Methods of Information in Medicine

N. F. de Keizer[1],
A. Abu-Hanna[1],
J. H. M. Zwetsloot-Schonk[2]

Understanding Terminological Systems I: Terminology and Typology

[1]Department of Medical Informatics, Academic Medical Center, Amsterdam, [2]Julius Center for Patient Oriented Research, Utrecht University Medical School, Utrecht, The Netherlands

Abstract: Terminological systems are an important research issue within the field of medical informatics. For precise understanding of existing terminological systems a referential framework is needed that provides a uniform terminology and typology of terminological systems themselves. In this article a uniform terminology is described by putting relevant fundamental notions and definitions used by standard organizations such as CEN and ISO into perspective, and interrelating them to arrive at a useful typology of terminological systems. This typology is illustrated by applying it to five well-known existing terminological systems.

Keywords: Terminology, Typology, Terminological Systems, Standards

1. Introduction

The use of medical data stored in computer-based patient records (CPRs) has increased the need for structured and controlled data entry and data representation [1]. The usual way to cope with this need is using standard terms from a terminological system to record medical data. Consequently, terminological systems are an important line of research in medical informatics. In this article the term "terminological system" is used as an umbrella term for the notions "classification", "thesaurus", "vocabulary", "nomenclature" and "ontology", which are further described in Section 3.

Before an existing terminological system can be used or a new system developed, a good understanding of the terminological system is essential. In the context of a project aimed at developing a terminological system for intensive care diagnoses [2], we found that the literature on terminological systems, including the most well-known systems, is hard to understand and is ambiguous. This is due to heterogeneity and indistinctness in the terminology used to describe the terminological systems, and by incomplete description of the structure and characteristics of the various systems.

Attempts to clarify terminological systems have been made by organizations such as the International Standards Organization (ISO) [3, 4] and the Comité Européen de Normalisation (CEN) [5]. Although the standards developed by these organizations are freely available, the support they provide for the understanding of terminological systems in order to assess their merits is minimal because they only enumerate dry definitions of notions. We feel an overall accessible framework for the understanding of terminological systems, in which definitions are placed in perspective and put in practice, is still lacking.

This paper comprises the first part of two articles and is a summary of an extensive technical report [6], which can be obtained from the authors. The goal of these two articles is to provide a referential framework, by which terminological systems can be characterized, understood and compared, and to describe our experience with the use of this framework. This description is restricted to terminological systems for medical diagnoses but the treatment is equally relevant for all terminological systems. The aim of this framework is to assist those interested in understanding and applying terminological systems.

This paper describes the first part of the referential framework in which definitions have been arranged and put into perspective, resulting in a typology of terminological systems. To illustrate this first part of the referential framework, we apply it to five existing medical terminological systems. The second part of the framework provides a basis for a formal and thereby unambiguous description of the *structure* of a terminological system. In the second paper [7] we also describe our positive experience with terminological system formalization in terms of (1) understanding existing terminological systems; (2) recognizing patterns in the structure of the systems and (3) developing a new terminological system [2].

In Section 2 of this article the building blocks for conceptualization of terminological systems (object, concept, term, code) and the different types of relations between concepts, are described. The typology of terminological systems is described in Section 3. In Section 4 we applied the typology to five existing terminological systems with this framework.

2. Terminology and Definitions

Before discussing terminological systems, some basic elements have to

be explained: objects, concepts and designations, which form the so-called semiotic triangle or meaning triangle [5, 8]. Reality can be conceived as consisting of *objects* (things), such as "heart valve" or more abstract things such as "pain". We use characteristics of objects to form cognitive constructs, called *concepts*, which are units of thoughts. Although objects and concepts are different notions, in the remainder of this article we only address concepts, because terminological systems mainly comprise general concepts, e.g. diseases, that are used to describe instances of patient's diseases (i.e. objects) recorded in patient's records.

