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is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.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,

system configuration and other operational factors that impacted the care delivery patterns and outcomes. However, this contextual information may not always be able to provide a clear and consistent basis for the automated aggregation of clinical data in a warehouse. Determination of patient morbidity in the warehouse may be more easily and effectively established by looking for multiple consistent diagnoses, or by using corroborating evidence from lab tests and medication history, than by parsing the contextual information around individual diagnoses.

Another important aspect of faithfulness is the preservation of original data values. If clinical or other codes are to be applied to textual data in order to add value to its subsequent analysis, the original text expressions should also be retained. If coded values from an originating system are to be stored along with the plain text rubrics for these codes, it is essential that these rubrics are taken from the same terminology, version and language as was used at the time of data entry. This is important so as to permit any spurious analysis results to be examined against the original data. This might mean that copies are taken of local coding schemes and other knowledge resources used in the source data systems. All reference tables from source systems and master codes tables should be loaded into the CDW, and the mapping of these source systems codes to the master codes should be performed before the first load of other data into the CDW. Changes in these reference and master tables should be part of the CDW load as well as transactional data.

If data values are missing, the CDW should ideally be able to indicate a reason for the data being absent, if this is known.

Some CDW data may be acquired through clinical trials. If so, the acquired data should contain references to the source study, as users of the database may need to review the study protocol, including data quality criteria, in order to be able to re-use these data in valid scientific ways.

If possible, the manner in which the CDW schema is constructed should allow for future schema evolution, and make provision for changes in medical knowledge, new terminologies and novel investigations and treatments which may emerge. The social and cultural environment in which clinical care is given (and EHR data is acquired) will certainly change over time, and varies between countries. Such culture does impact the way health and illness are perceived, and the way health care is given and documented. As CDWs are receiving more data from international sources, or are themselves federated internationally, it will become vital to know the source country of all data.

The metadata defining the schema and constraints on the data structures held in the repository should ideally be published in a standardised format such as that used for EHR standards, e.g. as archetypes (see prEN 13606-2).

5.2.2 Longitudinal and familial linkage

Health services are progressively adopting single unique patient identifiers as a means of combining data about individuals across information systems. However, the penetration of these identifiers within individual systems varies considerably: some have retained local identifiers within the local database and only map these to a national identifier on export for specific messages and reports.

Most CDWs will need to acquire data on individuals longitudinally, and will therefore need to be capable of brokering multiple identifiers from different data sources and over time. For example, a repository of cancer care data might need to integrate records from a general practioner, a local hospital, a specialist hospital, community nursing services and a cancer registry system, acquired incrementally over several years. Ideally, there needs to be a system for consistent person identification across systems and services internationally. However, unique identifiers are not a foolproof mechanism to ensure that all records pertaining to an individual are correctly linked to an individual's record of care: a significant amount of effort cleaning data and correcting errors in transactional systems may still be needed to ensure that correct data is associated with the right patient. In practical terms, it will be important to preserve a set of personally identifiable patient information, including such things as name, birthdate, gender, and address, and to establish matching processes and policies using this data. These policies will also need to observe requirements for personal information privacy.

For some kinds of research, it is important to study households or families, and therefore to be able to link records of individuals. There are as yet no agreed standards for representing or communicating such linkage

mechanisms, and the approach taken within any CDW might not therefore be replicable. Nevertheless, it must be recognized that “linkage” data is not stable and is not always documented correctly in EHR systems, e.g.

— marriage relationships and household memberships do change over time;

— new members are added to families;

— adoption status is not always known by young adults.

If a relationship between two individuals is asserted, the source and date of this assertion needs to be retained and be capable of subsequent revision.

5.2.3 Derived data

It may be appropriate to add derived information to original clinical data in order to optimize the database for secondary uses, sometimes known as “data cleaning”. This might be achieved by:

— making inferences from single or multiple data items,

— performing information extraction from narratives,

— abstracting,

— aggregating,

— flagging key milestones, or

— summarising episodes of care.

These derivations need to be managed in an accountable way.

These data need to be distinguished from the original clinical data (e.g. by tagging, but not necessarily stored separately). Data cleaning, re-coding and encoding of free text (derived data) should be stored additionally to the original clinical data and marked as such. These derived data should clearly reference the process (algorithms, reference tables, tools and their versions) used to create them. This processing might need to be repeated as new techniques are developed, i.e. derived data needs to be formally version managed. In general, new data cleaning processes should start from the original clinical data and not from derived data. A similar principle applies to any rounding of dates, times or numeric values.

This accountability is important because, for example, natural language processing (NLP) technologies are constantly improving and it is often critical to know how and when a narrative was parsed, in order to decide if the extracted codes are likely to be complete and accurate enough for a given secondary investigation.

If data cleaning is necessary to ensure data quality, the original data should not be permitted to contribute to analyses until that cleansing has taken place.

5.2.4 Reliability

Emphasis has been put on the importance of integrity from the perspective of faithfulness to the original purposes and contexts in which data were acquired in order to maximise the likelihood that secondary uses can be performed accurately, precisely and safely, and can allow for valid inferences. Integrity also needs to be managed at a technical level. If the CDW is being populated from feeder systems such as EHR systems, intervals when connections to those feeder systems were interrupted need to be known and available to a secondary user, particularly if there is a permanent loss of some data or if the process of deriving additional values has been delayed or skipped. The currency of a database (e.g. when it was last updated) must always be known and made available appropriately to secondary users.

There are a variety of errors that may need to be managed within a CDW, e.g.

— damaged or corrupt data from a feeder system which has been imported and now needs to be physically or logically deleted;

— incorrect mapping of patient identifiers or codes, resulting in false data being added to the repository and possibly used for some period of time;

— errors of recording in an EHR system that are subsequently corrected and are communicated to the CDW as a revision of previously provided data.

CDWs need to be formal version-managed repositories, in which revisions are audited and rollback facilities permit verification of the state of the database at any previous known point in time. Since imports or data cleaning are likely to occur as batch processes, the repository should include information about when the data were added (and optionally a batch identifier), so that a single imported or processed batch of records can be distinguished, if roll back is necessary.

The longitudinal reproducibility of results is also important for scientific credibility, to verify unexpected findings (e.g. before a public health alert is issued), and to enable appropriate audit of secondary users themselves. All changes to the CDW repository must be version managed, and not result in physical deletion from the database unless it is known that this will not have accountability consequences (e.g. running a correcting process before the erroneous data has been used).

The technical means to manage the integrity and availability of a database are not particular to health care or to health research, and are therefore not specifically dealt with in this Technical Report. They include measures such as backup, mirroring, and archiving policies. The reputation of a CDW will be dependent upon its reliability (trustworthiness), but as a general principle, integrity is probably more important than availability for secondary uses (i.e. it is more important to get the right answer to a query than to get any kind of answer rapidly).

5.2.5 Scalability

A clinical data warehouse can become more valued than was originally anticipated, and grow in size, complexity and rate of access. This can create requirements for capacity, distributed access and performance that go well beyond the initial engineering choices or budgets. Clearly it is impractical, and probably not cost effective, to plan for capacity beyond what can be realistically envisaged or budgeted. Additional requirements for capacity and access may be met by alternative systems rather than within the CDW itself. An “extensibility” plan could include export to other systems, and communications with other systems which retain the integrity of the data contained within the CDW.

5.2.6 Custodianship

The legal custodian of a CDW is often the legal organization funding its development or the one on whose premises it is located. It is often, but need not be, the owner of the hardware on which the CDW is located and is also often, but need not be, the employer of the staff managing and maintaining it. These important roles need to be formally defined, and it needs to be made clear who is responsible for creating and enforcing policies that specify how the database should be managed, used and maintained. This may include policies for granting licences or access privileges for secondary use of the CDW, and policies for appointing a governance authority for it. Role definition must include responsibility for developing and maintaining the technical integrity of the database. It has been observed that such policies are often not defined, and that for some CDW projects, it is not clear who ought to be responsible for creating them. In many situations, it is also unclear who should assume responsibility for the CDW when its period of funding has expired.

At a logical level, there might be different “stewards” taking responsibility for different kinds of data content and/or different data sources. Stewards might take responsibility for managing some of the policies described in this clause of the Technical Report. However, it will probably still be the responsibility of the legal custodian to ensure that stewards are appropriately appointed and to take overall responsibility for CDW governance.

The principles by which ethical policies should be defined are discussed below. However, the first step is to ensure that the party with legal responsibility for the CDW is unambiguous, i.e. that a legally recognized entity accepts responsibility to ensure that suitable policies for a CDW are defined and enforced, and that this entity is legally accountable for its management, use and possibly for its subsequent destruction.

There may also be a need for data sharing agreements between organizations contributing data to the CDW and the organization which manages the CDW.

