BMJ Health & Care Informatics

Routine data are not 'dirty' data: how to approach using routine data for research

Grace Duffy ,[1] Alice Cunningham,[1] Vishal Sharma,[1] Tom Lawton[2,3]

To cite: Duffy G, Cunningham A, Sharma V, *et al*. Routine data are not 'dirty' data: how to approach using routine data for research. *BMJ Health Care Inform* 2026;**33**:e102018. doi:10.1136/bmjhci-2025-102018

Received 16 December 2025
Accepted 29 April 2026

Check for updates

© Author(s) (or their employer(s)) 2026. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

[1]Improvement Academy, Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK
[2]Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK
[3]University of York, York, UK

Correspondence to
Dr Grace Duffy;
grace.duffy@nhs.net

ABSTRACT

The rise of electronic patient records has led to an abundance of routine healthcare data, and the development of NHS Secure Data Environments (SDEs) is transforming their potential for research. However, routine data differ fundamentally from research data; they are generated to support care processes rather than answer predefined research questions. This creates challenges when they are analysed without understanding the context in which they were produced.

Using real examples from practice, this paper illustrates the variety of ways routine data can be shaped and how this leads to misinterpretation if treated as merely 'dirty' research data. These examples underpin practical strategies for researchers, including analysis at or by source and working closely with local data providers to understand how a dataset was created.

We argue that SDEs must foster strong links between data users and creators to preserve data provenance and support accurate interpretation. By approaching routine data as meaningful records of healthcare processes—rather than flawed research datasets—researchers can fully realise their potential to drive healthcare improvement.

INTRODUCTION

With the rise of electronic health records (EHRs) across the globe,[1] routine data collected during the process of providing care to patients are now more accessible than ever to researchers. Easy access to plentiful, contemporaneous, low-cost data is likely to shift the landscape of healthcare research forever. To fully harness the potential of routine data, NHS England is developing Secure Data Environments (SDEs) across the country, which give approved users access to relevant and anonymised routine healthcare data for research and analysis purposes.[2] However, routine data are not the same as research data. This means that it is increasingly important that researchers understand how to correctly use and interpret routine data within their work.

In this perspective piece, aimed at both current and future users of routine data, we will reflect on the differences between routine and research data, the benefits and challenges of its use and explore real-world examples that we have encountered in our own practice to explain how we should approach the analysis of routine data to overcome these challenges.

WHAT ARE ROUTINE DATA AND HOW DO THEY DIFFER FROM RESEARCH DATA?

Healthcare systems generate large amounts of data as part of their day-to-day operation. Routine data are typically collected to support care and treatment, facilitate administration and, ultimately, document care processes.[3] They are synonymous with real-world data, although for the purposes of this piece, we will use the term routine data.[4] Examples include blood results, length of admission, resuscitation status, diagnoses and many more.[5] In contrast, research data for clinical studies are collected in a purposeful and targeted manner,[3] with researchers considering which data they require to test their hypothesis and designing specific data collection methods to capture this.

The key differences between routine and research data are explored further in the table below (table 1).

WHAT ARE THE ADVANTAGES OF USING ROUTINE DATA?

The use of routine data has many advantages: large quantities of data are available quickly and retrospectively without the costs of setting up data collection and with a diversity difficult to replicate in research populations, which may improve representativeness.[5 6] There is, therefore, tremendous potential for using routinely collected data to derive new insights about health, disease and treatments.[7 8] This potential is already being realised with projects such as Connected Bradford which links anonymised routine data across primary, secondary, community and social care for previous and current patients of Bradford Teaching Hospitals NHS Foundation Trust.[9] Routine data collected as part of this research programme have demonstrated

Table 1 Key differences between routine and research data

Routine data	Research data
Records a process	Records a fact
Ad-hoc recording	Recorded at defined time points
Varied aims, for example, communication, reporting and funding	Aims to answer a research question
Used in care processes	Specifically produced for a study
Structured around care processes	Structured for analysis

that reducing exposure to pollution is likely to have a big impact on improving health, particularly for those most vulnerable.[10] These findings have prompted collaboration between the Bradford Institute for Health Research and Bradford Council in the development of a Clean Air Plan, including the introduction of Clean Air Zones across the city.[10 11]

WHAT ARE THE CHALLENGES OF ANALYSING ROUTINE DATA?

