PHARMACOEPIDEMIOLOGY AND DRUG SAFETY (2011)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.2229

COMMENTARY

Guidelines for Good Database Selection and use in Pharmacoepidemiology Research†

Gillian C. Hall[1]*, Brian Sauer[2], Alison Bourke[3], Jeffrey S. Brown[4], Matthew W. Reynolds[5] and Robert Lo Casale[6]

[1]*Grimsdyke House, London, UK*
[2]*Salt Lake City VA IDEAS Centre & Division of Epidemiology, The University of Utah, Salt Lake City, UT, USA*
[3]*CSD Medical Research, London, UK*
[4]*Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA*
[5]*United BioSource Corporation, Lexington, MA, USA*
[6]*Department of Epidemiology, Merck & Co. Inc, West Point, PA, USA*

ABSTRACT

The use of healthcare databases in research provides advantages such as increased speed, lower costs and limitation of some biases. However, database research has its own challenges as studies must be performed within the limitations of resources, which often are the product of complex healthcare systems. The primary purpose of this document is to assist in the selection and use of data resources in pharmacoepidemiology, highlighting potential limitations and recommending tested procedures. This guidance is presented as a detailed text with a checklist for quick reference and covers six areas: selection of a database, use of multiple data resources, extraction and analysis of the study population, privacy and security, quality and validation procedures and documentation. Copyright © 2011 John Wiley & Sons, Ltd.

KEY WORDS—Health databases; research design; epidemiology; pharmaceuticals; guidance

Received 08 March 2011; Revised 04 July 2011; Accepted 19 July 2011

INTRODUCTION

The use of databases of routinely collected healthcare information in pharmacoepidemiology has expanded in the last decade as awareness has increased and more and larger resources have become available. The increased speed, limitation of biases, such as some recall and reporting bias, and lower cost afforded by such databases are important, for example, when there is pressure for timely information to allow prompt public health decisions. However, database research has its own challenges. These studies must be performed within the limitations of a resource not specifically designed to test the research hypothesis but the product of complex and evolving healthcare systems. When multiple databases are needed to increase numbers of subjects or breadth of information, issues of comparability and linkage must also be addressed. In all cases, researchers must work within local and regional policy and legislation designed to protect privacy. Ultimately, not all resources can answer all pharmacoepidemiology research questions, and in some cases, a randomised clinical trial or other ad hoc design may still be required.

The primary purpose of this document is to assist investigators in the selection and use of data resources for an observational study in pharmacoepidemiology by highlighting potential limitations and recommending tested procedures. It is also anticipated that it will be an additional aid for those evaluating studies based on multi-purpose data resources[1] or evaluating a registry as a data resource to answer a specific question.[2] This paper is not intended to deal with specific methodological and coding issues or the actual study design[3,4] and reporting,[4,5] which have been covered elsewhere. Only factors relevant to the selection and use of multi-purpose data resources in pharmacoepidemiology are considered here.

*Gillian Hall, Grimsdyke House, Ravenscroft Park, Barnet, EN5 4ND, UK. E-mail: gillian_hall@gchall.demon.co.uk
†This document has been endorsed by the International Society for Pharmacoepidemiology as policy and has been posted on their website as a preprint version.

CHECKLIST FOR INVESTIGATORS IN DATABASE RESEARCH

1 Database selection

Population covered: Does the resource include an appropriate population in terms of size, coverage and representativeness?
Capture of study variables: Are all exposures, outcomes and other study variables captured in sufficient detail, without bias and accessible for research?
Continuous and consistent data capture: Are there any breaks or changes in data collection over time for either individual patients or the whole population during the study observation period? Are there any inconsistencies in provision of healthcare or capture of study variables across the database population?
Record duration and data latency: Is the average patient record duration, and the time between the occurrence of the exposure and data collection, sufficiently long for the study event?
Database expertise: Is the expertise required to use the resource available: in-house or elsewhere?

2 Use of multiple resources

Multiple resources linked to increase breadth of patient information: Can data resources be linked?
Multiple resources linked to increase numbers: Are the data sources and data systems compatible in metrics, policy and terminology?
Linkage: Is reliable person-matching possible for a sufficiently large proportion of the database population? Are experience and techniques available, and can duplicates be identified?
Data storage and analyses: In multi-institutional studies, should a central or distributed system be used?

