

Model Formulation ■

Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework

DANIELLE G. T. ARTS, MSc, NICOLETTE F. DE KEIZER, PhD,
GERT-JAN SCHEFFER, MD, PhD

Abstract Over the past years the number of medical registries has increased sharply. Their value strongly depends on the quality of the data contained in the registry. To optimize data quality, special procedures have to be followed. A literature review and a case study of data quality formed the basis for the development of a framework of procedures for data quality assurance in medical registries. Procedures in the framework have been divided into procedures for the co-ordinating center of the registry (central) and procedures for the centers where the data are collected (local). These central and local procedures are further subdivided into (a) the prevention of insufficient data quality, (b) the detection of imperfect data and their causes, and (c) actions to be taken / corrections. The framework can be used to set up a new registry or to identify procedures in existing registries that need adjustment to improve data quality.

■ *J Am Med Inform Assoc.* 2002;9:600–611. DOI 10.1197/jamia.M1087.

Several developments in healthcare, such as progress in information technology and increasing demands for accountability, have led to an increase in the number of medical registries over recent years. We define a medical registry as a systematic collection of a clearly defined set of health and demographic data for patients with specific health characteristics, held in a central database for a predefined purpose (based on Solomon et al.¹). The specific patient characteristics (e.g., the presence of a disease or whether an intervention has taken place) determine which patients should be registered. Medical registries can serve different purposes—for instance, as a tool to monitor and improve quality of care or as a resource for epidemiological research.² One example is the

National Intensive Care Evaluation (NICE) registry, which contains data from patients who have been admitted to Dutch intensive care units (ICUs) and provides insight into the effectiveness and efficiency of Dutch intensive care.³

To be useful, data in a medical registry must be of good quality. In practice, however, quite frequently incorrect patients are registered or data items can be inaccurately recorded or not recorded at all.^{4–8} To optimize the quality of medical registry data, participatory centers should follow certain procedures designed to minimize inaccurate and incomplete data. The objective of this study was to identify causes of insufficient data quality and to make a list of procedures for data quality assurance in medical registries and put them in a framework. By data quality assurance we mean the whole of planned and systematic procedures that take place before, during, and after data collection, to guarantee the quality of data in a database. Our proposed framework for procedures for data quality assurance is intended to serve as a reference during the start-up of a registry. Furthermore, comparing current procedures in exist-

Affiliations of the authors: Department of Medical Informatics, Academic Medical Center, Amsterdam, The Netherlands (DGTA, NFdeK); Amphia Hospital, Breda, The Netherlands (GJS).

Correspondence and reprints: Danielle G. T. Arts, MSc, Department of Medical Informatics and Department of Intensive Care, J2-257, Academic Medical Center, P.O. Box 22700, 1100 DE, Amsterdam, The Netherlands; e-mail: <D.G.Arts@amc.uva.nl>.

Received for publication: 2/1/02; accepted for publication: 6/17/02.

ing medical registries with the proposed procedures in the framework should allow the identification of possible adjustments in the organization to improve data quality.

Methods

Literature Review

To gain insight into the concept of data quality, we searched the literature for definitions. For the development of the data quality framework we searched the literature for (a) types and causes of data errors and (b) procedures that can minimize the occurrence of data errors in a registry database. An automated literature search was done using the Medline and Embase databases. The following text words and MeSH headings were used in this search: "data quality," "registries," "data collection," "validity," "accuracy," "quality control," and combinations of these terms. The automated search spanned the years from 1990 to 2000. To supplement the automated search, a manual search was done for papers referenced by other papers, papers and authors known by reputation, and papers from personal databases and the Internet. To search the Internet we used the same terms as for the literature search. The manual search was not restricted to the medical domain or to a specified time period. Papers were considered relevant if they described the analysis of data quality in a registry database or a trial database in terms of types and frequencies of data errors or causes of insufficient data quality. In addition, we selected papers that described procedures for the control and the assurance of data quality in registry or trial databases. We considered registry and trial databases as systematic collections of a prespecified set of data items for all people with a specific characteristic of interest (e.g., patients admitted to intensive care) for a predefined purpose, such as evaluative research.

Methods for the Case Study

We analysed the types and causes of data errors that may occur in a registry by performing a case study at two ICUs that had collected data for the NICE registry for at least one year. The NICE dataset contains 96 variables representing characteristics of the patients and the outcome of ICU treatment. It includes demographic data, admission and discharge data and all variables necessary to calculate severity of illness scores and mortality risks according to the prognostic models APACHE II and III,^{9,10} SAPS II,¹¹

MPM II,¹² and LODS.¹³ Thirty-nine variables are categorical (e.g. presence of chronic renal insufficiency prior to ICU admission), 48 are numerical (e.g., highest systolic blood pressure during the first 24 hours of ICU admission), 7 are dates/times and 2 are character strings (Appendix 1).

One of the ICUs in the case study collected data automatically by extracting the data from their electronic patient data management system (PDMS)¹⁴ into a local registry database. The other ICU collected the data manually by filling in case record forms (CRFs) that were manually entered into a local registry database. Each month the data from the local databases from both ICUs are transferred to the central registry database at the NICE coordinating center. Data flows are shown in Figure 1.

For each ICU we retrieved from the central registry database the records of 20 randomly selected patients that had been admitted in September or October of 1999. To evaluate the accuracy and the completeness of the data we compared the data from the central registry database, the local database and the CRFs with the gold standard data. The gold standard data were re-abstracted from the paper patient record or the PDMS by one of the authors (DA). Registered values were found to be inaccurate if (1) a categorical value was not equal to the gold standard value or (2) a numerical value deviated from the gold standard value more than acceptable (e.g., a deviation in systolic blood pressure of > 10 mmHg below or above the gold standard systolic blood pressure). The appendix contains a complete list of variables and criteria. A data item was found to be incomplete when it was not registered, even though it was available in the paper-based record or the PDMS.

