Making the most of your discovery data

Discovery scientists in most pharmaceutical companies are struggling with incompatible legacy systems, a growing volume of data and the imminent advent of a product model that will demand much more of them. Data-driven drug discovery (4D) is a methodology for identifying a company's strategic goals in research and development, and aligning its IT infrastructure with those goals. It enables companies to measure the effectiveness of their IT systems and networks, assess the impact of any shortcomings, and design solutions that are tailored to their business objectives.

By Dr Nick Davies and Dr Tim Peakman

Pharmaceutical research is fundamentally about generating high-quality data and making sense of it to obtain new insights into disease and its treatment. But despite huge advances in information technology (IT), that task is steadily getting harder for discovery scientists. Mergers and acquisitions have left many large pharmaceutical companies struggling with legacy systems that cannot speak to each other, and the sheer volume of data is growing massively, as the new molecular sciences come on stream. The nature of the research the industry performs is also becoming ever more complex, as are the data it uses to make decisions, and the speed with which it must make decisions.

In fact, most pharmaceutical companies invest heavily in IT; according to META Group, the technology research firm, they spend between 4% and 5% of their annual gross revenues on hardware, software and related services[1]. But they often focus on technologies that will enable them to do more things rather than technologies that will help them to make sense of the data they possess – and this is what discovery scientists most need. We shall outline here a process for identifying the changes that are required to create such an infrastructure, and ensuring that it is aligned with a pharmaceutical organisation's key business objectives.

An increasingly data-intensive environment

Research from PricewaterhouseCoopers shows that between 1998 and 2002 (the latest year for which figures are available), there were 1,584 mergers and acquisitions in the pharmaceutical industry[2]. The vast majority of these deals left the companies concerned struggling to reconcile totally different IT systems. Many of them have now harmonised the technologies supporting back-office activities such as human resources and

Informatics
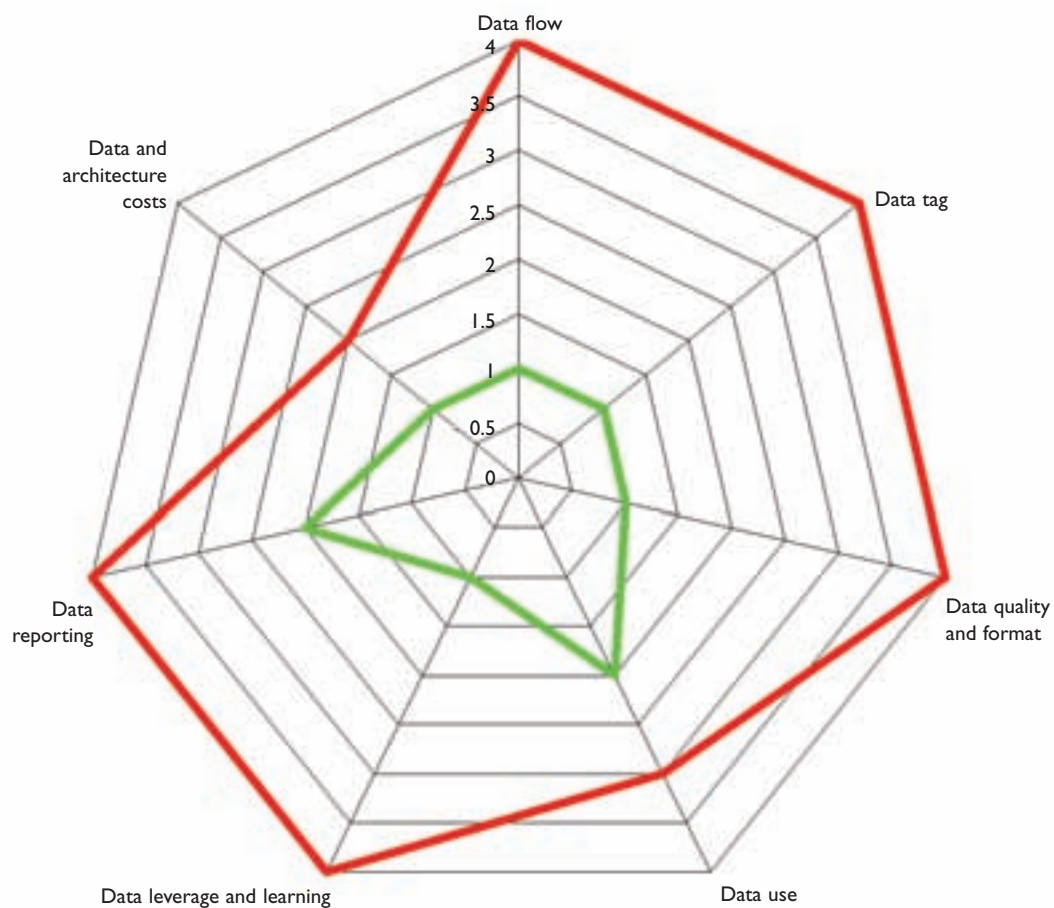
accounting, but integrating their discovery data is a far bigger challenge.