3 Extraction and analysis of the study population

Specification of extraction: Are the following specified in detail: how to extract the study population and variables, code lists and non-coded systems, retrieval and merging of additional external data, output and final analysis?

4 Privacy and security

Compliance with privacy and security policy: Have all relevant local, regional and national policies been complied with?
Limited use of identifying information: Are all direct identifiers removed or masked? Whose responsibility is it to ensure privacy?

Secure data storage and transfer: Is there a formal data security policy, and has this been adhered to?
Review of policy and procedures: Are regular privacy reviews adhered to? Has the use of a new database, collection of additional patient or physician data, use of multiple resources, or narrative data impacted confidentiality?

5 Quality and validation procedures

Overall database: Have appropriate general quality checks been completed?
Study population: Which study-specific quality checks are needed: the extraction process, data merging, study variables, assumptions, etc.? Has the annotated programming code been reviewed by an independent programmer?
Testing: The checks can be external, logical or internal and should be cross-sectional, longitudinal and up to date.

6 Documentation

Format: Are rules of Guidelines for Good Pharmacoepidemiology Practices followed, including storage and indexing?[4]
Specifics: Have extraction specification, output, quality testing, merging resources, responsibility for privacy and annotated programming code for data extraction and final analysis been documented?

DOI: 10.1002/pds

METHODS

The document was prepared by six volunteer members of the International Society of Pharmacoepidemiology's (ISPE) special interest group on database research. The group was from Europe and North America and comprised researchers from academia, industry and services as well as a database provider. A draft document of six sections was prepared based on the deliberations of the core group and literature reviews. A checklist was prepared from the document for simple access. This draft was reviewed by five volunteers from the same special interest group and then opened to the general ISPE membership for comment.

The term 'multi-purpose database' was defined as a healthcare database used in observational research but collected for other purposes and includes, but is not limited to, electronic medical records and healthcare claims or payment records. Other terms for these databases include observational datasets, linked databases and data resources. This guidance is presented as a detailed text with a checklist for quick reference and covers six sections:

1. Selection of a database,
2. Use of multiple data resources,
3. Extraction and analysis of the study population,
4. Privacy and security,
5. Quality and validation procedures and
6. Documentation

GUIDANCE

1. Selection of a database

Data resources used in research vary widely in terms of population covered, source health system, purpose of data collection, longitudinality and variables included.[6] The specific research question should dictate whether a database study is appropriate and the type of database needed. In some cases, studies cannot definitively answer the research question but, nonetheless, can contribute to an understanding of the issue. This section aims to provide key questions used to assess the suitability of a data resource for a particular study and to encourage the research team to fully understand the resource at the planning stage. The assessment will require input from someone with detailed understanding of the database and its source healthcare system.

1.1 **Population covered:** The required study population should be clearly defined then compared with that available from the source database.

1.1.1 *Study sample size:* A database has a fixed number of subjects available for study at a given time. A power calculation based on an estimated minimal sample size can inform if the sample is sufficiently large to meet the objectives. However, the impact of all inclusion and exclusion criteria on study numbers often can only be approximated until the study population has been extracted.

1.1.2 *Coverage:* Are required subsets of the general population included? Examples of subsets include paediatrics, those undergoing procedures, or people who are too ill to work. A database that excluding those who are too ill to work may not be suitable to study some drug therapies such as one indicated in end stage renal disease.

1.1.3 *Representative sample:* Investigators should understand the representativeness of the source database and resultant study population to determine how well analyses can be generalised to the wider population at risk. Any difference between the study population and the wider population should be investigated and reported. The impact of inclusion/exclusion criteria should be understood, for example, including only patients with a full suite of test results may result in selection of a population with more severe disease.

1.2 **Study variables:** Each study exposure, outcome, inclusion or exclusion criterion, potential confounder and other information required for the study should be defined and then compared with available fields in the database.

1.2.1 *Capture of study variables:* Is each study variable, or a marker for the variable, recorded? The percentage of patients in which each study variable is captured or, when multiple records are usual, the proportion of records captured should be considered against the study requirements.

1.2.2 *Accessibility:* Accessibility within and outside the database should be considered. Database content may be stored as a coded field, free text or scanned document, and the storage location and accessibility may vary. Where electronic data are to be supplemented with external data, as in a paper chart or questionnaire, then ease and comprehensiveness of access to the patient

DOI: 10.1002/pds

or physician, as well as cost and timelines, should be considered.

