LINK TO ORIGINAL ARTICLE

Broadening access to electronic healthcare databases

M. Soledad Cepeda, Victor S. Lobanov, Michael Farnum, Rachel Weinstein, Peter Gates, Dimitris K. Agrafiotis, Paul Stang and Jesse A. Berlin

Electronic health databases have become extremely valuable resources for pharmacoepidemiological and translational research, as noted in the recent Perspective article, Beyond debacle and debate: developing solutions in drug safety (*Nature Rev. Drug Discov.* **8**, 775–779; 2009)[1]. Through these databases, researchers can gain a better understanding of the short- and long-term impact of exposure to drugs and devices, identify populations at risk for adverse effects, estimate the prevalence and natural history of medical conditions, and assess drug utilization across different demographic groups[2–4].

However, the daunting size and complexity of these databases have made them inaccessible to all but a few experts with advanced data-management and statistical programming skills. Although simpler interfaces have begun to emerge[5,6], a truly integrative approach that combines convenient access with advanced analytics is still lacking.

Our Advanced Biological and Chemical Discovery (ABCD) system has demonstrated the potential of such an approach in drug discovery[7,8]. Recently, we extended our toolset and the underlying design principles to the field of outcomes research. The solution consists of four major components.

The first is a high-performance relational database with a specialized data organization and indexing strategy, which decreases the data processing time by orders of magnitude compared with traditional, file-based approaches. To facilitate cohort selection by distinct criteria such as exposure to medications or clinical diagnoses, it creates separate copies of the same data tables with different clustered indices (BOX 1), thus greatly improving performance.

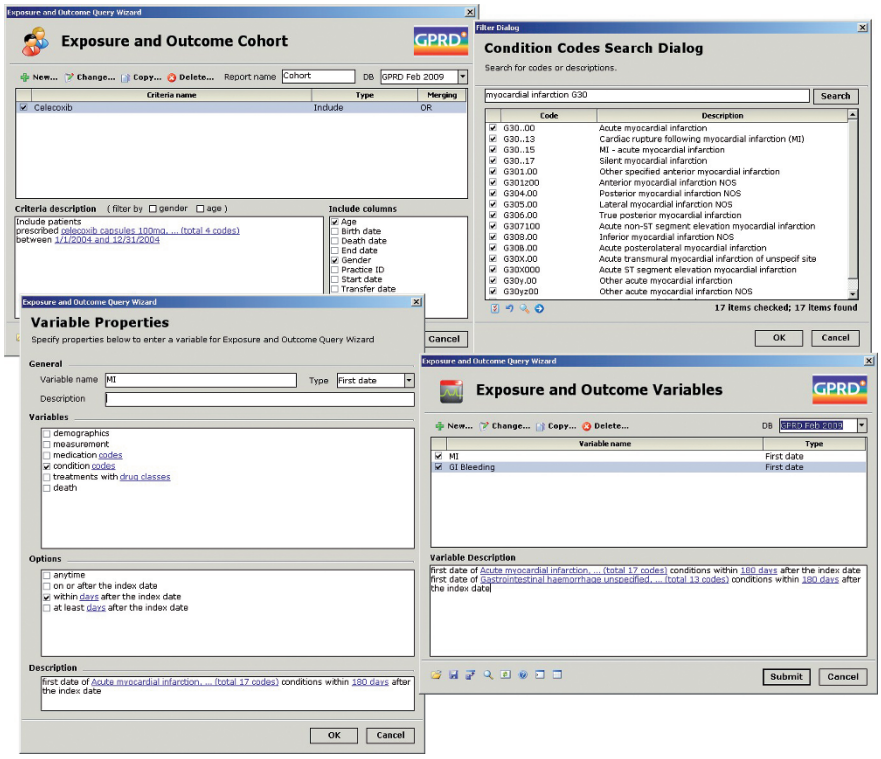

Figure 1 | **The wizard interface.** A mosaic of the dialogue boxes used to define the inclusion or exclusion criteria, the demographic characteristics of the subjects that the user would like to retrieve and the time windows for the events of interest is shown. In this example, we are using the UK's General Practice Research Database (GPRD) and selected subjects that have been exposed to celecoxib. From this cohort of subjects, we sought to identify how many developed myocardial infarction or gastrointestinal bleeding within 6 months after the first prescription of celecoxib in 2004 (the index date). The user can perform a text search to find the codes for the outcomes, as seen in the third dialogue box, the codes for the medications or can import an external file.

The second is a graphical interface that allows the user to define complex queries with multiple inclusion or exclusion criteria through a series of dialogue boxes, employing language and identifiers that are familiar to epidemiologists (FIG. 1). This interface, which was implemented as a plug-in to Third Dimension Explorer (3DX)[7], eliminates the need for programming expertise by hiding complex Structured Query Language (SQL) query generation, execution and post-processing.

The third is 3DX itself, an intuitive data analysis and visualization environment for interactive plotting, analysis, filtering and manipulation of the query results[7] (FIG. 2). Follow-up queries and more elaborate statistical assessments can be invoked through specialized routines or services available from within 3DX, or by exporting the data into an external statistical package.

The final component is a rigorous evaluation of the tool's performance, usability and versatility for different users and study types, and a strict validation of the results by comparing them with the current gold standard: customized SAS routines that are specific to the issue at hand. We found that queries requiring many hours

Box 1 | **Indices**

Indices are data structures that improve the speed of operations on a database table. Clustering is an indexing technique that re-arranges the raw data blocks in a way that matches the index, resembling an address book in which entries are ordered by last name. This technique leads to dramatic improvements when the data are accessed sequentially, in the same or reverse order of the clustered index, or when a range of items are selected, which is typical in cohort selection.

