JAMIA Open, 5(2), 2022, 1–6
https://doi.org/10.1093/jamiaopen/ooac026
Application Notes

Application Notes

Development and validation of *MicrobEx*: an open-source package for microbiology culture concept extraction

Garrett Eickelberg [1], Yuan Luo [1], and L. Nelson Sanchez-Pinto [1,2]

[1] Department of Preventive Medicine (Health & Biomedical Informatics), Feinberg School of Medicine, Chicago, Illinois, USA, and [2] Department of Pediatrics (Critical Care), Chicago, Illinois, USA

Dr. Yuan Luo and Dr. L. Nelson Sanchez are co-corresponding authors.

Corresponding Author: L. Nelson Sanchez-Pinto, Department of Preventive Medicine (Health & Biomedical Informatics), Feinberg School of Medicine, 750 N Lake Shore, Chicago, IL 60611, USA; Department of Pediatrics (Critical Care), 225 E. Chicago Avenue, Chicago, IL 60611, USA; lazaro.sanchez-pinto@northwestern.edu

Received 14 January 2022; Revised 25 March 2022; Editorial Decision 2 April 2022; Accepted 11 April 2022

ABSTRACT

Objective: Microbiology culture reports contain critical information for important clinical and public health applications. However, microbiology reports often have complex, semistructured, free-text data that present a barrier for secondary use. Here we present the development and validation of an open-source package designed to ingest free-text microbiology reports, determine whether the culture is positive, and return a list of Systemized Nomenclature of Medicine (SNOMED)-CT mapped bacteria.

Materials and Methods: Our concept extraction Python package, *MicrobEx*, is built upon a rule-based natural language processing algorithm and was developed using microbiology reports from 2 different electronic health record systems in a large healthcare organization, and then externally validated on the reports of 2 other institutions with manually reviewed results as a benchmark.

Results: *MicrobEx* achieved F1 scores >0.95 on all classification tasks across 2 independent validation sets with minimal customization. Additionally, *MicrobEx* matched or surpassed our MetaMap-based benchmark algorithm performance across positive culture classification and species capture classification tasks.

Discussion: Our results suggest that *MicrobEx* can be used to reliably estimate binary bacterial culture status, extract bacterial species, and map these to SNOMED organism observations when applied to semistructured, free-text microbiology reports from different institutions with relatively low customization.

Conclusion: *MicrobEx* offers an open-source software solution (available on both GitHub and PyPI) for bacterial culture status estimation and bacterial species extraction from free-text microbiology reports. The package was designed to be reused and adapted to individual institutions as an upstream process for other clinical applications such as: machine learning, clinical decision support, and disease surveillance systems.

Key words: concept extraction, information extraction, electronic health records, natural language processing, microbiology report

LAY SUMMARY

Microbiology culture reports are a type of medical laboratory report created by laboratory specialists to summarize their findings after detecting and characterizing bacteria and other organisms present in a patient sample (like blood, urine, etc). The data contained within large collections of microbiology reports can be helpful for numerous clinical and public health applications. However, extracting this relevant information from large collections can be time consuming and challenging as these reports are stored as text, and both the language and format of these reports vary widely across different report writers and clinical settings. This research sought to develop an open-source software tool to enable users to extract relevant information from microbiology reports automatically. Our software tool, *MicrobEx*, uses a variety of rule-based logic and text pattern collections to ultimately classify whether a bacterial infection is described in the text report (yes/no) and to catalogue all relevant bacteria mentioned in the report. *MicrobEx* was developed against data collated from Northwestern Medicine and subsequently validated against reports from 2 distinct institutions that had been manually reviewed by an expert. Overall, our results suggest *MicrobEx* can achieve improved performance over other methods and comparable performance to manual chart review.

