Quality assessment of healthcare databases

Flavia Carle[1], Lidia Di Minco[2], Edlira Skrami[1], Rosaria Gesuita[1], Luigi Palmieri[3], Simona Giampaoli[3], Giovanni Corrao[4]

(1) Centre of Epidemiology and Biostatistics, Polytechnic University of Marche, Ancona, Italy
(2) National Health Information System, Ministry of Health, Rome, Italy
(3) Department of Cardiovascular, Dysmetabolic and Aging-associated Diseases, Istituto Superiore di Sanità, Rome, Italy
(4) Department of Statistics and Quantitative Methods, Università di Milano-Bicocca, Milan, Italy

CORRESPONDING AUTHOR: *Flavia Carle - Centro di Epidemiologia Biostatistica e Informatica Medica, Facoltà di Medicina - Università Politecnica delle Marche - via Tronto 10/a, 60020 Torrette di Ancona, Ancona, Italy - f.carle@univpm.it*

DOI: ***10.2427/12901***

ABSTRACT

The assessment of data quality and suitability plays an important role in improving the validity and generalisability of the results of studies based on secondary use of health databases. The availability of more and more updated and valid information on data quality and suitability provides data users and researchers an useful tool to optimize their activities.
In this paper, we have summarized and synthesized the main aspects of Data Quality Assessment (DQA) applied in the field of secondary use of healthcare databases, with the aim of drawing attention to the critical aspects having to be considered and developed for improving the correct and effective use of secondary sources.
Four developing features are identified: standardizing DQA methods, reporting DQA methods and results, synergy between data managers and data users, role of Institutions. Interdisciplinarity, multi-professionality and connection between government institutions, regulatory bodies, universities and the scientific community will provide the "toolbox" for i) developing standardized and shared DQA methods for health databases, ii) defining the best strategies for disseminating DQA information and results.

Key words: Quality, healthcare utilization databases, electronic health record databases

INTRODUCTION

Healthcare utilization databases (HUDs), and other secondary data sources, are being used more frequently in observational studies to estimate the burden of disease and to assess health care interventions worldwide. Their increased popularity, as a research tool, can be attributed to the large patient populations they cover, the continuity of data provision over time, low cost, timely availability, and applicability for studying real world clinical practice.

It is equally accepted that randomized controlled clinical trials (RCTs), although universally recognised as the most robust "evidence generators", are inadequate for guiding the decision making process since they are intrinsically unsuited to capture and assess the impact of treatments in routine clinical practice. Complexity of treatment regimens, demographic and clinical heterogeneity of patients receiving treatments, and the long time frame of many treatments, explain the gap between the evidence generated in the controlled, but artificial, setting of RCTs

and its actual impact in the real world [1, 2].

Comparative effectiveness research and translational research look at secondary data sources as a useful tool for improving the usability of studies results for public health and health policy.

Countries and government-funded public health agencies use HUDs for building core indicators for monitoring and assessing changes over time in health status, health determinants and health systems [3, 4]. Core health indicators can be defined as a set of measures (direct or indirect) of health status, determinants and care, changing over time along with health status, determinants and care changes. Core indicators are also useful for international comparisons at meta-national level; well-known sets of core health indicators are the Millennium Development Goals (MDGs) 2015 for the United Nations member states [5] and the European Core Health Indicators for member countries of the European Union [6].

The proliferation and use of electronic health record databases (EHRDs) in the clinical setting provides a rich secondary source of clinical data that can be used to support research on patient outcomes, comparative effectiveness, and health systems research. Particularly, reusing EHRDs provides the distinct ability to study patients and interventions in actual clinical practice as they naturally occur [2], facilitating rapid translation of study findings back into practice.

Most research efforts now include EHRDs abstraction to support individual studies, or more generally to support aggregation of large volumes of data in disease specific registries or clinical data repositories [7].

There are, however, serious reasons to justify considerable scepticism towards the ability of HUDs and other secondary data sources to fulfil the requirements of clinical and translational research, and to be useful for building valid and reliable health indicators. This scepticism derives from the fact that HUDs are designed and maintained mainly for the purposes of managing claims for reimbursements for healthcare services, as well as of monitoring the rational use of healthcare [8].

The scope of EHRDs is the routine collection of clinical information in primary care or in specific disease settings (e.g. diabetes, cancer); routine clinical data are also collected for clinical and billing uses, not for research. Advantages and limitations of using healthcare databases in clinical research, monitoring and assessing quality of care are analysed and discussed by several Authors [9, 10, 11, 12, 13].

There are two strategies to improve the usefulness of healthcare databases as secondary data source: to guarantee high quality data and apply rigorous and standardized methodology for planning and conducting studies using healthcare databases.

High quality data are the prerequisite for better information, better decision-making and better population health; they improve and strengthen the application of standardized study methodology.

The knowledge of data quality level is one tool to help secondary data users to improve the validity and generalizability of study findings.

Data quality (DQ) is recognized as a complex, multi-dimensional concept concerning different information systems contexts and multi-disciplinary expertises. Data quality is context dependent, which means the same data elements or data sources may be deemed high quality for one use and poor quality for a different use [14, 15]. The current literature on DQ is inconsistent in the use of terms that describe the complex multidimensional aspects of DQ [16]; DQ assessment methodology is not standardized and not always effective [17, 18].

The secondary use of different types of healthcare data sources can entail new DQ dimension definitions and different DQ assessment methods.

