DATE: FEB 2020
PREPARED BY: KEVIN TRAN, ANA ALVAREZ

Medical Device Data Cleaning Executive Summary

Qualtrics Survey Deployment

After the Qualtrics survey was developed it was sent as part of a work package to three different data entry teams. Data entry teams also received a set of FDA 510(k) submission PDFs, each team was responsible for a different set of years. The survey captured the KNumber entered by the data entry team and ran it through the FDA's API to identify the corresponding Product Code and Company Name for that KNumbers. This information was checked against the manual entry by the team, as a way to flag potential errors.

Qualtrics Survey Data Error Rate Investigation

After conducting the survey, we investigated the error percentage of each survey to deduce which group performed better. We first extracted the survey results from all the Qualtrics surveys through Excel. Following this, we loaded the survey results within Jupyter Notebook through Anaconda Navigator and continued cleaning the survey results through Python by filling in the predicate device identification numbers, depicting whether or not the company name and product code was listed correctly, filling in the Tradename and predicate devices when empty, and updating the status of the K-numbers. From there, we parsed through the survey results to only keep entries that were fully completed, had the same amount of equivalent KNumbers listed, had the correct product code and company name, and were not duplicate entries with the same applicant device numbers. Keeping these entries, the only error remaining were for entries that had different predicate device numbers listed. This would then be checked in the manual review of entries.

Manual Review of Entries

To introduce an additional layer of error identification, groups were also given a subset of PDFs that had been entered by a different group. The majority of these entries were identical across both teams: these were interpreted to be non-errors. However, Kevin's script identified eighteen pairs of entries done by two different teams and did not match. Ana manually reviewed the PDFs for these Applicant KNumbers and compared the predicates entered by the groups to the information in the PDF. Ana recorded whether the entry by each group was correct. In some cases, neither group's entry was correct. For most of these pairs, the mismatched entries were actually done by the same group. These are thought to be entries that a group completed, and then re-did because they realized they had entered something incorrectly. The majority of these pairs by the same group had the second entry (still by the same team) correct. Only four pairs were identified as true errors, where neither team / neither entry was correct.

Summary of Error Rate Findings

Following both the error rate investigation and manual review of entries, we summarized our findings through extrapolating different analytics about the survey results. These included total errors, error entries, correct entries, duplicate entries, incomplete entries, incomplete entries resubmitted, entries with the same applicant devices but different inputs, unconfirmed and non-equivalent KNumber entries, and incorrect product code and company name. Reviewing these results, we found that overall there was an error percentage of 3.06% for group 1, 3.55% for group 2, and 20.09% for group 3. This all factored together had an overall error percentage rate of 4.50%.