Extraction of Methicillin-Resistant *Staphylococcus aureus* Data from the Nation’s Veterans Affairs Medical Centers

Makoto Jones, MD$^{1,2}$§, Scott L. DuVall, PhD$^1$§, Joshua Spuhl$^1$, Matthew H. Samore, MD$^{1,2}$, Christopher Nielson, MD$^3$, and Michael Rubin, MD, PhD$^{1,2}$

$^1$VA Salt Lake City Health Care System, Salt Lake City, UT

$^2$Department of Internal Medicine, University of Utah, Salt Lake City, UT

$^3$VA Reno Medical Center, Reno, NV

§Corresponding author

Email addresses:

MJ: makoto.jones@hsc.utah.edu

SD: scott.duvall@hsc.utah.edu
Abstract

Background
Accurate information is needed to direct healthcare systems’ efforts to control methicillin-resistant *Staphylococcus aureus* (MRSA). Assembling complete and correct microbiology data is vital to understanding and addressing the multiple drug-resistant organisms in our hospitals.

Methods
Herein, we describe a system that securely gathers microbiology data from the Department of Veterans Affairs (VA) network of databases. Using natural language processing methods, we applied an information extraction process to extract organisms and susceptibilities from the free-text data. We then validated the extraction against independently derived electronic data and manual record annotation.

Results
We estimate that the collected microbiology data is 98.5% complete and that methicillin-resistant *Staphylococcus aureus* was extracted accurately 99.7% of the time.

Conclusions
Applying natural language processing methods to microbiology records appears to be a promising way to extract accurate and useful nosocomial pathogen surveillance data. Both scientific inquiry and the data’s reliability will be dependent on the surveillance system’s capability to compare from multiple sources and circumvent systematic error. In time, the dataset constructed for this investigation could serve as a comprehensive infectious disease surveillance system.
Background

There is a pressing need for timely, reliable, and generalizable information to guide the infection control efforts directed against the spread of methicillin-resistant *Staphylococcus aureus* (MRSA) within hospitals. This microorganism frequently causes abscesses, bloodstream infections, post-surgical infections, and sometimes deaths; estimates from existing research and census data suggest that 17,000 attributable deaths occurred in 2008 [1-2]. With the objective of reducing the nosocomial transmission of MRSA, the Department of Veterans Affairs (VA) implemented the National MRSA Prevention Initiative in October 2007[3]. The program included VA-wide MRSA active surveillance for admission to, discharge from, and transfers between acute care wards; rules for contact precautions; hand hygiene; a change in culture to one of shared responsibility; and new reporting systems [3]. The VA Inpatient Evaluation Center (IPEC) gathers data to evaluate this program from chart reviews conducted on all identified MRSA clinical specimens. Assessing the effectiveness of the MRSA Initiative via this method of manual surveillance may be augmented and made more efficient with detailed electronic microbiology data. Such electronic data may also be used for algorithmic surveillance, which has the advantage of not having its own subjective bias to consider.[4]

Microbiology data are increasingly collected electronically and could provide a powerful means of infectious disease surveillance, but the synthesis and utilization of databases across large networks remains a daunting endeavor. Barriers include differing data models [5], messaging strategies, and security issues [6]. The VA medical centers have had an electronic medical record system for over 20 years. This includes, but is not limited to, microbiology data maintained at 152 hospitals currently active worldwide. Unfortunately, the data are not easily accessed or standardized.
Hence, our objective was to evaluate methods that permit extraction and validation of these microbiology data.

The VA stores patient-level microbiology and most other types of data in a hierarchical health information system called the Veterans Health Information Systems and Technology Architecture (VistA). VistA uses a programming language and database called MUMPS (Massachusetts General Hospital Utility Multi-Programming System). Although all VA medical centers use the same software programs, the underlying files have distinct naming conventions and some variation in data structure[7]. There has been some consolidation of VistA instances among medical centers, but most continue to maintain their own VistA system. Because a core system integrating microbiology data across VistA systems was not otherwise available, we utilized a system developed by VA Patient Care Services (PCS).

