Drilling Down to Success

Designing better clinical trials that increase value and minimise risk for patients is a far from straightforward task. Mohammad Afshar and Lilia Abtroun of Ariana Pharma review the benefits of data mining and systematic hypothesis generation

With an overall success rate of 19 per cent for the top 50 pharmaceuticals in the US, the clinical approval rate remains extremely low (1). Although differences exist between therapeutic classes, small versus large molecules and in-licensed versus in-house, it is generally believed that improving the design of clinical trials should have a significant impact on success rates.

IMPROVING TRIAL DESIGN & THE SEARCH FOR THE BEST HYPOTHESIS

A typical clinical trial case report form (CRF) captures several hundred items of information per patient, and the data is usually recorded through multiple visits. One or more hypotheses are explicitly stated and the trial is often run to validate (or reject) the given hypothesis. For example, in a randomised, double blind drug-versus-placebo efficacy trial for a cancer drug, the efficacy of a particular treatment (or procedure) may be tested in terms of overall survival or event-free survival, as well as the cohort of patients exposed to the drug compared to the placebo. Stating whether the product significantly increases survival can be a relatively simple statistical test.

Often, it is when the test fails that sponsors rush into deeper analysis in an attempt to understand what went wrong and to discover why the initial hypothesis did not hold true with the population that was exposed to the therapy. Ideally, they would then come up with an alternative hypothesis which would be the basis of the design of a new trial. Should we exclude patients that have not been exposed to a previous treatment? Are there any other preconditions that are linked to increased response rates? Also as importantly, are there any specific populations that should be excluded or monitored closely due to higher risk? Given the vast amount of data that is collected multiple times, this exercise rapidly resembles searching for a needle in a haystack, and the clinician's intuition, although invaluable, is often not able to guarantee that all equivalent hypotheses have been checked.

Systematic analysis remains important for successful trials, even though its necessity may not be as acutely felt by the sponsor. When transitioning between phases, it is critical to be aware of alternative hypotheses which are consistent with the data. This knowledge can be used to fine tune the design of the next phase, minimising risk. This remains true at the end of Phase 3 where even in large clinical trials the number of patients enrolled is often orders of magnitude smaller than the population size that the product would be exposed to in real life. In addition, there is often retrospective analysis of failed (and successful) clinical trials for the repositioning of clinical candidates or existing marketed drugs (2).

The need for systematic analysis to identify new hypotheses has led to the development of data mining technologies, with application going well beyond life sciences. These tools are essential for the full analysis of trials that include genetic and genomic data, as well as translational medicine approaches where, beyond traditional hypothesis testing statistics, the challenge is to associate genomic, proteomic and metabolic data with clinical data, including patient characteristics, demographics and treatment history. A good illustration of the limit of traditional analysis is provided by the genome-wide association studies (GWAS); the availability of more than one million single nucleotide polymorphisms (SNPs) across the genome has led to a large number of GWAS studies. However, most SNPs discovered via GWAS have a small effect on disease susceptibility (3). Successful analysis has required a more holistic approach, systematically searching for combinations of multiple SNPs and patient characteristics to identify disease susceptibilities.

'ASSOCIATION RULES' MINING

A growing number of authors demonstrate the use of data mining tools to determine associations between drugs, genes or protein targets, and possible mechanisms of actions or therapeutic efficacy which could be used to refine or direct new clinical studies (4). Several examples in oncology illustrate how association mining can be used to determine the relationship between a cancer type or drug and the symptoms or efficacies observed, which in turn optimises clinical trial design (5).

'Association rules' play an essential role in data mining techniques. They are constituted of 'if/then' statements that help uncover relationships between seemingly unrelated data in a trial database or other information repository. An example of an association rule would be: "If a patient has been

previously treated and his cholesterol is high, he has 80 per cent chance of responding to the drug."

An association rule has two parts: an antecedent (if) and a consequent (then). Association rules are created by analysing data for frequent if/then patterns and using the criteria support, confidence and relative probability (or lift) to identify the most important relationships. 'Support' is an indication of how frequently the items appear in the database. 'Confidence' indicates how often the consequent is true when the antecedent is true. The 'lift' captures the increase in the probability of the appearance of the consequent when the antecedent is true, compared to its unconditional probability.

