Safety in numbers

Traditional safety techniques developed for clinical trials and quantitative signal detection of spontaneous adverse events reports have their limitations. **Wayne Kubick** and **Sigfried Gold** report on a data mining approach which uses existing tools and techniques to visualise longitudinal health data

KEYWORDS: Pharmacovigilance; Spontaneous adverse events; Longitudinal data; Data mining; ICD9

Traditional drug safety practices in the pharmaceutical industry have relied primarily on the collection and analysis of safety data from clinical trials before and after approval, as well as passive surveillance based on the voluntary reporting of spontaneous adverse events (AEs) once a drug is approved for marketing. It has long been recognised that these data sources severely under-represent the population that is likely to be exposed to a given drug in the real world.[1] As a result, many government, academic and industry initiatives are being established to improve the process of safety management through active surveillance, often involving the potential to use longitudinal health data as a resource for examining the safety of a drug in action among a broader, more representative patient population. While exploitation of these healthcare data sources will require analytical approaches beyond those now used for clinical trials and spontaneous report data, past experiences in these areas may still provide useful insights that can benefit the study of longitudinal data.

There has been a recent groundswell of interest in identifying new techniques for improving drug safety. During the May 2008 Post-Approval Summit at the Harvard Medical School, Mark McClellan, former US FDA commissioner and current director of the Engelberg Center for Health Care Reform at the Brookings Institution and chairman of the Reagan-Udall Foundation, cited the reauthorisation of the FDA Amendments Act (FDAAA), which will improve upon the 'current bifurcated, costly and time-consuming pre-market/post-market process' to provide a more continuous 'improvement lifecycle approach.'[2] Similar sentiments were echoed at subsequent events, including the Brookings Forum on Drug Safety and Post-Market Evidence[3] and the DIA annual meeting in June 2008.[4] While there are still opportunities to learn much more from clinical trials and spontaneous report data, the prospect of using real-world medical evidence has especially captivated the interest of researchers and regulators.

Limitations of traditional methods

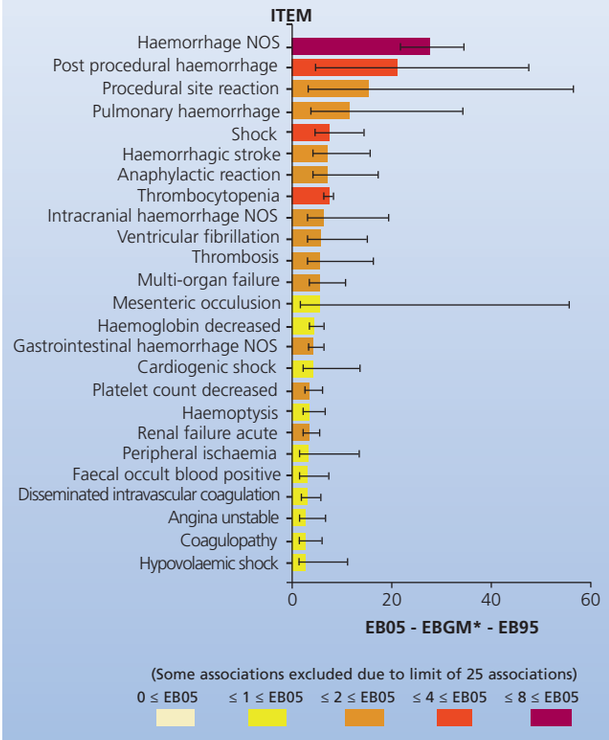
Although high-quality safety data are an essential output of the drug development process, clinical trials typically involve only a small percentage of the potential target patient population who have been carefully selected to meet specific conditions; thus potentially under-representing important subgroups such as women, the elderly and children.[5] Therefore, while clinical trials data are a key input to the regulatory approval process, it is insufficient to ascertain the full safety profile of a new drug product.

