A multi-part matching strategy for mapping LOINC with laboratory terminologies

Li-Hui Lee,1,2 Anika Groß,1,3 Michael Hartung,1,3 Der-Ming Liou,4 Erhard Rahm1,3

ABSTRACT

Objective To address the problem of mapping local laboratory terminologies to Logical Observation Identifiers Names and Codes (LOINC). To study different ontology matching algorithms and investigate how the probability of term combinations in LOINC helps to increase match quality and reduce manual effort.

Materials and methods We proposed two matching strategies: full name and multi-part. The multi-part approach also considers the occurrence probability of combined concept parts. It can further recommend possible combinations of concept parts to allow more local terms to be mapped. Three real-world laboratory databases from Taiwanese hospitals were used to validate the proposed strategies with respect to different quality measures and execution run time. A comparison with the commonly used tool, Regenstrief LOINC Mapping Assistant (RELMA) Lab Auto Mapper (LAM), was also carried out.

Results The new multi-part strategy yields the best match quality, with F-measure values between 89% and 96%. It can automatically match 70–85% of the laboratory terminologies to LOINC. The recommendation step can further propose mapping to (proposed) LOINC concepts for 9–20% of the local terminology concepts. On average, 91% of the local terminology concepts can be correctly mapped to existing or newly proposed LOINC concepts.

Conclusions The mapping quality of the multi-part strategy is significantly better than that of LAM. It enables domain experts to perform LOINC matching with little manual work. The probability of term combinations proved to be a valuable strategy for increasing the quality of match results, providing recommendations for proposed LOINC concepts, and decreasing the run time for match processing.

BACKGROUND AND SIGNIFICANCE

There is increasing use of standard terminologies (ontologies)1 for electronic health records2–4 and public health surveillance.5–10 One of these standard terminologies for laboratory and clinical observations is Logical Observation Identifiers Names and Codes (LOINC),11 12 which is used to facilitate information exchange—for example, for reporting infectious diseases to the Centers for Disease Control and Prevention (CDC). LOINC is a national standard in several countries such as the USA, Canada, Germany and Taiwan.13 However, hospitals usually have their own local laboratory terminologies,14 often focusing on certain aspects such as laboratory results,5 9 10 15–18 clinical reports,19 and radiology reports.20 21 This situation has led to a growing need to determine mapping between local and standard terminologies22 in order to correctly share data. Several semiautomatic LOINC matching algorithms and tools have been proposed to assist domain experts,4 7 9 10 16 19 23 but mapping from local terminologies to LOINC is still a resource-intensive and time-consuming problem.9

Each LOINC concept (term) comprises a fully specified name (ie, the formal LOINC name11) and an identity code.24 Each fully specified name consists of six parts—for example, Component/Analyte, Kind of Property, etc (see table I for several examples)—separated by colons (eg, ‘Rotavirus Ag:ACnc:Pt:Stool:Ord:Aggl’). LOINC is primarily a pre-coordinated terminology14 23 that a priori defines which combinations of atomic concepts are allowed. For instance, the six-part combination of ‘Rotavirus Ag’, ‘ACnc’, ‘Pt’, ‘Stool’, ‘Ord’ and ‘Aggl’ is meaningful for disease surveillance and has thus been included in LOINC with a unique code (S879-2). Its long common name is ‘Rotavirus Ag [Presence] in Stool by Agglutination’. For improved coverage and completeness, there is also the possibility to include additional concepts (post-coordination) in LOINC. In particular, users or developers can suggest new combinations of existing atomic concepts (proposed LOINC concepts) to the development organization to be considered for inclusion in LOINC.4 23 24 26 27

Existing approaches to mapping local terminologies with LOINC rely on extensive preprocessing of the local terminologies to facilitate automated match processing. Common preprocessing steps include normalization of local terminologies,10 18 adoption of synonyms,16 23 28 resolution of abbreviations,4 10 and augmenting local terms with definitions and annotations.9 25 Existing tools for matching the preprocessed local terminologies with LOINC usually calculate concept similarities by using the fully specified names—for example, ‘Rotavirus Ag:ACnc:Pt:Stool:Ord:Aggl’ (S879-2). For example, most previous studies used the Regenstrief LOINC Mapping Assistant (RELMA),4 5 9 16 which applies such a full-name strategy based on restrictive exact name matching. Only a few studies applied linguistic matching,10 19 where a match can also be identified for similar but slightly different names. Fidahussein and Vreeman investigated corpus-based matching to leverage a collection of previously matched terminologies.18 Lau et al23 considered multi-part matching with separate name matching for each part; such an approach can avoid false matches due to a name similarity in unrelated LOINC parts (eg, ‘SP’ is the abbreviation of ‘species’ in Component/Analyte as well as for ‘sputum’ in Sample Type). Moreover, Bodenreider proposed a multi-part matching technique for comparison of laboratory tests between LOINC and SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms).20


