TECHNICAL REPORT

ISO/TR 22221

First edition
2006-11-01

Health informatics — Good principles and practices for a clinical data warehouse

Informatique de santé — Principes et indications d'exploitation d'un entrepôt de données cliniques

Reference number
ISO/TR 22221:2006(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In exceptional circumstances, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide by a simple majority vote of its participating members to publish a Technical Report. A Technical Report is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TR 22221 was prepared by Technical Committee ISO/TC 215, *Health informatics*.

Introduction

0.1 General

A clinical data warehouse (CDW) is regarded as conceptually distinct from the clinical data repository of an operational electronic health record. It is as yet a largely under-implemented and under-exploited resource which, however, has many possible features with health care, education and research aspects. Such features include:

— quality assurance,

— feedback to individuals and teams of caregivers,

— infectious disease or medication surveillance, and

— evaluation of organizational continuity as patients move between organizations.

Such data are also a crucial link between individual care, organizational and public health needs. The CDW can provide a system view of different perspectives and levels of activity that cannot be provided easily and properly by an operational system; these different levels and perspectives can require different characteristics of the associated datasets.

This data access also has social, legal and ethical, epidemiological and informatics challenges, which may variably impact the use dimensions of a CDW. This will be of particular importance as pedigree and genetic data content of CDWs increases over time.

0.2 Purpose of this Technical Report

The data warehouse is not yet widely used by health organizations. There still is no common knowledge and understanding about the creation and exploitation of data warehouse features by health organizations. The purpose of this Technical Report is to enable the different CDW users to have a uniform understanding of a CDW, including both general principles and particular characteristics of different major use perspectives.

0.3 Benefits of this Technical Report

The CDW is presently a largely under-exploited resource of invaluable information for supporting the service, research and educative missions of the health system. It enables practice assessment as well as knowledge discovery, but it also has the potential to support more efficient and effective innovation, as well as being an essential tool for interdisciplinary collaboration. This Technical Report is intended to help orientate future developments by creating the preliminary work for a technical specification of a clinical data warehouse and leading to the development of standards for different use applications.

0.4 Target users

Target users include all stakeholders in the health system, public and private, including (but not limited to):

— clinicians and para-clinical personnel,

— administrators,

— educators,

— epidemiologists,

— economists,

— researchers,
— system developers,
— data and modelling specialists,
— accreditation organizations,
— citizen organizations, and
— policy makers.

Health informatics — Good principles and practices for a clinical data warehouse

1 Scope

The focus of this Technical Report is clinical databases or other computational services, hereafter referred to as a clinical data warehouse (CDW), which maintain or access clinical data for secondary use purposes. The goal is to define principles and practices in the creation, use, maintenance and protection of a CDW, including meeting ethical and data protection requirements and recommendations for policies for information governance and security. A distinction is made between a CDW and an operational data repository part of a health information system: the latter may have some functionalities for secondary use of data, including furnishing statistics for regular reporting, but without the overall analytical capacity of a CDW.

This Technical Report complements and references standards for electronic health records (EHR), such as ISO/TS 18308, and contemporary security standards in development. This Technical Report addresses the secondary use of EHR and other health-related and organizational data from analytical and population perspectives, including quality assurance, epidemiology and data mining. Such data, in physical or logical format, have increasing use for health services, public health and technology evaluation, knowledge discovery and education.

This Technical Report describes the principles and practices for a CDW, in particular its creation and use, security considerations, and methodological and technological aspects that are relevant to the effectiveness of a clinical data warehouse. Security issues are extended with respect to the EHR in a population-based application, affecting the care recipient, the caregiver, the responsible organizations and third parties who have defined access. This Technical Report is not intended to be prescriptive either from a methodological or a technological perspective, but rather to provide a coherent, inclusive description of principles and practices that could facilitate the formulation of CDW policies and governance practices locally or nationally.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1
clinical data repository
CDR
operational data store that holds and manages clinical data collected from service encounters at point of service locations

EXAMPLE Point of service locations include hospitals and clinics.

NOTE Data from a CDR can be fed to the EHR for that client, such that the CDR is recognized as a source system for the EHR. The CDR can be used to trigger alerts in real time.

2.2
clinical data warehouse
CDW
grouping of data accessible by a single data management system, possibly of diverse sources, pertaining to a health system or sub-system and enabling secondary data analysis for questions relevant to understanding the functioning of that health system, and hence supporting proper maintenance and improvement of that health system

NOTE A CDW tends not to be used in real time; however, depending on the rapidity of transfer of data to the data warehouse, and data integrity, near real time applications are not excluded.

