MINING ADVERSE DRUG REACTIONS WITH E-SCIENCE WORKFLOWS

V. Curcin[1], M. Ghanem[1], M. Molokhia[2], Y. Guo[1], J. Darlington[1]
[1]Department of Computing, Imperial College London, United Kingdom
[2]London School of Hygiene and Tropical Medicine, London, United Kingdom
e-mail: vc100@doc.ic.ac.uk

Abstract- **In this paper, we describe the challenges faced when conducting large-scale Adverse Drug Reaction (ADR) studies by mining large-scale primary care databases. We provide a high-level view of the design of a generic informatics infrastructure that is needed to address such challenges. We also present our realization of such an infrastructure using a re-usable workflow-based e-Science middleware, and describe our experiences using the system based on studying the risks associated of statin induced myopathy and myalgia from UK-based data sources.**

Keywords – **ADR, workflows, e-Science, methodology, case-control, case-crossover.**

I. INTRODUCTION

Adverse drug reactions represent a worldwide problem. In the UK, errors in drug prescribing and ADRs cost the National Health Service (NHS) as much as £1.1 billion in 2002 [1]. A forthcoming study by UK think-tank Compass estimates that the figure in 2008 is £1.9 billion [2]. In the United States, ADRs are responsible for 5.8 percent of all hospital admissions [3]. Globally, 3.7 percent of patients admitted with ADRs die as a consequence [4].

The authors have developed a novel methodology for modelling and understanding ADRs by mining existing large primary care data collections. We have already applied the methodology to study the risks associated of statin induced myopathy and myalgia in the period 1991-2006 using two independent care databases. The results of the study from a clinical perspective have been reported in [9]. The approach is generic, and can be easily applied in different contexts.

In this paper, we describe the challenges faced when conducting such large-scale ADR studies and provide a high-level view of the generic informatics infrastructure needed to address them. We also present our realization of such an infrastructure using a re-usable workflow-based approach. In Section II, we provide the background on ADR studies and the informatics challenges they pose. In Sections III and IV, we describe the generic data model for such studies and the suite of data analysis methods that needed. In Section V, we describe the prototype of our system. In Section VI, we demonstrate how it has been used to conduct our first study. In Section VII, we summarize our work.

II. BACKGROUND AND CHALLENGES

Detection of associations between drugs and adverse events, also known as pharmacovigilance, currently relies on three main approaches: clinical trials, prescription event monitoring and spontaneous reporting. All three approaches have limitations. Clinical trials, even the ones on a large scale, may fail to detect rare adverse events, for example, recent studies on glitazones and risk of CVD and heart failure suggested there may be adverse effects associated with these drugs based on meta-analyses of trial data involving thousands of patients. At the same time, individual trials had not consistently shown the same risks. Prescription event monitoring (PEM) seeks to detect an excess risk of adverse events in users of newly marketed drugs, but is limited by relatively short follow-up time and by the difficulty of selecting appropriate control. Spontaneous reporting systems, such as UK's Yellow Card system, suffer from low reporting rates, typically less than 10%.[4] An additional limitation of all existing pharmacovigilance systems is that they can never detect unsuspected protective effects that a drug may have on risk of an apparently unrelated disease.

Mining primary care databases present a novel approach to the challenge of investigating ADRs. Data stored in such databases provides sufficient breadth to design bias-free control sets for the study based on full historical data and follow-up. Furthermore, the confounding factors may be amended to the study at any point, since they are typically already present in the data. Two basic study types for investigating ADRs are cohort studies and case-control studies. In the former, two groups of patients are followed through to see which one will develop an adverse effect (AE). In the latter, patients known to have developed an AE are taken as the case set, and a control set is defined, matching them by known confounding factors – age, gender and other study-specific ones. The success and validity of this data mining approach relies on many factors.

First, and foremost, an important issue is the availability of enough high quality historical primary care data to cover the specific conditions being studied (in terms of patient numbers as well as time periods), especially if the condition is relatively infrequent. In practice, and since the analysis is conducted retrospectively, the data required for a particular study may need to be collected from multiple sources. For example, our first study, for detecting statin associated myopathy ADR in 93,831 patients in the period 1991-2006 required use of two independent primary care databases – The Health Improvement Network (THIN), and MediPlus. The design and computational time required for processing the data to set-up the study and then conduct the analysis, by using traditional desktop statistical analysis tools (such as STATA, SPSS or SAS) is in order of weeks.