Linguistic labels, called *terms*, are used to designate a concept. Concepts can be designated in many ways because there are many languages from which terms can be taken and one language can use many different terms for the same concept. *Definitions* are statements about the meaning of a concept. Codes (letters, numerals or a combination thereof) can be used to designate concepts in a computerized system. In general we can distinguish two types of codes: *significant codes* and *non-significant codes*. Non-significant codes are context free, meaning that the code's value is not related to the meaning of the concept, e.g. *sequential codes* such as aaa, aab, etc. and *random codes*. The advantage of using non-significant codes is that concepts can be altered in the structure of the terminological system without consequence for the codes, and that new concepts can easily be added to a terminological system, contrary to some cases in which significant codes are used. Significant codes are related to the characteristics of the concept and its place in the terminological system, for example *mnemonic codes* (codes related to a name describing the concept to aid the user to memorize the meaning of the code e.g. A900 = **A**natomical component "lung"); *hierarchical codes* e.g. viral pneumonia is coded by 480 and pneumonia caused by the adeno virus is coded by 480.1); and *juxtaposition codes* (composite codes, e.g. a code for pneumonia A900.P100 in which A900 = anatomical component "lung" and P100 = process "inflammation").

2.1 Concepts and their Characteristics

Concepts can be described by their characteristics. For example color ("red"), size ("5 cm^2") and shape ("ellipse") could describe the concept "ulcer". Some characteristics are considered essential, i.e. that the characteristics are part of the definition of the concept, e.g. "duodenum ulcer" is always located in the duodenum so the location "duodenum" is an essential characteristic of the concept "duodenum ulcer". These essential characteristics are called necessary conditions in the definition.

2.2 Relationships between Concepts

According to ISO and CEN standards [3-5], concepts can be related to each other by hierarchical and non-hierarchical relationships. In a hierarchical relationship an order is expressed between (at least) two concepts: a superordinate concept (e.g. heart valve) and a subordinate concept (e.g. mitral valve). Hierarchical relationships can be *generic* or *partitive*. In the past, generic relationships were called *logical* relationships in ISO standards. A generic relationship, the "Is_a" relation, is a relationship between a genus (superordinate concept in a generic relationship) and a species (subordinated concept in a generic relationship) where the *intension* of the genus is contained and extended in the *intension* of the species. The intension of a concept is the set of uniquely describing characteristics including relationships of that concept, e.g. the set {health problem with anatomical localization "liver", dysfunction "infection" and etiology "virus"} constitute an intensional definition of the concept "viral hepatitis". Hepatitis B (species) is a "viral hepatitis" (genus) and therefore it has the same intensional definition but extended with: etiology is "hepatitis B virus". The other type of hierarchical relationships are *partitive* relationships in which the superordinate concept denotes an object which represents the whole, and the subordinate concepts refer to its parts, e.g. a heart valve *is part of* the heart.

In the past, the ISO standard described *ontological* relationships, which nowadays are split up in partitive and non-hierarchical relations. Non-hierarchical relations describe a wide range of relationships between concepts such as spatial, temporal, causal or any arbitrary one, e.g., the relationship between an operative procedure and the organ to be operated on. In theory there is an unlimited number of relationships between concepts. However, in practice only a small portion of relationships, those determined by human's knowledge and experience, is widely useful. Therefore, ontological relations are called philological relations (distinguishing concepts and coherence between concepts as being determined by cultural history, i.e. subject to the constraints of language, art and science) by Hirs [9]. Figure 1 shows the above-described types of relationships between concepts.

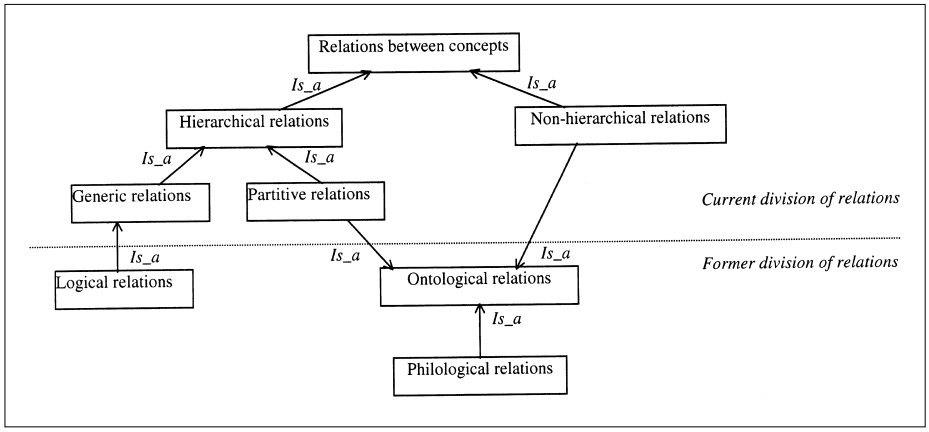

Fig. 1 Types of concept relations. Above the dotted line the current ISO and CEN standards are represented. Below the dotted line former ISO/CEN standards and Hirs' relation types are represented.

