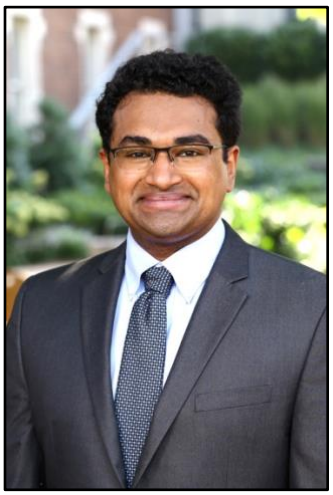
Avirath Vaidya

M.P.H. Candidate, Health Policy Track

avirath.u.vaidya@vanderbilt.edu



Practicum Site: Vanderbilt University Medical Center - Institute for Clinical and Translational Research

Practicum Site Supervisor: Laura Jones, M.P.H.

Integrating New Variables into the Research Derivative using a Probabilistic Matching Algorithm

Keywords: big data, record linkage, urology

Introduction: The Big Data Team at the Vanderbilt Institute for Clinical and Translational Research helps restructure healthcare data into formats more accessible for investigators. Understanding how to leverage data resources for research is a vital skill for public health professionals. The practicum primary objective was to develop an ingestion plan for new data from kidney stone patients into the Research Derivative (RD) of the Vanderbilt University Medical Center (VUMC) data pipeline. This data contained information for over 6,000 individuals, however lacked identifying information necessary to link to patient records. Therefore, a probabilistic matching algorithm was created to successfully integrate new variables into the RD.

Methods: Data dictionaries were created for the variables. A literature review for probabilistic matching techniques was performed to determine the best way to incorporate records with minimal identifying information available. Data standardization and manipulation was done in consultation with the research team. A matching algorithm was presented for approval to the project stakeholders. The algorithm was coded and executed by the programming team.

Results: The probabilistic algorithm resulted in 93% of patient records being successfully matched into the RD. The missing 7% of matches were attributed to patients' names changing and misspellings, however these records will be matched in the RD using chart review in the coming months. In addition to the data transfer being successful, a full summary of deterministic and probabilistic matching techniques was added to the Big Data Team website to serve as a reference for similar projects in the future.

Conclusions: The probabilistic algorithm developed during the practicum was successful. The kidney stone data will be invaluable for researchers as they work to better understand the factors that lead to kidney stone formation. The practicum experience reinforced the importance of creating tools to improve the quality of research at VUMC and medical centers around the world.