The first key challenge is thus *developing effective methodologies and systems* that reduce such long cycles that are incurred in the design of the study and executing the data processing and computational steps. Moreover, the scientific validity of the approach requires the reproducibility of the study and results using the original data sources, as well as allowing other researchers to verify the results on new collections. The second key challenge is thus *simplifying verification and reproducibility of the studies* through the use of an accessible representation of the data processing

and analysis steps and through enabling the re-use of the same executable methods by other researchers.

Namely, from an informatics perspective, what is needed to address these challenges are:

- A common data model for the ADR data analysis, allowing mapping of the analysis itself between different data sources and collections.
- A standardized way for representing the data processing analysis so that it can be easily modified and reused.
- A simplified framework that provides simplified abstractions for accessing multiple large data collections and for conducting the analysis over high-performance resources.

We address each of these challenges in the following sections.

III. ADR Data Model

A. Schema

An architecture for mining adverse drug reactions must rely on a vendor-neutral schema that is simple and efficient enough to provide the functionality needed for the analytical modules while being able to deal with the range of existing primary care data collections. The schema itself should also be independent of the coding used in the databases and codes used may also vary across studies. For example, READ and OXMIS codes are generally used for clinical events in the UK, as well as being the basis for SNOMED-CT ontology [6]. HL7 Reference Information Model [7] is popular in the United States and worldwide.

A high-level description of our chosen schema is shown in Figure 1. The central entity is that of a patient. Patient is uniquely identified by one or more attributes, with all non-time dependent information present in the table, such as gender, date of registration, etc.

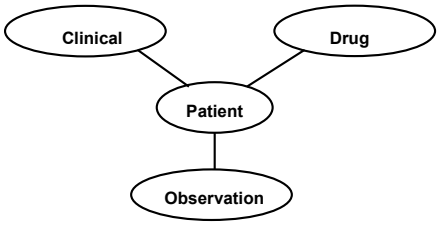

Fig. 1: Data model for ADR analysis.
Three types of events are all linked to the patient entity.

All other data in the schema is event-based, containing patient identifier and event date. Three distinct entities are present: drug prescriptions, clinical events and observations.

- Drug prescriptions represent any event in which a new or existing drug is prescribed to the patient, with dosage information.
- Clinical events represent instances where the medical practitioner asserts that the patient suffers from a certain condition.
- Observations capture all measurement events, including BMI, blood pressure, creatinine levels. These are typically separate from clinical events since they have associated with them a categorical or numerical value.

It is important to note that this abstract schema is optimized for analysis, not knowledge capture, and ignores issues such as reliability of information, data changing over time, conflicting data and inconsistent coding. Where necessary, these are resolved in the analysis itself.

B. Modelling Events

When conducting a study, we define a subset of events that are of interest to us. These event criteria may contain a presence of an event (prescription, clinical or observation) or a particular combination of events, such as occurrence of an event within 6 weeks of another event. Therefore, we distinguish between:

- Absolute events: Events that have occurred at a particular time and are used as selection criteria.
- Relative events: Events that are defined in terms of temporal constraints, e.g. occurring before, after or within a certain window of another event.

Specification of the events for a study is then a composition of multiple criteria that have to be satisfied. The common data model allows us to represent the criteria in the database-independent way and modify them during the analysis if needed.

C. Database Mapping

Once the schema is designed and the associated database constructed, the next step is to import and map data from existing data collections into the databases. This process, which occurs only once at the beginning of a study, can be implemented using either SQL code, or alternatively ETL (Extract-Transform-Load) tools. If the underlying storage for the new schema is in a relational database, indices are constructed on the patient and event identifiers. Depending on the study, additional indices may be constructed for certain attributes of any entity. The implementation we use in this paper is described in Section V, and is performed using standardized workflows that transform the original schema into the format required for the analysis. So far, we have used this schema successfully using data from both the UK's The Health Information Network (THIN) and General Practitioners' Research Database (GPRD) [5], as well as a number of custom checkouts from systems such as Vision and EMIS.