CrossMark
Although previous approaches have already proved quite useful, they still require a substantial amount of human effort, with 11–31% of the local terms requiring manual assignment to corresponding LOINC concepts. Moreover, 6–37% of the local concepts could not be mapped at all, since they do not exist in LOINC (ie, unmappable/uncodable concepts).

A further problem is internationalization—that is, support of local terminologies in different languages. While RELMA supports translations between some languages such as English, German and Simplified Chinese, there are more than 26 000 LOINC users from 157 countries and therefore many languages are not yet supported.

**OBJECTIVE**

We aimed to improve mapping of local terminologies to LOINC with respect to the amount of the remaining manual work. We built on recent advances in automated ontology matching (alignment), an area that has seen a large amount of research and development effort. Proposed match approaches and prototypes combine a variety of similarity measures considering the name, description and structural neighborhood of concepts, or further background information. In particular, we used the state-of-the-art match tool GOMMA (Generic Ontology Matching and Mapping Management). We also devised and evaluated a new multi-part match strategy. Specifically, we made the following contributions.

1. We studied full-name matching using the fully specified name in GOMMA to find the most applicable match approach for laboratory data. We considered different preprocessing, matching and selection methods as well as language translations.

2. We proposed a novel multi-part matching strategy that uses the occurrence probability of combinations of atomic LOINC concepts as a filter. The approach can also provide recommendations for possible (new) combinations of atomic concepts that could match unmapped local terms. To our knowledge this is the first work to automatically provide such recommendations, thereby supporting post-coordination.

3. We evaluated both strategies by comparing them with the RELMA Lab Auto Mapper (LAM) using datasets from three Taiwanese hospitals. The results show that multi-part matching with probability filtering can significantly outperform other approaches with respect to quality and execution time.

**MATERIALS AND METHODS**

We first describe the local laboratory terminologies used. We then define the assumed data model and present the full name and multi-part matching strategies. Finally, we outline our evaluation methods.

**Materials**

Because of a request from the Taiwanese Center of Diseases Control (Taiwan CDC), in the pilot project, we focused on their 49 most important pathogen-related tests for notifiable disease surveillance; the 49 pathogens are shown in box 1. Hence, only the tests related to these pathogens are considered in the study. To validate our algorithms, we used three local laboratory terminologies from Taiwanese hospitals. Two terminologies come from medical centers: Shin Kong Wu Ho-Su Memorial Hospital (SKH) with 919 beds and Mackay Memorial Hospital Hsin-Chu branch (Mackay) with 1118 beds. The third comes from a regional hospital, the National Taiwan University hospital Hsin-Chu branch (NTU) with 818 beds. The hospitals use different commercial laboratory information systems, TrakCare Lab, Amesdata and Tatung, which are also used in several other Taiwanese hospitals. Since the three systems do not have unified codes corresponding to LOINC codes, we collected their laboratory terminologies, including ‘test order names’, ‘sample type names’, ‘abnormal results and definitions’, ‘normal results and definitions’, ‘measurement units’ and ‘test device or reagent names’. In April 2013, the SKH, Mackay and NTU contained 452, 245 and 116 unique concepts, respectively, related to these pathogens. All the laboratory terminologies were converted into the LOINC data model. The pilot study started in May 2012. We used LOINC 2.38 (extracted through UMLS MetamorphoSys 2012AA for additional synonyms from Metathesaurus) released in December 2011 and took into account the assumed data model and present the full name of each pathogen.

