LEADING ARTICLE

Drug Safety 2004; 27 (12): 857-870
0114-5916/04/0012-0857/$31.00/0

Data Quality Management in Pharmacovigilance

Marie Lindquist

Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring, Uppsala, Sweden

Abstract

Pharmacovigilance relies on information gathered from the collection of individual case safety reports and other pharmacoepidemiological data. Even given the inherent limitations of spontaneous reports, the usefulness of this data source can be improved with good data quality management. Although under-reporting cannot be remedied this way, the negative impact of incomplete reports, which is another serious problem in pharmacovigilance, can be reduced.

Quality management consists of quality planning, quality control, quality assurance and quality improvements.

The pharmacovigilance data processing cycle starts with data collection and, in computerised systems, data entry; the next step is data storage and maintenance; followed by data selection, retrieval and manipulation. The resulting data output is analysed and assessed. Finally, conclusions are drawn and decisions made. The increased knowledge feeds back into the data processing cycle.

Focussing on the first three steps of the data processing cycle, the different quality dimensions associated with these steps are described in this review, together with examples relevant to pharmacovigilance data.

Functioning, well documented, and transparent quality management systems will benefit not only those involved in data collection, management and output production, but, ultimately, also the pharmacovigilance end users, the patients.

In pharmacovigilance, making the right decisions at the right time is critical. As in all risk assessment, a judgement has to be made based on available information. Whereas we have a good picture of a medicine's efficacy, based on carefully planned and controlled clinical trials, our knowledge of the negative aspects of therapy is limited, particularly for new drugs. Clinical trials give us some information, but only the most commonly occurring adverse reactions are likely to be known before a medicine is released onto the market, often for use in patient categories that differ from those included in the trials.[1]

The WHO definition of pharmacovigilance is "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible problems related to medicinal products".[2] In the early years, there were no systems in place for the routine collection of relevant data on all patients exposed to medicines, therefore pharmacovigilance decisions had to rely on post-marketing reporting of suspected adverse drug reactions (ADRs). Data collected in this way in most instances give no evidence of causality, nor can such data be easily quantified to estimate the level of the risk posed.[3] Over the last decades, complementary pharmacoepidemiological methods have been introduced, allowing for hypothesis-testing and incidence estimates. Prescription-event monitoring systems and longitudinal healthcare in-

formation database systems (registries) can be used both for signal detection and follow up, albeit in defined, relatively small populations. For the detection of new adverse reaction problems, particularly those that occur rarely, post-marketing reports by alert individuals remain a main source of information.[3]

Even given the limitations of this data source, the value of reports of clinical concerns should not be underestimated – after all, it is from such data collected over the years that we have learned most of what we know about the negative aspects of therapy involving medicines.[4] However, for any data to be meaningful in terms of leading to better information, increased knowledge, and sound decision making, the whole process from collection to retrieval and analysis must follow rigourous procedures to try to ensure the highest possible quality of data, and of its management, at every step in the process.

This paper gives an overview of data quality management in pharmacovigilance for those readers who are not primarily concerned with day-to-day data management, but it also gives some practical advice of interest to those who are in charge of the design, implementation and maintenance of database systems.

Data quality management is well described and discussed in the information technology and quality management literature.[5-10] The need for quality management in pharmacovigilance and pharmacoepidemiology has already been identified.[11-17] This paper puts general quality issues identified by other authors into a more detailed pharmacovigilance context.

1. Data in the Pharmacovigilance Process

Pharmacovigilance knowledge is built by the accumulation of individual case safety reports and their subsequent analysis, together with the results of further pharmacoepidemiological studies. Increased knowledge in turn influences reporting, which leads to new data and better information.[18] Access to up-to-date and accurate information is crucial – without it the process is slowed down, or at worst, halted.

As shown in figure 1, data processing can be seen as a series of loops; linked together in a chain. It is a cyclical process, with an input and an output side. When data is transferred between systems, a new, parallel cycle is started. Although much of the data processing today is done with the help of computerised database management systems (DBMSs), it should be pointed out that the term 'data processing' for the purpose of this paper denotes the wider concept of handling, analysing and interpreting facts or information; the data provides the information content, whether residing in a database or not.

Data processing in pharmacovigilance is complex; the process involves many players, from different settings and with different interests. This complexity is further accentuated on the international level. The steps of the data processing involve people with different roles, who can be identified as:

- data producers, generating the data which starts the processing cycle;
- data collectors, assembling original data and turning it into structured information;
- data custodians, responsible for data storage, and for maintenance and security of database systems;
- data output producers, selecting, retrieving and manipulating data to create data outputs; and
- data consumers, utilising and analysing data outputs.

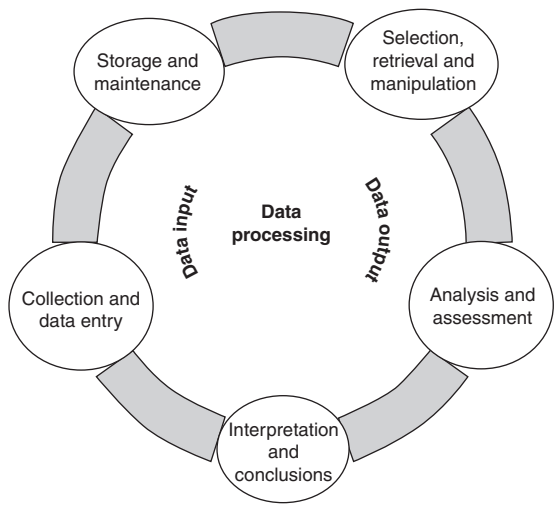

Fig. 1. The data processing cycle.

Table I gives an overview of the different steps in the pharmacovigilance process, the players involved, the data processing roles and the relevant data processing issues. As can be seen in the table, the overlap between these roles is often considerable, not only within one organisation or group, but also between these.