The volume and variety of data are also growing rapidly. Combinatorial chemistry, high throughput screening, genotyping and proteomic technologies, x-ray crystallography and other such tools have already generated numerous petabytes of data, but this is nothing compared to what is just around the corner. Where previously, for example, a company might generate 500 hits from high throughput screening and conduct assays on the most promising compounds, with high throughput profiling it can now conduct multiple secondary assays on all 500 hits, generating as many as 100,000 assays for one project alone. High throughput biology – genomics, proteomics, metabonomics and the like – will produce even more data. The genetic profile of a single person generates about two terabytes[3], and the number of different proteins in the human body is at least an order of magnitude greater than the number of genes.

These same sciences are changing the nature of the medicines that are made. They will eventually enable the industry to produce healthcare packages for specific disease pathologies, or targeted treatment solutions, as IBM dubbed them in its research paper Pharma 2010: The Threshold of Innovation[4]. But though targeted treatment solutions represent the most promising source of future revenues, discovering and developing them poses problems with which pharmaceutical companies have never formerly had to contend.

Making such treatments involves the simultaneous development of drugs, diagnostics and biomarkers, so it will substantially expand the scope of the discovery process. It will also blur the traditional boundaries between biology and chemistry, and between discovery and development, and accelerate the speed with which new products can be tested in man. The data that are used will thus span a wider range of disciplines and be more complicated than those required to support conventional drugs. Similarly, the decisions that are made on the basis of the data – both scientific decisions about whether to push a molecule further down the pipeline and practical decisions like how to micro-manufacture biologics for preliminary

clinical studies – will need to be made at a much earlier stage in the process.

The bottom line, then, is that targeted treatment solutions will demand much more of the discovery function. The requirement for data analysis will become even greater, as will the need to share data among a wider group of people – including the regulators, research, development and manufacturing partners, and in-house sales and marketing staff – much more rapidly than before.

Creating a maturity profile

Of course, every company is different, which is why it is essential to use a broad-based approach that starts with the corporate culture, specific business needs and goals of the organisation. Only then can a company determine the sort of IT infrastructure it needs.

4D provides just such a framework. It begins with initial interviews, in which senior discovery scientists and management from a wide range of project teams, as well as key individuals from the IT discovery function, identify what they want to achieve, what is stopping them from doing so, and what changes would help them. This exercise rapidly generates the information with which to create a basic profile of the company's discovery/IT organisation and infrastructure. We use seven key criteria, all of which can be measured, to assess its maturity (see table).

Assessing a company in such terms produces a picture of its individual strengths and weaknesses.

It also shows how the company compares with other organisations. **Figure 1** illustrates the typical profile of a large pharmaceutical concern (depicted in green) and a small biotechnology firm (depicted in red), based on our analyses of 12 pharmaceutical companies and five biotechnology companies. They have been ranked on a score from zero (weakness) to four (excellence).

