TECHNICAL SPECIFICATION

ISO/TS 29585

First edition
2010-05-15

Health informatics — Deployment of a clinical data warehouse

Informatique de santé — Déploiement d'un entrepôt des données cliniques

Reference number
ISO/TS 29585:2010(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 other circumstances, particularly when there is an urgent market requirement for such documents, a technical committee may decide to publish other types of document:

— an ISO Publicly Available Specification (ISO/PAS) represents an agreement between technical experts in an ISO working group and is accepted for publication if it is approved by more than 50 % of the members of the parent committee casting a vote;

— an ISO Technical Specification (ISO/TS) represents an agreement between the members of a technical committee and is accepted for publication if it is approved by 2/3 of the members of the committee casting a vote.

An ISO/PAS or ISO/TS is reviewed after three years in order to decide whether it will be confirmed for a further three years, revised to become an International Standard, or withdrawn. If the ISO/PAS or ISO/TS is confirmed, it is reviewed again after a further three years, at which time it must either be transformed into an International Standard or be withdrawn.

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/TS 29585 was prepared by Technical Committee ISO/TC 215, *Health informatics*.

Introduction

This Technical Specification furthers the work of ISO/TR 22221 by providing implementation guidance for a clinical data warehouse and describing general considerations of development and deployment, issues and applications of data aggregation and data modelling, and architecture and technology approaches.

The role of the clinical data warehouse is to enable data analyses in support of effective policies and decision-making, to improve quality of care, to improve health services organizations, as well as to influence learning and research. It will have relevance to both developing and more established health systems. It will enable meaningful comparison of programmes and outcomes.

Although data warehouse technologies are becoming increasingly used in non-healthcare sectors, their use in health is still at an early stage. ISO/TR 22221 had a primary goal of underpinning a coherent approach to the diverse and multi-stakeholder perspectives of secondary use of data from various health system sources. This Technical Specification is intended to have pragmatic relevance by indicating best practice in setting up a clinical data warehouse and in using it from data abstraction and architectural perspectives. The clinical data warehouse is distinguished by the complexity of the interactions of data and hence the challenges to provide adequate methods for evaluating process and outcomes of care for different populations and sub-populations. Currently such knowledge is relatively fragmented and it is too early to be integrated into an International Standard. A Technical Specification will however benefit progression to an International Standard by aligning emerging best practice from different international experience.

The clinical data warehouse is also, in health informatics, the place of the intersection of health services delivery, organization and epidemiological expertise concerned with adequate and effective data abstraction and presentation for different decision-making contexts as presented in ISO/TR 22221. Good use of the clinical data warehouse will depend on furthering common approaches to frequently used data abstractions that concern analysis of care delivery and organization. Effective data warehouse deployment will be enabled by promoting good practice in furnishing dynamically accessible, interpretable data combinations, which will depend on showing the relationship between clinical and health system need and the architectural properties of the data warehouse.

This technical specification complements the ISO 13606 series in that competent extended use of data beyond immediate care delivery depends on the effective organization of the original source data.

Health informatics — Deployment of a clinical data warehouse

IMPORTANT — The electronic file of this document contains colours which are considered to be useful for the correct understanding of the document. Users should therefore consider printing this document using a colour printer.

1 Scope

This Technical Specification has three sections, 1) general considerations of design and deployment, 2) data aggregation and data modelling and 3) architecture and technology, and is intended to provide an overall set of guidelines for clinical data warehouse deployment supported by useful descriptions concerning different data aggregation and modelling approaches as well as particular aspects of information architecture that contribute to successful deployment. The first section is of particular interest to healthcare decision-makers, including information technology managers, of requirements and procedures that support successful clinical data warehouse deployment. The second section supports the understanding, choice, instigation and evaluation of methods that ensure reliable selection and aggregation of primary data for adequate compilation and presentation to support decisions – this section is of particular interest to statisticians, epidemiologists, healthcare evaluation specialists and others. Section three is of particular interest to informaticians concerned with efficient architectures, data mining methods, dynamic data querying and visualization for clinical data warehouses.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/TR 22221, *Health informatics — Good principles and practices for a clinical data warehouse*

ISO/TS 25237, *Health informatics — Pseudonymization*

ISO 27799, *Health informatics — Information security management in health using ISO/IEC 27002*

3 Terms and definitions

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

3.1
clinical data repository
CDR
operational data store that holds and manages clinical data collected from service encounters at point of service locations

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

3.2
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.

3.3
dashboard
user interface based on predetermined reports, indicators and data fields, upon which the end user can apply filters and graphical display methods to answer predetermined business questions and which is suited to regular use with minimal training

3.4
data dictionary
database used for data that refer to the use and structure of other data, i.e. a database for the storage of metadata

3.5
data mart
subject area of interest within the **data warehouse** (3.6)

EXAMPLE An inpatient data mart.

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

3.6
data warehouse
subject-oriented, integrated, time-variant and non-volatile collection of data

NOTE See Reference [5].

3.7
data warehouse dimension
subject-oriented, often hierarchical business relevant grouping of data

3.8
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.

3.9
episode of care
identifiable grouping of healthcare-related activities characterized by the entity relationship between the subject of care and a healthcare provider, such grouping determined by the healthcare provider

[ISO/TS 18308:2004, definition 3.23]

3.10
health indicator
single summary measure, most often expressed in quantitative terms, that represents a key dimension of health status, the healthcare system, or related factors

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

[ISO/TS 21667:2004, definition 3.1]

3.11
metadata
information stored in the data dictionary which 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, interpretative notes and users of the database sources loaded in the data warehouse and the data warehouse itself. Metadata help users, developers and administrators manage and interpret information.

3.12
master data management
enablement of a program that provides for an organization's data definitions, source locations, ownership and maintenance rules

3.13
online analytical processing
OLAP
set of applications developed for facilitating the collection, analysis and reporting of multidimensional data

NOTE See Reference [7].

3.14
organization
group of people who have their own structure rules and culture in order to work together to achieve goals and/or to provide services through processes, equipment and technology, etc.

3.15
performance indicator
measure that supports evaluation of an aspect of performance and its change over time

3.16
persistent data
data in a final form intended as a permanent record, such that any subsequent modification is recorded together with the original data

3.17
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.

3.18
secondary data use
expression sometimes employed to describe the use of data for additional purposes other than the primary reason for their collection, adding value to these data

3.19
star schema
dimensional modelling concept that refers to a collection of fact and dimension tables

3.20
widget
standalone visualization component (e.g. a heat map, gauge or geographic map) that can be integrated with a data warehouse source and presented in an end-user dashboard

NOTE Custom widgets can be developed using a business intelligence vendor software development kit (SDK) and managed via the widget library.

4 Abbreviated terms

— DICOM℠ Digital Imaging and Communications in Medicine

— EHR Electronic Health Record

— HL7 Health Level 7

— ICD© International Classification of Diseases

— LOINC© Logical Observation Identifiers, Names and Codes

— SNOMED CT© Systematized Nomenclature of Medicine — Clinical Terms

5 Principle

The roles and capacities of operational databases and informational databases (data warehouses) are complementary. An operational database is designed to perform transactions in real time such as adding, changing or deleting patient data, or displaying current data for immediate care decision making. It has a limited capacity for data analysis and is focused on online support for care delivery. The exploitation of already existing and persistent data for other purposes, sometimes referred to as secondary use of data, typically involves data aggregation and/or linkage from multiple data sources. The concept of a clinical data warehouse here is an application of the notion of data warehouse (that is the bringing together of data relevant to the functioning of an enterprise), for clinical purposes understood in the broadest sense, including the ensemble of healthcare system factors that can influence patient care. Emerging issues such as semantic interoperability with research databases in the fundamental sciences are not considered in this Technical Specification.

Deployment of a clinical data warehouse

ISO/TR 22221 provides an informative description of the uses and principles of implementation of a CDW, including an overview of the issues that are further developed and addressed in this Technical Specification. A CDW allows many perspectives of use and is therefore of interest to many categories of stakeholders. The activities of CDW use in ISO/TR 22221 are considered under the headings of:

— quality assurance and care delivery;

— evaluation and innovation of health procedures and technologies;

— disease surveillance, epidemiology, and public health;

— planning and policy;

— knowledge discovery;

— education.

These titles give insight into the increasing relevance of the CDW to several aspects of the health system and its mission of effective healthcare.

6 General considerations of deployment of a clinical data warehouse

6.1 Overview

This subclause guides the setting-up, deployment and ongoing management of a clinical data warehouse. It should be of use to an array of management and deployment stakeholders by articulating the range of considerations pertinent for successful planning, project management and ongoing governance, as well as describing key issues of data sources and quality, choice of architecture and maintenance of privacy.

There is a distinction between an operational electronic health record system, designed to support direct patient care, as opposed to a clinical data warehouse which will typically combine data from a number of sources and/or organizations for analytical purposes, with a diverse group of users accessing highly sensitive personal information, where use of this information is governed by multiple pieces of legislation and policy. Sometimes the term “secondary uses” is used for the latter application.

A clinical data warehouse is typically used for many purposes such as planning, management, research, audit and public health. The intention is that information is automatically collected or abstracted from operational electronic health record systems and then organized and maintained for subsequent reporting. The range of reporting applications can be very wide, however, and as illustrated in Figure 1 the different purposes have different characteristics.

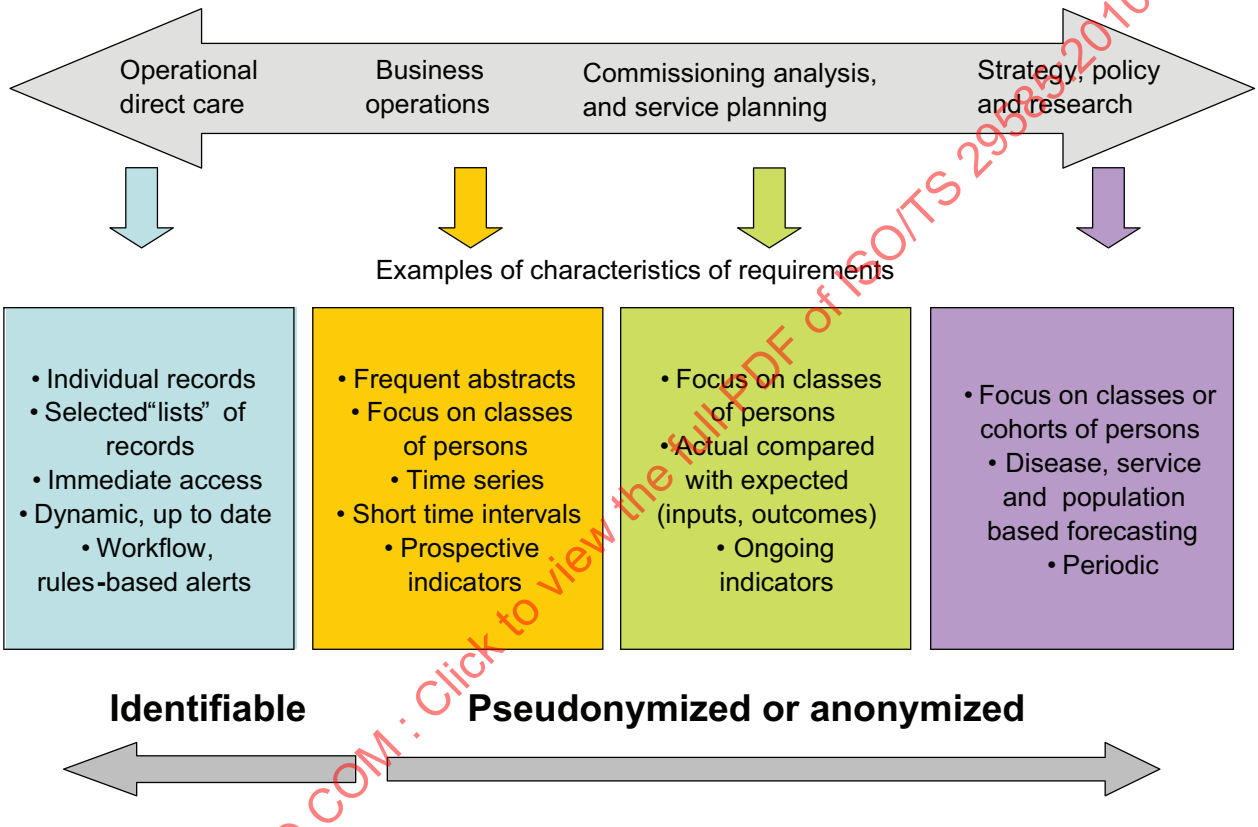

Figure 1 — Different types of information use

One of the issues, therefore, is to consider where to start. This subclause aims, therefore, to provide advice for those considering the development of a clinical data warehouse, to assist in:

— understanding the requirements;

— clarifying the scope;

— planning and implementation issues;

— design considerations;

— data and metadata;

— security and privacy.