Whilst data analysis and interpretation are vital steps, these are well described in standard textbooks on pharmacoepidemiology[19] and pharmacovigilance,[20] and in the existing pharmacovigilance literature.[21-24] Therefore, this paper will focus on the first three steps of the data processing cycle.

2. Quality Management Concepts and Definitions

The aim of a quality management system is to build quality into any product or process, through planned and systematic actions. When set up and managed properly, such a system provides confidence that the result of a process (the product), or the process itself, perform according to the defined quality criteria.

The International Organization for Standardization (ISO) 9000 identifies four fundamental parts of a quality management system (see table II for definitions of terms):[25]

- quality planning
- quality control
- quality assurance
- quality improvement.

When a quality management system is designed and implemented, the *quality planning* is the first phase. To start with, the goals must be defined – what are the quality criteria that must be met, and how should they be achieved? Next, one needs to determine the resource requirements, both in terms of personnel and equipment; and decide what processes should be set up and implemented in order to achieve the objectives.

Quality control and quality assurance are two distinctive, but related parts of the quality management system. A *quality control* programme provides the means to monitor and check the results of a process, or the characteristics of a product, against the preset requirements. The aim of the quality control activities is goal fulfilment – does the product/process meet the quality criteria? – whereas *quality assurance* aims at ensuring the actual effectiveness of the quality control programme. This is done through a continuing evaluation of the quality control activities and results, and by the implementation of corrective measures where necessary.

The fourth cornerstone of quality management is about striving for excellent quality by continuous *quality improvements*. Most processes are not rigid and unchangeable, but part of a dynamic system, and the notion of a 'steady state', where the quality of a process or product can be deemed perfect once and for all, is rarely achievable. No matter how good a quality control programme is, new and unanticipated problems cannot always be avoided, even in the absence of apparent changes that might influence, for example, a production process. However, with appropriate feed back, both effectiveness and efficiency of the quality system can be improved. This is achieved by ongoing critical appraisal and the ability to learn from mistakes. Failure to do so suggests complacency and arrogance, whereas quality improvements can lead to competitive advantages.[26]

An obvious reason for good quality management is cost reduction: the extra costs involved in quality control and quality assurance should be weighed against the direct and indirect costs of not having an adequate quality management system in place, e.g. the cost of the recall, repair or rework of faulty products as compared with a product that works satisfactorily.

Another argument is the effect of poor quality on trust – a flawed process or faulty product may result in anything from irritation to a complete lack of confidence in the process or product concerned, sometimes compounded by perceived or real harm caused.[21,26] This in turn may lead to increased costs. A company may have to pay compensation to dissatisfied or injured customers, or launch expensive promotion campaigns to regain lost trust; poor quality of public sector services will lead to extra costs for society as a whole, and so on.

Database systems or data warehouses store facts, or information, which can be communicated and lead to increased knowledge. Data retrieved from databases form the basis for analyses, planning and decisions in many critical areas of modern society.

Table I. The pharmacovigilance process – activities, players involved and type of data processing

The players	Data processing role	Data processing issue
Diagnosis of a suspected adverse drug reaction		
Prescribers, patients and other health professionals involved in patient care	Data producer	Data analysis and interpretation
Filling out a case report form or otherwise making a notification		
Prescriber or other health professional	Data producer/data collector	Collection and data entry
Filing report locally (e.g. in hospital or in pharmaceutical company database)		
Report originator or appointed pharmacovigilance responsible person	Data collector/data custodian	Collection and data entry Data storage and maintenance
Sending report to a regional pharmacovigilance centre		
Report originator or appointed pharmacovigilance responsible person	Data collector	Transfer between systems
Sending report to a national pharmacovigilance centre		
Report originator or appointed pharmacovigilance responsible person	Data collector	Transfer between systems
Collecting and entering report into a national pharmacovigilance system		
National pharmacovigilance centre	Data collector/data custodian	Data collection and entry Data storage and maintenance
Collecting and entering report into a regional pharmacovigilance system		
Regional pharmacovigilance body (e.g. European Medicines Evaluation Agency [EMEA])	Data collector/data custodian	Data collection and entry Data storage and maintenance
Collecting and entering report in the WHO international database		
Uppsala Monitoring Centre (UMC)	Data collector/data custodian	Data collection and entry Data storage and maintenance
Screening of collected data		
Pharmaceutical company, national pharmacovigilance centre, regional body, UMC	Data output producer	Transfer between systems Data selection, retrieval and manipulation
Preliminary analysis of evidence		
Pharmaceutical company, national pharmacovigilance centre, regulatory authority, regional body, UMC	Data customer Data output producer	Transfer between systems Data analysis and assessment Data interpretation and conclusions
Further studies		
Pharmaceutical company, national pharmacovigilance centre, regulatory authority, regional body, UMC, academia	Data customer Data output producer	Transfer between systems Data analysis and assessment Data interpretation and conclusions
Analysis of evidence from study		
Pharmaceutical company, national pharmacovigilance centre, regulatory authority, regional body, UMC, academia	Data customer Data output producer	Transfer between systems Data analysis and assessment Data interpretation and conclusions
Effectiveness–risk assessment		
Pharmaceutical company, national pharmacovigilance centre, regulatory authority, regional body, UMC, academia	Data customer	Transfer between systems Data analysis and assessment Data interpretation and conclusions

Continued next page

Table I. Contd

The players	Data processing role	Data processing issue
Changes in regulatory status		
Regulatory authority and/or pharmaceutical company	Data customer	Transfer between systems Data interpretation and conclusions
Follow-up and impact studies based on new knowledge		
Regulatory authority, pharmaceutical company, regional body, UMC, academia	Data customer	Transfer between systems Data analysis and assessment Data interpretation and conclusions
Individual benefit–harm assessment		
Prescriber, patient	Data customer	Data analysis and assessment Data interpretation and conclusions

Poor quality data therefore may have wide ranging impact, and may cause great harm to individuals and societies. Furthermore, once a database system fails to maintain good quality data, re-engineering is difficult and very costly.

