Using Natural Language Processing to Screen and Classify Pathology Reports

C. Moody, B.A., CTR; M. Scocozza, CTR; M. Brant, B.A., CTR; C. Remen; A. Sutliff, B.S.; Y. Chen, B.S.; T. Davison, MCP, B.Sc.

California Cancer Reporting and Epidemiologic Surveillance (CalCARES) Program*

Institute for Population Health Improvement, UC Davis Health System

Background

Annually, over 200,000 pathology reports are received at the California Cancer Registry (CCR) electronically via HL-7 messages which are referred to as "ePath" reports. These narrative pathology reports require regional staff to manually read each pathology report to determine report usability and then classify each reportable tumor on the CCR's data system in terms of histology, site, laterality, date of diagnosis and behavior. Estimating approximately 2 minutes to read each report equates to 6,666 hours to screen narrative ePath reports. In addition, estimating approximately 1 minute per reportable pathology report to classify information into the CCR database adds another 1,500 hours to the manual work effort for a total of 8,166 hours. The CCR entered into a contractual agreement with a natural language processing (NLP) analytics company, Health Language Analytics Global (HLA-G) for a pilot project to develop a solution to auto-receive narrative ePath reports from the CCR, apply their natural language processing solutions and auto-screen and classify the reports per California standards.

Pilot Project Process

Our pilot project consisted of providing Health Language Analytics Global LLC (HLA-G) with 10,000 representative pathology reports (5,000 reportable/5,000 non-reportable) that they used to develop an algorithm to determine reportability. The NLP vendor would also classify the following variables in ICD-0-3 format: Site, Histology, Behavior, Grade, Laterality, and Date of Diagnosis. They built a language model from a 'training set' of documents which are annotated manually. The language model was assembled using NLP analysis of text and annotated content was analyzed by a machine-learning algorithm. The goal was for HLA-G to reach a 90% accuracy rate for screening as well as classifying pathology reports. If that percentage were achieved, our pilot project would be implemented as an ongoing solution to screening and classifying pathology reports.

Project Scope

In Scope:

- ◊ Histopathological pathology reports (tissue)
- ◊ Cytopathological pathology reports (cells)
- ◊ Reportable cancer reports to extract the data items (listed to the right).

Data items in scope:

LATERALITY
SITE
GRADE
HISTOLOGY
BEHAVIOR
DATE OF DIAGNOSIS

Reports out of scope:

HISTORY ONLY
SCANNED IMAGES
CAP ECC
GENOMIC
IMMUNOHISTOCHEMICAL

Out of Scope:

- ◊ Immunohistochemical only reports
- ◊ Scanned/images of pathology reports
- ◊ CAP eCC pathology reports
- ◊ Genomic path reports
- ◊ History only reports

Resources provided to HLA-G

The following coding and staging resources were provided to HLA-G:

- ◊ Multiple Primary and Histology Coding Manual
- ◊ Hematopoietic and Lymphoid Neoplasm Database and Coding Manual
- ◊ ICD-O-3 Manual
- ◊ NAACCR Volume II, Appendix G Abbreviations
- ◊ California Volume 1 Pages:
  - Reportability Guide
  - Laterality Coding Instructions
  - Laterality - Paired Sites
  - Terms Indicating In Situ
  - Primary Site - Site specific Special Conditions
  - Grade Rules
  - Ambiguous Diagnostic Reportable Terms
- ◊ Breast Clock Positions
- ◊ Colon CM Measurements
- ◊ SEER Inquiry Database

Methodology

Pathology Report Selection Criteria

CCR Pathology reports were selected from previously manually screened and classified reports. The reports were selected by using simple random sampling based on the distributions of different criteria. The criteria are either single variable or combination of variables. The variables used were site codes (first 3 characters), histology, reporting source, and pathology type (Histopathological or Cytopathological). The combination of variables includes site code and histology, site code and reporting source, pathology type and reporting source, etc. Selected reports were from experienced coders. The number of reportable cases selected was 5,000, out of total 223,175, and the number of non-reportable cases selected was 5,000, out of total 103,750. Each set of 5,000 was sent in 500 report increments as long as the overall selection represented the distribution in the population files.