6.2 Requirements

6.2.1 Overview

Consideration is given to three levels of clinical data warehouses: national (global); regional; local. National uses might be for statistical collection and comparison purposes including at a global level; regional might (depending on the country) be state, province or regional health organizations; local might mean individual organizations or hospitals.

Figure 2 illustrates how data may be collected at local level and the level of granularity within existing fields can be abstracted and summarised for use at regional and national levels. See Reference [37]. Subclauses 6.2.2 to 6.2.8 consider potential uses for a clinical data warehouse at national, regional and local levels. While it is often appropriate to have CDWs at each of these levels, each of which is attuned to the particular information analysis and reporting requirements of the sponsoring organization, a coordinated strategy for developing and populating the CDWs recognises there is a great deal of commonality in the underlying source data.

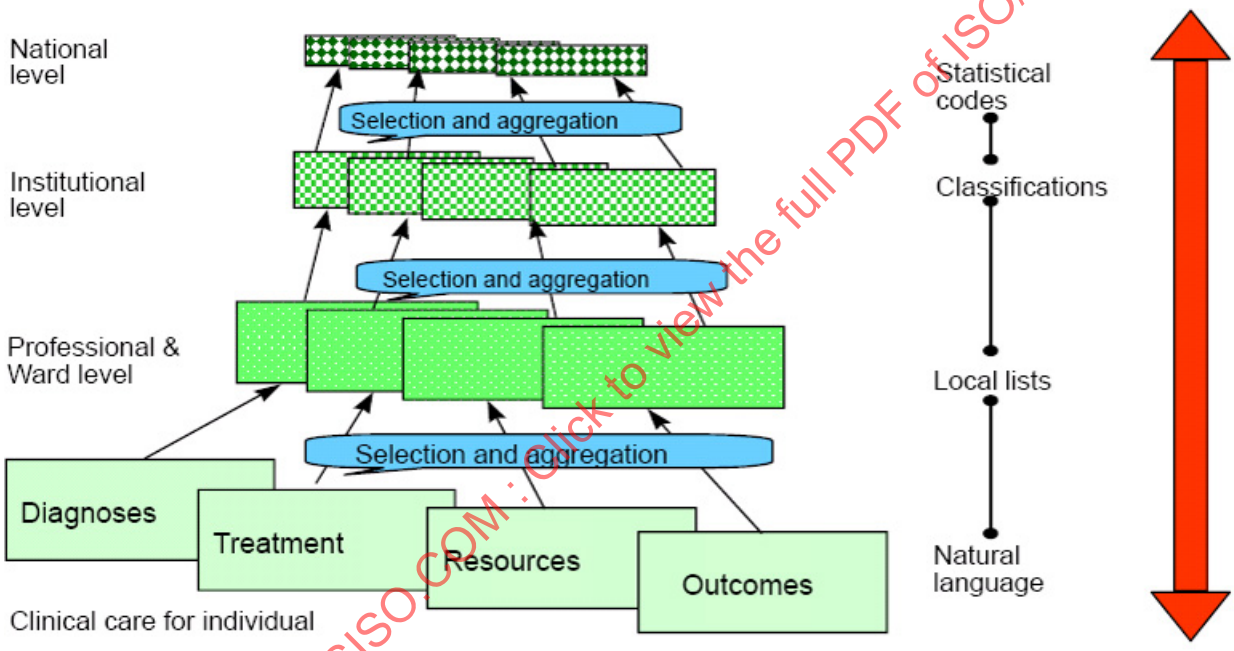

Figure 2 — Levels of clinical data aggregation

6.2.2 Users

There is a large range of potential users, which might include:

— national or regional government;

— government agencies, e.g. analysis and reporting centres;

— regulators;

— international organizations, e.g. World Health Organization (WHO);

— professional bodies, e.g. colleges;

— medical research and education;

— local care organizations, e.g. health providers;

— local government, e.g. environment, education, housing;

— other commercial users, e.g. pharmaceutical companies.

6.2.3 Requirements for local healthcare provider organizations

This subclause identifies typical requirements for healthcare providers, both as individual organizations and as part of a group or network of care service providers. Similar requirements might exist at a district or community level.

For individual providers, requirements might include:

— care service planning, monitoring and review analysis via

  — demand and capacity;

  — quality (access to services, etc.);

  — costs, efficiency and productivity;

  — outcomes and effectiveness, including clinical or care audit;

  — benchmarking and comparison with “peer” service providers;

  — strategic service planning;

— financial and contract management via

  — calculation of local costs (for reference cost comparison);

  — income estimation;

  — assignment of care provided to contracts;

  — monitoring of income against plans (budgets) and costs.

For provider networks, requirements might also include:

— care service planning, monitoring and review (as above but requires analysis of patient activity data which might have inputs from several networked providers);

— capability for spanning organizational and geographical boundaries where appropriate (e.g. cancer networks).

These functions require the capability to receive, manage, link and analyse data extracts from existing systems. Typically, this will require standards to be defined for these local extracts. The following functional components may be required:

— functionality to produce mandatory datasets;

— functionality to enable users to select and extract data from their operational systems (all elements of a patient's record);

— functionality to manage/store these extracts and combine them (via linkage) with data extracted from other systems;

— functionality to enable analysis and reporting of these data and to provide users with access to other specialist analysis tools;

— production of standard reports (both scheduled and ad hoc).

6.2.4 Regional requirements

This subclause identifies possible requirements for commissioners, insurers or purchasers of healthcare services for a large geographic area. In some countries the purchasers may be government organizations or insurance companies operating at regional level.

The requirements might include:

— public health and planning via

  — analysis of the prevalence of risk factors and disease;

  — analysis of the incidence of disease/conditions;

  — early outbreak detection, bio-terrorism surveillance;

  — analysis of the demand for care services;

  — analysis of the outcome of care services (in both the short term and the longer term);

  — analysis of access to services;

  — analysis of preventative care;

  — evaluation of the quality, efficiency and outcomes of traditional and alternative health service providers;

  — evaluation of the efficiency and effectiveness of alternative care approaches, service models and configurations, including alternative primary care provision and prevention services;

  — programme monitoring;

  — allocation of services in relation to health needs, resource inputs, access, quality, etc.;

— commissioning and contracting via

  — management of patient access to services, including achievement of access targets for pathways which span care providers;

  — expenditure forecasting;

  — allocation of resources;

  — assignment of care received against agreed contracts with healthcare providers;

  — reimbursement, contract and grant management;

  — monitoring of expenditure against plans (budgets) and provider service levels.

The functionality required includes the capability to receive, manage, link and analyse data extracts from relevant healthcare providers. The data requirements covering non-acute care events and settings might not be covered by universal standards. It will typically be necessary to have the ability to maintain local "reference data" relating to contracts, budgets, etc. and geographical analysis capability to support planning and purchasing activities.

6.2.5 National requirements

There are many potential uses for a national or supra-regional data warehouse:

— business and performance management;

— capacity and demand planning, commissioning linked to reimbursement;

— improving productivity, possibly through comparative analysis and benchmarking;

— evaluation of health procedures and technologies;

— standards and performance monitoring, looking at clinical indicators or other performance benchmarks, such as used by a regulator;

— monitoring and evaluation and international reporting;

— public health information, including screening, surveillance and epidemiology;

— research and development, including longitudinal studies and the monitoring of outcomes and effectiveness;

— support for broader health issues, including social care;

— international data comparisons.

6.2.6 Non-functional requirements

The requirements also include "non-functional" aspects. All of these features are important for the effectiveness of any CDW. They include:

— data volumes – capacity and scalability;

— timeliness of source data;

— timeliness of reporting feedback (which for local organizations might need to be close to real time, whereas for regional or national organizations reporting could be daily, weekly or monthly);

— data validation and data quality;

— robustness and resilience (e.g. fail over capability);

— strong security and privacy safeguards;

— deidentification services such as pseudonymization;

— usability and the flexibility to address the analytic needs of diverse stakeholders;

— performance requirements.

6.2.7 Services

A range of services should be provided. The planning for CDW will need to consider who is best placed to provide such services as:

— managing the ongoing intake and quality of the data and metadata;

— pseudonymization and management of access privileges;

— processing (e.g. derivations, aggregations, etc.);

— development of data dimensions, views and reporting;

— analysis and interpretation of the data and its derivations;

— tools (e.g. for data presentation);

— end-user education and support.

6.2.8 Benefits

The key benefits to be achieved by having a coordinated approach to the development of clinical data warehouses to support local, regional or national-level analysis and reporting include:

— consistency of data collection and analysis across a jurisdiction;

— comprehensive coverage of data collection;

— cohesion of information collection enabling, for instance, linkage of patient data across primary, community and acute settings for those receiving long-term care;

— timeliness of data that are collated directly from local sources on a regular schedule;

— a secure environment that enables patient privacy to be maintained;

— increased ability for sharing (particularly of aggregated data) for comparative purposes;

— a common approach to derivation of data.

However, achieving national or regional databases does require local organizations to submit data in a standardized format, in timely fashion and in accordance with agreed standards.

It should be noted that there also some potential “disbenefits” from the use of a clinical data warehouse. Aside from the obvious issues of security and privacy, it is possible that performance and monitoring information may be perceived to be commercially confidential or even a threat to local managers. Whilst there is an ethical responsibility to use data to manage the healthcare system, the confidentiality of the data must be protected.

6.3 Scope

6.3.1 Data content

There are various types of data that might be held:

— person-specific data, e.g. age, blood type, geographic residence, chronic conditions;

— person-specific activity data, e.g. health interventions, diagnostic test results, etc. across the continuum of care;

— other patient-related data, e.g. patient satisfaction surveys, patient-reported outcome measures;

— derived and support data, e.g. geographic data derived from post code/zip code;

— data on service providers, locations and care settings;

— data for non-health organizations;

— data other than that relating to care activities and experience (e.g. health determinants, workforce, finance, facility/equipment, etc.).

The data requirements need to be based around the purpose for which they are to be used, the ways in which they are collected and the types of output through which they will be reported. However, there are some important aspects to the consideration of these measures, including the potential sources of data, the types of data required and the characteristics of the data.

6.3.2 Data sources

There are various sources of the data, including extracts from EHR systems, data abstracts (acute care); assessment data, diagnostic data, patient surveys, etc. There are also areas (e.g. in the community) where there are still paper records. The sources of data are important in understanding what is available, and what will be available where new electronic record systems are being implemented. This analysis might indicate:

— those areas for which data are readily available, and hence early candidates for inclusion within the data warehouse;

— those areas for which data are not available electronically, and hence liaison is required with those responsible for planning the implementation or upgrade of operational systems to ensure both secondary and primary uses of the data for decision-making are contemplated when system investments are made.

Typically, detailed data are available from the acute sector, although datasets may provide measures of activity rather than outcome measures and hence can be limited in their value for purposes such as audit.

Areas such as mental or community health may have fewer data available. Many of the other data sources are based on aggregate or summary information. In some instances, e.g. finance, this may be appropriate in the form of returns that are required for national reporting purposes, especially when costs can be allocated to cost or activity centres. The inclusion of data relating to patient experience reflects the growing call for increased patient involvement and accountability to patients.

6.3.3 Timeliness of data

The types of data sources include, but are not limited to:

— detailed event datasets provided in “real time”;

— detailed event datasets provided in batch form;

— demographic data on the subjects of both care and healthcare providers;

— summary returns and contextual information on service availability, health determinants, costs, etc.;

— questionnaires on public, consumer and health provider viewpoints.

