MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Domain Analysis and Modeling to Improve Comparability of Health Statistics

Mihoko Okada[a], Hideaki Hashimoto[a], Takashi Ohida[b]

[a] *Department of Medical Informatics, Kawasaki University of Medical Welfare*
[b] *National Institute of Public Health*

Abstract

Health statistics is an essential element to improve the ability of managers of health institutions, healthcare researchers, policy makers, and health professionals to formulate appropriate course of reactions and to make decisions based on evidence. To ensure adequate health statistics, standards are of critical importance. A study on healthcare statistics domain analysis is underway in an effort to improve usability and comparability of health statistics. The ongoing study focuses on structuring the domain knowledge and making the knowledge explicit with a data element dictionary being the core. Supplemental to the dictionary are a domain term list, a terminology dictionary, and a data model to help organize the concepts constituting the health statistics domain.

Keywords:

health statistics; data element; data element dictionary; domain modeling; domain analysis; XML

Introduction

To facilitate flexible and effective use of health statistical data, a formal study on summary tables was made previously [1,2]. In this study, a general structure and meta data of summary tables were defined, and a statistical summary table management system was developed. The previous study focused on the use of summary tables for a given statistical survey. As an extension to the previous study, domain analysis of health statistics has been made and a domain model (in a broad sense) has been developed. The goal of the present study is to make health data more sharable and to improve comparability of health statistics by standard data elements derived from a domain model. In this paper, we present the method for domain analysis, and an obtained domain model on health statistics including domain knowledge description, domain terms, a data element dictionary, and a domain model in a narrow sense.

Two Levels of Health Data Sharing

Health information is expected to become more sharable as a result of a recent move from paper-based to electronic-based health records. It is of critical importance, however, to distinguish between the two levels of health information sharing (Figure 1). The first level is information sharing for clinical practice, and the second is that for medical research, epidemiology, management, policy making, and so on. Health informatics standards for exchanges of patient records, including standards for messaging, terminology, and medical images, are for the first level health information sharing. In the second level, interests are not in individual patients but in a group (a population in terms of statistics) as a whole. Although information sources overlap each other largely, the nature of and the required standards for the second level information sharing are entirely different from those of the first. Domain analysis was applied and a domain model including a data element dictionary was developed in order to promote data sharing of the second level.

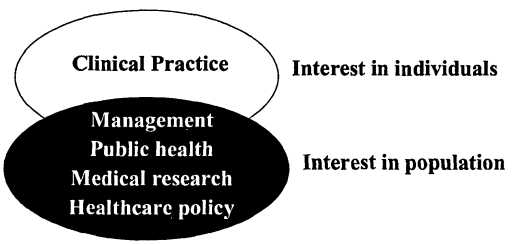

Figure 1 – Two levels of health information sharing

Domain Analysis of Health Statistics

Domain analysis is an activity developed in the field of systems engineering. Domain analysis, first introduced in the 1980s, is an activity within domain engineering and is the process by which information used in developing systems in a domain is identified, captured, and organized with the purpose of making it reusable when creating new systems [3]. The term "domain" is used to denote a set of systems or functional areas, within systems, that exhibit similar functionality. Domain analysis affects the

maintainability, understandability, usability, and reusability characteristics of a system or family of similar systems.

Process and Deliverables

To improve comparability of health statistics and to facilitate sharing of health information, domain analysis was applied. Several domain analysis methods exist, but there is no standard definition of domain analysis. However, there are common themes among the methods which are of particular interest for our study, including mechanisms to define the basic concepts (boundary, scope, and vocabulary) of the domain, describe the data that support the functions and state of the system or family of systems, identify relationships and constraints among the concepts and data within the domain. We have taken the following steps as a general process of domain analysis:

1) Definition of the domain boundary
2) Collection of domain knowledge and information
3) Analysis of the domain
4) Modeling of the domain

The results of the analysis and modeling process, collectively referred to as a domain model (in a broad sense), consist of the following:

- Domain description: information and knowledge specific to the domain, including narrative documents, figures and tables, definition of domain terms, and so on.
- Data element dictionary: a collection of data elements (metadata) of health statistics represented in a standard form.
- Domain model (in a narrow sense): a formal representation of the domain.

Domain Boundary Definition

The first step of domain analysis is the definition of the boundary (or the scope) of a domain. There is no established method for defining the domain boundary. The wider a domain is, potential usability might increase, on one hand. However, the complexity of the analysis may increase, validation of a model may become more difficult, and objectives of modeling and the use of a model might become ambiguous. The scope of our domain analysis is health statistics in general, but to define the boundary more clearly, we selected the following set of information and knowledge as input to our domain analysis:

1) National health and welfare statistical surveys
2) National healthcare delivery service system structure
3) National social security system

The scope of the domain was thus defined by the set. Sources include books, documents (reports, papers, etc.), and expert knowledge. The collected and summarized knowledge and information about the domain are called "domain description" which forms a part of the domain model in a broad sense.

