1. Introduction

Medical data are at the centre of many healthcare-related activities, ranging from providing medical care for each individual patient to supporting scientific research and education. This review offers an introduction to medical data structures and management. Examples of coding and classification systems are also presented.

2. Medical data

Medical data include a broad range of data types, including text-based alphanumeric data, recorded physiological signals and medical images. In addition to traditional clinical data, recent progress in medical science and technology has led to the accumulation of a tremendous amount of genomic data in support of individualised healthcare and personalised therapies. In this section, the characteristics of typical clinical data are introduced, followed by a brief introduction to genomic data.

2.1. Clinical data

In general, there are three types of clinical data in the practice of medical and health science [12,14].

- Alphanumeric data are typically recorded as unstructured, free text in the medical record. They account for a large portion of the information that is gathered verbally in the care of patients, such as patient demographics, a description of patients’ illnesses and responses to physician’s questions. Results of physical examination such as vital signs and laboratory tests, and other narrative descriptions, such as treatment plans and surgical procedures, also belong to this category.
Physiological signals, captured from the human body by using appropriate biomedical sensors, are usually represented by a series of values along two dimensions with the x-axis representing time and the y-axis displaying voltage values associated with the underlying physiological activity. Advances in physiological signal acquisition techniques have generated a vast quantity of physiological recordings, including electrocardiogram (ECG), electroencephalogram (EEG), gait, respiration, blood pressure tracings, and oxygen saturation. As illustrated in Fig. 1, an ECG records the electrical activity of the heart as a function of time. Each ECG beat represents one cardiac cycle. By convention it includes the P wave, the QRS complex, T wave, and sometimes a small U wave may be seen following the T wave.

Medical images, as illustrated in Fig. 2. With the ability to reveal the areas of interest in great detail, medical images are a valuable resource in clinical diagnosis and medical research. Radiological images like X rays, computed tomography and magnetic resonance imaging techniques are the most well known common types. Using the electromagnetic spectrum, radiological images provide a means to visualize the condition of internal organs or an area that is not externally visible. Optical acquired images may also be produced by skin photography and endoscopy. Instead of using electromagnetic radiation, ultrasound imaging is based on mechanical energy in the form of high-frequency sound waves.
2.2. Genomic data

Advances in biological science are fostering a new clinical practice, where clinical diagnosis and treatment will be supported by the information encoded in relevant genomic data, such as DNA (deoxyribonucleic acid) sequences, protein structures and microarray gene expression data. Examples of genomic applications in medical research can be found in the Programs for Genomic Applications (PGAs) funded by the National Heart, Lung and Blood Institute (NHLBI) of the NIH (National Institutes of Health) (http://www.nhlbi.nih.gov/resources/pga/).

Each type of genomic data has a unique data structure. For instance, DNA may be represented by a linear sequence from the alphabet \{A, G, T, C\}. Each of these characters refers to a nucleotide. A typical microarray experiment dataset includes expression levels of thousands of genes in a number of experimental samples (conditions). These samples may correspond to different conditions or serial time points taken during a biological process. An expression dataset can be summarised by a matrix, in which the horizontal rows represent genes, one gene in each row, and the vertical columns contain the various samples corresponding either to the time points or to various conditions, with one sample in each column. Thus, various models of electronic medical records have been developed to integrate traditional clinical data and diverse genomic data [11].

The strong interest in integrating genomic data into clinical research is driven by the hope that it will result in the prevention and diagnosis of complex genetic disease and the design of highly targeted therapies [2]. For example, with DNA microarray technology, scientists are able to measure gene expression of thousands of genes under selected physiological conditions at a given time. Thus, by comparing expression profiles of an individual gene under various conditions or by comparing expression profiles of samples to understand their similarities and differences and to find the relationships among genes and samples, many important physiological and medical phenomena can be studied at the molecular level.