2.3 Definitions of Concepts

Relationships (and characteristics) can be used to order and define concepts in a system. An intensional definition is a definition based on the intension of the concept (see Section 2.2). A *definition per genus et differentium* is the intensional definition of the genus (if it has one) extended with extra characteristics of the species. Instead of describing a concept intensionally, an *extensional definition* can be used, i.e. the set of all specific concepts (species) of a superordinate concept (genus) is enumerated. For example "Granulocytes" can be intensionally defined by "Leucocytes with abundant granules in the cytoplasm" or extensionally by the enumeration of "Neutrophil", "Eosinophil" and "Basophil", all the species of the genus "Granulocytes".

3. Typology of Terminological Systems

A terminological system relates concepts, of a particular domain, among themselves and provides their terms and possibly their definitions and codes. In literature, terms such as "terminology", "thesaurus", "vocabulary", "nomenclature" and "classification" are often confused. This section describes a typology of terminological systems based on literature and existing standards, including definitions and relationships between different types of terminological systems (see also Fig. 2).

A *terminology* is a list of terms referring to concepts in a particular domain. A thesaurus is a terminology, in which terms are ordered, e.g., alphabetically or systematically and in which concepts can possibly be described by more than one (synonymous) term. When a concept in a terminology or thesaurus is accompanied by a definition, it is called a *vocabulary* or *glossary*. A *nomenclature* is a system of terms composed according to pre-established composition rules or the set of rules itself for composing new complex concepts. *Classification* is an arrangement of objects or concepts (by the *is_member_ of* relation) based on their essential characteristics into groups of concepts, called classes. A *taxonomy* is an arrangement of classes according to the *Is_a* relationship from the subordinate class to the superordinate class. The CEN standards [5], however, do not distinguish between "classification" and "taxonomy". Therefore, we use the term "classification" in this article loosely to also include taxonomy. A *Nosology* is a classification of diseases. A terminology, thesaurus, vocabulary, nomenclature or classification is called a *coding system* when the system uses codes for designating concepts. Figure 3 is a model of different types of terminological systems and their relations. Each terminological system is at least a terminology with possibly additional characteristics, e.g., it is also a classification when the terminology consists of generic relations between concepts.

Sometimes the notion of *ontology* is used as a synonym for different types of terminological systems. An ontology is a (formal) specification of concepts, relations and functions in a domain [10] and hence focus on concepts. Concepts are important in terminological systems but they also focus on the *terminology* itself. An ontology is usually used to model consensus in understanding a domain between different partners.

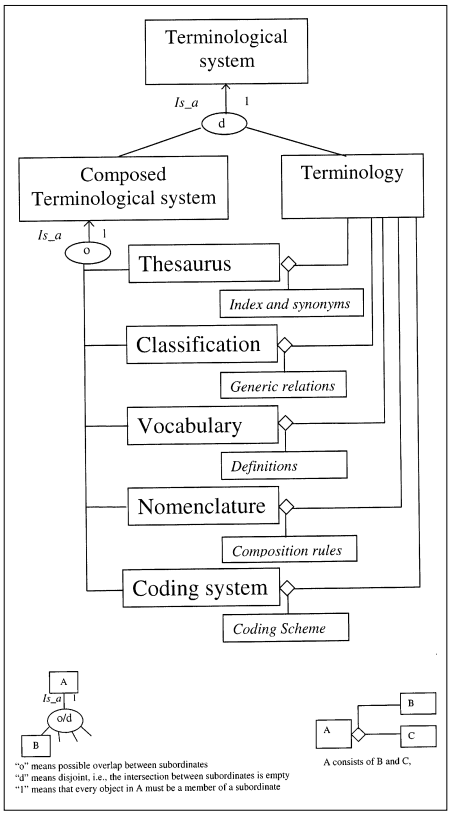

Fig. 2 Model of relations between terminology, thesaurus, vocabulary, nomenclature, classification and coding system. Each terminological system is either a terminology or includes a terminology as part of a thesaurus, classification, vocabulary, nomenclature or coding system.

