Piloting the EHR4CR Feasibility Platform across Europe*

J. Doods1; R. Bache2,3; M. McGilchrist4; C. Daniel5,6; M. Dugas1; F. Fritz1 on behalf of Work Package 7
1University of Münster, Münster, Germany;
2Department of Informatics, School of Natural and Mathematical Sciences, King’s College London, London, UK;
3Department of Primary Care and Public Health Sciences, King’s College London, London, UK;
4Health Informatics Centre, University of Dundee, Dundee, UK;
5INSERM, UMR S 1142, LIMICS, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, France;
6CCS SI Patient, AP-HP , Paris, France

Keywords
Protocol feasibility, clinical trials, electronic health records, secondary use, clinical research.

Summary
Background: Pharmaceutical clinical trials are primarily conducted across many countries, yet recruitment numbers are frequently not met in time. Electronic health records store large amounts of potentially useful data that could aid in this process. The EHR4CR project aims at re-using EHR data for clinical research purposes.

Objective: To evaluate whether the protocol feasibility platform produced by the Electronic Health Records for Clinical Research (EHR4CR) project can be installed and set up in accordance with local technical and governance requirements to execute protocol feasibility queries uniformly across national borders.

Methods: We installed specifically engineered software and warehouses at local sites. Approvals for data access and usage of the platform were acquired and terminology mapping of local site codes to central platform codes were performed. A test data set, or real EHR data where approvals were in place, were loaded into data warehouses. Test feasibility queries were created on a central component of the platform and sent to the local components at eleven university hospitals.

Results: To use real, de-identified EHR data we obtained permissions and approvals from ‘data controllers’ and ethics committees. Through the platform we were able to create feasibility queries, distribute them to eleven university hospitals and retrieve aggregated patient counts of both test data and de-identified EHR data.

Conclusion: It is possible to install a uniform piece of software in different university hospitals in five European countries and configure it to the requirements of the local networks, while complying with local data protection regulations. We were also able set up ETL processes and data warehouses, to re-use EHR data for feasibility queries distributed over the EHR4CR platform.

1. Introduction
Clinical studies are fundamental for expanding medical knowledge and improving healthcare. To achieve sufficient recruitment numbers of patients, studies are often conducted in multiple locations crossing national borders. However recruitment often fails to meet targets. Van der Wouden et al. [1] and McDonald et al. [2] show that in more than 50% of trials started, recruitment periods have to be extended. Electronic health records are increasingly accumulating large quantities of potentially useful patient data. This electronic information is already used to support recruitment for single sites [3, 4] and despite the many challenges [5] of using this data at a multinational level it also has great potential [6], not only for patient recruitment but also for protocol feasibility.

The Electronic Healthcare Records for Clinical Research (EHR4CR) [7] project aims to support clinical research on a European scale and is part funded by the Innovative Medicines Initiative (IMI) [8]. The project runs for four years (2011–2014) and comprises of 33 partners from the pharmaceutical industry, academia and small and medium-sized enterprises. The university hospitals in the project are located in five countries France (Assistance Publique – Hôpitaux de Paris AP-HP, Université de Rennes U936), Germany (Friedrich-Alexander-Universität Erlangen-Nürnberg FAU, Westfälische Wilhelms-Universität Münster WWU), Poland (Medical University of Warsaw MUW),...
Switzerland (Hôpitaux Universitaires de Genève HUG) and the United Kingdom (Kings College London KCL, University College London UCL, University of Dundee UNIVDUN, University of Glasgow UoG, University of Manchester UoM), European Federation of Pharmaceutical Industries and Associations (EFPIA) members of the project are AMGEN, AstraZeneca, Bayer Health Care, F. Hoffmann-La Roche Ltd, GlaxoSmithKline, Johnson & Johnson, Lilly, MERCK KGAa, Novartis Pharma AG and Sanofi-Aventis. EHR4CR addresses the following scenarios 1) clinical protocol feasibility, 2) patient identification and recruitment, 3) clinical trial execution and 4) adverse event reporting. The covered disease areas include: oncology, inflammatory diseases, neuroscience, diabetes, cardiovascular and respiratory diseases.