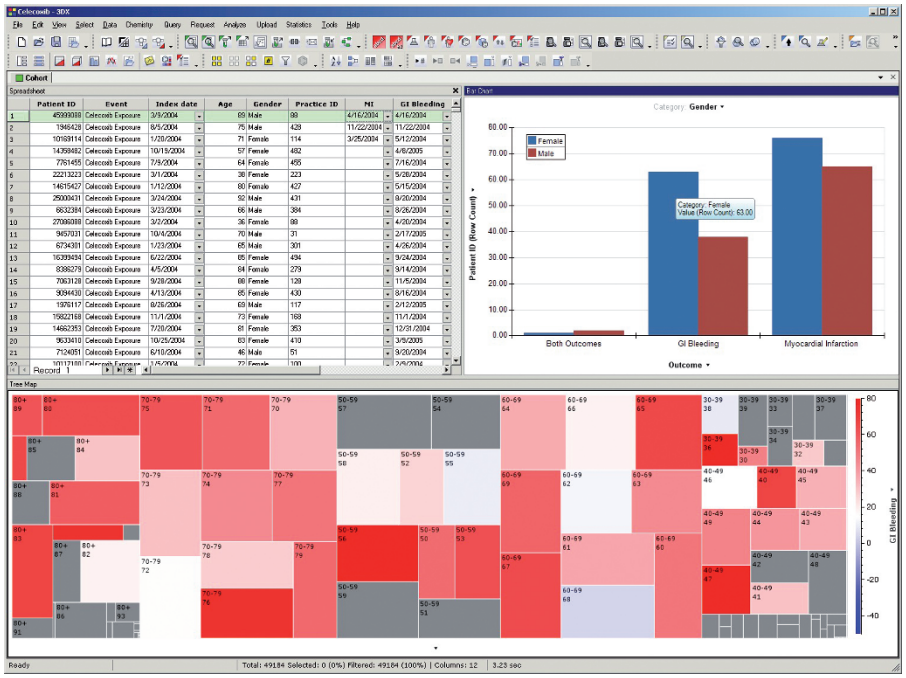

Figure 2 | **Third Dimension Explorer.** The screenshot shows the list of subjects exposed to celecoxib in 2004, their age and gender, and for the subjects who developed myocardial infarction or gastro-intestinal bleeding the date that the general practitioner registered the events. The application can also perform basic statistical analysis. The distribution by gender of subjects exposed to celecoxib in 2004 who developed myocardial infarction or gastro-intestinal bleeding can be seen in the histogram. From a total of 9,526,386 subjects present in GPRD, 49,184 were exposed to celecoxib in 2004. This query was executed in 30 seconds. The query for the outcomes was executed in 5 seconds.

or even days of painstaking programming and data processing were constructed and executed within a few minutes, yielding identical results.

Examples of queries that are greatly simplified by the current interface include retrieving records of patients who have been administered a specific medication within a particular time period and have developed specific undesirable adverse events; identifying subjects who have experienced a particular event of interest, regardless of exposure history; developing an inception cohort of subjects with a particular diagnosis; and a number of other largely descriptive analytic routines.

Our future plans include the development of interfaces that simultaneously query multiple databases for the same investigation, as well as the direct linking of discovery, clinical, biomarker and health outcomes data to address translational hypotheses.

M. Soledad Cepeda, Rachel Weinstein, Paul Stang and Jesse A. Berlin are at the Department of Epidemiology, Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 1125 Trenton Harbourton Road, Titusville, New Jersey 08560, USA.

Victor S. Lobanov, Michael Farnum, Peter Gates and Dimitris K. Agrafiotis are at Informatics, Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 665 Stockton Drive, Exton, Pennsylvania 19341, USA.

Correspondence to: M.S.C. and V.S.L.
emails: scepeda@its.jnj.com; vlobanov@its.jnj.com

doi:10.1038/nrd2988-c1

1. Ray, A. Beyond debacle and debate: developing solutions in drug safety. *Nature Rev. Drug Discov.* **8**, 775–779 (2009).
2. Suissa, S. & Garbe, E. Primer: administrative health databases in observational studies of drug effects–advantages and disadvantages. *Nature Clin. Pract. Rheumatol.* **3**, 725–732 (2007).
3. Schneeweiss, S & Avorn, J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J. Clin. Epidemiol.* **58**, 323–337 (2005).
4. Strom, B. L. How the US drug safety system should be changed. *J. Am. Med. Assoc.* **295**, 2072–2075 (2006).
5. IMS Health. *PharMetrics Integrated Database* [online], <http://www.imshealth.com>
6. GPRD [online], <http://www.gprd.com/services/online.asp>
7. Agrafiotis, D. K. *et al.* Advanced Biological and Chemical Discovery (ABCD): Centralizing discovery knowledge in an inherently decentralized world. *J. Chem. Inf. Mod.* **47**, 1999–2014 (2007).
8. Kirkpatrick, P. The ABCD of data management. *Nature Rev. Drug Discov.* **6**, 956–957 (2007).

Acknowledgements
We would like to thank C. Confoy, L. Bulusu, G. Griffin, V. Ogay and R. Chihevski for their contributions to this project.