In this paper we consider categories and management levels of healthcare databases, DQ dimensions and DQ assessment methods applied in the context of data sources secondary use. The aim is to pinpoint the key elements needed to use DQ assessment to improve the validity and generalizability of studies based on secondary use of healthcare databases.

THE GENERATION OF HEALTHCARE DATABASES

Healthcare databases can be classified into four broad categories:

A. those that collect information for administrative purposes, such as the payment of health services and/or monitoring the health service supply, denoted as administrative or healthcare utilization databases (HUDs) [9],
B. those that collect all personal health information belonging to an individual, e.g. the patient's medical records used by practitioners tracking health information on their patients, denoted as electronic health or medical record databases (EHRDs) [19, 20],
C. those that collect information for both epidemiological and clinical purposes on patients diagnosed with a disease or other health-relevant condition, or undergoing a particular procedure or therapy, or using a health care service; population-based disease registries, hospital-based disease registries and clinical quality registries are included in this category [21, 22, 23],
D. those that collect information for epidemiological purposes on population's health and life styles; health examination surveys (HES) and health interview surveys (HIS) are included in this category [24, 25].

Although the main dimensions of the data quality

control are considered for all healthcare databases, the process of generating the healthcare records, the observation unit and the origin population of each healthcare databases category have to be taken into account in planning the quality control steps.

HUD can be defined as an electronic system designed to store, on an ongoing basis, healthcare encounters data (e.g., filled prescriptions, professional services, outpatient visits, hospitalizations), included patient's demographic data; such data are increasingly collected routinely for the payment and administration of health services for a well-defined dynamic population, e.g., people covered by a public or private healthcare delivery system. [9].

EHRD is an electronic version of an archive of patients' medical histories, that is maintained by a practitioner over time; it stores all data relevant to persons followed by a general practitioner or a specialist, including demographics, progress notes, problems, allergies, medicines, treatment plans, life habits, past medical history, immunizations, laboratory data and radiology reports, and so on. [26-27].

Details on diseases registries and HES/HIS are reported by Palmieri et al. [28] and Di Lonardo et al. [29], in this volume.

Figure 1 shows the process generating healthcare record for the four categories of healthcare databases. Health services and health supplies produce data that are directly recorded in the HUD and EHRD; specific diagnostic criteria have to be applied to select disease cases before recording data in disease registries; in drug-users registries and in registries on a specific diagnostic procedure, data from diagnostic investigation, drug prescription and pharmacy are directly recorded; the subjects recruited for health examination or interview survey provide data that are directly recorded in HES/HIS databases. The observation unit is the health provision in HUDs and the person in the other healthcare databases.

LEVELS OF DATA MANAGEMENT

HUDs are generated at different hierarchical levels reflecting the healthcare system organization and the health information system structure, including the modalities to transfer data between levels; the hierarchical levels can be the hospital or the healthcare unit, region, country.

At the lower hierarchical level HUDs contain information on the persons accessing a specific health care provider, e.g. an hospital; at the upper two levels HUDs include data from all health care providers in a region or in a country and provide information on the people accessing any health service provider in their geographic or administrative area of residence.

The quality control of recorded data is carried out at each hierarchical level; the upper levels (region, country) also perform the quality control of data flows between the levels.

EHRDs are generated by practitioners representing the lower hierarchical level; each general practitioner or specialist is responsible of the EHRD of own patients, including data quality control.

To support research and comparative effectiveness research, Practitioners' HERD Networks and Central Data Warehouses have been developed; these data sources are also used for peer-comparing of healthcare performance to improve the patient care quality, e.g. reducing the incidence of medical errors, and at improving adherence to guidelines [30-35]. In this upper hierarchical level of data management, data partners (lower level) maintain control of own databases and their uses, and perform data quality control; a coordinating centre of data partners is the second level for data quality assessment. The coordinating centre harmonizes and distributes standard procedures for collecting and registering data and for data quality control; central data quality checking procedures are also implemented.

A strong collaborative relationship between members of the coordinating centre and individual partners is essential for identifying and correcting data errors [36].

HEALTHCARE DATABASES QUALITY DIMENSIONS

Data quality has been defined as the capability of data to be used effectively, economically and rapidly to inform and evaluate decisions [18].