INTRODUCTION

Microbiology culture reports are relied upon for myriad healthcare applications ranging from guiding clinical treatment decisions to global disease surveillance. In a clinical setting, microbiology culture reports are helpful in answering if an infection is present and what organisms are driving that infection.[1] Outside of the clinical setting, microbiology data are used to monitor disease outbreaks, improve healthcare operations (eg, monitor nosocomial infection rates), and are leveraged in a variety of observational studies.[2–5] Thus, the data within microbiology reports impact clinical treatment and public policy decisions and are therefore critical for secondary use.[2,6]

Unlike many other structured laboratory test results, microbiology culture reports are often complex, semistructured reports that pose unique challenges for large-scale secondary use applications. Samples sent to a microbiology laboratory routinely undergo numerous tests, such as gram stains and antibiotic susceptibility tests, each of which have different turnaround times, can produce more than a single result, and need to be linked to the original accession number.[1,2] Additionally, results from each test can include both quantitative and qualitative data, and need to be reported as they become available to facilitate treatment decisions.[1,2] Unfortunately, although there are efforts to standardize reporting and analysis of clinical microbiology data, the suitability of existing microbiology reports for secondary use are hindered by reporting variability and analysis practices.[7–9] Finally, microbiology reports contain varying amounts of protected health information as defined by the Health Insurance Portability and Accountability Act, thus limiting the flexibility of this data for data sharing projects.

OBJECTIVE

There is critical need for informatic tools that can navigate microbiology report data challenges and extract information to facilitate their secondary use. The goal of this study was to develop, validate, and release an open-source microbiology concept extraction (*MicrobEx*) system to facilitate secondary use of microbiology reports.

MATERIALS AND METHODS

Datasets

The 2 derivation datasets for this study were extracted from 2 source systems (Epic Systems Corporation and Cerner Corporation) within the Northwestern Medicine (NM) Enterprise Data Warehouse (EDW). Data from source systems 1 and 2 were extracted into separate respective derivation sets in order to reflect different world conditions and preserve their unique microbiology report structures and language characteristics. The regular expressions and logic flow of our extraction system were developed using 216 372 raw free-text microbiology reports extracted from critical care patients treated at 1 of 10 Northwestern Medicine intensive care units between January 1, 2010 and January 1, 2020. To define microbiology reports, we queried the NMEDW and manually curated 235 unique procedures associated with microbiology culture orders. The collection of microbiology reports had highly heterogeneous formatting and lacked consistent template features such as concept-value pairs and table structures. Additionally, our corpora contained full microbiology reports, as well as individual microbiology components such as gram stains and antibiotic susceptibility reports. To address these challenges, rules were crafted to separate reports into sections wherever possible. For cultures with multiple report entries tied to the same accession number, only the notes with the latest report update time were selected for downstream processing and analysis. Testing and validation of our extraction system was performed on 2 external datasets with 65 448 expertly annotated free-text microbiology reports from University of Chicago (validation 1) and Ann & Robert H. Lurie Children's Hospital (validation 2). The validation sets of microbiological culture results were part of prior study and details have been previously published.[6] The reports from both hospitals were annotated by the same senior clinical research coordinator. All 4 datasets included microbiologic cultures reports from blood, urine, respiratory, and cerebral spine fluid samples.

Algorithm overview

A summary of our algorithm workflow is presented in Figure 1. Our concept extraction algorithm uses a comprehensive set of rules, as well as context, keyword, and morphologic features that capture overall bacterial infection status and identify bacterial species present in a microbiology report. Rulesets and regular expressions were developed through an iterative process based on document structural and context features in addition to clinical criteria and domain knowledge. For bacterial species captures, we wrote regular expressions to capture the genus and species for bacteria present in a dictionary of clinically relevant organisms collated from knowledgebases.[1,10] Organisms captured were mapped to Observational Health Data Sciences and Informatics (OHDSI) and Systemized Nomenclature of Medicine (SNOMED) IDs via a dictionary