2.3. Characteristics of medical data

Apart from privacy and confidentiality requirements, medical data have important and unique features. For example, a patient record is a mixture of opinion and factual data generated from a wide range of clinical professionals. Some of these data, such as observations and interpretations, are difficult to record in a structured form. Recorded physiological signals and medical images are contaminated by noise generated from the machine recording the data or from the patient. For instance, an ECG record could be contaminated by physiological variability, baseline wander noise, 50 or 60 Hz power line interference, and muscle noise [9]. Due to the quantum noise inherent in photon detection, images generated by computed tomography-based techniques such as positron emission tomography (PET) and single photon emission tomography (SPECT) have a high noise content [10].

3. Data structures and management

The characteristics of medical data pose a great challenge to efficient data storage and management. In this section, electronic patient records are described. The examples of management and storage of medical images and physiological signals are introduced respectively.
Fig. 3. An EPR provides an integrated view of patient data and has the flexibility to generate task-specific views on all the available data to meet the needs of different tasks.

3.1. Electronic patient records

A patient record contains all the clinical data relevant to the care of one patient. Given the increasing volume and scope of medical data available nowadays, traditional paper-based patient record systems are not sufficient to address many issues related to efficient medical data management. The Electronic Patient Record (EPR) is revolutionising storage, management and communication of medical information. With EPR comes all related medical data, including patient history, physical examination, drug prescriptions, and recorded biological signals, linked into one record that can be easily accessed for different purposes. The ability to store all patient-related medical information in one record, which may be accessed at any time or from any place, ultimately contributes to improvements in healthcare delivery.

It has been shown that an EPR-based system offers several significant advantages in the process of data management and analysis [12,15]. One of the most obvious benefits is that the speed of data retrieval is significantly improved. A physician should be able to search through thousands of records for specific attributes such as medication or diagnosis with little effort. With structured data entry, medical data collected in an EPR are well structured and organised, allowing the electronic sharing of patient recording among different clinical professionals. Moreover, an EPR-based record system has the inherent flexibility to provide task-specific views on all the available patient data to meet the needs of different clinical tasks, as illustrated in Fig. 3.

One of the key components in an EPR is a database management system (DBMS) [15]. DBMS is a software system designed to manage the storage, retrieval, and organization of data in a database. Different data models have evolved in recent years to structure the logical view of the DBMS, including relational, hierarchical, text-oriented, and object-oriented data models. Among these, the relational data model is most widely used.

A relational data model, developed by E.F. Codd [1], is one in which the data and relations between them are organised into a series of tables. Each table is made up of columns and rows. All values in a given columns have the same data type, which could be numeric or textual. All values in a row
conceptually belong together. Typically, each row, known as a record, is uniquely identified by a primary key which may contain a single column or multiple columns in combination. Figure 4 shows an example of a relation database with two tables: PatientDemographics and PhysicalExamination tables.

3.2. Medical images – DICOM

Digital Imaging and Communications in Medicine (DICOM) is the healthcare industry standard for medical image data file storage and transfer. It was first conceived in 1983 by the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) and first known as the ACR-NEMA standard. This standard was revised in 1993 and named DICOM. The standard defines principles and methods for the storage and transfer of medical images (mainly radiological) in a multi vendor environment (http://medical.nema.org). Virtually all radiological imaging modalities are now digital and a large teaching hospital may generate around tens of Terra ($10^{12}$) bytes of image data per annum. The use of medical imaging is increasing, as is the resolution of imaging systems, therefore we can expect to see the computer network and storage requirements also increase.

The main DICOM functions are to transmit and store medical images, to enable searching and retrieval, to facilitate printing of hard copy, to support workflow management and to ensure quality and consistency of image presentation [4]. The DICOM file format is well defined and contains two specific sections. The first part of the file, called the header, holds patient demographics, hospital information and details of the referring physician. Also, the header contains detailed data on the actual specific imaging modality parameters for each patient scan. For example, in a magnetic resonance (MR) image data file, the slice thickness, field of view, pulse sequence details, table position, and scan date are stored. The header also contains hidden fields used by the equipment vendor to store proprietary information on image reconstruction algorithms. Another feature of the DICOM header is that it contains a unique identification number that identifies the scanner manufacturer, the individual scanner, the modality types and the patient study number. This is called the unique identifier (UID). DICOM data files written to storage media like DVD or to another computer will be named using the UID as part or the entire file name.
Figure 5 shows a screen shot of part of a DICOM image header from OSIRIS developed by The University Hospital of Geneva, Switzerland. Osiris is a free DICOM viewer available at http://www.sim.hcuge.ch/osiris/01_Osiris_Presentation_EN.htm. The second part of the DICOM data file is the image pixel data where pixel values are stored using between 4 and 32 bits, depending on the imaging modality. A typical file size for an MR image is 0.5 MBytes (one patient examination may contain up to 1000 images), whilst a Computed Radiography chest X-ray may be 32 MB. The DICOM standard includes the functionality to compress medical images using lossless or “lossy” techniques [5]. JPEG is the common standard achieving a lossless compression of approximately 3:1. Further compression is possible, however, there is some reluctance to reduce file size further as diagnostic integrity of the original image may be lost after significant compression.

