Managing Change for Good Pharmacoepidemiology Practice in Healthcare Databases and Related Tools

By Bourke A[1], Bate A[2], Sauer B[3], Brown JS[4], Hall GC[5]

Affiliations: [1] IMS Health, United Kingdom, [2] Pfizer, United Kingdom, [3] University of Utah School of Medicine, USA, [4] Harvard Medical School, USA, [5] Grimsdyke House, United Kingdom

Abstract

There is an increasing reliance on databases of healthcare records for pharmacoepidemiology and other medical research, and such resources are often accessed over a long period of time so it is vital to consider the impact of changes in data, access methodology and the environment. The authors discuss change communication and management, and provide a checklist of issues to consider for both database providers and users. The scope of the paper is database research and changes are considered in relation to the three main components of database research: the data content itself, how it is accessed, and the support and tools needed to use the database.

Background

Electronic healthcare and claims records are now a mainstay of pharmacoepidemiology and other medical research. One of the benefits of these resources is that they allow both interrogation of historic data collected over an ever increasing number of years[1 2] and prospective analyses of sometimes undefined duration, such as in post-authorisation safety studies.[3]

There is a tension between maintaining stability of data recording over a study period and allowing a database resource to develop and expand, not only as far as adding more recent data, but also in terms of model and tools, as, for example, new data elements are collected or medical knowledge evolves. If a database changes after a study is complete then there may follow reproducibility issues, and if the database evolves during the course of a study,

this may impact on analyses. Indeed, since research is usually a secondary use of information collected for other purposes, such as medical recording, administrative or billing purposes, neither the researcher nor the database provider has complete control over the data collection process and change is inevitable. It is therefore vital that the continuity and sustainability of records is fully understood, documented and communicated to all stakeholders. Those intending to use a particular resource must be confident that they will be aware of any changes and how to deal with them. Conversely, study design and conduct must incorporate a flexibility and robustness to maximise the consistency of results over time (within study and for comparisons across studies). Consideration of sustainability factors will also increase transparency to enhance validity, reproducibility and comparability.

The purpose of this document is to highlight issues of sustainability and suggest ways of minimising the impact on research of potential problems arising from changes in databases and the data storage and extraction methodology that surrounds them. It is intended to aid those creating, managing or selecting a data resource, access tools and / or data repository methodologies, perhaps using the check-list provided as a way to minimise the occurrence of problems in the future which can occur due to changes in data/systems. Therefore it will be relevant for both Analysts/Researchers want to utilise data as well as Data Custodians/Guardians (ie people responsible for access to and protection of data in line with the organisation's security policy and relevant information governance). It should be noted that other aspects of pharmacoepidemiology database research that relate to data source choice,[4] data quality (intrinsic or extrinsic) and its assessment,[5 6] methodological and coding issues, or actual study design[7 8] and reporting, [8 9] have been covered elsewhere and are not within scope of this paper.

Database System Components which are sensitive to change

There are a number of common core components required in order to undertake database research, namely - the raw data; data repository or database structure; look-up tables and support documentation to add meaning to the raw data; and computer programs to process or extract data. Additional components may be present such as access tools (tailored tools

to allow pre-programmed queries of the data) and common data models (CDM - a restructured “view” of the data to facilitate the use of common queries across disparate databases). In each of these components there is a need to consider how a study may be affected by changes over time and how reproducibility is impacted by the changes. Some of these considerations are common to all components within the system, and others are more specific to individual components.

Issues common to all system components

a. Version control – Adequate documentation and nomenclature as appropriate is needed to fix in time and function the version of the database, program, CDM, coding system etc that is being used, for example it may be the ‘OCT 2013’ version of a particular database, and that database contained only data from contributing sites in a specified particular locality.
b. Documentation of change - All changes to any components of the system, whether routine or ad hoc, should be clearly documented, and any existing documentation affected should be updated. For example if a new data field is added to a database.
c. Communication of change - All changes, whether routine or ad hoc, should be effectively communicated to researchers accessing the data, well in advance of the change. Ideally researchers working on ongoing studies and those being planned should be consulted about future changes to allow them to understand any study specific implications and plan adaptations if necessary.
d. Institutional memory – Institutional memory is the collective facts, concepts, experiences and know-how held by a group of people and this knowledge may not be formally documented, therefore reducing reliance on institutional memory is desirable. For example if a change in database structure occurred five years ago, someone within the organisation may remember why that change occurred even if it is not written down anywhere. For successful continuity, mechanisms and processes (e.g., training, documentation) should be put in place to harness and diffuse the expertise of individuals to those researchers who would also benefit from the knowledge as it relates to their study.

