One for All

The collation of clinical trial management system data between companies to create a common data-sharing environment will inevitably allow for a broader picture of drug development. Collaborating in such a way minimises the difficulty of selecting a suitable investigator, reducing time and cost of trial initiation

Claire Sears and Graeme Benson at DrugDev

One major cause of delay and spiralling costs within clinical trial initiation is that sponsors struggle to find appropriate, qualified investigators to manage their protocols. Each company relies on their own internal information and, consequently, has an incomplete view of the investigator world.

This article will focus on a collaborative case study in which investigator and site information is shared between five pharmaceutical companies – known as the Investigator Databank. One of the greatest struggles to launch was, and continues to be, development and evolution of shared data standards across companies, and the internal systems within them. Initial hurdles to integration were focused around identifying consistent data fields of interest across all the businesses, and establishing mechanisms for data transfer. Once the data were received, the next set of issues related to the development of algorithms for matching investigators and sites across companies and systems that often contain incomplete and/or contradictory data.

The evolution of data integration challenges this cross-pharma team has faced – and the strategies that have been employed to move towards a uniform database – will be discussed. This article will also outline the development of interoperability standards and golden investigator and site lists – a single common guide to investigators and sites, benefiting the entire industry.

Time and Money

We still live in a world where people die from incurable diseases and others live in discomfort – and we are working in an industry with lots of ideas for drugs that could make a difference, but are currently unavailable to patients. Why? Two of the main reasons for this, as with many things in life, are time and money. There are huge costs involved in completing the lifecycle of a trial, and it takes too long to go from drafting a protocol to writing a prescription.

There are plenty of efficiencies to be gained across the multiple processes involved in a clinical trial – contracts, budgets, project management and monitoring, for example. But there is a far more fundamental area to consider: doctors. If you are struggling to find clinical trial investigators, then you will not find patients, and you will not get your study started; poor recruitment adds additional cost and time to the initiation of a trial.

Driving Innovation

But what can be done? It is not about regulators reducing the burden, or people working harder, faster or longer. A company's greatest weapon is its data and the way in which it is utilised. People need to reject the current process norms and think disruptively instead.

Currently, most pharma companies and CROs restrict their ability to find start-up sites by centering on those with which they have established a previous relationship. In some cases, the time and associated cost of identifying, selecting, qualifying and engaging investigators is almost prohibitive. At best, trials are not being conducted at the optimal sites necessary to maintain protocol.

Three years ago, a few key people in some of the biggest global pharma companies came together to change this, and the Investigator Databank was born. The Investigator Databank is now a global collaboration between Janssen, Lilly, Merck, Pfizer and Novartis (with more companies to come) through which each business shares investigator information that they have on file. It aims to reduce administrative burden for investigators and increase visibility of qualified investigators to research sponsors, and is hosted by a trusted third party.

The companies involved in this project believe that the global community of clinical trial investigators is a common resource on which industry sponsors rely; sharing information about trial participation and recruitment metrics results in better matching of doctors and future protocols, thus providing a collective benefit to both investigators and sponsors. The Databank exists as a collective knowledge about sites and investigators, their facilities and past recruitment history.

The belief that site and investigator information should be shared was a huge first step for this project. However, many other hurdles had to be cleared to reach the present situation – five companies sharing each other's clinical trial management system (CTMS) data, as well as that from the third-party host's network of over 80,000 investigators. Currently, each company extracts data on a monthly basis, which is then imported into a shared repository and made available through a search interface.