Association rules are useful for analysing and predicting patient behaviour. Although many technologies co-exist for extracting association rules, formal concept analysis (FCA) has attracted particular attention in the field in recent years. FCA is an unsupervised method of data analysis and knowledge extraction. By grouping variables into concepts, FCA overcomes dimensional limitations by identifying strong association rules in a systematic way. Each concept contains the information of many association rules, and strong associations can be easily extracted and interpreted. An FCA can integrate heterogeneous sources of data, examine complex interplay of different factors and is robust to missing data, thus making the use of all available information possible. As a result, unexpected association rules can be discovered systematically and without any pre-established assumptions (6,7).

RESCUING & DESIGNING BETTER CLINICAL TRIALS

In the life sciences, the pharmacovigilance community has been the earliest adopter of data mining technologies. Current adverse drug event (ADE) databases receive tens of thousands of reports each year, accumulating to millions of reports. Challenged by their vast size and complexity, the traditional manual case-by-case review by clinical experts has been complemented by more efficient methods consisting of using automated and quantitative data mining algorithms. These associations, also referred to as signals, are not necessarily true ADEs, but rather hypotheses that warrant further investigation to qualify them as credible. They allow evaluators to peruse the large volume of reports and focus their attention on potentially important safety issues (8).

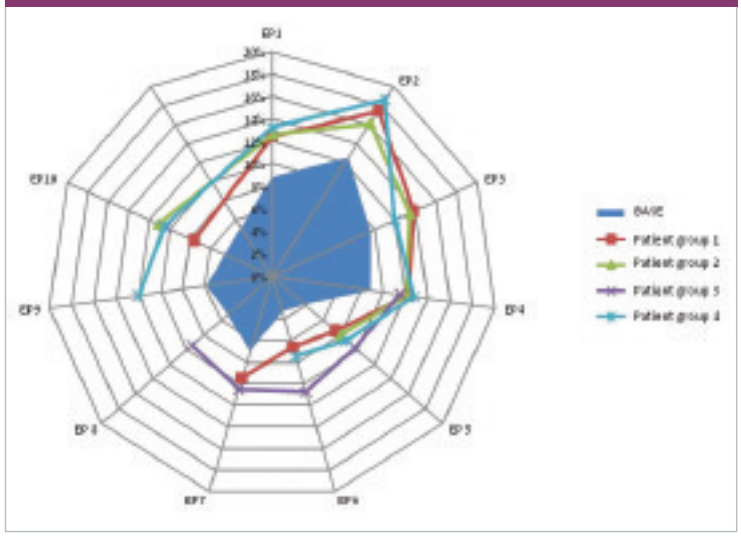
In contrast with pharmacovigilance, the use of data mining tools has not yet become a requirement for the analysis of clinical trial data, but there are clear signs that the industry is moving in that direction. Over the last seven years, several advanced association rule mining methods using FCA have been developed. The technology has been created in close collaboration with sponsors and regulatory agencies, and is actively used in clinical trial optimisation and biomarker investigation applications. It is able to systematically extract and quantify relationships between sets of patient characteristics and conditions or genes by describing the direction (causality) of the relationships. These associations are the source of the hypotheses that are the cornerstone of the design of any clinical trial. Responder sub-populations can be identified based on specific characteristics or phenotypes. This is not easily done using models generated by conventional classification methods, and can be a challenging task as small numbers of high responder profiles may be masked by larger groups of patients. Early detection of responder or non-responder profiles has a direct impact on the design of inclusion/exclusion criteria.

At present, many sponsors resort to data mining technologies in order to redesign and rescue a failed clinical trial. Although this appears to be leading to a growing number of successes, we would argue that the systematic search for association rules should be used earlier; in the initial design of clinical trials, particularly when transitioning between phases. Going beyond testing of the hypotheses put forward in the statistical plan, all patient characteristics and biochemical measures, along with any biomarker data, can be analysed in order to identify all responder sub-groups as well as all patient sub-groups with increased risk. This systematic analysis enables the clinicians to exhaustively identify all equivalent hypotheses. The clinicians can then explicitly select hypotheses, adding in their critical field knowledge, in order to maximise efficacy and safety.