By means of a structured interview with the physicians responsible for the data collection process, we gained insight into the local organization of the data collection. This information and discussion of discovered data errors helped us to identify the causes of insufficient data quality. The causes of insufficient data quality that we found through the literature review and the case-study were grouped according to their place in the data collection process.

Quality Assurance Framework

Procedures for data quality assurance were collected through literature review as described before. According to their characteristics, the procedures were placed in a framework that maps with the

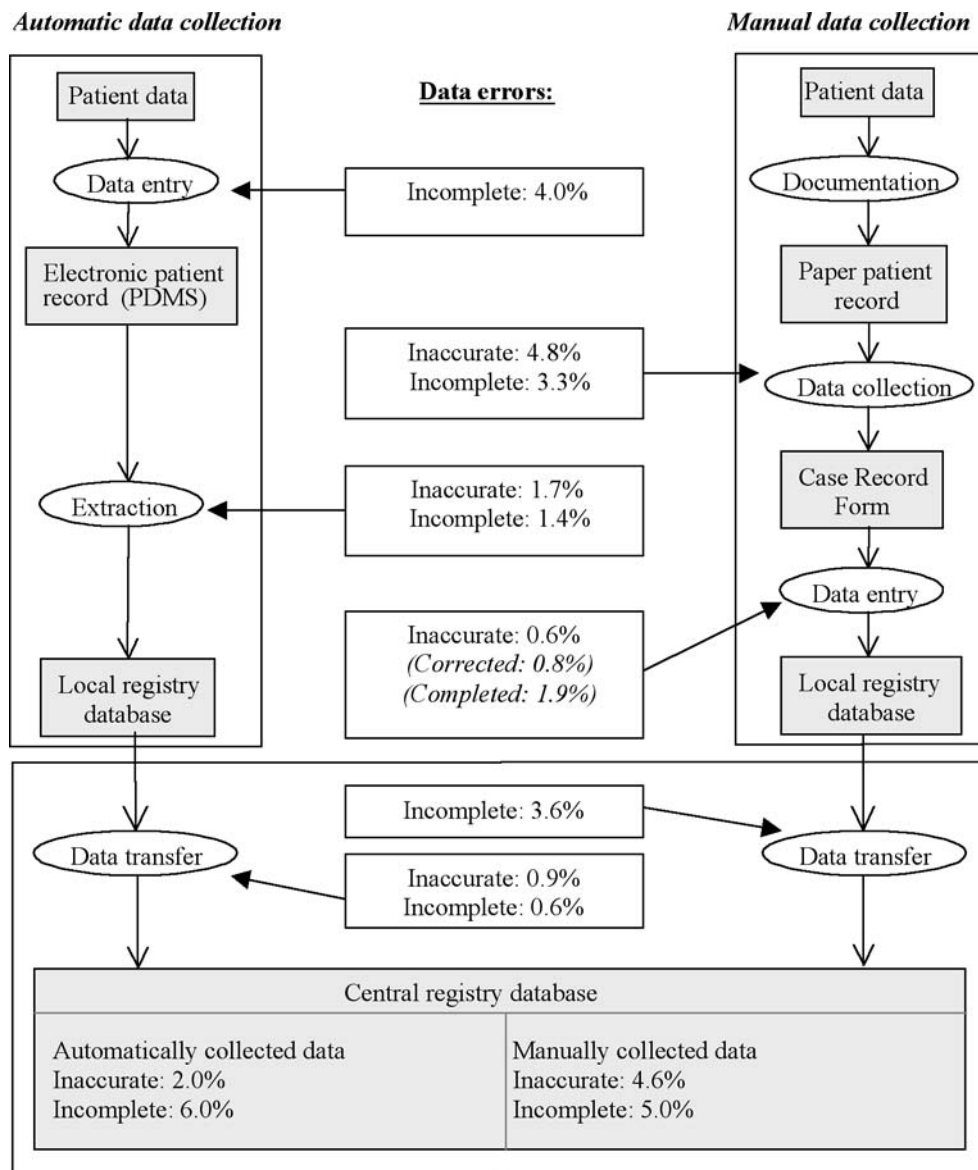


Figure 1 Results from the case study: types and percentages of newly occurring data errors at the different steps in the data collection process.

grouping of data error causes, obtained by the literature review and case study.

Results

Data Quality Definitions from Literature

According to the International Standards Organisation (ISO) definition, quality is “the totality of features and characteristics of an entity that bears on its ability to satisfy stated and implied needs” (ISO 8402-1986, Quality-Vocabulary) Similarly, in the context of a medical registry, data quality can be defined as “the totality of features and characteristics of a data set, that bear on its ability to satisfy the needs that

result from the intended use of the data.” Many researchers point out that data quality is an issue that needs to be assessed from the data users’ perspective. For example, according to Abate et al.,¹⁵ data are of the required quality if they satisfy “the requirements stated in a particular specification and the specification reflects the implied needs of the user.”

The review of relevant literature yielded a large number of distinct data quality attributes that might determine usability. Most of the data quality attributes in literature had ambiguous definitions or were not defined at all. In some cases, multiple terms were used for one single data quality attribute in different articles. The two most frequently cited data quality attrib-

utes were “accuracy” and “completeness.”^{6,7,15–31} From all data quality attributes found in literature, these two are the most relevant in the context of the case study. Based on the definitions found in literature, we formulated clear, unambiguous definitions for (1) data accuracy (the extent to which registered data are in conformity to the truth) and (2) data completeness (the extent to which all necessary data that could have been registered have actually been registered).