Technical Problem

It became apparent to members that the lack of a standard approach for sharing CTMS data would be a limiting factor to the success of the Databank concept. It would mean that the monthly process of extracting data from the various sources, converting it into a consistent form and importing it would become increasingly onerous. Manual effort could be applied to make all data look the same, but it did not necessarily have common meaning – leading to possible errors, as well as inefficiency, in the process. Furthermore, underlying CTMS database schemas differ, based on each company's vendor and local configurations within the business. This all contributed to the lack of standardisation, which was exhibited in a number of ways:

- Syntactic: different sets of data fields; varying nomenclature for similar data field labels; alternative vocabularies – for instance, international classification of diseases (ICD-9) (1), medical dictionary for regulatory activities (MedDRA) (2) or medical subject headings (MeSH) (3) – and terminology; and non-compliance with international coding – ISO, for example (4)
- Semantic: different meaning for the same fields (for example, site recruitment dates) and various definitions for types of data (such as study phase)

Figure 1 shows some examples of these inconsistencies. In each case, the left columns show the variations in terms

Figure 1: Local CTMS of global controlled list mappings

Recruitment status

All variations	Common term
Closed	Complete
Complete	
Completed	
Active, not recruiting	Ongoing
Actively recruiting	
Enrolment closed	
Enrolment in progress	
In progress	
In preparation	
Last patient entered trial	
Open	
Recruiting	
Start-up in progress	
On hold	Terminated
Recruitment on hold	
Site on hold – complete	
Site on hold – enrolment in progress	
Site on hold – start-up in progress	
Stopped	
Terminated	

Role

Allowable values (Person_Job_Type)	Map to
Principal investigator	Principal investigator
Sub-investigator	Sub-investigator
Co-investigator	

