# Probabilistic data standardisation of big heterogeneous Datasets in biomedicine

Alexia Sampri[1], Nophar Geifman[1], Philip Couch[1], and Niels Peek[1]

*[1] Division of Informatics, Imaging and Data Sciences University of Manchester, Manchester, UK*

Background

Putting data together from different sources into a homogeneous data resource would enable unprecedented opportunities to study human health. However, these disparate collections of data are inevitably heterogeneous and have made aggregation a difficult challenge. We focus on the issue of content heterogeneity in data integration. Traditional approaches for resolving content heterogeneity map all source datasets to a common data model that includes only shared data items.

Objectives

Our focus is on integration of structured data. We assume that each one of these datasets that needed to be integrated consists of a single table; and that each of these datasets describes a disjoint set of entities. Therefore, record linkage is not needed.

Methods

We propose the development of improved, probabilistic approaches for data integration, capable of advancing the timely utilisation of large-scale biomedical data resources. Our approaches aim to forego the need for perfect data standardisation by employing a probabilistic post-alignment of data items that is integrated with statistical inference. Using these approaches, missing or semantically ambiguous information is estimated from datasets potentially relevant for answering the research question.

Results

The MAximizing Sle ThERapeutic PotentiaL by Application of Novel Stratified approaches programme (MASTERPLANS) aims to improve care for Systemic Lupus Erythematosus patients by taking a precision medicine approach to identifying groups of patients that respond to biologic therapies. Based on dataset examples provided by MASTERPLANS we describe and evaluate the proposed probabilistic data integration approaches.

Conclusions

Our approaches insist on the future existence of health data heterogeneity. They strive for post alignment of Big datasets. As a post-alignment of heterogeneous data sources will be always imperfect and it is not a problem if we estimate the probability that they are. Our approaches are also pragmatic because they always provide an answer.

**Keywords**
Big Data, probabilistic data integration, data heterogeneity