4. Typology of Existing Terminological Systems

To further illustrate the theory of Sections 2 and 3, we describe the types of the terminological systems and the coding schemes used in the terminological systems ICD-9-CM/ICD-10, SNOMED, NHS Clinical terms, UMLS and GALEN. We have chosen these five terminological systems for their intended use to describe (among others) diagnoses in daily healthcare practice. Hence, we did not consider, for example, MeSH because this system is only intended to link concepts to medical literature. Table 1 summarizes the types of each terminological system and coding schema. Further explanation about the types of each terminological system can be found in Sections 4.1 to 4.6.

4.1 ICD-9-CM and ICD-10

The International Classification of Diseases (ICD), maintained by the World Health Organization, is perhaps the best known terminological system used in medicine. The ICD is a *classification* (recall that we do not distinguish here between classification and taxonomy) of generic-related diagnostic terms represented in 17 (ICD-9-CM) [11] or 21 (ICD-10) [12] "chapters", mainly arranged according to anatomical system or etiology. The ICD-10 also contains an alphabetical index, which makes it also a *thesaurus*. The ICD-9-CM further contains two extra classifications, the V-list for factors influencing health status and contacts with health services, and the E-list for external causes of injury and poisoning. It also contains a number of appendices, one of which is a very limited *nomenclature* that further specifies the morphology of

neoplasms of chapter 2 (this explains one "•" in Table 1). In the ICD-10 the V- and E-list have been added to the chapters and the nervous system chapter of the ICD-9-CM is split into three different chapters.

The codes of the ICD-9-CM and ICD-10 are hierarchical and hence significant. In the ICD-10 they are also mnemonic, e.g., all terms in the chapter "Diseases of the digestive system" start with a "K". In the ICD-10 some terms have two alternative codes: one marked with a dagger (†) belonging to the etiology, and the other marked with an asterisk (*) belonging to the localization or manifestations. For example "meningococcal meningitis" has two codes: A39.0† in the chapter "Certain infectious and parasitic diseases" (etiology) and G01* in the chapter "Diseases of the nervous system" (localization). In general, dagger codes are used for mortality registration, asterisk codes are used for morbidity registration.

4.2 NHS clinical terms

The NHS clinical terms, formerly called Read Clinical Classification [13, 14], were originally developed for automated description of clinical and administrative data in general practice, but evolved to a version which facilitates daily care practice in the entire field of health care. It forms a *classification* of generic-related medical concepts designated by a preferred term and some synonyms (if applicable) which are ordered hierarchically, qualifying it as a thesaurus. Since version 3.1 it is a *nomenclature* in which rules are given to modify some terms with qualifiers in a controlled way, e.g., "course of illness: (acute/chronic)" to qualify heart failure. Some of the attributes have the status "atom" which means that these are intrinsic characteristics of a concept, giving the NHS clinical terms a *vocabulary* character. Until version 2, the codes were significant, representing the relative place of the concept in the hierarchy. Concepts could be assigned by more than one code when they are placed in more than one part of the classification. From version 3 onwards each concept has a unique non-significant code consisting of characters and numbers.

Table 1 Types of well-known terminological systems.

Type	ICD-10	Read	SNOMED	UMLS	GALEN
Terminology	••	••	••	••	••
Thesaurus	••	••	•	••	•
Vocabulary	-	•	•	•	••
Nomenclature	•	•	•	-	••
Classification	••	••	••	•	••
Ontology	-	-	-	•	••
Coding schema	Significant	Non-significant	Significant	Non-significant	Non-significant

•• : acceptable for classification; • : partially acceptable for classification

4.3 SNOMED

In 1975 the College of American Pathologists published the Systematized Nomenclature of Medicine (SNOMED) to provide terms for a broad range of clinical domains. A couple of revisions resulted in SNOMED International [15, 16] which is intended to be incorporated into a computer-based patient record.

