MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190198

Interactive Machine Learning for Laboratory Data Integration

Nathanael Fillmore[a], Nhan Do[a,b], Mary Brophy[a,b], Andrew Zimolzak[a,c]

[a] *Massachusetts Veterans Epidemiology Research and Information Center, VA Boston Healthcare System, Boston, MA, USA*
[b] *Boston University School of Medicine, Boston, MA, USA*
[c] *Baylor College of Medicine, Houston, TX, USA*

Abstract

Laboratory data collected in the electronic health record as part of routine care can be used in secondary research. For example, the US Department of Veterans Affairs maintains a data warehouse covering over 20 million individuals and 6.6 billion lab tests. However, data aggregation in such a data warehouse can be difficult. In order to retrieve all or nearly all of one type of lab result with a high degree of precision, we perform clinical concept adjudication, which is the process of an expert determining which database records correspond to a target clinical concept. In this work, we develop an interactive machine learning tool to "extend the reach" of expert laboratory test adjudicators. Our tool provides access to automatic laboratory classification in a user-facing front end that covers all steps in an adjudication workflow, in order to lower barriers to collaboration, increase transparency of adjudication, and to promote efficiencies and data reuse.

Keywords:

Clinical Laboratory Information Systems; Systems Integration; Supervised Machine Learning

Introduction

Clinical laboratory data are crucial to medical research, including retrospective studies and clinical trials. Laboratory data collected in the electronic health record (EHR) during the care process can be used in research (termed "secondary use") [1]. For example, serum creatinine lab test results, collected during routine clinical care, serve as safety endpoints for a "point of care" trial comparing the efficacy of antihypertensives [2]. Research requires all serum creatinine results to be retrieved using a unique database identifier, but reality is not this simple.

The United States Department of Veterans Affairs (VA) maintains a data warehouse covering 20 million individuals and 6.6 billion lab tests (as of 2014) [3], but as others have noted "data aggregation across the VHA is highly problematic" [4]. As one example, if we search for "creatinine" in the warehouse's LabChemTest dimension table, we find >1000 matching lab test types, many of which are not truly serum creatinine. If the query requires "creatinine" followed by "serum," it retrieves a shorter (64 lab types) and much more specific list, but many true positives are missed. Therefore, in order to retrieve all or nearly all the serum creatinine results with a high degree of precision, we perform *clinical concept adjudication*, which is the process of an expert determining which database records correspond to a target clinical concept.

The clinical concept adjudication process currently used in our research center has been previously described [5] and is designed to harmonize lab test results from 144 independent VA clinical laboratories (Figure 1). In brief: (1) Clinical subject matter experts (SMEs) design a search for appropriate laboratory test names. For example, if hemoglobin is the target concept, search terms might include "HGB" or "Hemoglobin", and several Logical Observation Identifiers Names and Codes (LOINCs). (2) Database technicians retrive candidate database records. For each record, the technicians also pull associated metadata, including specimen types (e.g. whole blood, urine, cerebrospinal fluid), units, distribution of numeric results, and laboratory test names. Many SMEs do not have technical expertise to perform a database pull, or they do not have permission to execute such a pull. Database technicians deliver these results to SMEs using an Excel spreadsheet that can be sorted and filtered (Table 1). (3) Two SMEs label each record, evaluating for appropriate specimen types, units, distribution, and test names. (4) SMEs resolve disagreements, producing a consensus labelling. (5) Database IDs and consensus labels of