The typical biotechnology firm excels on five counts; the quality and quantity of data available on its systems is superb, as is the speed with which they are made available. The number of new insights the data generate and the ease with which results can be reported are likewise first-rate. The extent to which scientists actually use the data and the costs associated with producing them are not quite as good, but the scores even here compare very favourably with those elsewhere in the life sciences sector.

Conversely, the typical large pharmaceutical company has a much lower score on all seven counts. Its data usage and data reporting processes are relatively strong, but its data flows, data quality, data costs and ability to extract new insights are all quite weak. This is partly a factor of size; sharing data between thousands of people based in different sites and different countries is obviously much more difficult than sharing data between a few hundred people working out of the same location. But it also reflects the problems many big companies experience in dealing with numerous legacy systems and operating in a complex environment characterised by multiple applications.

The profile a particular organisation has can then be correlated with all the main components of data usage in discovery to measure the impact on its business. So, for example, poor data flows impede decision-making, restrict access inappropriately and delay the discovery process. Good data flows, by contrast, facilitate decision-making, by providing real-time access to data and drill-down tools for manipulating the data. They also ensure that access to data is dependent on need rather than organisational structure or seniority. They promote innovation and optimise planning, and they capture all the relevant data in a consistent format.

Homing in on the problems

Creating a maturity profile helps a company to identify its core problems very rapidly, but the 4D process does not stop there. The next stage is a series of detailed interviews and 'milestone' workshops to flesh out the picture and get to the root of each problem. This is important because many

---

The seven key criteria for assessing an organisation's IT maturity

DATA LAG: The time required to load the data, once they have been generated, and the ease with which they can be integrated and shared.

DATA QUALITY AND FORMAT: The number of errors in the data, the amount of redundant data and the degree of compliance with corporate policy.

DATA USE: The extent to which the data that are available actually get used.

DATA LEVERAGE AND LEARNING: The number of times better and greater access to data produces a 'win', such as prior research that accelerates the discovery process or enables scientists to kill a compound more quickly.

DATA PRODUCTIVITY AND REPORTING: The amount of time that is saved performing unproductive data-related tasks like logging on, finding data, integrating data, visualising and manipulating data; and making a decision.

DATA AND ARCHITECTURE COSTS: The costs associated with uploading, storing, cleaning, integrating, accessing and reporting, and sustaining the entire IT infrastructure.
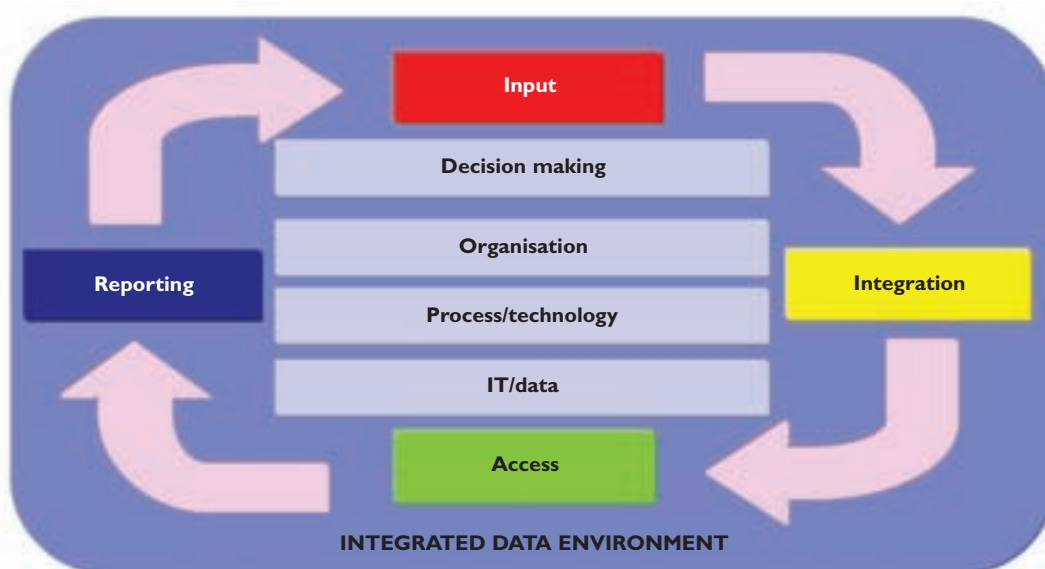
problems often stem from the same underlying cause. If a company has difficulty locating and analysing non-numeric data, for example, it may be because the data are not searchable, because the scientists who need it do not have access to the relevant database or even because there is no such database within the organisation. Trying to design a solution without identifying the real cause is rather like discovering a drug and then hoping to find a condition it can treat.

