ABSTRACT

Motivation: Mapping internal, locally used lab test codes to standardized logical observation identifiers names and codes (LOINC) terminology has become an essential step in harmonizing electronic health record (EHR) data across different institutions. However, most existing LOINC code mappers are based on text-mining technology and do not provide robust multi-language support.

Materials and methods: We introduce a simple, yet effective tool called big data-guided LOINC code mapper (BGLM), which leverages the large amount of patient data stored in EHR systems to perform LOINC coding mapping. Distinguishing from existing methods, BGLM conducts mapping based on distributional similarity.

Results: We validated the performance of BGLM with real-world datasets and showed that high mapping precision could be achieved under proper false discovery rate control. In addition, we showed that the mapping results of BGLM could be used to boost the performance of Regenstrief LOINC Mapping Assistant (RELMA), one of the most widely used LOINC code mappers.

Conclusions: BGLM paves a new way for LOINC code mapping and therefore could be applied to EHR systems without the restriction of languages. BGLM is freely available at https://github.com/Bin-Chen-Lab/BGLM.

Key words: big data, LOINC code mapping, electronic health records, multi-language support
LAY SUMMARY
Mapping internal, locally used lab test codes to standardized logical observation identifiers names and codes (LOINC) terminology has become an essential step in harmonizing Electronic Health Record (EHR) data across different institutions. In this study, we propose a novel tool called big-data guided LOINC code mapper (BGLM) for automatic LOINC code mapping. We validated the performance of BGLM with real-world EHR datasets and showed that high mapping precision could be achieved under proper false discovery rate control. In addition, we showed that the mapping results of BGLM could be used to boost the performance of Regenstrief LOINC Mapping Assistant (RELMA), one of the most widely used LOINC code mappers. Compared with existing methods, there are two significant advantages of BGLM. First, BGLM provides robust multi-language support; second, the performance of BGLM could be further improved with more EHR data integrated. BGLM has the potential to become a useful tool in large-scale EHR data harmonization procedures and would be of immense interest to both the scientific and clinical community.

INTRODUCTION
Laboratory test data stored in electronic health record (EHR) systems are valuable resources for clinical studies. Although more and more EHR datasets are being utilized for research, each institution uses its local coding system to refer to lab tests, and such inconsistency in lab test reference creates a significant barrier for downstream integrative analysis. Logical observation identifiers names and codes (LOINC) provides a universal coding standard for lab tests and is widely used in EHR data standardization processes. To enable lab test data to be integrable with data from other institutions, mapping internally used lab test local codes to LOINC codes is necessary.

Since the number of lab test local codes stored in an EHR system is large, manual mapping is time-consuming and not feasible at scale. Multiple tools have been developed to automate and accelerate the LOINC code mapping process. Although different in detail, most of them conduct the mapping based on the similarity between text-formatted metadata (such as lab test names, descriptions) of local codes and LOINC codes. These automatic LOINC code mappers have achieved high success; however, a few challenges remain and need to be addressed.

First, existing methods have limited multi-language support. Currently, almost all LOINC code mappers only accept English texts as their input. It is time-consuming for non-native English speakers to accurately translate the texts stored in their parent language to English and it is even infeasible when performing mapping in large-scale EHR systems.

Second, existing methods do not fully utilize patient data. The EHR system stores not only the metadata of lab tests but also their observed values in patients. The lab test metadata is relatively static and rarely updated while the amount of patient data rapidly increases as long as the EHR system is in use. These accumulated patient data are also valuable resources for LOINC code mapping; however, no strategy has been proposed to leverage them.

Here, we propose a novel big-data guided LOINC code mapper (BGLM), which simultaneously addresses the above two technical challenges. Different from existing methods, BGLM maps local codes to LOINC codes based on distributional similarity. We validated the performance of BGLM with in-house EHR datasets and showed that high mapping precision could be achieved under proper false discovery rate (FDR) control.

MATERIALS AND METHODS
The BGLM workflow is composed of three major steps (Figure 1).

Step I. Construction of empirical cumulative distribution functions
Given an EHR dataset (which we call the “query”), various lab test local codes (denoted as Ci, C2, … CN) are needed to be mapped. For each local code, 1000 reading values of the corresponding lab test are randomly drawn from the EHR system to construct its empirical cumulative distribution function (eCDF). Similarly, we construct the eCDFs of LOINC codes (denoted as L1, L2, … LK) using the reference EHR dataset (in which LOINC codes, instead of local codes, are used to refer to lab tests).

Step II. Generation of mappings
We form a N × K distance matrix D with its entry D[i, j] representing the Kolmogorov–Smirnov test statistic between the eCDFs of Ci and Lj. Then, the LOINC code mapping of Ci is 

\[ i^* = \arg \min_j D[i, j] \quad j \in \{1, 2, \ldots, K\} \]

Step III. Control of FDR
Given Ci, we derive a Z-score vector (denoted as Z) by scaling the i-th row of the distance matrix D. If Z[i] (Z-score of the mapping <Ci, Lj>) is smaller than the predefined cut-off, the mapping <Ci, Lj> is accepted; otherwise, it is considered a wrong mapping and dropped.