5.3 Perspectives of individual and social protection

5.3.1 Data protection legislation

5.3.1.1 Consent

The importance of protecting the privacy of citizens has been recognized by governments in many jurisdictions. The European Union (EU), the U.S.A. and a number of other countries have introduced, or are in the process introducing, laws and regulations to safeguard personal information by protecting privacy and rights over it. In Europe, a cornerstone EU Directive (Directive 95/46/EC [10]) has led to the passing of data protection laws in all European countries, such as the UK (the Data Protection Act [11]). In the U.S.A., the Health Insurance Portability and Accountability Act (HIPAA) [12] is broadly equivalent in this area.

Access to health record information by those involved in the direct provision of health care to an individual is deemed in most cases to be within the terms of the implied consent that exists when health care is sought. The use of personal health data by care providers is limited to medical purposes, i.e. direct care provision and basic quality assurance, unless specific consent is obtained for additional use. (There are usually particular clauses to permit in addition the disclosure of personal data in the vital interests of the patient or of society.) Medical research is usually considered a legitimate “medical purpose” for which health data may be used, but it is generally held that the use of identified patient data for research ought to be formally permitted through explicit consent. This is particularly important if the data might need to be accessed by persons who are not part of the clinical team delivering the health care through which the data were acquired, as in multi-centre research.

Research and epidemiology studies are also undertaken, and are of great value, when health records are analysed without any intention of influencing the care of those individual patients. New knowledge can be gained by looking for trends, frequencies and other patterns across populations of existing records, or previously collected research data, without requiring the new active participation of the patients themselves.

There are many situations where explicit consent cannot be gained for secondary use of health data. For example, if existing clinical databases are to be mined for novel research questions, it is often not possible to go back to former data subjects to obtain new consent, as the persons concerned might no longer be contactable or might have died.

Most CDWs are established on the premise that they will be used for a wide range of (often unanticipated) secondary purposes, without requiring any new consent from the data subjects whose data are stored in them. However, even if a CDW is populated from, located in and used by a single health care setting, the purposes for which the data will be used are unlikely to be compatible with the implied consent obtained for health care delivery. This is even true of education if a CDW is to be used as a learning resource. The key challenge is how to allow secondary use without prejudicing the privacy of individuals, and in particular without the impractical requirement for fresh consent to be obtained for every new analysis of the data.

In both the EU and the U.S.A. there are specific data protection exemptions for data that are anonymous, i.e. if the data subject cannot be identified from the anonymized data directly and the data cannot be linked to any other publicly available data that helps to identify the subject. This avenue is now regarded as an attractive pathway by which clinical information may be transformed into a suitably de-identified form, and then made available for secondary purposes.

BioBanks are being established in many countries: large populations of volunteers participate in studies to collect longitudinal health data and periodic tissue samples from them. These bioBanks are kinds of CDW, often established with strong governance arrangements and formalised consent processes. BioBanks are

progressively adopting relatively generic consent procedures for their participants, in order to ensure that the long-term secondary uses of the data are unhampered by the legislative restrictions outlined above. Nevertheless, it would seem sensible that these organizations adopt voluntary codes of practice that still seek to minimize the exposure of individuals through the adoption of de-identification measures. Since a CDW exists specifically to support uses of clinical information that are not directly in support of the care of those individuals, algorithms must be developed and available within the CDW – or within Extraction Transformation and Load (ETL) tools – to protect the identify of the subject from disclosure.

5.3.1.2 Privacy

Given the legislation and public concerns about the inappropriate disclosure of health data, and the increasingly expressed wish of the clinical and health informatics communities to widen access to longitudinal clinical data and to make secondary uses of them, it is now necessary to adopt models of good practice.

CDW custodians must recognize that no consent process will perfectly anticipate all future uses and users of the data, and no de-identification technique can be perfect.

CDW custodians and secondary users should adopt policies and measures to minimize the exposure of all identifiable or potentially disclosure-sensitive health data, both internally to the CDW management teams and to its secondary data users.

Any such measures will need to balance this protection against the practical difficulty of de-identifying the CDW, as well as against any obstacles that this introduces to achieving the purpose of the secondary data use according to the societal value of this purpose, and against the financial cost of implementing such measures.

Formal policy should specify how the CDW database and its contents will be handled during and after its project life-time. In brief, this should specify at least:

— WHY the data may be accessed/used,

— WHO is permitted to see the information,

— WHAT classes of data may be accessed,

— HOW the data is protected and accessed, and

— WHEN the data may be accessed, and what should happen to the data afterwards.

To enable personal data to be shared locally, nationally or internationally (for example if a CDW is to hold data from a multinational trial), there needs to be a harmonization of regulatory and health informatics processes. ISO 22857 provides important guidance on this. On the other hand, to take one example, legislative protection of the rights to privacy for deceased persons still varies considerably internationally. This makes it quite difficult to define appropriate policies that enable international secondary uses of a repository of deceased persons unless one adopts the (high) standards required for living persons.

Consideration should be given to acquiring explicit patient consent for secondary uses at the time, with the implication that consent is required for provision of health services. However, if and how this should be handled probably needs considerable debate.

5.3.2 De-identification policies and measures

5.3.2.1 Identifier management and pseudonymization

Most secondary users do not need identifiable information, but they do require longitudinal records that link the various episodes for each patient, preferably derived from real EHR data sources, to enable them to observe patients' histories as they evolve. Some uses are not possible if identifiers are stripped from the data, i.e. a CDW may need to retain person identifiers and demographic data if it is also acting as a primary repository for some studies or other data collection purposes, or if it is to provide feedback on quality of care

to particular health care teams. A careful balance between the benefit of the use of the data and the privacy of the individual is clearly needed.

There is also the occasional requirement to be able to re-identify specific patients in special circumstances, e.g. to effect public health measures, to warn groups of patients of new risks uncovered by research or in order to recruit patients for clinical trials.

Genetic studies are yet another area where linkage may be particularly important. Family studies are increasingly contributing to our understanding of many diseases. Losing the ability to link family members may have a significant adverse effect on the quality of the research. The sensitivities of work in this area require very careful safeguards to be designed into the procedures for such linkage.

It may be helpful to think of CDW repository data as falling under one of three broad categories:

a) **identifiable**, i.e. personally identifying characteristics form part of the database schema, and persons accessing the database might deliberately or incidentally access personal characteristics of the database subjects,

b) **anonymized**, i.e. sufficient personal characteristics and identifiers have been irreversibly removed such that individuals cannot be identified from these data or in combination with any other publicly available data, and

c) **key-coded** or **pseudonymized**: Lowrance [6] defines key-coding as the technique of separating personally identifying data from substantive data but maintaining a potential link by assigning an arbitrary code number to each data-identifier pair before splitting them; held securely and separately, the key makes it possible to re-associate the substantive data with the identifiers, under specified conditions, if that is ever necessary.

Pseudonymization is the general term used for reversible anonymization, and key-coding is the most common way of achieving this. With key-coding, the master table(s) of identifying information are secured by a trusted party, which may, for example, be the local CDW manager, a representative of the CDW custodian, an independent organization, or one of the health care sites providing the data. The CDW custodian needs to define the basis on which a key-holder is appointed, e.g. when an internal key-holder is acceptable and when an independent party must be used.

A significant challenge in any pseudonymization approach is distinguishing the set of characteristics that are personally identifying from those that are required for the secondary data use. Date of birth, for example, might fall into both categories, and a decision needs to be made as to whether this data item is to be part of the secured data with a transformed version (e.g. year of birth) in the CDW database. This challenge is complicated by a need to balance the degree of de-identification with the risk of making the data less useful. One way of addressing this challenge is to provide variable access to data using a combination of database views and the dimensional roll-up of data attributes, e.g. from birth date to birth month or birth year.

Key-coding does not remove the need to define a suitable access policy to the remaining data, such as a classification of data item sensitivity and a mapping of these to specified data users. If practical, identifiable data access should also be audited for staff performing or supporting the data matching, quality assurance, and data transformation processes, while source data is being loaded into the CDW.

Policies need to specify the kinds of parties to which data may be revealed, and specify the authorizations and bodies required to sanction additional releases that are not predefined in the policy. However, the use of key-coding also requires a formal approach to consider re-identification, i.e.

— which team members need identifiable data,

— whether external advisors (e.g. statisticians) need identifiable or de-identified data,

— which secondary users need identifiable data, and

— on what grounds exceptional re-identification should take place (e.g. for public health alerts, clinical trials recruitment).

A key-holder also needs to be audited, to ensure that key reversals are always appropriate and in accordance with agreed policies. A key-management policy therefore also needs to define a process for monitoring and escalation.

The CDW custodian also needs to define policies that specify the future of the key-coded data after the funded life of the CDW, and any subsequent grounds on which database access and re-linkage of individuals may be permitted. Alternatively, the key-holder may be given instructions on how the database is to be archived or destroyed. These instructions need to consider any backups or other copies of the data that might have been made during the course of the data use (these should be systematically logged so that they can be appropriately archived or destroyed), whom is to be the final custodian(s) of the data, and how restrictions on its use are to continue to be applied or enforced.

It should be noted that, in some countries, any personal data that is reversibly identifiable and/or allows linkage is not considered anonymized. This view might then regard such a repository as illegal if patient consent is not held for the data.

