L, Guttman A, *et al*. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLoS Med* 2015;12:e1001885.
- 2 Safran C. Using routinely collected data for clinical research. *Stat Med* 1991;10:559–64.
- 3 De Coster C, Quan H, Finlayson A, *et al*. Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. *BMC Health Serv Res* 2006;6:77.
- 4 Harron K, Goldstein H, Dibben C. *Methodological developments in data linkage*. Wiley, 2015.
- 5 Sayers A, Ben-Shlomo Y, Blom AW, *et al*. Probabilistic record linkage. *Int J Epidemiol* 2016;45:954–64.
- 6 Harron K, Dibben C, Boyd J, *et al*. Challenges in administrative data linkage for research. *Big Data & Society* 2017;4:205395171774567.
- 7 Doidge JC, Harron KL. Reflections on modern methods: linkage error bias. *Int J Epidemiol* 2019;48:2050–60.
- 8 Bohensky MA, Jolley D, Sundararajan V, *et al*. Development and validation of reporting guidelines for studies involving data linkage. *Aust N Z J Public Health* 2011;35:486–9.
- 9 Gilbert R, Lafferty R, Hagger-Johnson G, *et al*. Guild: guidance for information about linking data sets. *J Public Health (Oxf)* 2018;40:191–8.
- 10 Di Consiglio L, Tuoto T. When adjusting for the bias due to linkage errors: a sensitivity analysis. *SJ* 2018;34:589–97.
- 11 Lujic S, Simpson JM, Zwar N, *et al*. Multimorbidity in Australia: comparing estimates derived using administrative data sources and survey data. *PLOS ONE* 2017;12:e0183817.
- 12 Johnston MC, Crilly M, Black C, *et al*. Defining and measuring multimorbidity: a systematic review of systematic reviews. *Eur J Public Health* 2019;29:182–9.
- 13 Rankin J, Best K. Disease registers in England. *Paediatr Child Health* 2014;24:337–42.
- 14 Hafezparast N, Turner EB, Dunbar-Rees R, *et al*. Adapting the definition of multimorbidity-development of a locality-based consensus for selecting included long term conditions. *BMC Fam Pract* 2021;22:124.
- 15 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
- 16 Page MJ, Shamseer L, Tricco AC. Registration of systematic reviews in PROSPERO: 30,000 records and counting. *Syst Rev* 2018;7:32.
- 17 Popay J, Roberts H, Sowden A, *et al*. *Guidance on the conduct of narrative synthesis in systematic reviews*. University of Lancaster, 2006.
- 18 Cezard G, McHale CT, Sullivan F, *et al*. Studying trajectories of multimorbidity: a systematic scoping review of longitudinal approaches and evidence. *BMJ Open* 2021;11:e048485.
- 19 Eekhout I, de Boer RM, Twisk JWR, *et al*. Missing data: a systematic review of how they are reported and handled. *Epidemiology* 2012;23:729–32.

- 20 Pratt NL, Mack CD, Meyer AM, *et al.* Data linkage in pharmacoepidemiology: a call for rigorous evaluation and reporting. *Pharmacoepidemiol Drug Saf* 2020;29:9–17.
- 21 Schnier C, Duncan S, Wilkinson T, *et al.* A nationwide, retrospective, data-linkage, cohort study of epilepsy and incident dementia. *Neurology* 2020;95:e1686–93.
- 22 Folkerts K, Petruski-Ivleva N, Kelly A, *et al.* Annual health care resource utilization and cost among type 2 diabetes patients with newly recognized chronic kidney disease within a large U.S. administrative claims database. *J Manag Care Spec Pharm* 2020;26:1506–16.
- 23 Raffray M, Bayat S, Lassalle M, *et al.* Linking disease registries and nationwide healthcare administrative databases: the French renal epidemiology and information network (REIN) insight. *BMC Nephrol* 2020;21:25.
- 24 Howlett S, Florio T, Xu H, *et al.* Ambulatory mental health data demonstrates the high needs of people with an intellectual disability: results from the new South Wales intellectual disability and mental health data linkage project. *Aust N Z J Psychiatry* 2014;49:137–44.