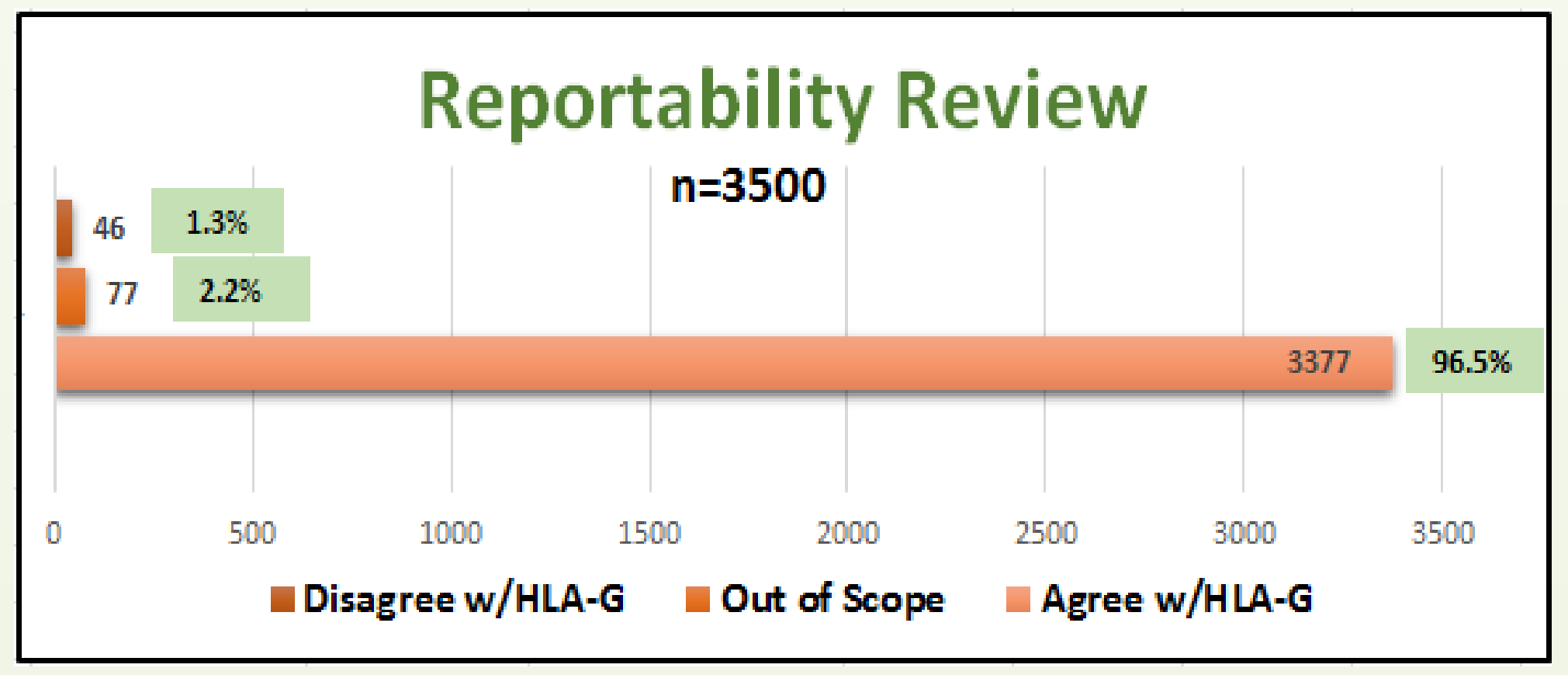
For the purposes of this project, the following were considered "Out Of Scope".

OUT OF SCOPE	
CCR - Designation	Designation-Definition
OOS-HX	History of
OOS-IHC	Immunohistochemistry ONLY
OOS-TX/NED	Treatment with negative Findings
OOS-SPO	Slide Prep Only – No path report
OOS-AMD	Amended Report - Reportable case already received
OOS-PORT	Port-A-Cath Placement
OOS-GEN	Genetic Test