In terms of the characteristics of data, there have been increasing concerns about the timeliness and relevance of data, with a constant refrain about the poor quality of the data. This implies the need for:

— higher quality of data;

— electronic feeds where possible, with automatic validation;

— consistent derivation;

— more timely data.

It may be that there are local requirements (e.g. for regulatory purposes, or for government reporting) that provide a basis for requiring the submission of timely and accurate data, with the potential for sanctions to be applied in the event of non-delivery.

The analysis of data collection needs to consider:

— the development and maintenance of a comprehensive information model and data dictionary;

— a more formalized approach for the agreement and appraisal of datasets, which requires agreement on standards;

— further work to improve data quality, with validation at source, and regular reporting on performance.

6.4 Planning and implementation

6.4.1 Introduction

One of the key issues for a CDW is to agree where to start. Some users may wish to collect all the data possible, and then consider what to do with them (although they would need to justify the purposes for which data were being collected). Others may start with a very specific application in mind, and run the risk of creating a dead-end which cannot be extended and built upon. This subclause aims to provide advice on this. Typically, CDWs need to be planned with specific purposes in mind, but they also need to be extensible; the questions then become how to start, and how to grow. Implementing a CDW should be viewed as the development of an ongoing programme and not as a project with a defined end-point.

This subclause considers the prioritization and planning of applications, and then describes the phases of deployment: definition, development, implementation and education and training.

6.4.2 Prioritization

The factors that might be taken into account when prioritizing requirements for a CDW include:

— policy and strategic reporting needs for the business, especially in addressing those “compelling questions” which existing data marts and silos cannot address;

— the need to support day-to-day business requirements for the targeted CDW stakeholders;

— the availability of data and corresponding metadata from source systems;

— the ability to ensure continuity of service, particularly where current arrangements will cease;

— the need to demonstrate that appropriate security and privacy facilities are in place;

— the ability to provide “quick wins” which would be of obvious benefits to users;

— the contractual requirements (if appropriate) for current information systems and service suppliers.

An example, taken from the National Health Service in the UK, illustrates a sequenced approach to the provision of demographic data. This agreed series of steps, linked to the national provision of a demographic service, is as follows.

Step 1: provision of a simple monthly copy of the demographic to enable basic counting (e.g. of numerators and denominators).

Step 2: a daily copy of the demographic data including history, plus cohort management facilities enabling groups of patients to be selected, marked and tracked.

Step 3: cohort management facilities to support a range of patient tracking, including linkage with the Government Statistics Office.

The longer-term vision then sees the patient demographic data within the data warehouse being used as the master for all demographic analysis, obviating the need for specific datasets to include demographic information other than the core patient identifier.

The important tenet with this area, as with others, has been to ensure that early design decisions are taken with a view to future development and requirements.

6.4.3 Definition stage

The definitional work includes the following.

— Detailing the target stakeholders and their business requirements, and addressing the scope and the purposes for which data are to be collected (e.g. patient care, performance management, commissioning, outcomes management, planning, reimbursement, etc.).

— Assessing the availability of the data required to meet the business requirements. This will be important in order to understand the implications of any proposed data collection — what coding scheme is used, whether sites rely on local coding, whether the systems are patient-based and whether they record details of patients' problems and community interventions, etc. There may be an important decision here: a short-term imperative might require significant re-use of current datasets, but such an approach might not meet the business requirements, and hence a more radical approach may be needed to subsequently develop a new source for the data.

— Dataset definition, including any required data dictionary reconciliation and documentation of the metadata. This could include reference to, and assessment of, current datasets. Table 1 is an example of some of the dimensions that might need to be considered.

— An agreed schema to transfer the datasets from local systems.

— Work with pioneer sites to test the feasibility of collecting the data.

Table 1 — Examples of dimensions

Who	Care professional identity definition, professions (main specialities/areas of work), group appointments
What	Clinical observations, findings, interventions, treatment function, codes and classifications, terminologies
Where	Organization identifiers, location type, mobile/transient locations
When	Scheduled start and end; actual start and end; episode structures
How	Care professional drop-ins/pre-arranged, medium used
Why	Referral definitions, pathway definitions, diagnoses/classifications for reasons for referral

The proposed dataset definitions and resulting schema, along with any data mapping requirements and limitations, will need to be appraised following piloting, considering a range of issues relating to the practicality and suitability for implementation of the proposed standard. Once the definition stage is completed, the requisite changes can be commissioned from national and local suppliers.

At this stage, it is helpful to identify standards requirements, either using existing standards such as SNOMED CT®, ICD®, HL7®, or identifying the need for specific standards to meet a particular purpose.

6.4.4 Development

The development stage could either take the form of a traditional system development life-cycle or be built around a more agile, prototyping environment. The development of initial proof-of-concept demonstrators or the subsequent development of reports could effectively be progressed using rapid techniques. The core build of an underlying warehouse (once requirements are formally documented) is a major task, and should be undertaken accordingly.

The commissioning of developments will need to be aligned with other information system or service provider developments. This might require management of national regional and/or local suppliers.

For a national solution, the steps might be:

— issue of change control note to supplier;

— specification of requirements (including the logic for the loading, mapping, processing, managing privacy/security and reporting of received data);

— supplier development;

— supplier testing and user assurance;

— implementation of central solution.

For local solutions, similar activities will be required:

— issue of change control note to suppliers;

— upgrade of local solutions;

— testing of data capture and submission with CDW supplier.

6.4.5 Implementation stage

Once development is complete, deployment and implementation can begin, including:

— integration testing to ensure national and local solutions work together;

— local deployment in each local community service;

— migration locally to submission of the updated version of the clinical datasets.

Once data is flowing centrally, then reporting activity starts.

Suppliers use different technical solutions, and this needs to be borne in mind when attempting to consolidate data. Some suppliers may not record patient interventions according to defined terms or do not provide functionality to support context-specific pick lists. These differences are important, as they may constrain the options (and the timing) for local organizations to implement collection systems.

Although testing will hopefully have proved the functionality of the system and (to an extent) its usability, test data can never replicate all live situations. This is a risk that needs to be managed. It may therefore be prudent to have a period of user assurance post go-live, in which a smaller number of users operate with full volume live data, but test the content, format and presentation of all reports. This might identify further defects and might also indicate areas where tuning can improve reporting performance.

6.4.6 Education and training

A comprehensive training programme is required for CDW programmes and should involve training for data providers, CDW operations staff, data analysts and consumers, since data quality and documentation need to be everyone's concern. The main topics, appropriately targeted to the respective audiences, should include:

— data quality “best practices”;

— approved uses of the data;

— data limitations and documentation;

— CDW processes and policies;

— access to reports and associated metadata;

— approved analytical techniques;

— analysis and use of reports;

— privacy and security of the data.

6.5 Design considerations

6.5.1 Overview

This subclause provides an introduction to some of the architectural principles to be borne in mind when planning a clinical data warehouse. Further details may be found in the architecture and technology clauses of this Technical Specification.

6.5.2 Enterprise architecture overview

The architecture in Figure 3 shows those components that are typical in a data warehouse environment. The main components (flow from left to right) are the source systems, extract, transform and load (ETL) and data architecture layer, presentation layer and security layer – these are complemented by supporting components, including the development and test environments, application and infrastructure management, information governance and scheduling. The detailed description of the architecture is found in the architecture and technology clauses of this Technical Specification.

Subclauses 6.5.3 to 6.5.6 provide key design principles and supporting information on source systems, the extract, transform and load process, presentation and supporting tools.

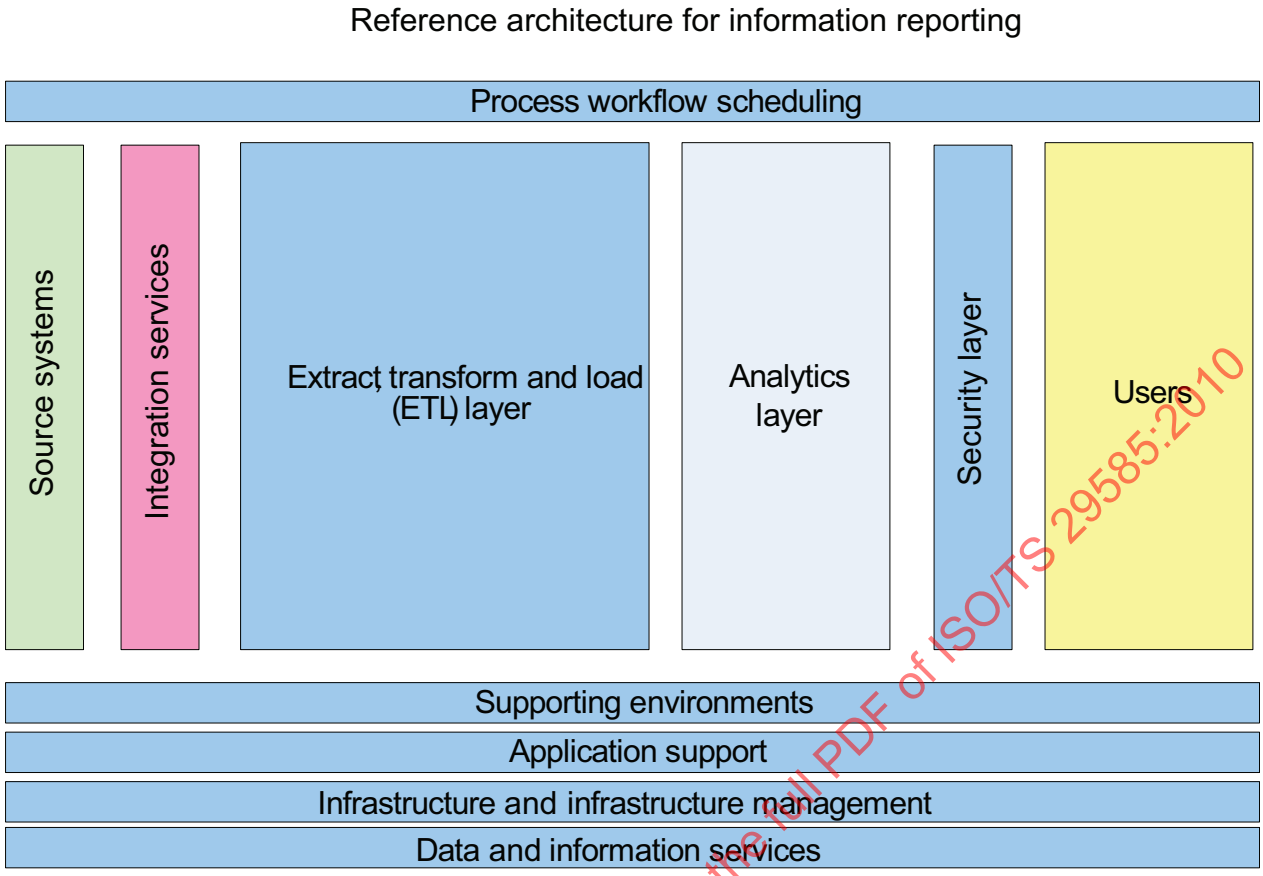

Figure 3 — Main components of a data warehouse

6.5.3 Source systems

Key principles for source systems and data architecture are:

— the data architecture layer should be both source system and data warehouse application technology independent;

— the data warehouse needs to be protected from both short and long-term source system changes and technology landscape changes;

— a clear audit trail of data from source to presentation must be available as it provides traceability and builds confidence in the organization's management of the data;

— data must not be mastered in the data warehouse as they are not trivia and would detract from the overall focus and operation of the data warehouse; data should originate from a robust, independent source system;

— similarly to mastering data, the CDW must not be the source for applying new business logic to datasets, e.g. hierarchical groupings.

The source systems in this context are all those “productivity” applications and other data sources that provide data for the CDW. These are likely to be all the clinical and non-clinical applications that are used on a daily basis to support the end-users' “day jobs”; these are also known as online transaction processing systems (OLTP). The CDW may also require data from non-OLTP systems, such as external data, e.g. mosaic codes for supporting demographic analysis. This type of information is usually extracted in a batch time frame, typically overnight when system resource demands are low. However, the CDW may also receive real or near-real-time data feeds from any of the source systems, providing the infrastructure supports it.

