Pharmacovigilance: A data mining approach to signal detection

Drug discovery and development processes are complex, lengthy, knowledge-intensive and highly expensive. Therefore, regulatory approval processes of international organizations closely scrutinize all aspects of a new drug regarding its chemistry, manufacturing, controls, safety and efficacy before granting marketing license for a drug. The preclinical and clinical studies that are conducted before approval can map the majority of serious adverse drug reactions (ADRs) but not all of them. Rare and unsuspected serious ADRs, which represent some of the unknown safety risk of a drug can be detected through a well-developed vigilance program in lieu of prohibitive huge clinical trials demanding sample sizes of tens of thousands of patients. In addition, sometimes the severity and frequency of known ADRs are in fact only partially known from the premarket data.

As the World Health Organization (WHO) defines it, pharmacovigilance (PV) is the science and activities relating to the detection, assessment, understanding and prevention of ADRs or any other possible drug-related problems. PV is usually practiced by agencies and pharmaceutical companies by focusing on signals of ADRs in large databases. These databases are of huge sizes, e.g. USFDA database, adverse event reporting system; WHO database, VIGIBASE: GlaxoSmithKline database, OCEANS – all with millions of individual case safety reports (ICSRs). Based on a study, the highest power for finding a true signal is achieved by combining those databases with the most drug-specific data. The PV database needs to be well integrated with the clinical data management software. Such a database may also keep track of postmarketing prescription utility and complaints data.[1]

“Signals” of ADRs, as referred above, are “reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously,” according to the WHO. Usually, more than a single report is required for signal detection (SD), depending on the seriousness of the event and the quality of the information. Once a signal is detected, one can then analyze and confirm it. In detecting signals from large ADR databases, however, one has to use a procedure that is sensitive (low false negativity) and specific (high true positivity) for the purpose. The PV program of India (PvPI), revived with a fresh mandate in 2010, has now over 100,000 individual reports. With such a number of records, pilot SD programs can be initiated.[2]

A whole range of statistical methods have been applied for data mining and SD in PV. Data mining essentially involves getting something useful from lots and lots of data. Although it might sound and appear so, the data mining methodology is not linear, as it involves building and assessing models, carrying out simultaneous as well as serial steps. Data mining recognizes an alert from any available source data from pre- or post-marketing studies, although in practice data from postmarketing safety databases are largely used. Procedurally more than a single report needed for a successful SD note that a signal suggests a drug-ADR (D-R) association and per se does not establish causality between the two variables. Typical goals for SD include low false positive signals, that is, D-R association should be real; low false negative signals (should not miss any D-R signal); early detection of signals if possible and zero or very low rates of false discovery. Examples of already discovered D-R associations are bupropion-seizures, olanzapine-thrombosis, pergolide-increased libido, risperidone-diabetes mellitus, terbinafine-stomatitis and rosiglitazone-liver function abnormalities. SD algorithms can be used to establish a dis-association as well, such as, isotretinoin and suicide.

For mining purposes, ICSRs from a suitable database are collected. The higher the number of ICSRs the better it is for mining. The free text collected from a database is usually converted into a structured format, and finally statistical methods are applied to calculate an actual measure of signals. A typical data display is done in the form of a 2 × 2 contingency table [Table 1]. In this contingency table, the elements counted as a, b, c, and d are ICSRs available in the database. Thus, a given ICSR may contribute to only one of the cells of the table event, if the individual case refers to multiple medicinal products or multiple adverse events.[3]

All statistical approaches of SD are based on the observation that a large actual frequency of a would be found over an expected frequency of a (i.e., when R = a/E (a) is “large” there is a possible signal) or disproportional reporting of the signal over noise. Primarily there are frequentist as well as Bayesian approaches to SD. A frequentist approach would look for a statistically significant reporting ratio with E (a) = nTD ∗ nTA/n; significant proportional reporting ratio with E (a) = nTD ∗ c/nOD; and/or significant odds ratio with E (a) = b × c/d. These tests would be followed by a χ² test for statistical significance. Bayesian methods, however, work better when a is rather small (rare), as the uncertainty is accommodated better by the former.

| Access this article online |
| Website: www.ijp-online.com |
| DOI: 10.4103/0253-7613.157102 |

Quick Response Code:
The USFDA uses a Bayesian data mining approach called multi-item gamma Poisson Shrinker (MGPS). WHO also uses a Bayesian method based on a Bayesian confidence propagation neural network (BCPNN). These estimates provide shrinkage towards zero of the observed to expected number of ADRs, e.g., the empirical Bayesian geometric mean or the information component (IC) of the BCPNN. These Bayesian estimators are robust measures of D-R association and methodologically appealing when very small numbers are involved and where there is a need of continuous reassessment of the probability of association with the acquisition of new data over time.

