International Journal of Medical Informatics 84 (2015) 1099–1110

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

MELLO: Medical lifelog ontology for data terms from self-tracking and lifelog devices

Hye Hyeon Kim[a], Soo Youn Lee[a], Su Youn Baik[a], Ju Han Kim[a,b,*]

[a] Seoul National University Biomedical Informatics (SNUBI), Division of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110799, South Korea
[b] Systems Biomedical Informatics National Core Research Center (SBI-NCRC), Seoul National University College of Medicine, Seoul 110799, South Korea

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form 6 August 2015
Accepted 11 August 2015

Keywords:
Ontology
Consumer health
Lifelog

ABSTRACT

Objective: The increasing use of health self-tracking devices is making the integration of heterogeneous data and shared decision-making more challenging. Computational analysis of lifelog data has been hampered by the lack of semantic and syntactic consistency among lifelog terms and related ontologies. Medical lifelog ontology (MELLO) was developed by identifying lifelog concepts and relationships between concepts, and it provides clear definitions by following ontology development methods. MELLO aims to support the classification and semantic mapping of lifelog data from diverse health self-tracking devices.

Methods: MELLO was developed using the General Formal Ontology method with a manual iterative process comprising five steps: (1) defining the scope of lifelog data, (2) identifying lifelog concepts, (3) assigning relationships among MELLO concepts, (4) developing MELLO properties (e.g., synonyms, preferred terms, and definitions) for each MELLO concept, and (5) evaluating representative layers of the ontology content. An evaluation was performed by classifying 11 devices into 3 classes by subjects, and performing pairwise comparisons of lifelog terms among 5 devices in each class as measured using the Jaccard similarity index.

Results: MELLO represents a comprehensive knowledge base of 1998 lifelog concepts, with 4996 synonyms for 1211 (61%) concepts and 1395 definitions for 926 (46%) concepts. The MELLO Browser and MELLO Mapper provide convenient access and annotating non-standard proprietary terms with MELLO (http://mello.snubi.org/). MELLO covers 88.1% of lifelog terms from 11 health self-tracking devices and uses simple string matching to match semantically similar terms provided by various devices that are not yet integrated. The results from the comparisons of Jaccard similarities between simple string matching and MELLO matching revealed increases of 2.5, 2.2, and 5.7 folds for physical activity,body measure, and sleep classes, respectively.

Conclusions: MELLO is the first ontology for representing health-related lifelog data with rich contents including definitions, synonyms, and semantic relationships. MELLO fills the semantic gap between heterogeneous lifelog terms that are generated by diverse health self-tracking devices. The unified representation of lifelog terms facilitated by MELLO can help describe an individual's lifestyle and environmental factors, which can be included with user-generated data for clinical research and thereby enhance data integration and sharing.

1. Introduction

Numerous types of mobile and wearable devices such as mobile phones and armbands are now available for recording what you eat, what you weigh, how many steps you take, and how well you sleep. Other real-time self-tracking devices such as Nike's FuelBand, Runkeeper, Fitbit, and the UP by Jawbone that have also been developed can be used in many ways such as for health self-tracking, lifelogging, self-quantification, and behavior monitoring [1,2]. All

* Corresponding author at: Division of Biomedical Informatics and Systems, Biomedical Informatics Research Center, Seoul National University College of Medicine, 28 Yongon-dong, Chongno-gu, Seoul 110799, South Korea. Fax: +82 2 747 8928.
E-mail address: juhan@snu.ac.kr (J.H. Kim).

http://dx.doi.org/10.1016/j.ijmedinf.2015.08.005
1386-5056/

of these devices make it possible for an individual to monitor personal parameters in her/his daily life that are related to health, such as body weight, blood pressure, activity levels, and medication adherence.

Expectations of utilizing personal lifelog data in medicine are increasing [3,4] as accurate and reliable monitoring becomes possible. This situation has prompted many developments in health self-tracking devices, but also has resulted in the use of a large number of uncontrolled terms in lifelogging, which is due to the terms from mobile health self-tracking devices having evolved in a haphazard and uncoordinated manner. For instance, 2 representative tracking tools (Fitbit and Runkeeper) have different representations for recording of running—*running at 5 mph (12 min/mile)* vs. *running on the treadmill with moderately activity*—and for types of activities—*running, Bicycling, and Walking* vs. *Run, Cycling, and Walk.* This heterogeneity of lifelog data represents a significant obstacle for individuals wanting to combine their own tracking data from disparate sources into a single summarized report or an integrated dataset.