An illustration is given by the following example. A sponsor performed a Phase 3 prospective, multi-centric, randomised, double-blinded, drug-versus-placebo trial with 2,300 patients. Three hundred baseline variables measured at inclusion and 10 endpoints were investigated. The results show statistically

"In the life sciences, the pharmacovigilance community has been the earliest adopter of data mining technologies. Current adverse drug event (ADE) databases receive tens of thousands of reports each year, accumulating to millions of reports. Challenged by their vast size and complexity, the traditional manual case-by-case review by clinical experts has been complemented by more efficient methods consisting of using automated and quantitative data mining algorithms"

Figure 1: Identification of patient sub-groups in a retrospective analysis showing the responses for multiple endpoints



Source: Ariana Pharma

significant drug-versus-placebo efficacy on endpoints one, two and three. The amplitude of the difference is moderate, leading to questions regarding patient benefits.

Before responding to the agency with additional information, the sponsor uses an associate rule mining method for exhaustive analysis of their data in order to identify large, significant and clinically relevant patient sub-populations maximising efficacy on multiple endpoints, in that syndromes improve rather than symptoms. The mining technology identifies large patient sub-groups (N>700) that show improved efficacy over eight endpoints. Using the output from the mining study, the clinicians choose the four most clinically relevant responder groups and provide the results as additional data for the successful registration of the compound (see Figure 1).

While data mining can improve success rates and minimise risk, it is important to emphasise that the design of the eCRFs also has an impact on the analysis that can be performed in subsequent stages. Tools such as associate rule data mining enable R&D experts to explore billions of combinations; however, that information needs to be present in the databases. We would therefore encourage sponsors to collect information beyond their initial hypotheses. Building on the wealth of information gathered, data mining technologies may also be able to identify unexpected relationships.

CONCLUSION

As both the safety and biomarker fields fuel the need for sophisticated data mining technologies to detect early strong signals from a sea of data, these technologies are also starting to be adopted by the clinical trial community. They positively impact the design and outcome of clinical trials. Often used in case of failure, and in uncertain transitions between Phase 2 and Phase 3, we believe they will soon become required use. Combining systematic analysis using data mining technologies with the critical insight of clinicians will help increase success rates and diminish risk in clinical trials.

References

1. DiMasi JA, Feldman L, Seckler A and Wilson A, Trends in risks associated with new drug development: success rates for investigational drugs, *Clinical Pharmacology and Therapeutics* 87: pp272-277, 2010
2. Padhy BM and Gupta YK, Drug repositioning: Re-investigating existing drugs for new therapeutic indication, *J Postgrad Med* 2: pp153-160, 2011
3. Moore JH, Asselbergs FW and Williams SM, Bioinformatics challenges for genome-wide association studies, *Bioinformatics* 26: pp445-455, 2010
4. Galustian C and Dalgleish AG, The power of the web in cancer drug discovery and clinical, *Cancer Informatics* 9: pp31-35 2010
5. Epstein RJ, Unblocking blockbusters: using Boolean text-mining to optimise clinical trial design and timeline for novel anticancer drug, *Cancer Informatics* 7: pp231-238, 2009
6. Sallantin J, Dartnell C and Afshar M, A pragmatic logic of scientific discovery, *Discovery Science* 6: pp231-242, 2006
7. Ben Yahia S and Engelbert MN, Special issue on Concept Lattices and their applications, S.I, World Scientific Publisher, 2008
8. Harpaz R, Haerian K, Chase HS and Friedman C, Statistical mining of potential drug interaction adverse effects, in the FDA's AMIA 2010 Symposium Proceedings, pp281-285, 2010

About the authors

Mohammad Afshar is the founder and CEO of Ariana Pharma, a leading decision support company based in Paris, France, focused on accelerating the development of novel therapeutics with the help of computational decision support technologies. Prior to joining Ariana in 2003, Mohammad was one of the founders and the Director of Drug Design at RiboTargets, Cambridge, UK. He holds a Medical Degree (DCEM), MPhil in Computer Science, a PhD in structural biochemistry and a Habilitation Doctorate from the Faculty of Medicine of the University of Montpellier, France. He is a member of the Scientific Committee of the French Cystic Fibrosis Association.
Email: m.afshar@arianapharma.com

Lilia Abtroun is an Application Scientist at Ariana Pharma. She focuses on developing and integrating expert data mining tools dedicated to biomarker discovery, as well as conducting projects for the Ariana biomarkers division. Lilia has also been involved in the analysis of various clinical projects and is working towards her PhD on formal concept analysis for biomarker discovery.
Email: l.abtroun@arianapharma.com