The MGPS ranks D-R combinations in a matrix. The ranking is done according to how “interestingly large” is the number of reports of that D-R combination compared with what would be expected if the drug and event were statistically independent. Unlike the IC, MGPS technique gives an overall ranking of D-R combinations, whereas IC gives a nonrelative measure for each D-R combination.

The Uppsala Monitoring Centre (UMC) for WHO databases uses BCPNN architecture for SD. These neural networks are highly organized and efficient and give a simple probabilistic interpretation of network weights, which is analogous to a living neuron with its multiple dendrites and a single axon. BCPNN calculates cell counts for all potential D-R combinations in the database, not just those appearing in at least one report. Technically it is carried out with two fully interconnected layers one for all drugs and the other for all ADRs. In BCPNN, the statistic IC is used to decide whether the joint probabilities of ADRs are different from independent D-R.

While generating or detecting a signal by data mining is a very crucial aspect, causality assessment is perhaps the most important aspect of PV. Has this drug in fact caused this target ADR in this patient? Causality assessment is not a simple task and thus there are various tools and approaches exist for this purpose. The WHO-UMC system for standardized case causality assessment is one of the tools for standardizing assessment of causality for ADRs. Another scale, versus the Naranjo scale consists of 10 questions that are answered as either yes, no, or “do not know.” Different point values (−1, 0, +1 or +2) are assigned to each answer, and a score of >8 is required to be certain of causality. Of the 10 questions, they key ones are; Did the adverse event appear after the drug was given; did the adverse reaction improve when the drug was discontinued or a specific antagonist was given; did the adverse reaction reappear upon readministration of the drug; did the reaction worsen upon increasing the dose: was the reaction lessened upon decreasing the dose; and was the adverse event confirmed by any other objective evidence?

Finally, PV is finally making its importance felt in India both among industrial stakeholders and the regulators. Rejuvenated PvPI marks this significant inflection point in this country. Its increased activities in data collection, some analysis and co-operation with WHO-UMC have opened up one of the useful modern ways of managing risks due to drugs. Sooner or later Indian regulators have to harmonize with international bodies like EMA and USFDA where registration of any drug product in the USA and European countries is impossible without a thorough postmarketing PV program along with premarketing risk management plan. Thus, capabilities of data mining for SD would be a welcome initiative in India.

**Table 1:**

A typical 2×2 contingency table showing the frequencies (numbers of ICSRs) in a large database of a target ADR associated with a target drug (a); other ADR associated with a target drug (b); target ADR associated with other drugs (c); and other ADR associated with other drug (d).

| Number of reports | Target R | Other R | Total |
|-------------------|----------|---------|-------|
| Target D          | a        | b       | nTD   |
| Other D           | c        | d       | nOD   |
| Total             | nTR      | nOR     | n     |

It is important to note that a target ICSR may contribute to only one of the cells of this table event, if the individual case refers to multiple drugs or multiple adverse events. TD and OD are target drug and other drug, respectively whereas TR and OR are target ADR and other ADR, respectively. ICSRs=Individual case safety reports, ADR=Adverse drug reaction.

**References**

1. Hammond IW, Gibbs TG, Seifert HA, Rich DS. Database size and power to detect safety signals in pharmacovigilance. Expert Opin Drug Saf 2007;6:713-21.
2. Pharmacovigilance Program of India (PvPI). National Co-ordination Centre, Indian Pharmacopoeia Commission, Ghaziabad, India. Available from: http://www.ipc.gov.in/PvPI/pv_amcs.html. [Last accessed on 2015 Feb 16].
3. Gavali DK, Kulkarni KS, Kumar A, Chakraborty BS. Therapeutic class-specific signal detection of bradycardia associated with propranolol hydrochloride. Indian J Pharmacol 2009;41:162-6.
4. The Use of the WHO-UMC System for Standardized Case Causality Assessment. World Health Organization (WHO) – Uppsala Monitoring Centre. Available from: http://www.who-umc.org/Graphics/24734.pdf. [Last accessed on 2015 Feb 16].
5. Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts EA, et al. A method for estimating the probability of adverse drug reactions. Clin Pharmacol Ther 1981;30:239-45.

Chakraborty BS. Pharmacovigilance: A data mining approach to signal detection. Indian J Pharmacol 2015;47:241-2.

Source of Support: Nil. Conflict of Interest: No.