**Box 1** Pathogen names

<table>
<thead>
<tr>
<th>Pathogen names</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haemophilus influenzae type b (Hib)</td>
<td>Vibrio cholera</td>
</tr>
<tr>
<td>Hepatitis A virus</td>
<td>Chikungunya virus</td>
</tr>
<tr>
<td>Listeria monocytogenes</td>
<td>West Nile virus</td>
</tr>
<tr>
<td>Neisseria meningitides</td>
<td>Rickettsia prowazekii</td>
</tr>
<tr>
<td>Respiratory syncytial virus</td>
<td>Bordetella pertussis</td>
</tr>
<tr>
<td>Rotavirus</td>
<td>Japanese encephalitis virus</td>
</tr>
<tr>
<td>Salmonella species</td>
<td>Legionella pneumophila</td>
</tr>
<tr>
<td>Shigella species</td>
<td>Leptospiro interrogans</td>
</tr>
<tr>
<td>Streptococcus agalactiae (Group B Strept)</td>
<td>Burkholderia pseudomallei</td>
</tr>
<tr>
<td>Streptococcus pneumonia</td>
<td>Clostridium botulinum</td>
</tr>
<tr>
<td>Streptococcus pyogenes (Group A Beta Hemolytic Strept)</td>
<td>Coxiella burnetii</td>
</tr>
<tr>
<td>Yersinia enterocolitica</td>
<td>Rickettsia typhi</td>
</tr>
<tr>
<td>Campylobacter species</td>
<td>Borrelia burgdorferi</td>
</tr>
<tr>
<td>Influenza virus</td>
<td>Francinsella tularensis</td>
</tr>
<tr>
<td>Parainfluenza virus</td>
<td>Oriente tutsugamushi</td>
</tr>
<tr>
<td>Measles virus</td>
<td>Varicella-Zoster virus</td>
</tr>
<tr>
<td>Mumps virus</td>
<td>Bartonella henselae</td>
</tr>
<tr>
<td>Rubella virus</td>
<td>Toxoplasma gondii</td>
</tr>
<tr>
<td>Enterovirus</td>
<td>Neisseria gonorrhoeae</td>
</tr>
<tr>
<td>Corynebacterium diphtheriae</td>
<td>Hepatitis B virus</td>
</tr>
<tr>
<td>Dengue virus</td>
<td>Hepatitis C virus</td>
</tr>
<tr>
<td>Plasmodium vivax, P malariae,</td>
<td>Mycobacterium tuberculosis complex,</td>
</tr>
<tr>
<td>P falciparum, P ovale</td>
<td>multi-drug resistant (MDR-TB)</td>
</tr>
<tr>
<td>Entamoeba histolytica</td>
<td>Mycobacterium tuberculosis complex (TB)</td>
</tr>
<tr>
<td>STEC (E. coli 0157:H7, E. coli 0157:NM)</td>
<td>Treponema pallidum</td>
</tr>
<tr>
<td>Hantavirus</td>
<td></td>
</tr>
</tbody>
</table>
account all LOINC concepts with class type ‘lab’ (45,896 LOINC codes).

Reference mappings (gold standard) between laboratory terminologies and LOINC were created manually by three independent experts familiar with these local laboratory concepts. We manually examined the accuracy and consistency of the preliminary reference mappings of each expert. In the case of any mapping inconsistency, these three experts and two more external experts were invited to discuss it conjointly and agree to reasonable matches and consistent correspondences. Note that the mappings also include unmappable concepts.

Data model
We assumed that an ontology (terminology) $O$ consists of $n$ sub-ontologies (parts) $SO_1, \ldots, SO_n$—for example, LOINC has six parts $COM, \ldots, MET$. Each part contains a number of atomic concepts, $c_i \in SO_i$ for $1 \leq i \leq n$, which can be used to build combined concepts, $c = (c_1, \ldots, c_n)$—for example, ‘Rotavirus Ag’ from $COM$ part, ‘Aggl’ from $MET$ part, etc. form the combined concept LOINC 5879-2. Thus, it is possible to build up to $32,564 \times 3 \times 1149$ combined concepts. For instance, the multiplication of the number of each LOINC part size results in more than 37 million combined concepts (for LOINC 2.38). However, only commonly used combinations $c \in C$ are defined as concepts in $O$—that is, only 68,350 combined concepts are allowed in LOINC 2.38. Thus an ontology can be described as follows: $O = \{(SO_1, \ldots, SO_n) \mid C \subseteq SO_1 \times \cdots \times SO_n$, $n \geq 1\}$. For instance, LOINC is described as $LOINC = ((COM, PRO, TIM, SAM, SCA, MET), C)$, where $C$ contains the 68,350 combined concepts. Other combinations $c \notin C$ are named as proposed (post-coordinated) concepts. Such combinations present practical events and can be suggested as new concepts to the ontology’s maintenance organization. According to table 1, (Rotavirus Ag, ACnc, Pt, Stool, Ord, Aggl) $\in C$ is a pre-coordinated concept in LOINC, composed of the atomic concepts Rotavirus Ag $\in COM$, ACnc $\in PRO$, etc. In contrast, the practically useful combination (Rotavirus Ag, ACnc, Pt, Stool, Ord, EIA) $\notin C$ is a proposed concept and could be submitted to the Regenstrief Institute for future inclusion in LOINC.