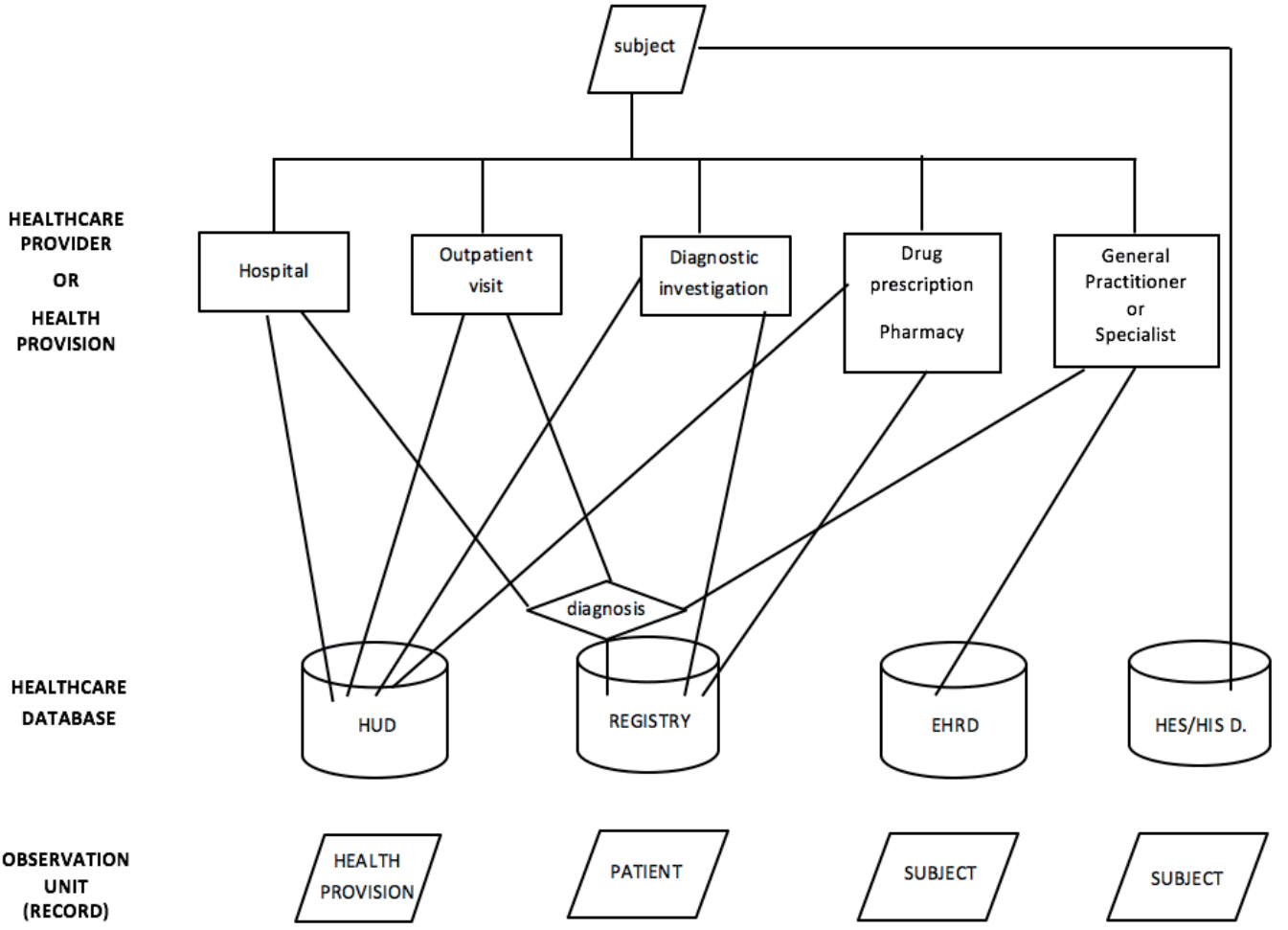
DQ is multi-dimensional, going beyond record-level accuracy to include such factors as accessibility, relevance, timeliness, metadata, documentation, user capabilities and expectations, cost and context-specific domain knowledge.

Defining DQ dimensions is the first step of DQ assessment process and helps addressing actions to improve DQ. Karr and Chen proposed to group the numerous data quality dimensions described in literature in three hyperdimensions, "data collection process", "data" and "data use" [17-18].

The dimension of "data collection process" refers to the generation, assembly, description and maintenance of data. The dimension of "data" focuses on data values or data schemas at record/table level or database level. The dimension of "data use", related to use and user, is the degree and manner in which data are used [17]. This approach has provided an exhaustive perspective to assess DQ quantitatively and qualitatively.

Data quality assessment methods are generally based on the measurement theory which links data attributes to concrete measures. Each dimension of data quality consists of a set of attributes. Each attribute characterizes a specific data quality requirement, thereby offering the standard for data quality assessment. Each attribute can be measured by different methods; therefore, there is variability in methods

FIGURE 1. Record generation process



used to measure data quality [17, 37, 38].

A feasible description of the data quality hyperdimensions are shown in Table 1. The reported definition of attributes describes the fundamental data quality features considered by most authors, even if named and classified differently. These features represent the information on the DQ necessary to optimize the choice and use of secondary sources. The reported associations between the attribute name and definition are the most used associations in DQ literature. We will refer to the contents of Table 1 below.

In the recent years, a number of reviews have been carried out to describe and compare methods for assessing data quality [17, 39-42]. The same concepts of data quality and the same variability in the definition of dimensions, attributes and measurements between different studies on HUDs or EHRDs were found.

Chen et al. [17] performed the review using the DQ hyperdimension framework and found that the three above-mentioned hyperdimensions were not given the same weight across the reviewed studies; the dimension of "data use" and the dimension of "data collection process" didn't receive adequate attention. Authors suggest this poor attention might reflect a lack of standardization and consensus on the definitions of dimensions, attributes and measures used in data quality assessment.

All authors of the reviews observed a high inconsistency of the terminology used in the reviewed studied. Completeness, accuracy, and timeliness were the three most-assessed attributes for both HUDs and EHRDs. Some studies represented completeness as the percentage of blank or unknown data, not zero/missing, or proportion of filling in all data in the facility report form; other studies measured completeness of data by the percentage of health facilities that completed data reports. The correspondence between data value in the database and right value was defined as "accuracy" or "validity" or "correctness" or "completeness" in the different reviewed studies. The fall of data value in exogenously defined and domain-knowledge dependent set of values was defined "validity", or "plausibility", or "attribute domain constraints". The availability of recorded data within a reasonable period of time following measurement was defined "timeliness" or "currency".

Inconsistent use of terms to describe DQ features makes it difficult to understand when similar or different DQ features are being discussed and compared.