The following three examples by English[27] illustrate different effects relating to poor data quality:

- A manufacturing company considered scrapping a $12 million data warehouse project due to inconsistently defined product data and poor data quality.
- A publishing company spent 4 years cleansing customer records that were being sourced from 12 disparate systems.
- Analysis of medical records from an insurance company's data warehouse identified an unusually high incidence of 'haemorrhoid' diagnosis in one particular region. It transpired that the region's claims manager confessed that this diagnosis was internally used as a term to identify claimants "that were a pain in the ass!"

Information technology is particularly affected by changes. New software versions are released frequently, offering better functionality; as a consequence of an upgrade of software, a corresponding upgrade in the hardware on which the software applications run is often necessary.

Pharmacovigilance relies heavily on data, and information technology solutions for storage and retrieval of the data.[12] The above arguments should convince the reader of the necessity of good quality management in this area.

3. Collection and Data Entry

Once a decision to report an ADR has been made, the relevant data items need to be recorded. Even in the information technology age this first step is still often paper based – an ADR reporting form is filled out with patient and reaction details, in pre-allocated spaces. Such data recorded on a paper form is later the basis for data entry into a computerised system, where such a system exists. In a local setting, good pharmacovigilance practices do not necessarily include computerisation; however, as larger amounts

Table II. Concepts relating to quality management – International Organization for Standardization (ISO) definitions of terms

Quality	Degree to which a set of characteristics fulfils requirements
Quality management	Coordinated activities to direct and control an organisation with regard to quality
Management system	System to establish policy and objectives and to achieve those objectives
Quality planning	Part of quality management focused on setting quality objectives and specifying necessary operational processes and resources to fulfil the quality objectives
Quality control	Part of quality management focused on fulfilling quality requirements
Quality assurance	Part of quality management focused on providing confidence that quality requirements will be fulfilled
Quality improvement	Part of quality management focused on increasing the ability to fulfil quality requirements
Effectiveness	Extent to which planned activities are realised and planned results achieved
Efficiency	Relationship between result achieved and resources used

of data are accumulated the analysis becomes increasingly difficult. In most countries, and certainly on the international level, computer-based pharmacovigilance is essential.[12]

Data collection and capture in a database management system involve creation, update and transformation of information. The implications of these processes on data quality will be discussed under the following headings:

- the data model
- the data entry process
- the data values.

3.1 The Data Model

The data model is a fundamental part of the design of a database system. It is a conceptual view of the structure and organisation of a database system, showing the tables, their content, and how they are linked. Quality dimensions of the data model discussed below relate to its *scope*, *consistency*, and *level of detail*.

3.1.1 Scope

First of all, the data model must be *relevant* for its purpose, as defined by domain and user needs. Secondly, the data model must be *comprehensive*, i.e. able to cater for current needs in terms of what information can be recorded, but should exclude excessive or redundant information. Ideally, it should also be *realistic* in terms of accessibility of the information to be captured. However, as regards the latter, there needs to be a degree of flexibility to accommodate anticipated future changes, so that frequent redesigns and the ensuing costly and time-consuming data conversion can be avoided. The ability to accommodate changes in a database without having to change the data model is a sign of *robustness*.

This was the reasoning behind the comprehensive listing of data items included in the report of the CIOMS Ia working group, convened to discuss which data fields could be considered for international exchange of individual case safety reports.[28] The subsequently developed International Conference on Harmonization (ICH) E2b guidelines are almost identical as to their coverage of data fields.[29] Due to the penetration of the ICH recommendations, not only in the countries that are directly involved in the ICH process (EU members, Japan, and the US), it is likely that more and more pharmacovigilance databases, when designed, will be based on the ICH fields.

It would, however, be a mistake to think that databases moulded on this 'ideal' data set will often, if ever, be completely populated with information (see also section 3.3.2). Surveys of the completeness of information in the WHO adverse reactions database (Vigibase) have shown that even key fields are often not filled in.[3] These include medication and ADR onset dates, indication for treatment and patient outcome.

3.1.2 Consistency

The consistency of a data model refers both to its contents and the nomenclature used. A data model must be clear, with well defined concepts. Overlapping or otherwise ambiguous data fields must be avoided. This is important both for data entry and retrieval. Data fields for identical or related concepts should be named consistently throughout a database. For example, the concepts 'date recorded' or 'date changed' commonly occur as data fields in several tables in a database; the same applies to concepts such as 'medicinal product' or 'gender', for which there are other possible designations (e.g. 'drug' and 'sex', respectively) – one should be chosen and all instances of this concept in the data model should have the same designation. Also, subordinate concepts should be labelled consistently with their parent concepts. Thus, an 'organisation' table should not contain the data fields 'company name' and 'company address' – but 'organisation name' and 'organisation address'.

3.1.3 Level of Detail

The requirements of a particular domain should decide the level of detail allowed for in a data model. For instance, a database system for recording and forecasting weather in a country with a temperate climate will need many different values for 'snow' attributes, whereas a corresponding data model in a tropical country very reasonably need only include one 'snow' attribute, if at all.

When defining data fields, one must avoid aggregated information in one field. For instance, dosage information like '200mg daily' should NOT be recorded in one field, since it contains three different

attributes, with a different set of permissible values for each. The aggregation, when needed, can easily be made at the output level.