6.5.4 Extract, transform and load

Key principles for the ETL component are:

— the data architecture must work to a set of common, conformed dimensions and measures;

— clear data architecture reduces end-user misunderstanding and silos of information;

— the CDW must be designed to support reconciliation of data from source to presentation;

— the CDW needs to demonstrate its accuracy and completeness in order to build and maintain trust with the end user; the easiest way to do this is to show that the data entering the system match the data being presented, with exceptions fully explained;

— the CDW needs to enable linkage of data (e.g. combining separate data elements relating to the same patient);

— the data warehouse should have no single points of failure;

— removing “bottlenecks” and failure points promotes availability and improves performance;

— the data warehouse must be scalable at each ETL level;

— a scalable CDW ensures that increases in demand and throughput, both short-term (potentially transient) measures and forecasted growth can be accommodated;

— the data warehouse processing must be balanced and resilient. A balanced system ensures the available resources are used optimally; a resilient system ensures that failures are handled well;

— the data warehouse processing must be protected from end-user report demands;

— the data warehouse performs two main but very different operations; the first is loading and updating and the second is searching (protecting the first from the second promotes availability through the reduction of resource conflict);

— the data warehouse must be implemented as a layered, component-based architecture, to improve reusability and to maximize deployment options;

— component architecture supports flexibility and therefore changes in the business and technology landscape;

— performance benchmark data must be recorded for both core data processing and end-user experience;

— the data warehouse must be designed to require no manual intervention for the normal loading and processing of source data;

— including manual steps in the data warehouse process can introduce delays and performance bottlenecks; processing should be automated by default and only manual when all other options have been exploited.

6.5.5 Presentation

Key issues include:

— the data warehouse must support web-based end-user access for both consumer and analyst profiles;

— the presentation facilities must allow end users a choice of access types;

— web-enabled deployments and their associated technologies are inherently easier to implement and deploy to a variety of devices;

— an architectural approach that supports the presentation of data at all levels (national, regional, local, etc.) and facilitates access to the associated metadata to inform the end user in interpreting the data being presented;

— a consistent set of standards for data use and data disclosure, in accordance with information governance and legal requirements;

— a basis within which other information service providers are able to access, use and add value to the nationally-produced data.

There are various forms of presentation required:

— pre-formatted outputs (e.g. monthly statements provided for senior management);

— direct access (e.g. for information analysts to carry out more detailed reports);

— dashboards to provide the ability to monitor performance and compare with others;

— analytic tools which might link with local data sources/facilities;

— output to other presentation areas.

Where there are various publication mechanisms currently in use, it would be sensible to rationalize these from the point of view of approach, intended audience, content and facilities. A publication scheme would also need to consider the work of organizations such as statistical centres and regulators.

The Information Governance Standards will be critical to assure the privacy and particularly of personal identifiable data, and it will be necessary to ensure local policies and procedures are in place to oversee access and use. The role of “honest broker”, i.e. a neutral individual or agent that works to mediate different interests in favour of the optimal privacy and security of personal health information, needs to be considered where different datasets are linked but then provided in pseudonymized form to other users.

The technical considerations for access include the possibility of using the same presentation tools for access. There is an important skills aspect as the expert analysis and interpretation of information becomes increasingly important. This is an aspect where a national, professional, lead is required.

6.5.6 Environments and supporting requirements

Key principles for the environment and supporting requirements are:

— the chosen data warehouse technologies should be proven commercial off-the-shelf (COTS) packages or mature open source solutions rather than bespoke coding development; the CDW, like other systems, benefits from COTS packages or mature open source solutions in terms of maintenance, product development and proven capability;

— access to the data warehouse must be in line with the security and privacy policy;

— the data warehouse will contain sensitive data therefore it is vital all access is controlled;

— the data warehouse must adhere to all statutory legislation and organizational policy detailed in the non-functional requirements (NFRs);

— data archiving must be in line with data protection requirements;

— archiving ensures that the live system only has relevant data to parse; however, the archive must stay accessible for retrieval, i.e. irrespective of software changes, hardware changes and data model changes;

— the data warehouse must have a clear strategy for dealing with disaster recovery;

— the CDW is considered a key system and therefore must be part of an overall business continuity plan.

6.6 Data and metadata

6.6.1 Approach to data and metadata management

A strategic approach to data management requires:

— an architectural approach that supports the collection and reporting of data at all levels: national, regional and local;

— a consistent set of standards for data, for data sharing and (by implication) pseudonymization and anonymization.

As indicated in 6.2.1, typically there will already be various data warehouses at different reporting or organization levels, and it will be important to have a process for coordinating and rationalizing these. Ideally, this would enable the federation or linkage of data warehouses to enable “summarization up” and “drill down” facilities.

There are also issues around the processing of data:

— to decide whether the intention is to provide transaction processing or data warehouse reporting and intelligence facilities;

— agreement on standards for areas such as algorithms, geographic and other identifiers, data derivations and the construction of indicators.

The management of data warehouses needs to be considered carefully, particularly in the light of information governance concerns.

The development of national approaches, while clearly beneficial, also raises the question of confidence in the processing and availability of data. It will be important to ensure that national and regional services have transparent performance regimes to demonstrate their effectiveness and reliability.

6.6.2 Master data management

Master data management (MDM) actually covers a much wider subject area than just the CDW, as it is not just the CDW that requires high-quality data, it is the entire organization. An MDM programme should provide the organization with a clear set of data definitions, source locations, owners and maintenance rules.

In the context of a CDW, the following are considered best practice.

— Where possible, identify a single source system for each and every data item that will reside in the warehouse. If a single source is not immediately possible then a plan should be developed to achieve a single source and in the interim clear rules should be defined to resolve conflicts.

— Resist the urge to maintain master data in the CDW. The effects of the real world sometimes means that the warehouse is the most appropriate place to master certain reference data items, but these are usually only where no business source system exists, e.g. external demographic lists used only in the CDW.

— Develop guidelines to ensure that master reference data are provided [perhaps through Service Level Agreements (SLAs)] and to deal with transaction data with late arriving reference data. Typically transactions with late arriving reference data can be handled through a suspense process, i.e. hold the transactions until the reference data arrives, or post the transactions for user reporting with an “unknown” dimension tag.

— Ensure that all reference data (to be clear, the items that can describe a transaction), which include lists of values, dimension attributes, hierarchies, etc., have an associated effective from and to date.

— Consider the use of a data hub for key datasets, e.g. patient. A data hub in this context would be another source system for the CDW but would act in a wider context, e.g. it could provide a consistent set of agreed data about the subject to all systems in the landscape, not just the CDW.

6.6.3 Metadata

Technical and business metadata are mandatory for the effective delivery of business intelligence (BI); they are equally important but distinct. An understanding of the technical and business meaning of a data element allows for more effective exploitation of the analysis of information related to the element. Technical metadata describe the transition from the operational systems to the CDW and from the CDW to the data marts, whereas business metadata describe the data to the business user.

Technical and business metadata should be independent of the architecture of the warehouse and of the application. Metadata must ensure that the same information produces the same results across marts. Business metadata for elements in the CDW must be consistent for the same elements within the data mart(s).

Metadata required for each field include values, alignment with business process, length, association to other fields, units of measure, special considerations, etc. (The metadata definition process is recognised as a valuable mechanism to identify business processes that require rework or enhancements.) Standard naming conventions for metadata should be adopted.

6.7 Security and privacy

6.7.1 What is information governance?

Governance can be defined as the action of developing and managing consistent, cohesive policies, processes and decision rights for a given area of responsibility. Information governance (IG) is therefore concerned with the ways and means that a system and its supporting services handle information, in particular personal and sensitive information. IG has to provide a framework to enable personal information to be dealt with legally, securely, efficiently and effectively, through bringing together all the requirements, standards and best practice that apply to the handling of personal information.

It is important to engage with stakeholders to agree the “rules of the road” around who gets access to what data, and what the key services and tools are with which we have to manage their privacy and security concerns, e.g. through managing the level of access to data, how data get published, how any consent directives are managed and how patient and provider identities are protected.

Information governance is a holistic approach to data protection and information management issues, using a variety of techniques covering areas such as:

— security;

— privacy;

— data integrity;

— data quality.

The subjects (individuals, the public, care providers and organizations), the data providers and the consumers of the clinical data warehouse are very concerned with these issues, and maintaining the confidence of these stakeholders is paramount if databases of sensitive information are going to be developed and deployed for uses outside the direct care process and the local health services delivery context. Hence it is important that there are strong and clear policies in place for managing the security, privacy, integrity and quality of the information as well as the necessary oversight and transparency so that the stakeholders are confident that the policies are being adhered to.

Standards and best practices are as applicable here as they are for EHRs and other health information systems. Nonetheless, there are a number of special considerations that are relevant to a clinical data warehouse. These are described in 6.7.2 to 6.7.8.

6.7.2 Governance

A respected governance structure provides oversight for the CDW and establishes clear rules and expectations for the providers of data, the managers of the CDW and the consumers of the CDW's services. The needs and interests of all parties, including the subjects of care and the public good, need to be carefully balanced so that all stakeholders are confident that the data warehouse respects the rights of individuals, while enabling improvements in the health system, by using information to enhance the quality of care that providers can deliver and the ultimate outcomes of health services for patients. A variety of mechanisms can be used for governing a CDW. Suggested best practices include having a governing committee composed of the key stakeholders including members of the public, publicly accessible policies through which information is managed, a clear set of goals and benefits of the CDW, etc.

6.7.3 Privacy

It is important that access to personally identifiable data (for both subjects and providers of care) is very tightly restricted to only those individuals who have the need and privilege/permission to know – any patient directives regarding access to their information must be respected within the parameters defined through legislation or policy. Rigorous approaches to deidentification and pseudonymization (see ISO/TS 25237). In addition, even though personal identifiers may have been removed from low level data, there is the risk that, by combining enough attributes of an individual in a small population cohort, identities can be inferred. Therefore, policies, practices or techniques should be employed to reduce these risks, e.g. by setting lower limits on cell sizes in response to queries or in published reports using data from the CDW.

6.7.4 Data quality and integrity

Stakeholder confidence in the CDW is strengthened if data integrity issues and procedures are well ordered and transparent. Data integrity challenges are significant since the CDW draws upon data from a number of varied external sources. Integrity can be compromised by errors in extracting data from the source systems, loading, transforming, or managing their integrity within the data warehouse or failing to document known gaps or limitations in the data through a robust metadata mechanism that is accessible to end users of the information outputs of this comprehensive health information resource. Data quality practices in the CDW are described in Clause 7.

6.7.5 Privacy principles

The following principles underpin the privacy measures for the clinical data warehouse:

a) The expectation would be that data for reporting should be provided in unidentifiable (aggregate or anonymized) form except where specific justification can be made and approvals provided; thus the default is the use of data that cannot be linked back to individuals.

b) Where the case is made for access to data relating to identifiable individuals, the informed consent of these individuals should be obtained wherever feasible.

c) Where use of identifiable data is required, and where patient consent cannot be obtained, a full justification would be needed.

   Specific justification is required for the particular use or there may be a “class” justification under local law. In all cases an approval process would be followed, with the involvement of an appropriate lay-person, and approval would not be automatic. It would depend upon the extent to which it had been demonstrated that the use of identifiable data would:

   — benefit current, former or future patients in either the short or long term;

   — benefit the population by improving the cost-efficiency of the provision of health and/or social care;

   — and, (in either of the above cases) where it is not reasonably feasible to achieve these benefits through consent or the use of anonymized data.

d) All users of data for secondary purposes should be subject to enforceable standards regarding privacy and security of data.

e) The process of determining and granting access to data should be transparent and follow principles of good communication with all parties in order to achieve the appropriate balance between individual privacy and public benefit. The involvement of patients and the public must be regarded as essential.

6.7.6 Security

Given the concentration of sensitive data in the CDW, particular care must be taken to secure the databases and infrastructure, and to ensure that any data being prepared for input, or being extracted from the CDW for analysis or reporting, are similarly well secured.

The main elements of a CDW which support security are access control, namely the combination of role-based access control (RBAC), definition and assignment of accessible functions, audit facilities and pseudonymization (a particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms).