At the clinical protocol feasibility stage the goals are to determine whether the planned in- and exclusion criteria are adequate for a trial protocol and to identify suitable medical centers or clinics. Currently paper questionnaires are sent out to physicians asking them how many patients they treat with certain criteria in a specific period of time. Those questionnaires are then collected and based on that information changes are be made to the criteria. To support this first scenario of EHR4CR, a platform was designed and implemented by ten informaticians at the different university hospitals sites and one SME, resulting in three major components (Workbench, Orchestrator and Endpoint) [9]. The Workbench was developed to create site feasibility queries. A set of eligibility criteria can be composed using a drag-and-drop graphical user interface utilizing a comprehensive query model [10]. The criteria in the query model are represented through codes from a central terminology. A query can consist of several inclusion and exclusion criteria, which can be linked by Boolean and temporal logic. Temporal constraints include: ‘diagnosis x within the last 6 months’ or ‘diagnosis x before procedure y’ and frequencies such as ‘laboratory finding z at least twice within the last year’ can be used to construct queries. The problem of executing complex temporal queries on diverse warehouse schemata was solved by separating Boolean and temporal logic from data extraction of the necessary clinical data from the warehouse [10]. The former was implemented as a generic algorithm in Java and the latter by a set of SQL templates that can be adapted for each warehouse schema. The aggregated results are visualized in different ways: by sites, individual criterion, age in decades and gender.

The Orchestrator is a message broker which receives encrypted feasibility queries from the workbench and information as to which sites the queries should be relayed. It retains those queries until the local software component, known as the query endpoint, requests them. The Endpoint handles the access to the database and is responsible for parsing the site feasibility queries. It polls the orchestrator for new queries at frequent intervals, which are then executed to yield counts of patients using anonymised patients data. Two clinical data warehouse (CDW) types are currently supported, an EHR4CR ‘native’ schema and a Integrating Biology & the Bedside (i2b2) [11] warehouse, depending on the sites’ preference or existing infrastructure. All communication between the different components is done with the Simple Object Access Protocol over Simple Mail Transfer Protocol (SOAP:SMTP).

To query multiple CDWs in different countries, a central set of clinical codes had to be agreed upon. It was decided to follow a pragmatic approach by identifying data elements from clinical studies of the EHR4CR EFPIA partners. These ‘feasibility data elements’ are in general fewer in numbers and less complex in comparison to data elements for patient recruitment [12]. Based upon the elements appropriate central codes were identified, for example SNOMED-CT code 271650006 for diastolic blood pressure, to which local codes at the sites are mapped. For pre-existing i2b2 CDWs a terminology service was implemented, to map the local codes to central ones on runtime, while native CDWs did the mapping during extract, transform and load (ETL) processes.

To send a federated query to university hospitals in different countries the different data protection legislations in each country have to be adhered to. In Germany for example obstacles for re-using patient data are high because of the right to ‘informational self-determination’. On the other hand, governance principles within the UK permit the existence of large databases, like the General Practice Research Datalink with access to data on more than 5.5 million subjects, which can be used by academia, governments and the pharmaceutical industry worldwide [13].

2. Objective

In order to run the clinical protocol feasibility scenario it is necessary to connect hospital sites to a central platform. We therefore assessed whether it is possible to install the same software components locally in different hospital networks and retrieve aggregated protocol feasibility numbers from EHRs across Europe.

3. Methods

In order to setup the local components and test whether aggregated patient counts are sent back to the workbench, the following steps were undertaken:

1. Approvals: To use EHR data the respective sites acquired necessary permissions and approvals from data controllers, governance authorities and ethical committees to get access to real EHR data and use it within the EHR4CR project.
2. Technical setup: Each site set up physical or virtual machines running Microsoft Windows or a Linux distribution. EHR4CR software components were deployed and configured according to the sites network requirements.
3. Warehouse and ETL: For the native CDWs MySQL databases were installed, while the i2b2 warehouses used Oracle. ETL procedures were set up to transform the source data to the corresponding schema and load it into the warehouses.
4. Mapping: Each hospital created mappings of local codes from their site to codes of a central terminology.
5. Integration test: During this technical test queries were created on the workbench and sent to the hospital endpoints to determine whether all sites are able to retrieve and process the queries.