5.3.2.2 Anonymization of the data

There is as yet no formal consensus on how to anonymize health information, and no “gold standard” by which any attempt can be judged. The HIPAA legislation in the U.S.A., for example, lists the following information that needs to be removed in order for it to be considered de-identified:

a) names (e.g. data subject, relatives, employer),

b) geographic addresses (e.g. residence, hospital, post code),

c) communication and residential addresses (e.g. post box, telephone, email),

d) dates relating to an individual (e.g. birth date, treatment date),

e) numbers relating to an individual (e.g. social security, health plan, prescription),

f) graphic representations (e.g. fingerprints, photographs), and

g) genetic profiles and investigation results.

Some of these items, such as date of birth or postal district, could not, however, be removed without damaging the utility of the data for secondary purposes. Even with removal of all of these HIPAA-listed items, it is still difficult to achieve complete anonymization, such that no data subject could be recognized, whilst retaining the integrity and value of the clinical data. In this regard, the following are important considerations:

— some nearly-identifying characteristics are very valuable in secondary data use, e.g. date of birth, postal district, ethnicity and occupation;

— some kinds of medical data may be absolutely identifying, such as a facial or body photograph, a voice recording, a genomic sequence;

— much of the clinically-rich data collected electronically today exists in the form of narratives, e.g. letters, reports, free-text boxes on forms, etc., which sometimes mix medical and social information, even within a single sentence; and

— clinical case histories are themselves unique, even if devoid of demographic and social information.

Techniques for anonymizing clinical data repositories derived from health record data are currently the subject of research (e.g. Reference [7]). This includes research on the automated extraction of structured clinical data from narratives, thereby reducing the need to access the narratives themselves during the research.

The CDW data schema needs to be analysed to determine how unique and how sensitive it is to disclose each data item, individually or in combination with others. An individual’s genetic profile, while unique, does

not on its own lead to the disclosure/identification of the individual in a ready manner. An occupation is not often unique, but is more likely to result in revealing whom the subject is when combined with other information in the database. In some repositories, just being recognized in the database confirms that the data subject has a particular medical condition.

Genetic data will prove to be a particular challenge because it is in part shared by other related persons: it may be possible to deduce family linkage within a de-identified database, and from this to infer other characteristics of a family that eventually help identify them.

Ultimately, the concern is the risk of harmful disclosure. Focus should not be restricted solely to data identifying the subjects, but to what might be revealed about the subjects. Data items need to be considered in their social context, i.e. the degree to which a fact can be linked back to an individual or through which the person becomes recognizable, and the potential harm that may ensue if the fact is thereby revealed. The impact of a disclosure is increased if socially embarrassing information is revealed. Pregnancy status may not be particularly unique or disclosure-sensitive (given that many female members of the population are pregnant at any one time), but if a person is identifiable in a database, the revelation of their pregnancy status may be personally harmful or upsetting.

For this reason, anonymization techniques must be regarded as part of the protection, and not the only protection, offered to individuals in respect of their personal health data. It is wise to consider anonymized data as if it still retains some small risk of re-identification, and to still take steps to minimize and partition access to the data.

5.3.2.3 Data item sensitivity

One approach of reducing the risk of inappropriate disclosure is to identify and limit access to those data items within the CDW database that are either relatively identifying or are likely to be considered harmful by the data subjects if they were to be revealed. Specifically, a simple hierarchical classification of sensitivity could be developed, perhaps comprising a limited number of categories (i.e. two to five), and this could be mapped in a matrix to secondary user organizations and staff on a need-to-know basis. It may be possible to design the overall database schema, and access control to it, such that only a few key people will access the data items of the highest sensitivity. This classification could contain a category of data that is not sensitive or that is purely statistical, to be made generally available.

5.3.2.4 Masking on extraction

A number of CDWs might provide access to a wide range of user groups working in different settings, on different projects or on different aspects of a complex single study. Not all of these secondary data users will need record-level detail: some may need to run frequent queries on fine grained values but not see the full data set on any one individual. Where these queries include the more sensitive data items, it may be possible to mask such data items in the result set even if they cannot be masked in the raw data. Masking involves some transformation of the data values to make them less distinctive, such as rounding numeric values to a few significant digits, shortening a postcode to postal district.

5.3.2.5 Statistical disclosure control

Statistical disclosure control (SDC) is a branch of mathematics/informatics dealing with calculating the probability that a particular data set may result in an unintended disclosure, e.g. of the identity of an individual or of a collection of facts that reveal attributes about a small group of persons. In situations where a CDW has secondary users that require detailed result sets, fine grained data values, linkage between result sets or the running of regular queries over time to look for changes, such a methodology should be considered. SDC might be applied as a risk assessment tool to review the database schema and its contents in order to recommend which data items should be considered most sensitive, or it may be implemented as a monitoring component within a CDW query interface in order to examine and approve incoming requests for data and to check result sets before they are released.

5.4 Policies about people

The skills, attitudes and commitment of the people involved in managing, maintaining and using a CDW are at least as important as the policies and technical measures used to protect the database and the privacy of its data subjects. This is widely acknowledged across all sectors, and investing appropriately in the people who will come into contact with a CDW is essential.

The training of staff who will handle CDW data involves the organization, management and supervision of both new and existing staff, whether they are processing data intentionally as part of the internal management of the warehouse, accessing it incidentally or periodically for either analysis or for technical support, or using it as part of data quality assurance or research, whether in-house or external.

Secondary users also need to train staff to use the CDW appropriately, to respect their access privileges and authentication mechanisms. These users need to recognize that even if the data they are receiving as a result of a mining query is aggregated or de-identified, these measures are not perfect and they must still properly use the data they receive.

It may be that with some CDWs holding quite sensitive information, or in situations where secondary users need quite detailed results, a formal process of accreditation is required so that such users can demonstrate organizational and individual staff competence. This should include audit and escalation policies, with evidence that policies are in effect, monitored and reviewed.

Policies for people also should be accompanied by clearly defined sanctions that will be put into effect for deliberate breach or carelessness.

In a distributed environment, such as a CDW with multiple stakeholders and widely distributed users, the CDW custodian must ensure that lines of accountability are clear and adequate, since the ultimate responsibility for all use, and in particular for unintended disclosures, will otherwise lie with the custodian.

5.5 Security review and audit

5.5.1 General

Policies risk being defined in good faith, communicated via training to staff when a CDW project is being set up, then gradually forgotten to “make life easier”, with a consequent erosion of good practice.

Policies should therefore not be so stringent as to be impractical to adopt. They should also be regularly reviewed to ensure that they are appropriate to the current state of the CDW, address relevant risks, and remain in keeping with legislation and professional or other guidance.

Audit is an essential and ongoing part of maintaining good practice in the same manner that clinical audit is known to help maintain quality of clinical care. The CDW custodian or an appointee should develop an audit plan and review its enforcement and results regularly. The precise specification of the manner in which such audits might be performed will depend upon many local factors and the nature of the CDW being protected.

CDW audit should include reviews of general patterns of data access and use, specific re-identification of pseudonymized data, security management processes and practices, and processes related to data quality and integrity.

5.5.2 Technical security measures

The approaches and policies described in this clause of the Technical Report will need to be complemented by other policies and procedures designed to protect the CDW itself and its interconnections. These technical security measures are not specific to a CDW repository server and its networks: most clinical and EHR systems have similar requirements, which are documented in other existing and forthcoming standards. The most pertinent of these is ISO 27799. This aspect of good practice will therefore not be discussed further hereafter, apart from the affirmation that a CDW needs to be protected to the same security standards as any clinical data repository or EHR system. Some further security policy guidance is given in Clause 7.

6 Architecture

The architectural underpinning for the CDW is substantial and can be drawn from best of breed industry practices from within the domain of data warehousing and business intelligence. The application of the unique characteristics of the CDW in terms of architectural facets framed within a background of generalized data warehouse practices, and how to accommodate these within a health information context, is the topic of this clause. The approach here is to, where possible, reference and draw on established effective industry practices within the data-warehousing arena.

6.1 Existing work on data warehousing

There are currently two primary methodology leaders within this field: the methodologies and guidelines as espoused by Inmon [1] and Kimball [2]. Over the last ten to fifteen years, these methodologies have established a solid core of successfully deployed data warehouses. While other methodologies are available [4] these two approaches, also known respectively as "Hub and Spoke Architecture" (Inmon [1]) and "Data Mart Bus Architecture with Linked Dimensions" (Kimball [2]), are predominant. The key requirement for any CDW is to be able to accommodate the highly dimensional and complex nature of health care data and its associated analysis. Both these methodologies facilitate this requirement. It should be noted that much of the architectural underpinning described below is drawn from these two sources.

A "data warehouse" (see definition in 2.6) has also been described as either a "subject-oriented, integrated, time-variant and non-volatile collection of data" (Inmon [1]) or, alternatively, as a "copy of transaction data specifically structured for query and analysis" (Kimball [2]). Kimball's definition is further expanded to define the data warehouse as a collection of integrated, atomic, dimensional data marts, glued together by an architectural bus composed of commonly defined linkage elements ("conformed dimensions").