Ontology matching
In this section, we present our two matching strategies: full name and multi-part. Figure 1 illustrates the general situation. Full name matching (blue line) directly compares names of combined concepts from source and target ontology ($O$ and $O'$). For example, the local concept ‘ROTA VIRUS:ACNC:PT:St:Ord:AGGL’ matches the LOINC concept ‘Rotavirus Ag:ACnc:Pt:Stool:Ord:Aggl’. The multi-part matching (red dash line) aligns names of atomic concepts from source and target sub-ontologies to finally find correspondences between combined concepts of $O$ and $O'$. For example, the local COM concept ‘ROTA VIRUS’ matches the LOINC COM concept ‘ROTA VIRUS Ag’, the local SAM concept ‘St’ matches the LOINC SAM concept ‘Stool’, etc. In general, ontology matching is the process of determining a set of semantic correspondences (mapping) between concepts of two related ontologies $O$ and $O'$ (eg., the local terminology and LOINC): $M = \{(c, c', \text{sim}) \mid c \in O, c' \in O', \text{sim} \in [0, 1]\}$ A correspondence is a connection between two combined concepts $c$ and $c'$, whereby the strength of the connection is represented by a similarity value sim ranging from 0 to 1.

Full name matching
GOMMA is freely available for research purposes and can be accessed via http://dbs.uni-leipzig.de/de/gomma. It has achieved top

---

**Full name strategy**

**Preprocessing**
- Normalizer
- Translator

**Matching**
- Name/Synonym

**Selection**
- Threshold
- Max1 or MaxN

---

**Multi-part strategy**

**Preprocessing**
- Normalizer
- Translator

**Matching**
- Probability calculation
- Name/Synonym

**Selection**
- Threshold
- Max1

---

**Recommendations Generation**
- Name/Synonym

**Selection**
- Threshold
- MaxN (TopK)
results in the 2012 benchmarking competition OAEI (Ontology Alignment Evaluation Initiative). Our full name matching is a linguistic matching approach using GOMMA based on comparing names and synonyms of combined concepts. Figure 2A shows the overall matching process. Before matching, names and synonyms of concepts are preprocessed, including normalization to lowercase and replacement of common delimiters, abbreviations and synonyms. (The same preprocessed local concepts' names were also imported to RELMA LAM to ensure comparability.) Some local concepts contain parts in Traditional Chinese. We translated these parts to English using MyMemory API and replaced them with translated English names. During the actual matching, we applied a name/synonym matcher determining the maximal string similarity for the names and synonyms per (combined) concept pair. We used trigram as the string similarity measure. Trigram tokenizes strings into tokens of length three and considers the overlap of trigram sets to compute a similarity value using the Dice metric:

$$\text{sim}_{\text{trigram}}(s_1, s_2) = \frac{2 \times \text{No. of overlapping trigrams between } s_1 \text{ and } s_2}{\text{No. of all trigrams in } s_1 + \text{No. of all trigrams in } s_2}.$$ 

For instance, the number of overlapping trigram tokens between ‘rotavirus acnc pt st ord aggl’ (including 31 tokens, eg, ‘rot’, ‘ota’, ‘tav’,...) and ‘rotavirus ag acnc pt stool ord aggl’ (including 37 tokens, eg, ‘rot’, ‘ota’, ...) is 29. Hence, the trigram similarity value is \(\frac{(2 \times 29)}{(31 + 37)} = 0.85\).

Note that an automatic match strategy may produce a set of correspondences for each (local) source concept. To provide domain experts with a manageable set of correspondences, we first applied a simple threshold filter to discard all correspondences below a minimal threshold (eg, 0.60). We then applied a second, more sophisticated selection strategy, such as MaxN. For each local combined concept, MaxN selects the top N correspondences. Max1 only selects the best correspondence, while N presents a larger set of match candidates, allowing expert users to select the correct match. These selection filters were applied from the source to the target ontology, since we aimed to find one or more LOINC candidate correspondences for each local concept.