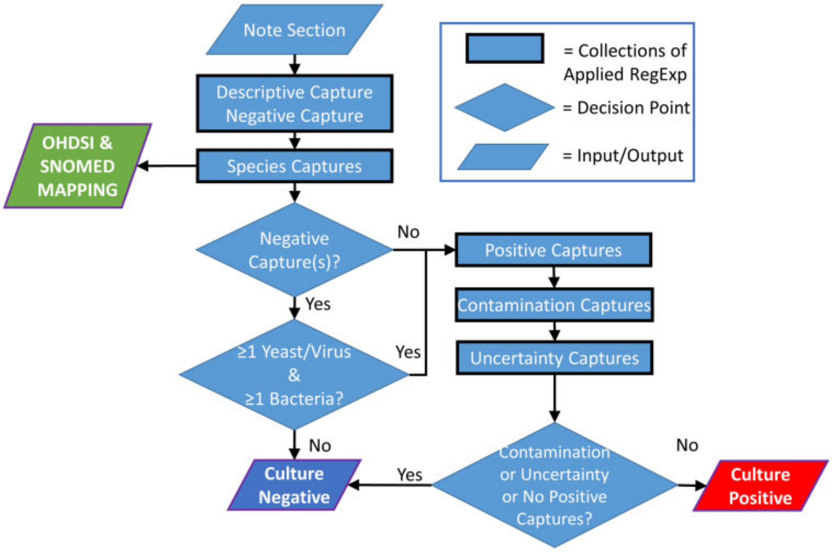


Figure 1. The MicrobEx algorithm structure. The input of our algorithm is a whole or parsed section of a free-text microbiology culture reports. Within the algorithm, a series of regular expression collections are applied to the text input and the captures are associated with bacterial absence (negative), bacterial presence (positive), microbiological species, potential bacterial contamination, and uncertainty. Bacterial species captured are subsequently mapped to both OHDSI and SNOMED concept IDs. Hierarchical decisions are applied to the regular expression collection captures to categorize the culture as positive or negative for bacteria.

included in the source code. The mapping dictionary for microorganism to OHDSI and SNOMED IDs was constructed by passing the collated microorganism list into Usagi software indexed on SNOMED vocabulary and restricted to class “ORGANISM” and domain “OBSERVATION”.[11] During each iteration, concept extraction performance was reviewed manually using a variety of different pattern occurrence-based audits on our training data sets. Customized regular expressions were created to capture remaining complex patterns. Each regular expression was developed with generalizability in mind to maximize dissemination and reusability. For all false positive and negative cases, we reviewed the associated case context, assigned a reason for misclassification. We addressed the cases by either refining existing rules or implementing new ones. This iteration process was repeated until all remaining uncaptured cases were caused by report noise, uncommon misspellings, or lack of report clarity.[12]

Validation

Figure 2 includes example reports annotated with extracted concepts, species, and estimated bacterial culture positive status. Both species extraction and binary bacteria positive culture status (yes/no) were evaluated as outcomes for validation of our algorithm and compared to the manually annotated results in the validation sets. For species extraction, we compared species captured across all report sections by our algorithm and the expert annotation. We encoded our species extraction binary outcome as positive only if *MicrobEx* captured every species identified by the expert. Cases where *MicrobEx* captured bacterial species not identified by the expert were manually reviewed and coded on a case-by-case basis. For positive culture status, MicrobEx assigned a binary classification to all report sections. A report was classified as culture status positive if any of the corresponding report sections were assigned a positive classification. The resulting report level classifications were then compared to expert annotation. Supplementary Table S1 presents a report-level bacteria culture status classification example.

Performance benchmark

In order to benchmark our algorithm’s performance against a well-established clinical natural language processing (NLP) tool, we applied MetaMap[13] to both validation sets and built a rule-based decision workflow to predict positive bacterial culture status and capture bacterial species.