In a CDW, user access control needs to be defined in the context of:

— specific sets of data and the business function the user is associated with;

— the level of identifiability of the data (i.e. aggregate, anonymized, pseudonymized or patient identifiable);

— the organization of the user if patient-level data are involved;

— particular CDW functionality (e.g. extract data, access online reports).

As indicated earlier, CDW has a wide range of data and functions and together these combine such that CDW requires support for a high level of granularity in order to provide access to specific data for specific purposes. This translates to a relatively large number of business functions being required to differentiate the uses and users of CDW. This shall run counter to the need in order to simplify the numbers of activities or business functions to be administered by registration authorities (RAs).

6.7.7 Auditing CDW and information governance

The auditing functions within the CDW should enable records of CDW users and usage to be generated to enable the audit question of “who has accessed CDW?”, considered as a “forensic” audit. Within CDW, functionality should record access and transaction content in terms of recording the queries and datasets accessed by users in order to answer the “what did they do?” question. Only results of queries (not standard extracts) or online accesses involving patient identifiable data should be stored.

A stratified approach to managing the logging of usage to meet the purposes set out above in relation to levels of “identifiability” is summarised in Table 2.

Table 2 — Logging access to CDW data

Identifiability level	Output	Implications
Aggregated and anonymized data	Extracted or non-extract (i.e. on screen)	Aggregated and anonymized data do not include data that allow identification of individual patients and therefore there is no necessity to log actions of users or the data subjects involved.
Pseudonymized data 1	Extracted or non-extract (i.e. on screen) result set greater than K (where K is to be defined)	Pseudonymized versions of patient-based activity records are created in order to minimize the identification of individuals whilst enabling linkage of individuals' records. Assuming the result set is of sufficient size, then it should be sufficient to audit the user's execution of the query (i.e. which dataset, parameters of the query, etc.). It should not be necessary to audit the consequent retrieval of all data comprising the result set.
Pseudonymized data 2	Extracted or non-extract (i.e. on screen) result set less than K (where K is to be defined) or is equal to 1	Access to a small dataset that could provide potentially readily identifiable results but access to which will not normally be covered by legal approval. Assuming the result set is small enough to risk the identification of one or more individuals from it (i.e. the result set size is less than K) then the audit should include both the user's execution of the query (i.e. which dataset, parameters of the query, date, time, etc.) and identify the data items that comprised the result set.
Depseudonymized (or reidentified) facility	Used against an extract or non-extract (i.e. on screen)	Such an extract or access should only be permitted within the terms of the security and confidentiality policies that apply to the CDW and/or by explicit legal approval. As the output will always provide access to numbers only, a record should be kept against the person who used the depseudonymization facility and the pseudonyms for which the facility was used, date, time of usage, etc.
Patient identifiable data for use within CDW	Data extract/access on screen outside the CDW realm	This would be to move data from one part of CDW to another. As this would be internal to CDW, it would only be necessary to record the query against the person who ran the query and not record the query against the people who formed the result set.
Patient identifiable data for use outside CDW	Data extract/access on screen outside the CDW realm	Such an extract or access may be covered by legal approval, although it may require a policy decision. It would only be necessary to record the query against the person who ran the query and not record the query against the people who formed the result set.

Storing the results of all transactions involving pseudonymized data would generate huge volumes of data, in effect replicating the CDW. The effort required would be out of proportion to the potential risk of compromising patient identifiable data. Transactions involving the move from pseudonymized data to patient identifiable data should be captured by the audit functionality referred to above.

The results of queries involving patient identifiable data for disclosure outside CDW would enable subject access requests to be met, if this is deemed relevant. The consequence of logging the people identified in the extract or query result set in order to support subject access requests would be to create a “subject access” data mart. The size of this will be dependent on the numbers and frequency of use of patient identifiable data.

In order to ensure only *bona fide* users of patient identifiable data are allowed continued access, a user housekeeping function should be established. This will enable lapsed users to be deleted from the licensed users list for extracts. Similar functionality is being developed to ensure that CDW-derived data distributed through CDW-based mailboxes are deleted after a suitable period.

External users should be subject to the same conditions in their contract for access and receipt of data.

6.7.8 Pseudonymization

To protect the identity of individuals, pseudonymization is an important privacy-enhancing tool in the CDW environment, in particular in situations where aggregate data are not sufficiently granular for an approved use of the data.

It is, however, also necessary to consider the impact of providing pseudonymized, as opposed to identifiable, data to support routine business processes. This would need to consider the extent to which each user (at the local, regional or national level) needs access to patient-level data that could be identifiable.

A study carried out in England concluded that there were legitimate reasons for local healthcare purchasing organizations (or "commissioners") to access patient identifiable data, but the volume of use could be severely reduced. This study was followed by two sets of pseudonymization pilots, in effect simulations, which sought to identify when and where patient identifiable data were needed, and by whom, in undertaking business processes. The outcome was to identify specific issues where controlled access by non-clinicians is required. These issues are set out below.

— Data quality: a fundamental issue for the use of pseudonymized data, in particular the assignment of derived data as well as the coding of administrative data. Plans to reduce the need for access to patient data include use of data from the demographics master file.

— Specialist healthcare purchasing: low-volume, high-cost cases, commissioned across a wider geographic area. The potential for gaining the patient's consent is being explored.

— Clinical support: some uses of CDW data are patient care related, such as selecting and identifying patients at the risk of rehospitalization (PARR) and patients that would benefit from active case management (ACM) by provision of community-based services instead of requiring acute care.

— Business process support: clarification of outliers or patients at risk requires identification of the relevant patients.

— Spatial analyses: these are used extensively to examine a wide range of subjects, such as optimum locations of service provision or clusters of conditions. These analyses rely on post code data, classed as sensitive information. The means of provision of spatial analyses need resolution.

6.7.9 Further issues

There are a number of issues that require solutions, including the following.

— Small number handling, where a particular query may generate a very small result set from which an individual's identity might be inferred. The Office of National Statistics in the UK has developed proposals for addressing this.

— External data linkage: policies need developing for a range of factors such as how usage can be commissioned by users; technical aspects like input data quality standards, formats, specification of keys, and key control; in areas like the need for pseudonymization of historic data and for extracts from CDW using a common key and in requirements placed on the submitter in the future use and handling of the "submitted clear" data.

— Back pseudonymization, where external organizations would need to replace existing patient identifiable datasets with pseudonymized versions into their data warehouses and analytic systems. An issue arises in how far back in time the pseudonymized versions need to go.

— Managing data access where there are consent directives from patients. Currently, it is felt that a major issue is identifying those records to which these restrictions will apply. Potentially the "free-standing" element of the pseudonymization service could provide the pseudonymization facilities for such records.

— Subject access requests: typically provided for under data protection legislation, these requests give patients the right to know who has accessed their records. It is important to clarify the implications of such requests for information in terms of whether this goes beyond accessing the CDW records in patient identifiable form or not.

7 Clinical data warehouse: data aggregation and data modelling

7.1 Introduction

This clause supports the understanding, choice, instigation and evaluation of methods that ensure reliable selection and aggregation of primary data for adequate compilation, and presentation to support decisions. This clause is of particular interest to decision-makers but also statisticians, epidemiologists, healthcare evaluation specialists and others.

7.2 Data and decision making

7.2.1 Overview

Aggregated data is a normal requirement of decision making. Decision making depends on the decision-making context often referred to as the business context. The decision-maker knows the major constituents of the business. In the health sector for a physician or an administrator this might be use of drugs, or tests, or use of hospital beds or many other possibilities. These constituents are candidates to be considered as the “dimensions” in the CDW data model.

7.2.2 Data quality

7.2.2.1 General considerations

Effective and responsible secondary use of data depends on knowing and defining the quality of the primary data. The more that health information standards enable primary data integrity and semantic interoperability, the more that analysis based on this data is supported.

Cleansing should take place within the source system and allow the normal CDW data processing (ETL) to pick up the cleansed records; however, it is normal practice to use the CDW processing to trap erroneous records and feed this information back to the business for action.

The CDW must also know if changes occur in the nature of the dataset during the history of the CDW, for example that a given laboratory test has a change of analytical method which has caused a change in its reference ranges. Such changes should be recorded in the CDW metadata.

7.2.2.2 Principles concerning source data quality

The following principles apply to source data quality:

— Business decisions should be made during development regarding the quality of existing historical information and the historical data to be made available.

— Business direction is needed for handling data quality situations related to business practice; for example, a null value may be acceptable in certain business scenarios and unacceptable in others. The CDW may handle some unacceptable values automatically, according to business direction.

— The business should provide rules to detect quality issues in the data staging area. These situations can be detected programmatically with an appropriate level of automatic reporting and notification of anomalies. All the business rules that have been used to trap erroneous records should be fully documented.

— Any data that is found to be erroneous should be amended in the source system and not the CDW; the CDW should pick up the change through its normal processing.

— Certain data elements will also be used in processes for pseudonymizing of data (see 6.7) and particular care needs to be given to the quality of these data elements given the impact on the assignment of pseudonyms.

7.2.2.3 Principles concerning CDW data quality

The following principles apply to CDW data quality.

— Complementary enrichment of data can be supported providing the original data are not altered.

— A key consideration in implementing a CDW is in maximizing the value provided back to the point-of-care provider who is providing the source data and being very attuned to their needs and motivations in wanting to improve the care they provide to their patients. The more tightly the loop can be closed between decision making at the point of care and the types of analyses that come out of the CDW, the more people who are providing the data will understand the importance and value of quality and consistency in the data the CDW is capturing from their source systems. This user-based experience has the potential to feed back to the designers and administrators of the primary data sources to enable improvement in primary data collection.

7.2.3 HL7 information modelling and terminologies

The progressive application of the HL7 RIM (Reference Information Model), the standardization of the information model and the use of standard terms, for example from SNOMED CT, support the specificity of the primary data and their relationships including the temporal relationships. This in turn supports the attribution to dimension hierarchies.

7.2.4 Atomic data and data aggregation

Original primary data ("atomic data") should be the preferred source for the CDW. However, it is possible that some data aggregation may occur prior to CDW transfer. Data aggregation may occur at the time of data transfer so that the aggregate is a new data field in the CDW, or occur during or following data query including the creation of graphs or tables or the application of statistics.

7.2.5 Dynamic querying and reports

Although in past practice the decision-maker may receive a static report of data about the business, the data warehouse enables dynamic data querying so that at the data warehouse terminal, the requested analysis may suggest other analyses and these can be immediately undertaken. The query may do a "drill down" of a given dimension. For example, instead of looking at activity per week, this could be activity by day or part of a day such as a work shift.

Reports based on these queries should be subject to the same criteria for static reports, clearly describing the units of the displayed data, the composition of any aggregate used, the date and time of the report, and including interpretative notes as metadata for the end user.

7.2.6 Data Interpretation, decision making and knowledge management

Reports from the CDW are used as evidence to inform decision making and policy development. They do not however substitute for interpretation. Similar reports in different contexts do not necessarily mean that the processes in those contexts are identical and different explanations may be required. Interpretation importantly requires the reference to the associated metadata and may involve consultation of persons with experience and expertise according to the context. Hence CDW reports insert value into a decision-making process and also future reports can show the impact of decisions made which may have been influenced by the information from these reports.

7.2.7 Local and regional analyses: dimensions and metadata

The CDW is important for local as well as regional and national analyses and decision making. As systems are increasingly networked, then certain data, useful at a local level, can be abstracted for use at higher levels. The more often that there is a direct relationship between local and higher levels, the more informative the information at a higher level, with the added possibility of returning to a lower level to better inform on that particular context. Nevertheless, certain local dimension descriptions may need to reflect particular local needs and their use at a higher level may not be possible or require adaptation for use from a higher level perspective.

Dimensions can be required at higher levels that are not required for local use. Programmatic reporting, for example, may require dimensions like programme area or funding mechanism not required at the point of care. A data element may acquire metadata and business rules at each level of the reporting hierarchy.