The successful use of routine data, as in the Connected Bradford project, requires researchers to have an understanding of the numerous factors which can affect the quality of routine data and their implications for data analysis. Below, we explore these factors through discussion of scenarios we have encountered within our own practice and uncover the challenges these present to researchers using routine data.

Purpose of data

Scenario

A group of researchers investigating the number of emergency department (ED) attendances at a local hospital noted a sudden large increase in attendances. On looking into this, they uncovered that the data were initially being used to test a new hospital command centre which had later gone live, explaining the apparent jump in attendances.

Challenges

Routine data are designed for operational rather than research purposes; therefore, it is vital to understand why the data are being collected. Data generated to test systems or provide staff training, for example, through test patients, can lead researchers to incorrectly interpret workload trends, resource pressures or patient outcomes if not identified prior to data analysis.

Variation in processes across systems and data inputters

Scenario

A research group wanted to analyse the proportion of patients who died in hospital in a specific region. On analysing patient records for the SNOMED code ‘died in hospital’, they found no records associated with the code. On further investigation, it was revealed that a death within hospital was coded as a discharge destination rather than as a SNOMED code.

Challenges

Identification of this variation in coding was relatively easy for these researchers, given the very unlikely probability that no patients had died in hospital. However, it could have been much more difficult if there had been variation in which code was used, leading to some patients being coded using the SNOMED code and others being coded as a discharge destination. This would have led to a lower but potentially more plausible number of records being associated with the SNOMED code, which would have been difficult to spot.

Processes are likely to vary between different healthcare organisations and between data inputters leading to certain information not being recorded or being recorded in different ways. Such variation can be difficult to spot if the data from various organisations are aggregated together, particularly if the organisation affected is small. This means that aggregating data into ‘data lakes’ could make it impossible to draw accurate inferences.

Data process issues

Scenario

Researchers sought to understand how the number of elective and non-elective admissions had changed over time at an English NHS hospital between 2015 and 2019. The graph produced is shown below (figure 1).

As can be seen, an increase in overall admissions occurred in late 2017 alongside a large spike in non-elective admissions and a smaller drop in elective admissions. This change coincided with a switch from paper records to a new EHR. Researchers were puzzled as to why the data appeared to show a sudden increase in the overall number of admissions and changes in the proportion of elective and non-elective admissions as other data confirmed that there had been no sudden, persisting changes to the overall number of admissions or to the proportion of elective and non-elective admissions.

Challenges

This scenario highlights the challenges of identifying whether variation in routine data reflects changes to reality or changes to the process the data measures. In this case, there was a clear change in the routine data produced following an alteration to the process (implementation of the EHR), suggesting the shift in the data represented this change in process rather than a change in reality. The initial suspicion was that elective admissions were being processed as non-elective admissions as this was a faster administration process. However, the graph also shows an apparent increase in the total number of admissions and the difference in non-elective and elective admissions continuing even following embedding of EHR processes. While the exact process changes responsible for this are unclear,

