A Rationale for Parsimonious Laboratory Term Mapping by Frequency
Daniel J. Vreeman, PT, DPT, MSc, John T. Finnell, MD, MSc,
J. Marc Overhage, MD, PhD
Regenstrief Institute, Inc. and Indiana University, Indianapolis, IN

ABSTRACT
Mapping local observation codes to a standard vocabulary provides a bridge across the many islands of data that reside in isolated systems, but mapping is resource intensive. To help prioritize the mapping effort, we analyzed laboratory results reported over a thirteen month period from five institutions in the Indiana Network for Patient Care. Overall, more than 4,000 laboratory observation codes accounted for almost 49 million results. Of the observations reported in the thirteen months, 80 codes (2%) accounted for 80% of the total volume from all institutions and 784 codes (19%) accounted for 99% of the volume from all institutions. The 244 to 517 observation codes that represented 99% of the volume at each institution also captured all results for more than 99% of the patients at that institution. Our findings suggest that focusing the mapping effort on this modest set of high-yield codes can reduce the barriers to interoperability.

INTRODUCTION
Despite mature standards in many areas, interoperable electronic health information exchange is hampered by the plethora of idiosyncratic conventions for representing identical concepts in separate electronic systems. Mapping local observation terms to a standardized vocabulary provides a bridge across the many islands of data that reside in isolated systems. A comprehensive information exchange must coalesce all of the various sources that produce health data to provide clinicians with complete information when and where they need it. Linkages across independent systems that yield semantic interoperability enable consolidation of a given patient’s data for clinical reports, decision support, public health, and research purposes. Too often, however, these linkages and mappings are not in place, and valuable data like laboratory results are unavailable to clinicians when they need it.

The Indiana Network for Patient Care is an early example of an operational local health information exchange (HIE) that has been operating in central Indiana for over ten years. The INPC includes data from five major hospital systems (over fifteen different hospitals and more than a hundred clinics), the state and county public health departments, Indiana Medicaid, and RxHub. The federated INPC repository now stores more than a billion discreet clinical observations.

The INPC has coalesced many of the various sources that produce and store data in our community, with emerging clinical and financial benefits. In the INPC collaborative, Regenstrief Institute serves as a neutral third party convener. Regenstrief receives all of the clinical messages streams from participating systems and accomplishes the task of integrating data these sources by mapping the idiosyncratic local terms to a common master dictionary based on LOINC® (Logical Observation Identifiers Names and Codes), a universal code system for identifying laboratory and other clinical observations. Presently, over one hundred source systems send HL7 clinical result messages to Regenstrief within the INPC.

Mapping the local observation codes from all of these data sources requires substantial effort and domain expertise. Laboratory data is particularly challenging to map because of the large number (2,000-5,000) of distinct test observations per laboratory, the short and often ambiguous names, and the relative lack of additional data in master files to inform the mapping process (as billing codes can for radiology tests). Our centralized approach consolidates the expertise and tools necessary for the task, but nevertheless we find the work of managing source system mappings a challenging aspect of operating a HIE.

Several studies have demonstrated that automated tools can help improve the efficiency and consistency of mapping. For example, Zollo and Huff described an automated mapping approach using extensional definitions for local codes based on data from repositories, and found that it reduced the mapping effort. Regenstrief has developed the freely available program called the Regenstrief LOINC Mapping Assistant (RELMA) that provides tools for mapping local laboratory and other observation codes to LOINC. Even with the best available automated tools, however, mapping between systems or to a standardized nomenclature like LOINC® is a complex process that often requires extensive human review by domain experts.

Frassica reported on the frequency of laboratory tests performed in the intensive care setting and noted
that a relatively small subset of laboratory tests (approximately 200) contributed to the 99% of all testing and could fulfill the needs of commonly used acuity scoring systems. We hypothesized that a modest number of laboratory tests would also account for a large proportion of the results in a community-wide HIE. We also suspected that many patients would only have results from the common tests that accounted for a large proportion of the volume. If this were true, then mapping only a modest number of tests could result in all of the data for many patients being mapped and most of the data for the remaining patients being mapped.

To evaluate these hypotheses, we leveraged our efforts to comprehensively map laboratory data in the INPC. Specifically, the purposes of this study were to 1) characterize the groups of laboratory observation codes that account for varying proportions of the results within the INPC, and 2) to determine what proportion of patients would have all of their results contained in these groups of codes.

METHODS

Data Sources
To identify the group of laboratory observation codes most frequently reported in the INPC, we extracted observations from five participating institutions. These five institutions represent autonomous and competing health care systems, which together provide the vast majority of acute medical care in the greater Indianapolis area. Two of the hospitals from these systems are designated level one trauma centers and receive major trauma by protocol. The INPC laboratory results for these institutions contain data originating from both the inpatient and outpatient setting and from point-of-care tests when they are reported through the laboratory. This study was approved by our local institutional review board.