7.3 Defining CDW dimensions according to business need and relation to process

7.3.1 General considerations

In ISO/TR 22221, the role of the CDW is to meet major health system perspectives of adding value to the data obtained in clinical practice. These perspectives are:

— quality assurance and care delivery;

— evaluation and innovation of health procedures and technologies;

— disease surveillance, epidemiology and public health;

— planning and policy;

— knowledge discovery;

— education.

The health system as an enterprise is in several ways more complex as a challenge to develop a fully functional data warehouse. A patient is not a stable object but an entity that is continuously changing and it is the interpretation of these changes that is of central interest. A basic unit of analysis often reflects upon a sequential process, that is the study of temporal relationships between events gives important information. The care of an individual patient is a key example of a process which can be considered as a process with intermediate and final outcomes. Exploitation of the CDW takes a population view and seeks to analyse, for comparable populations and sub-populations from different perspectives, how effective and efficient are the processes of care being applied to these populations. Analysis of a process classically distinguishes people, organizations, places and events. These four components are a major basis of the dimensions commonly defined in the CDW.

A process is a set of interrelated or interacting activities, which transforms inputs into outputs. Analysis of sequence and possible cause-and-effect relationships needs to be appropriately sensitive to time dependencies. Hence time is a further important basic dimension.

7.3.2 Defining dimensions, attributes and facts

Dimensions tend to have a hierarchical classification and clearly specification of the members of that hierarchy should be where possible consistent. For example, providers of care could be physicians who could be further classified according to their speciality and/or to their unit of service, and/or to their seniority, etc. These differences within a dimension become “attributes” of that dimension. When a particular instance of a dimension occurs in the database then this becomes a “fact” about that dimension and it is important that the units in which that fact is expressed are defined and consistent.

7.3.3 Perspective: quality assurance and the delivery of patient care

7.3.3.1 The care process

This perspective directly relates to the process of clinical care. Clinical care starts with a problem or set of problems that require diagnosis and then treatment. Table 3 suggests the dimensions linked to this process.

Table 3 — Dimensions linked to the process of clinical care (example)

Major process events/CDW dimension candidates	Dimension components	Data sources	Units and standardization
Diagnostic problem	Symptoms and signs	Physician record of EHR	Terminology, SNOMED CT/disease classification
Diagnostic tests	Clinical findings, laboratory images, pathology and other tests	Diagnostic services	Test quantitative and qualitative results; test units; LOINC and DICOM
Therapy	Nursing care, medication, surgery, radiation or other	Therapeutic services	Different classification systems
Outcome	Cure, morbidity, mortality, functional impairment, quality of life/wellbeing	Physician record/EHR – medical record (health information manager) personnel	Terminologies and classifications that can be disease dependent
Time	Time intervals	Time stamp	Actual or relative time units

In the health system a person (subject of care) is cared for by personnel (provider of care) at a given location (service delivery location). According to population characteristics such as age, location or other characteristic, or according to attributed diagnosis or therapy selection, or stages in progression of the disease such as disease remission, persons can become part of a diversity of populations or sub-populations of care (population of care). The underlined items are common CDW dimensions.

Care is a continuum throughout life and includes preventative as well as therapeutic interventions and is influenced by organizational as well as other determinants such as genetic, environmental and socioeconomic factors. Analysis of the continuum of care needs to take into account significant activity groupings (service events), for instance hospital versus community care, which may, for example, be considered as a "visit" in the service event dimension, with a "length of stay" in the time dimension. The notion of episode of care varies in its definition whether linked to service event such as a period of hospitalization, or to the overall care independent of where different parts of the care occurred linked to a particular care problem or sub-problem, for example a fractured hip in someone with osteoarthritis. This latter definition of episode of care is preferred as it can be linked to analysis of outcomes and effectiveness of care.

7.3.3.2 Temporal considerations

The time dimension is intimately connected to the notion of process and is a common basis for aggregation, particularly during analysis. The time stamp on individual data items enables consideration of relative time, such as before, during and after. It gives a temporal base to link events to a time period, whether organizational in nature such as during the morning shift on weekdays, or disease management related such as during a hospital visit or episode of care.

Time relationships provide information on:

— states;

— trends;

— sequence.

States include the presence of a risk factor as well as identifying a period in disease evolution such as a state of remission in cancer care or clinical stability in chronic heart failure management. Trends such as progressive change in a diagnostic test can indicate an acute change or precede a disease relapse. Assessing the sequence of events is essential for causal inference.

These considerations can be the basis of relatively complex aggregation. For example, initially data may be converted to simply abnormal or normal; a sequence of normal data over a specified time interval could enable the attribution of stable or instable to a disease condition. Much of this analysis could occur at the time of or after extraction of data from the CDW for more complex data mining, modelling, analysis and data presentation.

7.3.4 Perspective: evaluation and innovation of health procedures and technologies

7.3.4.1 General considerations

The CDW is an important opportunity to support evaluation of innovation of procedures and technologies (P/T). A new P/T is inserted into a process of care. Normally the P/T has been previously evaluated in relation to its specific purpose and validated as being a significant improvement. More complex is to assess how an individual innovation can influence the overall process and outcome of care. Furthermore, a specific evaluation is necessarily limited in its scope, and cannot cover all possibilities of its use in real clinical practice. The CDW potentially offers information on the whole process of care as detailed above, enables comparison before and after P/T introduction and can enable detection of sub-populations who have more benefit or more risk with this innovation.

7.3.4.2 Benefit and cost

Evaluation of technologies is a cost-benefit calculation. The addition of econometrics to the CDW suggests that individual items need to be associated with costs which require a microeconomic analysis which in turn can mean complex considerations for particular items. There are direct and indirect costs. For example, the cost of a single laboratory test clearly depends on a given technique, the personnel to do the test, but also the depreciation of the laboratory equipment and the cost of providing laboratories etc. Costs can be institutional and jurisdictional dependent and complex to standardize.

Macroeconomics and econometric indicators are not the direct consideration of this Technical Specification.

7.3.4.3 Process and economic modelling

From an economic perspective there is a need to relate cost to result. The sequential steps in a care process can be linked to a cost. Although each individual care pathway has its particular characteristics, normative considerations can apply to particular patient conditions. The progressive use of guidelines provides a basis for such normative considerations.

This approach is similar to case mix analysis whereby methods to quantify the severity of illness can be matched to population outcome and is used in the assessment of the organization of care.

A peer-reviewed process model that considers optimization of a care pathway and the relevance of each step in the context of the pathway enables a dialogue between the two levels of care, individual care delivery and organization of healthcare.

The granularity of data in the CDW thus should enable evaluation of care by care providers which in turn should influence evaluation of care organization by healthcare administrators. The peer-reviewed analysis of data of the process of care (care path analysis) by clinicians and its relation to patient outcomes should lead to optimization of practice and in turn change the case-mix data viewed by health administrators.

7.3.5 Perspective: disease surveillance, epidemiology and public health

Populations concerned in these perspectives may be institutional but can cross regions and jurisdictional boundaries with resulting implications for system interoperability and the extent of influence of the CDW. Important dimensions are population, location and defined dimension members such as specific diagnostic tests and prevention interventions or therapies. See Table 4. Pattern analysis considers not only the process of care of a given individual but also cross-sectional views as well as the process of change in relation to population, geography and time of the incidence of the selected disease markers and responses to therapy. Data presentation may use geographical-based techniques for displaying dispersion of disease incidence and how it changes over time.

Table 4 — Dimensions linked to surveillance (example)

CDW dimension candidates	Dimension components	Data sources	Units and standardization
Population	Age; disease category	Reporting centres	Postal code SNOMED CT
Exposure	Environmental hazard	Surveillance centres	Qualitative or quantitative measures of exposure
Public health laboratory results	Patient; environment tests	Laboratory measurements	Test quantitative and qualitative results; test units; LOINC
Interventions	Procedures, therapies	Reporting centres	SNOMED CT
Geography	Region, population density	Reporting centres; surveys	Postal code, measures of population characteristics
Time	Time intervals	Time stamp	Actual or relative time units

7.3.6 Perspective: planning, policy and links to large population databases

Planning and policy tend to take a relatively high level view of the business and may ask questions such as the relation of resources to both clinical and organizational outcomes. Other databases may be usefully linked such as survey databases of socioeconomic and education indicators, or geographical variations in environment.

High level indicators (see 7.4) such as accessibility and continuity of care require aggregation from different dimensions.

The evaluation of resources covers more than one dimension which can include personnel, costs of tests or treatment and bed occupancy. As noted before for costs, summarising costs can vary according to local or jurisdictional context and cost assessment methodology. This could lead to the creation of a composite dimension for resources and the components of the resource hierarchy need to be clearly specified in the CDW metadata as to their relation to the atomic data that are used as sources for the aggregations for each of these components.

7.3.7 Perspective: new knowledge discovery

New knowledge discovery can be hypothesis driven. The hypotheses may stem from different desires such as:

a) to compare practice with literature findings;

b) to identify sub-populations that show variation of treatment response, as yet unconfirmed in the literature;

c) to investigate patterns of pathophysiological reaction to new treatments.

Exploration is supported by the process analysis discussed in 7.2 and potentially benefiting from tools adapted to support this exploration.

Of increasingly recognised importance is the advantage of linking clinical practice to research databases. The precise characterization of sub-populations provides a specification of phenotype that can be matched to genotype evaluated in the same population as measured by different genomic techniques.

7.4 Health system indicators

7.4.1 General considerations

Health system indicators represent a means of measuring and comparing performance. They depend on data aggregation as well as predefined special measures and the CDW is a natural source of health indicator information as the interpretation of indicators is facilitated through the presentation, “drill down”, metadata and other features of the CDW. The principles of good use of health indicators are essentially similar to the principles of good use of the CDW.

The same indicators can serve a range of different interveners and clinical teams to support both clinical policy and organizational decision making. Of evident importance is the display and interpretation of the changes in these indicators over time.

7.4.2 Derivation of indicators

Table 5 has been taken from ISO/TS 21667 describing a conceptual framework for health indicators. Table 5 is a high level view of the framework. It serves to illustrate how different sections of the framework depend on different sources and dimensions of primary data. The use of the word dimension in Table 5 is pertinent to the dimensions of that framework but not necessarily directly equivalent to a data dimension in the CDW.

The choices of data for an indicator can be determined locally. However, given the role of an indicator in enabling performance evaluation many indicators should become the subject of consensus both by peers and by jurisdiction.

Table 5 — Conceptual framework for health indicators

Dimensions		Sub-dimensions					
1	Health status	Wellbeing	Health conditions		Human function	Deaths	Equity
2	Non-medical determinants of health	Health behaviours	Socioeconomic factors	Social and community factors	Environmental factors	Genetic factors	
3	Health system performance	Acceptability Continuity	Accessibility Effectiveness		Appropriateness Efficiency	Competence Safety	
4	Community and health system characteristics	Resources		Population		Health system	

7.4.3 Consensus and indicators

Indicator development and instruments for measurement of outcomes have a similar lifecycle where ideally proposed indicators for a given need are subject to extensive validation prior to use, and evaluation after use.

The organization interRAI (www.interrai.org) is a good example of international collaboration in developing performance indicators.

Such indicators will influence which primary data should be collected and when. In some circumstances aggregated data are used to define both the indicator and the methodology for collecting the underlying source data themselves.

8 Architecture and technology

8.1 Introduction

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 CDW's unique characteristics in terms of architectural elements framed within a background of generalized data warehouse practices, and how to accommodate these within a health information context, is the topic of this section. The approach here, where possible, is to reference and draw on established effective industry practices within the data warehousing arena.

8.2 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 have been drawn from a variety of operational sources and enriched by supplementary value added data. Specific requirements of the CDW stem from healthcare'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 healthcare 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 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 are 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, and service delivery location. Conformance and normalization in the use of codified values, e.g. in classifying interventions and diagnoses, are 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; in other words, data are generally stable in a CDW. More data are added but data are never removed.

— Its performance is important. Data access is optimized both to reduce input/output and to support analysis (i.e. through the use of indices, star schemas, parallelism). Healthcare data can be of 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 the 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 analytical demands of healthcare require data to be stored as an atomic grain, as opposed to being stored as an aggregate.