The PCS system uses a straightforward approach similar to the process that retrieves records during the course of clinical care. Healthcare providers access data using the CPRS (Computerized Patient Record System) graphical user interface. The CPRS interacts with the core MUMPS files through a number of established remote procedure calls (RPCs). The process for using a RPC is identical at all medical centers and it is highly reliable. Because it is necessary for care providers to access records at other VA medical centers, a web based-system, VistaWeb, was developed a number of years ago to access VistA data through RPCs. VistaWeb is in daily use nationwide and it provides a well-established, reliable infrastructure. [8]

The PCS system uses the VistaWeb interface to access RPCs at each medical center and then uploads the data to a Structured Query Language (SQL) relational database in a secure VA data center. The VA login and network security processes are based upon the same approaches required for providers system-wide. Because the
PCS system uses existing, reliable data processes, it requires no independent maintenance and can be run automatically at night.

Local RPCs can have a valuable role in maintaining data validity. The RPCs incorporate the meta-data necessary to make VistA data intelligible. If data structures change within an implementation of VistA, then RPCs are updated so that healthcare workers continue to have correctly displayed patient data. Healthcare workers thus participate in data curation and quality maintenance by reporting errors they see in the patient record. When raw VistA data are pulled into a central database, during a process temporally removed from RPC updates, the data must be carefully evaluated for changes in structure and semantics. This is critical because of the more complex, hierarchical structure of cultures, microorganisms, and susceptibilities in microbiology data. Although direct data retrieval into a warehouse can retain native data structure, differing data models dictate that relationships must then be scrutinized for each medical center. The semantic relationships between fields can change because data can be entered inappropriately. This is because 1) not all of the complex aspects of microbiology reports were known when the data model was developed; 2) microbiology reports from client laboratories may have formats incompatible with the data model; and 3) MUMPS has only one data type, so there are no data type checks. For example, laboratory addresses are routinely stored in antimicrobial susceptibility fields. The methodology we investigated retrieves VA-wide, patient-level data in the same format used by health care providers. The records are in a semi-structured, free-text form that is as "human-readable" as an official microbiology report, regardless of changes in the core VistA file structures. Because these data represent human-readable documents, the contents of each record have face validity. Even though the individual fields have been lost by using this format, the un-extracted record can be
inspected visually against extracted elements. As the data are already assembled and can be updated daily, they represent a valuable resource that needed to be formally validated.

**Methods**

*Description of Information Extraction Process*

Microbiology records from the beginning of 1990 through the end of 2009 were collected from all VistA systems for this task. Before information extraction, we removed MRSA surveillance tests from the microbiology data using a filter on cultured sample and specimen types. To develop this filter, 11,596 unique sample and specimen types were reviewed manually by one of the authors (MJ) and annotated as to whether they were consistent with a MRSA screen from the anterior nares. This removal was performed because MRSA surveillance test report formats are often dramatically different from that of routine culture and susceptibility tests.

While the reports gathered by PCS provided microbiology data from VA sites nationwide, the free-text format of these reports meant that the data could not be used efficiently without further processing. To identify the organisms mentioned in these reports along with their antibiotic susceptibilities, we employed a set of natural language processing (NLP) techniques for information extraction. The extraction creates a formal representation of concepts in the text that can be used in computer algorithms.

The NLP system was developed in the Apache Unstructured Information Management Architecture (UIMA), an open-source framework that provides a consistent data model and interface for handling annotations and meta-data associated with unstructured data such as text. The UIMA supports development of multi-stage applications where individual processes are used together in sequence to achieve a
final result. This pipeline of processes is cumulative, with each step using information added from previous steps to perform more complex tasks. This system was composed of four general tasks: section identification, organism detection, susceptibility detection, and MRSA inference (see Figure 2).

1. Section identification: Although the microbiology reports were free-text, the semi-structured format of the reports allowed for the decomposition of the full document into sections. The sections, constructed by RPCs, contained consistent information across VA sites. Specific rules were crafted to take advantage of these sections and to identify which types of concepts of interest may be found within. Using this approach, the system extracted the templated meta-data contained in each report, including accession number, site and specimen of the sample, and dates and times that certain tests were performed.