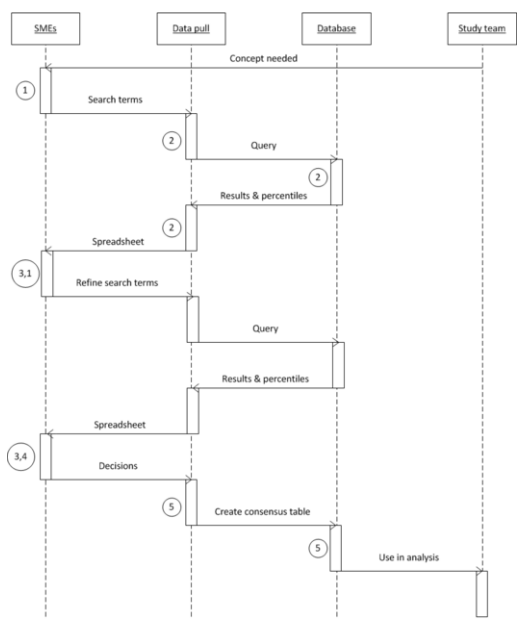

Figure 1– Sequence diagram. The workflow between subject matter experts, data technicians, the database, and the study team is illustrated.

Table 1– Example data used in adjudication. The task is to label each row, identified by the ID column, with a label, shown in the SME Label column, based on data in the remaining columns. The Lab Test Name, Topography, and LOINC columns identify the name and specimen type of the lab test. The Count column shows the number of lab test results associated with the ID, and p1, p50, and p99 (as well as others not shown) describe the first, 50th, and 99th percentiles of values of these results.

ID	SME Label	Lab Test Name	Topography	LOINC	Count	p1	p50	p99
1	Yes	SODIUM	SERUM	Missing	115053	126	140	149
2	No	RANDOM URINE SODIUM	URINE	Missing	734	6	52	194.1
3	No	SODIUM	URINE	Missing	89	5	49.5	155.9
8	Yes	SODIUM	SERUM	2947-0	126	133.2	140	144.7
9	No	SODIUM	URINE,24HR	2947-0	98	13.8	150	877.7
92	No	SODIUM	PERITONEAL	2950-4	10	124	132	138.8
461	Yes	SODIUM*IA	BLOOD	2950-4	714	125	139	170.9

"yes" or "no" for each record are stored in a new database table, and the spreadsheet is stored for future reference. Note that steps (1)-(3) are often iterated if the search terms need refinement.

Several drawbacks to this process exist. First, several steps in this process are time-consuming. Step (2) can take hours to run, because it may require computing on millions of values for each database ID. Step (3) requires SMEs ultimately to address each row (often about 1000) of a spreadsheet. The current process requires manual transmission of information and requires the SME and database technician to wait on each other. Second, the adjudicated concept can go out-of-date as new database records are added. Third, it is difficult for the final consumer (i.e. the study team performing analyses) to understand, validate, or adapt the new database table that contains the final adjudicated concept. This is partly because of the manual transmission of work products, which can result in the initial search terms (or other work products) being lost, or at least not available to final consumers.

Therefore, we sought to develop a machine learning tool to "extend the reach" of expert laboratory test adjudicators, so that they do not need to classify each data element manually. We aimed to provide access to automatic laboratory classification in a SME-facing front end for continued use by adjudicators. We intend our tool to cover all steps in the workflow (including the initial search for laboratory test names, the consensus process, and the database storage of SME decisions), in order to lower the barrier to collaboration among the group, to increase transparency of the adjudication process, and to promote efficiencies and reuse of adjudicated laboratory tests across the entire VA.

Prior work. Previous authors have faced similar lab result harmonization problems. For instance, the LOINC standard has been developed to identify clinical laboratory test results [6], and previous authors have described mapping their local data to this standard [7], but mappings of local laboratory tests to LOINC may be erroneous [8]. In another example, the Mini-Sentinel program received laboratory results from twelve data partners and encountered inconsistent units and LOINC availability, among other challenges, many of which were addressed by manual quality checking [9]. As an example from VA data, we found LOINC wrongly coded as body fluid sodium (2950-4) when it should be coded as serum/plasma (Table 1). This means that LOINC alone cannot be used to accurately retrieve all test types for a single clinical concept.