Multi-part matching

In contrast with full name strategy, the multi-part strategy (figure 2B) considers the atomic concepts of combined parts for matching. As depicted in figure 1 for LOINC, we matched atomic concepts from the parts COM, PRO, TIM, SAM, SCA and MET individually—that is, we applied different match approaches for the different parts and chose the best matches for each part. However, treating the atomic concepts independently from each other in the match process can cause problems. For instance, when one considers the best matches in each part for a combined concept, it is likely that the combination of these atomic concepts may not be a pre-coordinated concept in the target ontology. However, our aim was to match to pre-coordinated concepts whenever possible. For this purpose, we also considered the probability of possible combinations of target atomic concepts. We first describe how we determined the probability of combinations and then outline the overall algorithm for multi-part matching.
Each atomic concept of an ontology part has an inherent certainty, which can be derived from its occurrence probability with respect to all (pre-coordinated) combined concepts. For our example in table 1, this probability is 0.60 (three of five concepts) for the SAM concept ‘Stool’ and even 1.0 for the TIM concept ‘Pt’. We can also determine the certainty for combinations of atomic concepts from two or more ontology parts by determining their conditional occurrence probabilities. For example, the conditional probability for the COM concept ‘Rotavirus Ab’ given SAM concept ‘Stool’, P(COM=‘Rotavirus Ab’|SAM=‘Stool’) is 0.33 (one of three occurrences). In contrast, for the COM concept ‘Rotavirus Ab.IgG’, the conditional probability for the SAM concept ‘Stool’ is 0, since both atomic concepts do not occur together in any combined concept. In our multi-part match strategy described next, we used such predetermined conditional properties to focus on finding matches only for combinations of atomic concepts with non-zero conditional occurrence probabilities.

Algorithm
The detailed workflow of the multi-part match strategy (figure 2B) is displayed in figure 3. The pseudo code appears in online supplementary algorithm 1. The workflow (algorithm) shows how we matched the input source ontology, O, with the target ontology, O’. We computed the best matching concept in the target ontology O’=(⟨SO_1,…,SO_n⟩,C⟩ (eg, LOINC=((COM,..,MET), 68 350 combined concepts) for a given combined concept c = (c_1,…,c_n) of source ontology O— for example, (‘ROTA VIRUS’, ‘ACNC’, ‘PT’, ‘St’, ‘Ord’, ‘AGGL’). We further used different thresholds per sub-ontology t_1,…,t_n ∈ [0,1] to select matches for the atomic concepts c_1,…,c_n. The output is the best matching concept r’=(r'_1,…,r'_n) in O’—that is, (c, r’) forms a correspondence in the final mapping. For example, ((‘ROTA VIRUS’, ‘ACNC’, ‘PT’, ‘St’, ‘Ord’, ‘AGGL’), (‘Rotavirus Ab’, ‘ACNC’, ‘Pt’, ‘Stool’, ‘Ord’, ‘Aggl’)) is one of the correspondences in the final mapping.

Multi-part matching requires the specification of an order in which the different parts are handled. In our case for c = (c_1,…,c_n) of O, we first matched c_1 to SO_1, then c_2 to SO_2, etc. The order is provided as an input parameter. Users can manually provide (manual order (MO)) or use the following scheme (automatic order (AO)) to automatically derive an ordering. For AO, we considered the sub-ontology with the fewest concepts as SO_1 (SCA in the case of LOINC). We then iteratively selected the sub-ontology, SO_2, with the minimal number of distinct concept pairs with respect to the previously selected sub-ontology, SO_1. For LOINC, we thus generated the following order—SCA-TIM-PRO-MET-SAM-COM—since the number of atomic concept pairs for SCA-TIM is smaller than for SCA-PRO, SCA-MET, etc; the number of atomic concept pairs for TIM-PRO is smaller than for TIM-MET, TIM-SAM, etc. The scheme aims at maximal filtering to reduce the search space for finding correct matches. This is because we only considered (LOINC) atomic concepts c_i of the i-th sub-ontology as a match candidate if there existed a combined concept for c_i with the previously determined matches, c_1,…,c_{i-1}. The proposed AO ordering thus minimizes the number of match candidates, c_i, to consider. In the evaluation, we also studied alternative orders including a manually optimized order.