2. Organism detection: Lines of text that contained organisms were found using rules that leveraged the structural cues of common formats, such as numbered lists and indented lines in the Culture Results section and in tables in the Antibiotic Susceptibility Test Results section. Once these lines were found, the name of the organism (either a descriptor like rods or bacilli, a genus like *Staphylococcus*, or a full genus and species like *Staphylococcus aureus*) was split from descriptors such as quantity or concentration, descriptions of resistance (MRSA), and results of other identifying tests (gram negative, alpha-hemolytic). Separating these from the organism name allowed for straightforward comparisons and inferences that did not have to rely on lexical variants. The organism names were mapped to SNOMED-CT terms, but as the lookups took significantly longer than all other steps, mapping was done subsequent to note processing.
3. Susceptibility detection: Terms denoting susceptibility were detected and rules were used to determine whether the words surrounding these terms described an antibiotic or some other test for susceptibility. Susceptibility-associated terms occurred along with the organism names in free-text comments in the Culture Results section and in tables in the Antibiotic Susceptibility Results section. Susceptibility terms were mapped to logical values of resistant, susceptible, or indeterminate for computation.

4. MRSA inference: If *Staphylococcus aureus* (*S. aureus*) was one of the organisms detected in the microbiology report, the detected susceptibilities were used to determine whether the culture was positive for methicillin resistance.

Records were processed on the VA Informatics and Computing Infrastructure (VINCI), a high-performance computing environment that provides researchers a secure, central location for data access and application development.

*Reference Comparisons*

We used reference data sets from two sources: routine MUMPS extractions of VistA data into local data warehouses and manual annotation of raw PCS data. The first set was extracted and manipulated by data managers at their native sites as they saw fit and was used as a standard for assessment of completeness and correctness. The second set was only used for an assessment of extraction correctness. We assumed that the manual review of raw PCS data was appropriate when present because a sample of 142 VA-wide records verified that PCS data substantively matched reports accessed through VistAWeb. Having data derived from alternative, independent data extraction methodologies was critical to ensuring a valid estimate of data completeness and the correctness of our information extraction [9]. An ideal reference would be a health information system that provides all patient information
accurately, immediately, and unambiguously [10-11]. However, imperfections arise from the perceptions and opinions of health information recorders, who have logistical and technological limitations. The electronic health record, accessed through CPRS, is the accessible source of information closest to our ideal. Observations recorded through the CPRS system allow for review and correction of ambiguities or inaccuracies. However, CPRS is only accessible manually, so we made compromises with other extractions from VistA. In order to evaluate the PCS microbiology set, we have assessed the completeness, concordance, and correctness of MRSA data.

*Evaluation of completeness*

A complete microbiology data set would be ideal, but it is more important that the data not be subject to selection bias while still providing adequate sensitivity for outbreak detection. Estimates of completeness facilitate assessments of the data along these criteria. We compared PCS data to independently-derived data sets and evaluated completeness by linking microbiology tests in each set and measuring the concordance of their presence. We used the microbiology accession number (a nearly unique identifier for microbiology tests when coupled with the VistA site identifier) and collection time to perform the linkage between the sets. Because acquiring independent data from VA medical centers was difficult, we also estimated the lower bounds of completeness by evaluating meta-data contained in the PCS tables.

We used patient data from the VA Salt Lake City Health Care System (VA SLC HCS) (data from 2005-2008) and the Veterans Integrated Service Network (VISN) 4 (a network of ten VA hospitals, data from 1999-2006) data warehouses for comparisons. VISN 4 microbiology organism and susceptibility had been coded independently by VISN 4. Because data warehouse and PCS data were retrieved by
entirely independent processes and methodologies, concordance provides a reliable indication that these data accurately reproduced the original VistA database.