4. Results

In October 2012 a proof of concept was conducted, to see whether all sites could install the software and set up CDWs. It included an end to end test in which queries were generated at the workbench and transmitted to the endpoints and aggregated numbers sent back to the workbench.

1. All sites started obtaining the necessary approvals to use real EHR data for the feasibility scenario and at the time of the test seven out of the eleven university hospitals had succeeded. The processes varied depending on situation at the site and country legislations. Some did have general approval to use EHR data in the project while most had to obtain approvals from data controllers, ethical committees and/or data protection officers. A representative clinical data set in an anonymized form was provided for those sites that had not finished the approval process for using de-identified data.

2. Physical or virtual machines were set up with the required Java Runtime Environment 7. We successfully deployed the same endpoint software, written in Java, in Tomcat application servers at all eleven data provider sites and configured the components according to the local network requirements which took in general less than half a day. Those sites which used i2b2 additionally deployed a terminology service in Tomcat.

3. Seven sites set up databases with the native database schema and four sites used i2b2 CDWs. All eleven sites were able to load the test data into their warehouses and those sites with approvals also loaded real, de-identified data coming from their local EHRs into the CDWs (Table 1). ETL operations imposed restrictions on units, scores of other kinds and timestamps associated with clinical events and dates of birth and death. Dates would be obfuscated and shifted up to 356 days back in time on a per-subject basis. There was significant variation in the software tools used to perform the ETL, for example Talend Open Studio or 4th/3rd generation language applications were used.

4. Each of the eleven sites did a mapping of 111 data elements from local codes to the central terminology. These elements were identified by analyzing 12 clinical trials of interest from the EFPIA companies in the consortium. The mapping task was done manually and took each site about a week. The central clinical terminology used for defining queries comprised the following standard coding systems: the International Statistical Classification of Diseases and Related Health Problems 10th revision (ICD10-WHO), Anatomical Therapeutic Chemical (ATC) classification system, The Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT), Logical Observation Identifiers Names and Codes (LOINC) and the Anatomic Pathology Lexicon (PathLex). The data elements cover the areas of demographics, diagnoses, procedures, medication, findings, laboratory findings, medical history and scores. The initial list is available in the supplementary material.

5. With the mappings in place queries were created centrally at the workbench. Queries contained eligibility criteria like ‘More or equals to 18 years old’ AND ‘Insulin-dependent diabetes mellitus’ AND ‘Hgb A1c Bld HPLC > 6%’ AND NOT ‘Non-insulin-dependent diabetes mellitus’ and were send to all eleven university hospital sites. The local endpoints could all successfully poll the queries from the Orchestrator, execute them and send back the aggregated results within less than five minutes. As seen in Figure 1 components were deployed in different locations throughout Europe. All eleven university hospitals hosted endpoints, while the orchestrator was situated in Belgium and the workbench was hosted in France.

5. Discussion

Using data from routine documentation for clinical protocol feasibility has great potential, as it can speed up the process significantly. It can for example help in creating better study protocols, when the potential study patient numbers of different criteria can be compared and adjusted accordingly in a matter of minutes rather than weeks.

Obtaining approvals for data use and obtaining the consent from ethics committees to use the EHR4CR platform is a time consuming task. The four sites that did not have approvals at the time of the test did not run into problems, as no patient information leaves the hospital and only aggregated counts get send back to the workbench, but rather had to wait for decisions from committees, which in general only meet every few weeks. Although each country has its own data protection laws and regulations, we were already able to have real data available for the feasibility scenario from seven sites at the time of the test.

With our proof of concept we showed, that it is possible to install one uniform piece of software and configure it to the requirements of different hospital networks. During the installation and preparation of the platform we identified issues with the network configurations at UK sites however. The sites could not host the endpoint software in the National Health Service (NHS) networks, due to restrictive firewall configurations. This blocked the SMTP communication and as a conse-

<table>
<thead>
<tr>
<th>No. of/Type of data used</th>
<th>Demo data</th>
<th>De-identified EHR data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sites using the EHR4CR DB schema</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Sites using the i2b2 CDW</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Sites in total</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1: Overview of warehouse schemas and data used for the proof of concept.
differing permissions had to be obtained, to use de-identified data in the university hospital network as a temporary measure. A more permanent solution requires that the communication be changed to the Hypertext Transfer Protocol (HTTP) instead of SMTP in the future. The communication between software components is handled through the Java Message Service in newer versions of the platform.

