Heterogeneity and Accuracy Issues in Federated Patient Data Repositories

Sarah N. Lim Choi Keung, MSc¹, Edward Tyler, MSc¹, Adel Taweel, PhD², Theodoros N. Arvanitis, DPhil¹, Brendan Delaney, MD², F. D. Richard Hobbs, MB ChB¹
¹University of Birmingham, UK; ²King’s College London, UK

Abstract
The federation of patient data repositories is an essential precursor of analysis and reuse in clinical research. In the United Kingdom, primary care data originates from GP systems with syntactic and semantic differences. Identifying and recruiting eligible patients to clinical studies rely on the ability to search these repositories, despite data heterogeneity. In this work, we discuss the heterogeneity issues of data federation from all widely adopted GP systems to create a unified repository.

Introduction
Patient cohort identification and recruitment is a time-consuming and costly process but essential to ensuring adequate numbers of research subjects. In the UK primary care setting, much of this process is done by clinical research staff and practice staff using direct searches of the electronic health records (EHRs). The aim of our work is to facilitate the patient cohort identification and recruitment by semi-automating this process. All General Practitioner (GP) practices use EHRs¹; however, this data is not directly accessible to research staff to anonymously search for patient cohorts.

Regional Federation of Primary Care Data
In the UK, Primary Care Trusts (PCTs) are regional organizations, each responsible for a number of GP practices. Some PCTs have local initiatives to integrate patient data from GP systems into a unified EHR repository for several reasons, including commissioning and improved patient care. The federation of patient data has created the opportunity to facilitate the identification of patient cohorts through access to these regional data repositories.

Analysis of Data Quality and Accuracy
The PCT under study in this paper provides healthcare services for more than 300,000 patients in 78 GP practices. The unified repository federates data from 63 practices. The data is coded using Read Codes, version 2, a 5-byte clinical terminology ubiquitously used in the UK primary care. However, the data coding does not strictly follow the standardized Read Code, but also uses several local coding approaches devised by propriety EHR systems. From the data extract, 42.8% of patient record items have incorrect 5-byte Read codes, i.e. not found in the published Read Code terminology. These incorrect codes have been further analyzed and categorized in Table 1.

<table>
<thead>
<tr>
<th>Error Category</th>
<th>Codes</th>
<th>Practices</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMIS GP System</td>
<td>46.1%</td>
<td>51</td>
<td>65.8%</td>
</tr>
<tr>
<td>7-byte Read Codes</td>
<td>32.0%</td>
<td>6</td>
<td>7.8%</td>
</tr>
<tr>
<td>Other ‘local’ codes</td>
<td>24.8%</td>
<td>6</td>
<td>12.3%</td>
</tr>
</tbody>
</table>

Table 1. Categories of incorrect Read codes.

GP systems allow propriety local codes to be created if no appropriate Read Codes are found in the standard terminology. Although many equivalent codes are now available in the reference Read Code terminology, a notable number of local codes are still used, presumably as no strategy has been deployed to update these codes. Some other GP practices use 7-byte Read Codes, which are made up of the normal 5-bytes codes extended by a 2-byte code for synonymous terms. The proportion of incorrect Read Codes decreases to 27.3% from 32.0% after considering the correct 7-byte Read Codes.

Conclusion
The proportion of incorrect Read Codes is a significant problem for data quality. Further analysis is required to examine the quality of the data and its accuracy for patient cohort identification feasibility studies. Various data quality optimization techniques may need to be applied on the data in an attempt to improve data re-use for clinical research.

References