Types and Causes of Data Errors

Literature Review

Review of relevant literature^{4,17–19,32–38} resulted in a number of types of data errors. Van der Putten et al.³⁶ divides data errors into interpretation errors, documentation errors and coding errors. Other authors divide data errors into systematic (type I) errors and random (type II) errors.¹⁷ Knatterud et al.³⁵ additionally mentions bias as a category of data errors, which can cause random as well as systematic errors. Causes of systematic data errors include programming errors,^{32,34} unclear definitions for data items,^{4,33,36} or violation of the data collection protocol.^{34,35} Random data errors, for instance, can be caused by inaccurate data transcription and typing errors^{18,32,35,39} or illegible handwriting in the patient record.^{19,34} Clarke³² describes only two causes of data errors, inaccurate data transcription and programming errors in the software used. These two causes of errors are most frequently cited in literature. Inaccurate data transcription occurs during the actual data collection process. Alternatively, programming errors are part of the procedures that precede the actual data collection process. Other examples are the lack of clear definitions for data items and guidelines for data collection^{4,33,36} or insufficient training of participants in using the data definitions and guidelines.^{34,37}

We have now identified several possible types and causes of data errors through literature review. To analyse the types and causes of data errors in real practice we performed a case study at the NICE registry.

Case-study of Data Quality in the NICE Registry

Figure 1 displays the frequencies of inaccurate and incomplete data as they occurred during the different steps in the data collection process. Because some variables were not documented in the PDMS, for all patients, 4.0% of the data was not available in the extraction source. After the extraction of the data from

the PDMS another 1.4% of the data was incomplete because the extraction software lacked queries for some data items. Of the extracted data, 1.7% was inaccurate because of programming errors in the extraction software. Because of programming errors in the software used for transferring data into the central registry database, another 0.6% of the data was incomplete and 0.9% was incorrect. For example, one extraction query did not return the longest prothrombin time in the first 24 hours of ICU admission but in the entire ICU stay. Finally, the central registry database contained 2.0% inaccurate and 6.0% incomplete data for the hospital with automatic data collection.

In the hospital using manual data collection, after transcription of data from the paper patient record to the CRF 4.8% of the data was inaccurate and 3.3% was incomplete. We identified several causes of these errors, such as inaccurate transcription of the data and inaccurate calculations of derived variables such as daily urinary output and alveolar-arterial oxygenation difference. A relatively frequent error cause was nonadherence to data definitions. Contrary to the definitions, data outside the first 24 hours of ICU admission were frequently registered. The next step in the data collection process is the manual entry of the data from the CRFs into the local registry database. Inaccurate typing was a relatively infrequent cause of data errors (0.6%). During entry of data in the local database, some data that were inaccurate or incomplete on the CRFs were corrected or completed. Programming errors in the software used for transferring the data into the central registry database caused another 3.6% incomplete data items. Finally, the central registry database contained 4.6% inaccurate and 5% incomplete data for the hospital with manual data collection.

The causes of inaccurate data that we found through the literature review and the case study, and the types of data errors that they cause are presented in Table 1. The causes are grouped according to the related stage in the registration process and subdivided into causes at central and at local level. Central level refers to the coordinating center that sets up the registry and transfers data sets from all participating centers into the central registry database. Local level refers to the sites where the actual data collection takes place.

Proposed Framework with Procedures for Data Quality Improvement

Many different quality assurance procedures have been discussed in literature. Whitney et al.³⁸ discuss

Table 1 ■

Causes of Insufficient Data Quality in Medical Registries at the Different Stages in the Registry Process and Type (Systematic or Random) of Data Errors

Central coordinating center	Data errors	Local sites	Data errors
Set up and organisation of registry			
Unclear / ambiguous data definitions	Systematic	Illegible handwriting in data source	Random
Unclear data collection guidelines	Systematic	Incompleteness of data source	Systematic
Poor CRF lay-out	Systematic/ random	Unsuitable data format in source Data dictionary not available to data collectors	Systematic/ Systematic/ random
Poor interface design	Systematic/ random	Lack of motivation	Random
Data overload	Random	Frequent shift of personnel	Random
Programming errors	Systematic	Programming errors (data entry module/ extraction software)	Systematic
Data collection			
No control over adherence to guidelines and data definitions	Systematic	Non-adherence to data definitions Non-adherence to guidelines	Systematic Systematic
Insufficient data checks	Systematic/ random	Calculation errors Typing errors Insufficient data checks at data entry	Systematic/ random Random Systematic/ random
		Transcription errors Incomplete transcription Confusing data corrections on CRF	Random Random Random
Quality improvement			
Insufficient control over correction of detected data errors locally	Systematic/ random	No correction of detected data errors	Systematic/ random
Lack of a clear plan for quality improvement	Systematic/ random		

data quality in longitudinal studies. They make a distinction between *quality assurance* procedures and *quality control* procedures. Quality assurance consists of activities undertaken before data collection to ensure that the data are of the highest possible quality at the time of collection. Examples of quality assurance procedures are a clear and extensive study design and training of data collectors. Quality control takes place during and after data collection and is aimed at identifying and correcting sources of data errors. Some examples of quality control procedures are completeness checks and site visits.³⁸

Knatterud et al.³⁵ describe guidelines for standardisation of quality assurance in clinical trials. According to the authors, quality assurance should include prevention, detection, and action, from the beginning of the data collection through publication of the study results. Important aspects of prevention are the selection and training of adequate and motivated personnel and the design of a data collection protocol. Detection of data errors can be achieved through routinely monitoring the data, which means that they are compared with data in another independent data

source. Finally, action implies that data errors are corrected and causes of data errors are resolved.