Dataset customization

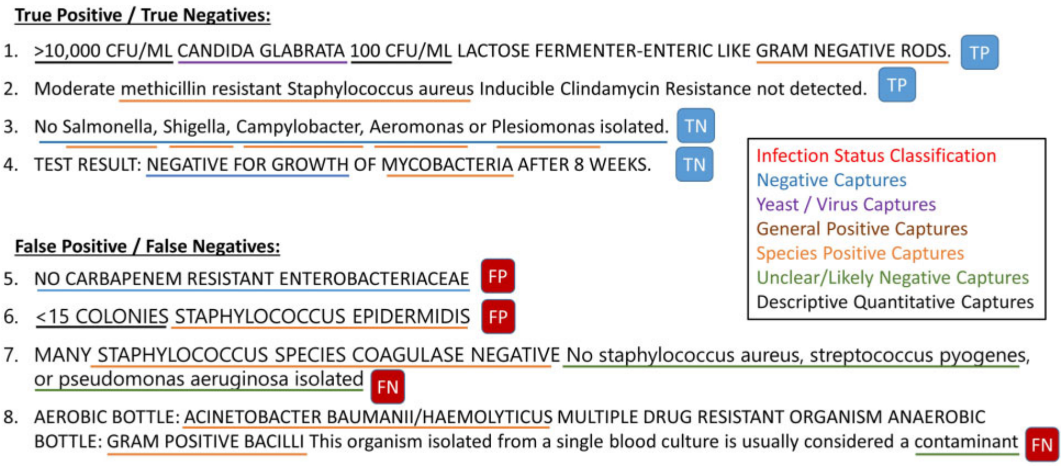
The detailed code, documentation, and Python package installation instructions have been made available at: https://github.com/geickelb/microbex. To identify and address dataset-specific patterns capable of causing misclassifications, we audited our workflow as described in the github documentation prior to final validation. See the *AlgorithmDetails* section for Regular expression examples and the *audit_example* section for a description and example on how to deploy and customize our package to a new dataset. Algorithm details are additionally presented in Supplementary Section C.

RESULTS

Validation

Table 1 summarizes the distribution of positive bacterial culture status in the 4 datasets. The ratio of positive to negative cases across our training set predictions is consistent with that seen in the 2 curated validation sets.

Table 2 summarizes the validation results across both species and positive bacterial culture status classification tasks. The algorithm had excellent and consistent performance, with validation sets 1 and 2 having F1 scores of 0.99 and 0.96 for positive culture classification and species capture, respectively. To estimate the improvements made by introducing customized regular expressions from the data audits, each validation set was reanalyzed using a codebase with the associated regular expressions deactivated. From this, we estimate that culture positivity classification increased from 0.93 to 0.96 and 0.69 to 0.96 for validation sets 1 and 2, respectively. The



True Positive / True Negatives:

1. >10,000 CFU/ML CANDIDA GLABRATA 100 CFU/ML LACTOSE FERMENTER-ENTERIC LIKE GRAM NEGATIVE RODS. TP
2. Moderate methicillin resistant Staphylococcus aureus Inducible Clindamycin Resistance not detected. TP
3. No Salmonella, Shigella, Campylobacter, Aeromonas or Plesiomonas isolated. TN
4. TEST RESULT: NEGATIVE FOR GROWTH OF MYCOBACTERIA AFTER 8 WEEKS. TN

False Positive / False Negatives:

5. NO CARBAPENEM RESISTANT ENTEROBACTERIACEAE FP
6. <15 COLONIES STAPHYLOCOCCUS EPIDERMIDIS FP
7. MANY STAPHYLOCOCCUS SPECIES COAGULASE NEGATIVE No staphylococcus aureus, streptococcus pyogenes, or pseudomonas aeruginosa isolated FN
8. AEROBIC BOTTLE: ACINETOBACTER BAUMANII/HAEMOLYTICUS MULTIPLE DRUG RESISTANT ORGANISM ANAEROBIC BOTTLE: GRAM POSITIVE BACILLI This organism isolated from a single blood culture is usually considered a contaminant FN

