A Practical Guide to Clinical Data Warehousing

Over the past decade we have been involved in Clinical Data Warehousing and one thing we still get asked quite frequently is "what is a clinical data warehouse?". We recently presented a web seminar on the subject and this prompted us to write this article, where we hope to dispel some of the mystique around the subject and provide a practical and pragmatic explanation of what exactly a clinical data warehouse is.

Perhaps before we take a look at a *clinical* data warehouse, we should establish what a typical data warehouse is. There are many definitions of a data warehouse but perhaps the simplest is that it is a **"database used for reporting and analysis".**

Typically, data warehouses are of a fixed design where data is extracted from one or more source systems, transformed and then loaded into a central data model that has been optimised for analysis. The traditional forms of analysis that occur with a typical data warehouse are decision support, market research and analysis, data mining and other commercially biased analysis. The first column in Table 1 below lists the common attributes of a typical commercial data warehouse.

A clinical data warehouse, in practice is often a far more loosely defined system, and perhaps even a misnamed one at that. It certainly is possible to design a clinical data warehouse that follows the model of a traditional data warehouse with a single well-defined data model into which clinical data are loaded. However, it is more common to load loosely related data into a central (but not single data model), well-organised, robust, secure, compliant "clinical data repository" for multiple reporting and analysis uses, allowing companies immediate access to their greatest asset, information.

Table 1 further illustrates the core differences between a classic business data warehouse compared to a clinical repository. This distinction is often mis-understood by IT departments who have too often started a clinical warehouse project in the belief that it is a typical warehousing project, only later to understand the complexities and variability of the clinical world.

Let's take a look at some of the distinguishing features of a clinical repository/warehouse:

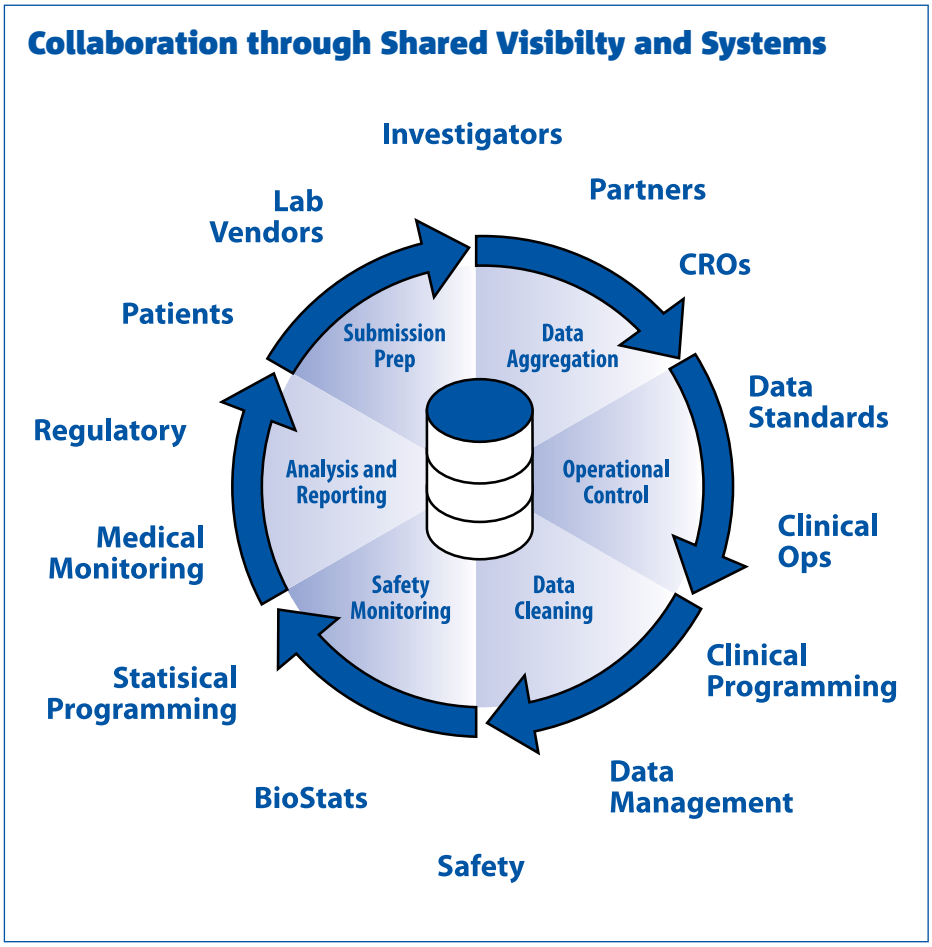
Firstly, a clinical data warehouse can be used for many tasks, some of which do not traditionally fall into the analysis and reporting category.