The other method to estimate completeness relied on the standard practice of sequentially ordering microbiology accession numbers at each VA. Typically, microbiology accession numbers are made up of a sample component (for example “bc” for blood culture), a date component, and a sequential number. We encountered two types of sequential numbering strategies: one incremented accession numbers with any type of culture and the other separated incremental counters for each culture type. When separate incremental counters were used, we used the dominant sample type for our analysis. Since there is variation in sample naming and numbering conventions between medical centers, it was necessary to perform the analysis separately at each site. Gaps in sequence were identified between cultures. When a sequence gap was found, this was taken to be the maximum number of cultures potentially missing from the database. Unfortunately, sequence gaps occur for reasons other than missing data, so this method only produces a lower bound estimate. This analysis was performed for each data warehouse data set, as well as, a sample of ten randomly picked VA hospitals.

Evaluation of the extraction and inferencing of MRSA

The original PCS report text was used as a reference to evaluate the accuracy of the NLP system. Ten thousand randomly selected microbiology documents containing the strings “staph”, “coag”, “mrsa”, “orsa”, “oxacillin”, or “aureus” were queued for manual review. Clinical nurses were trained to use a secure web-based application to annotate the presence of S. aureus, MRSA, and whether the record was an MRSA surveillance test. Reviewers were blind to the NLP results, but were free to consult with each other and with the authors. Synonyms for S. aureus included
accepted abbreviations, MSSA, MRSA, and coagulase-positive S. aureus. Our standard for the definition of methicillin resistance was based on Clinical Laboratory and Standards Institute guidelines. Historically, not all VAs have reported oxacillin, cefoxitin, penicillin binding protein 2A or mecA testing, so resistance to cefazolin, cephalothin, nafcillin, or simply a statement of methicillin resistance were permitted as well. Screening studies from the manually derived set were identified by reviewers, but removed during evaluation of information extraction accuracy. Separate training and test sets were generated from this sample by randomly selecting half to go into each set.

Microbiology data from VISN 4 and the Salt Lake City Health Care System, which had already been annotated, were also used for comparison. Half of the records were randomly chosen to become part of the training set, while the other half became part of the test set. Sensitivity and specificity were calculated separately when PCS data were compared with manual and electronic data sets. Confidence intervals were calculated assuming normal distributions. Because of the enormity of the set and known issues with manual annotation by single reviewers and electronic coding of organisms, we anticipated the need to re-review records that were preliminarily categorized as discordant between extracted records and the reference standard. We also planned to review 200 concordant records (50 at the S. aureus and MRSA levels for each modality) anticipating that true concordance was >=98%.

Results

Description of the data

Using PCS extraction methods to build the data set, microbiology data from January 1, 1990 to December 31, 2009 were available for analysis. The data set
included 33,024,796 unique records from 128 VistA sites representing 152 currently active acute care hospitals and 170 total hospitals during the entire time frame.

Completeness of the data

We estimated PCS data completeness through comparison with DW patient data, which were available from eleven VA sites with average yearly admissions of 2,842 (range from 356 to 7080). Matches on microbiology accession numbers were found to represent 98.5% [95% CI 98.5-98.6%] of the whole. As concordance was high and serves as a lower bound for report completeness, further investigation of the discordant set was not pursued.

By analyzing microbiology accession sequences, we estimated PCS data completeness for the same sample of eleven hospitals. Five hospitals were found to be missing more than five percent of accession numbers, but we found that these sequence gaps were almost entirely attributable to quality control and other non-patient sampling. An additional ten hospitals were randomly selected to examine sequence gaps. Four of ten additional randomly selected hospitals demonstrated missing accession numbers of greater than 5%, but their data warehouse data were not available for comparison. We also attempted to assess temporal gaps and combinations of sequence and temporal gaps to assess completeness. However, because we observed the presence of long gaps, particularly at nights and on weekends, we did not pursue the identification of single or small numbers of missing cultures with this method.

Accuracy of the data

As mentioned, MRSA screens were removed from the microbiology data by means of a string-searching algorithm we developed. This algorithm identified MRSA
surveillance screens among our 10,000 manually annotated charts with a 99.4% sensitivity and 97.9% specificity.