Early work has shown that lexical comparison of test names followed by subject matter expert review can integrate a more detailed set of laboratory test codes (LOINC) with a less detailed one (SNOMED procedures) [10]. To facilitate the mapping of local lab terms to LOINC, The Regenstrief LOINC Mapping Assistant (RELMA) was developed [11,12]. As a practical example of lab test harmonization, a large group of 44 European hospitals developed an internal dictionary of laboratory observations and mapped it to LOINC, but they found that some tests were not referenced in LOINC [13]. More advanced work has been been done to share results for LOINC mapping, to take advantage of "crowdsourcing" [14,15], and to perform automated mapping to LOINC using machine learning, which assigned about 70-80% of local terms to the correct LOINC [16]. Finally, work has been performed to address laboratory unit conversions [17] and standardizing laboratory test *results* across multiple institutions, as distinct from standardizing simply the type of test [18].

In our data warehouse, lab tests are already mapped to LOINC, but imperfectly. We also often need to aggregate several granular LOINCs into one more general concept. Our current lab adjudication process accomplishes this, but with many roles and manual steps (Figure 1).

Methods

We developed a machine learning system to predict whether a candidate lab test type clinically belongs within the concept of interest, based on a set of lab tests already labelled by SMEs as belonging or not. Seven lab tests were studied: alkaline phosphatase (ALP), alanine transaminase (ALT), albumin (ALB), high-density lipoprotein cholesterol (HDLC), sodium (Na), magnesium (Mg), and hemoglobin (HGB). We chose the following initial features for use by these algorithms. We used a bag-of-words encoding for longer textual fields like the test name, topography, component, and specimen. We used a categorical encoding for short textual fields, including the VA hospital site identifier, the lab result units, and the LOINC. We used as-is the numerical fields describing the number of results and their distribution. We included as a feature a Kolmogorov-Smirnov statistic that compares, for one lab test type, the distribution of its results to the distribution of all positive training examples' results [19].

Table 2– Machine learning results (accuracy) for each of 7 lab tests, using 3 different learning methods: LASSO (LR), support vector machines (SVM), and random forests (RF). Also shown are the total number of database IDs, the number of IDs labelled positive in the "ground truth" SME labelling, and the number of associated observations (lab test results).

Data Set	Number of Database IDs	Number of Positive Database IDs	Number of Observations	LR	SVM	RF
ALP	1588	747	38,297,131	0.98	0.97	0.98
ALT	716	341	84,180,150	0.98	0.94	0.92
ALB	3351	601	114,613,056	0.97	0.92	0.98
HDLC	742	358	79,767,824	0.98	0.91	0.98
Na	322	153	7,821,081	0.97	0.98	0.99
Mg	2066	502	17,768,685	0.97	0.95	0.99
HGB	1737	830	135,382,741	0.97	0.95	0.99

We evaluated this basic system using seven datasets that had been already adjudicated by VA experts (Table 2), and using three algorithms: logistic regression with an L1 penalty (also known as the least absolute shrinkage and selection operator, or LASSO), support vector machines (SVM), and random forests. Initial development and testing of learning algorithms was performed in Python using the scikit-learn package [20]. We used 10-fold cross validation to evaluate the accuracy of the system using each algorithm.

After the basic learning system was developed and validated, an interactive adjudication interface was implemented as a web application, as described in detail below. This system is available to VA SMEs in a hosted environment, as well as to the public at https://github.com/nfillmore/active-adjudication.

Results

Machine learning

Ten-fold cross-validation accuracy was high for all seven laboratory tests, and for all three methods (Table 2). For 6 out of 7 laboratory tests, random forests achieved the top cross validation accuracy, and LASSO achieved second place or better. Random forests achieved the fastest convergence per training example, and LASSO consistently achieved >90% cross validation accuracy at 100 or fewer training examples (Figure 2).

Operationalizing as web application

After the basic learning system was developed, an interface was designed as a web application, with a front-end written in JavaScript (Figure 3). We implemented the back-end (including LASSO) entirely in Java, which we found removes the need for any research group to install their own machine learning pipeline. The application covers the the laboratory adjudication workflow from initial assignment of the lab concept to a SME and generating search terms, to saving the final consensus adjudication for study use. The web application is hosted in a secure environment where VA SMEs can access it, and its source code is available for use by the public at https://github.com/nfillmore/active-adjudication.

