LOINC Codes That May Contain Personally Identifiable Information

Eugene M. Sadhu, MD[1], Malarkodi J. Samayamuthu, MD[1], Shyam Visweswaran MD, PhD[1]

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Introduction

Providing clinical data for research often requires suppressing personally identifiable information (PII) and, specifically for patients, via the removal of protected health information (PHI). Structured electronic health record (EHR) data are often coded with standard terminologies such as the International Classification of Diseases (ICD) or the Current Procedural Terminology (CPT) that do not contain PHI and do not require deidentification. Logical Observation Identifiers, Names, and Codes (LOINC)[1], a standardized coding system widely used in laboratory information systems and in EHRs differs from the ICD and CPT systems in that, unlike them, it may be used to code PII. An earlier effort, which was mostly manual, resulted in the creation of a disallow list of LOINC codes that may contain PII[2]. In contrast, we used a combination of rule-based search and manual review to identify LOINC codes that may contain PII. Our method is repeatable, and new LOINC releases require only incremental effort to search and review.

Methods

We downloaded LOINC version 2.74 from the LOINC website and loaded the data into a SQLite database. We constructed a broad set of inclusion criteria and implemented them as database queries to extract a list of LOINC codes that potentially contain PII; the queries were designed to be more sensitive than specific in identifying appropriate codes. We included only codes that had the following values in the LOINC Property field (since we determined that codes with other values cannot contain PII): Clock time, Date, URI, Date and Time Range, Extended Identifier, Scope.modifier.document template, "-", Address (HL7 data type), Angle, Arbitrary, Bibliographic Citation, Class, Complex, Description, Finding, History, Identifier, Impression/interpretation of study, Instructions, Location, DateRange, Person name, Presence or Identity, Scope.modifier.selection item, Scope.modifier.time window, Telephone number (HL7 datatype), Time (e.g., seconds), Time Stamp (Date and Time), TQ2 (HL7 datatype), Text, Extended Address (HL7 datatype), and Email Address. We further restricted the selection to codes containing one of the strings in the Long_Common_Name column as shown in Table 1 (first column).

Table 1. Counts of LOINC codes.

Inclusion criterion string	Count before review	Count after review
"address"	105	35
"number"	299	100
"name"	266	115
"phone"	166	27
"fax"	3	2
"email"	14	7
"city"	78	9
"age"	347	1
"social security"	8	4
"longitude"	6	4
"latitude"	7	4
"identifier"	1848	93
"date"	747	581
Total codes	**3,751**	**930**

Results

Application of the inclusion criteria reduced the number of codes from 99,687 to 3,751, and after review, all three reviewers agreed on 930 LOINC codes as likely to contain PII. The number of LOINC codes corresponding to each inclusion criterion before and after review are given in Table 1. In comparison, the earlier effort identified 601 LOINC codes[2].

Discussion and Conclusion

We have created a preliminary list of LOINC codes that potentially contain PII available publicly at https://github.com/shyamvis/LOINC-PII. While the three reviewers obtained consensus on 930 LOINC codes, there were 187 LOINC codes that two of the three reviewers labeled as likely to contain PII and 69 LOINC codes that only one reviewer labeled as likely to contain PII. We plan to discuss the disagreements and perform an additional round of review to obtain consensus. In the future, we plan to distinguish between date and ZIP code containing LOINC codes from the other PII containing LOINC codes so that the lists can be useful for deidentifying to HIPAA-limited and safe harbor standards. Furthermore, we plan to maintain this list and update it as new versions of LOINC are released.

References

1. Huff SM, Rocha RA, McDonald CJ, et. al. Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998 May-Jun;5(3):276-92.
2. Pfaff ER, Haendel MA, Kostka K, Lee A, Niehaus E, Palchuk MB, Walters K, Chute CG. Ensuring a safe(r) harbor: Excising personally identifiable information from structured electronic health record data. J Clin Transl Sci. 2021 Dec 9;6(1):e10.