Similarly, while spontaneous report data provide clearly identifiable patients, drugs and events, and drive many fundamental pharmacovigilance processes, the data's utility is also limited by concerns such as underreporting,[6] uneven quality and lack of adequate exposure information.[7]

Some of these shortcomings have been traditionally addressed by conducting epidemiological studies once a drug is approved for marketing.[8]

Limitations of epidemiological methods

The Guidelines for Good Pharmacoepidemiology Practices[9] have been published by the International Society for Pharmacoepidemiology, These describe the trusted but highly labour- and time-



The bars in this figure illustrate a quantitative signal score for the combination of a particular drug with a series of diagnoses or events with confidence interval bars using the EBGM Bayesian statistic. The colours indicate the strength of the signal based on the lower 5% confidence limit

***EBGM Empirical bayes geometric mean**

Figure 1: Example of a graphical 'patient profile' display

intensive methods of pharmacoepidemiological research through the design and conduct of studies. However, as Dr Hugh Tilson of the University of North Carolina stated at the 2008 DIA annual meeting in Boston, progress is limited due to a critical shortage of trained epidemiologists, whom he called 'an endangered species'.[4] The recent surge of interest in further exploring longitudinal data is based on the assumption that it should be possible to exploit a broader variety of longitudinal data sources in new ways to discover relevant safety information. The processes of risk identification, analysis and verification based on longitudinal data sources promise faster results with great power and fewer resources – assuming the development of a robust and efficient analytical toolset to support these purposes.

While there are many similarities between clinical trials and epidemiology studies, such as the selection of patient treatment groups in clinical trials versus epidemiological cohorts and the availability of drug exposure information, the applicability of quantitative signal detection concepts to the study of longitudinal data may not be quite as self-evident. But, as this paper will demonstrate, some of the experience gained in the development of tools and techniques for detecting spontaneous report data may also be applied to longitudinal data sources.

Clinical trial coding systems	
MedDRA	Medical dictionary for regulatory activities
ICD9	International classification of diseases, 9th revision
CPT	Current procedural terminology
SNOMED	Systematic nomenclature of medicine
LOINC	Logical observation identifiers names and codes

Healthcare data challenges

While the FDAAA has mandated that the FDA be able to access 25 million patient records for drug safety analysis by 2010 and 100 million records by 2012,[10] the challenges are daunting. Since the logistics, ownership, access and privacy issues surrounding the accumulation of such a vast amount of data make it impractical to gather the data into a single physical warehouse, the FDA's Sentinel Network Project[11] has been considering a federated data warehouse. Under this approach, data would remain with their owners – large insurers and healthcare systems – who would accept queries in some standard form and feed summary results back to Sentinel. (Other similar pilots are being sponsored by the likes of the Foundations for the National Institutes of Health and the eHealth Initiative Foundation). They face an imposing list of challenges, such as:

- How to get the owners of the healthcare data of one third of the US population to make their patients' medical histories available in a timely manner
- Defining a common, standard format for the data and common query specifications
- Setting policies to determine which governmental, academic, medical and pharmaceutical organisations could make use of the system
- Governance and sufficient funding for infrastructure, data services, researchers and tools
- Privacy and data protection restrictions, such as the US Health Information Portability and Accountability Act (HIPAA).

Beyond these, the prospect of mining or analysing patient medical histories for drug safety purposes raises a host of thorny problems even once data standardisation and access issues are resolved. For example, it's often difficult to get lists of in-patient and out-patient medications combined in a consistent, reliable format. Reducing the plethora of drug names and forms to a standard set of generic ingredient names is a challenge. It's difficult to identify and remove drugs that may be mentioned in a medical history but not prescribed or prescribed but not taken – or to establish is a patient is taking over-the-counter or prescribed drugs outside the given institution.

And greater difficulties arise when attempting to identify AEs in a stream of patient history data. Clinical data in the electronic medical record (EMR) relevant to drug safety may exist in the

form of diagnoses, symptoms, signs or procedures, most of which may be described largely as narrative, or in the form of lab results. Clinical trial and spontaneous report data generally code AEs in MedDRA, which is seldom used by EMR systems – except in the extremely rare case when a doctor submits an AE report directly through a specialised feature of the EMR system. While MedDRA's multi-axial hierarchical structure and support for Structured MedDRA Queries is directly intended to support drug safety analysis, other coding systems, such as ICD9, CPT, SNOMED, or LOINC (see Box on page 22) are intended to support other purposes such as claims processing, billing and health data interchange, and there is often no easy or reliable translation from one coding system to another.