Support for search term generation

Recall that in the current process for lab adjudication, data pull personnel must wait on SMEs to generate search terms, and SMEs then wait to receive an Excel spreadsheet. In order to avoid this inefficiency, we integrated the initial search into our interactive tool. The SME specifies search terms and can see results immediately. We accelerate this process with pre-computed value percentiles for all lab test identifiers in the VA system (this took over a month of wall clock time). After refining the search terms, the SME can begin labelling examples, although the query can be modified later without losing any labels already input.

Support for adjudication

The labeling interface allows SMEs to view both the original table of data elements and the feature matrix for the learning algorithm. These tables can be viewed separately or side-by-side, with support for sorting by any column. As the SME labels examples, statistics are recomputed in real time. The default feature matrix is as described above for the basic system, but SMEs may add or remove any features (a process termed *feature engineering*), including features corresponding to a

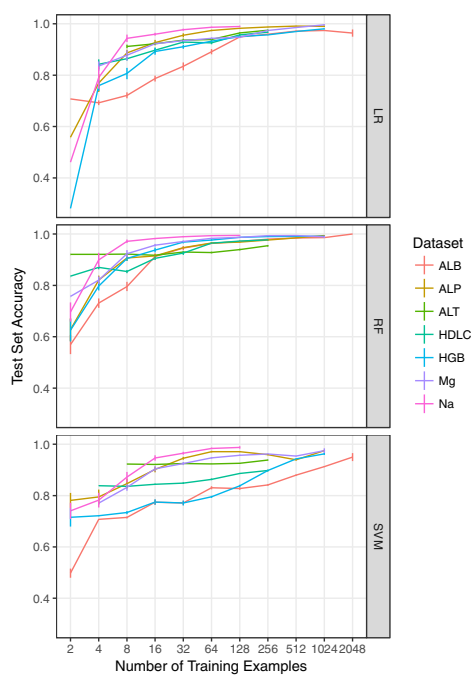

Figure 2– Learning curves for LASSO (LR), random forests (RF) and support vector machines (SVM)

single word or phrase. For example, if the SME needs to retrieve measurements of serum immunoglobulin free light chains, the phrase "free light" is more specific than either "free" or "light" by itself, and thus the SME may want to add this two-word feature, which is not available to the learning system by default.

Support for consensus labelling

The current adjudication process requires SMEs to resolve labelling conflicts. Our application facilitates this by displaying any conflicts and prompting for a consensus label. As in the main labelling interface, the application uses machine learning to predict the consensus label.

Support for finalizing and sharing results

Because lab adjudication does not involve any protected health information, adjudicated labels can be freely shared among different researchers, but there is currently no simple mechanism to do this within VA's nationwide network of researchers. Therefore, our tool offers such a mechanism. Each adjudication can be published to a searchable list of "Completed Adjudications", tagged with the target lab, the group that carried out the adjudication, the date of completion, and other information. These adjudications can be viewed and exported for use in individual research projects. In some cases, an adjudication might be close to what a research group is looking for, but not an exact match for their target concept. Thus, the tool also allows published adjudications to be "forked", i.e., copied into a new adjudication instance, where labels and search criteria can be modified as needed, exactly as if the adjudication had been created from scratch by the new researcher. The lineage of forking is tracked and displayed to users, to promote appropriate credit as well as transparency. To promote division of labor, even incomplete adjudications are viewable and forkable by other researchers, but these are clearly marked as incomplete.

Discussion

We have developed a tool that uses interactive machine learning to assist lab adjudication experts. Among the three learning methods tested, LASSO performed nearly as well as the others, with the advantage that it is easier to interpret the model's predictions. We found no large differences in classifier performance among the seven lab tests for which we simulated annotation. Because of the rapid convergence of the learning methods, only about 100 lab data elements need to be adjudicated, and the rest (about 1000, depending on the lab test) can be inferred with high accuracy, which is theoretically a 10-fold improvement in adjudication time. In practice, the expert could label all data elements with > 1000 observations and allow the system to predict the rest.

