A Query Tool Enabling Clinicians and Researchers to Explore Patient Cohorts

Sarah N. LIM CHOI KEUNG[a,1], Omar KHAN[a], Ali ASADIPOUR[a], Huseyin DERELI[a], Lei ZHAO[a], Tim ROBBINS[a,b] and Theodoros N. ARVANITIS[a]
[a] *Institute of Digital Healthcare, WMG, University of Warwick, Coventry, UK*
[b] *University Hospitals Coventry and Warwickshire, Coventry, UK*

Abstract. Due to the increasing amount of health information being gathered and the potential benefit of data reuse, it is now becoming a necessity for tools, which collect and analyse this data, to support integration of heterogeneous datasets, as well as provide intuitive user interfaces, which allow clinicians and researchers to query the data without needing to form complex SQL queries. The West Midlands Query Tool consists of an easy-to-use graph-based GUI, which interacts with a flexible middleware application, able to connect to multiple heterogeneous data sources and return cohort counts satisfying query criteria.

Keywords. Query Tool, Service-Oriented Architecture, Heterogeneous datasets.

1. Introduction

The big data health revolution is under way, with an increasing amount of information being gathered. This availability of data necessitates the reuse the health data, based on the *collect once, use many* paradigm [1]. This enables clinicians to more meaningfully access to their patients' records in order to enhance decision making and support researchers to find eligible research participants and pose feasible research questions.

The West Midlands Query Tool (WMQT) project aims to adapt existing query tools for cohort identification and patient dataset exploration in the West Midlands region of the United Kingdom. In this paper, we present our approach and the design of the tool, enabling clinicians themselves to explore health records to support clinical decision making and research. The tool has been developed with a focus on finding cohort counts. The user-friendliness of the tool is an important feature, while the underlying platform is generic and can be used on different data sources.

1.1. Related Work on Query Building and Health Data Exploration

A number of query tools, such as i2b2 [2], TRANSFoRm Query Workbench [3] and TrialViz [4] have been developed by different teams to try and make full sense of routinely collected data, for clinical research in particular. They allow the end-user to build a query from a criteria set and return the count of matching patients in the dataset. The TRANSFoRm Query Workbench is the only approach to query heterogeneous datasets, without first converting them into a common schema. Some of the limitations

[1] Corresponding Author. E-mail: S.N.Lim-Choi-Keung@warwick.ac.uk

of the above-mentioned tools concern the visual element to facilitate users to define their query and also see the results, at different query execution points.

1.2. Example Clinical Scenario

An example scenario demonstrating WMQT in a clinical setting would be the exploration of a new medication approved for the treatment of Type 2 Diabetes Mellitus. This example medication is demonstrated through clinical trials to improve long-term glucose control (Hba1c) in a subset of patients who are aged over-55, already on two other common anti-diabetes medications (metformin and either gliclazide or glibenclamide), yet whose disease remains poorly controlled (Hba1c above 48 mmol/mol). The medication is only safe in patients with a normal renal function (eGFR >60 ml/minute/1.73m^2). WMQT would enable a clinician to rapidly interrogate their electronic health record system to identify which of their patients would both be eligible and benefit from such a medication. They would be able to flag such patients so at subsequent clinic attendances they can discuss, and potentially start, the new medication. The tool would also be useful to researchers in identifying the potential population size for the new medication intervention, prior to and during development.

2. Methods

2.1. Design Overview

Reuse of medical record data for clinical research presents many challenges [5], particularly in the case of accessing and integrating data from heterogeneous databases. The lack of intuitive query tools can further impede access of non-technical users. Many existing solutions, such as i2b2 [2], rely on creation and maintenance of a central data warehouse, which often requires complex Extract, Transform, Load (ETL) procedures to transfer data from different representations into a common model. While this approach is efficient for the analytic workloads for which the database is optimised, it cannot easily integrate external data sources where providers retain control on the database and restrict use of ETL. The WMQT data integration approach, instead, relies on data virtualisation [6], where heterogeneous data sources are encapsulated under a unified abstracted query application programming interface (API). This query API is a set of protocols and tools to build software applications. Data virtualisation is flexible and agile as new virtualised views can be created dynamically without a change in the data source. WMQT is platform independent, and is implemented as an HTML5 web application on a service-oriented architecture.

2.2. Graphical User Interface and Visual Query Builder

The design of the WMQT graphical user interface is based on the concept of visual data flow programming [7]. Users design queries by building a visual data flow diagram, which represents a step-by-step eligibility decision process. All query values, parameters and relationships are represented in this flowchart-style diagram, a unique approach which focuses on simplifying the user's process. Each node in the flowchart represents an eligibility criterion or subset. In each step, users can use source supported

terminologies and temporal criteria to filter eligible patients. Relieving the pain of debugging complex SQL statements, a diagrammatic representation visualises the decision-making process and helps users comprehend and reason about the query logic.

As the process of setting up eligibility criteria normally involves a large amount of data input when specifying the criteria and their ranges, an enhanced user interface is provided to ensure that entering this information does not become burdensome to users. The appropriate use of colours, icons, shapes, and drag-and-drop features allow users to focus on their task instead of the technical aspects of the tool, as they assist the processes of human perception, memory, attention, and decision-making.