Distinguishing between the proximate and ultimate causes of a problem also makes it possible to cluster superficially unrelated problems into similar themes. Many pharmaceutical companies, for example, are concerned about the cost and effectiveness of their IT; the need for better inter- and intra-corporate collaboration, as they participate in a growing number of partnerships with academic institutes and biotech firms; and the speed with which they can upload and share the data they generate. Other common concerns include integrating multiple forms of data from multiple sources; reconciling data that have been generated using different standards and nomenclature; reading data in context; accessing historical data or data produced prior to a merger; reporting data; and tracking projects.

Some of these problems are technical, but others are organisational. The discovery and IT functions often have different investment cycles, for example. The discovery function identifies its main business objectives and which technologies it wants to buy, but it does not call in the IT function until it needs support for those new technologies. The IT function therefore has no opportunity to establish how the new technologies can be integrated with the existing infrastructure or even, indeed, whether they were really necessary, until after the investment has been made.

Similarly, most discovery functions suffer from the 'happy hacker' syndrome; frustrated by the limitations of the systems they are using, individual scientists frequently develop bespoke tools for solving a local problem but when they leave the organisation that knowledge is lost. Such ad hoc solutions, however ingenious, are also notoriously difficult to integrate with the remaining IT architecture.

Almost all the data problems from which pharmaceutical companies suffer actually fall into one of four categories: how important data are loaded into the corporate systems (input); how data are shared across the organisation (integration); how scientists access, visualise and manipulate the data (access); and how the data are assembled and presented in order to facilitate decision-making (see **Figure 2**). But tackling these issues requires an approach that simultaneously addresses all the key areas of data usage in discovery – including the way in which the function makes decisions; the way in which it is organised, both in itself and in its relations with external bodies; what processes and technologies it uses; and what sort of data environment it operates in.

Perfection isn't necessary

In a perfect world, the four stages via which data flow through a pharmaceutical organisation would all be completely integrated. The data would be uploaded in real time in a format that

provides universal access. Multiple data types from multiple sources would be rapidly assimilated. The data (both current and historic) would be promptly and easily available via desktops and user-friendly querying, visualisation and manipulation systems. And, lastly, a simplified, customised suite of reporting tools would enable users to produce reports with minimal cutting, pasting or re-inputting – so that management could make decisions safe in the knowledge that the information on which it was acting was both accurate and comprehensive.

Moreover, this perfect world is now quite achievable, since the IT components required to create an integrated data environment have all been developed. Grid platforms and server farms have provided the computing power to process vast quantities of data at great speed. Open standards and sophisticated middleware wrappers have provided the means with which to integrate applications and different data sources, and share data. Biometric authentication and encryption technologies have provided a secure way of sharing data,

and the most recent portals and data mining tools are sufficiently advanced to handle complex scientific information.

In practice, however, perfection is not necessary. Discovery functions naturally differ in their ability to use data to support the scientific process; they range from organisations that recognise their needs but have done little to address them to those – a tiny minority – that operate in a seamless IT environment. But in our experience, it is enough for most companies to occupy an intermediate position, in which data can be shared between different project teams.