Health Statistical Surveys as Input to Domain Analysis

Domestically, a number of health statistical surveys are carried out including surveys on health institutions, health professionals, patients, disease, medical expenditure, health of people, pharmaceuticals, and so on. Among the domain knowledge and information, health statistical surveys are fundamental to the domain analysis since health statistics is the most prominent health information which can be shared, and significant data items are collected about people, health, and society. Hence health statistical surveys are used as input and are analyzed in full detail (by "input", we mean not statistical data themselves but information about the surveys, i.e., the purpose, the subjects of a survey, and the items surveyed). Table 1 shows some of the major surveys we used. As an example, a survey on "people and life" is carried out to obtain fundamental statistics about people, health, families and households, and so on. Items include type of households, expenditure and income, health insurance, employment status, hospitalization (or going to hospitals), tax, and so on.

Table 1 Sample Health Statistical Surveys

Survey	Data items
Health institutions	name, owner, number of beds, departments, inpatients and outpatients, diagnostic equipment, etc.
Doctors, dentists, and pharmacists	name, address, sex, age, type of licensure, professions, departments, place of clinical practice, etc.
Patients	sex, date of birth, disease, department, method for payment, inpatient/outpatient, treatment, length of stay, etc.
Vital statistics	population movement, including birth, death, fetal death, marriages and divorces, etc.
Vital statistics & Social economics	Social economic factors affecting vital statistics (birth, death, fetal death, marriages and divorces)
People and life	household by type, income, tax, expenditure, health insurance, pension, employment, health condition, etc.

Data Element Dictionary for Health Statistics

Health Statistics Metadata

When a survey is carried out, the results are usually summarized in the form of summary tables. In general, a summary table is obtained by classifying individuals and describing some numerical properties of groups of individuals. In a national survey on clinical institutions, for example, the individuals are hospitals and clinics. They are classified according to the location, bed-size, providers running hospitals, etc., and some numerical properties describing the groups of hospitals such as the numbers of health professionals, the number of in-patients, the length of stay, etc., are summarized in tabular form. The attributes used to classify individuals such as providers running

hospitals, bed-size, types of hospitals, etc. are called category attributes. The attributes which describe numerical properties of groups of individuals, such as the number of in-patients, average length of stay, etc., are called summary attributes. In our previous study on summary tables, a formal representation and meta data of summary tables were established [1,2]. Data that describe "what information a given summary table represents" are called metadata, and we defined category attributes and summary attributes with their representational scheme as meta data of summary tables.

Health Statistics Data Elements

If the values that a given data item may take are defined by individual applications, organizations, and/or regions, then the results obtained from different surveys are hardly comparable. By defining standards for data items before hand, and by adopting the standards when applicable, comparability of statistics could be highly enhanced. To define standards for basic data items of health statistical surveys, information about surveys were analyzed, and data elements were extracted. The extracted data elements were put into a data element dictionary. The data element dictionary is a collection of metadata of health statistics. Each entry of the dictionary is a set of attributes about data definitions represented according to international standards. The data element dictionary is an essential component of the domain model (in a broad sense).

If a data element is a classificatory attribute, then a standard classification system(s) is defined. If a data element is a numeric attribute, then a unit, data types (integers, floating), number of digits, and so on, are defined. The metadata standards have been based on ISO/IEC 11179 "Specification and Standardization of Data Elements" [4]. According to the standards, data elements are considered to correspond to columns of a table of a relational database, attributes of a relation in relational model, attributes of classes in object-oriented models. Data elements are specified by their attributes. The basic attributes are classified into: identifying; definitional; relational; representational; and administrative. Some of the representational attributes are :

- Data type of data element values: e.g., character, ordinal number, integer, real, etc.
- Layout of representation: integers may be indicated with 'n' for decimal mark; the number of characters before and after the decimal mark are specified as n(5).n(33). For code representations, if the data element has a code representation structure consisting of two alphabetic characters followed by a four-digit number for example, then "layout of representation" is AA9999.
- Permissible data element values: The set of representations of permissible instances of the data element. The set can be specified by name, by reference to a source, by enumeration of the representation of the instances, or by rules for generating the instances.

When the permissible data element values are an enumeration of coded representations, each data element value and instance should be presented as a pair. For example, permissible values of "health professionals" are:

1 doctors (full time) 2 doctors (part time) 3 dentists
4 pharmacists 5 public health nurses 6 midwives ...