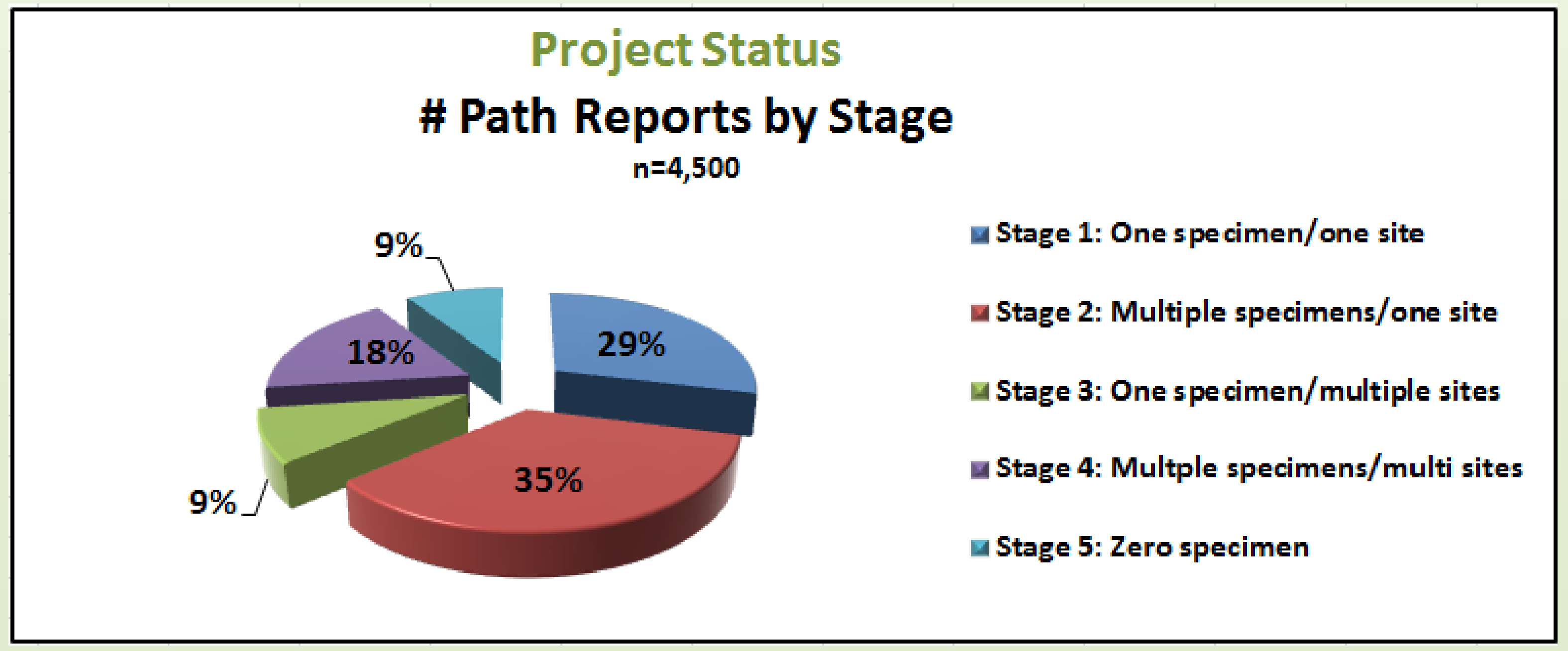
Analysis

One of the preliminary functions performed by CCR staff (CTR Business Analysts) was to define unknown terms determined by the HLA-G algorithm. A list of 4,194 terms and abbreviations that HLA-G were unable to recognize were provided to the CCR for resolution. The CCR staff were able to resolve 1,488 as well as determine that many of the remaining were physician names, initials, acronyms, or other terms not used in classifying the data elements.

Our analysis then turned to confirming reportability. The CCR sent a batch of 3500 pathology reports to HLA-G for review. They returned a spreadsheet identifying the discrepancies between HLA-G's results and original coder's results. The discrepancies were reviewed by internal CCR staff with the following results as shown in the graph on the right.

