Data generated by electronic health records, mobile and wearable technologies, and the numerous medical studies conducted worldwide have the potential to be extremely beneficial to the scientific and patient communities by generating new knowledge about health, disease, and treatment. This promise is well recognised by research communities but it remains the case that many biomedical datasets are underutilised. To realise the potential of such datasets, these must be integrated with other existing data, to generate large-scale, readily usable, cross-cutting, informative and therefore, useful data resources. However, in most cases, the process of integration is cumbersome and requires significant investment of time and effort. Different datasets will likely have non-uniform data schemas, employ different terminologies and naming conventions, apply different levels of abstraction, and suffer from missing annotation, making integration and comparison of data from different studies difficult.

Traditional approaches for data integration either rely on better coding of information at the source, which is rarely implemented adequately, or on the prospective alignment of datasets which is wasteful, labour-intensive and often impossible to achieve completely. The heterogeneity of data in biomedicine has been addressed using information models and coding standards (e.g ICD, HL7v3, SNOMED, UMLS, caBIG etc), which often, are hard to implement or lead to information loss. Generally speaking, databases can vary in their schema; table decomposition; metadata representation; structuring of fields (e.g coded vs. free text); and representation. With regards to in this paper we focus on representational heterogeneity, which occurs when different databases have identical real-world semantics but represent information differently. Combining multiple datasets is crucial to achieve generalisability (where findings from a sample population can be extended to larger populations), more accurate and larger denominators (to enable different research questions to be answered), and to better understand the different aspects of heterogeneity (such as in names, geographical locations, and health care services).

The MAximizing Sle ThERapeutic Potential by Application of Novel and Stratified approaches (MASTERPLANS) programme aims to improve care for SLE patients by taking a precision medicine approach to identifying groups of patients that respond to particular biologic therapies. Through MASTERPLANS we have gained access to data from three cohort studies: ALMS, LUNAR, and EXPLORER. The three studies were conducted across different sites, with different coding frames and employing different entry and exclusion criteria. Furthermore, study protocols, the evaluated medication, recruitment numbers, and duration of follow-up varied across the studies with only partial overlap.

We propose the development of an improved, probabilistic approach for data integration, capable of advancing the timely utilisation of large-scale biomedical data resources. Using this approach, missing or semantically ambiguous information is estimated from datasets potentially relevant for answering the research question. Thus, datasets alignment is learned from the data itself, rather than imposed on it. By adopting a top-down approach that starts with research questions rather than data collections, successfully integrated datasets may be created from data stemming from different sources. This phase includes familiarisation with existing data resources to be used in the research. Probabilistic approaches for data harmonisation are being developed and assessed by comparison to current integration gold standards. Also, to demonstrate its utility, the developed approach will be applied to real-world biomedical and health datasets including studies in Lupus and asthma. Each stage of the research is likely to yield high-impact publications. The methodological approach will be generalisable to many fields and as such will be targeted to translational and informatics methodology journals.