1.2.3 *Level of detail:* The appropriate level of granularity must be considered for each research question. For example, is 'ischaemic' or 'haemorrhagic' stroke recorded rather than 'stroke', or can the start or end of an episode of disease or treatment be identified? For drug exposure, details of formulation, dosage or duration may be required, or it may be important to understand if drug information is captured when prescribed or dispensed. This assessment should be based on actual data recording.

1.2.4 *Biased capture:* Is the capture of variables systematically biased? For example, datasets can record abnormal test results but not always normal findings.

1.3 **Continuous and consistent data capture:** Does data capture vary over the study period either for individual patients or the whole population, or are there differences in recording between subgroups? The effect of local and national policies on capture of study variables or interpretation of data should be considered.

1.3.1 *Breaks or changes in data collection:* These can result from changes that affect the total population, such as updates to coding systems or collection software, or changes to policy on health service provision or reimbursement. Alternatively, a safety alert may suddenly modify physicians' recording of medical events. Individual patient records also can be affected, for example, data can be missing if healthcare coverage is stopped for the period when an individual is not working, or an outpatient record can stop while someone is an inpatient.

1.3.2 *Shared care:* Are patient records incomplete because data are captured at one provider, facility or level of care? Examples include when patients are entitled to enrolment by more than one healthcare provider, when certain care or treatments (e.g. antenatal care or renal dialysis) are provided by a third party or if a centre records data on only secondary care diagnoses and procedures.

1.3.3 *Exclusion of specific information:* Are variables missing because some services are free and not included in the billing data or because details are considered particularly sensitive such as an AIDS diagnosis and not recorded on the electronic record? In addition, pharmacy or prescription databases may not capture exposure details if the individual pays directly.

1.3.4 *Inconsistencies in data capture:* These may occur because of variation in healthcare provision such as no drug coverage for some groups or a policy of more detailed investigation of specified chronic diseases like diabetes.

1.3.5 *Local drug policies:* Formulary policies and treatment guidelines may impact data quality. Examples include whether the study drug is: (i) included on any formulary or treatment guidelines or (ii) limited in use, for example, to second-line or severe cases so not used in a representative population.

1.4 **Record length and latency:** The researcher should have an understanding of the mean and variability in both the duration of patient records and the time between the occurrence of the event and its availability for analysis.

1.4.1 *Record duration:* Record duration for the overall database and mean observation for individual patient groups should be considered. Is sufficient history recorded to allow identification of first ever use and exclusion of prevalent events? Is the required length of follow-up consistently available for an event with a long latency period, such as a malignancy?

1.4.2 *Data latency:* The frequency of data collection and database uploading can affect data latency. Not all components may be uploaded at the same time. This can be important in the study of a newly marketed medicine when time from product launch to data collection should be considered.

2. Use of multiple data resources

A review of database suitability may show that a study requires data from more than one source either to enhance data available through linkage of disparate sources or to expand the size of the population through combination of similar data sources.

2.1 **Data resources containing different information on the same patients:** Person-level linkage of disparate databases can allow a more robust

DOI: 10.1002/pds

evaluation by providing a more complete picture of patient care and characteristics. Common linkages include the combination of inpatient, outpatient or pharmacy data or linking cancer, death or immunisation registries to medical records and may be within or across institutions.

2.1.1 *Reliable person-level linking*: When within-patient linkage between databases is needed, in the best scenario, each dataset will include enough common relevant patient descriptors to allow a high-probability match (e.g. based on medical record number or other standardised person-level identifier, date of birth, residence). In general, the more linkage variables available the better. Patient privacy is a concern when conducting linkages and new approaches for anonymous linkage include secure hashing algorithms. Linkages across data sources typically require a probabilistic or deterministic linkage algorithm to account for ambiguity, for example, slightly different spelling of names or addresses. The choice of linking method should be based on expertise in usage of the approach, previous linkage of the databases (if any), and the acceptable balance of false positives and false negatives realising that some linkages will be incorrect and some will be missed. Additionally, the overlap in populations should be assessed because low linkage will impact sample size.

Validated linkage algorithms: Whenever possible, researchers should use validated linkage approaches specific to the databases being joined. If a validated approach is not available, a process for validation should be completed whenever feasible, and sensitivity analyses included to evaluate potential linkage errors.