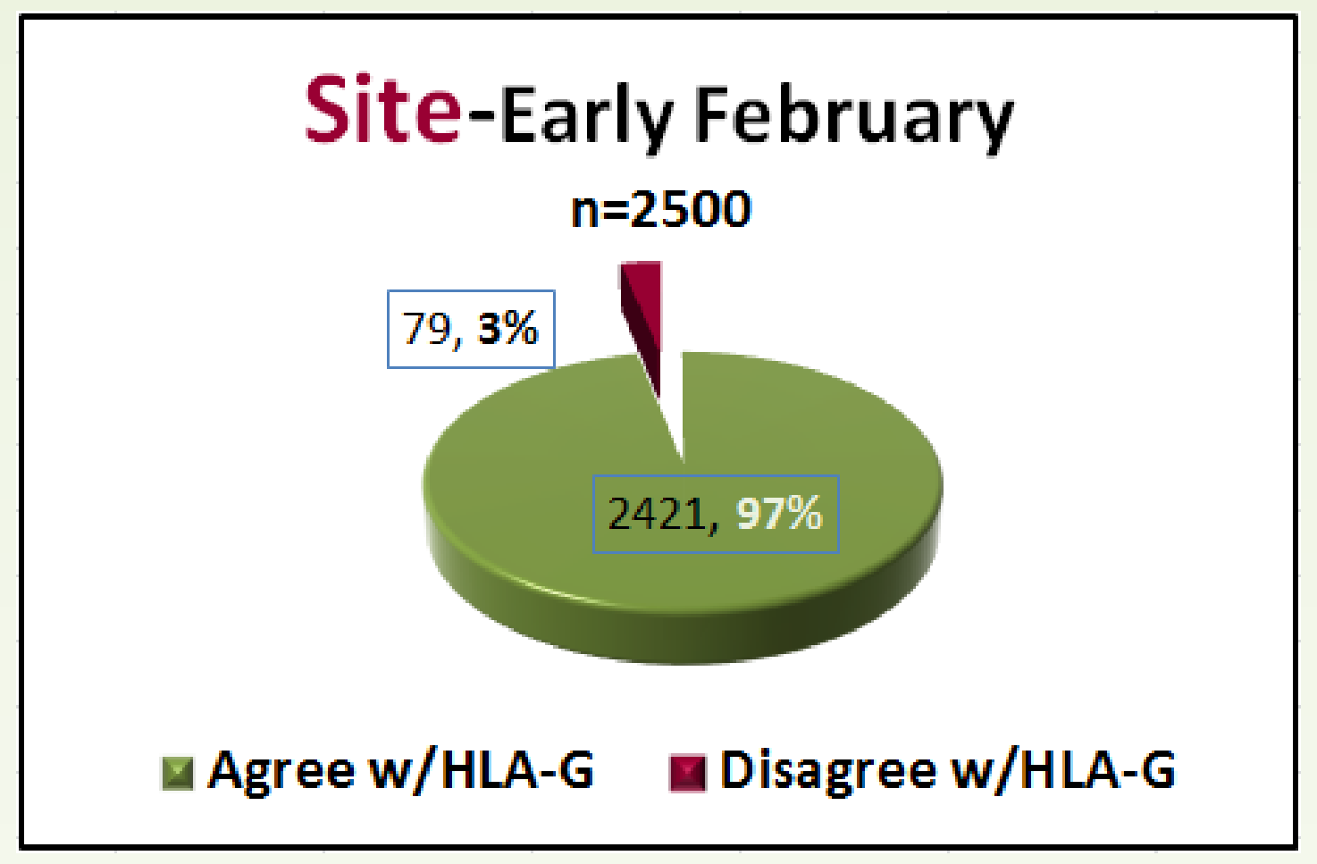


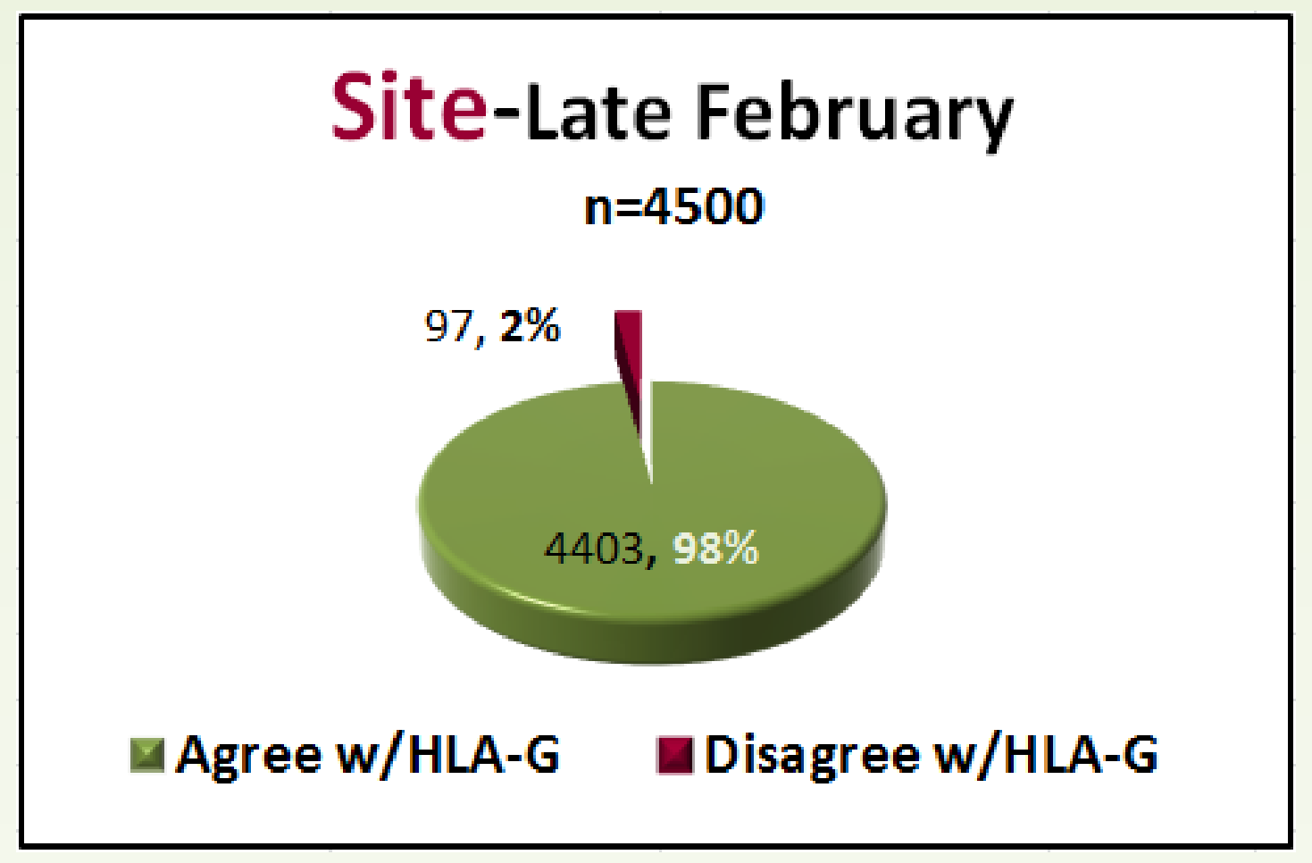
HLA-G determined that due to the complexity and variability of pathology reports, their preference would be to classify based on pathology report complexity. Five Stages of classification were defined for classification excluding "Out of Scope". Stages are outlined in the graph below :