Its structure consists of eleven modules, also called axes or dimensions (e.g. topography, disease and diagnosis, procedures, etc.) which can be conceived as distinct classifications. Almost all ICD-9-CM terms and codes can be found in the "Disease and Diagnosis" module. By linking terms of the various modules one can compose new complex medical concepts. The discrete terms, together with the information about the cross-reference relations between the different axes, provide (at least a partial) definition of each concept, giving SNOMED its *vocabulary* character. Although SNOMED is an acronym including the term nomenclature, it is questionable whether SNOMED is a *nomenclature* that can generate sensible compositions. It is a nomenclature in the sense that new concepts can be composed, but *rules* which define which concepts may be *sensibly* linked to compose new concepts are missing. Due to the lack of suitable composition rules it is possible to represent the same clinical concept by different composites and even to arrange clinically irrelevant composites, e.g., a fractured left lung encoded by T-28500 (Left lung); M12200 (Fracture, open). SNOMED-RT [17], which is under development, will address these problems.

Codes in SNOMED are significant. Within a module they are hierarchical and mnemonic, e.g., topographical codes start with a "T". Newly composed concepts get juxtaposed codes by joining the codes for the associated terms from the different modules.

4.4 UMLS

In 1987 the National Library of Medicine developed the Unified Medical Language System (UMLS) [8, 18]. The goal of the UMLS is to facilitate the retrieval and integration of information from multiple machine-readable biomedical information sources such as patient record systems and bibliographic databases.

The UMLS consists of four knowledge sources, among which the Metathesaurus and the Semantic network are the most relevant in the context of this paper. As the name indicates, the Metathesaurus is a *thesaurus* in which concepts are linked to (synonymous and preferred) terms which are alphabetically ordered. The Metathesaurus is manually enriched with hierarchical relations between concepts from established terminological systems such as ICD-9-CM, Read Codes, SNOMED and MeSH which give it characteristics of a *classification*. As each concept has an attribute "definition", the metathesaurus is also a *vocabulary*. The Semantic Network provides information about concepts on a high aggregation level (semantic categories) and their relations. All concepts represented in the Metathesaurus are categorized into semantic categories of the Semantic Network which regulates (together with the Metathesaurus) the mapping of

concepts from different sources to each other. Although in theory semantic relations defined in a semantic network could aid a user to make new composites, in practice the UMLS does not support the composition of new concepts.

Each concept and term in the Metathesaurus has a unique UMLS code. Although these codes are somewhat mnemonic (concept codes start with a "C" and term codes with an "L") they are not significant in the sense that the numbers next to the starting character are non-significant.

4.5 GALEN

The goal of the Generalized Architecture for Languages Encyclopedias and Nomenclatures in Medicine (GALEN) project [19-21] is to formally describe and model the medical domain by which the interchangeability of electronic medical data of different data sources can be supported. The "Terminology Server", the implementation of GALEN's goal, integrates three modules: the Concept Module, the Multilingual Module and the Code Conversion module.

The Concept Module implements a formal language, the GALEN Representation and Integration Language (GRAIL) which is used to represent concepts and their characteristics and relations in the Concept Reference (CORE) model. The CORE model is an *ontology*, which currently comprises the definitions of approximately 5,000 concepts (e.g. bones, organs, and fractures) and 1,000 relations (e.g. HaveLocation, HaveComplexity). The GRAIL formalism allows developers of terminologies to create models containing these concepts and relationships, and to derive new concepts which are valid compositions of existing ones. The composition rules included in GRAIL make the system a *nomenclature*, the definition rules make it a *vocabulary*.

The mapping of concepts to synonym and preferred terms *(thesaurus)* in multiple languages is managed by the Multilingual Module. The mapping of concepts to and from existing coding systems is managed by the Code Conversion Module. Each concept in the CORE model has a unique non-significant code. Different applications can communicate with the terminology server. By this direct communication it is possible to generate random unique numbers for new composites which are saved at the server.

5. Summary and Discussion

Historically, medical data were coded mainly for (retrospective) statistical, epidemiological and administrative purposes. Nowadays with the electronic availability of medical data, the importance of these data in daily care practice, and especially the importance of these data in interdisciplinary communication and for clinical research, has increased. This shift in use of medical data implies new requirements on terminological systems concerning, for example, the level of detail and the structure of the terminological systems.

Although a major goal of terminological systems is the standardization of terminology to improve communication, the notions used in the literature to describe terminological systems themselves are not uniform, which makes it hard to communicate their underlying ideas.