Identification of duplicates: Some linking scenarios require the identification of duplicate records or information for the same individual; for example, two sources could have demographic or clinical information. In those instances, the acceptance/rejection rules for which potentially duplicate information is kept must also be specified.

2.2 **Data resources containing similar information on different patients:** Many pharmacoepidemiology studies require very large populations to address questions when the population of interest is small (e.g. cutaneous T-cell lymphoma), exposures are uncommon (e.g. safety surveillance of new treatments), or where outcomes are rare (e.g. Guillain–Barré syndrome). When no single database is large enough to address such research topics in a timely and adequate way, multiple resources, such as data from more than one healthcare provider, will be needed to identify an adequate study population. Examples include Vaccine Safety Datalink Project (http://www.cdc.gov/vaccinesafety/ Activities/VSD.html) and EU ADR (http://www.alert-project.org/).

2.2.1 **Comparability of data capture and data systems:** In this context, comparability of data sources refers to the way in which the data are captured and recorded so that the data can be reasonably combined with respect to data capture and terminology. Comparability should be assessed qualitatively through detailed understanding of the data source, and quantitatively across all relevant metrics, to ensure that information from the different sources can be combined. For instance, public and private health insurer claims databases may be comparable in that denominators can be well defined, the data may be captured via a standardised reimbursement system, and the information is recorded using standardised coding schema. However, a claims database is less comparable with electronic health record systems that may record information using clinical concepts rather than reimbursement-defined schema. For example, one may record prescriptions written, whereas the other records prescriptions dispensed.

There may be important differences between comparable data sources, and substantial care should be taken to quantitatively evaluate seemingly comparable sources to identify less obvious variability or subtle differences in data capture and coding. For example, even though claims data sources may be comparable with respect to data capture and terminology, differences in policies and practices may introduce cross-site variability that would need to be addressed analytically.

2.3 **Data storage:** A unique issue with multi-institutional studies is whether to create a central data warehouse that physically combines data from all participating institutions or to store data in a distributed fashion.[7] With a distributed model, data owners maintain physical control of their data in adherence to their privacy and security rules. In a centralised model, owners transfer data to a single location for analysis, thereby giving up control. A mixed model is also possible.

DOI: 10.1002/pds

2.4 **Data analysis:** In a centralised data warehouse, anyone with access to the data warehouse can conduct analyses. In a distributed model, there are two analytic options. The first involves *independent analytics*, which does not require the data to conform to a common data model. In this approach, a common study protocol is created, each data owner independently implements it locally and the results are returned to a central coordinating centre for aggregation or meta-analysis. For data that are stored in a common data model, a *coordinated or distributed analytic* approach may be used. In this approach, the common data model is specified and implemented, and a common study protocol is created. The central study team then creates analytic code based on the protocol and data model which is distributed to each data owner for execution.[7–9] The benefits and limitations of these approaches have been described elsewhere.[7] Both approaches should include comprehensive data characterisation analyses to evaluate variability across the data partners with respect to overall cohort metrics, such as age and sex distribution, and study-specific metrics, such as exposure and outcome rates by age, sex and year.

3. Data extraction

The first step in using multi-purpose database is to extract a specific study population (e.g. all patients over 18 years old with a diagnosis of diabetes) either by request from the data warehouse or by direct access. All predefined study variables for this study population will then be extracted for the final analysis. Much of the investigational work to understand data generation, capture and storage should be carried out prior to developing a data extraction protocol for the specific study as described in section 1. This section describes the stages in data extraction.

3.1 **Specification of data extraction:** All aspects and steps of the data extraction should be planned based on the final protocol and clearly defined in the specification document before programming commences. The extraction process will require specialist input from others unless expertise on the resource is available in-house. The specification document helps crystallise the researchers' needs and serves as a communication tool between the researchers and programmers; it also documents extraction procedures to support reproducibility. The document should include all decision rules, algorithms, assumptions and so on, taking into account that the same information can be stored in more than one area of the database or format. It may be difficult to anticipate all data idiosyncrasies and a heuristic process may be needed. For this reason, the programmer should be encouraged to annotate and modify the specification document with the goal of recording all decisions and assumptions in the document. This will allow the investigator to identify changes and deviation from the specification document and determine if the data had been handled as intended. Documenting changes and rationale will protect the investigator and institution during research compliance audits. More than one specification document may be needed for a variety of reasons including a complex study design such as the use of multiple resources and/or company policy or standard operating procedure.