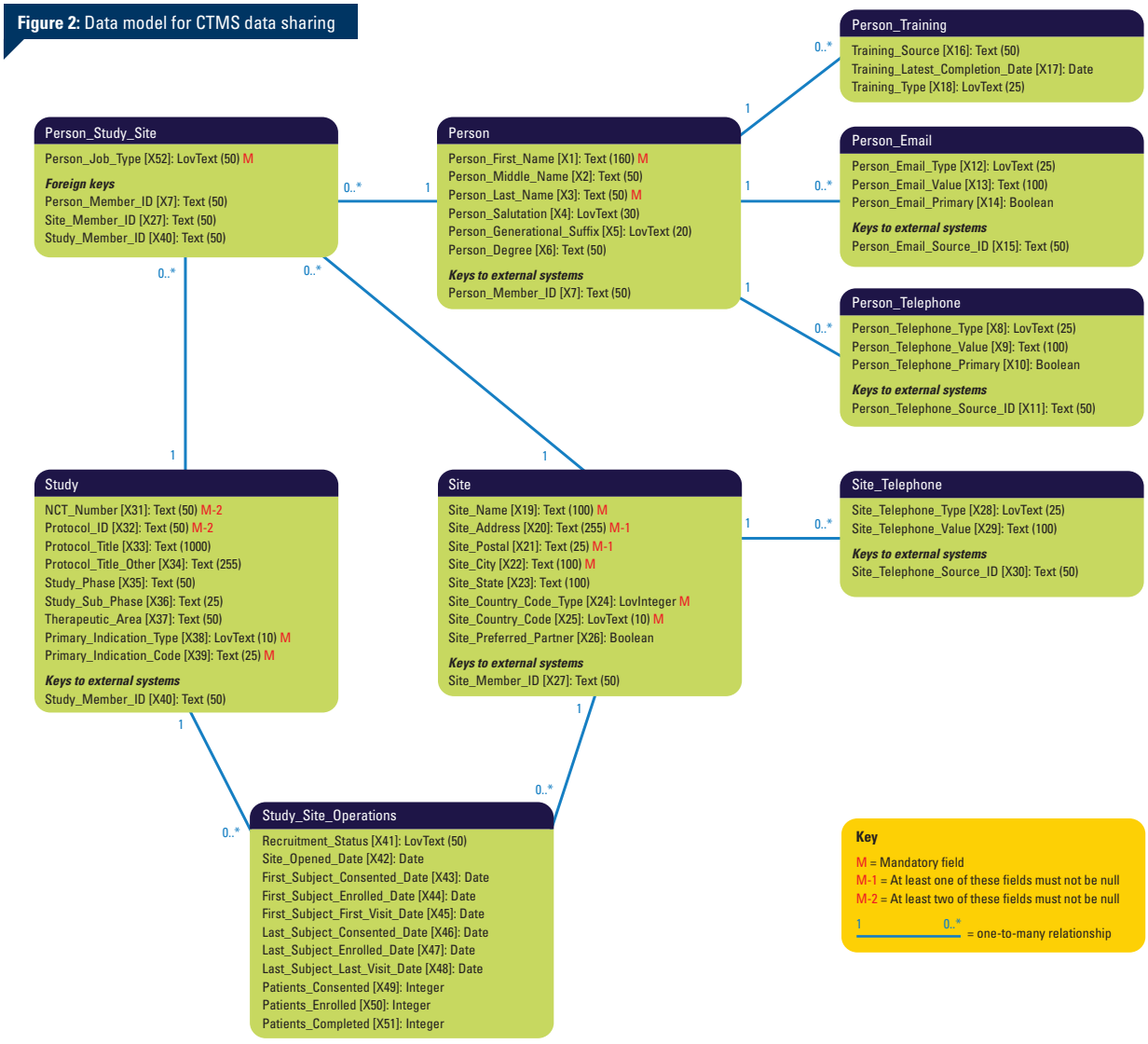
Primary indication

All variations	Common term
ICD-9	MeSH
MedDRA	
Proprietary	

Phase

Allowable values (Study_Phase)	Allowable values (Study_Sub_Phase)	Map to
1	null	Phase I
1	A	
1	B	
1	P	
I	null	
IA	null	
IB	null	
Phase I	null	
Phase IA	null	
Phase IB	null	
Phase I clinical trial	null	
Phase IB clinical trial	null	
Ph.I	null	
Ph.IB	null	
I/II	null	Phase I/II
IB/II	null	
Phase I/II	null	
Phase I/II clinical trial	null	
Phase I/IIa clinical trial	null	
Phase I/IIB clinical trial	null	
Phase IB/II	null	
Ph. I/IIA	null	
2	/1	
2	null	Phase II
2	A	
2	B	
02-Jan	/2	
II	null	
Phase II	null	
Ph. IIA	null	
Ph. IIB	null	
Phase IIb clinical trial	null	
Phase IIa/IIb clinical trial	null	
Phase II clinical trial	null	
Ph. II	null	
Phase IIIa clinical trial	null	
Phase IIIA	null	
Phase IIIB	null	
3	/2	Phase II/III
II/III	null	
Phase II/III	null	
Phase IIb/III clinical trial	null	
Phase II/III clinical trial	null	
3	null	Phase III
3	A	
3	B	
III	null	
IIIB	null	
Ph. IIII	null	
Phase IIII clinical trial	null	
Phase IIIIa clinical trial	null	
Phase IIIIb clinical trial	null	
Phase IIII	null	
Phase IIIIA	null	
Phase IIIIB	null	
III/IV	null	Phase III/IV
Phase III/IV	null	
Phase IIIB/IV	null	
4	null	Phase IV
4	A	
4	B	
IV	null	
Phase IV clinical trials	null	
Ph. IV	null	
Phase IV	null	
Phase IV local	null	

Figure 2: Data model for CTMS data sharing



that can be provided from different CTMS sources. The right column shows the agreed terms to which the CTMS data should be mapped for the purposes of consistent meaning and display.

In addition to the inherent differences of syntax and semantics, the data within each CTMS system contains many data quality inconsistencies, mainly as a result of having been created by a large number of users from different geographies over a long period of time. Missing, incomplete or incorrect data represents the greatest challenge.

Collaborative Solution

Collaboration between the senior technical representatives from each member company and the third-party host allowed for an agile and iterative approach to the generation of standards.

The objective was to create a set of three documents – the first two defining the syntax, and the third the semantics:

1. An Excel workbook that defines all of the data fields, their purpose, format and allowable values
2. An extensible markup language (XML) schema definition (XSD) (5) that describes the specific technical definition of the related XML schema (6)
3. A Word document that explains the rules that should be applied when importing the data files in order to ensure common meaning

After six months of bi-weekly, one-hour conference calls, a candidate standard file specification was drafted, setting a uniform data specification and its associated mappings of vocabularies and controlled lists.