We were able to set up different CDWs and fill them with demo data as well as real, de-identified EHR data. The disparate nature of the site EHR systems offered a good test of the native schema as a useful target for ETL operations. Although the real data was de-identified, for example by removing patient IDs, some sites required further measures to prevent re-identification of patients. Obfuscating dates by shifting them back a set number of days in time or setting them to the first day of the month or quarter was done, similar to approaches from the research warehouse in Vanderbilt and the SHARPn project [14, 15].

A further concern that was brought up by the sites using the native CDW is that performing terminology mapping during the ETL leads to the problem of altering the ETL scripts every time a new data element is added or the terminology is updated. Thus, in future, the native CDW will be populated with the local codes and mapping will be performed at query time as is the case with the i2b2 CDW.

We were able to send one centrally created feasibility query and execute it at 11 sites to receive aggregated feasibility numbers with two different types of database schemas from hospitals in five European countries.

5.1 Related Work

The Strategic Health IT Advanced Research Projects (SHARPn) have similar goals to EHR4CR, that is to say the re-use of EHR data to improve patient safety and medical outcome and research [16]. Some concepts are the same for both projects like the use of a terminology mapping service for semantic interoperability, but there are differences as well. Different in SHARPn are the use of natural language processing and storing anonymized data centrally, while in EHR4CR solely structured data are used and data is stored in multiple local CDWs. Rea et al. [15] also reported similar issues we identified when dealing with interoperability for example that mapping efforts of local to central terminologies is a laborious task.

The Shared Health Research Informatics Network (SHRINE) [17] is an undertaking with similar goals to those of EHR4CR as well, namely the re-use of routine data for research. SHRINE enables queries across hospitals and aggregation of patient numbers. Queries can be created by applying Boolean logic and constraints like dates and number of occurrences on demographics, diagnoses, medication and laboratory tests data. Similar to EHR4CR and SHARPn a terminology service with mapping from local to central terminologies is in place, but it also supports query expansion, which is currently not the case for EHR4CR. Differences are that complex temporal constraints besides dates are not supported, for example ‘diagnosis x before medication y’. SHRINE networks are used in the US and Europe successfully already, but only within national borders. In comparison EHR4CR allows querying hospitals across different countries.

5.2 Limitations and Future Work

Our proof of concept was a technical feasibility test to see if the platform could work across countries and different network settings. In this work we did not intend to describe the implementation of the components, as such are already published elsewhere [10; 18] or are in the process of being published with a more technical
focus. Two evaluations are currently being conducted. The first one evaluates whether the endpoint algorithm works correctly and identifies the same subjects IDs of the test data set across multiple sites. The second evaluation of the EHR4CR feasibility platform conducted measures whether our approach improves clinical protocol feasibility process as such. It will focus on time frames and validity of retrieved numbers for the current and EHR4CR processes. Another major success factor is the sustainability of the EHR4CR platform. To ensure this an exhaustive business model is being developed and will be published in the near future.

The mapping of local to central codes was done manually for a limited set of central codes and independently of ETL processes. This was done as a pragmatic approach to see if the platform works. However, during testing, the mappings were not generated or verified by clinical staff and not all central codes were mapped successfully. Resource requirements for ETL and terminology mapping varied substantially between sites depending on pre-existing resources and operations. When the number of central codes increases, creating and maintaining the mapping in a manual way will not be suitable anymore and methods will have to be identified to automatize this task.

6. Conclusions

The EHR4CR feasibility system has been shown to work technically across different hospital sites in Europe. A query can be composed at the workbench using the central terminology and executed at local sites with local terminologies of different EHR systems. These queries can be executed on individual database schemas as has been shown here for the EHR4CR ‘native’ and i2b2 schema. Aggregated patient counts can be sent back as results and displayed at the workbench. Both technological and governance issues have been overcome to demonstrate this proof of concept.

Acknowledgments

We would like to thank all the partners who developed the EHR4CR feasibility platform and also the site representatives who installed the software components and prepared the sites.

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement nº 115189, sources of which are composed of financial contribution from the European Union’s Seventh Framework Programme (FP7/2007–2013) and EFPIA companies’ in kind contribution.

References