To standardize data collection in a registry, the data items that need to be collected should be provided with clear data definitions, and standardized guidelines for data collection methods must be designed.^{18,19,32–35,38} Many authors recommend training of persons involved in the registry.^{5,18,33–35,38,40–42} Issues in training sessions are the scope of the registry, the data collection protocol and data definitions. Participants should be trained centrally to guarantee standardization of the data collection procedures across participating centers.

From two studies^{20,43} it appeared that data quality improves if the CRF contains less open-ended questions. To reduce the chance of errors occurring in the data during collection several authors recommend collecting the data in space close to the original data source as soon as the data are available.^{19,40,44,45} Ideally, the data should be entered by the clinician or obtained directly from the relevant electronic data source (e.g., a laboratory system).¹⁹

Table 2 ■

Framework of Procedures for the Assurance of Data Quality in Medical Registries

Central coordinating centre	Local sites
Prevention during set up and organisation of registry	
<p>At the onset of the registry</p> <ul style="list-style-type: none"> compose minimum set of necessary data items define data & data characteristics in data dictionary draft a data collection protocol define pitfalls in data collection compose data checks create user friendly case record forms create quality assurance plan <p>In case of new participating sites</p> <ul style="list-style-type: none"> perform site visit train new participants <p>Continuously</p> <ul style="list-style-type: none"> motivate participants communicate with local sites <p>In case of changes (e.g., in data set)</p> <ul style="list-style-type: none"> adjust forms, software, data dictionary, protocol, training material, etc. communicate with local sites 	<p>At the onset of participating in the registry</p> <ul style="list-style-type: none"> assign a contact person check developed software for data entry and for extraction check reliability and completeness of extraction sources standardise correction of data items <p>Continuously</p> <ul style="list-style-type: none"> train (new) data collectors motivate data collectors make data definitions available place date & initials on completed forms keep completed case record forms data collection close to the source and as soon as possible use the registry data for local purposes <p>In case of changes (e.g., in data set)</p> <ul style="list-style-type: none"> adjust data dictionary, forms, software, etc. communicate with data collectors
Detection during data collection	
<p>During import of data into the central database</p> <ul style="list-style-type: none"> perform automatic data checks <p>Periodically and in case of new participants</p> <ul style="list-style-type: none"> perform site visits for data quality audit (registry data <> source data) and review local data collection procedures <p>Periodically</p> <ul style="list-style-type: none"> check inter- and intraobserver variability perform analyses on the data 	<p>Continuously</p> <ul style="list-style-type: none"> visually inspect completed forms perform automatic data checks check completeness of registration
Actions for quality improvement	
<p>After data import and data checks</p> <ul style="list-style-type: none"> provide local sites with data quality reports control local correction of data errors <p>After data audit or variability test</p> <ul style="list-style-type: none"> give feedback of results and recommendations resolve causes of data errors 	<p>After receiving quality reports</p> <ul style="list-style-type: none"> check detected errors correct inaccurate data & fill in incomplete data resolve causes of data errors <p>After receiving feedback</p> <ul style="list-style-type: none"> implement recommended changes communicate with personnel

Several different methods can be applied to detect errors at data entry. For example the registry data can be entered twice, preferably by two independent persons. Detected inconsistencies, that indicate data-entry errors should be checked and re-entered.^{19,38,41,44,46} Automatic domain or consistency checks on the data at data entry, data extraction or data transfer, can also detect anomalous data.^{8,19,32,38,42,44,45} Not all data errors can be detected through automatic data checks. Data errors that are still within the predefined range will not be uncovered. Therefore, in addition to the automatic checking of the data, a visual check of the entered data is recommended.^{8,44} Analyses of the data (e.g., simple cross-tabulation) could also help to uncover anomalies in data patterns.^{32,35,44} The coordinating center of

a registry can control data quality by visiting the participating centres and performing data audits. These audits imply that a sample of the data from the central registry database is being compared with the original source data (e.g., in the paper patient record).^{34,35,38,42}

Based on causes of insufficient data quality from the case study and on experiences described in literature we made a list of procedures for data quality assurance. These procedures have been placed in a framework (Table 2) in which quality assurance procedures are divided into “central” and “local” procedures. Central and local procedures are further subdivided into three phases: (a) the prevention of insufficient data quality, (b) the detection of inaccurate or incom-

plete data and their causes and (c) corrections or actions to be taken to improve data quality. This grouping for the data quality assurance procedures resembles the grouping of causes of data errors in Table 1. Preventive procedures are aimed at the causes of inaccurate data due to deficiencies in the set-up or the organization of the registry. Detection of inaccurate data takes place during the local data collection process and during the transfer of data into the central database. Finally detected data inaccuracies and their causes should lead to actions that improve data quality.

Discussion

The definitions of data quality that we found clearly give “the data requirements that proceed from the intended use” a pivotal position. Thus, the intended use of registry data determines the necessary properties of the data. For example, in a registry that is used to calculate incidence rates of diseases, it is essential to include all existing patient cases. In other cases (e.g., registries used for case-control studies), it is essential to record correctly characteristics of registered patients, such as diagnoses; the exact number of included patients is of minor importance.¹⁷

Definitions of data quality and data quality attributes in literature are frequently unclear, ambiguous or unavailable. Before designing a plan for quality assurance of registry data, a clear description of what attributes constitute data quality is necessary. Additionally, standard definitions of data quality and data quality attributes are necessary to be able to compare data quality among registries or within a registry at different points in time.