e. Shared learning and communication – related in some way to institutional memory, but not restricted to an institution, there will be more effective use of databases, programs, CDMs and tools if communities are developed to discuss change issues. In the past this may have been in the form of "User Groups", but increasing technology advances have provided a plethora of on-line community platforms.
f. Impact of changes - Resources to deal with changes that arise, for example in terms of time, training needs and possible procedural/programming/ Quality Control impacts will need to be considered.
g. Audit trail – In order to be able to provide additional information on change and its consequences, there is need for a clear audit trail on who has implemented changes and run queries. This applies both to database maintenance by the database custodians, and also to researchers extracting and analysing data.
h. Backup and disaster recovery - Obviously it is always essential that data collected, look-up tables and programs written are secured, and can be recreated in the event of extreme adverse conditions. This applies even more in situations of flux where it may be necessary to step back to a system version prior to a modification.

Change issues relating to individual components

1. Databases

Considerations within a database

The following issues may not be known at the outset of a study, and they may not be within the control of the database custodian, but it is worth making an assessment on the robustness of the infrastructure in place now and in the future. This applies to both database providers and database users.

- Motivation and intention of database providers - There are many reasons for data to be collected and made available for research. It is worth considering the motivation and intention of the data custodians and whether those driving forces will sustain in order to ensure that data are accessible until the end of your study. A national database, such as the Taiwan National Health Insurance Database or the Health

Insurance Review and Assessment Service (HIRA) of South Korea will not be susceptible to the effects of acquisitions and mergers that some commercial databases may suffer.

- Privacy - The maintenance of privacy and confidentiality is paramount in medical records research. Therefore this is an ongoing issue and one that can have a major impact. For example it may be necessary to change dates to make them less specific. This it is a particularly difficult area as rules and laws on privacy and confidentiality evolve over time, geography and data collection setting.

- Funding of database - Putting in place a stable infrastructure to collect and maintain a database takes time, resources and commitment. When creating a data source or choosing a database for research, consider whether the database provider has the financial security to continue maintaining data collection at least for the length of time that data are needed.

- Maintenance of data flow - It may be difficult to predict the continuation of the data flow, especially when data may be collected from several streams, however technical systems for the collection of the data must continue to be operational over time.

- Data content – Although changes to data structure are covered above, over time there may be "softer" changes. For example, the database may include/exclude certain subpopulations, by specific location/clinic specialty/insurance provider. It is obviously vital to most studies that any changes in population or geographic make up of the database are understood.

- Contract – It is important to ensure that the database custodian is able and willing to enter into a contract that ensures that you have sufficient rights to make available or access the data throughout the study and appropriate access afterwards.

Considerations outside the database

- The healthcare system - Changes in policy and /or funding may have fundamental impacts on the content of the data collected. For example a change in payment to the healthcare provider may impact data recording (ie Quality Outcomes Framework

– QOF in the UK: http://www.hscic.gov.uk/qof). It may not be possible to anticipate or manage such a change, or to document this metadata in its entirety, but any important issues that may be local or national have direct, significant impact on the data for a study should be recorded and an analysis should be undertaken of how the change potentially affected the study.

- Additional information providers – there may be changes in the third party provision of certain data elements within the record. For example, when an institution changes a specific laboratory provider, thereby altering the way lab data gets integrated into the EHR and the significance of reference values.

- Legal - Valid legal systems/documents to collect and process the data must be in place and likely to endure throughout the study period. National and international law may have an influence on the database.

- Ethical Permissions - Valid ethical reviews and permissions to collect and process the data must have been established and endure throughout the study period.

- Coding systems /classification - Some code lists are updated regularly while others have a major change at longer intervals (e.g. ICD for disease diagnosis). Additionally there may be changes in the actual coding / classification system, and these are usually known and can be planned for well in advance. As an example, in the UK, many primary care computer systems used the OXMIS clinical coding system in the 1980s, the Read coding system from the 1990s, and they will soon move to the SNOMED system.

2. Access to Data and Analysis programs

- Program language & documentation – The use of Open Source programs –e.g. Observational Medical Outcomes Partnership (OMOP) programmes, should be considered. The difficulty may be balancing the encouragement and focus on innovation while also protecting ongoing or completed projects as there might be less incentive to protect what has already been developed rather than focus on further iterations. If an Open Source tool is used but you are unable to be active in a collective, you may not get input to developments and maintenance issues. Even if

you are active in the collective, it may decide to make changes that you do not agree with. This is different from a more traditional “paid for” product where there is an expected right for software to be maintained.