LOINC Codes That May Contain Personally Identifiable Information

Eugene M. Sadhu, MD, Malarkodi J. Samayamuthu, MD, Shyam Visweswaran MD, PhD

Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

INTRODUCTION

Providing clinical data for research often **requires suppressing personally identifiable information (PII)** and, specifically for patients, the removal of protected health information (PHI). Structured electronic health record (EHR) data are often coded with standard terminologies such as the International Classification of Diseases (ICD) or the Current Procedural Terminology (CPT) that do not contain PHI and do not require deidentification. Logical Observation Identifiers, Names, and Codes (**LOINC**), a standardized coding system widely used in laboratory information systems and EHRs differs from the ICD and CPT systems in that, unlike them, it **may be used to code PII**.

We used a combination of **rule-based search and manual review** to identify LOINC codes that may contain PII. Our method is **repeatable**, and new LOINC releases **require only incremental effort** to search and review.

Table 1: Counts of LOINC Codes

Inclusion criterion string	Count before review	Count after review
"address"	105	35
"number"	299	100
"name"	266	115
"phone"	166	27
"fax"	3	2
"email"	14	7
"city"	78	9
"age"	347	1
"social security"	8	4
"longitude"	6	4
"latitude"	7	4
"identifier"	1848	93
"date"	747	581
Total codes	**3,751**	**930**

METHODS

We downloaded LOINC version 2.74 from the LOINC website and loaded the data into a SQLite database. We constructed a **broad set of inclusion criteria** and implemented them as database queries to extract a list of LOINC codes that potentially contain PII. The **queries were designed to be more sensitive than specific** in identifying appropriate codes. We included only codes that had the following values in the LOINC Property field (since we determined that codes with other values cannot contain PII):

Address (HL7 data type)	Email Address	Scope.modifier.document template
Angle	Extended Address (HL7 datatype)	Scope.modifier.selection item
Arbitrary	Extended Identifier	Scope.modifier.time window
Bibliographic Citation	Finding	TQ2 (HL7 datatype)
Class	History	Telephone number (HL7 datatype)
Clock time	Identifier	Text
Complex	Impression/interpretation of study	Time (e.g., seconds)
Date	Instructions	Time Stamp (Date and Time)
Date and Time Range	Location	URI
DateRange	Person name	-
Description	Presence or Identity	

We further restricted the selection to codes containing one of the **strings** in the **Long_Common_Name column** as shown in Table 1 (first column).

RESULTS

Application of the inclusion criteria reduced the number of codes from **99,687 to 3,751**, and after review, all three reviewers agreed on **930 LOINC codes as likely to contain PII**. The number of LOINC codes corresponding to each inclusion criterion before and after review is provided in Table 1.

DISCUSSION and CONCLUSION

We have created a preliminary list of LOINC codes that potentially contain PII available publicly at https://github.com/shyamvis/LOINC-PII. While the three reviewers obtained consensus on **930 LOINC codes**, there were 187 LOINC codes that two of the three reviewers labeled as likely to contain PII and 69 LOINC codes that only one reviewer labeled as likely to contain PII. We plan to discuss the disagreements and perform an additional round of review to obtain consensus. An earlier, mostly manual effort by another team of researchers resulted in the creation of a disallow list of 567 LOINC codes that may contain PII[2].

In the future, we plan to distinguish between date and ZIP code containing LOINC codes from the other PII containing LOINC codes so that the lists can be useful for deidentifying to HIPAA-limited and safe harbor standards. We will also flag institution-identifying codes. Furthermore, we plan to maintain this list as a resource for the Informatics community and update it as new versions of LOINC are released.

REFERENCES

1. Huff SM, Rocha RA, McDonald CJ, et. al. Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998 May-Jun;5(3):276-92.
2. Pfaff ER, Haendel MA, Kostka K, Lee A, Niehaus E, Palchuk MB, Walters K, Chute CG. Ensuring a safe(r) harbor: Excising personally identifiable information from structured electronic health record data. J Clin Transl Sci. 2021 Dec 9;6(1):e10.