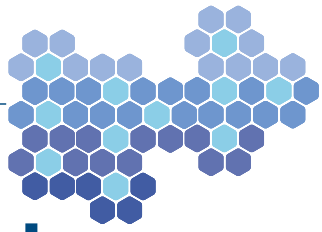

# Behind Closed Doors

Big data is transforming many industries, but the pharmaceutical sector remains steadfast in secrecy. What are the benefits, challenges and considerations surrounding the sharing of proprietary data, and how can an 'honest broker' help?

By Katharine Briggs at Lhasa

Pharmaceutical research and development has historically been shrouded in mystery – a secretive activity conducted behind closed doors to protect commercial advantage. However, as big data continues to transform many industries – from sales and marketing to banking and manufacturing – why does the pharma sector remain so reluctant to share data?

It is well documented that the average pharma company spends $350 million getting a single drug to market. One of the challenges in medical research is the scarcity of real-world data available to help develop new and improved drugs.

A large proportion of the drug development cost is spent on the research and discovery of new compounds and the lengthy biological and chemical testing of their properties in the laboratory using cells on plates (tissue cultures) and animals. This means every pharma company is sitting on a goldmine of big data. Analysis of that data could significantly reduce the product development lifecycle, yet a reluctance to collaborate still remains. That is not to say that data sharing does not happen in the industry, but it is not yet standard practice and stays the preserve of special projects.

Keywords

Big data
Data capture
Data sharing
*In silico* systems

## ChEMBL Database

One example of data sharing is the ChEMBL database. Hosted by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI), ChEMBL is a vast online database containing bioactivity data on more than 1.6 million drugs, as well as small drug-like molecules and their targets. Originally developed as a private resource by a biotechnology firm, it was acquired by EMBL in 2008 and has become a valued public resource for virtual screening, drug design and product development.

ChEMBL is utilised by academics and industries of all sizes, strengthening innovation from new research and the discovery of new treatments and drugs benefiting human health and agriculture. In the Strategic Vision for UK e-infrastructure report, Professor Dominic Tildesley of Unilever identified the use of the ChEMBL database as a crucial part of the company's development of antiperspirants. Unilever used the database to identify active components for antiperspirants and the ChEMBL data to build a model of their inhibition activity. Similarly, chemists from agrochemicals business Syngenta use ChEMBL in their product development. Mark Forster from Syngenta said that "ChEMBL has links between both chemistry and biology data which makes it searchable in ways that the underlying literature would not be. People at the EMBL-EBI do a fantastic job in making a vast amount of data of different types openly available to researchers, and without the EMBL-EBI resources in general, I'm sure life science research would be greatly hindered" (1).

## Increasing Collaboration

However, increased collaboration and dissemination of data not only benefits public health, it is also increasingly required by funding organisations and plays a vital role in reducing animal testing, which, aside from the ethical benefits, delivers savings in terms of time and money. The data and knowledge gained in data sharing also enables more informed decisions about what substances to test and what experiments to perform. An initiative led by the NC3Rs and the Medicines and Healthcare Products Regulatory Agency – involving 32 organisations sharing data for 137 compounds and 259 studies – identified that the use of recovery animals could be reduced by up to 66%, saving thousands of animals globally each year.

A fundamental aspect of the EU Registration, Evaluation, Authorisation and restriction of CHemicals regulation is the requirement to share data from studies involving vertebrate animal testing through substance information exchange forums to avoid the unnecessary duplication of tests. Additionally, the Cosmetics Regulation prohibits the use of animal testing for products marketed in the EU and their ingredients, as well as requiring data on toxicological properties to be included in the product information file.

A key obstacle to data collaboration is the perceived need within the industry to protect proprietary information. However, organisations need to be clear about how much of a competitive advantage they will lose by sharing data versus the knowledge they will gain. Consideration should also be given to the risk of not taking part in data sharing as organisations that participate will have a competitive and economic advantage over those that do not.

It is often the case that only regulatory bodies have ready-access to pooled datasets from multiple companies and, therefore, the opportunity to identify these broader patterns by performing cross-company analyses. This can present problems when businesses submit a new drug application, as broader regulatory knowledge can lead to challenges and assertions, resulting in delays and the need for additional data generation for the company.

Unlike other industries where research data can quickly become obsolete, in the pharma industry, data can remain valuable for many years, and fresh eyes can often reveal new insights. Additionally, new research topics and fields are emerging between the boundaries of traditional disciplines. By sharing data, companies can gain from external expertise in the same or different fields, opening up the data to be explored and used in new ways. Academics, small biotechs, small- and medium-sized enterprises and contractors can be counted as collaborators, further broadening the skills and experience, thereby creating new relationships. An opportunity to improve data quality also exists, as providing access to other experts will help identify errors and inconsistencies. As the costs of generating the data are also shared, it opens up the possibility for exploratory research that otherwise might not be commercially viable.

### *In Silico* Systems

As *in silico* systems move towards the prediction of more complex phenomena for which datasets of an appropriate size, quality and coverage are limited, maximising the accessibility of data will become increasingly important.