8.3 Existing work on data warehousing

8.3.1 General considerations

The two prominent proponents of data warehouse architecture are Bill Inmon [5] and Ralph Kimball [6]. Over the last ten to fifteen years, these methodologies have established a solid core of successfully deployed data warehouses. While other methodologies are available (see Reference [9]) these two approaches, also known respectively as “Hub and Spoke Architecture” (Inmon [5]) and “Data Mart Bus Architecture with Linked Dimensions” (Kimball [6]), are predominant. The key requirement for any CDW is to be able to accommodate the highly dimensional and complex nature of healthcare data and their 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 has also been described as either a “subject-oriented, integrated, time-variant and non-volatile collection of data” (Inmon) or, alternatively, as a “copy of transaction data specifically structured for query and analysis” (Kimball). 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”).

For further details on dimensional modelling, see “Dimensional modelling” section; for further details on star schema, see “Star Schema” section. In contrast, “hub and spoke architecture” is composed of an integrated, normalized, relational and atomic hub surrounded by “dependent data marts” drawn from this hub. Dependent data marts are marts that receive their data from a centralized enterprise data warehouse (hub). Access to data is through these data marts (and on occasion directly against the hub). While less prevalent, other architectural configurations are found (see Reference [9]), 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.

While the practice of data warehouse architecture is somewhat polarized between the approaches of Inmon and Kimball, 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 include:

— drug claims;

— inpatient discharge data;

— health expenditures;

— any other data that could respond to user requirements.

The examples given in 7.3 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.

8.3.2 Data modelling

8.3.2.1 Data modelling and approaches

Data models are necessary to fully describe both source data and target CDW constructs. A standard approach to data modelling is required to define with precision how data will be arranged within the CDW and as a means to communicate requirements to stakeholders. Although different data warehousing methodologies place different emphasis on the approach to modelling, best practice requires that at least a conceptual model or reference model be developed for the enterprise of concern. This conceptual model should use a standard notation and entity definitions should be clearly defined and endorsed by the organization's data governance. The development of a conceptual model will assist in data conformance and/or master data management (MDM) (see 6.6.2) through an organization-wide ratification of common healthcare-related concepts.

The following refers primarily to logical and physical data models.

The development of physical models for the CDW is required; the development of logical models is recommended. At this level, most CDW design and deployment efforts will involve dimensional modelling techniques in some instances combined with traditional entity/relation (E/R) normalized models.

Existing examples of CDWs suggest common themes in terms of arrangements of data. Normally these would be modelled and physically deployed as a dimensional model (star schema, constellation — a constellation being a collection of star schemas linked together by common dimensions). Standard E/R models, including object-oriented variants [i.e. a unified modelling language (UML) class diagram] normalized to third normal form (3NF) or above, can also be used throughout the design and implementation phase of a CDW development, particularly as mentioned during 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 should be dimensional in nature. Certainly, specialized requirements may necessitate 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. The more common themes found in CDWs using dimensional modelling are outlined in this subclause.

8.3.2.2 Dimensional modelling

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

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;

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

8.3.2.3 Fact and dimensions

In data warehousing, fact tables store the measures of a business. These measures are quantitative or factual data about a subject. The measures are generally numeric and correspond to the “how much” or “how many” aspects of a business question. Dimension tables contain attributes that describe fact records in the fact table. Some attributes provide descriptive information such as code descriptions. Dimension tables also contain

hierarchies of attributes that aid in the summarization of measures. For example, a dimension containing geographic information would often contain a hierarchy that separates geography into categories such as country, province/state, region, etc. until the lowest level of geography (i.e. postal code) contained in the data warehouse is reached.

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 an element that can be both a dimension (e.g. group by age) or a fact (e.g. average age).

8.3.2.4 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 physical data mart design (with a data mart being a subject-specific subset of a data warehouse), although some methodologies do allow for the direct use of third normal form data. Star schemas are highly descriptive and 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 for the writing of queries that cross more than one fact table. A snowflake schema is an arrangement used to specifically aid in performance in high cardinality (high row count) dimensions, such as the subject of care geography.

The following informally describes the dimensions and facts in the generic illustrative model in Figure 4.

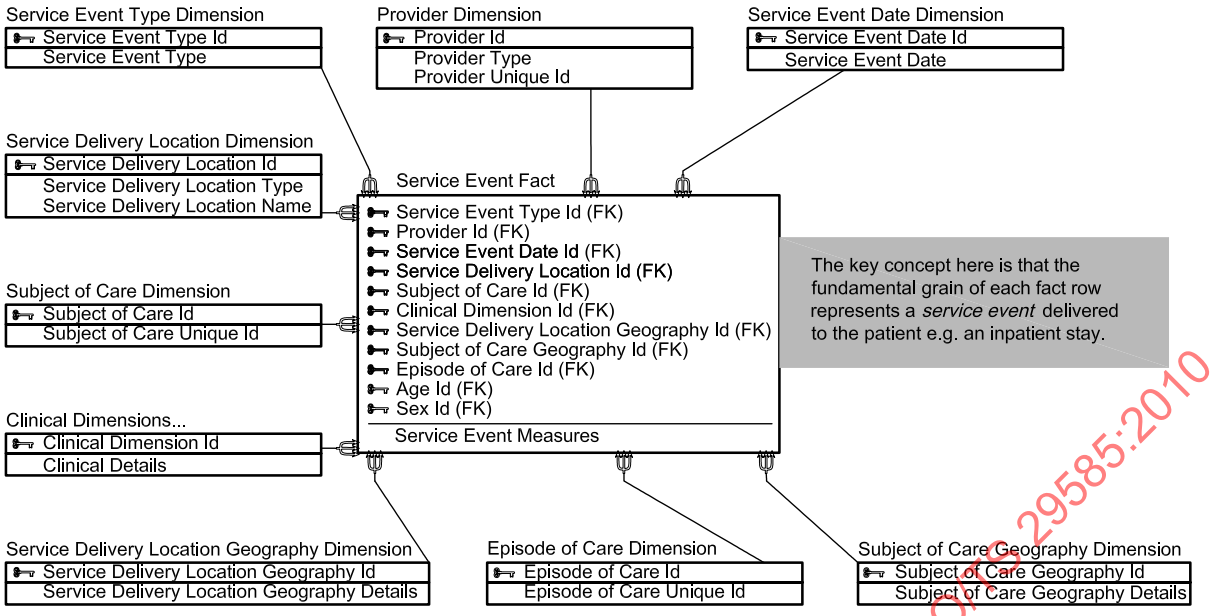

Figure 4 — Example of star schema/dimensional model based on service event

For the purposes of the following example, a “service event” is an interaction between one or more healthcare providers with one or more persons for assessment, care, consultation and/or treatment:

— 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 table 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);

— subject of care geography dimension: describes the geographical location of the subject of care/patient;

— service 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 analyse the fact data);

— 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.

8.3.2.5 Constellation schema

A constellation schema is a collection of star schemas linked together by common dimensions. The following example is based on 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 shown in Figure 5 links a simplified population star schema to the example service event schema in Figure 4. Two additional common dimensions, age and sex, have been added to the service event model 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 analyses 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 5:

— 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);

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

NOTE Common dimensions are sex dimension, age dimension and subject of care geography dimension.

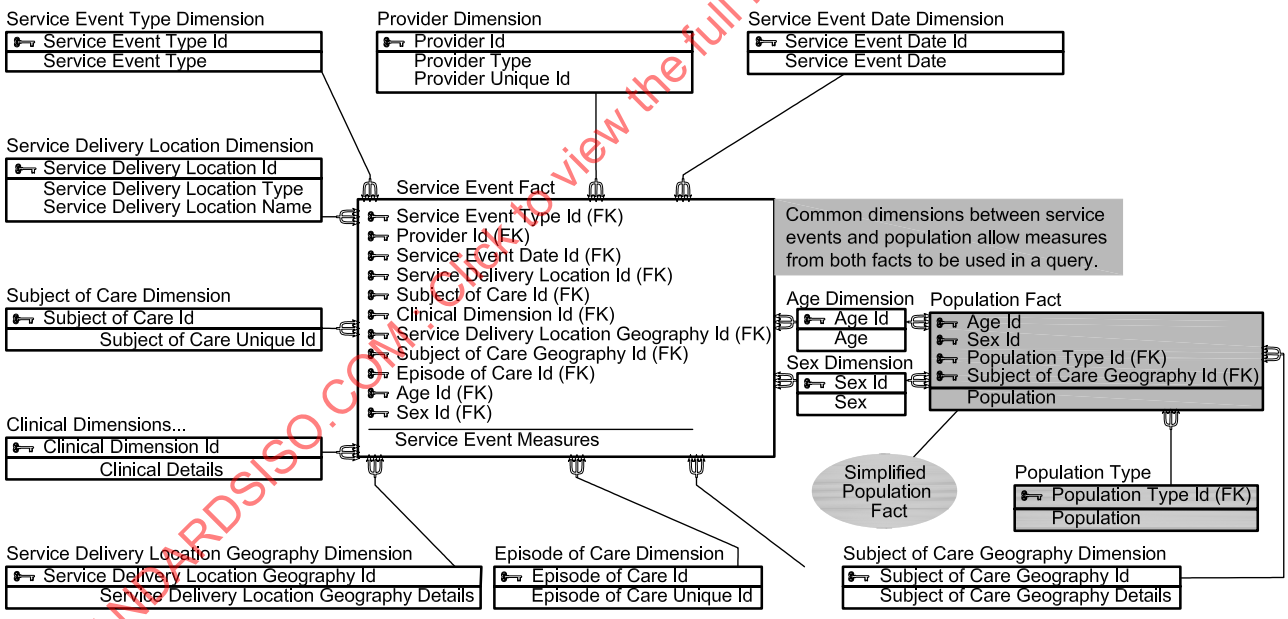

Figure 5 — Constellation schema example linking population and service event

8.3.2.6 Surrogate keys

In most instances the primary key of a dimension table (the column that uniquely identifies each distinct natural key) should 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. (A surrogate key is an artificial or synthetic key that is used as a substitute for a natural key.)

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.

8.3.2.7 Granularity/atomic data

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 should 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 measures 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 (defaulting the value to 1 for each row);

— length of stay;

— distance travelled from home to facility.

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

Regarding “number of discharges”, alternatively this could be a “factless fact” (see Reference [7]) with the number of discharges being determined by a SQL COUNT operation.

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.

To answer the full breadth of healthcare questions, fact data should be stored at an atomic or record level.

8.3.2.8 Conformed dimensions and integration strategies

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 dimensions include date and location. 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.

The evolution of conformed dimensions or master data should be informed by an over-arching information architecture or health information framework.

8.3.2.9 Codes

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 rigour 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; with certain "best of breed" BI tools permitting the binding of codes and their associated descriptions at a metadata level, removing the need for explicit combination fields.

8.3.2.10 Hierarchies

Hierarchies should be formally specified as part of the dimensional modelling process. 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 dependent to a degree on the analytical or BI tool employed.

8.3.2.11 Roles/logical views

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 (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 content, the only difference being the context in which the dates are 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;

— the place where treatment was delivered.

The use of roles as opposed to separate physical tables should be practised in order to improve standardize data within the CDW.

8.3.2.12 Many-to-many relationships

Many healthcare 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 5); an alternative approach is to create linking tables associated with a single fact table that resolves the many-to-many relationships (see Reference [7]).

8.3.2.13 Conformed dimension grid

During the initial stages of gathering CDW requirements, a conformed dimension grid should be developed. This should be in line with an over-arching information architecture or health information framework. 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 grid. 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 6. An "X" in the cell indicates that the dimension is available within the data source.

Table 6 — Example of conformed dimension grid

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

8.3.2.14 Derived values

Derived values (data elements) are a fundamental value-add for any CDW. They represent data that are not contained in whatever source systems are used to populate the CDW but rather are unique to the clinical data warehouse. However, often the components that are used that create the derived values are present in the CDW.

Derived values can be found in either fact tables (in the form of metrics) or in dimension tables in the form of dimensional attributes. The benefits of derived values include:

— a common definition and method of calculation across the entire CDW;

— efficiency for the CDW user, since the values are already present in the CDW there is no need for users to calculate them;

— the derived values are likely indexed as part of the ETL process thereby providing improved query efficiency.