The successful extraction of *S. aureus* and methicillin resistance from PCS data was assessed by comparison to DW data and a manually annotated random sample from the PCS data. Four thousand and two records were identified as screens in the manually reviewed data set; they were removed before the information extraction analysis. We included half of the remaining 5,998 records in our training set from a random nationwide sample that we manually reviewed; 5,967 records were documented for both *S. aureus* status and methicillin resistance from MUMPS extracted data from ten VISN 4 hospitals. In addition, 53,627 microbiology records from VA SLC HCS that documented the presence or absence of MRSA were included. When determining the categorization of *S. aureus* in the training set, the sensitivity was 99.6%, specificity 99.9%, and PPV 99.9%. Of records in which *S. aureus* was correctly identified, we found a sensitivity of 99.8%, a positive predictive value (PPV) of 99.7%, and a specificity of 99.9%, with respect to the correct assignment of methicillin resistance.

We then made comparisons using our validation set of a second set of 5,927 electronic and 3,092 manually derived records. Sensitivity, specificity, and PPV for identification of *S. aureus* were estimated at 100%, 99.9%, and 99.9% compared to electronic data and 98.3%, 99.7%, and 99.6% compared to manually annotated data. Susceptibilities were analyzed in a similar fashion. Among records that successfully identified *S. aureus*, we found that there were 57 discordant records in the electronic set and 53 in the manual set. Discordant records were re-reviewed manually (by MJ) on both the manually and electronically derived test sets. Forty-seven records were resolved in favor of information extraction in the electronic data set, while 18 records
were resolved in favor of information extraction in the manual data set. After this re-
review, there was a concordance of 99.8% with the electronic and 98.9% with the
manual data-sets. Sensitivity, specificity, and PPV were estimated at 100%, 99.9%,
and 99.9% compared to electronic data and 99.2%, 99.4%, and 97.9% compared to
manually annotated data. Results are summarized in Table 1. Of the 200 concordant
records re-reviewed for concordance, all records were found concordant.

Discussion
The methods used to create and extract patient information from each VistA
site’s microbiology files are meant to permit well-powered and generalizable
operational and research studies. The ability to do so depends upon, among other
things, the quality of extracted data. The absence of gold standards to evaluate health
information systems has long been acknowledged [9]. In accordance with established
frameworks to evaluate data quality given these limitations [9-10, 12], we assessed
the PCS database with respect to accuracy.

Our assessment of this large database was limited not only by the absence of a
valid reference standard, but also by limitations on the amount of accessible data
available for comparison. A random sample across more of the nation’s VA data
warehouses would have improved our analysis of completeness, but logistically
would have been difficult. Instead, we did analyze microbiology accession number
sequence gaps at ten VA hospitals and demonstrated that roughly half of the included
VA sites had significant gaps. However, when we examined sites where data
warehouse data were available, we found that in roughly half of the hospitals these
gaps were largely due to environmental samples. We cannot conclude from these data
that we have complete data from the unsampled VA sites; however, our results
suggest that many VA sites are complete and that some, if not most, of the accession
number gaps observed do not represent true data loss. It appears from our analysis that missing data are unlikely to be systematically related to elements analyzed in clinical studies.

The analysis of correctness was assisted by the determination that in a nationwide sample PCS records were exact text copies of the CPRS reports. We reviewed a large sample of microbiology data from all available VA medical centers to capture the variability that may have occurred in the microbiology reports over time and place. Even with 10,000 random samples, our sample size from each of the individual VA sites was not large, particularly from small VA facilities. Thus, we were unable to report VA site-specific estimates of accuracy. The high accuracy scores we have reported are reassuring that the natural language processing methods are robust and will serve well in the future, as well as, against other organisms.

Our data warehouse comparisons highlighted the fact that not all routine data warehouse pulls are validated. As data are consolidated for large studies – particularly those in which individual effects are analyzed – there should be significant concern about systematic errors. The capability of comparing data from multiple sources, as is planned by VINCI, will be critical to ensuring data integrity and solid scientific analysis.