RESULTS
The publicly available EHR dataset MIMIC-III was used as the reference to construct eCDFs of LOINC codes. Our in-house EHR dataset derived from Spectrum Health West Michigan was used as the query. There were 113 local codes with more than 1000 data points and their manually mapped LOINC codes (which are considered as true mappings) were within MIMIC-III. We utilized BGLM to generate the mappings for them and observed that the derived Z-scores of correct mappings were significantly lower than those of incorrect mappings (P-value = 1.192e−08, Supplementary Figure S1a), which provides strong evidence to use the Z-score to control FDR. Since there is randomness in our method, we ran BGLM (on the same datasets) 30 times to get an average performance curve (Supplementary Figure S1b). The precision value is 0.33 if all the mappings are accepted; however, it could be boosted to 0.79 using Z-score cut-off −5.

We also performed the mapping with Regenstrief LOINC Mapping Assistant (RELMA), one of the most widely used LOINC code mappers. Although the performance of RELMA is better than BGLM (precision = 0.43, Supplementary Figure S1b), it does not provide any criteria to help users control FDR. We therefore
designed an ensemble LOINC code mapper based on the rule shown in Figure 2 and found its precision achieved 0.5, suggesting the results of BGLM could be used to improve the mapping quality of RELMA. More details of the ensemble LOINC code mapper are included in the Supplementary Material.

**DISCUSSION AND CONCLUSIONS**

The LOINC code mapping in BGLM is solely dependent on distributional similarity and such characteristic enables it to be applied to any EHR systems without considering the internally used language. BGLM paves a new way of LOINC code mapping and achieves ideal performance on our in-house and the public MIMIC-III datasets. Meanwhile, we notice that there are limitations which should be overcome to further enhance BGLM.

First, only the MIMIC-III dataset was used to construct the eCDFs of LOINC codes and this may bring biases to the mapping process. In the future, we will integrate multiple publicly available EHR datasets (such as NIH’s All of Us researcher workbench, UK Biobank, and N3C) to construct more accurate reference eCDFs of LOINC codes, which should improve the performance of BGLM.7,8 Prior to our study, Bradwell et al9 proposed a revolutionary data-driven approach to harmonize units and values of quantitative data elements in a large nationally pooled EHR dataset, suggesting the validity and feasibility of such strategy in heterogeneous EHR data harmonization.

Second, the current version of BGLM does not check the consistency of “unit of measure” between local code and its mapped LOINC code, which may result in false positive mappings. We will introduce additional matching processes to overcome such limitation in the next release of BGLM.

**DATA AVAILABILITY STATEMENT**

The source code and tutorial of BGLM are freely available at https://github.com/Bin-Chen-Lab/BGLM. Our testing in-house data are available upon reasonable request from Spectrum Health.

**FUNDING**

This work is supported by Spectrum Health-MSU Alliance Fund. The content is solely the responsibility of the authors and does not necessarily represent the official views of sponsors.

**AUTHOR CONTRIBUTIONS**

K.L. conceived the study. K.L. and M.W.-L. performed the majority of computational analysis. D.C. and B.C. supervised the study. All authors contributed to the writing, reviewing, and editing the manuscript and approved the manuscript.

**SUPPLEMENTARY MATERIAL**

Supplementary material is available at JAMIA Open online.

**CONFLICT OF INTEREST STATEMENT**

None declared.

**REFERENCES**

1. Vreeman DJ, McDonald CJ, Huff SM. LOINC®: A universal catalogue of individual clinical observations and uniform representation of enumerated

---

**Figure 1.** Detailed workflow of BGLM. $C_i$, the $i$-th local code; $L_j$, the $j$-th LOINC code; eCDF: empirical cumulative distribution function; and KS-statistic, Kolmogorov–Smirnov test statistic.

**Figure 2.** Design of the ensemble LOINC code mapper. Given a local lab test code C, suppose the LOINC mapping generated by BGLM and RELMA are B and R, respectively. If B and R are identical (or the $Z$-score of B is smaller than $-3.80$), the mapper mapped C to B; otherwise, it mapped C to R.
collections. *Int J Funct Inform Personal Med* 2010; 3 (4): 273–91. doi:10.1504/IJFFIPM.2010.040211

2. Parr SK, Shotwell MS, Jeffery AD, et al. Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. *J Am Med Inform Assoc* 2018; 25 (10): 1292–300. doi:10.1093/jamia/ocy110

3. Fidahussein M, Vreeman DJ. A corpus-based approach for automated LOINC mapping. *J Am Med Inform Assoc* 2014; 21 (1): 64–72. doi:10.1136/amiajnl-2012-001159

4. Kim H, El-Kareh R, Goel A, et al. An approach to improve LOINC mapping through augmentation of local test names. *J Biomed Inform* 2012; 45 (4): 651–7. doi:10.1016/j.jbi.2011.12.004

5. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 1–9. doi:10.1038/sdata.2016.35

6. Abhyankar S, Demner-Fushman D, McDonald CJ. Standardizing clinical laboratory data for secondary use. *J Biomed Inform* 2012; 45 (4): 642–50. doi:10.1016/j.jbi.2012.04.012

7. Lyles CR, Lunn MR, Obedin-Maliver J, et al. The new era of precision population health: Insights for the all of us research program and beyond. *J Transl Med* 2018; 16 (1): 5–8. doi:10.1186/s12967-018-1585-5

8. Sudlow C, Gallacher J, Allen N, et al. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; 12 (3): e1001779. doi:10.1371/journal.pmed.1001779

9. Bradwell KR, Wooldridge JT, Amor B, et al.; the N3C Consortium. Harmonizing units and values of quantitative data elements in a very large nationally pooled electronic health record (EHR) dataset. *J Am Med Inform Assoc* 2022; 29 (7): 1172–82. doi:10.1093/jamia/ocac054