TABLE 1. Dimensions, attribute and measurements of healthcare databases quality: a non-exhaustive portrayal

	HYPERDIMENSION		
	data collection process	**data [1]**	**data use [2]**

Attributes	• metadata documentation:
	  • procedures for collecting, recording and transmitting data;
	  • information on DB's structure and variables;
	  • quality data reports;
	• confidentiality
	• security
	• training
		Record level:
		• Completeness
		• Accuracy
		• Validity
		• Consistency
		
		Database level:
		• Identifiability
		• Joinability
		• Relational integrity
			• Accessibility
			• Integrability
			• Relevance
			• Timeliness
			• Rectifiability

Measurements	• Quality scores from questionnaire checklists of guidelines and procedures; • Number and frequency of quality audit with staff and stakeholders	• Descriptive statistics • Frequency tables • Sensitivity and specificity indices • Concordance indices • Correlation coefficients	• Number and trend of data requests; • Number and trend of publications that used databases; • Qualitative analysis and quantitative scores from user interviews • Presence of helpdesk

Notes:
DB= database
[1] Completeness: all fields of the record are filled;
Accuracy: the data value is the right value; accuracy is measured comparing the value in DB (e.g. hospital discharge diagnosis) and value from another source of information as gold standard (e.g. medical chart);
Validity: data value is defined be valid if it falls in exogenously defined and domain-knowledge dependent set of values; this definition include the agreement between data value and the pre-specified data format;
Consistency: it concerns the intra-relationship among variables in a DB; e.g. surgical intervention time must be precede discharge time; gender must be coherent with diagnosis;
Identifiability: each record in a database must have an unique identifier (primary key);
Joinability: the value that identifies the observation unit (e.g. patient) is the same in different databases (e.g. hospitalization DB, drug DB) including data related the observation unit (link key);
Relational integrity: it compares elements from one database to related elements in another database (e.g., every person identifier in the hospital DB must have a record in the vital statistics DB; two elements recording the same information for a single patient have the same value in different DBs).

[2] Physical and structural Accessibility: the data are available for secondary uses; are accessibility rules clear and public? Is the timing of the accessibility process declared and respected?
Integrability: DBs are designed to be integrated each other and procedure and tools for integration are used by database owner;
Relevance: the data in the DB are the data that users want; relevance may be change over time; a periodic feedback from users should be obtained;
Timeliness: the data have to be available in time for reaching the use aims; e.g. the healthcare performance assessment needs data on the healthcare that has been supplied close to the date of assessment;
Rectifiability: the establishment of procedures for users to request corrections or information on potential data anomalies.

DATA QUALITY ASSESSMENT

Data quality assessment (DQA) methods relate data quality attributes to measurable items (measurements) which can be calculated exactly; e.g. the ratio of total number of missing values to the total number of records in a database can be a measurement related to the attribute "completeness".

Data quality attributes and corresponding measurements need to be precisely defined in the reports of DQA results.

In the healthcare databases quality assessment process we can identify 3 different sequential stages depending on the use of data.

DQA - stage 1

The database is evaluated using a "fit-for-main use" perspective, without considering the potential relations with other databases. Attributes of the three quality hyperdimensions (Table 1) are routinely measured at the lowest hierarchical levels of data management, except accessibility, joinability, relational integrity and integrability. The evaluation is performed considering the main use for which the informative flow was activated. For example, the accuracy and validity assessment of the discharge diagnosis is mandatory in the HUD aimed at monitoring the resources absorbed by hospital services; if it is necessary to measure the burden of a specific disease

on hospital and pharmaceutical care, joinability will be able to be occasionally assessed in the same database.

The "fit-for-main use" DQA approach may also produce different quality degrees among recorded data; to give an example, we consider the HUDs designed and maintained mainly for the administrative purposes, as the reimbursements for provided healthcare services. In these HUDs, different levels of completeness can be accepted for the main diagnosis of discharge and the educational level, considering the last one is not included in the algorithm to classify hospital admission on the basis of the hospital resources consuming.

When more than one level of data management are involved, different DQ attributes are measured and different tools are used at each level. Accuracy is measured at the lowest hierarchical level, e.g. the hospital for HUD or the physician for EHRD, comparing the value in the DB (for example diagnosis of hospital discharge) with the value of another source of information considered as a gold standard (e.g. patient's chart). A random sample of records is used in the comparison with the gold standard. At upper hierarchical levels, the accuracy assessment rules are defined including sampling criteria and timing; the adherence to the rules is controlled in order to harmonize DQ of the databases stored at the lowest level.

At the upper hierarchical levels specific guidelines to control syntactic and semantic variability among the same databases managed in different sites, e.g. hospitals, are also provided.

Syntactic variability is an aspect of validity; it concerns data variability across databases caused by differences in the representation of data elements as format and units, and in the data position within the record. For example, body weight may be recorded and stored in different locations within a DB and in different formats or units. Semantic variability is an aspect of relational integrity of databases and concerns data variability caused by differences in the meaning of data elements. Differences in data collection, extraction methods, or measurement protocols across databases can result in semantic variability; for example, failure to distinguish between fasting and random blood glucose, finger-stick or venepuncture sampling, or serum or plasma measurements would result in glucose values that do not represent the same concept [40].

In the HUDs context, an interactive feed-back among the hierarchical levels is needed. The upper level, e.g. Ministry of health or national Agency for health information systems, provides lower levels with lists of controls and results of application of them. Central controls can cause an immediate action, e.g. the automatic refusal to include a record because it contains not-valid values, or warning messages on consistency of the recorded information; in addition, tables with descriptive statistics and other quality attribute measurements are provided at the lower hierarchical levels after the inclusion of records in the national database.