NOTE 1 For further details on dimensional modelling, see 6.2.3; for further details on star schema, see 6.2.3.3.

In contrast, "hub and spoke architecture" is composed of an integrated, normalized, relational and atomic hub surrounded by "dependent data marts" [1)] drawn from this hub. Access to data is through these data marts (and on occasion directly against the hub) [2)].

NOTE 2 For further details on normalization, see 6.4.8.3.

While the practice of data warehouse architecture is somewhat polarized between the approaches of Inmon [1] and Kimball [2], there is a growing sense that compromises and trade-offs between them are useful, and that they complement each other in many ways.

It should be noted that data marts are subject area specific. Clinical examples [3)] include:

— drug claims,

— inpatient discharge data,

— health expenditures, and

— any other data that could respond to user requirements.

1) "Dependent data marts" receive their data from a centralized enterprise data warehouse (hub).

2) While less prevalent, other architectural configurations are found (see Reference [4]), e.g. access to the hub in a similar manner to the method above without recourse to a dependent data mart, or direct access to an operational data store (ODS), which may handle both query and transaction processing.

3) The examples of data marts given in 6.1 are very specific to data types, and possibly data sources. In addition to basic marts of discharge data, medication data, etc., it may be desirable to build derived or consolidated marts for specific areas of analytic focus, such as chronic disease patient groups like diabetics. Each of these derived marts, which may exist only for purposes of a limited research project or may be maintained on a long-term basis, is made up of data drawn from a variety of the type- or source-specific marts.

Conformed dimensions are the descriptive attributes that span more than one subject area and are defined in a standardized manner. The common conformed dimension examples in the CDW would be a subject of care (e.g. patient, recipient), provider, and service delivery location (e.g. hospital, facility). By defining a standard for these common dimensions, (conforming) discrete subject area data (e.g. inpatient discharge data and costs) can be linked together, thus expanding the scope of questions addressable in the CDW.

6.2 Characteristics of a clinical data warehouse

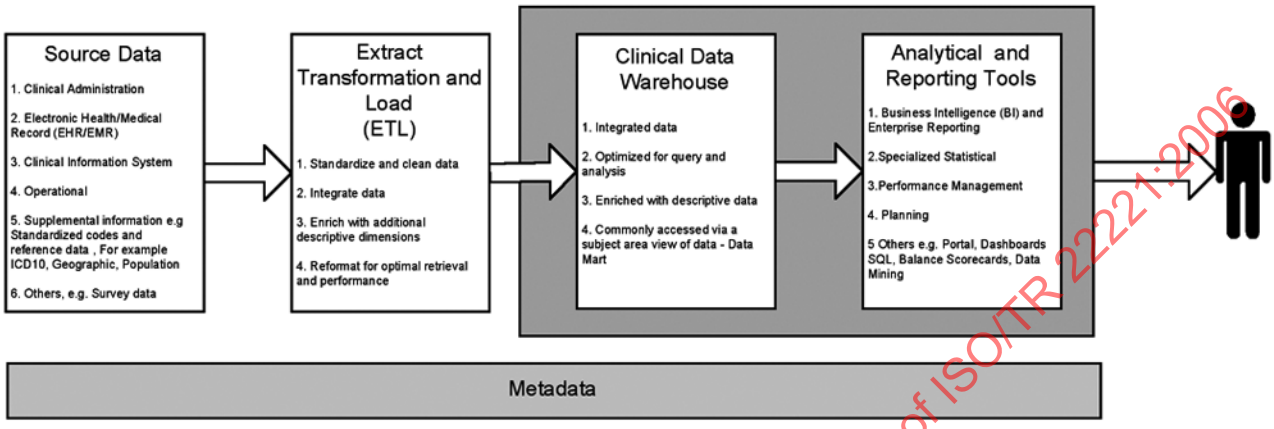


Figure 1 — CDW overview

6.2.1 General characteristics

Operational clinical systems are concerned mostly with the individual subject of care transactions, such as the point of care and administrative functions, whereas the predominant function of a CDW is at an aggregate level. This does not preclude individual-level longitudinal or trending analysis, but rather provides a broad differentiator in the role of a CDW as a secondary system. As with generalized data warehouses, the CDW encompasses data that has been drawn from a variety of operational sources and enriched by supplementary value added data. Specific requirements of the CDW stem from health care's rich dimensionality, broad subject areas, high volume, requirements for comprehensive linkage (particularly across the continuum of care) and the necessity to hold data at an atomic level to fully allow often complex health analysis to be satisfied. The complexity and critical nature of clinical data are also considerations.

The term “data warehouse” can often confuse a non-technical user because of the implication that one is dealing primarily with the arrangement of data. The terms “analytical environment” or “analytical portal” can be used as a substitute for “data warehouse”. These terms more fully express the scope of the product, which is to effectively deliver health care information to service a wide range of decision-making and research questions. To achieve these objectives, a complete solution is required, composed of database (data warehouse, data mart), business intelligence (BI) and analytical tools, metadata, interfaces (e.g. *ad-hoc* query builder, interactive reports, dashboards, portal), together with educational material.

From a data perspective, the objective of the CDW is similar to that of any other data warehouse in that it is structured specifically for effective query and analysis purposes. Its key characteristics might be categorized as indicated below.

— It is **integrated** and **standardized**. Data is gathered into the warehouse from a variety of sources and merged into a coherent whole with multidimensional components. This process is primarily achieved through standard definitions of key common dimensions, notably subject of care (patient), provider, service delivery location. Conformance and normalization in the use of codified values, e.g. in classifying interventions and diagnoses, is also essential to data integration and standardization in any data warehouse.

— It is **time-variant**. All data in the data warehouse are identified with a particular time period and/or episode of care. Data are in most instances non-volatile.

— Its **performance** is important. Data access is optimized both to reduce input/output and support analysis (i.e. through the use of indices, star schemas, parallelism, etc). Health care data can be very high volume, e.g. drug prescriptions data, and therefore the CDW must be able to deliver information in a timely manner in order to be effective.

— It is **subject-oriented**. Data perspectives are on information about a particular subject, instead of information on the multiple transactions associated with the operational context of the primary data. The dimensions reflecting the different structures and functions of the organization support intuitive manipulation in the analytical process. This also allows for complex measurement questions across the subject of care continuum of care to be answered, without an impact on pre-existing legacy operational systems.

— It has **atomic data**. Although the majority of queries on the CDW will involve aggregation, the complex analysis demands of health care data require data to be stored as an atomic grain, as opposed to being stored as an aggregate. This fact precludes the use of certain data warehouse methodologies.

6.2.2 Data sources

6.2.2.1 Primary sources: operational clinical systems

Primary data acquisition in the CDW is mostly from operational systems within health care organizations. As mentioned above, operational clinical systems are mostly concerned with individual subject of care transactions and are primarily for “front line” support or administrative purposes.

6.2.2.2 Secondary sources: use of third party or supplementary data

In general, data sources should be identified with regard to user requirements and the anticipated use of the data warehouse. In order to better leverage data originating within the organization (source data), third party data sources are often required. Examples of typical third party data include:

— population data for the calculation of crude and/or standardized rates, and

— geographic data that allows clinical data to be aggregated at levels of geography not found in the source data.

Third party data sources are highly desirable, as they can be used to leverage the source data, allowing for analyses that would not be possible otherwise. Population health analysis is at present highly dependent on data from the jurisdiction’s census organization that can provide, in addition to raw census data, information such as vital statistics, e.g. mortality, which when linked to primary data allow powerful measures of clinical outcome. Geospatial analysis is made possible with geo-referencing primary source data, allowing powerful insights into, for example, disease distributions and patterns.

In order to incorporate these third party data sources into the CDW, there needs to be one or more common attribute(s) between the source data and third party data. For example, if population data are available at the level of a postal code, then the postal code (i.e. with respect to the subject of care and/or facility) also needs to be available in the source data. A thorough understanding of the intricacies of the third party data is required so that it can be appropriately incorporated into the CDW. This incorporation into the CDW would often be done during the ETL process.

6.2.3 Dimensional modelling

Dimensional modelling is the key technique by which data structures within the data warehouse are designed and, to a degree, implemented.

NOTE See 6.4 for a brief description of alternative approaches to dimensional modelling in the CDW.

A dimensional model separates descriptive elements or dimensions and facts (sometimes called measures) into dimensions and fact tables respectively. This division allows a data structure to be:

— highly descriptive,

— easy to understand,

— quick (good query performance),

— easy to change,

— simple, and

— flexible, allowing integration of separate data from legacy operational systems.

6.2.3.1 Self maintainability

Data sources change over time. Changes in data sources imply maintenance activities in the CDW. The ETL components of the CDW should handle some of the data source evolutions, e.g. adding new laboratory analysis to source databases should be automatically updated in the CDW. It is also desirable to have mechanisms that create alerts for conditions that require manual intervention, such as data quality issues associated with missing values, or mapping issues across codes, etc.