Some key principles were agreed and implemented within the data standard:

- Existing international standards to be used – for example, ISO country codes (3166) and date (8601) (7), and internet engineering task force email format (8)
- MeSH is the standard set of terms within the Investigator Databank
- Where a recognised international standard does not exist, a global vocabulary for common terms was

agreed, against which all of the variations from sponsors are matched
- Member companies do not each have to complete mapping within their own CTMS – they simply export their terms and codes (ICD-9, MedDRA, and so on) and these are mapped on import
- XML is the preferred format for the standard file specification, but comma-separated variable (CSV) is acceptable

The iterative specification period was followed by a further three months of rigorous validation using real data. During this validation process, the draft business rules were refined; these rules include governance of allowable sharing of data between companies, and the calculations for key analytics and metrics.

The business rules were implemented through a set of algorithms that match investigators and sites across companies and systems, de-duplicating data and creating a single record for each investigator and site. The algorithms use the uniform data specification and mappings, applying a variety of fuzzy and semantic pattern-matching techniques to canonicalise and clean the imported data.

Figure 2 shows the data model that all member companies and the host company are using; it is the basis of the CSV structure and XML schema. Shown here is the long form of the entity and attribute names, as well as the abbreviated alternatives which help minimise the size of the data files.

End Result

Since April 2014, each of the five pharmaceutical companies has been extracting compliant files, and the hosting third party has been importing, matching and canonicalising the data into a common sharing environment. This has created an accessible technology and process platform, built from a scalable and extensible design.

A resultant by-product of this has been the creation of 'golden lists' – one list of investigators and one of sites. These are clean, accurate lists of core data, each with a unique global identifier. They can be used throughout the whole industry, revolutionising the way we track and contact investigators and sites. In real terms, the benefits are as follows:

- Access to more investigators – member companies have access to over three times more investigators with research experience since 2008
- Reduction in non-performing sites – better matching of the right site to the right protocol due to access to more investigators and more robust site-level metrics
- Decrease in time for feasibility and site ID – access to more up-to-date contact information
- Expanded access for investigators to clinical research opportunities by making investigators and sites known to a broader range of trial sponsors

Essentially, the Databank provides an ever-expanding meeting place in which sponsors can collaborate with investigators who they have not yet met.

Next Steps

The standardisation work undertaken for this project will form the first part of a broader interoperability standard, encompassing other data sources – such as public data, electronic medical records, country-level data, and other third-party data sources. The Databank also intends to provide open access to the golden lists so that the whole industry can benefit from a set of common identifiers, pointing to the truth about investigators and sites. This will allow companies to share in the efficiencies gained from the improved communications.

References

1. Visit: www.cdc.gov/nchs/icd/icd9.htm
2. Visit: www.meddra.org
3. Visit: www.nlm.nih.gov/mesh/overview.html
4. Visit: www.iso.org/iso/country_codes.htm
5. Visit: www.w3.org/tr/xmlschema11-1
6. Visit: www.w3.org/xml/Schema
7. Visit: www.iso.org/iso/home/standards/iso8601.htm
8. Visit: http://en.wikipedia.org/wiki/email_address

About the authors

Dr Claire Sears is Director, Investigator Engagement at DrugDev, and is based in its UK office. With a PhD in Cardiovascular Physiology, she also acts as Scientific Consultant to the site services team. Previously, Claire spent eight years at AstraZeneca in both clinical and commercial departments, where she worked in scientific communications and medical affairs roles on products in all phases of development and across a variety of therapy areas. Claire also served as a Royal Society Dorothy Hodgkin Research Fellow in the Department of Cardiovascular Medicine at the University of Oxford.
Email: claire.sears@drugdev.com

As Chief Information Officer, Graeme Benson is responsible for shaping DrugDev's overall strategy, while simultaneously ensuring that all daily operations are tracking to the company's long-term vision and goals. His duties include specifying software and infrastructure functional requirements, as well as managing their implementation and support. Prior to DrugDev, Graeme was in charge of the creation and implementation of information strategies and data standards for the UK Government, the NHS and a major London hospital. He received his MSc and PhD from Queen's University in Belfast.
Email: graeme.benson@drugdev.com