In order to ensure the alignment between the databases managed by the various hierarchical levels, the corrections of the anomalies detected by the central control (e.g. regional agency) are only performed at the level of data production (e.g. hospital).

In the EHRDs contest the DQ variability within and between DBs may be higher than HUDs context; electronic health record data are gathered during routine practice by individuals with a wide range of backgrounds and with different levels of commitment to data quality. Differences in measurement, recording, information systems, and clinical focus, increase the variability of electronic health record data quality and of DQA methods applied at the lowest hierarchical level (e.g. the practitioner) [16, 20, 39].

EHRDs networks have a strategic role in developing and making available shared standardized DQA methods and tools, not only for assessing data quality in secondary use of EHRDs but also in DQA-stage 1.

DQA-stage 2

The stage 2 is performed when the achievement of main-use goals requires aggregation of data on multiple sites or integration of different healthcare databases.

Monitoring and assessing healthcare performance at national level require the comparison of indicators from regional or other territorial units HUDs; peer-comparing of healthcare performance e.g. to reduce the incidence of medical error and improve adherence to guideline by clinicians, needs to aggregate data from EHRDs.

The comparisons can bring out differences among healthcare providers, unexpected clinical patterns or inconsistency across sites, trends, and cross-variables relations; it's needed to verify that this variability reflects true differences in healthcare performance or clinical practice and doesn't depend on differences in data quality. In addition to the quality controls carried out in the DQA-stage 1, active clinical audit and feedback programmes involving physicians and data managers working at lower hierarchical level can be activated by the upper levels of data management both of HUDs an EHRDs [36, 43-44].

The integration of healthcare databases including different health data on the same person is essential in epidemiological and comparative effectiveness research and in the assessment of integrated care [12, 45, 46].

The joinability, relational integrity, and integrability of healthcare databases (Table 1) provide information on capability and quality of DBs integration. These attributes are measured at the hierarchical level of data management having the responsibility of the DBs integration.

The definition and management of the link-key is the main operational aspect of the databases linkage; the procedures for creating a single anonymous identification code to be use as link-key in all the DBs to be linked can be different in each hierarchical level of data management.

The availability of guidelines or mandatory rules for anonymization procedures from the upper hierarchical level to the lower level can simplify the linking process and improve the quality of the DBs integration [47, 48]. In DQA-stage 2, the accessibility and the ease of use of these instructions are measurements of the attributes of quality hyperdimension "data collection process" (Table 1).

DQA-stage 3

The stage 3 is strictly related to the secondary use of healthcare databases in clinical, epidemiological, comparative effectiveness research and translational research.

Since healthcare databases are designed and maintained mainly for the purposes out of research tasks, it will be important to have a full understanding of the expendability and adequacy of an already-assembled dataset for testing the hypothesis of interest [40, 41, 49].

The HUDs may not include all the variables needed for the study; generally a HUD doesn't contain information such as prescribed drug dosages, laboratory test findings, lifestyle, etc. In EHRDs simplified classification or categorical scale may be used for some variables as body mass index, glycaemia, disease severity. Misclassification of exposure and outcome may occur, or approximations and classificatory algorithms defining study cohorts and groups of cases and controls have to be applied [12, 13].

The adequacy to the objectives of the study of the attributes evaluated in DQA-stage 1 is also taken into account; for example, random distribution of missing data must be verified, because, if not verified, a misclassification bias will be able to occur.

Finally, study-specific data checks are performed. These checks investigate exposure, outcome, and covariates of interest in detail, to identify potential biases and methods to control their effects.

In DQA-stage 3 the database is evaluated using a "fit-for-study purpose" perspective and the data quality assessment seems more exactly a data suitability assessment (DSA) to be performed for each research task.

Several author reviewed clinical, evaluative and epidemiological HUDs-based studies, to evaluate the data quality assessment applied methods. The aim was to define and propose standardized methods and frameworks to evaluate suitability of HUDs as secondary sources for research [15]. The reviews showed large variability among different studies, not only in quality assessment methods but also in quality terminology; they underlined the difficulty of proposing standardized and validated protocol for DQA usable for all research questions [41].

DSA of healthcare databases is firmly related to the methodology for planning and conducting observational studies and for analyzing data. DSA definitions and methods come from methodological research aiming to identify and control the pitfalls of observational research when based on HUDs [50-53].

In the secondary-sources-based studies, all three DQA-stages are needed but only the DQA-stage 3 must be performed for each study; DQA-stage 1 and DQA-stage 2 have to be performed independently by secondary use of DB and their results should be disposable for researcher and all data users. If DQA stage 1 and stage 2 aren't performed or their results aren't shared, researcher will have to carry out all DQA stages for guaranteeing the validity and solidity of the study.

Data quality assessment is typically conducted "behind the scenes" and the results aren't shared by public reports; some authors suggested to include information about the data quality approach and results as part of the standard comparative effectiveness research reporting template [36].

DISCUSSION

Data quality is an old-but-new problem; its dimensions, assessment methods and maintenance strategies change together with the development of new types of data collection, storage and use; modern research on DQ improvement is creating a large set of scientific, technological and process control challenges [18].

In recent years, the increasing use of healthcare utilization databases and other secondary data sources in medical research, and the possibility of interconnecting and aggregating different secondary sources of data, has reignited the attention towards the data quality dimensions and DQ assessment methods in healthcare and in medical research contexts.

While there is a general agreement on DQ multi-dimensional nature, there is no apparent consensus on definitions and measurements of data quality dimensions.