A data element named "birth weight (in grams)" may contain values which are positive integers within a range 0 to 9999. For a data element named "name of a hospital" has a domain that is defined by the national register of hospitals.

Not only data elements which take actual values, but also elements which represent concepts are defined as data elements (will be called conceptual elements) and are put into the data element dictionary. For example, an attribute "hospitalized or going to hospitals" is a data element which takes a value "yes/no." On the other hand, "health conditions" is a conceptual element which describes a person's health status. It consists of various attributes describing a person's health. A conceptual element may have one or more sub-conceptual elements, and a sub-conceptual element in turn may have its own sub-conceptual elements, and so on, and thus a hierarchy of conceptual elements may be formed. Data elements may appear only as leaves.

Element with Multiple Classification Systems

Whenever permissible values for a given data element are defined as dejure standards such as ISO and JIS (Japan Industrial standards), the standards are adopted. For example, we have JIS X0401 codes for "prefectures", X0407 for "education", X0409 for "relation to the house holder", and so on. Otherwise, permissible values of data elements are defined according to national surveys carried out by the government or government related agencies. If multiple definitions exist, they are all placed in a data element dictionary.

For a data element which takes categorical (classificatory) values, a classification system is not necessarily unique. For example, three classification systems, "the detailed, the intermediate level, and the broad" are used for providers running hospitals in the publication of the Japanese Ministry of Health and Welfare. Table 2 shows the three simplified classification systems. As another example, a number of classification systems are used for the types of households in a national survey on "people and life." When there are multiple classification systems conventionally used, the data element is defined to have multiple domains

Table 2 Classification systems of providers running hospitals

Classification System 1	Classification System 2	Classification System 3
Ministry of Health and Welfare	Ministry of Health and Welfare	Public
Ministry of Education	Other national	
Labor welfare corporation		
Other national		
Prefectures	Prefectures	
Cities, towns, villages	Cities, Towns, Villages	
Medical Foundations	Medical foundations	Private
School foundations	Other foundations	
Other foundations		
Companies	Other private	
Private persons		

of permissible values, and the relationship(s) between classification systems, if there exist any, are also represented. For example, three classification systems of "providers running hospitals", there is a hierarchical relationship.

Implementation of the Data Element Dictionary

We consider that the data element dictionary should be made open and be updated in an open manner. For maintainability of the dictionary and for representational flexibility of the structure of data elements, XML is used to represent the data element dictionary. The data elements can be viewed through a standard browser where XSL is used to show the contents. As an example, the data element "health professionals in hospitals" viewed through a browser is shown in Figure 2 (the data elements are defined in Japanese only at the moment, and the element in Figure 2 was translated into English for this paper). A set of software utilities was developed for a web-based data element dictionary including automatic XML instance creation, for automatic unique code assignments to permissible values of a categorical data element, and so on.

A Domain Model in a Narrow Sense

A domain model in a narrow sense is a formal representation of the domain. For representation of the details, we use Entity-Relationship (ER) models, and for the overall picture of the model, UML is adopted.

Each conceptual element is represented as an entity in ER notation. Figure 3 shows a conceptual element "individual." It is described by a set of attributes listed in the box. "sex" and "date of birth" are the data elements which take actual values. The others (underlined) are conceptual elements themselves. The attributes of "individual" are summarized in Table 3 where "income inf", "insurance inf", and"education inf" are omitted.

Table 3 Attributes of "individual"

Sex	1 male 2 female 9 unknown
date of birth	CCYYMMDD
Address	a sub-conceptual element of a conceptual element "location"
household inf	a conceptual element describing the relation between "individual" and "household"
health conditions	a conceptual element consisting of sub-elements
occupation inf	a conceptual element describing the relation between "individual" and "occupation"

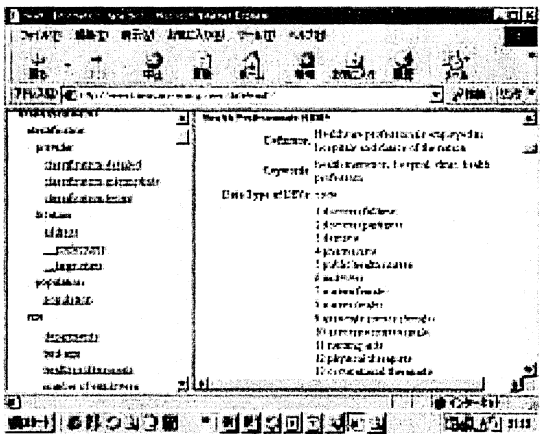

Figure 2 Data element representing health professions

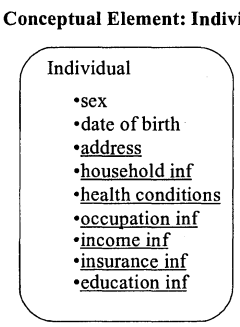

Figure 3 Conceptual element "individual."