A method for achieving flexible levels of granularity (the degree to which an attribute can be described) is to use hierarchical classifications.[12] A single data field can be used to capture information *to the level of precision available*. By linking the recorded data to the corresponding classification, the data can be aggregated and analysed at different levels of precision. In pharmacovigilance such classifications are used i.e. for the recording of medical and medicinal product information. For instance, the WHO Drug Dictionary[30] allows for many levels of precision:

- Anatomic, Therapeutic, Chemical (ATC) classification level, denoting the main indication for which a medicinal product is used; the ATC is in itself a hierarchy, with five levels;
- ingredient/combination of ingredients level;
- pharmaceutical product level (combination of ingredients/form/strength); and
- medicinal product level (referring to the named product marketed and sold in a particular country).

Similarly, medical classifications such as the WHO Adverse Reaction Terminology (WHO-ART), International Classification of Diseases (ICD) and the Medical Dictionary for Regulatory Activities (MedDRA)[31-33] allow for groupings and aggregation of data on different levels, from broad system-organ classes to individual signs and symptoms.

3.2 The Data Entry Process

Data entry is the process by which data is captured by a computerised system, e.g. a database. Data entry can be done manually: a person types characters on a keyboard directly into data fields in a database table; or by an automated procedure by which a set of already computerised data is entered into the appropriate data fields. The quality of data entry can be measured by the degree to which the recorded data is representative of the original information, when retrieved and presented as data output.

3.2.1 Data Entry Through a User Interface

The more user friendly a data recording interface is, the more likely it is that high quality can be achieved as a result of the data entry process. An interface is a computer application that presents the data fields, usually in a form view, together with menus, toolbars, field labels and guiding text/instructions for the user. Characters typed by the user are transferred into the data field in the database table to which the interface is linked.

To optimise data entry, it is essential to have user input into the design of a data entry interface, since the *layout* of the interface, and its *functionality*, influence the accuracy and speed with which data can be entered. For example, if the end user prefers to use the keyboard instead of the mouse (an inclination often ignored by our colleague computer programmers), the interface must allow for keyboard shortcuts to move around the screen and to execute commands. Problems with bad layout can range from the use of a colour that irritates the user, or a cluttered screen which is not user friendly, to incorrect entries being caused e.g. by confusingly or wrongly labelled data fields.

The data quality is also influenced by the *degree of judgement* which the user has to apply when recording the information. A typist can be perfectly adequate for transferring exact copies of values from a form onto a computer, whereas this might not be the case when the data entry requires transformation of the original information. In this instance, the quality of the recorded information is directly related to the qualifications and skills of the person who enters the data. For example, the translation of a diagnosis, written in free text, to a relevant adverse reaction term or a set of terms, chosen from a controlled vocabulary or a classification linked to the data field, should be done by a person with adequate medical training. The use of so-called auto-encoders has become more widespread in recent years. Although these may facilitate data entry, by highlighting acceptable values in free text, and presenting suggestions for terms based on e.g. lexical or medical equivalence, it must be stressed that they do require human judgement in their application. An example is an auto-encoder identifying medical terms from a case history; the conditions leading to treatment would be identified along with

the adverse reaction descriptors as valid terms in the medical classification used. The user must decide where these different types of information should go in the database.

3.2.2 Batch Data Entry

When a set of data is entered without direct human-computer interaction there is no need for a data entry interface. On the other hand, much thought has to be given to setting up a process which delivers the right result without human intervention. This process must, like data entry using an interface, involve procedures for checking the data values entered. This is relatively straight-forward provided that no transformation requiring human judgement is needed. When this is the case, the data entry process must include a stage at which human intervention takes place, so that data can be checked and edited, or entered manually.

3.3 Data Values

The quality of data recorded can to a large extent be improved by appropriate validity and coherence checks. An interface application allows for checks to be made at the point of data entry. These checks are executed through programming code in the interface.

A database usually resides on a server which is separate from the computer running the interface. Recent versions of DBMSs allow checks to be run from within the database itself. By storing programming logic in the server database, network traffic is minimised, and speed increased. In addition, the integrity of the database can be protected by statements that are run whenever a user attempts to modify data in a table. Such functions built in to the database can prevent erroneous, inconsistent or unauthorised data changes, and not only be executed when data is accessed through a programmed user interface.

It is for the above reasons recommended to move programming logic relating to the quality of the *data* from the interface application to the database, whereas the design of the data entry interface should concentrate on the quality of the *data entry process*. The argument is the same for batch data entry.

3.3.1 Accuracy

The accuracy of data values can be divided into *syntactic* and *semantic* accuracy. Misspellings or alternative spellings/representations of data values are examples of syntactic inaccuracy; although a data value may be correct in the sense that it conforms to a true value, as given in the data source, it may give rise to inaccurate, or not intended, computer interpretation if it is not a permissible value according to the specification for the domain. Semantic inaccuracy, on the other hand, occurs when a permissible value has been recorded, but which in this case does not correctly represent the information given in the data source.

A human can see that ‘benzodiazepines’, ‘benzodiazepine’ and ‘BZDs’ mean the same thing, but if ‘benzodiazepine’ is the only permissible value, the other two variants are examples of syntactic inaccuracies. Thus, if the assumption is made that all benzodiazepine reports are recorded as ‘benzodiazepine’, a count of the number of such reports would not yield the correct result, since not all existing reports would be included. Similarly, a duplicate check on adverse reaction reports would not recognise two reports as being identical if the format used for report identity numbering has changed from ‘000000123’ to ‘123’.

Syntactic accuracy can be obtained using controlled vocabularies: entered values are compared and checked against reference dictionaries (lists of permissible data values). These lists can be incorporated into a database system as *lexicon tables*, containing only predefined, allowed values, expressed as formatted text or codes. When linked to a field in a database table, the lexicon table ensures that a value entered in that field matches an existing value in the lexicon table. An associated look-up function facilitates data entry, in that the user can select the value from e.g. a drop-down box on a data entry screen, rather than typing in the text.