2.3. Service-oriented Architecture

Behind the front-end web application sits a web service application, which executes the queries on selected source datasets. Web services are commonly used to improve the design of distributed system applications [8]. Features such as interoperability, communication reliability, enhanced security and ease of cross-platform access to data entities are key advantages of this approach. The high-level stages involved in executing a query and presenting the results involve a query request message from the visual query interface, which is then parsed to dynamically generate query scripts for each node or filter in the flowchart, mapping the parameters to corresponding schema information in real-time. All query scripts are executed, in sequence, on a target database and logical connections between nodes are applied on the results. Finally, eligible patients for each query node are sent to the frond-end application in a response message. A customised XML messaging template was designed to standardise the representation of the visual query and the communication between the applications.

3. Results

In the example scenario, a clinician can start building a query using the eligibility criteria after logging in to the web application. After selecting the data source, a query diagram area is displayed allowing the user to: add new criteria; remove a selected criteria node; and execute the query. Every new filter is represented by a node: blue nodes for inclusion and red for exclusion criteria. Nodes can be moved around in the diagram area, making it easy for users to organise the query to suit their needs. Users can edit any parameters by double clicking a node. Popups in the same window guide the users through selecting parameters and entering required values. Once ready, the query can then be executed, as described in Section 2.3. A history of results is kept for every time a query is run. Example results are shown in Figure 1**Error! Reference source not found.**. Starting with a population of 10000 patients, up to 90 patients meet the full criteria.

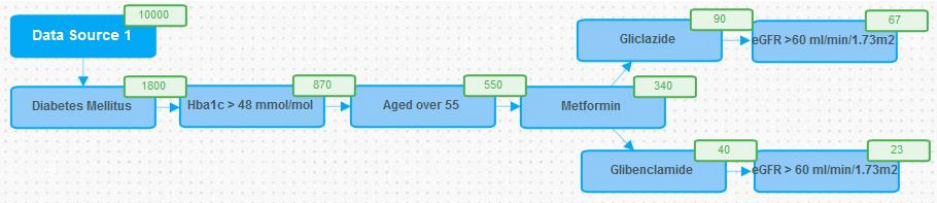

Figure 1: Screenshot of user interface for example clinical scenario.

4. Discussion

The graphical user interface allows easy building of queries without the use of technical querying languages and, as cohort counts are returned for each node, it facilitates modification of these queries and supports hypothesis generation. As the system is based on a generic data integration platform, it can be extended to support other application scenarios, such as analytics for different user groups. In our scenario for instance, WMQT can be extended to allow analysis of the likely drug cost and potential benefits from reduced complication rates, due to the improved glycaemic control within the patient population. Later in the drug's lifespan, WMQT would enable interrogation by research teams, exploring possible interactions or novel associations amongst the prescribed population. The tool would enable the comparison of cohorts within research databases who are prescribed the medication, compared to a matched control group. To extend the benefits of the tool and to facilitate further analysis of cohorts, the tool could also be extended to export data and also allow users to add new data sources and data virtualisation through the graphical user interface.

5. Conclusions

The WMQT consists of a well-designed web application, with an easy-to-use visual query builder, which supports the construction and modification of complex, multi-level and branching queries for identification of patient cohorts, comparison of cohorts with control groups and analysis of results at each level of the query. The highly adaptable platform, upon which the WMQT is built, allows the system to be customised for a diverse range of healthcare professionals. It enables querying of various heterogeneous datasets, both research and clinical, without the need to convert to a common schema.

Acknowledgements: This work has been supported by the West Midlands Academic Health Science Network, UK.

References

[1] J. J. Cimino, Collect once, use many. Enabling the reuse of clinical data through controlled terminologies, *J. AHIMA*, 78(2), (2007), 24–29; quiz 31–32.

[2] S. N. Murphy *et al.*, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), *JAMIA*, 17(2), (2010), 124 –130.

[3] L. Zhao *et al.*, EU FP7 TRANSFoRm Project: Query Workbench for Participant Identification and Data Extraction, in *2015 Joint Summits on Translational Science*, 2015.
[4] A. R. Tate *et al*., Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface, *JAMIA*, 21(2), (2014), 292–298.
[5] H. U. Prokosch and T. Ganslandt, Perspectives for medical informatics. Reusing the electronic medical record for clinical research, *Methods Inf. Med.*, 48(1), (2009), 38–44.
[6] R. van der Lans, *Data Virtualization for Business Intelligence Systems: Revolutionizing Data Integration for Data Warehouses*. Elsevier, 2012.
[7] W. M. Johnston, J. R. P. Hanna, and R. J. Millar, Advances in Dataflow Programming Languages, *ACM Comput Surv*, 36(1), (2004), 1–34.
[8] M. Parhi *et al.*, Discovery of sensor web registry services for WSN with Multi-layered SOA framework, in *2nd Int. Conf. on Computer and Communication Technology (ICCCT)*, 2011, 524–530.