Health data quality domain is fragmented and concerns information technology, statistics, epidemiology and medicine.

In this paper, we have summarized and synthesized the mainly aspects of DQA applied in the field of secondary use of healthcare databases, with the aim of drawing attention to the critical aspects having to be considered and developed for improving the correct and effective use of secondary sources. The following four aspects should be considered and developed.

Standardizing DQA methods

The reviews of studies on quality measures and studies using secondary sources have shown high variability in the terminology and in the assessment methods. Several authors have proposed different frameworks aiming at harmonizing and standardizing terminology and methods, but this

strategy doesn't seem to have solved the problem yet.

In the era of big data, which brings together new data sources with widely varying data characteristics, new DQ concepts, measures, and methods will emerge, resulting in expansion or revision of the current terminology and methodology [16].

The challenge is not so much the production of standardized lists of terms and indicators, but rather the definition and dissemination of a shared methodology aimed at assessing the quality of healthcare databases. All potential uses of healthcare databases should be considered in the definition of DQA methods.

Data quality is recognized as a multi-dimensional concept covering large information systems contexts, specific knowledge and multi-disciplinary techniques [14, 17, 18]. Biostatisticians, epidemiologists, biomedical computer scientists, data managers, should work together to define DQA methodology; scientific societies and universities can play an important role promoting interdisciplinary projects and targeted educational events.

Reporting DQA methods and results

Detailed documentation of the rationale for conducting the data quality assessment and the results of these assessments is essential.

The organizations and institutions responsible for HUDs should disseminate the results of DQA-stage 1 and -stage 2 in periodic reports easily accessible on web. These reports should include the definition of quality dimensions and their attributes, DQA methods, the warning on changes over time in databases and in the data storage and extraction methodology. The results on the quality of data collection process should also be included, as well as the contact details of those responsible for the DQA.

Information on DQ and DQA could form a public and mandatory *"data quality metadata dossier"* for every HUD. The mandatory status should be guaranteed by Governmental Organizations and supported by the data users; if data sources with data quality metadata dossier are used, the efficiency of DQA-stage 3 and the speed of execution will be increased.

The same DQ metadata dossier can be also propose for EHRDs. If mandatory status cannot be applied, the EHRDs with DQ metadata dossier will be able to receive an accreditation by scientific societies and Governmental Organizations. This accreditation will be useful to address data users in the choice of data sources.

Disseminating and sharing these documents should also facilitate the harmonization of the definitions of dimension and attribute and the methods among data sources.

In the past years, literature on information science proposed the use of *"data quality metadata tags"* attached to every database. Information on data provenance, privacy permissions and restrictions, summary of values for a pre-defined standardized list of data quality measures should be included in these tags; informatics tools measuring data quality automatically could be developed [40].

Nevertheless experience has suggested a *"one size fits all"* set of data quality measures is not a solution [37]. Every data quality assessment plan is a compromise between time and resources and the desire for the highest possible data quality.

DQA-stage 1 and -stage 2 are closely related to technical, organizational, behavioral and environmental settings and sustainability of local routine health information system (for HUDs) or local/specific clinical context (for EHRDs). Different data quality decisions can be made by database managers and programmers on the same data sources.

Assessing data quality is an on-going effort requiring awareness of the application of the fundamental principles underlying the development of subjective and objective data quality measures in the specific context.

While the methods and results of DQA-stage 1 and -stage 2 are specific to each HUD and EHRD, the DQA-stage 3, called the Data Suitability Assessment (DSA), is specific and must be performed for each study based on use of secondary data sources.

The reporting of DSA in published studies using secondary sources is an important tool to evaluate and discuss strengths, weaknesses and generalizability of the studies [12, 36, 53].

Nevertheless this report is often inadequate or aimed at discussing some key aspects of the study objective; studies on the same research questions and using same secondary sources can use different terminology reporting DSA methods and results [53].

Creating a common format for DSA reporting in published studies would be beneficial, as researchers would be encouraged to provide complete information to make comparison of studies easier. The format should be comparable to the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE Statement) [15, 54].

Synergy between data managers and data users

Data quality assessment is a continuous dynamic process; the main use and the secondary use of data sources suggest updates and integrations of the databases inducing new data quality challenges.

Databases users have a role to play in both providing feedback on data quality issues uncovered during the in-depth data analysis, and partnering with data managers on improving accessibility, relevance, timeliness and rectifiability (Table 1).

The organizations and institutions responsible for HUDs and for EHRDs networks should initiate audit processes with data users via web tools such as structured

questionnaires and discussion forums.

Role of Institutions

Governmental institutions and regulatory bodies have particularly important roles to play in improving data quality and data quality assessment.

Regulation for DQA reports should be established; agreements protocols among HUDs providers, EHRDs owner and other non-health organizations should be developed to support and facilitate the interconnection among data sources.

More financial investment should be dedicated to data quality area; particular attention should be paid to identifying research funding for the development of data quality assessment methodology, also promoting initiatives calling for cooperation with private companies, for studies based on secondary use of healthcare databases.

In conclusion, the assessment of data quality and suitability plays an important role in improving the validity and generalisability of the results of studies based on secondary use of health databases. The availability of more and more updated and valid information on data quality and suitability provides to data users and to researchers a useful tool to optimize their activities.

Interdisciplinarity, multi-professionality and connection between government institutions, regulatory bodies, universities, and the scientific community will provide the "toolbox" for i) developing standardized and shared DQA methods for health databases, ii) defining the best strategies for disseminating DQA information and results.