There is, however, also a danger in limiting the choice to predetermined values. Unless one is absolutely sure that all possible useful values have been identified when the lexicon table is created, it should always contain an additional value ‘other’. The direct use of this value is to indicate that another value than any of those listed would have been chosen, had it been included. This can be an important

pointer to an omission made in the lexicon table. A suggestion is to include a prompt in the data entry procedure for the user to add information in a separate text field whenever the choice 'other' has been made. The entries made into this extra field can then be analysed regularly to identify new values that are entered frequently. These can be added to the lexicon table, making it more comprehensive.

Classifications are helpful, not only in allowing variable granularity (as described above in section 3.1), but also in achieving syntactic accuracy. Unlike lexicon tables, which are linked to one data field, a classification may be linked to one or many fields. For instance, by selecting an ID number for a particular medicinal product, all corresponding information such as active ingredients, product name, marketing authorisation holder etc, can be filled in automatically into separate fields in an individual case safety report. (This is useful not only for data accuracy, but also for data entry speed.)

An additional advantage of controlled vocabularies is that they allow for language independence of both data entry and presentation. If the data in a field is stored in form of a code value, corresponding text values can be displayed in different languages from linked lexicon tables or classifications.

Something to consider, whenever controlled vocabularies are used, is the addition of a corresponding free text field to capture the whole information. This was out of the question in the early days of computerisation; now hard disk space is not a big issue. However, one should be aware that the analysis of free text fields is problematic since it is not amenable to counting, nor readily aggregated. This is supported by recent experience from exchange of individual case safety reports (ICSRs) in the ICH E2b format, which has made it clear that there is a need for controlled vocabularies for a number of ICH E2b data fields, where today free text is allowed.

In contrast to syntactic accuracy, semantic accuracy cannot readily be audited. However, if the same data instance is recorded in several data sources, a check can be made comparing the data values in the different sources. Here, a warning is in its place: when doing these comparisons one must remember that *concordance* does not necessarily equal *correctness*. This is particularly true if the data value is the result of an assessment made, rather than an objectively measurable fact.

3.3.2 Completeness

A major problem in pharmacovigilance is the level of under-reporting, which has been suggested as being as high as 95%.[34-37] It follows that the information which is captured in a reporting system only represents a minority of the adverse events actually occurring. The effects of under-reporting and how it could be remedied, although a critical issue to pharmacovigilance as a whole, are however outside the scope of this paper, the purpose of which is to describe the management of *available* data, whatever its shortcomings.

Apart from under-reporting, lack of completeness of received reports is a prevailing problem in pharmacovigilance. The impact of this can to some extent be reduced by good data quality management practices. These include routine production of statistics on data content and feed-back to those providing the data; problems with incomplete reports can be addressed, and sometimes solved, if the completion of key data items and changes over time are monitored regularly.

In a database system, as new data are entered, it is recommended that a 'marker' of completeness is generated. As an example, the WHO database includes a field 'documentation grading' which scores each adverse reaction report according to a predefined algorithm based on the completeness of information in key data fields. The documentation grading is used to identify problems of missing data in reports received from national pharmacovigilance centres. When investigating why there was a sudden change in the overall documentation grading scores from one country it was revealed that the 'indication' field was left empty. This was due to a change in the national database system, which had not been reflected in the data export files to the WHO database. The situation was quickly remedied.

When assessing completeness, one must differentiate between *mandatory* and *optional* data fields. For mandatory fields a value must be recorded. The permissible values must exclude 'unknown' or 'not applicable'. For optional data fields, on the other hand, these values are not only permissible, but should *always* be included as a possible choice.

Otherwise, when confronted with a blank field, it is impossible to know if the information was not available, not relevant in this particular case, or simply not filled in. If it is immediately clear that a certain piece of information is not available, a time consuming and fruitless inquiry to the data provider can be saved. On the other hand, such an investigation could be time well spent if there is an obvious likelihood that the information might be available from the original source, although not submitted.

Much of the problem of interpreting missing data could be avoided if database systems are designed in a way that eliminates the possibility of leaving optional fields blank. The database solution, however, postulates that the appropriate information is available from the data source used. This is unfortunately not often the case in pharmacovigilance, where much of the original information comes from manually filled in paper forms. Therefore, more pragmatic solutions are necessary, in which data is not rejected simply because a certain field is not filled in. It is still important, though, to always include 'unknown' and 'not applicable' as possible choices in optional fields, so that this information, when available, is not lost.

3.3.3 Consistency

Data inconsistencies occur when values in two or more data fields are in conflict. By introducing logical rules, many of these inconsistencies can be avoided. Examples in a pharmacovigilance database include date fields related to the start and stop of medicine treatment, or the adverse reaction. A logical constraint should be introduced which only allows a 'stop' date if it succeeds the corresponding 'start' date. Another example is a field 'duration of pregnancy' which is applicable only to females, and, in addition, only to a subset of pregnant females. A constraint could be introduced, only allowing a numerical value in the 'duration of pregnancy' field if: (i) the field 'gender' equals 'female'; and (ii) the field 'pregnant' equals 'yes'.

3.3.4 Currency

Data currency (the 'age' of a data value) is of particular importance for those data fields that involve information that may change over time. A measure of data currency can be made if each record, when entered, is assigned a 'time-stamp'. Another aspect of data currency, which is relevant to pharmacovigilance, involves medical terminologies and drug classifications, which are regularly updated. Whenever a code value is recorded from any of these types of sources, a corresponding field 'source version' should be filled in, to avoid possibly incorrect translations from codes to text which could occur if the wrong version of the source is used for reference.

4. Data Storage and Maintenance

Once data has been entered into a database, using appropriate quality control and quality assurance measures, this could be assumed to be a static system, in which nothing can change. However, maintaining the quality of data that has been stored poses its own challenges. In the first place, one must secure the data against partial, or complete loss. Secondly, the integrity of the data must be protected.