From each institution’s INPC repository, we extracted a one month (January 2007) and a one year (January 1st, 2006 – December 31st, 2006) sample of all clinical results. The observations from each of these institutions have been mapped to the INPC’s common concept dictionary. The data in each extract contained a unique patient identifier for that institution, our local common dictionary identifier, date, and an identifier for the institution. The patient identifier enabled us to identify all tests performed at that institution for a particular patient, and the dictionary identifier enabled us to aggregate results of the same test performed at different institutions. While the INPC does contain a global patient identifier to link across institutions, we constrained our analysis to patients within each institution separately.

The extracted data from each institution were parsed using Perl to identify observations from laboratory tests and then loaded into a PostgreSQL database (http://www.postgresql.org) for processing. We first restricted the observations by the internal term classes from our concept dictionary that represented primarily laboratory values. Second, we manually reviewed this subsequent set of codes, and excluded codes that were misclassified in our master dictionary as true laboratory observations (e.g. echocardiogram findings).

We then used SQL statements to obtain subsets of the data for analysis. All statistical analyses were performed with R (http://www.r-project.org).

Measures
We analyzed the extracted data from each institution and in aggregate to determine the laboratory observations that represented 80%, 90%, 95%, 98%, and 99% of the testing completed for these core INPC institutions. We assessed the degree to which the lists of most frequently occurring observations remained stable by comparing lists generated from the one month extract with lists generated from the one year extract within each institution.

Each of these INPC institutions represents a healthcare system that serves a different, though somewhat overlapping, patient population. We therefore sought to characterize the set of observations that account for the majority of results by determining which observations were shared and which were unique among institutions.

In order to determine what proportion of patients would have all of their results contained in these groups of codes, we identified the percentage of patients for whom all of their laboratory test results were present in the lists of observations accounting for 80%, 90%, 95%, 98% and 99% of the tests completed during one year for each institution.

RESULTS

There were 4,086 unique observation codes in the final data set from all institutions. These codes represented 48,913,006 laboratory observations from all institutions for the entire thirteen month period. The number of patients from each hospital system with observations in this set ranged from 153,865 to 408,402.
The distribution of observation codes in this aggregate set was used to construct Figure 1, which illustrates the cumulative proportion of results represented by the observation codes ordered from most common to least common. Of particular note in Figure 1 is the exponential nature of the curve. The long tail represents the large number of observation codes that would have to be mapped to cover 100% of the reported results.

Figure 1. Laboratory observation codes (ordered by descending frequency) that account for cumulative observation volume.

In order to better understand how many observations account for a specific percentage of the overall volume of results reported, we chose cut-offs that represented 80%, 90%, 95%, 98%, and 99% of the results reported for the individual institutions over the one month and one year timeframes. (Table 1) For each institution, one month data was highly representative of one year data from the same source; very few new observations were reported over the course of a year. Pearson’s correlation coefficient between one month and one year of data from the same source was 0.99 for all institutions.

Figure 2 illustrates the number of unique observation codes by institution and the number shared by all institutions in the list of codes that accounted for 99% of observations in one year. Over the course of a year, across all institutions, we found there were 97 observation codes common to all five institutions.

Figure 2. Number of observation codes that account for 99% of observations in one year that are common to all and unique among institutions.

Table 1: Number of laboratory observation codes that account for a given percentage of observation volume by institution

<table>
<thead>
<tr>
<th>% of Observations</th>
<th>Institution A</th>
<th>Institution B</th>
<th>Institution C</th>
<th>Institution D</th>
<th>Institution E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 mo</td>
<td>1 yr</td>
<td>1 mo</td>
<td>1 yr</td>
<td>1 mo</td>
</tr>
<tr>
<td>80</td>
<td>52</td>
<td>53</td>
<td>51</td>
<td>49</td>
<td>52</td>
</tr>
<tr>
<td>90</td>
<td>84</td>
<td>85</td>
<td>73</td>
<td>72</td>
<td>86</td>
</tr>
<tr>
<td>95</td>
<td>128</td>
<td>133</td>
<td>106</td>
<td>104</td>
<td>129</td>
</tr>
<tr>
<td>98</td>
<td>209</td>
<td>225</td>
<td>176</td>
<td>173</td>
<td>207</td>
</tr>
<tr>
<td>99</td>
<td>298</td>
<td>333</td>
<td>247</td>
<td>244</td>
<td>276</td>
</tr>
</tbody>
</table>

Figure 3 demonstrates the percentage of patients at an institution who would have all of their laboratory results mapped with varying levels of levels of mapping effort. The number of observation codes that account for a given percent of the observational volume for an institution is a proxy for the effort of mapping. Using Table 1, the number of observation codes needed to map 99% of an institution’s volume for one year ranged from 244 to 517. Mapping this small subset of codes would enable more than 99% of patients at any institution to have all of their laboratory results mapped.

