

International Journal of Population Data Science

Journal Website: www.ijpds.org



Probabilistic Linkage Pipeline Improving Linkage Quality and Explainability in Healthcare

Jonny Laidler¹, Amaia Imaz Blanco¹, and Divya Balasubramanian¹

¹NHS England, Leeds, United Kingdom

Objectives

The current methods used for data linkage or indexing in the healthcare industry follow deterministic algorithms that are not transparent to the end users and with often sub-optimal outcomes. Our work focuses on building and implementing a probabilistic algorithm that improves quality as well as explainability of the linkage.

Methods

In this project we have used Splink, a UK Ministry of Justice probabilistic linkage Python package, to build a pipeline that links any health data set to the Personal Demographics Service (PDS), a dataset containing information about all patients registered to a GP in England, which acts as our linkage “spine”. Our work has involved thorough investigation and evaluation every step of the process to ensure the quality of linkage is assured. We have taken considerations for data set ingestion, preprocessing, blocking rules, distance metric hierarchies, and explainability, including how to present the data appropriately to users of linked data.

Results

Throughout the developing process we have made comparisons to the existing deterministic linkage algorithm, clerically reviewing results that differed between the models, aiming to improve it. This allowed us to also build a ground truth dataset of records we reviewed and whether they were a true link or not. Furthermore, bias analysis is performed for evaluation purposes. Whilst we aim to continue improving the model, preliminary results show that this new methodology has improved linkage quality by up to 19% in comparison to the existing methodology. We are building an in-house capability to deliver this methodology at scale. Our mode, available in a public repo, provides additional metrics such as linkage probability made available to the end-users as part of improving transparency and explainability.

Conclusion

Our research and implementation provide evidence that probabilistic linkage algorithms are more sustainable indexing methods for continuous improvement of data quality, particularly in the healthcare industry. We believe the additional explainable measures will allow end-users to make informed decisions in their product creation ultimately improving patient health.