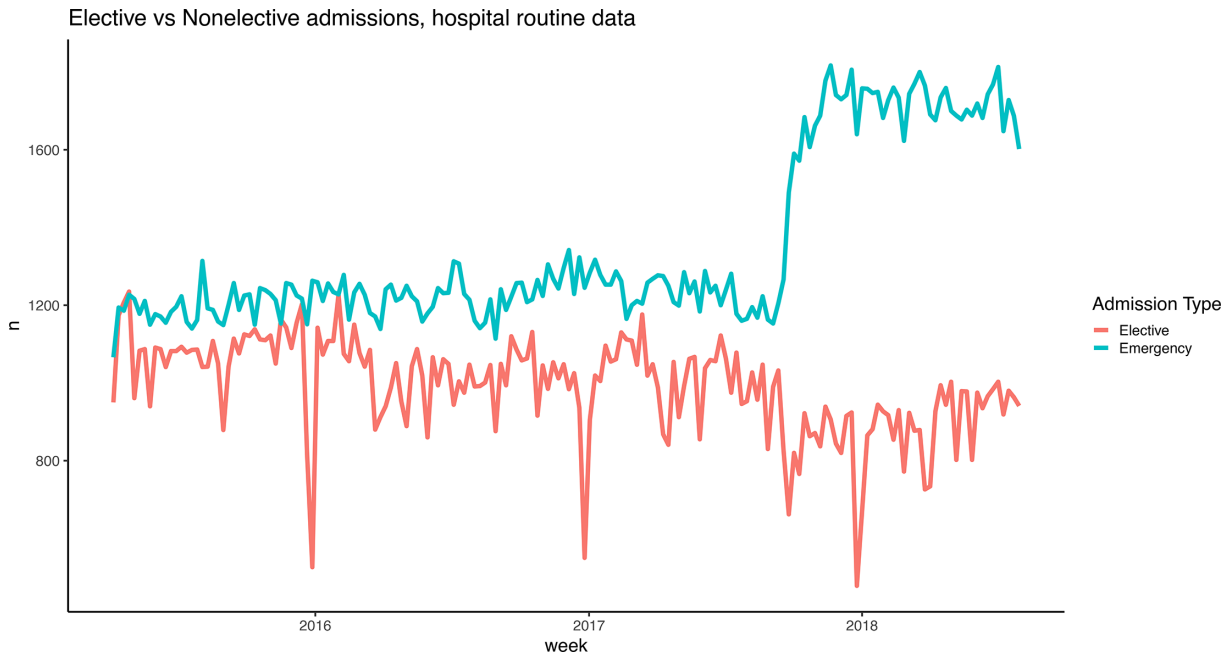


Figure 1 Elective versus non-elective admissions to an NHS hospital between 2015 and 2019.

one possible explanation may be novel recording of patients who were not formally admitted using the previous process.

Faced with this graph, it is tempting to ask 'which side is correct?'—our hope is that this paper will help to explain our belief that both sides are 'correct' in that the routine data displayed is not attempting to document an underlying constant ground-truth reality, but rather the processes in place at that hospital. Those processes changed between the two periods, and the graph accurately reflects that change. Thus, while it may be tempting to view these shifts as 'dirty' data which require 'cleaning', they may reflect changes to process rather than reality.

While some technical cleaning steps may be appropriate, such as removing duplicates or out-of-range values, routine data ultimately reflect the processes being recorded. This requires understanding the behaviour of data creators documenting both coded/structured fields and narrative free-text entries, and how they are shaped, constrained or encouraged by the system and their workflow. All forms of routine data therefore trace the recording process as much as the underlying clinical activity that we, as researchers, want it to measure.

Issues with ordering of data

Scenario

A research group was trying to identify which factors predict the tests required by patients in ED. Feeding a spreadsheet of data into artificial intelligence (AI) algorithms, the results initially appeared very promising but on review of the individual parameters the biggest predictive factor was whether they were admitted or discharged—a decision that is only made towards the end of their stay.

Challenges

Understanding when data are entered during the process can be vital to its interpretation and usefulness. An AI algorithm trained on a dataset will pick the best predictive factors, even if those factors would be unknown at the point when the output of the algorithm might be useful. As the fundamental purpose of research is to improve the real world, a model which incorporates 'future' data which would not be available at the point of use, while mathematically superior, might not be as useful as one which restricts itself to data that would be available at the point of use. Furthermore, it is relatively easy to understand that the discharge status would be unknown in this example, however, in other scenarios the timeline of data availability may not be so easily interpreted without a clear understanding of the process. This can impact on understanding the relationships between data items, for example, determining cause and effect.

Issues with timing of data

Scenario

Researchers wanted to use routine data to identify whether high blood pressure affected the risk of having a myocardial infarction. When they looked at the routine data, some patients had frequent blood pressure readings while others had no or very few blood pressure readings. The researchers struggled to agree which blood pressure readings to use and how to address missing blood pressure readings.