3.2 **Extraction of the study population:** The specification should define extraction of the study population in terms of patient restrictions, inclusion and exclusion criteria, and those sub-sections of the database to be used. The study population often is extracted in more than one step. The initial extraction could be of everyone eligible for care in 2010 with a prescription for a glucose-lowering medication in that year. This would be followed by stepwise inclusion and exclusion based on other criteria completed on this more manageable subset. Alternatively, the extraction code is developed for a subset of the database and then repeatedly run on further subsets. This stepwise approach allows development of an attrition table or figure.

3.3 **Extraction of study variables:** The extraction procedure for each variable should specify exactly which database files and fields should be searched, how to identify the information (a code list, a free text string search, a flag for yes/no, algorithm, etc.) and any relevant dates. For many variables, a combination of fields may be included in the definition. For example, if body mass index is required, weight readings during pregnancy may be excluded, so definitions for weight, height and dates of pregnancy would be required.

DOI: 10.1002/pds

3.4 **Output:** At each stage of the extraction, the researcher should specify exactly what fields are required, and the output format should be established (e.g. character versus numeric, length of character fields and date format). If attrition numbers (i.e. reduction because of inclusion/exclusion criteria) are required, this output should also be established.

3.5 **Extraction from coded systems:** When data are stored in a coded classification system, a code list is required for every term included. If an established code list is used, it should be reviewed for applicability to the current study and database. The dictionary version used to generate the list must match that which formed the study population. Codes can be added and removed over time, and the meaning can change. For example, diagnosis-related group codes have changed substantially over time and an individual code may have a different meaning depending on the time of data capture. The code list should be linked to the appropriate step in the extraction procedure document. Details of any data manipulation (e.g. removal of leading spaces or zeros) required to import an established code list should be documented in the specification document.

3.6 **Extraction from non-coded systems:** The extraction strategy and associated assumptions (e.g. use of wildcards, SAS SPEDIS, negation, annotation, etc.) should be documented, including the source and list of terms. Details of any manual extraction (if necessary) should be documented in the specification document.

3.7 **Retrieval of additional data:** Where the population is to be supplemented with additional data, such as patient scores or information from non-electronic records, both the collection of data and linkage to the population dataset should be specified.

4. Privacy and security

Privacy concerns exist whenever identifiable information is collected and stored. The challenge is to share data for research while protecting personally identifiable information. In addition to adhering to all relevant policy and legislation, it is recommended that researchers access the most limited amount of identifiable information possible, house the data in a secure environment and perform scheduled review of policy and procedures.

4.1 **Compliance with data use policy:** The regulations protecting individual rights to data privacy vary around the world. It is the responsibility of researchers to be aware of, and compliant with, all relevant local, regional and national policies. These include, but are not limited to, rules for patient consent to use data, policy for the acquisition, storage, transmission (including international transmission) and destruction of protected data; and administrative requirements.

4.2 **Limited use of identifying information:** An important method for protecting the privacy of individuals is to limit the use of personally identifiable information. However, pharmacoepidemiological studies typically require the use of some of these data. An alternative is the use of a data set where direct identifiers, such as name, address and identification numbers, are removed or recoded to non-sense codes, whereas indirect identifiers, such as age and treatment dates, are used. Keeping the research data separate from direct patient identifiers and using the least amount of other identifying information are important methods for protecting sensitive data.

4.3 **Secure data storage and transfer:** When possible, sensitive data should be stored on a centrally managed server, with access rights restricted to those authorised to use the data. When server storage is not possible, data should be housed on local machines stored in a physically secure location requiring unique logon with strong passwords. Security logs should be generated and reviewed. As a general rule, protected health information should never be transmitted over the internet without IT security-approved encryption.

4.4 **Review of policy and procedures:** Scheduled reviews should be conducted to determine if all personnel and procedures are compliant with local, regional and national data security and privacy policy. An environment that supports the review of outstanding security risks should be encouraged. During the review process, researchers should pay special attention to situations that can increase the chances of exposing direct patient identifiers. These include, but are not limited to, the following: (i) use of a new resource or one not normally used for research; (ii) collection of

DOI: 10.1002/pds

additional patient or physician data; (iii) use of narrative data; (iv) analyses that extract small numbers (e.g. very rare diseases or very old patients); and (v) merging data from more than one source.