IV. ADR Study Design

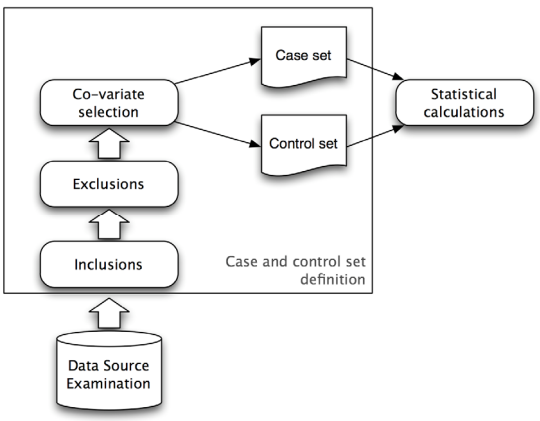

Fig. 2: ADR Study design steps

The design of ADR studies typically involves three main steps. These steps, summarized in Figure 2, are:

978-1-4244-2695-9/08/$25.00 ©2008 IEEE

1. Data examination: Investigation of data set properties, with visualisation of key attributes to identify potential confounders, factors which may be associated with either of the events observed.
2. Case and control set definition: Specifying the criteria for data subsets to be used in the analysis. Typically utilizing event criteria and matching of confounders – attributes that are required to have identical distributions in both sets.
3. Denominator calculation and statistics: These are study specific, but may include odds ratio, risk ratio, cumulative hazard, Kaplan-Meier test and hazard ratios [8].

The case-control specification consists of the following:

- Inclusion criteria: These define patient properties and event requirements for the entries to be included in the study, e.g., diabetes sufferers over the age of 55.
- Exclusion criteria: These define the patients who are not included in the analysis for having certain properties that would bias the study. For example, HIV sufferers are on retro-viral drugs that have known side effects that can interfere with the study.
- Covariate selection: This defines the attributes from the patient schema, or aggregated attributes from the event data, that are of interest in the study. For example, patient's smoking status, or the last creatinine measurement before an event:
- Case-set definition: This defines the selection of patients or events that are of interest, following inclusion, exclusion and containing all the needed covariates.
- Control-set definition: Based on the case selection, a control set is established, potentially matched for confounders that could affect the study, such as age, gender and any other study-specific ones.

During the implementation of the ADR study, each of the steps needs to be implemented using a combination of SQL code and statistical analysis routines. In the following section we present the design of our integrated environment that supports the implementation of such a system.

V. ADR Mining Environment

Figure 3 provides a high-level view of the ADR Mining Environment based on the Discovery Net workflow model [10] for coordinating computational services. The implementation uses the InforSense[1] KDE workflow system. It is the properties of the workflow model, rather than a particular tool, that enable us to address the key challenges highlighted in Section I for mining ADRs from multiple primary care databases. Our choice of the InforSense platform is governed by the maturity of the tool, however, conceptually, there are no major factors that prevent implementing the environment using other systems such as Taverna [11] or KNIME[2].

Briefly, a workflow provides an abstract description of data processing, statistical analysis or data mining steps required for executing a particular task and the information flow between them. Each workflow is represented as a graph with nodes as data processing activities. The arcs joining the nodes denote the dependency relationship between the activities in terms of data flow or control flow operations. A workflow system provides the mechanisms that enable users to access and integrate data from remote sources. It also provides the mechanisms for orchestrating the execution of computations over remote servers. For data mining applications, the workflow paradigm is a natural way of describing and managing distributed data mining tasks – data integration, processing and analysis.