Challenges

If a data collection process has not been designed to answer a specific research question, then there may not be data available from a required time period or it may be difficult to unpick which data points are relevant in terms

of timing if multiple are available. Further to this, missing data cannot be ignored as it may mean something, for example, data may be absent due to lack of access to healthcare associated with socioeconomic deprivation. Without understanding its significance, there is a risk of skewed interpretation of data, which could lead to negative outcomes such as widening health inequities.

Lack of requirement for data input

Scenario

Researchers wanted to investigate the proportion of patients who accept or decline lipid-lowering medication following discussion with their GP. They planned to do this by searching the GP record for consultations with the code which denotes discussions regarding commencement of lipid-lowering medications (XaYtu). However, when reviewing the data, they found far fewer instances of the code than they had expected, especially given the number of patients who are prescribed lipid-lowering medications. On discussion with GPs, they were told that there is no financial incentive to coding such a discussion and also that there is an assumption that a discussion did occur if the medication is prescribed.

Challenges

As described above, one of the key differences between research data and routine data is that research data are designed to capture the data required to answer a posed question while routine data are produced as part of a process within healthcare. If a process does not require data to be recorded, it may well be missing. This absent data reflect the data requirements for the process to occur rather than the reality of the situation. In this scenario, the lack of coding does not describe a reality in which discussions around lipid-lowering medication did not occur but rather reflects that its documentation has not been necessary to the process of shared decision-making between clinician and patient. Such discrepancies between process and reality can make drawing accurate insights from routine data difficult or impossible. In this case, researchers were aware that the routine data likely did not reflect reality and were able to investigate this; however, in cases where researchers are not aware of the discrepancy between process and reality, the use of routine data may lead to inaccurate conclusions. While consideration may be given to ways to improve the accuracy and completeness of routine data, for example, by using financial incentives for coding, this is not feasible for all outcomes of interest to researchers and represents a further administrative burden for healthcare professionals.

Compulsory requirement for data input

Scenario

A research group wishes to assess the impact of using contrast dye during CT scans on renal function. They find inconsistencies between the renal function data they derive from blood test results and those documented on the CT requests.

Challenges

If a process inappropriately forces data entry, then the data entered may be meaningless. In this scenario, the process of requesting a CT scan required entry of the patient's most recent renal function. If the data are not available to the healthcare professional, they may either choose to 'guess' a likely figure, use a historic result or add an intentionally erroneous number, for example, 999, to 'outsmart' the system and allow them to request the scan, thereby impacting the accuracy of the data.

Missing data

Scenario

Researchers were asked to assess if the patient demographics of a service were comparable to the wider population served. A review of the patient records identified that information such as ethnicity was not routinely captured, with 60% of patients having no ethnicity data recorded.

Challenges

If data are not required for care to occur or its documentation is not incentivised by other means, for example, financial gains, there is a risk that data may not be recorded. In this instance, staff did not consider capturing ethnicity to be important for the purpose of providing care and, therefore, did not document it. Worse, the 'missingness' might not be random—people from certain groups may be less likely to have their ethnicity or other data recorded, leading to serious bias in results that cannot easily be 'papered over' using techniques like multiple imputation. Identification of these gaps, the reasons for them and the likely effect of excluding or imputing data is essential to effective analysis.

IT glitches and system errors

Scenario

Researchers investigating average length of stay in a local ED note it is longer than expected. On examining the data, they note it is skewed by a small number of patients who appear to have stayed in ED for several months. On discussion with the ED team, they report they have been unable to remove these patients from the system and that the IT team are working on a solution.

Challenges

IT glitches and system errors can affect the accuracy of routine data. In this example, these errors skewed average length-of-stay calculations, making ED performance appear worse than it really is. If such errors and glitches are not spotted, investigated and corrected, inaccurate conclusions may be drawn, thus undermining trust in the data.

HOW SHOULD I APPROACH THESE CHALLENGES?