Infection Status Classification
Negative Captures
Yeast / Virus Captures
General Positive Captures
Species Positive Captures
Unclear/Likely Negative Captures
Descriptive Quantitative Captures

Figure 2. Examples of annotated reports for validation and error analysis on validation set reports. Colored underlines correspond to parts of the report captured by the associated regular expression collection. For bacterial culture positive status classification, the concepts captured in each block are considered in a hierarchical decision structure according to Figure 1. Examples 1 to 4 demonstrate algorithm annotation on cases found to be correctly classified as positive and negative. Examples 5 to 8 depict 4 examples representative of common misclassifications. FN: false negative; FP: false positive; TN: true negative; TP: true positive.

Table 1. Bacterial culture positive status distribution

	Positive bacterial culture	Negative bacterial culture
Derivation set 1	14 376 (20.7%)	55 065 (79.3%)
Derivation set 2	23 549 (16%)	123 382 (84%)
Validation set 1	2184 (14.5%)	12 916 (85.5%)
Validation set 2	7391 (14.7%)	42 957 (85.3%)

addition of customized regular expressions was found to cause little-to-no effect on species capturing across both validation sets.

Supplementary Table S2 presents the results from our customized MetaMap based benchmarking algorithm against both validation sets. Across both positive culture classification and species capture, *MicrobEx* matched or surpassed the benchmark algorithm performance. These results suggest that our task-specific classifier can outperform more general-use clinical NLP tools like MetaMap. Supplementary Section D presents our *MicrobEx* "Run Report" for validation sets 1 and 2, detailing report- and report section-level data regarding regular expression captures, binary classification decision data, and descriptive statistics.

Error analysis

In the error analysis we identified a collection of 5 patterns in which our concept extraction workflow had the majority of errors. Figure 2 presents annotated visual examples of the classification hierarchical logic for the different patterns observed, with examples for both correct classifications as well as misclassifications. Examples 5 and 6 depict the 2 most common types of false positive patterns and examples 7 and 8 present the most common patterns found in false negatives in the validation sets. We can summarize these patterns as a combination of multiple positive and negative organisms where the negative regex capture supersedes the positive captures, and the use of the term "contaminant" leading to a false negative classification.

DISCUSSION

In this study, we developed and validated an open-source, rule-based framework to extract and map clinical concepts from microbiology reports to standardized terminologies to facilitate secondary use of microbiology reports. Our main finding is that our algorithm can reliably estimate binary bacterial culture status, extract bacterial species, and map these to SNOMED organism observations when applied to semistructured, free-text microbiology reports from different institutions with relatively low customization.

Top performing rule-based concept extraction applications commonly employ a well-established clinical NLP tool that can map mentions to a corresponding medical concept(s) for broad medical corpora, such as cTAKES[14] and MetaMap.[13] Like the well-established tools, *MicrobEx* performs concept matching by leveraging existing microbiology knowledgebases as described in Materials and Methods section. In contrast to these tools however, *MicrobEx* uses custom rules and regular expressions tailored to microbiology reports for negation detection. *MicrobEx*'s higher performance on bacterial positive culture status prediction suggests that for this classification task, *MicrobEx*'s more tailored approach provides advantages over an out-of-the-box approach using a well-established NLP tool. To illustrate, validation set 1 had a language pattern ($n = 88$ report-level occurrences) where the results from the antibiotic susceptibility report were mentioned alongside the microorganism summary (eg, "Many methicillin resistant Staphylococcus aureus Inducible Clindamycin Resistance not detected"). Our MetaMap benchmarking algorithm, which used a Negex negation detection engine, classified culture status negative while MicrobEx correctly classified culture status positive for such cases. By including specific regular expressions to distinguish between susceptibility and resistance detection from microorganism detection (eg, "*(?<!resistance)(?<!susceptibility)\s+not\sdetected|indicated*"), MicrobEx was able to correctly classify binary bacteria culture status. To further improve *MicrobEx*'s prediction performance, additional institution-specific customized rules could be added. Figure 2 depicts 4 representative examples of cases misclassified for positive culture status that could be addressed with institution-specific custom rules.