Funding

The activities described in this paper were conducted in the framework of the Project 'Creazione e sviluppo del Network Italiano a supporto del progetto europeo BRIDGE-Health finalizzato a dare strutturazione e sostenibilità alle attività europee nel campo della Health Information (HI)' (Creation and development of the Italian Network to support the European project BRIDGE-Health aimed at structuring and providing sustainability to European activities in the field of Health Information), funded by the Italian Ministry of Health, Centre of Disease Control (CCM), and in the framework of the European project 'BRIDGE Health - bridging Information and Data Generation for Evidence-based Health Policy and Research' funded by the European Commission/DG Santè (Agreement no-664691 – BRIDGE Health).

Statement

The findings and conclusions provided in this paper are those of the authors, who are responsible for their contents; the findings and conclusions do not necessarily represent the views of the European Commission/DG Santè or the Italian Ministry of Health or the Istituto Superiore di Sanità. Therefore, no statement present in this report should be considered as official position of the European Commission/DG Santè or the Italian Ministry of Health or the Istituto Superiore di Sanità.

References

1. Dieppe P, Bartlett C, Davey P, et al. Balancing benefits and harms: the example of non-steroidal anti-inflammatory drugs. Br Med J 2004;329:31–34.
2. Sox HC, Goodman SN. The methods of comparative effectiveness research. Ann Rev Public Health 2012;33:425-445.
3. Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312:1215-1218.
4. World Health Organization. Framework and Standards for Country Health Information Systems; World Health Organization: Geneva, Switzerland, 2008.
5. The Millennium Development Goals Report 2015. Available online: http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf (last consultation May 6, 2018).
6. European Commission. European Core Health Indicators. Available online: http://ec.europa.eu/health/indicators/echi/index_en.htm (last consultation May 6, 2018) .
7. Andrew P. Reimer, Alex Milinovichb, Elizabeth A. Madigan, Data Quality Assessment Framework to Assess Electronic Medical Record Data for use in Research Int J Med Inform. 2016 June ; 90: 40–47.
8. Corrao G. Towards the rational use of Healthcare Utilization Databases for generating real-world evidence: new challenges and proposals. Epidemiology Biostatistics and Public Health - 2014, 11(3), e10328-e10328-6
9. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323–337.
10. Strom BL. Pharmacoepidemiology. 4th edn. Chichester: John Wiley & Sons Ltd; 2005.
11. Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects—advantages and disadvantages. Nature Clin Pract Rheumatol. 2007;3:725–732.
12. Corrao G. Building reliable evidence from realworld data: methods, cautiousness and recommendations. Epidemiology Biostatistics Public Health 2013;10:e8981-1-40.
13. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al., Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical Care, 2013. 51(Suppl 3): S30-S37.
14. Wand Y, Wang R. Anchoring data quality dimensions in ontological foundations. Comm ACM. 1996; 39:86–95.
15. Kahn MG.; Brown JS; Chun AT; Davidson BN; Meeker D; Ryan PB; Schilling LM; Weiskopf NG; Williams AE; Zozus MN. Transparent Reporting of Data Quality in Distributed Data Networks, eGEMs (Generating Evidence & Methods to improve

