Challenges in data warehousing for personalised medicine

Benjamin Jefferys, Peter Coveney
University College London

Introduction

A data warehouse is a collection of data intended for analysis and generating reports. Such a warehouse has clear utility in studying and treating human disease, and is especially useful in personalising healthcare. Large standardised sets of electronic health records, clinical trial data and related information will enhance analytical methods by providing a greater volume of data, and improve stratification by providing a greater breadth of data, leading to more personalised treatment.

However, information about disease treatment and resulting from clinical research is hard to collect into a warehouse for several reasons: it is highly fragmented; its use is hindered by numerous legal and ethical constraints; it is highly sensitive; it is generally stored without reference to universal standards of terminology, language or schema; and it is used for scientific analyses which must be accountable and reproducible.

The p-medicine project is constructing a data warehouse for analysis of data on treatment of three different cancers, with a long-term vision of creating infrastructure for personalised medicine in general. We describe existing solutions and our proposed response to the challenges presented above.

Collection

Our data warehouse will be distributed across several sites for the purpose of ensuring high quality of services and data availability. This also allows for the possibility of deployment within hospital data security environments: however, the primary use case, to maximise the scope of analysis within p-medicine, is that data is pushed to warehouses *outside* the hospital, and therefore available to our analysts.

We have clinical partners directly funded by p-medicine, who will proactively feed data from their initiatives into the warehouse. Data managers are employed on the project by each clinical partner. This will give an incubator to develop our ideas for data collection, such that they are ready for deployment to users not directly involved in p-medicine. The p-medicine clinical partners are obliged to use and test the data submission interfaces we develop, and give detailed feedback.

We propose the following methods to encourage clinicians to submit data: provision of the p-medicine infrastructure only to clinical groups who provide data; provision of a clinical trial management system called Obtima (1), which will integrate data submission into day-to-day trial management tasks; and highlighting the utility of integrating fragmented data sources within a particular institution, for use within that institution

Legal and ethical issues

Although, to an extent, technical solutions can enforce particular usage patterns, there will always be opportunities for data to be abused which are beyond the control of the p-medicine infrastructure. A dedicated legal entity, the Center for Data Protection (2) based in Belgium will act as data controller for p-medicine. It will have contracts with data providers and data consumers, ensuring data is used within legal and ethical constraints. This avoids the need for individual contracts between all data providers and consumers.

Anonymisation, authorisation and auditing

For most analysis purposes, personally identifiable information is not needed. To minimise the possibility of a leak of sensitive information, it can be anonymised by removing personal information.

However, it is useful to be able to trace back data to a particular individual, in order to add new information regarding them to the warehouse, and in order to solicit new information from them based upon a given analysis. Therefore, data is pseudonymised by a third party, replacing information which identifies an individual with a warehouse identifier, which is related to a particular person in a database maintained by the third party.

There are several challenges associated with this process. A date of birth, diagnosis or treatment might be sufficient to identify an individual. There are different ways of anonymising this information, and the best one to use depends upon the subsequent analysis required. A surprisingly small amount of data is required to identify an individual, especially if combined with information gleaned from social networking sites such as Twitter, or personal blogs. Finally, pseudonymised data is considered, by the EU, in terms of data protection, equivalent to personal data, and must be treated with the same care.

We must still, therefore, take care to secure the data and only allow access to authorised users bound by contracts with p-medicine. A role-based authentication and authorisation system will give a basic level of security, all transactions will be logged, and standardised automated security tests will be regularly performed by servers external to the warehouse.

Standardisation

Standards for the collection and recording of data on patient treatment and clinical trials are still fairly basic, and even then are not widely used. Some simple terminologies are used, usually where the medical infrastructure requires it - for example, use of HL7 (3) in reporting for insurance claims in the USA). Even within a particular hospital, there are no standard methods for recording data. Therefore, integrating data into the warehouse requires a process of semantic standardisation.

We are developing a new, very general ontology for annotating data, and tools to allow data managers in trial centres and hospitals to perform the annotation. We do not expect that annotation will be error-free, and the annotation-integration process is designed with continuous revision by data managers and p-medicine curators.

Provenance and reproducibility

The warehouse will be used for analysis that may lead to journal publications and systems for clinical decision making. It is important that these outcomes are justifiable, including the information which leads to them. We will store information on the source of all data using the Open Provenance Model (4). We will also allow all query and analysis operations on the warehouse to be performed on a particular (historical) version of the data, not just the latest version. This is so that analyses can be repeated and checked during review of a journal article or trial of a new procedure.

Existing solutions

Several solutions already exist for the collection of clinical data for analysis: however they all fall short in some respect when assessed against the criteria above. IMENSE (5) is a platform for managing clinical data created for a previous EU project, ContraCancrum (6). Whilst it provides a secure platform and a method for integrating data, p-medicine requires ontology-based integration which is less dependent upon creation of a specific schema. Additionally it does not support analysis on a specific historical snapshot of the data. Commercial solutions developed by companies such as IDBS (7) and Aridhia (8) have the same constraints, and in addition they are mainly focused upon ongoing management of patients. This is considered to be a separate concern in p-medicine, where data is occasionally pushed to the warehouse for separate analysis tasks.

Bibliography

1. *The ObTiMA System - Ontology-based Managing of Clinical Trials.* **Holger Stenzhorn, Gabriele Weiler, Mathias Brochhausen, Fatima Schera, Vangelis Kritsotakis, Manolis**

Tsiknakis, Stephan Kiefer and Norbert Graf. Cape Town, South Africa : s.n., 2010. Proceedings of the 13rd World Congress on Health (Medical) Informatics (Medinfo 2010).
2. Center for Data Protection. [Online] [Cited: 15 3 2012.] https://cdp.custodix.com/.
3. Health Level Seven International. [Online] [Cited: 15 3 2012.] http://www.hl7.org/.
4. *The Open Provenance Model core specification (v1.1).* **Luc Moreaua, Ben Clifford, Juliana Freireb, Joe Futrellec, Yolanda Gild, Paul Grothe, Natalia Kwasnikowskaf, Simon Milesg, Paolo Missierh, Jim Myersc, Beth Plalei, Yogesh Simmhanj, Eric Stephank and Jan Van den Bussche.** 6, s.l. : Future Generation Computer Systems, 2011, Vol. 27.
5. *IMENSE: An e-infrastructure environment for patient specific multiscale data integration, modelling and clinical treatment.* **Stefan J. Zasada, Tao Wang, Ali Haidar, Enjie Liu, Norbert Graf, Gordon Clapworthy, Steven Manos, Peter V. Coveney.** s.l. : Journal of Computational Science, 2011, Vol. In Press.
6. ContraCancrum. [Online] [Cited: 15 3 2012.] http://www.contracancrum.eu.
7. IDBS analytics software. [Online] [Cited: 15 3 2012.] http://www.idbs.com/.
8. Aridhia Multi-disciplinary healthcare informatics. [Online] [Cited: 15 3 2012.] http://www.aridhia.com.