4.1 Protection Against Data Loss

There are many stories about data lost through accidents, thefts, or carelessness. In an instant the results of years of work can be wiped out, unless copies have been made, and stored elsewhere. Therefore at the quality planning stage it is important to make sure that the database system is protected against data loss.

In addition to protecting the data itself, the programme code and instructions and documentation must also be protected against loss. The frequency with which back-ups are made should be determined by the rate of change of the information concerned. A data back-up might be needed on a daily basis, or more frequently, whereas programme code and documents need only be backed-up once, until a change is made. Data back-ups are often separated into complete, and incremental, back-ups. First, a complete back-up is made, then, for a time, incremental back-ups including only added or changed data are made, until a new complete back-up is made. The advantage of incremental back-ups is that they take less time, but that has to be weighed against the risk of not having specified the back-up correctly to include all modifications.

Back-ups should be stored on secure media (need one say that it is no point in making a back-up, if one

is later to find that the information cannot be retrieved?); in a separate location from the original; and in a place that is reasonably protected against fire and theft.

4.2 Protection of Data Integrity

An important function of a database manager is to see to that a database is protected against intrusion by outsiders. Skilled hackers can get into almost any system, so there is no absolute protection against this kind of theft or data manipulation, but a so-called *firewall* is a cost-effective investment to avoid such violations.[38] A firewall can be seen as an impenetrable wall, in which openings are made that only allow for traffic in one direction, from inside-out (the same function as a one-way valve). It is possible to make openings which allow two-way traffic, but only from a predefined address (a computer IP number) to a predefined address inside the firewall.

Another security aspect relates to internal access to data in a multi-user environment. Although database management systems have built in password protection features, these are not always activated, or applied optimally. Passwords should be set for the database as a whole, with specified table rights and field rights to provide different levels of access privileges for different users in a group. For instance, passwords can be used to allow a particular user to view data but not change it; another user may update a record, but not delete it.

5. Data Selection, Retrieval and Manipulation

The production of useful data output involves the transformation of raw data into a refined representation, which should remain truthful to the source of information, and be appropriate for analysis. When the data resides in a DBMS, the process includes the definition and selection of a subset of the total database and the retrieval of this data. This is usually followed by additional aggregation, the production of summary reports, and other manipulations needed to produce the desired output. Achieving this requires detailed knowledge of the characteristics of the data set, and a thorough understanding of the strengths and weaknesses of the recorded information.

The common technique for selection, retrieval and aggregation of data in a database involves the use of so-called 'query languages'. Query commands can be executed by direct access to the appropriate tables, or through specially designed search interfaces. The former method gives full flexibility, but requires advanced computer skills. Search interfaces, on the other hand, offer search capacities to less experienced data customers, but usually with a limited freedom. Whichever method is used, the result must always be examined with prudence, since mistakes in the definition and execution of database queries are not always immediately apparent.

For example, a search linking two tables which contain different kinds of patient information, with a common field 'patient ID' as the link field, will exclude patient records from table A for which there is no corresponding 'patient ID' in table B, unless the search specifies an inclusive link.

In recent years, automated methods for screening of pharmacovigilance data have been developed, using data mining techniques together with statistical measures for the identification of disproportionate reporting frequencies.[39-43] These methods have been shown to be useful in signal detection, in that they provide added value as compared with standard data outputs.[44] With these methods, there is always a risk of aimless searching for dependencies and patterns until 'something interesting' is found ('data-dredging').[13] By using predefined algorithms, and critical appraisal, this risk can be reduced. However, as described by Orre[45] the relative complexity of data mining techniques warrant caution, to avoid mis- or over-interpretation of results, e.g. making causal attributions based on statistical correlations in a data set.

The quality of data output can be judged by its *relevance* for its purpose, its *accuracy*, *precision*, *completeness* and *timeliness*. The first three are influenced by the data extraction method; the latter also by the data itself (the data currency).

5.1 Relevance

The relevance of data output is determined by its ability to provide adequate information to address a particular problem. In the definition of the output, it is essential to consider not only what *data items* to include, but also the appropriate *level of specificity*. A selection that is too narrow might not provide the full picture of the issue in question, and can therefore be misleading, even though the information included is correct.

A solution is to assemble data outputs at different levels of specificity. This allows for both the broader overview, as well as an in-depth analysis. Such aggregation of pharmacovigilance data is supported by hierarchical classifications for medical terms and products, although the proper use of these require intimate knowledge of their structure and content.[46]

5.2 Accuracy

Accuracy refers to the degree to which the data output correctly represents the source information. When information is derived from a database, the accuracy of data output should be measured, and checked, against the actual database content – this is imperative. If the data extraction is made through a search interface, the user can rely on accuracy of the result, but this presumes that the appropriate testing and validation has been made before the implementation of the search application/software. If ad hoc querying is used, it is wise to produce, and test, a set of standard query scripts that can be combined, and edited. For instance, the likelihood of achieving a correct result is higher if one only has to exchange a medical term for another in an otherwise identical search script, rather than rewriting the whole script.

5.3 Precision

Precision is related to, but not synonymous with accuracy. Lack of precision can be a problem with any data, whereas in data outputs, the problem is often the reverse; the presented output displays a level of precision with cannot be supported by the data. This can lead to misinterpretations and must be avoided. A simple rule for numerical output is to not include more decimals in the result than that of any of the ingoing values. Particularly when contrasting values with an inherent uncertainty there is an apparent risk of over-interpretation of a small difference between values if the variance is not taken into account.

5.4 Completeness

A data output is complete if it includes all the data items and data values that have been defined in the selection. Like accuracy, it should be checked against the data source from which it was retrieved. In this context, a distinction must be made between *missing data* due to data being missed in the search or lost in the processing; and *non-existing data*. For example, a cross-table showing number of reports for drugs and adverse reactions, listing all drugs on one axis and all reactions on the other, will show a blank, or zero, for the majority of drug-ADR combinations, since there simply are no reports stored for most of them.