Figure 3. Mapping effort required to achieve a specified percent age of patients who have all of their laboratory results mapped. The percent of institutional observation volume represents a subset of observation codes from that institution.
The INPC common dictionary contains several ‘miscellaneous’ test terms. One of these terms was present in the list accounting for 90% of the total volume. This “Miscellaneous Test” term accounted for 0.3% of the total test volume for one year.

**DISCUSSION**

Overall, more than 4,000 different laboratory observation codes accounted for almost 49 million results during thirteen months of information exchange in the INPC, but a relatively small number of observation codes accounted for the vast majority. The modest number of commonly reported observations also represents all of the laboratory results for the majority of patients.

Of the more than 4,000 observation codes reported in the thirteen months, 80 codes (2%) accounted for 80% of the total volume from all institutions and 784 codes (19%) accounted for 99% of the volume from all institutions. Within an institution, the number of observations codes that accounted for 80% of that institution’s volume in one year ranged from 49 to 68, and the number of codes that accounted for 99% of the volume ranged from 244 to 517. These 49 to 69 observation codes represented all of the results for more than 91% of the patients from that institution, and the 244 to 517 codes represented all of the results for more than 99% of patients.

We also found that the lists of frequently occurring observation codes in one month were very similar to those encountered over one year. We observed a relatively small number of additional observation codes transmitted during one year and found a high correlation of the one month set and one year set.

Because of the current state of idiosyncratic local naming conventions for laboratory data, semantic interoperability across sources is only feasible by mapping to a common vocabulary, or *lingua franca*, such as LOINC®. As we and others expand our information exchange efforts towards the goal of an interoperable national health information network, the mapping burden can be a significant impediment.

Given limited mapping resources, our findings support the strategy of focusing the effort on the small subset of observations that account for the majority of volume. Mapping the observation codes that cover 99% of the reported results would ensure that all of the results for more than 99% of patients would be mapped. Mapping even the few (49 to 68) observation codes accounting for 80% of reported results would cover all results for 91.98% of patients.

Our findings also suggest that a one month sample of results from an institution may be a reasonable place from which to identify the high priority observations for mapping when a larger set is not available.

A prioritization-based mapping strategy has the disadvantage of initially leaving some observations unmapped. A pragmatic approach for the unmapped observations might be to assign a temporary miscellaneous code that allows the results to be displayed for clinical care. When interest (e.g. for decision support, public health, bioterrorism surveillance, or research) and resources are available, these observation codes could be remapped to terms with appropriate fidelity. Some unmapped observations will likely be rare elements of commonly reported profiles or panels (e.g. complete blood count). Expanding the prioritized list of to-be-mapped observations to include these rare elements as well as other items of high interest would also be reasonable.

The clear benefit of mapping prioritization is to direct scarce resources towards the most salient observation codes. Even with automated tools, the process of mapping to a standard vocabulary is resource intensive and thus limits progress towards the goal of health information exchange. Prioritizing the mapping effort on the modest subset of most common laboratory codes (that also cover all results for most patients) is a convincing way to lower the barrier to interoperability; it can reduce the effort without sacrificing the rewards. Such an approach may help us move more quickly towards the goal of consumer-centric and information-rich care.

Our study has some important limitations. First, like most manually curated resources, our master concept dictionary and its mappings are not perfect. Our extract may not have captured all laboratory results that would be important to map. Additionally, there is some inherent variability in the mappings from various source systems to our master dictionary, introduced both in the initial system mapping and subsequently as changes in sources terms occur over time.17 The data sources we chose for this study include laboratories in healthcare systems. Some of the most commonly reported results in our samples were from point-of-care testing, which represents a growing proportion of clinical laboratory testing.18 Presently, we only receive these results, however, when they are reported through the hospital laboratory. Differing mechanisms for capturing these results may influence the distribution of observation reporting in other settings. Moreover, the INPC also receives some laboratory results from referral labs.
separately, and the pattern of results reporting and mapping implications for these referral laboratories may differ from hospital-based systems.

CONCLUSION

In order to reap the benefits of a regional health information exchange, data from source systems must be mapped to a common framework. This study demonstrates that approximately 20% of laboratory observations (less than 800 codes) account for 99% of all laboratory test results reported over the course of year from five separate and distinct health care systems. By mapping this small subset of codes, more than 99% of patients would have all of their laboratory results mapped.

A wise man once said: “You can spend the rest of your life getting the last 5% right.”

References


Acknowledgements

The authors thank Lawrence Lemmon for his prowess in data extraction. This work was performed at the Regenstrief Institute, Inc and was supported in part by the National Library of Medicine (N01-LM-3-3501).