The benefit of DICOM is the ability to connect a number of different vendor medical imaging modalities to a Picture Archive and Communication System (PACS). This is the main image database and transmission system in the health environment. This system enables multiple copies of images to be available at any time and at any place that has the appropriate network connections and permissions.
### SCP-ECG data structure: a 6-byte record header + 12 data sections

- **Record header**: consists of a 2-byte CRC followed by a 4-byte record length
- **Section 0 (mandatory)**: contains pointers to the start of the following each section
- **Section 1 (mandatory)**: contains information of general interest
  - Patient demographics: name, age, date of birth, sex, ...
  - ECG acquisition data: acquiring institution, analyzing machine ID, ...
- **Section 2 (Optional)**: contains all the Huffman tables used in encoding of ECG data
- **Section 3 (Optional)**: ECG lead definition
- **Section 4 (Optional)**: QRS locations if reference beats are encoded
- **Section 5 (Optional)**: encoded reference beat data if reference beats are stored.
- **Section 6 (Optional)**: "residual signal" that remains for each lead after reference beat subtraction.
  - Otherwise the entire rhythm signal
- **Section 7 (Optional)**: global ECG measurement data and a list of pacemaker spike measurements
- **Section 8 (Optional)**: textual diagnosis from the "Interpretive" device
- **Section 9 (Optional)**: manufacturer specific diagnosis and overriding data from the "Interpretive" device
- **Section 10 (Optional)**: lead measurement results
- **Section 11 (Optional)**: universal statement codes resulting from the interpretation

Fig. 6. Structure of a SCP-ECG record.

PACS also enables the creation of large digital libraries for teaching and research.

### 3.3. SCP-ECG

The SCP-ECG (the Standard Communications Protocol for Computer-Assisted Electrocardiography) was developed by the Comité Européen de Normalisation Technical Committee 251 (CEN/TC251) project team in 1993 exclusively for the purpose of the exchanging, encoding and storage of 12-lead ECGs [3]. The data level in the SCP-ECG standard includes the ECG signal data, patient demographics, ECG measurements, ECG interpretation results, as well as schemes for ECG waveform compression. In the SCP-ECG standard, various parts of data related to an ECG are specified in different data sections with different encoding forms as illustrated in Fig. 6.

### 3.4. XML based systems

Due to its ability and flexibility to define markups for specific type of data, XML (eXtensible Markup Language, http://www.w3.org/xml/) and its related techniques have gained great attention in medical data representation and management. Currently, various organizations within healthcare such as HL7 (health level 7) and ASTM (American Society for Testing and Materials) are working on recommendations of XML-based e-healthcare technologies within the medical domain. For example, HL7 has published the Clinical Document Architecture (CDA), which uses XML as the representation format for medical and administrative data (http://xml.coverpages.org/ni2004-08-20-a.html). Along with the CDA, HL7 version 3 messaging standard has also been developed using XML as the standard exchange format. The use of XML syntax for the exchange of electronic patient records was illustrated in the
EU project Synapses (https://www.cs.tcd.ie/synapses/public/) and its implementations [8]. The application of XML in DICOM can be found in several related projects [13]. The U.S Food and Drug Administration (FDA) Centre for Drug Evaluation and Research has proposed recommendations for the XML-based techniques for the exchange of time-series data. In December 2003, the FDA XML format was finalized and the Annotated ECG (aECG) format is now part of the HL7 family of standards (http://www.hl7.org/v3annecg/foundationdocuments/welcome/index.htm).