Operationally, this data set could eventually serve as a comprehensive surveillance system for infectious diseases. Its practical utility in this role will ultimately benefit from assessments of other elements, including simplicity, flexibility, acceptability, representativeness, timeliness, and stability in accordance with existing guidelines [11]. We plan to assess these elements as our information extraction becomes integrated with data extraction in order to provide a continuous, coded live-feed.
Conclusions

This study demonstrates that the PCS strategy of data collection coupled with information extraction can deliver surveillance data from over a twenty-year time span in one of the largest healthcare systems in the United States. Furthermore, the PCS system appears to be able to completely and correctly represent nationwide VA microbiology data. The ability of these natural language processing tools to facilitate information extraction is a promising advance, as we expand our capabilities to other important nosocomial pathogens. Sometimes duplicative, but independent data pulls can serve to improve data quality assessment and validation. As hospital systems merge into larger entities, each with their own electronic medical record, it will be important to consider processes, such as this one, to smooth the transition into one integrated medical record system. In the future, better data models and coding may supersede this method of extraction, but data that retain the semantic structure and human-understandability of reports, and have been reviewed by clinicians will likely continue to serve to an important purpose in data validation and quality assessment procedures.
Competing interests

None of the authors have received reimbursements, fees, funding, or salary from an organization that may gain or lose financially from this publication. There are no outstanding patents to which any of the authors have any interest. We have no other competing interests. All of the authors do have affiliations with the Department of Veterans Affairs.

Authors' contributions

MJ was involved with data collection, acted as a domain consultant for organism extraction, developed the analysis, and drafted the manuscript. SLD was involved with developing and performing information extraction, participated in the analysis, drafted the information extraction methods of the manuscript, and helped with the writing of the manuscript in general. JS was involved with data collection, manipulation, analysis, and helped draft the manuscript. MHS was responsible for the study’s conception, participated in the design, and helped draft the manuscript. CN performed data extractions from VistAs, participated in data modeling, drafted the data extraction methods, and helped with the draft of the manuscript in general. MR participated in the study’s conception, design, analysis, and manuscript drafting.

Acknowledgements

This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding support from the VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204 and the Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374. We also gratefully acknowledge VA Patient Care Services for developing the data acquisition mechanism and for providing the data and support for this project.
References


Figures

Figure 1 - Data Network Diagram
RPC - remote procedure call. CPRS - Computerized Patient Record System. RDW-regional data warehouse. CDW-corporate data warehouse. PCS – VA Patient Care Services. NLP- natural language processing. Figure courtesy of Kiyoshi Jones.

Figure 2 - Information Extraction Strategy
VINCI -VA Informatics and Computing Infrastructure. SNOMED – Systematized Nomenclature of Medicine. RxNorm – a standardized nomenclature of clinical drugs developed by the National Library of Medicine.
# Tables

## Table 1 - Information Extraction Accuracy

### Training Set

<table>
<thead>
<tr>
<th>Staphylococcus aureus</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99.62</td>
<td>99.98</td>
<td>99.98</td>
<td>99.63</td>
</tr>
</tbody>
</table>

### Validation Set

#### Electronic Records Reviewed: 5,927

<table>
<thead>
<tr>
<th>Staphylococcus aureus</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>99.91</td>
<td>99.89</td>
<td>100</td>
</tr>
<tr>
<td>Methicillin Resistance</td>
<td>100</td>
<td>99.96</td>
<td>99.86</td>
<td>100</td>
</tr>
</tbody>
</table>

#### Manual Records Reviewed: 3,092

<table>
<thead>
<tr>
<th>Staphylococcus aureus</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>98.25</td>
<td>99.65</td>
<td>99.56</td>
<td>98.62</td>
</tr>
<tr>
<td>Methicillin Resistance</td>
<td>99.15</td>
<td>99.37</td>
<td>97.91</td>
<td>99.75</td>
</tr>
</tbody>
</table>

PPV - positive predictive value, NPV - negative predictive value
Figure 2