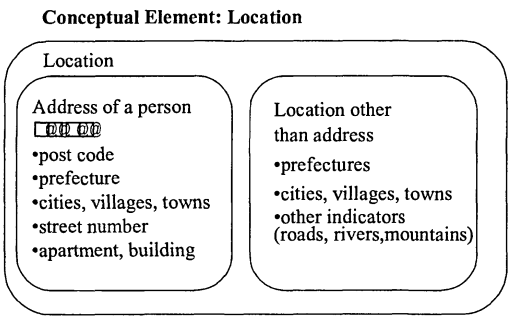

Figure 4 Conceptual element "location"

The attribute "address" of "individual" is a sub-conceptual element of "location" which is a more general concept than "address." In ER notation, this is modeled as shown in Figure 4. "address" is a conceptual element which is described by the attributes "post code, prefecture, cities, ...," all of which are data elements. Considering the hierarchy of conceptual elements, the data elements appear only at the leaf level, and a leaf node may have multiple parents. When multiple conceptual elements have the same attribute which is a data element, there should be only one entry of that data element in the data element dictionary.

The over all model is represented in UML. At the moment, the entire model consists of eight parts namely: healthcare delivery; illness/injuries; social security; people and life; labor and industry; vital statistics events; education; and geography. As described at the beginning of this paper, all information is modeled from a view point of second level of information sharing, and hence "illness/injuries," for example, does not represent the concepts or knowledge from clinical point of view but from statistical point of view. A simplified part of "people and life" is shown in Figure 5.

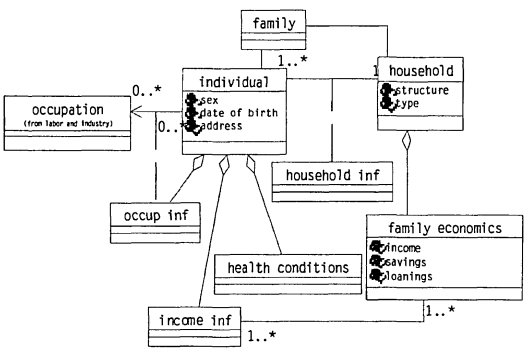

Figure 5 A model view of people and life

It shows that "individual" and "household" are conceptual elements, and an attribute "household inf" of "individual" is a conceptual element describing the relation between "individual" and "household."

Discussion and Conclusion

Modeling has been applied widely and extensively in the filed of heath informatics including an application to national health information modeling in Australia [5]. In our study, domain analysis and domain modeling are applied to health statistics. In the study, the scope of the domain was defined by the information and knowledge about health statistics, in particular the health statistics surveys. A domain model in a broad sense including a data element dictionary and a domain model in a narrow sense were developed. A domain model in a narrow sense was represented in ER and UML notations. Domain modeling is an iterative process which goes around domain analysis, identification of data elements, and modeling. A domain model may refine the data element dictionary, and refined data elements may improve a domain model. Development of the data element dictionary is highly facilitated by the domain model, and the domain modeling is guided by the data elements.

The domain of heath statistics is related to statistics of other areas such as labor, economics and industry, and some conceptual elements and data elements are common to multiple areas. For example, "health conditions", "income", and "households" appear in labor statistics as well. Without standard representation of concepts and standards for data elements, it is not possible to share statistics among different organizations, surveys, regions, and disciplines.

The standard data element dictionary would make health data more sharable and improve comparability of health statistics. The study will not only contribute to quality health statistics domestically, but also will help to improve comparability of health statistics across the countries.

References

[1] Okada, M., Takaba, M., et al.: Formal Representation of Summary Tables for Health Care Statistical Database Management, Comput. Biomed. Res., 31: 426-450, 1998.

[2] Okada, M., and Takaba, M.: A Formal Study on Summary Table Handling with Application to Health Care Statistical Databases, Proc. 1996 AMIA Annual Fall Symposium, 448-452, 1996.

[3] Prieto-Diaz, R.: Domain Analysis for Resusability, COMPSAC'87, 23-29, 1987.

[4] ISO/IEC 11179: Information Technology- Specification and Standardization of Data Elements, part1-part6.

[5] http://www.aihw.gov.au/knowledgebase/index.html

Address for correspondence

Mihoko Okada

m-okada@mw.kawasaki-m.ac.jp