4. Coding and classification

Clinical concepts can be represented and described in many ways. Consequently, standards for codes/terminology have become an essential element in the development of electronic patient records. In this section, several widely-used coding and classification systems are introduced.

4.1. The International Classification of Diseases

The International Classification of Disease (ICD) is a coding system published by the World Health Organization (WHO, http://www.who.int/en/). Its main purpose is to allow morbidity and mortality data to be collected from around the world in a standard format. Since its first edition was published in 1900, ICD has become the most widely used pathology classification system and de facto reference point for many healthcare terminologies [2]. The code has been updated approximately every 10 years to reflect the medical advances in the past years and the current version is ICD-10.

The classification of ICD-10 is structured into 21 chapters, identified by 21 Roman numerals: I, II, III, IV, V, VI, VII, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI, XVII, XVIII, XIX, XX, and XXI. Among these, Chapters I-V, XV-XVII and XIX cover special diseases, such as neoplasms, pregnancy, childbirth and the puerperium, and injuries, poisoning and certain other consequences of external agents. Diseases associated with body systems are found in Chapters VI to XIV. For example, Chapter VI covers diseases of the nervous system, and diseases of the genitourinary system are assigned to Chapter XIV. Symptoms, signs and abnormal clinical and laboratory findings are given in Chapter XVIII. Other external factors affecting morbidity and mortality are included in Chapters XX and XXI.

Each chapter is divided into blocks and categories. Unlike ICD-9, in which diseases are coded with three digits, category codes in ICD-10 start with an alphabetical character followed by two numeric characters, ranging from A00 to Z99. For instance, code I01 represents rheumatic fever with heart involvement. This greatly increases the number of available codes, which can be used to code common diseases in more detail.

Most 3-character categories in ICD-10 are further subdivided into up to 10 subcategories by adding a fourth digit after a decimal point. This can be used, for instance, to classify varieties of a disease, e.g. as illustrated in Fig. 7, acute rheumatic pericarditis is coded with I01.0, while acute rheumatic endocarditis is represented by I01.1.

4.2. The clinical terms (Read Codes)

Read Codes were initially introduced into the UK in the middle 1980s as a 4-byte set terminology to summarise patient treatment, including clinical process and administrative data, in primary care. This version only consists of about 40000 codes. Its subsequent Version 2 released in 1990 was based on the content of ICD-9 and Classification of Surgical Operations and Procedures, Fourth Revision (OPCS-4).
Read Codes in this version were extended to 5-digit alphanumeric characters, each representing one level of hierarchy. While being easy to understand and implement, such a rigid hierarchical classification structure poses some difficulties in the representation of clinical terms. For example, without giving two Read Codes, a clinical term can only be located within one hierarchy.

Due to increasing demands from clinical professionals, Read Codes were further expanded by the Clinical Terms Project, a joint project established by the UK’s Conference of Medical Royal Colleges and the government’s National Health Service (NHS). As a result, the Clinical Term Version 3 (CTV3) was developed and first released in 1994.

In CTV3, a concept refers to any clinical disorder, procedure or observation and can be described by more than one clinical term. Like its previous version, CTV3 adopts a 5-digit alphanumeric code, giving more than 650 million possible codes to represent clinical concepts. However, CTV3 implements a new, more flexible structure to address the problems caused by its early version. For example, the coded-based hierarchical structure has been abandoned. A Read Code is only used to uniquely identify a concept and no longer represents its hierarchy position. Instead, hierarchical relationships between concepts are defined by a set of parent-child links described in a Hierarchy file.