The increased recognition of the importance of *in silico* systems led to the creation of the eTOX consortium – a seven-year public-private partnership within the framework of the European Innovative Medicines Initiative. Aimed at developing innovative *in silico* strategies and novel software tools to better predict the toxicological profiles of small molecules in the early stages of drug development, the backbone of the project was a database hosted and curated by Lhasa, who acted as the honest broker – an organisation trusted by all partners – for the project. The database consisted of pre-clinical toxicity data for drug compounds or candidates, extracted from previously unpublished legacy reports from 13 European pharma companies. Enhanced by publicly available, high-quality toxicology data that was collected by the European Bioinformatics Institute, the database also incorporates the RepDose database donated by Fraunhofer.

Early eTOX-use cases included the investigation of the relevance of specific histopathology findings (confirmed to be target-related and species-specific), identification of potential target-related effects (leading to the inclusion of specific target organs in early *in vivo* studies) and the implementation of a framework of four key approaches (similarity of structure, pharmacology or adverse effects and use of *in silico* prediction) as part of an early small molecule drug development pipeline. The eTOX project led to the formation of eTOXsys: a software solution that delivers improved early drug candidate safety assessment through access to proprietary toxicology data and predictive models.

### Reaping Rewards

So how can pharma businesses overcome the challenges

> Creating an end user license that requires agreement to conditions of use – including specific authorisation from the data owner and limiting access to certain users – can mitigate risk

and concerns relating to data collaboration to reap the rewards of projects such as eTOX? Regulations to protect the privacy of personal health information can be seen as a barrier to data sharing due to the risk of accidental, malicious or compelled disclosure. However, by redacting it to strip out individual identifiers, statistically altering it in ways that do not compromise secondary analysis and placing restrictions on access, data can still be shared securely.

Information misuse is another concern. Nevertheless, creating an end user license that requires agreement to conditions of use – including specific authorisation from the data owner and limiting access to certain users – can mitigate risk.

Storing data in disparate repositories and formats and using potentially incompatible data types presents another significant technical challenge. Overcoming this by converting the data to an agreed format is achievable, but can add cost. Opting for platform-independent file formats for exporting and importing data, such as XML, CSV or SDF – which can be opened using several software applications – makes sense. However, using the same format for exporting and importing data does not avoid differences in what and how data are captured – for example, as a number, text, etc. Here, data standards such as Standard for Exchange of Nonclinical Data can ensure the compatibility of the data being captured.

## Capturing Data

When capturing qualitative data, a controlled vocabulary is preferred to avoid problems in spelling and terminology. The use of ontologies offers additional benefits, as relationships – synonyms, meronyms/homonym, hyponyms/hypernyms – between terms can be captured. Ontologies were developed as part of the eTOX data sharing project to help with cross study data analysis where pathology findings could be reported as different levels of granularity, such as gastrointestinal tract versus colon.

Quantitative data should ideally be captured using standardised units to simplify data mining and analysis. This is not always practical and can lead to an increase in the number of errors introduced during data entry. Additionally, when designing the schema, an assessment also needs to be made as to whether precise figures will always be given or if greater than/less than values and number ranges also need to be captured.

Sensitivity of data can also change due to the repurposing of drugs and drug candidates. One of the eTOX project participants was able to elaborate a procedure for obtaining general permission for full or restricted sharing, dependent on the status of the compound, such as whether it was marketed, terminated, under current development (excluding new formulations, new indications or combinations of marketed drugs) or subject to product liability claims.

Of course, legal and intellectual property departments often take responsibility for deciding if data can be shared. The disadvantage of this is that they only see the risks and, being risk adverse, say no by default. Demonstrating how the data will be used before it is donated is also difficult. The eTOX project participants highlighted the need for a project summary to be shared with upper management and departments involved in granting authorisation to increase publicity and facilitate decision-making.

In the case of confidential data, an honest broker can help to protect the security of sensitive data by controlling access for the other partners. A not-for-profit or academic organisation is likely to be preferred over a commercial one for this reason.

## The Way Forward

The past decade has seen data sharing within the pharma industry evolve from being virtually non-existent to a landscape where most companies will have gained experience through one or more initiatives. However, to truly benefit, data collaboration needs to be incorporated into 'business as usual', rather than remaining the preserve of special projects.

Data still exists within silos, and the people who could do something useful with that information often do not have access to it. A fear in the sector remains that sharing data gives away commercial advantages when, in fact, sharing information could significantly reduce overheads and speed up the development of new drugs. With the rising cost of clinical trials and health data, the industry needs to view collaboration as the way forward. Sharing information is not without its challenges, but, with the right partners, the benefits far outweigh the risks.

Reference

1. Visit: www.ebi.ac.uk/sites/ebi.ac.uk/files/groups/external_relations/Documents/ChEMBL_CaseStudy%20v4.pdf

Katharine Briggs is Research Leader at Lhasa. She joined the data team in 2006 and is involved in managing several data sharing initiatives. Katharine holds a Bachelor of Science (Special Honours) in zoology and a Master of Science in information studies from Sheffield University, UK.
Email: info@lhasalimited.org