While ICD9 codes are probably the most prevalent in longitudinal data, they are not ideally suited for drug safety analysis either. Beyond the difficulty of grouping ICD9 codes according to health outcomes of interest, like those recognised by Structured MedDRA Queries, ICD9 codes in a patient's record may provide a jumbled and inaccurate picture of actual clinical history.

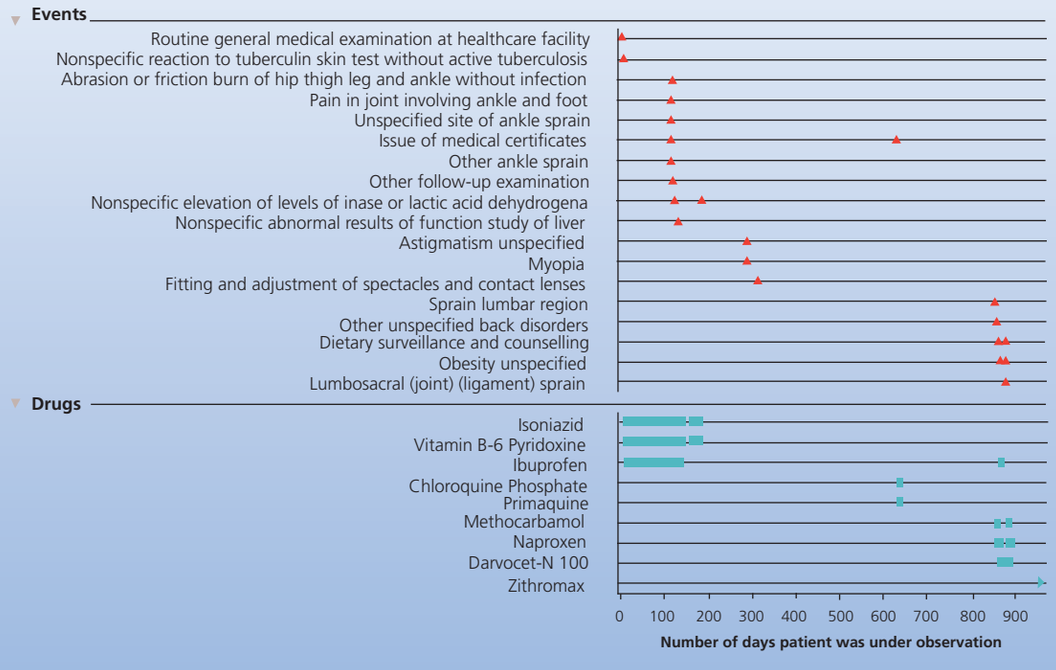
Another pervasive problem with ICD9 and other codes is the evolution of a diagnosis over time, as more information is obtained and/or a condition progresses. For instance, on day one the patient may be given a diagnosis of haematochezia (blood in stool). A colonoscopy on day eight, scheduled to determine the cause of the bleeding, may provide a new diagnosis of a polyp in the sigmoid colon, which in turn may be followed a few days later by a pathologic diagnosis of dysplasia. All of these will show up as different codes on different dates, even though they may represent a single medical event. If a medication is given in the middle of these dates, it may be difficult to determine whether each subsequent diagnosis was related to, and part of, the original event or resulted from the new drug being administered.

It is possible to transform a longitudinal healthcare record to mimic the data structure of a series of adverse event reports, so these can be further explored using other available tools

Mitigation of these and other data issues is a critical prerequisite to any sort of productive mining or analysis activity.

A data mining approach

As stated previously, one way to provide rapid visibility into longitudinal data to identify and verify safety signals is to capitalise on the experiences gained from the mining of spontaneous AE data. Such an approach can leverage certain existing tools and techniques already developed for quantitative signal detection, and by mimicking known safety data structures, may provide a familiar environment



The red triangles indicate when an event initially occurred (colours can indicate the severity of an event). The lower half of the figure illustrates the periods of time when prescriptions for specific drugs were assigned to the patient. In this particular case, the subject experiences abnormal liver results after initial administration of a drug and vitamin B-6 therapy.

Figure 2: An example of a patient profile for a patient experiencing a hepatic adverse event.

to risk managers already experienced in this area. This approach is most suitable for 'hypothesis-free' exploration of data and the generation of hypotheses for further testing, but it can also support more detailed exploration of the patient data that may comprise a signal.