6.2.3.2 Fact and dimensions

As stated in Reference [2], a fact table is “the primary table in each dimensional model that is meant to contain measurements of the business”, and “the most useful facts are numeric and additive” [4]. A dimension table is “one of a set of companion tables to a fact table”. Kimball [2] goes on to further describe a dimension table as a “table containing attributes that are the basis for constraining and grouping within data warehouse queries”.

EXAMPLE 1 Dimensions:

— subject of care (patient),

— provider,

— location (hospital, subject of care),

— diagnosis,

— intervention,

— date (admission, discharge).

EXAMPLE 2 Facts/measures:

— length of stay (mean, median and total),

— procedure time (mean, median and total),

— number of cases.

NOTE Age is an example of a element that can be both a dimension (e.g. group by age) or a fact (e.g. average age).

4) The requirements for aggregate measures in the CDW often go well beyond simple sums, totals, and means. For instance, “complication rate” may be a desired aggregate measure for a given intervention type or group, and this measure would be a more useful basis for comparison if it were standardized by age, gender, morbidity, or a combination of these. While it is desirable that the dimensional structures within the CDW support measures of this complexity, this is beyond the capability of many data access and analysis tools that make use of the dimensional model. This is improved to a degree by recent advances in the functionality of leading BI tools, which are allowing increasingly more sophisticated queries of dimensional structures to be resolved (e.g. star schema, etc.).

6.2.3.3 Star schema

A star schema is a dimensional modelling concept that refers to a collection of fact and dimension tables. Star schemas are commonly the foundation of data mart design, although some methodologies do allow for the direct use of third normal form data (see 6.4). Star schemas are highly descriptive, easily maintainable, perform well with high data volumes and are often deployed directly in relational database systems. To allow effective analysis in the CDW, the comprehensive linkage of multiple star schemas is required. This is often achieved through a constellation arrangement that is essentially a collection of star schemas linked together through common elements (e.g. subject of care, provider). This effectively allows drill across the continuum of care. A snowflake schema is an arrangement used to specifically aid in performance in high cardinality dimensions, such as the subject of care geography.

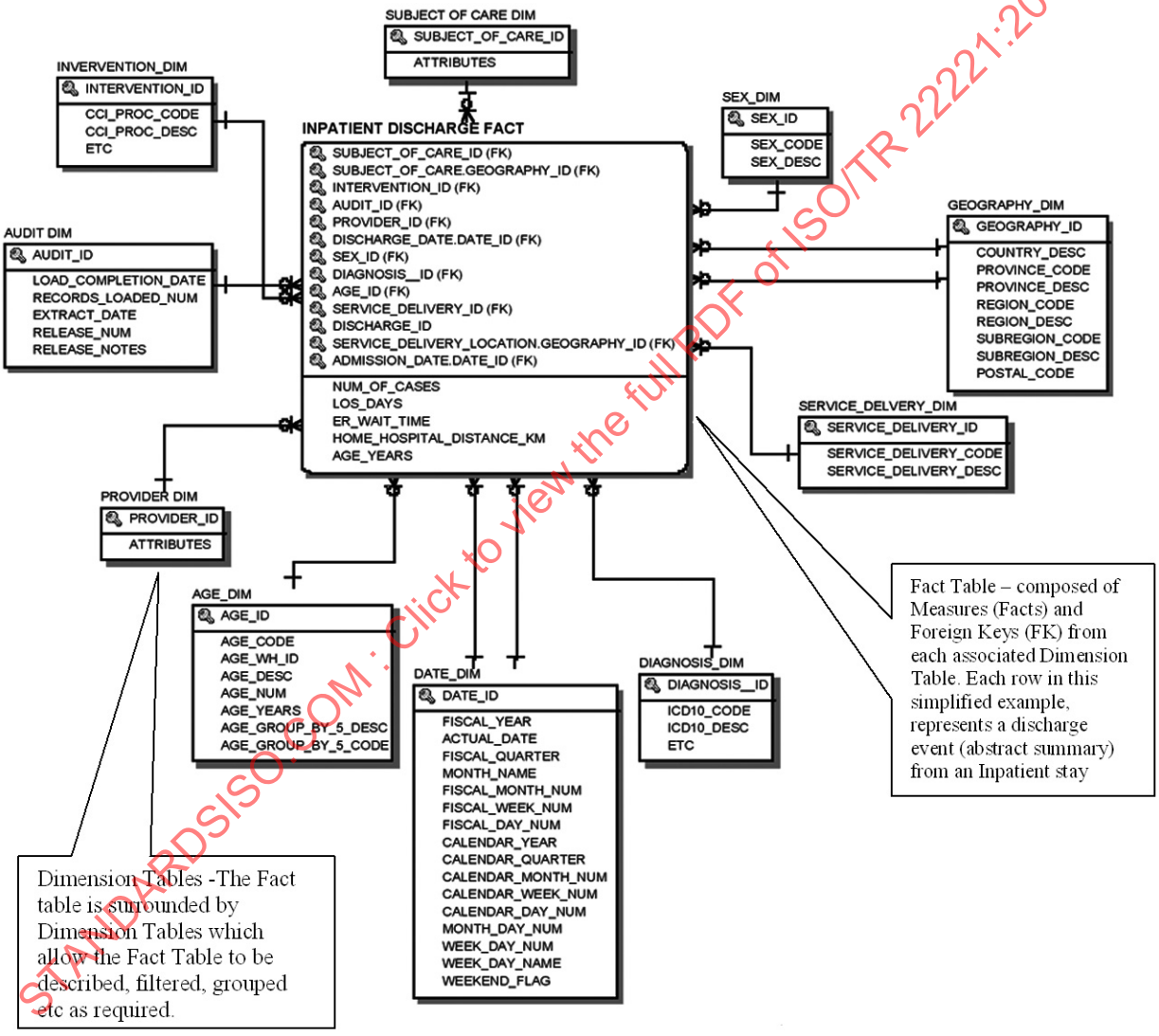


Figure 2 — A simplified example of star schema for inpatient discharge data

6.2.3.4 Surrogate keys

It is a generally accepted dimensional modelling convention that the primary key of a dimension table (the column that uniquely identifies each distinct natural key) be a meaningless number, sometimes referred to as a "meaningless but unique number" (MBUN). The reason for this is that a surrogate key greatly simplifies maintenance if the natural keys should ever change their domain values [5].

6.2.4 Declared grain

The term "grain" is used to describe the lowest level at which data are captured within facts and dimensions. For example, if a date dimension table had its lowest level of data as a single day, then that would be the declared grain of the dimension. The same concept also applies to fact tables as well. In order to answer any key questions using the CDW, fact table data must be atomic, in most instances, and able to create links between themselves, i.e. they must contain rows that represent a single transaction of interest, e.g. a row for each inpatient discharge transaction. In a practical sense, accurately declaring the grain of a fact table is a critical step in the design of any dimensional model. Ultimately, the declaration of grain is directly related to the questions that the model is designed to answer. Furthermore, this is dependent on the facts or measure one is interested in. To take our simple example of an inpatient discharge, this may involve many facts at different grains.

EXAMPLE 1 For inpatient discharge, example facts are:

— number of discharges [6],

— length of stay,

— distance travelled from home to facility.

The grain is: one row for each inpatient discharge transaction.

EXAMPLE 2 For inpatient discharge intervention, example facts are:

— number of interventions (procedures),

— intervention length (minutes).

The grain is: one row for each intervention associated with an inpatient discharge transaction.

EXAMPLE 3 For inpatient discharge prescribed medication, example facts are:

— number of prescribed medications,

— dosage,

— frequency.

The grain is: one row for each prescribed medication associated with an inpatient discharge transaction.

With each of these examples, the grain of the fact table is determined by identifying foreign keys associated with primary keys from dimension tables that are in context when working with fact data at a particular grain.

6.2.5 Conformed/common dimensions

As the CDW grows and data marts are added, there is usually an opportunity to create dimensions that are common to more than one of the data marts. These common dimensions (CD) act as a "data bridge" between fact tables, allowing queries to be written that leverage data from multiple facts. Typical examples of common dimensions include date and location.

---

5) In certain circumstances, extraction into statistical analysis tools, spreadsheets or other tools that are not designed to exploit star schemas may require, as an exception condition, the inclusion of natural keys within the fact tables to facilitate query following the extraction process.

6) This is a "factless fact" (see Reference [2]). This fact can be determined by a SQL COUNT operation or by defaulting this measure to a value of 1 for each row. Here the number of discharges are being counted.

In some cases, it might be necessary to standardize the codes within a common dimension such that it is equally applicable to multiple data marts.

6.2.6 Verbose codes and descriptions

Where available, codes should be paired with their descriptions. CDW customers find these descriptions very valuable, often choosing to report by descriptions rather than codes. This ease of use point should be placed in context, and supplemented by the usage of consistent standardized codes (e.g. ICD-10, SNOMED CT) within the CDW. The objective would be to balance the rigor of code-only filtering with the ease of use of textual descriptions. Much of this is dependent upon the choice and capacity of the analytical tools used to query the CDW [7)].