2.3
dashboard
user interface based on predetermined data fields that facilitate domain-specific data queries, and suited to regular use with minimal training

2.4
data dictionary
database used for data that refers to the use and structure of other data, i.e. a database for the storage of metadata

[ISO/IEC 11179-1:2004]

2.5
data mart
subject area of interest within the data warehouse

EXAMPLE An inpatient data mart.

NOTE Data marts can also exist as a stand-alone database tuned for query and analysis, independent of a data warehouse.

2.6
data warehouse
subject-oriented, integrated, time-variant and non-volatile collection of data

NOTE The term "data warehouse" is attributed to Inmon [1].

2.7
drill down
exploration of multidimensional data which makes it possible to move down from one level of detail to the next depending on the granularity of data

EXAMPLE Number of patients by departments and/or by services.

2.8
episode of care
identifiable grouping of health care-related activities characterized by the entity relationship between the subject of care and a health care provider, such grouping determined by the health care provider

[ISO/TS 18308:2004]

2.9
health indicator
single summary measure, most often expressed in quantitative terms, that represents a key dimension of health status, the health care system, or related factors

NOTE A health indicator is to be informative and also sensitive to variations over time and across jurisdictions.

[ISO/TS 21667:2004]

2.10
metadata
information stored in the data dictionary that describes the content of a document

NOTE In a data warehouse context, metadata are data structure, constraints, types, formats, authorizations, privileges, relationships, distinct values, value frequencies, keywords, and users of the database sources loaded in the data warehouse and the data warehouse itself. Metadata help users, developers and administrators for information management.

2.11
online analytical processing
OLAP
set of applications developed for facilitating the collection, analysis and reporting of multidimensional data

NOTE The term "OLAP" is attributed to Codd [3].

2.12
organization
group of people that have their own structure rules and culture in order to work together to achieve goals and/or to provide services through processes, equipments and technologies, etc.

2.13
performance indicator
measure that supports evaluation of an aspect of performance and its change over time

2.14
persistent data
data in a final form intended as a permanent record, such that any subsequent modification is recorded together with the original data

2.15
roll up
method of regrouping and aggregating multidimensional data to move up the hierarchy into larger units

EXAMPLE Weekly count of patients aggregated by quarter or by year.

2.16
secondary data use
use of data for additional purposes other than the primary reason for their collection, adding value to this data

2.17
star schema
dimensional modelling concept that refers to a collection of fact and dimension tables

3 Data warehouse features for a health organization

3.1 General

The roles and capacities of each of the operational databases and informational databases or data warehouses are complementary. An operational database is designed to perform transactions such as adding, changing or deleting a patient. It has a limited capacity for data analysis supporting online care delivery. Secondary data use refers to the exploitation of already existing persistent data. The concept of a clinical data warehouse refers to a set of secondary data for analytic purposes relevant to a health organization. As health care takes place in different organizations, including home care, family practice and care in institutions with different missions, the notion of organization can apply to just one of these entities or to a group of entities, e.g. a regional, provincial or national system of care. An organization uses different data sources, e.g. finance data is usually separate from patient data. For certain purposes, it is appropriate to link finance and patient data to analyse resource use. This clinical-administrative interface is one feature of a clinical data warehouse. A data warehouse can accept data from several different databases, including from other human services organizations such as social services or from technical devices, to facilitate different analyses pertinent for one or more of the organizations. As described in more detail in Clause 7, and as is the case for all data warehouses, there is a preliminary need to address different aspects of data quality prior to its transfer to the data warehouse. This clause describes the use of a clinical data warehouse from different important perspectives.

3.2 Quality assurance and care delivery

The predominant paradigm for quality assurance is a cycle consisting of problem definition, data collection, data analysis, and planning for problem resolution. The step of data collection often depends on searching for this data in a paper record, which is both time-consuming and possibly frustrating, depending on the quality of the record's maintenance. Although the paper record will not completely disappear, at least for some time, with the advent of the EHR and increasing use of electronic data collection, the CDW should dramatically reduce the time for data access and analysis. It should enable quality control teams to return from abstracted data analysis to the original data, to explore and ask related questions to obtain additional data to strengthen the evidence on the nature of the problem. The CDW is also a source of prospective data for monitoring improvement. It can be used to establish trends, identify changes and provide alerts. Knowing in advance the data categories that could be followed over time enables the creation of tailored interfaces, sometimes known as a dashboard, which enable checking of updated data as well as drill down to detailed data for a particular sub-question.