5. Quality and validation procedures

The pharmacoepidemiologist is responsible for ensuring that the data are of a sufficient quality to complete and interpret a particular analysis. Data quality is normally reviewed at two stages: when an appropriate database is selected for an investigation and after a study population has been created. At both stages, checks should consider accuracy and completeness of study variables at the source and data integrity after initial extraction. Where key study variables have not already been validated, this step should be included in the study protocol. In addition, all programming code should be verified.

5.1 **Overall database:** The completeness and accuracy of key study variables should be assessed when a resource is selected with consistency of data capture over time considered. This may need to be very specific, such as in a study where it is important to know whether patients smoke, but this variable may only have been recorded in more recent years. Quality testing results and methodology may already be available from either the database provider or the published literature. These quality checks should be consistent, transparent and available for scrutiny. In addition, databases often are dynamic and continually updated, so checking may need to be a continual process if analysts are not working with a static dataset.

5.2 **Study population:** Even if the entire multi-purpose database meets basic criteria for quality, it is possible that specific sub-populations or variables will have different characteristics that may impact data quality. For example, the measurement of 'height' will generally be more frequently measured and, therefore, recorded in children than in adults. Quality control of the study population gives the researcher an opportunity to validate study-specific variables and sub-populations, any algorithms, assumptions etc. All data extraction procedures, manual and automated, should also be assessed. Does everyone in the cohort meet the data extraction criteria? Is the study population size consistent with expectations, or could the extraction have been incomplete?

5.3 **Multiple resource studies:** When more than one dataset is used in a study, the quality of each should be assessed separately.

5.4 **Programming:** Programme review/testing should be completed at each stage of extraction and analysis. It is preferable to agree on a program specification and a test plan before programming is commenced. Testing can be performed on at least two levels: (i) the performance of the programming code/algorithm; and (ii) that the extracted data meet the inclusion/exclusion criteria using that programming code/algorithm. The test plan should include a list of all the programs to be tested and exactly how each test should be performed, for example, frequency and type of data used to run the test (active versus dummy). It is often best for the programme testing to be performed by someone other than the original programmer.

5.5 **Testing procedures:** Similar testing procedures are used to test the source database and study population:

5.5.1 **External validation:** It may be possible to validate database records against external documents such as clinical notes or death certificate registries.[10] Although this is expensive and time consuming, often necessitating a sample approach, it can be essential in certain settings as in claims data where 'working' rather than final diagnoses may be recorded. False negatives, as well as positives, should be considered. Where the electronic medical record *is* the source record, other validation methods should be used such as comparison of rates in the database with external figures, such as mortality statistics or meta-analyses from large clinical trials or review of the individual record for appropriate treatment/procedures. The positive predictive value of the algorithms/methods used to identify key outcomes can be assessed.

5.5.2 **Logical checks:** This includes checking if values in a field are plausible or range checking to look for outliers, missing values or unexpected occurrences. For example, has there been a check that no one appears to be aged over 140 year or have duplicates been reported? Is

DOI: 10.1002/pds

range checking performed at data capture? It is preferable to perform range checking at data capture so that errors can be corrected at the source. However, all systems are likely to have some errors, many of which may be mistakes in data entry, and some will be undetectable if they are clinically plausible.

5.5.3 **Internal validation:** The interdependence of variables within a case may be examined. Do various items of information contradict each other, or does one variable highlight an omission elsewhere such as when there is an administrative record of death but no cause of death captured?[11] It can be automatic and logical, such as testing that there has been no medical intervention after death. Often, simple logistic testing has to suffice as a proxy for data capture completeness, for example, where one would expect a series of tests or visits associated with certain conditions, or a diagnosis captured before prescription for a specific therapy.

5.5.4 **Consistency checks:** Simple consistency in data capture over time can also be measured by counting frequency of records or field occurrence to identify blocks of missing data or trends, or shifts in volume of recording.[12] As these may be dependent on external forces such as changes in data collection software, targeted motivation factors for the data recorders, and trends in disease/coding (e.g. there has been increased recording of autism since the 1990s), it is important to look at trends within sites as well as those over the whole database and between sites.