A good understanding of terminological systems is essential before one can assess whether an existing terminological system is appropriate for use in certain circumstances, or when one has to develop a new system. Therefore, a referential framework for understanding terminological systems is needed. Such a framework includes at least two components. First a terminology and typology of terminological systems and second a uniform (formal) representation of the structure of the terminological system. Existing standards such as ISO and CEN [3, 5] only describe the first part. Moreover, this is restricted to a rather dry enumeration of definitions about notions in the field. Therefore this article describes the first part of a framework for understanding terminological systems and summarizes the notions and definitions used by standard organization such as ISO and CEN, but enriched with interrelations between these notions, including a typology of terminological systems. This typology is illustrated by applying it to five well-known medical terminological systems including diagnoses.

This first paper, plays a facilitating role for the second paper [7] which includes our positive experience with the application of a conceptual and formal representation formalism to describe the structure of terminological systems. The aim of these papers is to help researchers to interpret the merits and limitations of existing terminological systems and to build on existing work in the field.

Acknowledgements

We thank Ronald Cornet for his valuable discussions and helpful remarks on an earlier draft of this manuscript.

REFERENCES

1. Moorman P, van Ginneken A, van der Lei J, van Bemmel JH. A model for structured data entry based on explicit descriptional knowledge. Yearbook of Medical Informatics 1995: 195-204.
2. De Keizer N, Abu-Hanna A, Cornet R, Zwetsloot-Schonk J, Stoutenbeek C. Analysis and Design of an Ontology for intensive care diagnoses. Method Inform Med 1999; 38: 102-12.
3. Coding Methods and principles. ISO 1989.
4. Terminology vocabulary. ISO 1990.
5. European prestandard. Medical informatics – Categorical structures of systems of concepts – Model for representation of semantics. Brussel: CEN 1997.
6. De Keizer N, Abu-Hanna A. A framework for understanding Terminological Systems. Amsterdam: Academic Medical Center, Department of Medical Informatics 1999.
7. De Keizer N, Abu-Hanna A. Understanding terminological systems II: Experience with conceptual and formal representation of structure. Method Inform Med 2000; 39: 22-9.
8. Campbell K, Oliver D, Spackman K, Shortliffe K. Representing Thoughts, Words, and Things in the UMLS. J Am Med Inform Assoc 1998; 5: 421-31.
9. Hirs W. Standaardclassificaties voor medische en niet-medische gegevens. Groningen: Rijksuniversiteit Groningen 1987. 141 p.
10. Gruber T. Towards principles for the design of ontologies used for knowledge sharing. Int J Hum-Comput Stud 1995; 43: 907-28.
11. International Classification of Diseases, manual of the International Statistical Classification of diseases, injuries and causes of death: 9th revision. WHO 1977.
12. International Classification of Diseases, manual of the International Statistical Classification of diseases, injuries and causes of death: 10th revision. WHO 1993.
13. Read J, Sanderson H, Drennan Y. Terming, coding and grouping. In: Greenes R, ed. Medinfo 95, 1995: 56-9.

14. Schulz E, Price C, Brown P. Symbolic Anatomic Knowledge Representation in the Read Codes Version 3: Structure and Application. J Am Med Inform Assoc 1997; 4: 38-48.
15. Rothwell D. SNOMED-Based knowledge representation. Method Inform Med 1995; 34: 209-13.
16. Rothwell D, Coté R. Managing Information with SNOMED: Understanding the model. SCAMC, 1996: 80-3.
17. Spackman K, Campbell K. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. J Am Med Inform Assoc Annual Symposium, 1998: 740-4.
18. Lindberg D, Humphreys B, Mc Cray A. The Unified Medical Language System. Method Inform Med 1993; 34: 281-91.
19. Rector A, Solomon W, Nowlan W, Rush T, Zanstra P, Claassen W. A Terminology Server for medical language and medical information systems. Method Inform Med 1995; 34: 147-57.
20. Rector A, Glowinski A, Nowlan W, Rossi-Mori A. Medical-concept models and medical records: an approach based on GALEN and PEN&PAD. J Am Med Inform Assoc 1995; 2: 19-35.
21. Rector A, Bechhofer S, Goble C, Horrocks I, Nowlan W, Solomon W. The Grail concept modelling language for medical terminology. Artif Intell 1997; 9: 139-71.

Address of the authors:
Nicolette F. de Keizer,
Department of Medical Informatics, J2-256,
Academic Medical Center,
P.O. Box 22660
1100 DD Amsterdam,
The Netherlands
E-mail: n.f.keizer@amc.uva.nl