Data are not the endpoint for research. To produce clinically useful evidence for decision-making from data, these

must be transformed through the knowledge hierarchy from data to information, knowledge and wisdom.[12] The interpretation of data within their context and relationships transforms the unorganised facts of data into information.[12] Knowledge is then produced through understanding of this information in terms of patterns. Finally, this knowledge can be extrapolated to determine fundamental principles or wisdom. The transformation from data to knowledge is achieved through analysis of routine data within scientific studies designed to test hypotheses or answer research questions, thus leading to the generation of real-world evidence.[13] Through such analysis, routine data have been able to provide valuable real-world evidence of treatment safety and adverse effects, disease epidemiology and treatment outcomes among others, with such evidence influencing real-world decision-making.[5 14–16]

The first step in transforming data into information and ensuring accurate analysis of routine data for real-world evidence is understanding the data within its context: the processes used to create them. In order to do this, we make the following recommendations.

Keep data at source

Robust data provenance maximises the ability to interpret the data in context. Ideally, data should remain at source and be analysed locally with an understanding of the processes through which it is generated.

Develop relationships with local data providers/collectors

Understanding the processes involved in creating routine data requires communication with local data collectors and providers. Development of close relationships with these teams enables researchers to reach out to local data providers/collectors to gain insight into how the data were created and raise any questions they may have about the local processes behind it.

Work with local data providers/collectors to understand the processes involved in creating the data

To understand routine data accurately, we must understand the processes through which they were created and their vulnerabilities. Liaw *et al* advocate for researchers to liaise with local data providers/collectors to understand six key areas of the data: the purpose of the data, the processes used to collect them, errors, accuracy, traceability and curation.[17] Understanding these characteristics is vital to identifying whether changes in the data reflect a change in reality or a change in the processes being measured. The risks of interpreting routine data without this knowledge have been amply demonstrated through our above examples.[18]

Complete local analysis before aggregating data from different sources

Aggregation of different data sources prior to analysis can lead to discrepancies between data sets being missed. One solution to this is the use of federated data, where analysts write code to run locally at each local data source, with results reviewable separately before aggregation later. This is greatly enhanced by Common Data Models such as the Observational Medical Outcomes Partnership (OMOP) model, meaning that code can be simply reused across 'technically harmonised' datasets.[19 20] The individual review provides a 'sense-check' of the data and its interpretation, identifying any inconsistencies before collation of results.

These guiding principles of routine data analysis are summarised in the routine data analysis checklist below (table 2).

WHAT DOES THE FUTURE OF ROUTINE DATA IN RESEARCH LOOK LIKE?

The development of multiple subnational SDEs and a single national SDE places the emphasis very firmly on the growing importance of routine data for research. However, the designs of these SDEs vary, which can have consequences for researchers' ability to understand the provenance of the data, which has been demonstrated as being so vitally important to accurate analysis in our above examples. Some platforms bring data together in one location as a 'data lake', which potentially removes the ability to trace data back to their original source. This may risk misinterpretation of changes in processes as changes in reality due to inability to understand the changes in context.

When analysing routine data, SDEs should act as the link between the data providers and the data users to allow researchers to ask questions about how the data was created. The solution to this is to ensure that SDEs are designed in a way in which data are still attributable to the producing organisations, with clear links to the data creators to enable researchers to understand their processes. This ensures that data users can liaise with data providers to interpret data alongside experts in the local organisations in which it was created. This allows

Table 2 Routine data analysis checklist

Routine data analysis checklist	
Keep data at source	✓
Develop relationships with local data providers/collectors	✓
Work with local data providers/collectors to understand the processes involved in creating the data	✓
Complete local analysis before aggregating data from different sources	✓

researchers and analysts to engage directly with the data providers to understand the context and processes under which the data were collected and how this might impact any planned analysis. Adopting such an approach enables data from multiple teams, organisations and/or regions to be combined with confidence, as there is assurance that the data are sufficiently similar to enable large scale analysis.

CONCLUSION

The increasing availability of routine data through EHRs and NHS SDEs offers a wealth of opportunities to carry out quick, low cost, relevant research which is representative of the populations we serve.[7] However, in order to take full advantage of this data, we need to recognise that routine data are not just 'dirty' research data. Rather than making misguided attempts to clean it, we need to take a different approach. This will require engagement with local teams who produce the data to ensure that we have a clear understanding of which processes they measure, how they are collected and to consider together whether changes in the data reflect a change in process or a change in reality. However, to achieve this, our SDEs need to be designed in a way that promotes local engagement with data providers and inputters by keeping data at source rather than conflating it into a 'data lake'. In this way, we can turn the significant potential benefits of routine data into a new reality.