5.5.5 **Data extraction:** When testing that an extraction program works, it often is advisable to create an artificially engineered sample of complex patient data that include every permutation of the extraction: a 'dummy dataset'. When the extraction program is run over the dummy dataset, it can be assessed whether the known appropriate patients are extracted correctly. Once the data extraction program has been tested, the 'live data' can be used.

6. Documentation

Each stage of the database selection and use should be documented. The Guidelines for Good Pharmacoepidemiology Practices describes good practice in archiving and documentation including the provision of secure storage and indexing.[4] Guidelines on program coding also are published.[13] The following documentation is specifically relevant to good database use but should follow the general guidance:

- Privacy policy to inform users at every stage how data should be handled and protected, including data use transfer agreements, monitoring and control of access to information and computing systems (passwords, network, firewalls, encryption and intrusion detection).
- Specification of each phase of data extraction and output, including the data source version and a list of all codes used to define treatment, outcome and covariates, and extraction strategy for non-coded systems. These specifications would also include any formal study protocols and post-project implementation protocol and analysis amendments with noted rationale for their addition.
- If multiple resources are used, each additional stage, such as compatibility checks and matching, should be documented.
- Complete population attrition numbers including each step of cohort eligibility during study database construction and each step of analysis.
- All aspects of quality assessment.
- Annotated programming code for data extraction and manipulation.
- End-user acceptance when data are prepared for a third party.
- Availability and access to the final research database.

CONFLICT OF INTEREST

The authors declare no conflict of interest. This is an ISPE policy. As such, the members and board have reviewed and approved the document. This work has been submitted by ISPE. All authors are participating ISPE members, and the views expressed do not necessarily represent positions of their government, institution or corporation.

ACKNOWLEDGEMENTS

The authors would like to thank Andrew Bate, Karin Benoit, Alexander Cole, Liang Huifang and Nigel Rawson, for their review of the draft guidance, and Corey Parker and Valerie Wiltshire for their help with the manuscript.

DOI: 10.1002/pds

REFERENCES

1. Motheral B, Brooks J, Clark MA, Crown WH, Davey P, Hutchins D, et al. A checklist for retrospective database studies--report of the ISPOR Task Force on Retrospective Databases. *Value Health* 2003; **6**(2): 90–7.
2. Gliklich RE, Dreyer NA eds. Registries for Evaluating Patient Outcomes: A User's Guide. April 2007: http://www.effectivehealthcare.ahrq.gov/repFiles/PatOutcomes.pdf.
3. Berger ML, Mamdani M, Atkins D, Johnson ML. 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.
4. Guidelines for good pharmacoepidemiology practices (GPP). *Pharmacoepidemiol Drug Saf* 2008; **17**(2): 200–8.
5. Elm Ev, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies 10.1136/bmj.39335.541782.AD. *BMJ* 2007; **335**(7624): 806–08.
6. ISPOR. (2010). "International Digest of Databases." Retrieved June 2010, from http://www.ispor.org/DigestOfIntDB/CountryList.aspx.
7. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care* 2010; **48**(6 Suppl): S45-51.
8. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf* 2010; Aug;**19**(8): 848–57.
9. Brown JS, Lane K, Moore K, Platt R. Defining and Evaluating Possible Database Models to Implement the FDA Sentinel Initiative. Report to the U.S. Food and Drug Administration, Contract No. HHSF223200831315P; May 2009. From http://www.regulations.gov/search/Regs/home.html#documentDetail?R=090000648098c282
10. Herrett E, Thomas SL, Schoonen WM, *et al*. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* 2010; **69**: 4–14.
11. Hall GC. Validation of death and suicide recording on the THIN UK primary care database. *Pharmacoepidemiol Drug Saf* 2009; **18**(2): 120–31.
12. Hennessy S, Bilker WB, Weber A, Strom BL. Descriptive analyses of the integrity of a US Medicaid claims database. *Pharmacoepidemiol Drug Saf* 2003; **12**(2): 103–11.
13. Winn TJ. Guidelines for coding of SAS® programs. In Proceedings of the Twenty-ninth Annual SAS® Users Group International Conference. SAS Institute Inc.: Cary, NC and Montréal, Canada, Paper 258–29 2004.

DOI: 10.1002/pds