5.5 Timeliness

The timeliness of data output relates to it containing up to date information, and being available on time. Both are critical quality aspects in pharmacovigilance. Some delays are outside the control of data custodians and data producers, but time efficient data processing is certainly not.

6. Discussion and Conclusions

There are a number of data quality dimensions involving data processing and its results. The maintenance of good quality must be upheld throughout the processing cycle, and in the computerised systems employed as supporting platforms. Although a truism, the statement ‘no chain is stronger than its weakest link’ is clearly valid. A problem at one level may not only be perpetuated, but exacerbated, at a later stage. For instance, if there is duplication of information at the input side, so that three individual case safety reports will be stored and counted as six, the latter figure might trigger a signal based on a quantitative threshold, whereas the correct figure of three might not. If not corrected in time, this could lead to flawed decision making and unnecessary alerts.

A major problem for pharmacovigilance, and a criticism of spontaneous reporting schemes, is that there is considerable under-reporting. Although the

level of under-reporting may be relatively low for serious events and for events that are specific for certain medicines, it would obviously be much better if the situation could be improved. With targeted, and repeated, education efforts, the resulting greater awareness by the public, patients, and doctors and other health professions would aid this aim. It has been shown that good feedback to doctors is very important in encouraging reporting.[47] There is a link between high reporting rates per capita, and the effort which has been put into feedback with useful information on the case in question; and an active education and publication policy.[47-50]

Given the resources required for improving the quantity, and quality, of reports, it seems likely that, for the foreseeable future, pharmacovigilance also will have to contend with both incomplete reporting and heterogeneous data. With a good awareness of its deficiencies, such data, nevertheless, when analysed and weighed carefully provides essential information, and can point the way ahead for further studies that may clarify the situation.

Although most countries have regulatory requirements in place determining who can or must report, what should be reported and when, the quality management of the data actually received is to a large extent left to those in charge of the database systems where this information is normally stored. Even if database systems are of high quality in respect to the data, this does not mean that the whole pharmacovigilance process, in which the data is merely one component, has been subject to critical appraisal and optimisation. With an increasing exchange of data, on the national and international level, one must also consider the implications of transfer between systems for data quality.

The risks of erroneous and misinterpreted data can be reduced, if not fully avoided. The introduction of the concept of 'good pharmacovigilance practice' is a step on the way to achieve higher quality of the pharmacovigilance process in the broader sense.[51] Another encouraging sign is that conferences on the topic of data quality in the broader pharmaceutical area have emerged in recent years, which should enable information exchange and possibilities to learn from each other's experiences.

To further improve the situation, the implementation of data quality management practices in pharmacovigilance needs to be put high on the agenda. The aim should be to have transparent, well documented and functioning quality systems in place. In this instance, it should be said that the author, regrettably, is fully aware that the system she is most familiar with does not yet live up to the standards put forward in this paper.

Together with close collaboration between all those involved in data management and analysis, good data quality management practices will directly benefit the custodians of pharmacovigilance data, and those who analyse it and base their decisions on the information (the data customers). However, this is not an endpoint in itself – all of those involved in pharmacovigilance should have the ultimate beneficiaries of our work in mind – the patients who are taking the medicines. They are the ones who are harmed if we do not do our job properly.

Acknowledgements

No sources of funding were used to assist in the preparation of this review. The author has no conflicts of interest that are directly relevant to the content of this review.

References

1. Strom BL. What is pharmacoepidemiology? In: Strom BL, editor. Pharmacoepidemiology. 3rd ed. Chichester: John Wiley & Sons Ltd, 2000: 3-15
2. WHO. The importance of pharmacovigilance: safety monitoring of medicinal products. Geneva: World Health Organization, 2002
3. Lindquist M. Seeing and observing in pharmacovigilance: achievements and prospects in worldwide drug safety [doctoral thesis]. Nijmegen: University of Nijmegen, 2003
4. Edwards IR. Adverse drug reactions. In: van Boxtel CJ, Santoso B, Edwards IR, editors. Drug benefits and risks: International text book of clinical pharmacology. Chichester: John Wiley and Sons Ltd, 2001: 195-209
5. Mecella M, Scannapieco M, Virgillito A, et al. Managing data quality in cooperative information systems. In: Proceedings of the 10th International Conference on Cooperative Information Systems (CoopIS 2002); USA: Springer-Verlag, 2002; 486-502
6. Redman TC. Data quality for the information age. Norwood: Artech House Inc, 1996
7. Uehling MD. Is data quality really job 1? [online]. Available from URL: http://www.bio-itworld.com/news/021003_report1988.html [Accessed 2003 Sep 25]
8. Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. Commun ACM 1996; 39 (11): 86-95
9. Strong DM, Lee YW, Wang RY. Data quality in context. Commun ACM 1997; 40 (5): 103-10
10. Tull PG. Managing data quality within BT: a feasibility study [master's thesis]. Leicester: Leicester University, 1997