It should be strongly encouraged to persist all derived values in the CDW based on an agreed-upon algorithm (definition) provided by the business. This ensures that all users are looking at the same value for an agreed-upon measure and that two users are not calculating the same value in different ways.

8.3.2.15 Missing value handling

If data values are missing, the CDW should ideally be able to indicate a reason for the data being absent, if this is known. Common domain values for replacement of missing values are, for example: “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. Also refer to 6.6.

8.3.2.16 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.

8.3.2.17 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 8.3.2.6). 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 hospitalization).

Although it falls outside the scope of this Technical Specification, where a consistent and unique subject of care identifier is not available in the CDW, 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.

8.3.3 Performance considerations

8.3.3.1 Defining acceptable performance

The simplest definition of acceptable performance is that the data warehouse is able to meet all of the business requirements in the required time scales. With this definition it is important to categorize the business requirements into those that are reasonable and those that are not. In addition, service level agreements are often negotiated between the data warehouse users and the team responsible for the operation of the warehouse.

Even the best data warehouse initiatives will fail, or have very limited success, if query performance is considered to be unacceptable by its intended users. That being said, acceptable performance usually comes with a cost and should hardware budgets limit the size of the data warehouse machine, then all parties should be aware and modify their expectations accordingly.

Performance service objectives, or if possible, service levels should be developed for the CDW.

8.3.3.2 Data warehouse sizing

Key to ensuring acceptable performance is having an accurate estimation of the size of the data warehouse. This includes its initial size after any legacy data are loaded, and its rate of growth.

In addition, having an estimate of user activity is also important. Questions to ask include: how many users are anticipated to be running queries at any given time (concurrency), at what time of day, and query complexity (what will be the mix of complex and simple queries).

8.3.3.3 Hardware considerations

When configuring a data warehouse hardware environment, it is important to ensure that all related components are balanced. What this means is that all active components (CPU, memory and I/O) are all effectively used and when one component reaches its maximum operating range, then all the other components approach their maximum at the same time.

System scalability is also an important consideration. Even though the environment may have sufficient power to meet initial performance requirements, it needs to be scalable to meet anticipated future requirements, and will it be necessary to acquire a new technical infrastructure?

8.3.3.4 Data organization

How the data are modelled is a key component to ensuring good performance. There are two prevailing data architectures today in data warehousing. One is a Kimball styled data warehouse (facts and dimensions) and the other is an Inmon style warehouse. See 8.3.1 for a discussion on the Kimball and Inmon data warehouse approaches.

8.3.3.5 Indexing

To a large degree, indexing options will be limited by the relational database management system (RDBMS) that the data warehouse is to be deployed on. However, the following are indexing guidelines that should be generally applicable regardless of the RDBMS.

— Index all frequently searched columns. If a column is going to be often used in SQL where the clause exacts it should be indexed.

— Use appropriate index types. There are several common index structures including b-tree and bitmap. They are intended for different types of data. Bitmap indexes work best on data with low to medium cardinality such as gender. If a column has high cardinality such as health number then a b-tree index will be more appropriate.

— Make sure that whatever statistics the RDBMS collects are kept up-to-date.

If available within the RDBMS, use function-based indexes where appropriate. A function-based index includes columns that are either transformed by a function, such as the UPPER function, or included in an expression, such as column1 + column2. With a function-based index, you can store computation-intensive expressions in the index thereby greatly reducing query runtime.

8.3.3.6 Parallelism

Parallelism is the ability of the RDBMS and the hardware environment (multi-CPU servers) to execute multiple queries at the same time (in parallel). Parallel query execution can often reduce runtimes. There are typically many tuning parameters with parallelism so it is important that the various options be well understood and various combinations be tested for optimal performance.

8.3.3.7 Summary management (data aggregation)

Implemented properly, summary management (i.e. data aggregation) can give rise to one of the single biggest improvements to data warehouse performance. Again, the exact implementation options will be RDBMS dependent. However, the following are general guidelines to be considered.

— Conceptually, data aggregation involves storing the result set of a query in a table. When a user-generated query is found to be compatible with the query that generated the data aggregation (in other words, the user query can be resolved with the contents of the data aggregation), the RDBMS query optimizer will rewrite the user query to use the data aggregation. This technique improves the execution of the user query, because most of the query result has been pre-computed.

— Facts of different grains can also be used to increase query performance. Similar to data aggregation described above, having facts of different grains is also a viable option to increase performance. For example, if user queries are often written by state or province, having facts with metrics already at this grain can improve query performance especially if it greatly reduces the size of the root fact tables.

8.3.3.8 Databases and data warehouse appliances

Up until recently servers running RDBMS software, such as Oracle or Microsoft's SQL Server, have been the standard means of implementing a data warehouse. Database servers continue to be widely used, but over the past several years data warehousing “appliances” have come into the mainstream. Rather than the more generic database servers which can be used for any database-related application (OLAP – On-Line Analytical Processing, OLTP – On-Line Transaction Processing), data warehousing appliances are designed specifically to host data warehouses. Their technical and data management architectures are optimized for data warehousing queries. Vendors of these appliances claim significant performance improvements as well as lower support costs [less database administrator (DBA) time required] over traditional database servers. Things to consider when looking at data warehousing appliances include compatibility with ETL tools and front-end business intelligence applications as well as retraining costs for support staff.

8.3.3.9 Ongoing monitoring

Once in a production environment, it is important for the DBA to routinely monitor the resources for the database environment and to look for any resource limitations. With ongoing monitoring, the DBA will detect if the database performance level has reached a plateau. To increase the performance beyond this plateau may require the addition of more hardware resources or reconfiguration of the system software (operating system, RDBMS or BI application). The following are some of the more common types of resource issue:

— hardware related (CPU, memory, disk I/O): these potential performance issues are typically easy to detect and straightforward (if sometimes expensive) to resolve;

— RDBMS related: unlike hardware issues, RDBMS performance issues are typically more difficult to identify and resolve; they usually require the availability of a knowledgeable DBA (preferably with data warehouse tuning expertise).

8.3.4 Extract, transform and load

8.3.4.1 General description

Extract, transform and load (ETL) in common database parlance, and particularly in data warehousing, involves extracting data from data sources, transforming those data to meet operational needs and then loading those data into the data warehouse.

8.3.4.2 Tools-based ETL versus hand-coded ETL

There are two fundamental approaches to creating the necessary ETL processes to load and maintain a clinical data warehouse. One approach involves having developers manually write the ETL scripts (hand-coded) and the other involves the use of an ETL tool which essentially writes the ETL scripts based on how a developer sets up the relationships between source and target data and any required data transformations via a graphical user interface (GUI).

There is no straightforward answer as to which approach is best. An informed decision can best be made by having a thorough understanding of the data warehouse project requirements, skill level of the resources available and budgetary constraints. It should be noted that, as a general rule, hand-coded ETL solutions are typically better suited to smaller-scale CDW implementations.

The following summarises the general high-level advantages of each approach.

Tools-based ETL:

— typically development is simpler, faster and less expensive; tools costs may be offset in large and/or complex projects;

— many ETL tools have integrated metadata repositories;

— ETL documentation can often be generated automatically from the metadata repository;

— pre-built “connectors” for heterogeneous environments;

— typically good performance even for very large datasets;

— change impact analysis.

Hand-coded ETL:

— automated unit testing tools are available in a hand-coded system, but not with all tools-based approaches;

— to more directly manage metadata in hand-coded systems but the systems (repositories, reporting) need to also be built;

— not necessarily limited by the abilities of the ETL tool; by its nature, hand-coding can be done with any programming language(s), but all mainstream ETL tools allow “escapes” to standard programming languages in isolated modules;

— hand-coded ETL provides unlimited flexibility and allows for the use of one or more programming languages.

8.3.4.3 Source to target mapping

Source systems for data warehouses are typically transaction processing applications. For example, a hospital patient management system would typically serve as a data source for that hospital's clinical data warehouse.

Target systems include all the necessary database objects such as tables and views which contain the various dimensional attributes and measures of the data warehouse.

Mappings are a series of operations that extract data from sources, transform the data, and load them into targets. They represent the flow of the data and the operations performed on those data.

The extraction and transformation process can encounter many challenges. For example, if the source system is complex and/or poorly documented, then determining which data to extract can be difficult. Also, typically source systems cannot be modified, nor can performance or availability be adjusted. To address these challenges it is advisable to first start by understanding the source system(s) as much as possible. Once there is a thorough understanding of the source data, an appropriate ETL strategy can be designed.

A source to target map should be produced during the ETL phase of the CDW's development.

8.3.4.4 Process flows

After mappings have been designed which define the operations for moving data from sources to targets, process flows are typically defined. Process flows describe dependencies and inter-relate mappings between the ETL system and external activities such as email, FTP – File Transfer Protocol, and operating system commands.

Each process flow begins with one activity and concludes with another activity for each stream in the flow. For example, if a mapping completes successfully, an email notification is sent and another process is launched. If the mapping were to fail, an email notification would be sent indicating the failure and the process flow would end.

8.3.4.5 Performance considerations

Typically, ETL processes are structurally complex. Often they are composed of numerous sources and targets and complex data transformations. Also, ETL performance tuning is highly dependent on the ETL technology being used. In a general sense, however, the following items should be considered:

Set-based versus row-based operations:

— Set-based is typically a single SQL statement that processes all data and performs all operations; although processing data as a set improves performance, the auditing information available is limited. In set-based operations you cannot view details on which rows contain errors;

— row-based inserts each row into the target separately so full auditing capabilities exist; use row-based along with fast, set-based operations to extract and transform the data but need extended auditing for loading the data, which is where errors are likely to occur.

Database commit options:

There are two basic approaches to committing data during an ETL mapping. Manual commit control enables users to specify when data are committed. This means that users can view the number of rows inserted and other auditing information before issuing the commit or rollback command. Automatic commit loads and then automatically commits the data based on the mapping design.

8.3.4.6 ETL-related metadata

Metadata are important to all those involved in a clinical data warehouse project. For business users metadata are used to help understand the warehouse's data structure and the business rules used to create and maintain it. They help users understand the origins of the data, the sources the data came from and any applied transformations.

For those responsible for building and maintaining the data warehouse (ETL developers, ETL administrators, data quality staff), metadata are key to change management. Metadata are used to evaluate the impact of changes to the data structures and the loading processes. Thorough metadata greatly reduce the maintenance workload of the IT team and assist with the evolution of the data warehouse.

ETL tools typically have repository browsers. These browsers allow users to query and report on the metadata contained in the repository. Typical reporting topics include summary reporting, detailed reporting, implementation reporting and lineage and impact analysis reporting. When evaluating ETL tools, repository reporting capabilities should be a key consideration. As mentioned in the “tools-based ETL versus hand-coded ETL” section, anyone opting for hand-coded ETL must also build their own metadata repositories and associated browsers.

8.3.5 Load frequency

One operational question common to all data warehousing initiatives, be they clinical or not, is how frequently to update the data warehouse. The correct answer to that question should be based on business need.

The delay from the time an event of interest occurs (for example, a patient discharge from a hospital) to the time that information is made available in the data warehouse is typically referred to as data “latency”.

Two common terms used in the data warehousing industry when discussing appropriate latency are “real-time” and “right-time”.

“Real-time” referrers to zero-latency. Near zero-latency would be more accurate, however, since there would always be some latency from the time the event of interest occurs to the time the data are available in the CDW (even if that delay is only a few milliseconds).

“Right-time” refers to a latency period that is based on the needs of specific business processes.

Not surprisingly, it is generally considered best practice to base CDW latency requirements on business need. Focusing on “right-time” is generally thought to lead to more effective implementations than technology-driven solutions striving for the fastest “real-time” data. However, it is important to consider future needs in the design of the data acquisition architecture for a data warehouse. A well designed architecture will allow for evolving business requirements which typically will require decreased latency.

It is important to implement a scalable solution to support decreasing data latency as business processes mature. Rewriting or rearchitecting a data acquisition infrastructure with the goal of decreasing data latency due to lack of foresight can be a significant undertaking. However, an over-engineered initial implementation can be just as large an undertaking. The ideal solution is a scalable architecture (preferably with no code rewrite required) that allows for the correct amount of data loading capacity at each stage of an organization's CDW evolution.