Investigating the causes of data errors is a prerequisite for the reduction of errors. The literature search that we performed was not conducted fully in accordance with the methodology of a systematic review. Nevertheless, we believe that we captured most of the relevant articles. In addition to the literature search, we performed a case study. We analyzed the types and causes of data errors in the NICE registry at the different steps in the data collection process, from the patient record to the central registry database. We did not question the quality of data documented in the patient record, either electronic or paper-based. Nevertheless, ample evidence in the literature indicates that the patient record is generally not completely free of data errors.^{47–49} For our case study, however, the paper-based patient record or the PDMS was the most reliable source available. The case study

showed missing values for some variables (e.g., admission source, ICU discharge reason) for each patient in the PDMS, because these variables were not configured in the PDMS. This was a relatively large source of systematically missing data for the hospital with automatic data collection. The paper-based records also missed some of the data that were obligatory in the registry. This however was a minor cause of randomly missing data in the registry.

From our case study, it appeared that in case of automatic data collection, data errors are mostly systematic and caused by programming errors. The disadvantage of these programming errors is that they can cause a large number of data errors. On the other hand, once the programming error is detected, this cause can easily be resolved. In case of manual data collection, errors appeared to be mostly random. Most data errors occurred during recording of the data on the CRFs due to inaccurate transcription or non-adherence to data definitions. The fact that in this case most data errors occur during transcription of data to the CRF corresponds to the results of other studies.³⁹ The percentages of inaccurate and incomplete data in case of manual data collection are comparable to those found in other studies.^{7,39,42} The literature search yielded no articles presenting results of data quality analysis for automatically extracted data to compare with the results of our case study.

It is unrealistic to aim for a registry database that is completely free of errors. Some errors will remain undetected and uncorrected regardless of quality assurance, editing, and auditing. Implementation of procedures for data quality assurance can merely lead to an improvement of data quality. As evident from the literature review, the assurance of data quality can entail many different procedures. Procedures were selected for our framework when they were practically feasible and when they seemed likely to prevent, detect, or correct frequently occurring errors. This implies that the procedures in the framework can be expected to be effective in improving data quality.

Procedures in the framework are divided into “central” and “local” procedures. This division was applied because most registries consist of several participating local centers that collect the data and send it to the central coordinating center that sets up and maintains the registry. Prud’homme et al.⁴² similarly described quality assurance procedures separately for all parties involved in a trial, such as the clinical centers, the coding center, and the data coordination center. Central and local procedures in our frame-

work are further subdivided into the prevention of insufficient data quality, detection of (causes of) insufficient data quality and actions to be taken. Knatterud et al.³⁵ and Wyatt¹⁹ made similar divisions. The framework of procedures for data quality assurance fits in the table with causes of data errors. If, for example, the local causes of data errors lie mainly in the data collection phase, quality assurance procedures from the local detection section in the framework should be implemented.

For the development of the framework we took two methods of data collection into consideration, manual and automatic data collection. We did not consider the use of alternative data entry methods such as the OCR/OMR scanning of registry data. Since manual and automatic data collection are the two most commonly used methods, the framework should be applicable for reviewing the organization in most medical registries.

The developed quality assurance framework for medical registries will also be useful for reviewing procedures and improving data quality in clinical trials. Good clinical practice guidelines for clinical trials state that procedures for quality assurance have to be implemented at every step in the data collection process.⁵⁰ To ensure the quality of data in clinical trials, the framework described in this article could be a suitable addition to the good clinical practice guidelines.

We believe that the framework proposed in this article can be a helpful tool for setting up a high quality medical registry. Our experiences with the NICE registry and with three other national and international registries support this opinion. In existing registries it can be a useful tool for identifying adjustments in the data collection process. Further research, such as pre- and postmeasurements of data quality, should be conducted to determine whether implementation of the framework in a registry in fact reduces the percentages of data errors.

The authors thank Martien Limburg and Jeremy Wyatt for their valuable comments on this manuscript.

References ■

- Solomon D, et al. Evaluation and implementation of public health registries. *Public Health Rep.* 1991;106(2):142-50.
- van Romunde L, Passchier J. Medische registraties: doel, methoden en gebruik [in Dutch]. *Ned Tijdschr Geneeskd.* 1992;136(33):1592-5.
- de Keizer N, Bosman R, Joore J, et al. Een nationaal kwaliteitssysteem voor intensive care [in Dutch]. *Medisch Contact.* 1999;54(8):276-9.
- Goldhill DR, Sumner A. APACHE II, data accuracy and outcome prediction. *Anaesthesia.* 1998;53(10):937-43.
- Lorenzoni L, Da Cas R, Aparo UL. The quality of abstracting medical information from the medical record: The impact of training programmes. *Int J Qual Health C.* 1999;11(3):209-13.
- Seddon D, Williams E. Data quality in the population-based cancer registration: An assessment of the Merseyside and Cheshire Cancer Registry. *Brit J Cancer* 1997;76(5):667-74.
- Barrie J, Marsh D. Quality of data in the Manchester orthopaedic database. *Br Med J.* 1992;304:159-62.
- Horbar JD, Leahy KA. An assessment of data quality in the Vermont-Oxford Trials Network database. *Control Clin Trials.* 1995;16(1):51-61.
- Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: A severity of disease classification system. *Crit Care Med.* 1985;13(10):818-29.
- Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults [see comments]. *Chest.* 1991;100(6):1619-36.
- Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993;270(24):2957-63.
- Lemeshow S, Teres D, Klar J, et al. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA.* 1993;270(20):2478-86.
- Le Gall JR, Klar J, Lemeshow S, et al. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. *JAMA.* 1996; 276(10):802-10.
- Weiss Y, Sprung C. Patient data management systems in critical care. *Curr Opin Crit Care.* 1996;2:187-92.
- Abate M, Diegert K, Allen H. A hierarchical approach to improving data quality. *Data Qual J.* 1998;33(4):365-9.
- Tayi G, Ballou D. Examining data quality. *Commun ACM.* 1998;41(2):54-7.
- Sørensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. *Int J Epidemiol.* 1996;25(2):435-42.
- Hilner JE, McDonald A, Van Horn L, et al. Quality control of dietary data collection in the CARDIA study. *Control Clin Trials.* 1992;13(2):156-69.
- Wyatt J. Acquisition and use of clinical data for audit and research. *J Eval Clin Pract.* 1995;1(1):15-27.
- Teperi J. Multi method approach to the assessment of data quality in the Finnish Medical Birth Registry. *J Epidemiol Community Health.* 1993;47(3):242-7.
- Hosking J, Newhouse M, Bagniewska A, Hawkins B. Data collection and transcription. *Control Clin Trials.* 1995; 16:66S-103S.
- AHIMA. The American health information management association practice brief: Data Quality Management Model. June, 1998. Available at <http://www.ahima.org/journal/pb/98-06.html>. Accessed Sept. 15, 2001.
- DISA. DOD Guidelines on Data Quality Management. 2000. Available at <http://datadm.disa.mil/dqpaper.html>. Accessed Aug. 15, 2001.
- Shroyer AL, Edwards FH, Grover FL. Updates to the Data Quality Review Program: The Society of Thoracic Surgeons Adult Cardiac National Database. *Ann Thorac Surg.* 1998; 65(5):1494-7.
- Golberg J, Gelfand H, Levy P. Registry evaluation methods: A review and case study. *Epidemiol Rev.* 1980;2:210-20.
- Teppo L, Pukkala E, Lehtonen M. Data quality and quality control of a population-based cancer registry. Experience in Finland. *Acta Oncol.* 1994;33(4):365-9.

