# Parsing complex microbiology data for secondary use

**Protiva Rahman, BS[1], Albert M. Lai, PhD[1,2], Courtney L. Hebert, MD, MS[2,3]**
**[1]Department of Computer Science and Engineering, [2]Department of Biomedical Informatics, [3]Department of Internal Medicine, The Ohio State University, Columbus, OH**

## Introduction

Clinical microbiology culture and antibiotic susceptibility data are a rich source of information required for a variety of secondary use applications such as epidemiologic studies of antimicrobial resistance. Our study team is working on using clinical microbiology data to develop antibiotic prescribing clinical decision support.[1] Unfortunately, culture and antibiotic susceptibility results are often reported in complex, semi structured or free text form, making it difficult to use them for analysis. Cultures from abdominal-biliary (AB) sources are especially complex because there are multiple types of cultures results (e.g. anaerobic, fungal) and a wide variety of organisms. Prior work in this area has focused on rule-based approaches, natural language processing, and often on blood cultures.[2,3] The objective of this project was to parse and annotate each unique organism name from a free text field of AB cultures and identify any antibiotic susceptibility results or evidence of resistance mechanisms (e.g. presence of penicillinase production) that occur in this free text field. By annotating the organisms with a unique concept identifier we are able to add additional, useful information to the dataset, such as Gram staining, type of growth patterns, and family.

## Methods

Adult patients admitted to an inpatient unit with a diagnosis of AB infection, with a positive culture from a relevant site in the first 4 days of their hospitalization to The Ohio State Wexner Medical Center from 1/1/2009 to 1/1/2014 were included for a total of 625 patients and 1023 unique cultures. The UMLS metathesaurus[4] was used to identify organism name. Our method included: 1) Tokenizing the result text and removing. 2) Using regular expressions to extract colony count and other culture data (e.g., resistance mechanisms). 3) If a token did not match any predefined regular expression, we checked to see if it matched the concept type of 'bacteria, virus or yeast' in the UMLS. If it did we continued consuming tokens to find the longest matching string that matched the semantic type identifier for a pathogen. 4) In the case where an organism was not fully speciated (e.g. Escherchia/Citrobacter), we assigned the code for the common parent. 5) If the token matched the UMLS semantic type of "Antibiotic," we look at the surrounding tokens with a span of two to see if they match a set of positive words: [e.g., confirmed, produces, covers] or a set of negative words: [e.g., negative, resistant] and accordingly assigned true or false. We developed our algorithm on 300 cultures and tested it on the remaining 723. Our subject matter expert (CH) manually validated a set of 200 parsed results, for which we report the precision (true positive/(true positive + false positive)), recall (true positive/(true positive+false positive)) and F1 score (2*precision*recall / (precision + recall)).

## Results

For identifying the correct organism precision was 0.91, recall 0.95, and F1 0.93. Penicillinase production was identified and related to the correct organism with a precision of 1, recall 0.95, and F1 0.97. Free text antibiotic susceptibilities were identified and related to the correct organism with a precision of 0.92, recall 0.93, and F1 of 0.93. The major challenges we faced were in dealing with not fully speciated organisms and relating parsed data to the correct organism in a poly-microbial infection.

## Conclusion

We achieved high F1 scores using our technique of longest matching and regular expressions customized to our data. While some techniques are specific to our data, the ideas can be replicated for other free text microbiology data. This kind of work is vital in order to use these complex data for secondary use applications.

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the NIH under Award Number R01AI116975

## References

1. Hebert C, et al. Infect Control Hosp Epidemiol. 2012 Apr;33(4):381-8
2. Matheny et al. AMIA Annu Symp Proc. 2009 Nov 14;2009:411-5.
3. Yim et al. AMIA Jt Summits Transl Sci Proc. 2015 Mar 25;2015:471-5. eCollection 2015.
4. http://www.ncbi.nlm.nih.gov/books/NBK9684/