6.2.7 Multi-language support

For those CDWs requiring the ability to query and report in more than one language, dimensional modelling makes it very easy to incorporate code descriptions from many languages. Moreover, some of the more sophisticated BI tools make it fairly simple to incorporate multi-language support.

6.2.8 Hierarchies

In the context of dimensional modelling, a hierarchy is a powerful concept that allows users to analyse data along a predefined continuum, e.g. in a date dimension, a typical hierarchy would be year-quarter-month-week-single day-hour-single minute. Such a hierarchy would allow users to aggregate data at any one of these levels and drill down or roll up along the hierarchy as desired. “Drill across”, which involves traversing dimensions or subject areas (e.g. in tracking the subject of care groups from emergency to acute to rehabilitation services), is also necessary and will be dependant to a degree on the analytical or BI tool employed.

6.2.9 Roles/re-use

In dimensional modelling, a single physical dimension may often take on several logical roles. A good example would be a date dimension, i.e. a CDW probably collects several dates [8)] (date of admission, date of discharge, date of birth, etc.). Through the use of database views, it would be possible to re-use a single date dimension several times by simply creating one view from the date dimension table for each date in the data mart. Each dimension would have exactly the same contents, the only difference being the context in which the date is collected (i.e. date of admission as against date of discharge). Another example of a dimension which can take on multiple roles is location. In this case, typical roles could be:

— the place where the facility is located,

— the place where the subject of care lives, and

— the place where treatment was delivered.

6.3 Methodology for clinical data warehouse development

6.3.1 General

As stated in 6.1, many of the methodologies and techniques for the development of a successful CDW are already well established by Bill Inmon [1] and Ralph Kimball [2] in their efforts to quantify best practices around data warehousing.

---

7) Certain “best of breed” BI tools permit the binding of codes and their associated descriptions at a metadata level, removing the need for explicit combination fields.

8) The roles that a date or time dimension may play in a CDW are rich and varied, and the examples given are not exhaustive.

6.3.2 Assessing business requirements

The data warehouse life cycle requires that relevant expertise be involved in determining user requirements, data sources, security and quality requirements, in the data analyses and in the recommendations based on the analyses. This coincides with needs for team and interdisciplinary decision making and for life-long learning.

Health care organizations responsible for the delivery of services often tend to have non-standard processes for data collection, administration and analyses. The challenge with this type of organization is often standardizing the various data sources and analysis requirements into a set of user requirements that can service the organization as a whole (see Clause 9 for details).

6.4 Basic data models

Existing examples of CDWs suggest common themes in terms of arrangements of data. As mentioned above, normally these would be modelled and physically deployed as a dimensional model (star schema, constellation [9], or similar). Standard entity/relation (E/R) models, including object-oriented variants [10] normalized to third normal form [11] or above, can also be used throughout the design and implementation phase of a CDW development, particularly the conceptual and logical stages. The degree to which E/R modelling will be used will depend on the methodology followed during the development of the CDW. In most instances, the final modelling and deployment in terms of a data mart accessible for query purposes is dimensional in nature [12]. The more common themes found in CDWs using dimensional model examples are outlined in this subclause.

6.4.1 Common characteristics of clinical data

Four of the most common examples [13] of subject areas, or centric views, associated with clinical data found within CDWs, are as follows:

— clinical administrative and operational [subject of care (patient) centric/service event], including the EHR,

— health expenditures (cost-centric),

— health human resources (provider-centric), and

— population health (population focus/census/geospatial).

6.4.1.1 Clinical administrative and operational (subject of care (patient) centric/service event)

Clinical administrative and operational data in a CDW is primarily centred on what is sometimes referred to as a “service event”. This may be loosely defined at its most granular level as the delivery of health care service(s) to a subject of care by a provider at a point in time with an expected result [14], e.g. this may be

9) A constellation is a collection of star schemas linked together by common dimensions.

10) For example, a unified modelling language (UML) class diagram.

11) Third normal form (3NF) is a data normalization step in which dependencies on non-key fields are removed (i.e. non-key fields are mutually interdependent) and all non-key fields are dependent on the primary key of the table.

12) Certainly, specialized requirements may require non-dimensional structures. In these instances, a common approach is to extract from the final dimensional model/star schema to alternative structures, e.g in the case of creating flat structures for use by specialized statistical tools, etc.

13) There are clearly other examples, such as environmental data, specialized research data, genomic data, etc. These are included for illustrative purposes only and are not at all exhaustive or definitive. A more detailed examination of CDW reference models may be considered as a future topic.

14) Note that this is only a loose definition: a detailed definition of the variety of temporal events (events, service events, episode of care, etc.) and specialized groupings (cohorts, etc.) may be better addressed in a future report on CDW reference models, in which sufficient depth and coverage may be given to these complex areas.

recorded at the point of discharge from an inpatient setting. The “service” in this instance is the inpatient event, including interventions, diagnoses, etc., and the “point in time” is admission and discharge dates. Each row in the fact table is defined by this service event. Joining service events across the continuum of care is supported via typing through a service dimension or a separate fact table for each service event type (e.g. inpatient, outpatient, emergency, etc.) [15]. This type of subject area is primarily subject-of-care-centric though it also involves providers (and service delivery locations) to a lesser degree. In a sense, it is the combination of these common dimensions that ultimately describes the service event in addition to the date/time and other dimensions mentioned.

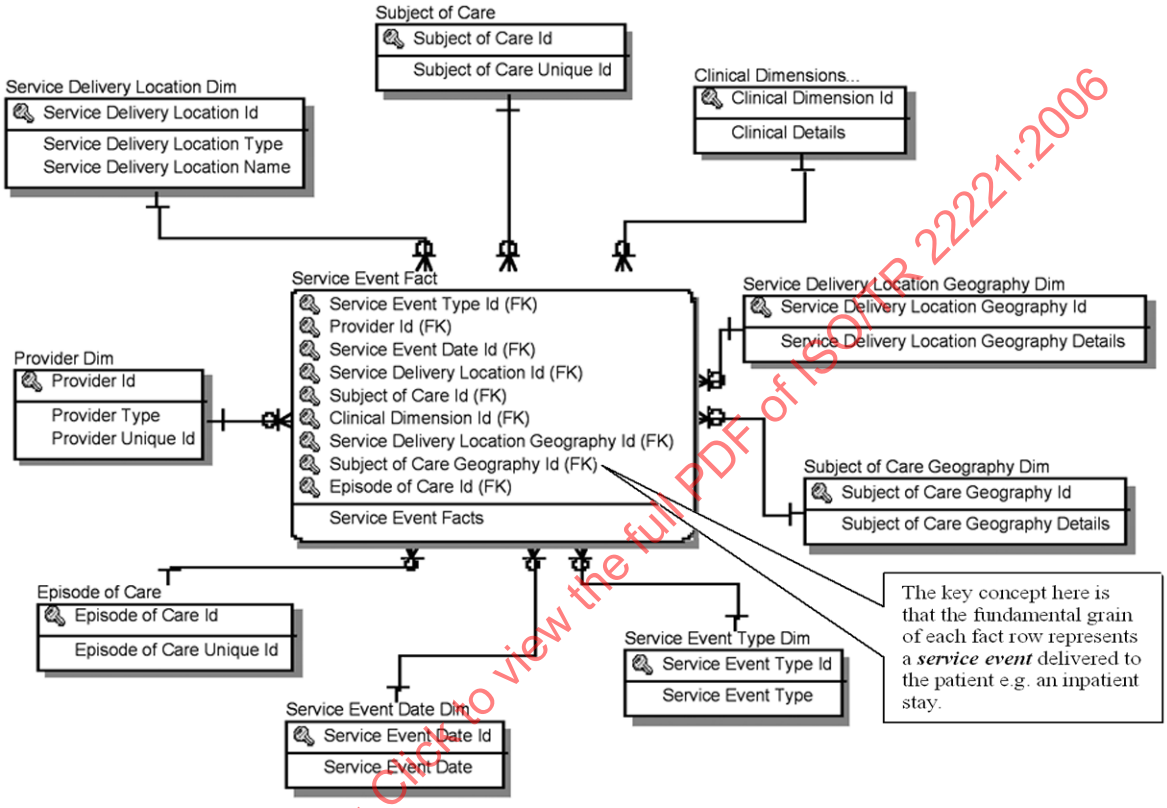


Figure 3 — Service event: example of core dimensional model

The following informally describes the dimensions and facts in the generic illustrative model in Figure 3:

— Subject of Care dimension: synonymous with patient dimension, describing a subject of care;

— Service Delivery Location dimension: loosely synonymous with location or facility, describing the location where the service was delivered;

— Provider dimension: describes the provider of the health service;

— Episode of Care dimension: a grouping or sub-grouping of fact rows associated with a subject of care and provider, usually determined by a specific methodology;

— Service Event Date dimension: describes the date/time at which the service was delivered;

— Service Event Type dimension: describes the types of service delivered (e.g. inpatient, outpatient);

15) An alternative design approach, which allows linkage across the continuum of care in this instance, would be to use an SQL UNION statement containing individual fact tables.