27. Grover FL, Shroyer AL, Edwards FH, et al. Data quality review program: the Society of Thoracic Surgeons Adult Cardiac National Database. *Ann Thorac Surg*. 1996;62(4):1229–31.
28. Wand Y, Wang R. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*. 1996; 39(11):86–95.
29. Maudsley G, Williams EM. What lessons can be learned for cancer registration quality assurance from data users? Skin cancer as an example. *Int J Epidemiol*. 1999;28(5):809–15.
30. Mathieu R, Khalil O. Data quality in the database systems course. Sept, 1998. Available at <http://www.dataquality.com/998mathieu.htm>. Accessed Sept. 15, 2001.
31. Kaomea P. Valuation of data quality: A decision analysis approach. Sept, 1994. Available at <http://web.mit.edu/tdqm/www/papers/94/94-09.html>. Accessed Aug. 15, 2001.
32. Clarke PA. Data validation. In *Clinical Data Management*. Chichester, John Wiley & Sons, 1993, pp 189–212.
33. Clive RE, Ocwieja KM, Kamell L, et al. A national quality improvement effort: Cancer registry data. *J Surg Oncol*. 1995; 58(3):155–61.
34. Gassman JJ, Owen WW, Kuntz TE, et al. Data quality assurance, monitoring, and reporting. *Control Clin Trials*. 1995;16(2 Suppl):104S–136S.
35. Knatterud GL, Rockhold FW, George SL, et al. Guidelines for quality assurance in multicenter trials: a position paper. *Control Clin Trials*. 1998;19(5):477–93.
36. van der Putten E, van der Velden JW, Siers A, Hamersma EA. A pilot study on the quality of data management in a cancer clinical trial. *Control Clin Trials*. 1987;8(2):96–100.
37. Stein HD, Nadkarni P, Erdos J, Miller PL. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository [see comments]. *J Am Med Inform Assoc*. 2000;7(1):42–54.
38. Whitney CW, Lind BK, Wahl PW. Quality assurance and quality control in longitudinal studies. *Epidemiol Rev*. 1998; 20(1):71–80.
39. Vantongelen K, Rotmensz N, van der Schueren E. Quality control of validity of data collected in clinical trials. EORTC Study Group on Data Management (SGDM). *Eur J Cancer Clin Oncol*. 1989;25(8):1241–7.
40. Wang R, Kon H. Toward total data quality management (TDQM). June, 1992. Available at <http://web.mit.edu/tdqm/papers/92-02.html>. Accessed Aug. 15, 2001.
41. Neaton J, Duchene AG, Svendsen KH, Wentworth D. An examination of the efficiency of some quality assurance methods commonly employed in clinical trials. *Stat Med* 1990; 9(1–2):115–23, discussion, 124.
42. Prud'homme GJ, Canner PL, Cutler JA. Quality assurance and monitoring in the Hypertension Prevention Trial. Hypertension Prevention Trial Research Group. *Control Clin Trials*. 1989;10(3 Suppl):84S–94S.
43. Gissler M, Teperi J, Hemminki E, Merilainen J. Data quality after restructuring a national medical registry. *Scand J Soc Med*. 1995;23(1):75–80.
44. Blumenstein BA. Verifying keyed medical research data. *Stat Med*. 1993;12(17):1535–42.
45. Christiansen DH, Hosking JD, Dannenberg AL, Williams OD. Computer-assisted data collection in multicenter epidemiologic research. The Atherosclerosis Risk in Communities Study. *Control Clin Trials*. 1990;11(2):101–15.
46. Day S, Fayers P, Harvey D. Double data entry: what value, what price? *Control Clin Trials*. 1998;19(1):15–24.
47. Aronsky D, Haug PJ. Assessing the quality of clinical data in a computer-based record for calculating the pneumonia severity index. *J Am Med Inform Assoc*. 2000;7(1):55–65.
48. Hogan W, Wagner M. Accuracy of data in computer-based patient records. *J Am Med Inform Assoc*. 1997;5:342–55.
49. Wagner M, Hogan W. The accuracy of medication data in an outpatient electronic medical record. *J Am Med Inform Assoc*. 1996;3:234–44.
50. World Health Organization. WHO Technical Reports Series. Guidelines for Good Clinical Practice (GCP) for Trials on Pharmaceutical Products. 1995, Annex 3.