For each atomic concept, c_i, of source concept, c, the algorithm computes the best matching atomic concept, c_i’, in SO_i (figure 3 steps 1–3 or see online supplementary algorithm 1 lines 2–14). For the first sub-ontology (eg, SCA), we computed similarity between the local atomic concept (eg, ‘Ord’) with all the LOINC (eg, SCA) atomic concepts (eg, ‘Qn’, ‘Ord’, ‘OrdQn’, ‘Nar’, ‘Doc’, and ‘-’) using GOMMA matchers as described in the full name approach (step 2.1). However, for all further sub-ontologies, we first determined the conditional probability of the current atomic concept, c_i (eg, c_i), based on the occurrence of already matched parts in the result (eg, ‘Ord’, ‘-’) (step 2.2.1 or see online supplementary algorithm 1 line 5)—that is, we calculated P(SO_i=c_i|SO_1=r'_1,…,SO_{i-1}=r'_{i-1}) (eg, P(TIM=c_i’|SCA=‘Ord’)). Only if the conditional probability is greater than zero, we consider c_i’ (eg, c_i’=‘Pt’, or c_i’=‘24H’) and computed the similarity between c_i (eg, c_i=‘Pt’) and c_i’ (eg, c_i’=‘Pt’). For c_i’=‘24H’) (step 2.2.2 or see online supplementary algorithm 1 line 7) using GOMMA matchers as described in the full name approach. The resulting similarity was evaluated—that is, we looked for the atomic concept, c_i’ (eg, c_i’=‘Pt’), with the maximum similarity (Max1 selection). This atomic concept is regarded as the best matching concept for the particular part, SO_i (eg, LOINC TIM) (step 3 or see online supplementary algorithm 1 line 12). If we did not find any best matching concept (ie, the similarity was too low or the probability was zero for all atomic concepts of SO_i), we left this part empty (0) in the result.

To assess the impact of probability filtering, we also performed multi-part matching without this step (skipping steps 2.2.1 and 2.2.2 or online supplementary algorithm 1 lines 5–6).

For unmapped concepts, one can generate possible recommendations (dashed line rectangle in figure 2) for such parts by applying computeRecommendations (its detailed flow is shown in figure 4).

**Figure 4** Algorithm for computeRecommendations.
in Figure 4; the pseudo code appears in online supplementary algorithm 2), which creates correspondences to proposed concepts. In particular, for each empty part we computed the k best matching atomic concepts (using GOMMA matchers and MaxN selection) as recommendations (steps 2 and 3 or see online supplementary algorithm 2 line 3). Users can afterwards judge these recommendations and accept them or not for their final mapping. For instance, if there were a missing SCA concept in result \( r’ = \text{‘Rotavirus Ag:ACnc:Pt:Stool: \theta :Aggl’} \), the algorithm could recommend the SCA atomic concepts, ‘Ord’, ‘Nom’ and ‘Qt’, based on the concepts from Table 1.

We executed a multi-part strategy algorithm (Figure 3 or see online supplementary algorithm 1) for each combined concept in \( O \) to create a mapping \( M \) between \( O \) and \( O’ \). The result contains correspondences to pre-coordinated concepts. In contrast with the full name strategy, one can optionally use computeRecommendations to create correspondences to proposed concepts for unmapped parts and unmappable concepts as well.

Evaluation methods
We first present the basic information of the source and target (sub) ontologies by calculating the number of concepts and atomic concepts as well as the percentage of mappable (\( \text{covmap} \)) and unmappable (\( \text{covunmap} \)) local concepts with respect to all local concepts. We then evaluate the match algorithms RELMA v5.6 LAM (using LOINC 2.38 and enabling the option ‘Prefer Common Tests Results’), GOMMA and the multi-part strategy for the following six configurations:

1. F-Max10-RELMALAM: the full name strategy with Max10 selection by RELMA LAM
2. F-Max10-GOMMA: the full name strategy with Max10 selection by GOMMA
3. F-Max1-RELMALAM: the full name strategy with Max1 selection by RELMA LAM
4. F-Max1-GOMMA: the full name strategy with Max1 selection by GOMMA
5. MP-Max1-GOMMA-AO: the multi-part strategy with probability filtering and Max1 selection by GOMMA using automatic ordering (SCA-TIM-PRO-MET-SAM-COM)
6. MP-Max1-GOMMA-MO: the multi-part strategy with probability filtering and Max1 selection by GOMMA using optimized manual ordering (SCA-TIM-PRO-MET-COM-SAM)

Table 2 Statistics about laboratory terminologies and LOINC 2.38

<table>
<thead>
<tr>
<th>Number of combined concepts or atomic concepts</th>
<th>Source (sub) ontologies</th>
<th>Target (sub) ontologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>SKH</td>
<td>Mackay</td>
<td>NTU</td>
</tr>
<tr>
<td>Related 49 pathogens</td>
<td>37</td>
<td>28</td>
</tr>
<tr>
<td>Six-part combinations</td>
<td>452</td>
<td>245</td>
</tr>
<tr>
<td>COM</td>
<td>36</td>
<td>32</td>
</tr>
<tr>
<td>PRO</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>TIM</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SAM</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>SCA</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>MET</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Mappable combinations (( \text{covmap} ))</td>
<td>403 (89%)</td>
<td>184 (75%)</td>
</tr>
<tr>
<td>Unmappable combinations (( \text{covunmap} ))</td>
<td>49 (11%)</td>
<td>61 (25%)</td>
</tr>
</tbody>
</table>

LOINC, Logical Observation Identifiers Names and Codes.