11. Peachey J. From pharmacovigilance to pharmacoperformance. Drug Saf 2002; 25 (6): 399-405
12. Lindquist M. The WHO programme for international drug monitoring: the present and future. In: Mitchard M, editor. Electronic communication technologies. Bufffalo (NY); Interpharm Press Inc., 1998: 527-49
13. Bortnichak EA, Wise RP, Salive ME, et al. Proactive safety surveillance. Pharmacoepidemiol Drug Saf 2001; 10 (3): 191-6
14. Golden MS. An incident reporting system: documented at the point of service. J Healthc Risk Manag 1998; 18 (2): 18-26
15. Lumpkin MM. International pharmacovigilance: developing cooperation to meet the challenges of the 21st century. Pharmacol Toxicol 2000; 86 Suppl. 1: 20-2
16. Stergachis AS. Record linkage studies for postmarketing drug surveillance: data quality and validity considerations. Drug Intell Clin Pharm 1988; 22 (2): 157-61
17. Spindler P, Seiler JP. Quality management of pharmacology and safety pharmacology studies. Fundam Clin Pharmacol 2002; 16 (2): 83-90
18. Edwards IR. Spontaneous ADR reporting and drug safety signal induction in perspective: to honour Professor Jens Schou. Pharmacol Toxicol 2000; 86 Suppl. 1: 16-9
19. Strom BL, editor. Pharmacoepidemiology. 3rd ed. Chichester: John Wiley & Sons Ltd, 2000
20. Mann R, Andrews E, editors. Pharmacovigilance. 1st ed. Chichester: John Wiley & Sons Ltd, 2002
21. Girard M. Data quality in post-marketing surveillance. Adverse Drug React Acute Poisoning Rev 1986; 5 (2): 87-95
22. O'Neill RT. Biostatistical considerations in pharmacovigilance and pharmacoepidemiology: linking quantitative risk assessment in pre-market licensure application safety data, post-market alert reports and formal epidemiological studies. Stat Med 1998; 17 (15-16): 1851-8
23. Edwards IR, Wiholm B-E, Martinez C. Concepts in risk-benefit assessment. Drug Saf 1996; 15 (1): 1-7
24. Meyboom RHB, Egberts ACG, Edwards IR, et al. Principles of signal detection in pharmacovigilance. Drug Saf 1997; 16 (6): 3355-65
25. International Organization for Standardization. ISO 9000. Quality management systems: fundamentals and vocabulary. 2nd ed. Geneva: ISO, 2000
26. Redman TC. Why care about data quality? In: Data quality for the information age. Norwood: Artech House Inc., 1996: 3-16
27. English LP. Help for data quality problems (ensuring data quality in data warehouses). Information Week 1996 Oct 7; (600): 53
28. CIOMS. Harmonization of data fields for electronic transmission of case-report information internationally. Public report. Geneva: CIOMS, 1995
29. Management of the ICH guideline on clinical safety data management: data elements for transmission of individual case safety reports [online]. Available from URL: http://www.ich.org/pdfICH/e2bm.pdf [Accessed 2003 Sep 20]
30. WHO drug dictionary 2nd quarter 2003. Uppsala: The Uppsala Monitoring Centre, 2003
31. WHO Adverse Reaction Terminology Dec 2002. Uppsala: The Uppsala Monitoring Centre, 2002
32. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf 1999; 20 (2): 109-17
33. WHO. International statistical classification of diseases and related health problems, 1989 revision. Geneva: World Health Organization, 1992
34. van der Klauw MM, Stricker BH, Herings RM, et al. A population based case-cohort study of drug-induced anaphylaxis. Br J Clin Pharmacol 1993; 35 (4): 400-8
35. Rawlins MD. Pharmacovigilance: paradise lost, regained or postponed? The William Withering Lecture 1994. J R Coll Physicians Lond 1995; 29 (1): 41-9
36. Belton KJ. Attitude survey of adverse drug-reaction reporting by health care professionals across the European Union. The European Pharmacovigilance Research Group. Eur J Clin Pharmacol 1997; 52 (6): 423-7
37. Begaud B, Martin K, Haramburu F, et al. Rates of spontaneous reporting of adverse drug reactions in France [letter]. JAMA 2002; 288 (13): 1588
38. Wack JP, Carnahan LJ. Keeping your site comfortably secure: an introduction to Internet firewalls. NIST Special Publication 800-10 [online]. Available from URL: http://csrc.nist.gov/publications/nistpubs/800-10/main.html [Accessed 2003 Sep 20]
39. van Puijenbroek EP, Bate A, Leufkens HG, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002; 11 (1): 3-10
40. Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 1998; 54: 315-21
41. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat 1999; 53 (3): 177-90
42. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001; 10 (6): 483-6
43. van Puijenbroek EP. Quantitative signal detection in pharmacovigilance [Doctorate]. Utrecht: Utrecht Institute for Pharmaceutical Sciences, 2001
44. Lindquist M, Ståhl M, Bate A, et al. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. Drug Saf 2000; 23 (6): 533-42
45. Orre R. On data mining and classification using a Bayesian Confidence Propagation Neural Network [Doctorate]. Stockholm: Royal Institute of Technology, 2003
46. Brown EG. Methods and pitfalls in searching drug safety databases utilising the Medical Dictionary for Regulatory Activities (MedDRA). Drug Saf 2003; 26 (3): 145-58
47. Coulter DM. The New Zealand intensive medicines monitoring programme. Pharmacoepidemiol Drug Saf 1998; 7: 79-90
48. Scott HD, Thatcher-Renshaw A, Rosenbaum SE, et al. Physician reporting of adverse drug reactions. Results of the Rhode Island adverse drug reaction project. JAMA 1990; 263 (13): 1785-8
49. Orsini MJ, Orsini PA, Thorn DB, et al. An ADR surveillance programme: increasing quality, number of incidence reports. Formulary 1995; 30 (8): 454-61
50. McGettigan P, Golden J, Conroy RM, et al. Reporting of adverse drug reactions by hospital doctors and the response to intervention. Br J Clin Pharmacol 1997; 44 (1): 98-100
51. Concept paper. Risk assessment of observational data: good pharmacovigilance practices and pharmacoepidemiological assessment [online]. Available from URL: http://www.fda.gov/cder/meeting/groupIIIfinal.pdf [Accessed 2003 Sep 20]

Correspondence and offprints: Dr *Marie Lindquist*, Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring, Stora Torget 3, Uppsala, 752 37, Sweden.
E-mail: marie.lindquist@who-umc.org