— Subject of Care Geography dimension: describes the geographical location of the subject of care/patient;

— Services Delivery Location Geography dimension: describes the geographical location of the service delivery location;

— Clinical dimensions: clinical descriptive dimensions (e.g. interventions, diagnosis as necessary to analyze the fact data); and

— Service Event fact: measures the occurrence of a health event associated with the delivery of a health service, e.g. the discharge of an inpatient.

6.4.1.2 Health expenditures (cost-centric)

Individual transactions at this level often revolve around costs associated with "functional centres". Patient billing systems are an exception to this (i.e. they are dependent on the service event). Independent examples include management information systems (MIS) which are used for a variety of fine grain cost measurements, (e.g. accountability reporting by managers for resource use), resource allocations, etc. Macro health expenditure, as in the case of national health expenditure, is another example of this perspective. Both of these examples are primarily independent of the service event.

6.4.1.3 Health human resources (provider-centric)

The health human resources focus is commonly on the health provider.

As in the example above, this can be both dependent and independent of the service event. An example of this view is a national database of registered nurses and their distribution, skills, etc. (independent of the service event). An example dependent on the service event would be a national database of physician activity.

6.4.1.4 Population health (population focus/census/geospatial)

A population health focus requires a model that is centred on population census, geospatial and survey data. The example below (see Figure 4) links a simplified population star schema to the example service event schema in Figure 3. Two additional common dimensions, age and sex dimension, have been added to the service event model and, along with the subject of care geography dimension, these link together service event and population data. The common dimensions (sex, age and subject of care geography) are indicated. This particular example (a constellation schema) would allow measures from both fact tables to be available in a query. For example, this could allow crude and direct age standardized rates to be calculated at different geographic and age grains for different subject of care groups from the service event star schema.

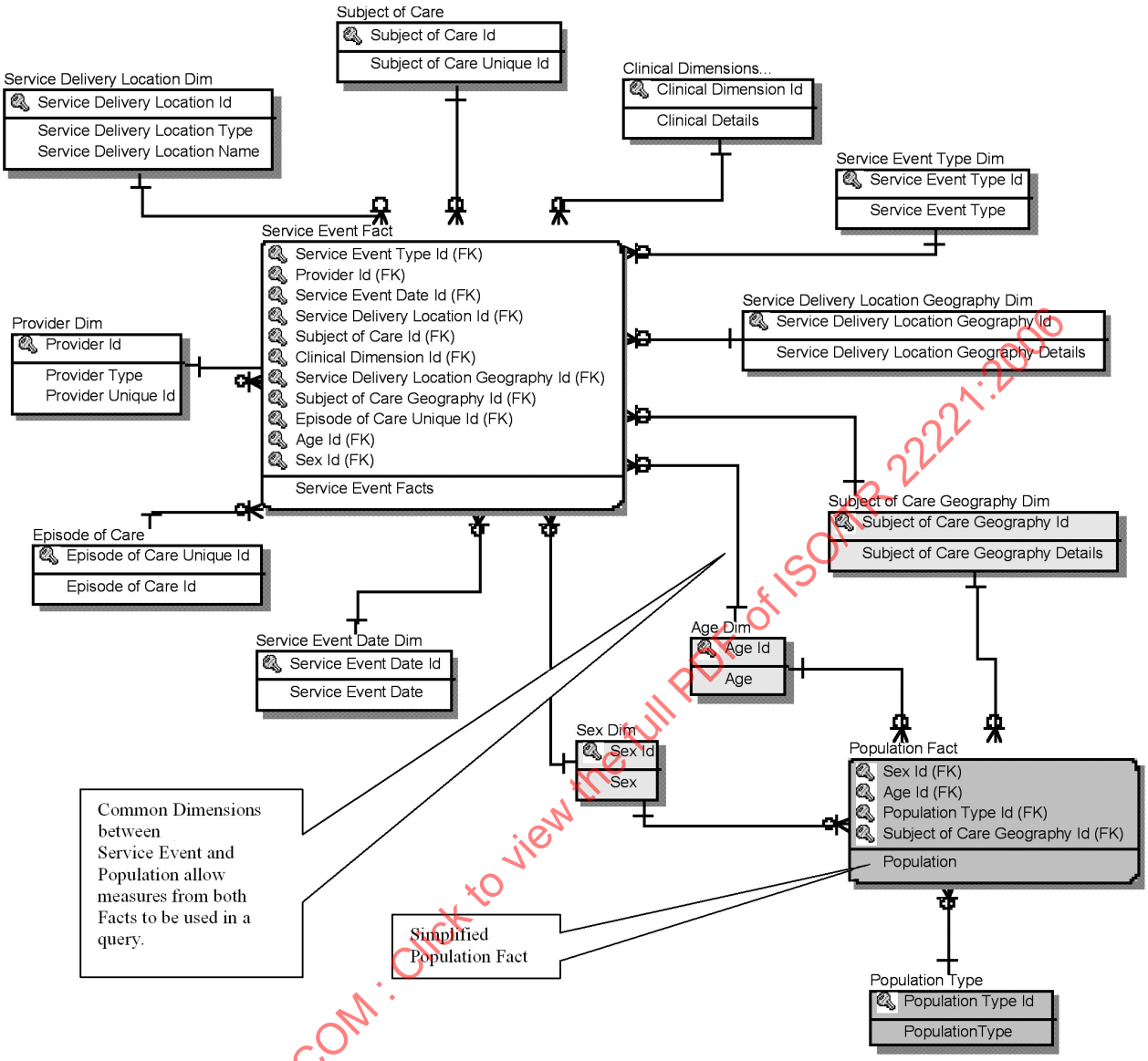
The following informally describes the additional dimensions and facts in Figure 4:

— Population fact: measures the population at a particular geographic, age and sex grain;

— Age dimension: describes age (common to both population and service event);

— Sex dimension: describes sex (common to both population and service event); and

— Population Type dimension: population type, e.g. a vintage associated with a census population estimate.



NOTE Common dimensions are Sex Dim, Age Dim and Subject of Care Geography Dim.

Figure 4 — Linking population and service event

6.4.2 Granularity and dimensionality

To answer the full breadth of health care questions, fact data are stored at an atomic or record level. Schema design and analytical tools must support highly dimensional user questions to be effective. The high dimensionality of health care data is a differentiator when it comes to comparing the CDW to other data warehouse subject areas.

6.4.3 Many-to-many relationships

Many health care questions require the resolution of queries that involve many-to-many relationships. For example, there are often many diagnoses related to an inpatient discharge, and these diagnoses are in a many-to-many relationship with the inpatient discharge fact (type of service event fact). Resolution of these many-to-many relationships can be achieved in the CDW by joining fact tables together through common dimensions, such as the service event or subject of care dimensions. This creates what is known as a "constellation schema" (see Figure 4); an alternative approach is to create linking tables associated with a single fact table that resolves the many-to-many relationships (see Reference [2]).

6.4.4 Conformed dimension grid

During the initial stages of gathering CDW requirements, a very useful exercise is the development of a conformed dimension grid. Reviewing the dimensional nature of the various organizational data holdings and looking to see which dimensions are common between data holdings create these types of grids. Once the grid is complete, a clear picture emerges as to which dimensions are candidates for conformance and which could act as bridges between data holdings. A simplified conformed dimension grid, with typical common dimensions, can be seen in Table 1. An "X" in the cell indicates that the dimension is available within the data source.

Table 1 — Example of conformed dimension grid

Data source	Dimensions			
	Date	Subject of care	Provider	Location
Data source A	X	X	X	
Data source B	X			X
Data source C	X	X	X	X
Data source D	X			

6.4.5 Unique subject of care identifier

A unique subject of care identifier, such as a health card number, should be collected in the CDW where available. This technical ideal will often have to be balanced with the jurisdictional availability of a unique subject of care identifier and local privacy legislation. In most instances, sensitive data such as this will require encryption within the database and the use of surrogate keys (see Clause 5 for details). If collected in the various CDW data marts, the presence of such an identifier will allow longitudinal and familial subject of care linkages across the various data marts within the CDW. These linkages help to leverage greatly the data found in the individual data marts and allow CDW users to follow subjects of care across a continuum of care (i.e. from the acute care visit to the possible subsequent acute care hospitalisation).

Although it falls outside the scope of this Technical Report, readers wishing to do longitudinal subject of care linkages should familiarize themselves with the principles of probabilistic and deterministic record linkage techniques. Linkage algorithms are often incorporated into the more sophisticated ETL tools and are commonly referred to as customer matching algorithms.

This notion is similar to an enterprise master patient index (EMPI) used in operational systems. An EMPI, commonly acquired as a commercial product, is a database that contains a unique identifier for every patient in the enterprise.