We compare each approach with the gold standard established manually with respect to precision, recall and F-measure (harmonic mean of precision and recall) for mappable local concepts as well as execution run time. For the unmapped local concepts, the multi-part strategy can additionally produce recommendations. We show how the percentage of correctly mapped local concepts (number of true positives divided by total number of local concepts) changes for a different number of recommended atomic concepts. The experiments were performed on an Intel W3520 machine (4×2.67 GHz, 4GB RAM).

RESULTS
Table 2 shows that a significant number of combined concepts can be created by atomic concepts. For example, the 45 896 LOINC codes comprise about 20 000 distinct atomic concepts. The local terminologies have 116–452 concepts, whereas there are only 1–37 atomic concepts in six parts. Some (8–25%) of the local concepts are unmappable (\( \text{covunmap} \)).

The achieved match quality of the six match configurations considered is shown in Figure 5. The two Max10 configurations aim at high recall by offering up to 10 match candidates per local term. Here, GOMMA has a better recall (84% vs 68%) than RELMA LAM for the NTU terminology but performs similarly (~94%) for the other two cases. Precision and thus F-measure are very low, since only 1 of the 10 proposed candidates is probably correct. Using the Max1 configuration, precision (F-measure) improves significantly by 57% (52%) in RELMA LAM and 66% (63%) in GOMMA, while recall is lower than for Max10 (average values for three datasets). For all three mapping problems, the multi-part strategies MP-Max1-GOMMA-AO and MP-Max1-GOMMA-MO achieve significantly better results than the best strategy so far, GOMMA (Max1). Using MO (AO), precision increases by 18% (14%), while recall improves slightly by 5% (2%), leading to an F-measure improvement of 12% (8%) on average. They also performed in the shortest execution time of 29–97 s, while RELMA LAM required 2–5 min—that is, about four times longer.

We now turn our attention to which shares of the local terminologies could be correctly mapped to LOINC. These shares are restricted not only because of less-than-perfect recall, but also because of unmappable terms (according to the reference mapping). Overall, the following percentages of local concepts (averaged over the three datasets) could be correctly mapped to pre-coordinated LOINC concepts: 72% (F-Max10-RELMALAM), 76% (F-Max10-GOMMA), 67% (F-Max1-RELMALAM), 73% (F-Max1-GOMMA), 74% (MP-Max1-GOMMA-AO) and 77% (MP-Max1-GOMMA-MO).

Figure 6 shows these percentages for MP-Max1-GOMMA-MO for the three local terminologies as green bars. The figure also shows to what degree the recommendation algorithm of MP-Max1-GOMMA-MO could find additional correspondences when varying the number k of recommended atomic concepts per part from 0 to 10 (horizontal axis). The red bars indicate the improved percentage of local terms, with a correct recommendation for mappable concepts (\( \text{covmap} \)).

The figure shows that the recommendation approach achieves an improved mapping coverage of 0%, 2% and 5% in the first case (red bars). The inclusion of proposed LOINC concepts further increases the number of correctly mapped local concepts by 10%, 18% and 4% (blue bars) to 95%, 90% and 87%, respectively—that is, 91% on average.
DISCUSSION

Full name strategy

Full name matching results are different in RELMA LAM and GOMMA, since the former adopts exact name matching, whereas the latter uses a trigram similarity function. This approximate similarity measure allows more robust and flexible matching—that is, names/synonyms of concepts with small differences can still be matched within a certain threshold. For the Max10 selection, the focus is on high recall, and GOMMA performs slightly better than RELMA LAM in this respect (90% vs 86% recall averaged over the three matching tasks). For Max1, F-Max1-GOMMA outperforms RELMA LAM not only with respect to recall (86% vs 78% averaged) but also for precision (75% vs 69%) and F-measure (80% vs 73%).