To our best knowledge, 3 previously published studies have applied clinical concept extraction methods to microbiology notes.[15–17] Jones et al[16] applied a set of crafted rules to blood culture reports from the Salt Lake City Healthcare system to extract organism information, antibiotic susceptibilities, and infer if methicillin-resistant *Staphylococcus aureus* (MRSA) was present. An evaluation was performed against approximately 10 000 expertly annotated reports to

Table 2. Infection classification and species capture performance across validation sets

	True negative	False positive	False negative	True positive	Precision	Recall	NPV	F1
Validation set 1								
Species capture	12 463 (82.54%)	2 (0.01%)	209 (1.38%)	2426 (16.07%)	0.998	0.921	0.984	0.958
Positive culture status	12 909 (85.48%)	7 (0.05%)	22 (0.15%)	2162 (14.32%)	0.995	0.990	0.998	0.992
Validation set 2								
Species capture	42 391 (84.20%)	4 (0.01%)	68 (0.14%)	7885 (15.66%)	0.999	0.991	0.999	0.995
Positive culture status	42 950 (85.31%)	7 (0.01%)	606 (1.20%)	6785 (13.48%)	0.998	0.918	0.986	0.956

NPV: negative predictive rate.

measure successful identification of MRSA. Yim et al[15] and Matheny et al[17] used hybrid and rule-based systems to capture combinations of microorganisms species and antibiotic susceptibilities from blood and multiple sample types, respectively. Our algorithm is notably different from the previously published systems in the following ways: (1) we estimate positive bacterial culture status, (2) our algorithm was designed to work with a variety of disparate microbiology report formats from different institutions, (3) we performed external validation on 2 expertly annotated microbiology datasets, and (4) our software is entirely open-source and available as a python package that can be further adapted to the reports of other institutions as described in our github documentation and supported by our results.

We recognize several limitations of our study. First, for users of this software, classifying positive culture status is the prediction task with the largest potential error. Compared to species extraction, which is largely string matching, estimating infection status requires significantly more complex logic. The hierarchical logic involved with positive bacterial culture status estimation is potentially susceptible to syntactic heterogeneity and report complexity, as depicted in Figure 2. Additionally, we focused on bacterial cultures for the development and validation of the algorithm given the importance of antibiotic stewardship, antibiotic resistance, and bacterial sepsis in hospitalized patients. While our algorithm captures other microorganism species (including fungal and viral species), we did not validate the performance on those. Finally, we included logic to extract relevant quantitative and semiquantitative concepts, however the performance of this was variable due to syntactic heterogeneity. As a result, we continue to provide quantitative captures as a feature of the *MicrobEx* algorithm, however these were not included in our validation.

CONCLUSION

In this article we detail the development, validation, and use of our open-source microbiology concept extractor (*MicrobEx*) algorithm and package. Our workflow achieved excellent performance in 2 independent validation sets with minimal customization, improved performance versus a well-established alternative, and comparable performance to manual chart review by an expert. Our concept extraction Python package is designed to be reused and adapted to individual institutions as an upstream process for other clinical applications such as machine learning, clinical decision support, and disease surveillance systems.

FUNDING

This research is partly supported by grants U01TR003528 & R01LM013337 from the National Institutes of Health (Luo), grant 5T32LM01220304 from the National Library of Medicine (Eickelberg), and grant R01HD105939 from the National Institute of Child Health & Human Development (Sanchez-Pinto).

AUTHOR CONTRIBUTIONS

All authors contributed to study design and funding acquisition. GE performed all software development, validation, data analysis, data visualization, and manuscript draft writing. LNSP and YL are co-corresponding authors and were involved in data curation, project administration, and manuscript edits.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

CONFLICT OF INTERESTS STATEMENT

None declared.