Appendix 1 ■

List of NICE variables, the severity-of-illness- models in which they are used, their data type, and the criteria used for analysis of data accuracy. Data accuracy was analysed by comparing the registry data to the gold standard (g.s.) data.

Admission data	Used in	Type	Criterion
Hospital number	—	Numerical	Must be consistent with g.s.
ICU number	—	Numerical	Must be consistent with g.s.
Admission number	—	Numerical	Must be consistent with g.s.
Patient number	—	Numerical (encrypted)	Must be consistent with g.s.
Surname	—	String (encrypted)	Must be consistent with g.s.
Maiden name	—	String (encrypted)	Must be consistent with g.s.
Date of birth	—	Date	Must be consistent with g.s.
Age	APACHE II & III, SAPS II, MPM _{0,24}	Numerical	Must be consistent with g.s.
Gender	—	Categorical	Must be consistent with g.s.
Length	—	Numerical	Deviation from g.s. must not exceed 10 centimetres.
Weight	—	Numerical	Deviation from g.s. must not exceed 10 kilograms.
Hospital admission date	—	Date	Must be consistent with g.s.
ICU admission date	—	Date	Must be consistent with g.s.

continued on following page

Appendix 1 (continued) ■

Admission data	Used in	Type	Criterion
ICU admission time		Time	Deviation from g.s. must not exceed 1 hour.
Referring specialty	—	Categorical	Must be consistent with g.s.
Admission source	—	Categorical	Must be consistent with g.s.
Admission type	APACHE II & III, SAPS II, MPM _{0,24}	Categorical	Must be consistent with g.s.
Planned admission	—	Categorical (yes/no)	Must be consistent with g.s.
ASA score at admission	—	Categorical	Must be consistent with g.s.
Diagnostic data	Used in	Type	Criterion
Cardio Pulmonary Resuscitation	MPM ₀	Categorical (yes/no)	Must be consistent with g.s.
Dysrhythmia	MPM ₀	Categorical (yes/no)	Must be consistent with g.s.
Cerebrovascular accident	MPM ₀	Categorical (yes/no)	Must be consistent with g.s.
Gastrointestinal bleeding	MPM ₀	Categorical (yes/no)	Must be consistent with g.s.
Intracranial mass effect	MPM _{0,24}	Categorical (yes/no)	Must be consistent with g.s.
Chronic renal insufficiency	MPM ₀	Categorical (yes/no)	Must be consistent with g.s.
Metastatic neoplasm	APACHE III, SAPS II, MPM _{0,24}	Categorical (yes/no)	Must be consistent with g.s.
AIDS	APACHE III, SAPS II	Categorical (yes/no)	Must be consistent with g.s.
Chronic dialysis	APACHE II	Categorical (yes/no)	Must be consistent with g.s.
Haematological malignancy	APACHE III, SAPS II	Categorical (yes/no)	Must be consistent with g.s.
Cirrhosis	APACHE II & III, MPM _{0v} , MPM ₂₄	Categorical (yes/no)	Must be consistent with g.s.
Chronic cardiovascular insufficiency	APACHE II	Categorical (yes/no)	Must be consistent with g.s.
Respiratory insufficiency	APACHE II	Categorical (yes/no)	Must be consistent with g.s.
Immunological insufficiency	APACHE II & III	Categorical (yes/no)	Must be consistent with g.s.
Mechanical ventilation at admission	MPM ₀	Categorical (yes/no)	Must be consistent with g.s.
Mechanical ventilation during first 24 hours	SAPS II, LODS II, MPM ₂₄	Categorical (yes/no)	Must be consistent with g.s.
Confirmed infection	MPM ₂₄	Categorical (yes/no)	Must be consistent with g.s.
Acute renal failure	APACHE II & III, MPM _{0,24}	Categorical (yes/no)	Must be consistent with g.s.
APACHE II Reason for admission	APACHE II	Categorical	Must be consistent with g.s.
Glasgow Coma Scale	Used in	Type	Criterion
Eye reaction at admission	MPM ₀	Categorical	Must be consistent with g.s.
Eye reaction after the first 24 hours of admission	MPM ₂₄	Categorical	Must be consistent with g.s.
Lowest eye reaction in the first 24 hours	APACHE II & III, SAPS II, LODS II	Categorical	Must be consistent with g.s.
Motor reaction at admission	MPM ₀	Categorical	Must be consistent with g.s.
Motor reaction after the first 24 hours of admission	MPM ₂₄	Categorical	Must be consistent with g.s.
Lowest motor reaction in the first 24 hours	APACHE II & III, SAPS II, LODS II	Categorical	Must be consistent with g.s.
Verbal reaction at admission	MPM ₀	Categorical	Must be consistent with g.s.
Verbal reaction after the first 24 hours of admission	MPM ₂₄	Categorical	Must be consistent with g.s.
Lowest verbal reaction in the first 24 hours of admission	APACHE II & III, SAPS II, LODS II	Categorical	Must be consistent with g.s.
Physiologic & laboratory data	Used in	Type	Criterion
Highest heartrate within 1 h after admission	MPM ₀	Numerical	Deviation from g.s. must not exceed 10 beats/min.