- 25 Lawson CA, Zaccardi F, McCann GP, *et al.* Trends in cause-specific outcomes among individuals with type 2 diabetes and heart failure in the United Kingdom, 1998–2017. *JAMA Netw Open* 2019;2:e1916447.
- 26 Meier UC, Ramagopalan SV, Goldacre MJ, *et al.* Risk of schizophrenia and bipolar disorder in patients with multiple sclerosis: record-linkage studies. *Front Psychiatry* 2020;11:662.
- 27 Nedkoff L, Knuiman M, Hung J, *et al.* Concordance between administrative health data and medical records for diabetes status in coronary heart disease patients: a retrospective linked data study. *BMC Med Res Methodol* 2013;13:121.
- 28 Wotton CJ, Goldacre MJ. Associations between specific autoimmune diseases and subsequent dementia: retrospective record-linkage cohort study, UK. *J Epidemiol Community Health* 2017;71:576–83.
- 29 Cooper LB, Lippmann SJ, Greiner MA, *et al.* Use of mineralocorticoid receptor antagonists in patients with heart failure and comorbid diabetes mellitus or chronic kidney disease. *J Am Heart Assoc* 2017;6:e006540.
- 30 Ooba N, Setoguchi S, Sato T, *et al.* Lipid-lowering drugs and risk of new-onset diabetes: a cohort study using Japanese healthcare data linked to clinical data for health screening. *BMJ Open* 2017;7:e015935.
- 31 Okosieme OE, Taylor PN, Evans C, *et al.* Primary therapy of graves' disease and cardiovascular morbidity and mortality: a linked-record cohort study. *Lancet Diabetes Endocrinol* 2019;7:278–87.
- 32 Chou CC, Huang MH, Lan WC, *et al.* Prevalence and risk of thyroid diseases in myasthenia gravis. *Acta Neurol Scand* 2020;142:239–47.
- 33 Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, *et al.* The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. *Eur J Epidemiol* 2009;24:659–67.
- 34 Relph S, Elstad M, Coker B, *et al.* Using electronic patient records to assess the effect of a complex antenatal intervention in a cluster randomised controlled trial-data management experience from the design trial team. *Trials* 2021;22:195.
- 35 Simera I, Moher D, Hoey J, *et al.* The EQUATOR network and reporting guidelines: helping to achieve high standards in reporting health research studies. *Maturitas* 2009;63:4–6.
- 36 Choi HS, Rhee CK, Park YB, *et al.* Metabolic syndrome in early chronic obstructive pulmonary disease: gender differences and impact on exacerbation and medical costs. *Int J Chron Obstruct Pulmon Dis* 2019;14:2873–83.
- 37 Islam JY, Rosenberg PS, Hall HI, *et al.* Abstract 5302: projections of cancer incidence and burden among the HIV-positive population in the United States through 2030. *Cancer Res* 2017;77:5302.
- 38 Pakpoor J, Goldacre R, Goldacre MJ. Associations between clinically diagnosed testicular hypofunction and systemic lupus erythematosus: a record linkage study. *Clin Rheumatol* 2017;37:559–62.
- 39 Woodhead C, Ashworth M, Broadbent M, *et al.* Cardiovascular disease treatment among patients with severe mental illness: a data linkage study between primary and secondary care. *Br J Gen Pract* 2016;66:e374–81.
- 40 McDonald HI, Thomas SL, Millett ERC, *et al.* CKD and the risk of acute, community-acquired infections among older people with diabetes mellitus: a retrospective cohort study using electronic health records. *Am J Kidney Dis* 2015;66:60–8.
- 41 Pelucchi C, Galeone C, Polesel J, *et al.* Smoking and body mass index and survival in pancreatic cancer patients. *Pancreas* 2014;43:47–52.
- 42 Singh B, Mielke MM, Parsaik AK, *et al.* A prospective study of chronic obstructive pulmonary disease and the risk for mild cognitive impairment. *JAMA Neurol* 2014;71:581–8.
- 43 Bello A, Padwal R, Lloyd A, *et al.* Using linked administrative data to study periprocedural mortality in obesity and chronic kidney disease (CKD). *Nephrol Dial Transplant* 2013;28 Suppl 4:iv57–64.