DATA AVAILABILITY

The derivation and validation data underlying this article cannot be shared publicly due to patient privacy and confidentiality concerns given the personal health information contained in the data. The data will be shared on reasonable request to the corresponding author.

REFERENCES

1. Miller JM, Binnicker MJ, Campbell S, *et al.* A guide to utilization of the microbiology laboratory for diagnosis of infectious diseases: 2018 update by The Infectious Diseases Society of America and the American Society for Microbiology. *Clin Infect Dis* 2018; 67 (6): e1–94.
2. Rhoads DD, Sintchenko V, Rauch CA, Pantanowitz L. Clinical microbiology informatics. *Clin Microbiol Rev* 2014; 27 (4): 1025–47.
3. Graham PL, San Gabriel P, Lutwick S, Haas J. Saiman L. Validation of a multicenter computer-based surveillance system for hospital-acquired bloodstream infections in neonatal intensive care departments. *Am J Infect Control* 2004; 32 (4): 232–4.
4. Bellini C, Petignat C, Francioli P, *et al.* Comparison of automated strategies for surveillance of nosocomial bacteremia. *Infect Control Hosp Epidemiol* 2007; 28 (9): 1030–5.
5. Eickelberg G, Sanchez-Pinto LN, Luo Y. Predictive modeling of bacterial infections and antibiotic therapy needs in critically ill adults. *J Biomed Inform* 2020; 109: 103540.
6. Sanchez-Pinto LN, Stroup EK, Pendergrast T, Pinto N, Luo Y. Derivation and validation of novel phenotypes of multiple organ dysfunction syndrome in critically ill children. *JAMA Netw Open* 2020; 3 (8): e209271.
7. Vuokko R, Makela-Bengs P, Hypponen H, Lindqvist M, Doupi P. Impacts of structuring the electronic health record: results of a systematic literature

review from the perspective of secondary use of patient data. *Int J Med Inform* 2017; 97: 293–303.

8. Turner P, Fox-Lewis A, Shrestha P, *et al.* Microbiology investigation criteria for reporting objectively (micro): a framework for the reporting and interpretation of clinical microbiology data. *BMC Med* 2019; 17 (1): 70.
9. Chaitram JM, Jevitt LA, Lary S, Tenover FC; WHO Antimicrobial Resistancce Group. The world health organization's external quality assurance system proficiency testing program has improved the accuracy of antimicrobial susceptibility testing and reporting among participating laboratories using NCCLS methods. *J Clin Microbiol* 2003; 41 (6): 2372–7.
10. c. Wikipedia. List of clinically important bacteria. https://en.wikipedia.org/wiki/List_of_clinically_important_bacteria. Accessed August 12, 2021 05:32 UTC. Revision history statistics.
11. Moinat M, Schuemie M, Rijnbeek P. Usagi. https://github.com/OHDSI/Usagi. Accessed August 1, 2021.
12. Fu S, Chen D, He H, *et al.* Clinical concept extraction: a methodology review. *J Biomed Inform* 2020; 109: 103526.
13. Aronson AR, Lang FM. An overview of metamap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010; 17 (3): 229–36.
14. Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical text analysis and knowledge extraction system (CTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
15. Yim WW, Evans HL, Yetisgen M. Structuring free-text microbiology culture reports for secondary use. *AMIA Jt Summits Transl Sci Proc* 2015; 2015: 471–5.
16. Jones M, DuVall SL, Spuhl J, Samore MH, Nielson C, Rubin M. Identification of methicillin-resistant *Staphylococcus aureus* within the nation's veterans affairs medical centers using natural language processing. *BMC Med Inform Decis Mak* 2012; 12 (1): 34.
17. Matheny ME, Fitzhenry F, Speroff T, *et al.* Detection of blood culture bacterial contamination using natural language processing. *AMIA Annu Symp Proc* 2009; 2009: 411–5.