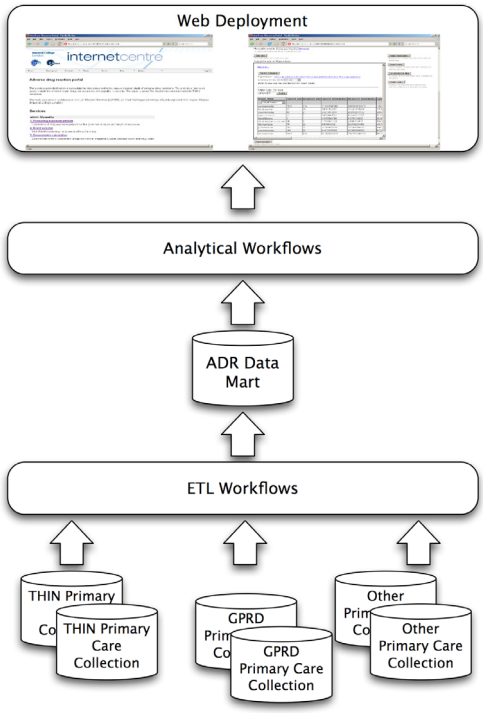

Fig 3: Overview of the ADR Mining Environment

The first step in Figure 3 describes how multiple data collections are transformed into the analytical schema using ETL workflows, such as the one in Figure 4. The aggregated data mart is then used as a data source for the analytical steps, also implemented using workflows, which are then further packaged as a set of portals and web services that are presented to the user.

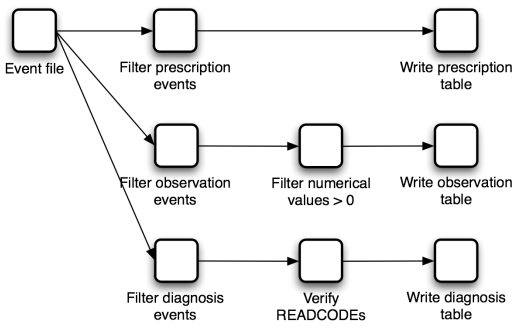

Fig 4: Portion of an ETL workflow for generating the schema

[1] www.inforsense.com
[2] www.knime.org

978-1-4244-2695-9/08/$25.00 ©2008 IEEE

VI. CASE STUDY

In this section we present an evaluation of the ADR Mining Environment in light of our experiences of using it in a study of risks associated with statin-induced myopathy and myalgia**.**

A. Summary of the clinical study and results

The clinical study itself has been reported in [9]. It is based on the use of a case-crossover design in detecting statin- and fibrate- associated myopathy ADR in 93831 patients, using two independent sets of primary care data from 1991-2006. The risk was analysed by drug class, by disease code and cumulative year and explored different cut-off exposure times and confounding of myopathy by temporality.

In the case-crossover design used, the control set consisted of the same patients in the time periods when they were not exposed to the drug observed. Therefore, we were observing individual events in the patient timelines, and classifying them as exposed or unexposed, with the unexposed serving as the control ones. The advantage of this approach is that basic confounding is automatically accounted for since the same patients are being analysed.

The results of the study indicate that using a 12 and 26 week exposure period, large risk ratios (RR) are associated with all classes of statins and fibrates for myopathy: RR 10.6 (9.8–11.4) and 19.9 (17.6–22.6) respectively. At 26 weeks, the largest risks are with fluvastatin RR 33.3 (95% CI 16.8–66.0) and ciprofibrate (with previous statin use) RR 40.5 (95% CI 13.4–122.0). AT 12 weeks the differences between cerivastatin and atorvastatin RR for myopathy were found to be significant, RR 2.05 (95% CI 1.2–3.5), and for rosuvastatin and fluvastatin RR 3.0 (95% CI 1.6–5.7). After 12 months of statin initiation, the relative risk for myopathy for all statins and fibrates increased to 25.7 (95% CI 21.8–30.3). Furthermore, this signal was detected within 2 years of first events being recorded. This data suggests an annual incidence of statin induced myopathy or myalgia of around 11.4 for 16, 591 patients or 689 per million per year.

B. Implementation Summary

The steps of the study have been implemented using a number of reusable computational services implemented using the workflow model. The data from the THIN and GPRD databases have been imported into the ADR Mining Environment based on using ETL workflows as described in the previous section. A number of analysis services have also been implemented.