3.3 Evaluation and innovation of health procedures and technologies

An extension of the concept of quality assurance is the assessment of the impact with the introduction of a new technology or a change in procedure. The paradigm for new technology development is a series of steps that start in the research context and move progressively through

— development,

— performance, robustness and safety testing,

— controlled clinical trials, and

— market release and market surveillance.

The CDW has two roles: one at the beginning and one at the end of this process. The CDW is increasingly a source of information on existing patterns of care, and especially the relative importance of particular investigations and treatments. Indeed, it is this process which is under continual examination as part of quality assurance. Companies and research groups can use this information to direct their development choices, selecting areas of testing and treatment where significant improvement might be obtained. At the end of the process, following the introduction of a new technology, the CDW becomes a source of data for surveillance of optimal use and also for evaluation of the impact of its use, as well as unexpected findings. The importance of post-market surveillance for ensuring appropriate uptake and early awareness of unexpected benefit or risk is already well known for new pharmaceuticals.

3.4 Disease surveillance, epidemiology, and public health

The CDW is a rich source of information that can profile communities and assess the health status to assist in planning, expose changes in patterns of care, or trends in use of procedures, or disease profiles including infections. The need for a CDW has been particularly promoted by epidemiologists and health services researchers, who need to understand a population profile of health and disease, aiming for disease prevention and risk minimization, as well as evaluation of variation in population outcomes and their causes. A major impediment is always the access to quality data, and the need to rely on imperfect data from a mix of sources with heterogeneous data organization. It is still common to come across a population data set which provides a clue of disease variation, but where the next step of getting more detailed data that might explain this variation is practically impossible. The CDW should be able to link to data sets or use indicators from other human services organizations, such as justice, education, social services, etc., for public health to analyze population health and related community needs. Depending on access, networking and permission, the CDW represents a new opportunity to delve finely into causes of variation and to link data between intervention and outcome, e.g. to better assess whether a preventative procedure results in improved outcomes. Furthermore, the CDW could be a source of information to understand probability distributions for different health care activities. The patterns could be used to develop simulation models for macro or micro system components to explore different options.

3.5 Planning and policy

Administrative and policy decision making depends on access to objective data, usually in abstracted form. In common with clinical decision makers, there can be a need to explore the data and to move from abstracted to particular data. Abstracted analysed data from the CDW may become a main way that data is shared between decision makers with different roles, such as between clinicians and administrators, and form a basis of negotiation: hence data should as far as possible be clearly presented and interpretable for a given purpose. Health and performance indicators are increasingly used for quality and economic reasons, as metadata can be, describing the way in which the indicators are derived. Their effectiveness depends on efficient data access and continual examination of validity, which can be supported by analysis of related data from the CDW, including comparison across systems of care. Certain abstractions are subject to coding, increasingly using semi-automated methodologies dependent on the quality of primary data. These codes can be available in the CDW.

3.6 Knowledge discovery

As well as providing evidence for quality assurance and to support technology assessment, the CDW using different analytical methodologies could be a source of unexpected new knowledge about disease evolution and treatment response, similar to that previously discussed concerning post-market surveillance. This should most probably be in a sub-population where manifestations are uncommon and the CDW provides the opportunity to analyse these cases in detail in comparison to the population to which the sub-population belongs, a task which was previously very difficult because of variable data quality and access difficulties.

3.7 Education

The CDW is a window on actual health care practice. It is an opportunity to study disease and practice variation, and hence a repository of teaching material of clinical cases and case management that can be correlated to the teaching of best practice. The teaching of quality assurance is variably practiced at present. Being a key resource for quality assurance, including query tools such as the dashboard, the CDW should provide an enhanced quality assurance education environment.

4 Description in detail of each category

4.1 General

The more detailed descriptions in this clause provide an appreciation of the different processes and roles related to the perspectives of CDW use. Security and privacy issues, as well as different analytical tools to support the CDW in these different perspectives, are considered in subsequent clauses.

4.2 Quality assurance and care delivery

4.2.1 Description of business processes involved

Data about patient care are collected as a function of an area of concern identified to or by a quality assurance professional or team. Detailed analysis may lead to a requirement for additional evidence before a correction plan is proposed and adopted. Regular data collection can check subsequently whether the situation is stable.

4.2.2 Sources and sorts of data linked to these processes

Data sources include:

— electronic health records,

— administrative databases (which may already be linked to the EHR source), or

— other institutional databases, such as resource allocation or material costs.