Multi-part strategy

The results are further improved by the proposed multi-part match strategy. A key to its success is the applied filtering with respect to conditional probability of concept parts, because it successfully avoids mapping to non-existing combinations of atomic LOINC concepts. Using the automatically computed order (MP-Max1-GOMMA-AO), we already achieved very good average precision (88%), recall (89%) and F-measure (88%). Based on expert knowledge, we manually optimized the order (MP-Max1-GOMMA-MO) leading to excellent match quality, with high precision (93% averaged), recall (91%) and F-measure (92%) results. Moreover, the generated recommendations leave only a small portion of the local terminologies unmapped (5% of SKH, 10% of Mackay, 13% of NTU) and thereby reduce the manual effort required to complete the mapping.

For comparison, we also tested other orders as well as multi-part matching without probability filtering, but observed significantly lower match quality as well as increased execution times (to 6–27 min) in all these cases. For example, the use of the LOINC-specific ordering of sub-ontologies (COM-PRO-TIM-SAM-SCA-MET) leads to a significantly lower F-measure value of 82%. For multi-part matching without probability filtering, the F-measure was even lower (71%). This is because independently matching concept parts often leads to concepts whose atomic concept combinations are non-existent in LOINC, so that only a few local concepts could be matched.

Unmappable local concepts

We classified the unmappable local concepts into several types. First, a local concept (eg, ‘Virus identified:Prid:Pt:?:NOM:Culture’) has a sample type (eg, Conjunctiva=’Cnjt’), which is not included in the current LOINC. Next, the data representation of a test result (eg, for ‘Escherichia coli O157:H7:Agnc:Pt:Stool:?:Organism specific culture’) is different between local designers (eg, used nominal ‘Nom’) and the LOINC developer (eg, used ordinal ‘Ord’). Third, a few unmappable cases are that the combined local atomic concepts in PRO-SCA, PRO-SCA-MET or COM-PRO-SAM-MET parts are not
In this study, we used translation for Traditional Chinese to terminology. We plan to evaluate other local laboratory datasets.

The correct mapped concepts in rate (%) contained in LOINC. We plan to further investigate the reasons for generating proposed LOINC concepts to help mapping adaptation for new LOINC versions.

**Limitations and future work**

The proposed multi-part matching process uses a relatively strict Max1 selection for atomic concepts. In future work, we plan to provide users with the possibility to choose an atomic concept out of a list of suggestions. This could be useful in cases where experts have background knowledge on complex decisions for some sub-ontology. Similarly, we would like to add more user interaction during recommendations of proposed concepts.

We have proposed a simple and effective automatic ordering strategy for the multi-part strategy. However, we found that using expert knowledge to manually adapt this can lead to even better results. For future work, we thus plan to study further automatic techniques for determining optimized sub-ontology orders.

We analyzed datasets for three hospitals and were able to produce very good results. However, a limitation of this study is that only a small, specialized subset of LOINC was evaluated (813 local concepts associated with tests for 49 microbiology pathogens), making it difficult to generalize the conclusions to different classes of LOINC laboratory tests or the full LOINC terminology. We plan to evaluate other local laboratory datasets. In this study, we used translation for Traditional Chinese to English, but translation from other languages to English is also worth investigating. Moreover, we plan to evaluate our approaches to other pre-coordinated terminologies such as SNOMED CT and ICNP (International Classification for Nursing Practice).

**CONCLUSION**

A new multi-part matching strategy with conditional occurrence probability filtering for pre-coordinated LOINC terminology is proposed. The mapping quality of the proposed strategy is significantly better than RELMA LAM and the use of full name matching. The proposed recommendation of new combinations of atomic concepts has been shown to improve the mapping coverage for local terminologies and to automatically support post-coordination.

**Contributors**

L-HL, AG, and MH conceived and designed the study, performed the analysis, and wrote and edited the manuscript. L-HL created the strategy and evaluated the results. D-ML coordinated cooperative organizations, collected data, and input practical concerns to the study. ER coordinated the study and revised the manuscript critically for important intellectual content.

**Funding**

The work was supported by ‘NSC-DAAD Sandwich Program, 2012 Spring’ from the Taiwan National Science Council (NSC) and the German Academic Exchange Service (DAAD) under grant number NSC101-2911-I-010-001-2 and partly supported by Taiwan Centers for Disease Control (CDC) under grant number DOH101-DC-1014.

**Competing interests**

None.

**Provenance and peer review**

Not commissioned; externally peer reviewed.

**REFERENCES**