6.4.6 Extraction, transformation and load (ETL)

The data warehousing extraction, transformation and load (ETL) process essentially consists of a series of programming steps that manipulate source data into a state such that they can be loaded into a data warehouse. The overall process of data warehousing includes:

— identification of data sources based on user requirements, at short and long terms, for extraction and modelling,

— cleaning, formatting, transformation, encryption and, possibly, encoding of data before and/or after warehousing, and

— improvement or change of data sources by the use of relevant information, such as rules generated by OLAP tools, decision support systems (DSS) such as dashboards, or data mining.

The statistical imputation of values, such as the probabilistic linkage of subject of care records, is performed within the ELT process, as is encoding.

6.4.7 Update strategy

In updating, the high volume of data to be loaded in a clinical data warehouse makes updating the data a considerable and complex task. In this case, incremental updating is generally the most appropriate updating method used for a large volume of data, since it is almost impossible to reload the whole data warehouse. User needs may involve real time, short-term as well as longitudinal long-term trends. The update frequency should take into consideration those needs, particularly to keep the data warehouse as near to real time as possible. Availability of the data warehouse should not be interrupted during an update. This is why it is important to choose technology that can offer this feature.

6.4.8 Extraction, transformation and load tools

ETL tools are a class of tool intended to automate the data preparation, formatting and loading tasks commonly associated with data warehousing. While not absolutely necessary for the maintenance of a data warehouse, this class of tool can be very beneficial for more complex warehousing applications with a heterogeneous data source environment and complex data transformation requirements.

6.4.8.1 Controlling the data quality

Controlling the data quality is an important issue in data warehousing, and tools that support the quality control of data are required, e.g.

— to check for null values,

— to recover incomplete data and missing values,

— to check for relationships between tables (primary keys versus foreign keys), and

— to issue alerts during the loads about erroneous or missing data.

6.4.8.2 Slowly changing dimensions

The concept of a “slowly changing dimension” is fundamental to a data warehouse based on dimensional modelling. Essentially, this applies in cases where the attribute for a record varies over time. There are three types of slowly changing dimensions, as described below.

— Type 1, which occurs when the new information simply overwrites the original information, i.e. no history is kept. A typical CDW example would be assigning a new name to an existing diagnosis or surgical procedure, without changing the actual meaning of the diagnosis or procedure.

— Type 2, which occurs when a new record is added to the dimension to represent the new information, such that both the original and the new record will be present. The new record gets its own primary key. A typical CDW example would be when a patient changes residence to a new address, but information about the old address is kept to support historical analysis about health or service in the context of geography.

— Type 3, which occurs when there are two columns to indicate the particular attribute of interest, one indicating the original value, and one indicating the current value. There is also a column that indicates when the current value becomes active. Type 3 dimensions are rarely used in practice.

6.4.8.3 Normalization

Normalization is a database design technique used to remove redundancy and enforce update consistency. It is primarily associated with getting data into a database. The systematic removal of redundancy and increased consistency associated with functional dependency between data items is associated with increasing normal forms. Most operational databases are at least third normal form (3NF). While often useful in the early stages of data warehouse design, these structures are not good at supporting query-only activity, as found in the CDW, where the objective is optimal retrieval and descriptive power.

6.4.8.4 Data quality

Due to the sensitivity and the critical nature of clinical data, data quality has become one of the most labour-intensive activities in data warehouse development and maintenance. Data quality processes are complex and should ideally be performed during and after ETL processing. During the ETL process, there are certain types of data-quality-related transformations that typically take place. These include the following:

— the generation of error reports and/or logs, e.g. the replacement of missing values (i.e. a subject of care record with no date of birth) with standard, data-warehouse-wide, missing values (such as “not applicable”, “not collected”, “not available”, “not self-reported”, “invalid”, “to be determined”, etc.): such missing value standards are important since they remove any ambiguity on the user’s part as to what a blank value might represent;

— sparseness issues in health care data, such as low occurrences of unique conditions (e.g. contagious diseases such as Ebola) which require that continuous validation between source and target data and imputation of values is done rarely: after the ETL process, data availability, reliability, integrity and confidentiality should be checked;

— the generation of “audit” dimensions, which capture details on the ETL process, such as load, dates and metrics on the volume of records loaded: information from the audit dimension can be incorporated as metadata in query results.

6.4.9 Performance

6.4.9.1 Strategies to reduce input/output

Rapid performance is a key factor in user satisfaction. As is common to all data warehousing, minimizing input/output is the key to good performance. Approaches to achieve this include:

— the use of the star schema by cost-based optimization in relational databases,

— the use of parallelism, e.g. parallel query execution,

— indexing, especially bitmap indexes,

— table partitioning,

— data caching,

— summary management/aggregate tables,

— derived fact tables or data marts to eliminate frequently used or complex joins, and

— dedicated data warehouse appliance.

6.4.9.2 Online analytical processing

OLAP provides a multidimensional conceptual view of the data, including full support for hierarchies and multiple hierarchies, as a logical way to conduct analysis. CDW would favour a relational online analytical processing (ROLAP) approach, due to the dimensional nature of health care data (see Clause 8 for details).

6.5 Security and privacy

When dealing with any form of person-identifiable health information, it should be of paramount concern to ensure that the data are accessed only in the manner intended (Clause 5 covers this topic in detail). From an architectural perspective, the appropriate use of data in terms of protecting the privacy of individuals cannot be achieved by technological/architectural means alone, and must always work in concert with legal, policy and procedural measures. With this in mind, the discussion below outlines key architectural features prevalent in the CDW.

A core principle in data security is that data is only made available on a "need to know" basis. CDW customers should be given access to the data they need to do their work, but to no additional data beyond that.

Star schema dimensions offer a very flexible architecture for implementing this principle within a CDW. The dimensions can be populated with all available data (including sensitive data elements), but then the ability to query those data elements can easily be controlled using standard relational database techniques, such as database views, and/or BI tool features, such as role-based security and limiting the level at which users can drill into the data. The ETL process used to load the CDW can also assist in hiding sensitive items while facilitating analytical capacity. Sensitive attributes are often linked to related non-sensitive descriptive attributes as part of the loading of CDW dimensions, e.g. the subject of care postal code (a potentially sensitive item) would be linked with related coarser grained geographical descriptors such as a health region, province, etc., in the load of a subject of care geography dimension. These types of coarser attributes would then be used to query data without the necessity of exposing the sensitive item (in this case, the postal code).

The following are common examples of technical safeguards that should be considered in the planning and deployment of the CDW:

— role-based security access control based on end-user role: selective permissions at table, column, row and drill-depth level are often required,

— user authentication processes,

— encryption methods for the transmission and storage of data,

— small cell suppression and threshold techniques,

— network protection mechanisms, firewalls and intrusion prevention/detection systems (IPS/IDS),

— threat and risk assessments (TRA) and ethical hacking, and

— auditing and monitoring.

For detailed guidance on security as it relates to health care data, including the technical safeguards above, reference should be made to ISO 27799.

7 Metadata and education

For the CDW to be leveraged in an effective manner, an appropriate education program needs to be provided to users. Furthermore, this needs to be accompanied by access to up-to-date metadata. Metadata, which is commonly defined as data about data, comes in many forms. Typical metadata items include:

— attribute names,

— lengths,

— domains, and

— descriptions of all kinds.

Further examples include data dictionaries, data models, data quality notes, data lineage, methodology notes, etc.

Metadata is often stored in a data dictionary and/or a specialized metadata repository. There is a substantial body of existing work associated with metadata and, as in the case of CDW architecture, this can be taken as the starting point for any metadata initiative.

NOTE For example, see Reference [8].

7.1 Importance of metadata

The key focus for the CDW end user would be associated with analytical usage, particularly as relates to using data in an appropriate way to support the query undertaken. Metadata would also be of interest to the technical team(s) supporting the CDW, as well as other areas such as data quality, methodology experts, etc.

7.2 Collection mechanisms

CDW metadata can come from many sources. CDW source systems will often have metadata that can be leveraged by CDW users.

By its very nature, a data warehouse blends data from numerous data sources. A CDW is no different. This blending of data sources often means that the most valuable metadata probably does not already exist and has to be authored by the CDW team. This would include information on data transformations, derived analytical data elements and known differences between the CDW and its source systems.

Additional common sources of metadata used in CDWs include ETL and data modelling tools. Reports can often be generated from these tools which document how the data are stored, manipulated, etc. This type of metadata is valuable to CDW users since it helps document the lineage to the data they are using. Although ETL and modelling tools are not the only sources of metadata, it should still be considered good practice to capture basic elemental data using these tools. It should be noted that there are also specialized metadata repository tools which allow various forms of metadata to be organized and accessed in an effective way. It is also possible to develop a custom metadata repository. A number of existing standards support this work, e.g. ISO/IEC 11179-1 and ISO/IEC 11179-3.

7.3 Ownership

To be most effective, overall metadata ownership should reside with a single authority. Larger organizations will often have a metadata repository team whose mandate includes collecting organization-wide metadata. In the absence of such an authority, the CDW team will often become the CDW metadata owners out of necessity.

Ownership for the individual pieces of metadata (i.e. individual documents) should ideally reside with the subject matter experts. In other words, ownership should reside with those closest to the data being