External sources of comparative data could be included.

4.2.3 Role-based use of this data

The principle users of this information are teams responsible for programme quality assurance. Usually such teams are composed of the care providers for that area and their students, possibly with the professional assistance of specialists in quality assurance and associated data and document management. They can use denominalised data for detecting trends. However they may need to be able to identify individual patients and individual practitioners following security and privacy guidelines. Other stakeholders include those persons responsible for institutional quality management, however the access to nominal data should be carefully restricted.

4.3 Services and technology evaluation and innovation

4.3.1 Description of business processes involved

Innovation is a constant process involving the discovery and application of new procedures, tests, equipments, medications and other matters. Innovation may be piloted by a researcher seeking new knowledge, or by a professional expert transferring research knowledge into practice by adopting or adapting published information. It may involve association with industry and the outcome may be subject to government regulation. In many cases, research ethics committee/institutional review board approval of a prepared protocol is required prior to the evaluation, and there needs to be accountable documentation. There is a continuum between quality assurance, practice optimization and innovation. In some cases, the prior evidence of the innovation may warrant its introduction into practice without ethics committee approval, but committee chair approval might nevertheless be sought. An impact review may be required to understand efficiency, effectiveness, safety, quality and cost implications before adopting the product or findings of a research study as part of normal health care delivery operations.

4.3.2 Sources and sorts of data linked to these processes

Specific data is collected according to the study. No innovation occurs independent of a current practice, hence the aim is to show the advantage of the innovation with respect to current practice. If the study database feeds to the CDW, either separately or through the institutional EHR, which is also capable of accepting the results of innovation studies, the CDW becomes a source both of study data and comparative data of the same or similar populations prior to the innovation. The CDW in this scenario might show more clearly the impact of changes to different variables on the outcomes being accessed. The CDW is also a source of information to distinguish the characteristics of sub-populations, which might benefit from the innovation.

4.3.3 Role-based use of this data

This data is of interest to:

— researchers,

— clinical decision makers,

— managers, and

— policy makers responsible for introducing, developing and/or regulating innovation.

The CDW is a source of information on the effectiveness and unexpected risks of the innovation after its introduction into real practice.

4.4 Disease surveillance, epidemiology and public health

4.4.1 Description of business processes involved

Epidemiology is concerned with the health of populations in different settings, and public health includes the larger perspective of overall health of a population. Both disciplines require data not always readily accessible because of availability, different formats or jurisdictional restrictions in order to monitor health, study disease patterns and to measure change over time, and in relation to other major perspectives, such as geography or employment. The CDW offers the opportunity to study the relationships between data, e.g. the relation between antibiotic use and the emergence of antibiotic resistance, and to put into place detection mechanisms to warn if there is a change in pattern. There is a relation to quality assurance in the provision of surveillance procedures for risk detection, such as adverse drug events.

4.4.2 Sources and sorts of data linked to these processes

Data is obtained from both healthy and sick populations and can be collected over long periods of time. Sources include population surveys, information from other human services agencies and the electronic health record at all levels of care and all sectors of health care. Other socio-economic data provide additional information. These questions may be restricted to an organization or regional health system and benefit from an associated CDW. A federated set of CDWs with defined rules of data sharing could support the study and tracking of disease of major public concern, so that early preventative decisions might be taken.

4.4.3 Role-based use of this data

These data are of particular concern to:

— health authorities developing community profiles and population needs assessments,

— institutional teams concerned with infection control and prevention,

— surveillance, community and public health specialists, and

— epidemiology researchers.

The general public is interested in this data particularly in the form of intelligible summaries.

4.5 Planning and policy

4.5.1 Description of business processes involved

Strategic assessment and decision making in relation to organizational mission, vision and values builds objective data into analysis and plan formulation. The CDW can reach different parts of the organization, identifying relationships between events and trends, providing a tool for managers and teams to explore, and suggesting explanations and solutions for different data-based findings. The organization might define CDW-based performance indicators for periodic peer review and determine economic priorities.

4.5.2 Sources and sorts of data linked to these processes

The relationship between clinical and organizational data, both within the organization and externally with its clients, is of particular concern. Certain data are regularly required by regional, provincial and nationwide organizations for health system assessment. The CDW linked to the EHR and other health system databases can provide a care process level of detail and hence more meaningful assessment across these different levels of abstraction. Aggregate/summarized statistics may obscure underlying patterns that only become apparent when a more detailed analysis is done of sorting out sub-components and contributing factors.