Core release files included in CTV3 are Concept file, Term file, Descriptions file, Hierarchy file, key file, Specialty files, Cross-mapping files, Template file, Redundant codes mapping file and Description Change file. The structures and examples of Concept file (describing concepts), Term file (describing terms), Descriptions file (linking terms to Read Codes) and Hierarchy file (describing a hierarchy) are illustrated in Fig. 8. The reader is referred to [6] for a more detailed description of CTV3.
Fig. 8. The structures and example of (a) Concept file; (b) Term file; (c) Descriptions file; and (d) Hierarchy file.
4.3. Systematized Nomenclature of Medicine

The Systematized Nomenclature of Medicine (SNOMED), developed and maintained by the College of American Pathologists, is generally regarded as one of the most comprehensive standardized medical coding systems available today (http://www.snomed.org/). It uses a hierarchical, multi-axial method to encode clinical data, allowing for the representation of various aspects of a disease. For example, SNOMED International is organized around 11 independent axes: T (Topography), M (Morphology), L (Living organisms), C (Chemical), F (Function), J (Occupation), D (Diagnosis), P (Procedure), A (Physical agents, forces, activities), S (Social context), G (General syntactic). Each axis represents a unique hierarchical classification system. Thus, a disease in SNOMED can be described using a morphologic code, a topographic code, an etiological code, and a function code. Moreover, these codes across different axes can be cross-referenced, leading to a better understanding of each code. For example, the disease pneumonia could be coded as T-28000 (topology code for lung), M-4000 (morphology code for inflammation) and L-25116 (etiological code for Streptococcus pneumoniae).

To reflect the advances in medical informatics, SNOMED has evolved further with the release of SNOMED RT (Reference of Terminology) in 2000. To provide a unified international language that supports the electronic patient record and decision support system, the College of American Pathologists and the UK NHS announced their intention to unite SNOMED RT and CTV3. As a result, SNOMED CT (Clinical Term), which incorporates the content and structure of CTV3, was released in 2002.

One of the important features of SNOMED CT is that it utilizes description logic to describe the scope of a concept. Its core content includes concepts, descriptions and the relationships between them [2]. As illustrated in Fig. 9, a concept in SNOMED CT can be described by terms in one or more descriptions. SNOMED CT contains more than 300000 concepts with unique meanings and formal logic-based definitions, which are organized into top-level hierarchies such as Clinical findings and Body Structure. The relationships between concepts are described in relationship tables. There are two types of relationships in SNOMED CT. While IS-A relationships connect concepts within the same hierarchy, Attribute relationships describe the relationships between concepts located in different hierarchies. For example, the relationships between the disease concepts: Arthritis of knees, Arthritis and Arthropathy belong to IS-A relationships. Examples of Attribute relationships in SNOMED CT include Finding site, Procedure site, Associated morphology, and Method. The relationship between the concepts Appendicitis that is a disease concept and Inflammation that belongs to the Body structure hierarchy can be described using Associated morphology.

The current version of SNOMED CT includes more than 300,000 clinical concepts, 770,000 descriptions and 900,000 semantic relationships. SNOMED CT can also map to other medical classification systems such as ICD-10.

5. Case study: An XML-based representation of ECG data

The ecgML model, a markup language for supporting representation and coding of ECG data, was first published in 2003 [17]. Based on XML, ecgML offers several advantages over existing ECG coding systems. For example, it includes all the components required for ECG representation. The terms and structure used can be further expanded or reviewed at any stage. This protocol can be used as a common coding platform between different ECG acquisition devices and visualization programs.
5.1. Hierarchical presentation of ecgML

The main components included in ecgML are one PatientDemographic, optional element MedicalHistory, one or more Record components, and an optional Diagnosis element, as illustrated in Fig. 10. To facilitate the inclusion of multiple time-related patient’s ECG data, each Record element, which consists of zero-or-one RecordingDevice, zero-or-one ClinicalProtocol, and one-or-more RecordData, is uniquely identified by AcquisitionDate and AcquisitionTime. As a key component in ecgML, each RecordData includes three main subcomponents (Waveforms, Annotations and Measurements) to represent original ECG waveform data, the annotations and measurements respectively.

ecgML incorporates several international standards to describe relevant information. For example, it applies the DICOM lead labeling format to define channel names associated with RecordData element. The Unified Code for Units of Measure (UCUM) scheme is used to define measurement units when appropriate. Moreover, different coding and classification schemes can be incorporated to define medical terms at different levels. One example is to utilize terminologies included in SNOMED CT to describe the diagnostic interpretation of the ECG (Diagnosis element) and medical history of patient’s clinical problems (MedicalHistory element).