The first service provides an automated summary of the data set for the properties selected and is used by the researcher to achieve full understanding of the data. The user submits the parameters, and the underlying workflow is adapted to produce the required data. The next service defines the case and control events observed in the study. The workflow implemented for the case-crossover study consists of five steps:

- Medical exclusions: Taking the list of known medical conditions that may cause myopathy, the patients suffering from those conditions at any point in their event histories are excluded from the event list.
- Drug exclusions: Similarly, some drugs are known to cause myopathy as a side effect of the treatment. Patients on those drugs are excluded from the event list. Exclusion workflow is shown in Figure 5.
- Exposure period calculation: The prescription events are extracted, matching the set of drugs that are analysed - in our case statins and fibrates. Some basic data cleaning is performed to make sure that events are usable, such as checking that quantity is non-zero and prescription events are within the patient registration period. Exposure end date is then defined as being 26 weeks after the start of prescription, defining a set of prescription windows.
- Event selection: First occurrences of myopathy events are extracted from the table of all medical events. Each event is assigned a unique identifier.
- Event exposure calculation: Two data sets are joined, and it is observed for each event if it falls into one of the exposure periods. If so, it is marked as exposed, otherwise it is unexposed. The exposed set constitutes the case set and the unexposed serves as control.

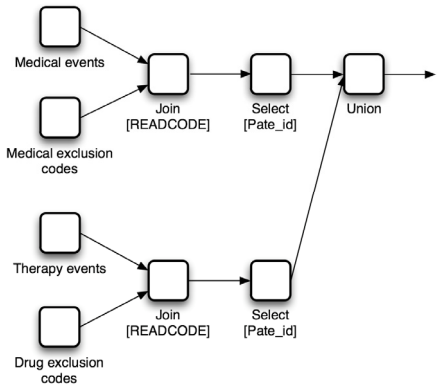

Fig 5: Concrete workflow for patient exclusions

Denominators in the study are taken to be the time from start of exposure until the event, which are grouped by each drug category, disease code and year. Rate ratios are produced to determine the significance of exposed against unexposed events in each grouping.

TABLE I:
RATE RATIOS FOR 12 WEEK EXPOSURE PER DRUG TYPE

Class of Drug	**Exposed events**	**Un-exposed events**	Rate Ratio 12 weeks	Standard error
Atorvastatin	1170	314	8.3 (7.4-9.4)	0.06
Cerivastatin	45	21	17.0 (10.1-28.5)	0.26
Fluvastatin	79	22	21.5 (13.4-34.4)	0.24
Pravastatin	313	78	15.7 (12.3-20.1)	0.13
Rosuvastatin	108	29	7.1 (4.7-10.7)	0.21
Simvastatin	1519	404	9.5 (8.5-10.6)	0.06
All statins	**3234**	**868**	**10.0 (9.3-10.8)**	0.04

The outputs of the study were the rate ratios and error rates calculated for the specified exposure period (12,26 and 52 weeks) and per drug code, disease code and year, producing

978-1-4244-2695-9/08/$25.00 ©2008 IEEE

nine tables in all. The example of a result for 12 weeks' exposure per drug is presented in table 1.

C. Web Portal Deployment

The analytical workflows used in the study have also been deployed as a web portal that presents a front-end for the researchers who want to investigate the data further or vary some of the parameters of underlying workflows. The exposed parameters include the disease and drug codes used for determining relevant diagnosis and prescription events, as well as the length of time between them that classifies as exposure. For example, the workflow performing the denominator calculation, the user sets the required parameters, and submits the workflow for execution. The front page of the ADR Portal, as deployed on HPC resources of Internet Centre at Imperial College is shown in Figure 6.

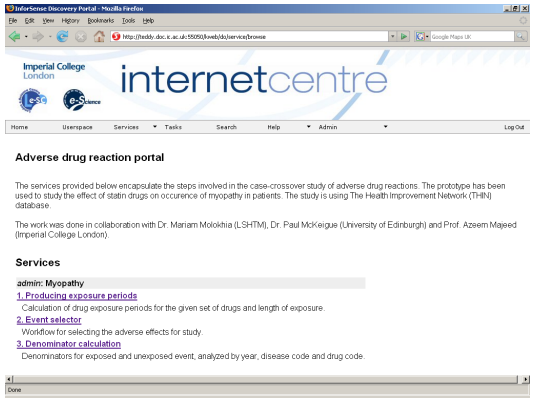

Fig 6. Portal front page listing available services www.internetcentre.imperial.ac.uk

VII. Summary and Conclusions

In this paper, we presented our experience of designing an extensible and re-usable environment for conducting Adverse Drug Reaction (ADR) studies based on mining large-scale primary care databases. The approach is characterized by:

a) the design and use of an ADR data mart and associated schema;
b) the use of ETL workflows for populating the data mart from the primary care databases;
c) the use of analytical workflows for constructing re-usable statistical analysis and data mining services; and
d) the deployment of the analytical services through a web portal interface for access by clinical researchers.