SMEs within the VA now have access to machine learning assistance for laboratory data integration, without the need to implement a learning pipeline. In addition, we have precomputed and stored percentiles of numerical laboratory values, so these are available rapidly to SMEs. Our web application covers all steps of the laboratory adjudication workflow (Figure 1, Figure 3), including capturing the initial query, and writing to the database for future use.

Our system can also be used by SMEs outside the VA, after creation of an appropriate backing database of precomputed metadata and percentiles of laboratory values. Creating such a database is straightforward and instructions are available with the software.

Figure 3– Screenshots of our tool: (A) Adjudication task creation. (B) Interactive annotation and machine learning for adjudication. (C) Sharing adjudicated results. Details on these steps are in the text.

We envision several feasible improvements to our system. First, a quantitative study of SME time saved could be undertaken. To facilitate this, we have already incorporated time-stamp logging of user actions into the software. Second, we could adapt the system to monitor the database for updates and ask for new labels as appropriate to keep concepts up-to-date. Third, we could use an approximation scheme for streaming computation of percentiles, so that the precomputed numerical distributions do not go out of date as new observations are collected [21].

In comparison to RELMA, we incorporate collaboration among multiple roles (e.g. SME, database technician), and we incorporate the aggregation of multiple potential LOINC code points. Machine learning has been applied to lab data cleaning [16], but to our knowledge *interactive* machine learning has not. Our web application covers all parts of the clinical laboratory test adjudication workflow, greatly decreasing the work required of SMEs, and it introduces further advantages by capturing work products that otherwise could be lost, increasing transparency, and encouraging sharing.

Conclusions

In this work, we developed an interactive machine learning tool for laboratory test adjudicaton. Our tool provides access to automatic laboratory classification in a SME-facing front end for continued use by adjudicators. The tool covers all steps in an adjudication workflow, and in doing so, lowers barriers to collaboration, increases transparency of adjudication, and promotes efficiencies and data reuse.

Acknowledgements

This material is based upon work supported by the United States Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Cooperative Studies Program (CSP) and the Veterans

Affairs/National Cancer Institute Big Data-Scientist Training Enhancement Program (BD-STEP). The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

References

[1] MIT Critical Data, Secondary Analysis of Electronic Health Records, 2016. doi:10.1007/978-3-319-43742-2.

[2] F.A. Lederle, W.C. Cushman, R.E. Ferguson, M.T. Brophy, and L.D. Fiore Md, Chlorthalidone Versus Hydrochlorothiazide: A New Kind of Veterans Affairs Cooperative Study., *Ann. Intern. Med.* **165** (2016) 663–664. doi:10.7326/M16-1208.

[3] S.D. Fihn, J. Francis, C. Clancy, C. Nielson, K. Nelson, J. Rumsfeld, T. Cullen, J. Bates, and G.L. Graham, Insights from advanced analytics at the Veterans Health Administration., *Health Aff. Proj. Hope*. **33** (2014) 1203–1211. doi:10.1377/hlthaff.2014.0054.

[4] B.P. Giroir, and G.R. Wilensky, Reforming the Veterans Health Administration — Beyond Palliation of Symptoms, *N. Engl. J. Med.* **373** (2015) 1693–1695. doi:10.1056/NEJMp1511438.

[5] S. Raju, Y.-L. Ho, A. Zimolzak, B. Katcher, K. Cho, and D. Gagnon, Validation of Laboratory Values in a Heterogeneous Healthcare System: The US Veterans Affairs Experience., *31st Int. Conf. Pharmacoepidemiol. Ther. Risk Manag. ICPE Boston 822-262015*. (n.d.).

[6] A.W. Forrey, C.J. McDonald, G. DeMoor, S.M. Huff, D. Leavelle, D. Leland, T. Fiers, L. Charles, B. Griffin, F. Stalling, A. Tullis, K. Hutchins, and J. Baenziger, Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results., *Clin. Chem.* **42** (1996) 81–90.