4.5.3 Role-based use of this data

This data is used by the following groups and individuals (often in collective consultation and negotiation):

— clinical teams and managers,

— resource and organizational managers, and

— executive teams and councils.

4.6 Knowledge discovery

4.6.1 Description of business processes involved

Knowledge discovery from the CDW is an as yet unassessed source of new knowledge. The greater the quality of data, the better the probability of distinguishing unusual events, including drug side effects, sub-populations resistant to treatment, or rarer patterns of disease presentation and other associations between factors that were previously unknown.

4.6.2 Sources and sorts of data linked to these processes

All data in the CDW might be involved.

4.6.3 Role-based use of this data

Potential role bases for data usage include:

— clinical specialists,

— quality and risk surveillance teams, and

— infection prevention and control and overall health system analysts and planners.

4.7 Education

4.7.1 Description of business processes involved

Different health education organizations can benefit from CDW data from any of the aforementioned perspectives to be analysed to show evidence of real practice. This could be in the form of creating a library of case examples, or it could be a request to a student to prepare material directly from the CDW ready for a teaching session.

4.7.2 Sources and sorts of data linked to these processes

All CDW data could be useful according to the teaching need. Compiled data, graphical representations and full use of the analytical tools, as well as the possibility of creating a library of material are important elements that support the educational process.

4.7.3 Role-based use of this data

The following should benefit from this opportunity:

— clinical specialists and researchers,

— students of the different health professions,

— further education organizations, and

— other educators concerned with the health system.

5 Governance and ethics considerations of clinical data

5.1 General

This clause considers the governance issues of responsible data organization, management and use. Such consideration is important partly because of the intrinsically sensitive nature of personal health data, which require suitable protection of privacy, and partly to ensure that the database contents and the means of interrogating it can be trusted to be fit for purpose and that the results of using it are as scientifically correct as possible. The key to good governance is the identification of responsibilities, the incorporation of good practice within policies, as well as the employment of measures to ensure that policies are followed, audited and reviewed, and where necessary that suitable escalation policies are in place. This clause of the Technical Report identifies a range of good practices that should be included in such policies.

5.2 Governance requirements for data integrity and management

5.2.1 Completeness, preservation of context and longitudinal utility

CDWs are usually constructed with a formal scope that defines the clinical, scientific and managerial domain(s) of interest, sometimes tightly and sometimes quite broadly specified. A CDW should ideally be capable of storing all of the potential classes of clinical or other data that fall within that scope, and not be limited in design to the data structures that are initially envisaged to be collected. Clearly not all CDWs will need to manage images, signals or genomic data. All CDWs should be designed to expand over time to receive data from extra feeder systems or to store additional data items.

Users need to be aware of any known limitations in data storage. These will include known limitations in the source systems providing the data, the extent to which longitudinal and familial linkage is supported, and the currency or otherwise of any semantic links or pointers to knowledge resources. For example, if a drug database is linked to a CDW with prescription data, the name and release version of drugs in that database ought to be available to users.

There is considerable evidence that data collected for one purpose in one setting cannot always be reliably re-used in another. CDWs will most commonly be secondary repositories, fed by clinical, EHR and other systems by a variety of push, pull, real-time and non-live approaches. If CDWs are to support secondary uses successfully and faithfully, they need to preserve as much as possible of the original context in which each data element was acquired. For clinical data, this context is well specified within contemporary electronic health record interoperability standards, since the communication of EHR context is vital for safe shared clinical care.

The following examples illustrate the importance of the original context.

a) Uncertainty expressed about a diagnostic finding must be retained with the diagnosis in the CDW.

b) If a clinical diagnosis was asserted on the basis of a cursory clinical assessment, perhaps for good clinical reasons at the time, this must not be confused with a diagnosis made on the basis of a thorough clinical work-up and/or made by an expert.

c) Proposals for treatment are not always put into practice, and must be distinguished from those that have been implemented.

d) Information about relatives must not be confused with information about the subject of care.

Architects of CDWs are strongly recommended to review EHR-related research and standards, such as Kalra [5], ISO/TS 18308 and ISO 13606-1, in order to identify relevant aspects of context that ought to be incorporated. This contextual information may need to be complemented to provide a clear and consistent basis for the automated aggregation of clinical data in a warehouse. For example, determination of patient morbidity in the warehouse context may be corroborated by looking for multiple consistent diagnoses, or by complementary evidence from lab tests and medication history. Missing context may contribute to wrong assumptions about the data collection and data meaning, as well as a lack of understanding of the policies,