continued on following page

Appendix 1 (continued) ■

Physiologic & laboratory data	Used in	Type	Criterion
Lowest heartrate in the first 24 hours	APACHE II& III, SAPS II, LODS II	Numerical	Deviation from g.s. must not exceed 10 beats/min.
Highest heartrate in the first 24 hours	APACHE II& III, SAPS II, LODS	Numerical	Deviation from g.s. must not exceed 10 beats/min.
Lowest respiratory rate in the first 24 hours	APACHE II & III	Numerical	Must be exactly equal to g.s.
Highest respiratory rate in the first 24 hours	APACHE II & III	Numerical	Must be exactly equal to g.s.
Lowest systolic blood pressure within 1 h after admission	MPM ₀	Numerical	Deviation from g.s. must not exceed 10 mmHg.
Lowest systolic blood pressure in the first 24 hours	SAPS II, LODS II	Numerical	Deviation from g.s. must not exceed 10 mmHg.
Highest systolic blood pressure in the first 24 hours	SAPS II, LODS II	Numerical	Deviation from g.s. must not exceed 10 mmHg.
Lowest mean blood pressure in the first 24 hours	APACHE II & III	Numerical	Deviation from g.s. must not exceed 10 mmHg.
Highest mean blood pressure in the first 24 hours	APACHE II & III	Numerical	Deviation from g.s. must not exceed 10 mmHg.
Lowest body temperature in the first 24 hours	APACHE II & III, SAPS II	Numerical	Deviation from g.s. must not exceed 0.5 degrees Celsius.
Highest body temperature in the first 24 hours	APACHE II & III, SAPS II	Numerical	Deviation from g.s. must not exceed 0.5 degrees Celsius.
Longest prothrombin time in the first 24 hours	LODS II, MPM ₂₄	Numerical	Must be exactly equal to g.s.
Lowest urine output in 8 hours	MPM ₂₄	Numerical	Deviation g.s. must not exceed 0.250 Liter.
Urine output in the first 24 hours	APACHE III, SAPS II, LODS II	Numerical	Deviation from g.s. must not exceed 1 Liter.
Vasoactive drugs	MPM ₂₄	Categorical (yes/no)	Should be consistent with g.s.
Lowest PaO ₂ in the first 24 hours	MPM ₂₄	Numerical	Must be exactly equal to g.s.
Lowest PaO ₂ / FIO ₂ in the first 24 hours	SAPS II, LODS II	Numerical	Must be exactly equal to g.s.
FIO ₂ to calculate the highest A-aDO ₂ in the first 24 hours	APACHE II	Numerical	Must be taken from sample that results in highest A-aDO ₂ according to g.s.
PaO ₂ to calculate the highest A-aDO ₂ in the first 24 hours	APACHE II & III	Numerical	Must be taken from sample that results in highest A-aDO ₂ according to g.s.
PaCO ₂ to calculate the highest A-aDO ₂ in the first 24 hours	APACHE II	Numerical	Must be taken from sample that results in highest A-aDO ₂ according to g.s.
Highest A-aDO ₂ in the first 24 hours	APACHE II & III	Numerical	Must be exactly equal to g.s.
PH from same bloodgas sample as A-aDO ₂	APACHE II & III	Numerical	Must be taken from sample that results in highest A-aDO ₂ according to g.s.
Lowest white blood cell count in the first 24 hours	APACHE II & III, SAPS II, LODS II	Numerical	Must be exactly equal to g.s.
Highest white blood cell count in the first 24 hours	APACHE II & III, SAPS II, LODS II	Numerical	Must be exactly equal to g.s.
Lowest serum creatinine level in the first 24 hours	APACHE II & III, LODS II	Numerical	Must be exactly equal to g.s.
Highest serum creatinine level in the first 24 hours	APACHE II & III, LODS II	Numerical	Must be exactly equal to g.s.
Lowest serum potassium level in the first 24 hours	APACHE II, SAPS II	Numerical	Must be exactly equal to g.s.

continued on following page

Appendix 1 (continued) ■

Physiologic & laboratory data	Used in	Type	Criterion
Highest serum potassium level in the first 24 hours	APACHE II, SAPS II	Numerical	Must be exactly equal to g.s.
Lowest serum sodium level in the first 24 hours	APACHE II & III, SAPS II	Numerical	Must be exactly equal to g.s.
Highest serum sodium level in the first 24 hours	APACHE II & III, SAPS II	Numerical	Must be exactly equal to g.s.
Lowest serum bicarbonate level in the first 24 hours	APACHE II & III, SAPS II	Numerical	Must be exactly equal to g.s.
Highest serum bicarbonate level in the first 24 hours	APACHE II & III, SAPS II	Numerical	Must be exactly equal to g.s.
Highest serum urea level in the first 24 hours	APACHE III, SAPS II, LODS II	Numerical	Must be exactly equal to g.s.
Highest bilirubin level in the first 24 hours	APACHE III, SAPS II, LODS II	Numerical	Must be exactly equal to g.s.
Lowest haemoglobine level in the first 24 hours	APACHE II & III	Numerical	Must be exactly equal to g.s.
Highest haemoglobine level in the first 24 hours	APACHE II & III	Numerical	Must be exactly equal to g.s.
Lowest serum albumin level in the first 24 hours	APACHE III	Numerical	Must be exactly equal to g.s.
Highest serum albumin level in the first 24 hours	APACHE III	Numerical	Must be exactly equal to g.s.
Lowest level of platelets in the first 24 hours	LODS II	Numerical	Must be exactly equal to g.s.
Lowest glucose level in the first 24 hours	APACHE III	Numerical	Must be exactly equal to g.s.
Lowest glucose level in the first 24 hours	APACHE III	Numerical	Must be exactly equal to g.s.
Discharge data	Used in	Type	Criterion
ICU discharge date	—	Date	Must be consistent with g.s.
ICU discharge time	—	Time	Deviation from g.s. must not exceed 1 hour.
Discharge destination	—	Categorical	Must be consistent with g.s.
Reason for discharge	—	Categorical	Must be consistent with g.s.
Patient died at ICU	—	Categorical (yes/no)	Must be consistent with g.s.
Patient died in hospital	—	Categorical (yes/no)	Must be consistent with g.s.
Hospital discharge date	—	Date	Must be consistent with g.s.