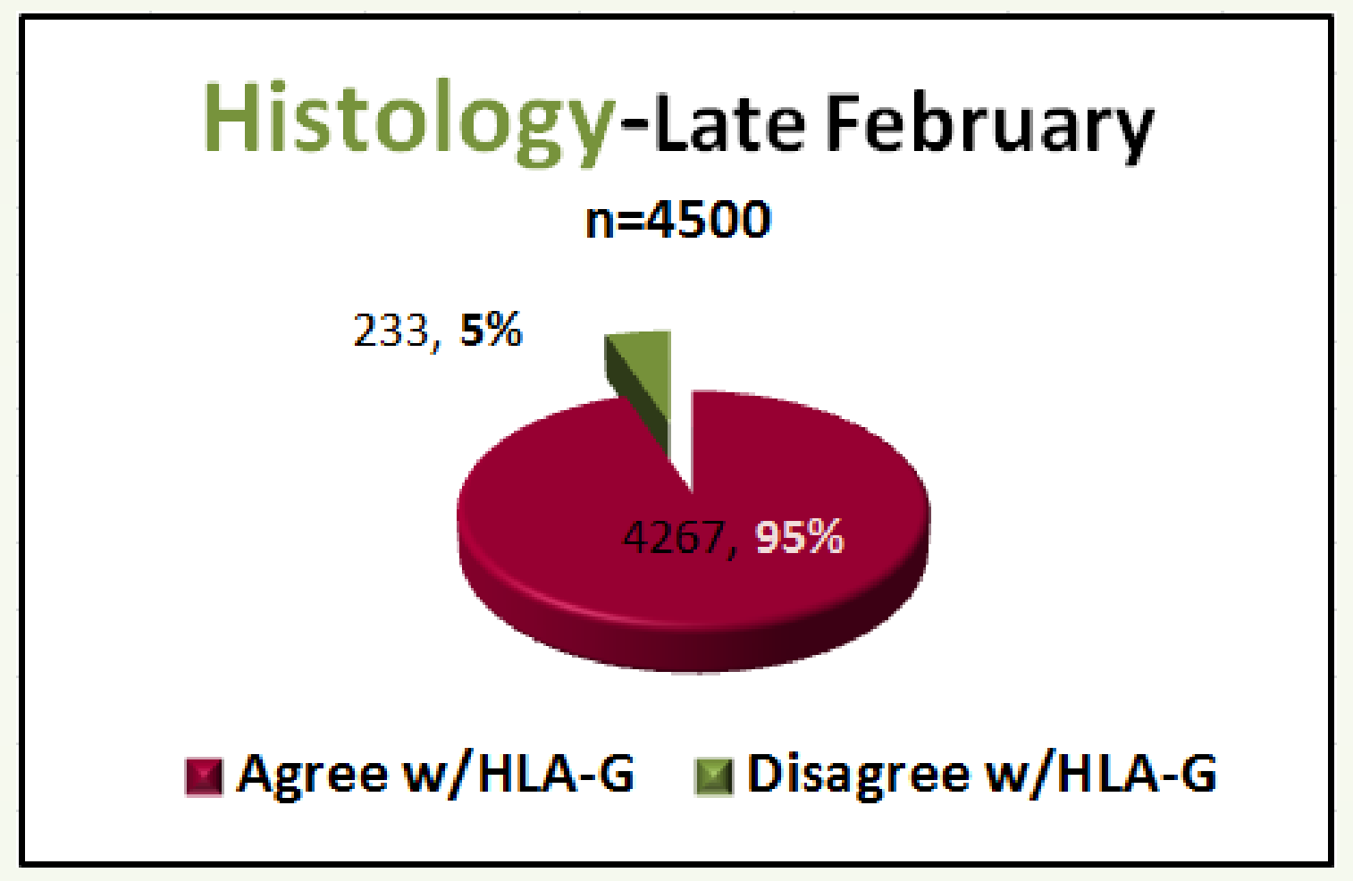
In early February, 2500 pathology reports were processed by HLA-G. 149 reports were reviewed by CCR staff where the NLP algorithm did not determine the same result as the original coder. This review was on site code only. Adjustments were made to the HLA-G process based on the Business Analyst feedback. In late February, HLA-G reprocessed the 2500 pathology reports and added another 2000, making the total processed 4500. 275 reports were returned with differences in both site and histology (see tables below). CCR reviewed the algorithm results and provided feedback to HLA-G on correct codes as well as the rationale for determining them. CCR staff provided supporting documentation for their codes citing standard setter resources originally provided to HLA-G. Discussions between HLA-G and CCR to resolve complex coding issues occurred to provide additional clarification on coding rules.

Pathology Report Review - Site:





Pathology Report Review - Histology:



Each time the algorithm is modified based on CCR feedback and the patterns identified have been resolved by HLA-G, the results become even more favorable.

Project Status

As of March 2017, HLA-G has been provided with a total of 10,000 pathology reports (5,000 reportable; 5,000 non-reportable). HLA-G reviews pathology reports provided and creates annotation tags. In order for the algorithm to learn efficiently and effectively, the annotation done must be accurate, and relevant to the task of screening and classifying path reports. For this reason, the discipline of language annotation is a critical link in fine-tuning the algorithm and requires an iterative process between HLA-G and CCR staff. To-date, they have annotated a total of 93,000 words or word combinations.

Although all stages are being processed to some degree, Stage 1 (One Specimen with One Site) is due to be completed by March 30th. Each subsequent stage is estimated to take 2 weeks from beginning of the process to completion. Completion is reached when accuracy is at or above 90%.

Conclusion

While the project was initially anticipated to be completed in January, 2017, the degree of complexity and variability within pathology reports was underestimated. The CCR agreed to a project extension of March for Stage 1, with 2 week intervals for each subsequent stage and project completion by June, 2017. Results appear favorable based on reported progress to-date. The CCR is looking forward to moving out of the project phase and into a production mode utilizing HLA-G's natural language processing tool for screening and classifying path reports.

Anticipating the success of this project, the CCR staff is investigating other possibilities for utilizing natural language processing tools to reduce manual work activities and increase efficiencies.

UCDAVIS
INSTITUTE FOR POPULATION HEALTH IMPROVEMENT

*The CalCARES program partners with the California Department of Public Health (CDPH) to manage the operations of the state mandated California Cancer Registry program

We would like to thank the following contributors: Public Health Institute, Jon Patrick. PhD, CEO-Health Language Analytics Global LLC (HLA-G); Rajiv Singh Kairon, B.Sc.