We had two key motivating factors behind designing our environment. Our first motivation has been to reduce the long design and execution cycles required when conducting these studies using traditional desktop data analysis tools (such as STATA, SPSS or SAS). Using these traditional tools the cycle for the clinical researcher is typically in order of weeks for each study since their use requires specialized expertise when manipulating the data, and high performance computing resources when conducting the data processing and reuse is rare. Our second motivation has been to simplify the verification and reproducibility of the studies through the use of an accessible representation of the data processing and analysis steps and through enabling the re-use of the same executable methods by other researchers. Our system effectively addresses these two goals in two ways.

Firstly, in contrast to traditional approaches, applying our ADR mining environment on new ADR studies using other data sets of the same scale by the clinical researchers requires only several hours, as opposed to weeks. This has been tested and verified in further studies that are currently being prepared for publication.

Secondly, the analysis routines are available as parameterized executable services through a portal interface enabling other researchers to verify and repeat the analysis easily using either the same data sets, or new ones. Furthermore, the representation and storage of the ETL steps and analysis steps as visual workflows enhances their understandability and the ability of re-purposing them for other studies.

The approach presented here is generic and can be implemented using a number of workflow system infrastructures and, thus, is not tied to any particular system or tool.

References

[1] L. Eaton, "Adverse reactions to drugs increase," British Medical Journal, vol. 324, no. 8, 2002.

[2] Compass, "Adverse drug reactions wastes NHS £2bn", 2008 [Online] Available: http://www.compassonline.org.uk/article.asp?n=1551

[3] N. Muehlberger, S. Schneeweiss, and J. Hasford, "Adverse drug reaction monitoring – cost and benefit considerations part I: frequency of ADRs causing hospital admissions," Pharmacoepidemiological Drug Safety vol. 6, supp. 3, pp S71-S77, 1997.

[4] M. Stephens, "Introduction," in: J. Talbot and P. Waller, ed. "Stephens' Detection of New Adverse drug Reactions 5th ed," John Wiley & Sons Ltd, pp. 10-17, 2004.

[5] L. Wood and C. Martinez, "The general practice research database: role in pharmacovigilance," Journal of Drug Safety, vol. 27, pp. 871-881, 2004.

[6] K. Giannangelo and L. Berkowitz, "Snomed CT helps drive EHR success," Journal of American Health Information Management Association, vol. 76, no. 4, pp. 66-67, April 2005.

[7] R. Dolin et al, "The HL7 clinical document architecture," Journal of American Medical Informatics Association, vol. 8, no. 6, pp. 552-569, 2001.

[8] S. Evans, "Statistics: Analysis and Presentation of Safety Data," in: J. Talbot and P. Waller, ed. "Stephens' Detection of New Adverse drug Reactions 5th ed," John Wiley & Sons Ltd, pp. 301-328, 2004.

[9] M. Molokhia, P. McKeigue, V. Curcin, and A. Majeed, "Statin Induced Myopathy and Myalgia: Time Trend Analysis and Comparison of Risk Associated with Statin Class from 1991–2006," PLoS ONE vol. 3. no. 6, 2008.

[10] A. Rowe, D. Kalaitzopoulos, M. Osmond, M. Ghanem, and Y. Guo, "The Discovery Net system for high throughput bioinformatics," Bioinformatics, vol. 19, supp 1, 2003.

[11] T. Oinn et al, "Taverna: a tool for the composition and enactment of bioinformatics workflows," Bioinformatics, vol. 20, no. 17, pp. 3045-3054, November 2004.

978-1-4244-2695-9/08/$25.00 ©2008 IEEE