- Collaborative program development - Cumulative knowledge production may involve leveraging work completed by others. The best opportunity for success is to share program code designed to execute analytic procedures or steps in the database programming process.

  Both Mini Sentinel (MS)[10] and OMOP[11] have shared analytic programs, diagnosis and outcome code lists and data models within their communities but these are relatively closed networks around specific data sources and analytic problems (i.e., drug safety surveillance).

  Other groups such as the Academy Health Electronic Data Methods group, referred to as the EDM forum, and NIH have sponsored the development of collaborative research environments. The NIH Collaboratory provides collaboration space and a knowledge repository for distributed networks to collaborate on pragmatic clinical trials. https://www.nihcollaboratory.org/Pages/default.aspx

  Another example is the EDM Collaborative Informatics Environment for Learning on Health Outcomes (CIELO). This is a workspace environment that supports development of a user profile, uploading code and data bundles, adding publications and connecting to other CIELO users based on similar use of code or data. This type of collaborative sharing is expected to improve the transparency and reuse of database programs and code. (http://cielho.bmi.osumc.edu/help)

  Collaborative environments such as CIELO often use github (https://github.com) to maintain version control and logging changes to code. Other users can comment and validate code if deemed necessary.

  There is a balance between open source and control, which may depend on funding source, continuity of research network, and other issues.

3. Common Data Models (CDMs) and Associated Tools including Code Mappings

- Transparency – It is important to know how the CDM conversion from the source data was conducted and why. This can be managed with documented assumptions (for example around the ETL – Extract Transform Load regarding systematic use of ETL approach, searchable accessibility - preferably generated automatically) and maintenance over time. In addition, a useful parameter to document is "goodness of fit" by which we mean how well data elements of a given database version can be converted to the data elements of the Common Data Model and the level of information loss if any for each data element's conversion. Clarity is vital in accounting for differences in results when using the same data and different CDMs and associated tools[12, 13].

  Any major changes in CDM approach and/or validity of output should be described and, as far as possible, be predictive of the criteria in advance which are likely to lead to significant differences in the goodness of fit of such an approach.

- Availability - The desired aim is to maximise reasonable availability but keep consistency over time, and as far as possible prospective warning is needed when accessibility changes.

- Support – Consideration should be given to how advice on implementation and routine use of a CDM will be provided and how shared, whether it is free, and whether the support will be sustained longer term.

- Maintenance – It is very important to know how the CDM will be maintained and by whom. This includes clarification on how improvements to the CDM and associated tools will be decided and implemented.

- Governance – All governance requirements, where in place, concerning the application to use and maintain CDM conversions of given databases should be clear. This may include agreement on how the CDM will be used for analysis by each partner in a network, for example right and reasons for refusal to conduct a specific analysis. As an example, the AsPEN (Asian Pharmacoepidemiology network) group uses CDM to facilitate pharmacoepidemiology research mainly in Asian countries and

has published statements on governance (http://aspennet.asia/pdf/GovernanceStructuresForAsPEN.pdf).

- CDM outputs - There may be restrictions or guidelines on how CDM outputs should be provided and these may change over time, for example level of detail, what level and accessibility of audit trail etc. Tools and processes for continual checking of quality[14], and there perceived reliability, may be even more important when running across a network of data where a researcher cannot readily access the underlying raw data [15,16].

- Purpose built tools for CDMs – Any tools created for use with CDMS should be clearly described in terms of both their functions and their limitations so that any alteration will be obvious.

- Future developments - Clarity on plans for future improvements of CDMs and time lines is vital, as well as maintenance not just of the CDM but the surrounding system including tools.

- It should be noted that a CDM is only one tool which can be used when converting raw data into specific database structures. The points above equally apply to all methods for translating raw data to specific database structures in general, of which a CDM will be one of the options.

Checklist of issues to consider

Which components are sensitive to change?