[7] A.N. Khan, S.P. Griffith, C. Moore, D. Russell, A.C. Rosario Jr, and J. Bertolli, Standardizing laboratory data by mapping to LOINC, *J. Am. Med. Inform. Assoc. JAMIA*. **13** (2006) 353–355. doi:10.1197/jamia.M1935.

[8] M.-C. Lin, D.J. Vreeman, C.J. McDonald, and S.M. Huff, Correctness of Voluntary LOINC Mapping for Laboratory Tests in Three Large Institutions, *AMIA Annu. Symp. Proc. AMIA Symp.* **2010** (2010) 447–451.

[9] M.A. Raebel, K. Haynes, T.S. Woodworth, G. Saylor, E. Cavagnaro, K.O. Coughlin, L.H. Curtis, M.G. Weiner, P. Archdeacon, and J.S. Brown, Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel., *Pharmacoepidemiol. Drug Saf.* **23** (2014) 609–618. doi:10.1002/pds.3580.

[10] R.H. Dolin, S.M. Huff, R.A. Rocha, K.A. Spackman, and K.E. Campbell, Evaluation of a "lexically assign, logically refine" strategy for semi-automated integration of overlapping terminologies., *J. Am. Med. Inform. Assoc. JAMIA*. **5** (1998) 203–213.

[11] C.J. McDonald, S.M. Huff, J.G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook, W. Williams, J. Case, and P. Maloney, LOINC, a universal standard for identifying laboratory observations: a 5-year update., *Clin. Chem.* **49** (2003) 624–633.

[12] Regenstrief Institute, Inc, Common LOINC Laboratory Observation Codes, (n.d.). http://loinc.org/usage/obs.

[13] P.-Y. Vandenbussche, S. Cormont, C. Andre, C. Daniel, J. Delahousse, J. Charlet, and E. Lepage, Implementation and management of a biomedical observation dictionary in a large healthcare information system., *J. Am. Med. Inform. Assoc. JAMIA*. **20** (2013) 940–946. doi:10.1136/amiajnl-2012-001410.

[14] B.E. Dixon, J. Hook, and D.J. Vreeman, Learning From the Crowd in Terminology Mapping: The LOINC Experience., *Lab. Med.* **46** (2015) 168–174. doi:10.1309/LMWJ730SVKTUBAOJ.

[15] D.J. Vreeman, J. Hook, and B.E. Dixon, Learning from the crowd while mapping to LOINC., *J. Am. Med. Inform. Assoc. JAMIA*. **22** (2015) 1205–1211. doi:10.1093/jamia/ocv098.

[16] M. Fidahussein, and D.J. Vreeman, A corpus-based approach for automated LOINC mapping., *J. Am. Med. Inform. Assoc. JAMIA*. **21** (2014) 64–72. doi:10.1136/amiajnl-2012-001159.

[17] R.G. Hauser, D.B. Quine, and A. Ryder, LabRS: A Rosetta stone for retrospective standardization of clinical laboratory test results., *J. Am. Med. Inform. Assoc. JAMIA*. **25** (2018) 121–126. doi:10.1093/jamia/ocx046.

[18] R.G. Hauser, D.B. Quine, A. Ryder, and S. Campbell, Unit conversions between LOINC codes., *J. Am. Med. Inform. Assoc. JAMIA*. **25** (2018) 192–196. doi:10.1093/jamia/ocx056.

[19] N. Smirnov, Table for Estimating the Goodness of Fit of Empirical Distributions, *Ann Math Stat.* **19** (1948) 279–281. doi:10.1214/aoms/1177730256.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, Scikit-learn: Machine Learning in Python, *J Mach Learn Res*. **12** (2011) 2825–2830.

[21] T. Dunning, and O. Ertl, Computing Extremely Accurate Quantiles using t-Digests, (2017). https://github.com/tdunning/t-digest/blob/master/docs/t-digest-paper/histo.pdf.

Address for correspondence

Nathanael Fillmore (Nathanael.Fillmore@va.gov) and Andrew Zimolzak (Andrew.Zimolzak@va.gov)