Acknowledgements Our thanks to John Birkinshaw, Dan Mason and Mai Elshehaly for taking the time to provide their feedback and suggestions on the article.

Contributors Conceptualisation, TL; writing—original draft preparation, GD, AC, VS, TL; writing—review and editing, GD, AC, VS, TL. Guarantor: TL.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests AC and TL are Professionals Engagement Leads for the Yorkshire and Humber SDE. VS is Associate Director of the Yorkshire and Humber SDE. GD has no conflicts of interest.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

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: https://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD

Grace Duffy https://orcid.org/0009-0004-9346-664X

REFERENCES

1 Slawomirski L, Lindner L, Bienassis K, *et al*. *OECD health working papers 160. Progress on implementing and using electronic health record systems: developments in OECD countries as of 2021*. Paris: OECD, 2023.
2 Care S. Data saves lives: reshaping health and social care with data. London: Department of Health and Social Care; 2024. Available: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data
3 Clarke GM, Conti S, Wolters AT, *et al*. Evaluating the impact of healthcare interventions using routine data. *BMJ* 2019;365:l2239.
4 Dang A. Real-World Evidence: A Primer. *Pharmaceut Med* 2023;37:25–36.
5 Chodankar D. Introduction to real-world evidence studies. *Perspect Clin Res* 2021;12:171–4.
6 Powell AE. Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls. *BMJ Qual Saf* 2003;12:122–8.
7 Peek N, Rodrigues P. Three controversies in health data science. *Int J Data Sci Anal* 2018;6:261–9.
8 Handoko M, Man F, Brugts J, *et al*. Embedding routine health care data in clinical trials: with great power comes great responsibility. *Neth Heart J* 2024;32:106–15.
9 Sohal K, Mason D, Birkinshaw J, *et al*. Connected Bradford: a Whole System Data Linkage Accelerator. *Wellcome Open Res* 2022;7:26.
10 Bradford Clean Air Zone Charging Order 2022. 2024. Available: https://www.bradford.gov.uk/breathe-better-bradford/news-events-and-other-projects/clean-air-zone-consultation/
11 Coulson J. Clean Air Zone health impact 'encouraging'- study. BBC News; 2024. Available: https://www.bbc.co.uk/news/articles/clyvnq1lnrpo
12 Bellinger G, Castro D, Mills AD. Information, knowledge, and wisdom. 2004. Available: https://homepages.dcc.ufmg.br/~amendes/SistemasInformacaoTP/TextosBasicos/Data-Information-Knowledge.pdf
13 de Lusignan S, Crawford L, Munro N. Creating and using real-world evidence to answer questions about clinical effectiveness. *J Innov Health Inform* 2015;22:368–73.
14 Corrigan-Curay J, Sacks L, Woodcock J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *JAMA* 2018;320:867–8.
15 Schmidt M, Schmidt SAJ, Adelborg K, *et al*. The Danish health care system and epidemiological research: from health care contacts to database records. *Clin Epidemiol* 2019;11:563–91.
16 Heerspink HJL, Karasik A, Thuresson M, *et al*. Kidney outcomes associated with use of SGLT2 inhibitors in real-world clinical practice (CVD-REAL 3): a multinational observational cohort study. *Lancet Diabetes Endocrinol* 2020;8:27–35.
17 Liaw S-T, Pearce C, Liyanage H, *et al*. An integrated organisation-wide data quality management and information governance framework: theoretical underpinnings. *Inform Prim Care* 2014;21:199–206.
18 Cruz-Correia R, Rodrigues P, Freitas A, *et al*. Data quality and integration issues in electronic health records. In: Hristidis V, ed. *Information discovery on electronic health records*. Boca Raton: Chapman and Hall/CRC, 2009: 73–114.
19 Observational Health Data Sciences and Informatics. Standardised data: The OMOP Common Data Model. Available: https://www.ohdsi.org/data-standardization
20 Cheng C, Messerschmidt L, Bravo I, *et al*. A General Primer for Data Harmonization. *Sci Data* 2024;11:152.