Microsoft HealthVault [5] is a platform for storing and maintaining health and fitness information. It works with connected apps and devices, and helps users to share their information with other people. Health and fitness data can be transferred from HealthVault-enabled wireless devices stored in multiple data exchange formats or personal health records, including using industry standards such as the Continuity of Care Record (CCR) [6] from the American Society for Testing and Materials and the Continuity of Care Document (CCD) [7] from Health Level 7 International (HL-7), with the aim of supporting interoperability. HealthVault tried to solve the problem of heterogeneous lifelog data that has made it difficult to integrate and analyze by providing a lifelog-related vocabulary, called the HealthVault vocabulary (HVV). HVV is probably the first lifelog-related vocabulary to be used by more than 200 health self-tracking devices. HVV was originally developed in order to encourage users to use common terms among HealthVault-enabled devices, thereby facilitating efficient data sharing. However, only certain lifelog terms are supported by HVV, and these do not have precise definitions and their interrelationships are unclear.

The various vocabularies and ontologies developed by the medical informatics community include a large number of clinical concepts and standards for reporting clinical events and healthcare activities. We assumed that these medical vocabularies are designed to include lifelog data as physicians obtain status information about patients, such as that related to exercise, diet, and symptoms by filling out medical questionnaires. We examined the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) [8,9] in the Unified Medical Language System (UMLS) to investigate how lifelog terms are represented. SNOMED-CT was chosen as a representative medical ontology since it contains not only clinical findings used in electronic medical record systems with comprehensive clinical scope but also the activities of daily living. It is freely accessible by registered users of UMLS. SNOMED-CT covers the highest percentage (81%) of the UMLS concepts [10,11] among the numerous ontologies and terminologies in the UMLS Metathesaurus.

UMLS concepts with 420 atomic unique identifiers (AUIs) were matched to 372HVV lifelog terms (Supplementary File 1). Each SNOMED-CT concept is designated with one Concept Unique Identifier (CUI). The CUIs and AUIs have a one-to-many relation, in that every CUI (concept) is associated with at least one AUI. We used the AUI because it is being used for representing the SNOMED-CT hierarchical structure in the MRHIER table in UMLS. We found that the 372 HVV lifelog terms were distributed in 11 of 22 first-level nodes of SNOMED-CT. These lifelog terms were very sparse and widely dispersed over the hierarchical structure of SNOMED-CT (Fig. 1) such that the definition and description of the semantic relationships between the lifelog terms were not rich enough to organize them by using SNOMED-CT hierarchy into a meaningful structure.

The widely scattered distribution of lifelog terms along with the poor definition and description of the relationships between concepts in the currently existing vocabularies indicate the need to develop a standard lifelog term set and/or ontology. The present study proposed the Medical Lifelog Ontology (MELLO), which has been developed by identifying lifelog concepts, determining the relationships between concepts, and providing clear definitions and descriptions by using a standard ontology development method. MELLO aims to provide a uniform way of representing and classifying lifelog terms for use by mobile and wearable devices. MELLO permits inference and reasoning of the relationships between lifelog terms and concepts, and it also addresses the ambiguity problem of lifelog terms by including UMLS concept names and CUIs.

2. Materials and methods

MELLO was developed using the General Formal Ontology method [12,13] with a manual iterative process comprising five steps: (1) defining the scope of lifelog data, (2) identifying lifelog concepts, (3) assigning relationships among MELLO concepts, (4) developing MELLO properties (e.g., synonyms, preferred terms, and definitions) for each MELLO concept, and (5) evaluating representative layers of the ontology content.

2.1. Defining the scope of MELLO

To determine the scope of lifelog data, we first prepared 3 kinds of source data: (1) proprietary lifelog terms from 25 popular health self-tracking devices and mobile apps, which were selected as they were the 25 best-selling apps in Health & Fitness category of Apple App Store at June 2013 (Supplementary Table S1), (2) the representative HVV [14,15] in the 70 subclasses that are classified by subjects such as medication-related, status-related, and disease-dependent (Supplementary Table S2), and (3) the SNOMED-CT and NCI thesaurus (NCIt) in UMLS that are flagged (as 'SNOMEDCT_US' and 'NCI') in the source abbreviation column of the MRCONSO table.

2.2. Identifying lifelog concepts

To obtain a concrete and precise understanding of the range of lifelog terms, we first defined a lifelog term as any term that was generated and recorded for an individual by using practical health self-tracking devices during his/her lifetime. More specifically, we classified lifelog terms into two components: (1) MELLO primary terms, which have independent meanings and describe fundamental lifelogging activities such as *walking* and *measuring body weight* and (2) MELLO secondary terms, which describe the context of lifelogging situations for primary lifelog terms. For example, given *My body temperature taken by mouth is 36 °C*, *Body temperature* is classified as MELLO primary term and *°C* and *mouth* are classified as MELLO secondary terms. MELLO concepts are composed of MELLO primary terms and MELLO secondary terms. The following sections address the systematic process that is illustrated in Fig. 2 for identifying MELLO concepts.

2.3. Systematic process for identifying primary lifelog terms

By using 2 sets of source data, we collected 250 primary seed terms: (1) 50 lifelog terms that are used in real-world practice were manually selected from 25 popular health self-tracking devices and mobile apps and (2) 200 lifelog terms were collected from the 6 subclasses of HVV (*icd9cm-reactions*,