patient outcomes): 2015, Vol. 3: Iss. 1, Article 7. DOI: http://dx.doi.org/10.13063/2327-9214.1052 ,Available at: http://repository.academyhealth.org/egems/vol3/iss1/7
16. Kahn MG; Callahan TJ; Barnard J; Bauck AE; Brown J; Davidson BN; Estiri H; Goerg C; Holve E; Johnson SG; Liaw ST; Hamilton-Lopez M; Meeker D; Ong TC; Ryan P; Shang N; Weiskopf NG; Weng C; Zozus MN.; Schilling LA. Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs (Generating Evidence & Methods to improve patient outcomes): 2016, Vol. 4: Iss. 1, Article 18.
17. Chen H, Hailey D, Wang N, Yu P. A Review of Data Quality Assessment Methods for Public Health Information Systems. Int. J. Environ. Res. Public Health 2014, 11, 5170-5207.
18. Karr, A.F.; Sanil, A.P.; Banks, D.L. Data quality: A statistical perspective. Stat. Methodol. 2006, 3, 137–173.
19. World Health Organization, Electronic Health Records: Manual for Developing Countries; WHO Western Pacific Regional Publications, Manila, Philippines, 2006.
20. Krish Thiru, Alan Hassey, Frank Sullivan Systematic review of scope and quality of electronic patient record data in primary care BMJ. 2003 May 17;326(7398):1070
21. Miquel Porta JM Ed. A dictionary of epidemiology. 6th edn. Oxford University Press, 2014.
22. Gliklich RE, Dreyer NA, Leavy MB. Registries for Evaluating Patient Outcomes: A users guide, 3rd edition. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 Apr. Report No.: 13(14)-EHC111. AHRQ Methods for Effective Health Care.
23. Hoque DME, Kumari V, Hoque M, Ruseckaite R, Romero L, Evans SM Impact of clinical registries on quality of patient care and clinical outcomes: A systematic review. PLoS ONE 2017, 12(9): e0183667. https://doi.org/10.1371/journal.pone.0183667
24. Mindell JS, Giampaoli S, Goesswald A, Kamtsiuris P, Mann C, Männistö S, Morgan K, Shelton NJ, Verschuren WMM, Tolonen H, and on behalf of the HES Response Rate Group, Sample selection, recruitment and participation rates in health examination surveys in Europe – experience from seven national surveys. BMC Medical Research Methodology. 2015, 15:78
25. Aromaa A, Koponen P, Tafforeau J, Vermeire C; HIS/HES Core Group, Evaluation of Health Interview Surveys and Health Examination Surveys in the European Union. Eur J Public Health. 2003 Sep;13(3 Suppl):67-72.
26. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform. 2008 May;77(5):291-304
27. CMS.gov - U.S. Centers for Medicare & Medicaid Services. 7500 Security Boulevard, Baltimore, MD 21244 https://www.cms.gov/Medicare/E-Health/EHealthRecords/index.html (last consultation May 6, 2018).
28. Palmieri L, Veronesi G, Ferrario MM, Corrao G, Donfrancesco C, Carle F and Giampaoli S. Acute myocardial infarction and stroke registries. The Italian experience. EBPH 2018
29. Di Lonardo A, Donfrancesco C, Iannucci L, Gargiulo L, Palmieri L, Carle F and Giampaoli S. Ad hoc surveys: how to measure and report quality methods. EBPH 2018
30. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al., The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Genomics, 2010. 4(1): 13.
31. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed Health Data Networks A Practical and Preferred Approach to Multi-Institutional Evaluations of Comparative Effectiveness, Safety, and Quality of Care. Med Care 2010;48: S45–S51
32. Pacaud D, Szypowska A, Witsch M (ed. by), SWEET Project, Pediatric Diabetes 2016, Volume 17, Issue Supplement S23: 1–52
33. Rossi, M.C., Candido, R., Ceriello, A., Cimino A., Di Bartolo P, Giorda C, Esposito K, Lucisano G, Maggini M, Mannucci E, Meloncelli I, Nicolucci A, Pellegrini F, Scardapane M, Vespasiani G. Trends over 8 years in quality of diabetes care: results of the AMD Annals continuous quality improvement initiative. Acta Diabetol 2015, 52: 557
34. The Pedianet Project, http://www.pedianet.it/en (last consultation May 6, 2018)
35. The Healthsearch Project, https://www.healthsearch.it/ (last consultation May 6, 2018)
36. Brown J, Kahn M, Toh S, Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013, 51(8 0 3): S22–S29
37. Pipino LL, Lee YW, Wang RY. Data Quality Assessment. Commun. ACM 45, 2002, 4: 211-218.
38. Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 2009, 41, 1–52
39. Wong KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev. 2010; 67:503–527.
40. Kahn MG, Raebel MA, Glanz JM, et al. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012; 50(Suppl): S21–29.
41. Weiskopf NG, Weng C, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20:144–151
42. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, Jalaludin B, Yeo AE, Talaei-Khoei A. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. Int J Med Inform. 2013, 82(1):10-24.
43. Wong K1, Huang SH, O'Sullivan B, Lockwood G, Dale D, Michaelson T, Waldron J, Bayley A, Cummings B, Dawson LA, Kim J, Liu G, Ringash J. Point-of-care outcome assessment in the cancer clinic: audit of data quality. Radiother Oncol. 2010, 95(3):339-43.
44. Baus A1, Hendryx M, Pollard C dentifying patients with hypertension: a case for auditing electronic health record data. Perspect Health Inf Manag. 2012;9:1e. Epub 2012 Apr 1.
45. Schneeweiss S, Seeger JD, Jackson JW, Smith SR. Methods for comparative effectiveness research/patient-centered outcomes research: from efficacy to effectiveness. J Clin. Epidemiol. 2013; 66 (8 supp): S1-4
46. Anne Marie Lyngsø, Nina Skavlan Godtfredsen, Dorte Høst, PT, Anne Frølich, Instruments to assess integrated care: a systematic review. Int J Integr Care 2014, 14; Jul–Sep; URN:NBN:NL:UI:10-1-114794.
47. Mohammed, N., Fung, B. C. M., Hung, P. C. K., and Lee, C.-K. 2010. Centralized and distributed anonymization for high-dimensional healthcare data. ACM Trans. Knowl. Discov. Data. 4, 4, Article 18

(October 2010), 33 pages. DOI = 10.1145/1857947.1857950. http://doi.acm.org/10.1145/1857947.1857950.

48. Corrao G, Cesana G, La Vecchia C, Vittadini G, Catapano A, Mancia G for the Scientific Board; Brignoli O, Filippi A, Cantarutti L for the General and Paediatric Practitioner Board, Merlino L, Zocchetti C, Carle F for Regional and Central Health Authorities. The CRACK programme: a scientific alliance for bridging healthcare research and public health policies in Italy. Epidem Biostat Public Health 2013, 10(3).DOI: 10.2427/8990.
49. Logan JR, Gorman PN, Middleton B. Measuring the quality of medical records: a method for comparing completeness and correctness of clinical encounter data. Proc AMIA Symp 2001:408e12.
50. Berger ML, Mamdani M, Atkins D, et al. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part I. Value in Health 2002;12:1044-1052.
51. Cox E, Martin BC, Van Staa T, et al. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research. Practices for retrospective database analysis task force report—Part II. Value Health 2009;12:1053-1061.
52. Johnson ML, Crown W, Martin BC, et al. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources. The International Society For Pharmacoeconomics and Outcomes Research Good Research. Practices for retrospective database analysis task force report—Part III. Value Health 2009;12:1062-1073.
53. Chen H, Yu P, Hailey D, Wang N. Methods for assessing the quality of data in public health information systems: A critical review. Stud Health Technol Inform. 2014, 204:13-8.
54. vonElm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, and Vandenbroucke JP, The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Annals of Internal Medicine, 2007, 147: 573-577.

*