A portion of an ecgML-based ECG recording generated from the MIT-BIH Arrhythmia Database is illustrated in Fig. 11.
5.2. Encoding ECG waveform data in ecgML

ECG waveform data are a key component of an ECG record. Given that a wide range of ECG recording devices are available today, the size of waveform data can vary dramatically. Moreover, they are usually expressed in binary format. In order to accommodate the full spectrum of ECG data, ecgML provides the following three options to handle ECG waveform summarized in Fig. 12.

- For large data files, typical of Holter recording, ECG data are maintained in an external file, which can be referenced in ecgML using a FileLink element;
- Using RealValue elements to directly include actual waveform values in ecgML as the content of the element;
- To encode binary waveform data using a BinaryData element. ecgML is based on XML techniques to encode and represent ECG records. Due to the valid-character restriction posed by XML specification, simply embedding binary ECG waveform data within ecgML may cause the parser to encounter invalid sequences and fail to achieve its intended goal. Thus, each BinaryData element is associated with a specified encoding scheme, which may be Base64 or hexadecimal.

5.3. Accompanying tools

A series of Java-based, user-friendly tools are being developed to assist users in exploiting ecgML-based application [16,18]. These include ecgML generator, ecgML parser and ecgMLBrowser.
The ecgML-generator produces ecgML-based ECG record from existing ECG databases. While the ecgML-parser allows the user to read the ECG record and access their contents and structure, ecgML-browser provides onscreen display of the collected waveform data as shown in Fig. 13. The hierarchical structure of the ecgML-based ECG record is displayed on the left hand side. It can be expanded and shrunk at any level. The waveform data are shown on the right hand pane. The ecgML-editor allows an authorized user to modify the contents of an ecgML-based ECG record. Some components such as the raw waveform data are not allowed to be changed.

Fig. 11. An example of an ecgML-based ECG representation.
Fig. 12. Framework for handling ECG waveform data in ecgML.

Fig. 13. A screenshot of ecgML browser.
6. Summary

The above coding systems are developed for a specific purpose – they are goal-oriented. For example, ICD takes a pathophysiological and aetiological view of medicine and classifies diseases primarily according to the organ system involved with some important exceptions such as infectious diseases [7]. It is unable to fulfill the needs of all users.

Another problem is that the terms are changing over time, thus the meaning of a same term can be different over time, new terms are created and old terms may be replaced. When the meaning of a standard code is changed, the data will be interpreted incorrectly. When a standard code is removed from the coding system, the data are no longer interpretable. Additionally there is a lack of comprehensive, standardized medical terminology, and highly skilled experts are required to interpret the systems.

Despite these limitations, coding and classification systems have been successfully applied in health care. A widely accepted coding system is essential for storing and exchanging patient records efficiently.

Glossary

ACR   American College of Radiology
ASTM  American Society for Testing and Materials
CDA   Clinical Document Architecture
CEN/TC251 Comité Européen de Normalisation Technical Committee 251
CTV3  Clinical Term Version 3
DBMS  Database Management System
DICOM Digital Imaging and Communications in Medicine
DNA   DeoxyriboNucleic Acid
ECG   ElectroCardioGram
EEG   ElectroEncephaloGram
EPR   Electronic Patient Record
FDA   U.S Food and Drug Administration
HL7   Health Level 7
ICD   International Classification of Disease
MR    Magnetic Resonance
NEMA  National Electrical Manufacturers Association
NHLBI National Heart, Lung and Blood Institute
NHS   National Health Service
NIH   National Institutes of Health
OPCS-4 Classification of Surgical Operations and Procedures, Fourth Revision
PACS  Picture Archive and Communication System
PET   Positron Emission Tomography
PGA   Programs for Genomic Application
SCP-ECG Standard Communications Protocol for Computer-Assisted Electrocardiography
SNOMED Systematized Nomenclature of Medicine
SNOMED CT Systematized Nomenclature of Medicine Clinical Term
SNOMED RT Systematized Nomenclature of Medicine Reference of Terminology
References