Using the maturity profile to design and implement the right solutions

Once a company has measured the maturity of its IT infrastructure, created a profile and mapped out the impact of any shortcomings on its business, the information can be used to define what its scientists need and what IT solutions will enable them to do their jobs as effectively as possible. Using an iterative approach to design, the company and its

technical advisers can then design and implement solutions that align with its strategic goals and address the underlying causes of the difficulties it is experiencing. They can also determine the order in which those solutions should be implemented, since no organisation can make all the changes that are required in one fell swoop. Lastly, they can install each solution in a logical progression that helps the company to build the IT environment it needs to support its key business aims.

One big pharmaceutical operation that has gone through this process found, for example, that it had 10 main data management problems. In common with many other multinationals in the sector, its research facilities were scattered across various sites and countries; it had conducted a number of acquisitions; and it was suffering from a decline in research and development productivity that had been exacerbated by several recent failures in the pipeline.

The 4D process rapidly established that data from different sites could not be compared, and data from different continents could not be shared, so working on projects on a global basis was extremely difficult. The use of numerous different tools and systems was compounding these challenges and driving up costs. Indeed, some research scientists had to use 22 separate passwords to access the data they required. Meanwhile, most non-numerical data were inaccessible, the systems used to visualise all data were very complex, and the absence of an agreed ontology and context for much of the data made them very hard to mine. As if this were not bad enough, most of the scientists did not receive sufficient training in the tools they were using, and tracking the progress of projects was almost impossible.

The company has now begun to redesign its IT infrastructure so that all its research scientists, wherever they are based, can access and manipulate all the discovery data it owns. It will introduce a common set of standards, common platforms and tools. It will also simplify the supporting architecture, with the integration of its biological and chemical data, single-password access, a single browser program for viewing all the data, and location-independent querying.

Clearly, no two organisations have the same starting point, so there is no one formula for sorting out their problems. But the 4D methodology is a valuable tool for rationalising and simplifying a company's IT environment and that, in turn, has several advantages. It reduces the costs associated with data management by as much as 30%. It minimises the amount of unproductive time spent min-ing, managing and reporting on data. It improves decision-making both within and across projects. Lastly, it helps companies to maximise the value of the data they possess and increase their chances of discovering good new drugs – the key measure for determining whether pharmaceutical companies and their share prices rise or fall.   DDW

Dr Nick Davies is a senior consultant in IBM Business Consulting Services' Pharmaceutical Discovery Practice. He gained a PhD in molecular biology and immunology from Cambridge University and completed a post-doctoral fellowship in the Department of Biochemistry and Molecular Genetics, Imperial College. He subsequently worked in drug discovery at Novartis and AstraZeneca, before joining PwC Consulting (now part of IBM).

Dr Tim Peakman completed his PhD thesis on the regulation of gene expression in anaerobic bacteria and also has an MBA. He worked in drug discovery at The Wellcome Foundation and Glaxo Wellcome before joining PwC Consulting (now part of IBM) to lead the pharmaceutical discovery group. He is currently Operations Director at UK Biobank.

References
1 META Group, Worldwide IT Benchmark Report 2004, Vol. 1 2004 IT Spending & Staffing Analysis: Pharmaceuticals & Medical Equipment (2003), p.2.
2 PricewaterhouseCoopers, Pharmaceutical Sector Insights: Annual Report 2002 (2003), p.4.
3 Cobbs, C (2002). "Machines and Genes: Superfast Computers aid Today's Genetic Advances". Published in Orlando Sentinel (July 15, 2001), p.G.1.
4 IBM Business Consulting Services, Pharma 2010: The Threshold of Innovation (2003), p.11. Copies available at http://www.ibm.com/industries/healthcare/doc content/resource/thought/390 030105.html