- ☐ the raw data
- ☐ data repository or database structure
- ☐ look-up tables
- ☐ support documentation
- ☐ programs

For these elements consider

- ☐ version control
- ☐ documentation of change
- ☐ communication of change
- ☐ institutional memory
- ☐ shared learning and communication
- ☐ impact of changes
- ☐ audit trail
- ☐ backup and recovery

Within a database consider

- ☐ sustainability of motivation and intention of all parties
- ☐ privacy
- ☐ funding
- ☐ data flow
- ☐ data content
- ☐ contract

Outside the database consider

- ☐ healthcare system
- ☐ additional information providers
- ☐ coding /classification systems
- ☐ legal issues
- ☐ ethical approval
- ☐ data and analysis program access
- ☐ data and analysis program development

Where tools such as common data models (CDMs) are used consider

- ☐ transparency
- ☐ availability
- ☐ support
- ☐ maintenance

- ☐ governance
- ☐ outputs
- ☐ purpose built tools

Conclusion

In a fast moving scientific environment such as pharmacoepidemiology, the only certainty is that nothing will remain stable. Databases of healthcare emerge and grow and develop, and alongside the methodologies and technologies that support access and analysis will also flourish and evolve. As more researchers interact with this ecosystem it is essential to develop robust documentation and communication systems (possibly via websites) to anticipate record and adapt to changes over time. These responsibilities should be split between the data providers and the data users. Database custodians have a responsibility to provide warning and documentation on any changes that they are aware of, while researchers accessing databases have a role to play in both providing feedback on issues uncovered during in-depth data analysis, and on partnering with data guardians on the use of access and support tools.

References

1. Hall GC, McMahon AD, Carroll D, et al. Observational Study of the Association of First Insulin Type in Uncontrolled Type 2 Diabetes with Macrovascular and Microvascular Disease. *PLoS ONE* 2012;7(11):e49908, 10.1371/journal.pone.0049908.
2. Irizarry MC, Webb DJ, Boudiaf N, et al. Risk of cancer in patients exposed to gabapentin in two electronic medical record systems. *Pharmacoepidemiology and Drug Safety* 2012;21(2):214-25, 10.1002/pds.2266.
3. Giezen TJ, Mantel-Teeuwisse AK, Straus SM, et al. Evaluation of post-authorization safety studies in the first cohort of EU Risk Management Plans at time of regulatory approval. *Drug Saf* 2009;32(12):1175-87, 10.2165/11318980-000000000-00000.
4. Hall GC, Sauer B, Bourke A, et al. Guidelines for good database selection and use in pharmacoepidemiology research. *Pharmacoepidemiology and Drug Safety* 2012;21(1):1-10, 10.1002/pds.2229.
5. Kahn MG, Raebel MA, Glanz JM, *et al.* A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical care* 2012;**50 Suppl**:S21–9. doi:10.1097/MLR.0b013e318257dd67
6. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Medical care* 2013;**51**:S22–9.
7. Berger ML, Mamdani M, Atkins D, et al. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report--Part I. *Value Health* 2009;12(8):1044-52, 10.1111/j.1524-4733.2009.00600.x.
8. International Society for Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practices (GPP). *Pharmacoepidemiol Drug Saf* 2008;17(2):200-8, 10.1002/pds.1471.
9. Elm Ev, Altman DG, Egger M, et al. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies 1136/bmj.39335.541782.AD. *BMJ* 2007;335(7624):806-08.
10. Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, Brown, JS (2012). Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiology and drug safety*, *21*(S1), 23-31.
11. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, Welebob E, Scarnecchia T, Woodcock J. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Annals of internal medicine* 153, no. 9 (2010): 600-606.
12. Xu Y, Zhou X, Suehs BT, Hartzema AG, Kahn MG, Moride Y, Sauer BC, Liu Q, Moll K, Pasquale MK, Nair VP, Bate A. A Comparative Assessment of Observational Medical Outcomes Partnership and Mini-Sentinel Common Data Models and Analytics: Implications for Active Drug Safety Surveillance. *Drug* Saf DOI 10.1007/s40264-015-0297-5
13. Gagne JJ. Common Models, Different Approaches. *Drug Saf* DOI 10.1007/s40264-015-0313-9
14. Hartzema AG, Reich CG, Ryan PB, Stang PE, Madigan D, Welebob E, Overhage JM. Managing Data Quality for a Drug Safety Surveillance System. Drug Saf. 2013 Oct;36 Suppl 1:S49-58. doi: 10.1007/s40264-013-0098-7.

15. Ross TR, Ng D, Brown JS, *et al.* The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 2014;**2**. doi:10.13063/2327-9214.1049
16. Mini Sentinel Data Quality Review and Characterization Programs v3.2 http://www.mini-sentinel.org/data_activities/distributed_db_and_data/details.aspx?ID=131