In essence, an AE report includes one or more events of interest: the drugs the patient was exposed to at the time of the event(s), the indications for those drugs along with relevant medical history, concomitant medications and other reference information. While the field of EMRs and personal health records is rapidly evolving, current implementations typically include prescription data and diagnostic codes from claims databases supplemented with general patient data, such as demographics, labs and medical history. Where a spontaneous event report provides a list of drugs and MedDRA codes for a particular event at a point in time, longitudinal healthcare data, at least for now, provide for each patient a stream of prescriptions and diagnoses (represented as ICD9 codes) along with their associated dates.

In a spontaneous event report, a physician (or some other reporter) has already determined that the patient was exposed to the drugs listed at the time the event occurred. However, with healthcare data it is probably necessary to define the temporality of exposure – by including the length of the prescription plus a period, such as two weeks. It will also be helpful to define a 'wash-out' period at the beginning of each patient's longitudinal data or between drug/event occurrences so that existing conditions are not counted as AEs.

Despite these and other complexities, it is possible to transform a longitudinal healthcare record to mimic the data structure of a series of AE reports. Under this approach, the user sets parameters to determine how the transformation is performed. Such transformation parameters may include:

- Determining whether the 'report' is based on the first occurrence of a medical event or any occurrence of such an event (with some wash-out period), including drugs within a decreed time period, or whether it should be based on first or any occurrence of a drug including events within a designated time period
- Deciding whether the prescription must occur before the event or not
- Excluding certain drug/event combinations, particularly known indications.

Once the longitudinal data have been restructured into event reports, these can be further explored using other available tools that have been created for quantitative signal detection and management. These include:

- Performing data mining runs based on drugs or events of interest with demographic filters and stratification factors where desired
- Computing signal scores using signal detection algorithms, such as the MGPS empirical Bayesian algorithm, and sorted by score[12]
- Defining case series based on results which can be used as the basis for further data mining runs
- Using graphical visualisations to illustrate patterns or significant findings from the data, such as the following example based on spontaneous report data (See figure 1).
- Performing a drilldown analysis on the results and viewing specific patient records with a graphical 'patient profile' display (See figure 2).

The approach described above builds on existing software, techniques and experience of risk managers to enable a rapid, credible introduction to signal detection based on longitudinal data. While it does not necessarily take advantage of all of the richness in EMR data, such as the ability to account for length of exposure or to look for rechallenge/dechallenge events, it does allow for the identification of off-label uses and exclusion from signal exploration along with other indications. More importantly, it provides a familiar platform for risk managers to begin working with longitudinal data and to ask the questions that will motivate the development of a more robust set of tools optimised to fully exploit the potential of these data.

Conclusions

While traditional safety analysis and quantitative signal detection techniques developed for use with clinical trials and spontaneous report data have not always proved sufficient to minimise all drug safety risks among a broad patient population over an extended time, they provide a broad base of experience that may be applicable to longitudinal data sources as well. This paper discusses some of the opportunities and challenges in using longitudinal data for pharmacovigilance and risk management, and presents a specific example of how methods developed for quantitative signal detection of spontaneous report data may prove useful to safety researchers conducting hypothesis-free analyses. As a wide range of ambitious pilot projects using longitudinal data move into active use, other approaches for both hypothesis-driven and hypothesis-free analyses are likely to be developed, which may also benefit from past experiences with safety analysis and pharmacovigilance methods. GCPj

References

1. RP Van Manen, D Fram, W DuMouchel. 'Signal detection methodologies to support effective safety management', *Expert Opinion on Drug Safety,* 6(4), pp451–464, 2007.
2. C Varmazis. McClellan Envisions Lifecycle Approach to Drug Surveillance, Bio-ITWorld.com, 23 June 2008 (http://www.bio-itworld.com/ecliniqua/2008/06/23/post-approval-summit-mcclellan.html).
3. Brookings Institution Forum on Drug Safety and Post-Market Evidence (http://www.brookings.edu/events/2008/0613_drug_safety.aspx).
4. DIA annual meeting 2008. The Impact of FDAAA on Drug Safety, 24 June 2008 (http://www.diahome.org/NR/rdonlyres/7B62B60C-32EB-434B-963A-F9DA947B3493/1141/2008_multitrackplenarysessionflyer.pdf).