If we start at the beginning of a clinical trial, the first major area a clinical data warehouse may come into play is in performing "data management" type tasks. Data can be loaded into a clinical data warehouse from many sources. Clinical data management and EDC systems often provide much of the source data, along with third parties such as laboratories, IVRS systems, PK/PD and many other source systems.

The common theme amongst all of this data is that it has some level of checking (cleaning) that needs to be performed before it can be considered ready for analysis. In the case of data from a clinical data management system with comprehensive internal data checking tools, the level of checking may be minimal, perhaps limited to counting the numbers of records loaded into the clinical data warehouse. For data from other sources, the level of data checking may need to include metadata checking, data consistency checking and reconciliation checks

Table 1

Typical Data Warehouse	Clinical Data Repository
• Slow variance over time in source systems	• New studies brought on-line continuously to be integrated
• Constrained set of input variations	• Local variations in input content to support study data needs or providers
• Non-transactional in nature	• Active data processing, enrichment, adjustment and even creation
• Single governing schema	• Different schemas for different stakeholders (eg raw, SDTM-like, SDTM, ADaM)
• Well-controlled dimensional master data	• Multiple sources of master data to align, not always fully controlled (eg Sites)
• "As-is' data currency	• Historical "as-was" views must be preserved for audit, even as data changes
• Binary security rules	• "Shades of grey" security to preserve study blinds while allowing access



against other data sources. Clinical data warehouses are increasingly including built-in capabilities, such as discrepancy management modules, for performing these data management tasks.

Once raw data are loaded into a clinical data warehouse, it is often converted into one of a number of standard data models. The most common of which are the CDISC SDTM and ADaM standards. Clinical data warehouses support this activity in many ways. They provide an environment in which you create version-controlled data manipulation programs. They provide an execution environment where you can test and then execute the programs over production data. And most importantly, they provide data traceability such that it is possible to trace back from any version of the data through multiple layers of contributing programs and data to the original raw data. This enables prompt responses to regulatory enquiries. The features of the clinical data warehouse can be applied to create a robust and regulatory compliant security structure, facilitating data blinding.

As alluded to earlier, one of the key differentiation points in a clinical data warehouse is the fact that you are dealing with multiple clinical studies which are, by definition, even if only slightly, different. This adds an interesting complexity to dealing with the data in clinical data warehouses. You either need to create totally generic programming to allow for the differences between trials, or take generic programs and modify them slightly for each study, or impose strict standards for certain parts of trials such that generic programs can work for all studies. Typically, a clinical data warehouse contains libraries of re-useable generic code that can be applied to multiple clinical trials without modification. Taking advantage of this pre-validated code can reduce the validation efforts, and can potentially accelerate trial setup.

> “THE CHALLENGE IS BUILDING A FLEXIBLE PLATFORM WHICH CAN CONSUME, AGGREGATE, TRANSFORM AND ENRICH THE DATA IN SUCH A WAY THAT ITS MEANING REMAINS TRUE, AND THAT CAN BE ACCESSED VIA A VARIETY OF TOOLS TO ENABLE FULL DATA EXPLORATION.

Once data is conformed to a common, controlled data model, automated programs can pool data from different trials in like structures. The advantages of such a pool are numerous, and can be used across the business, allowing modelling and simulation groups to evaluate the viability of a prospective clinical trial, streamlining signal detection, and enabling meta-analyses across large cohorts.

Further, this data pool can be accessed by many visualisation, business intelligence, and analysis tools, from within the security framework provided by the clinical data warehouse, allowing in-stream review of data without compromising data blind or access control. Support for in-stream medical review and patient safety profiling should be supported by the clinical data warehouse, reducing programming input. Reporting is something that is common to all types of warehouses. Clinical data warehouses are often used as a data and code repository for regulatory bio-statistical reporting. The standard tool for this type of reporting is SAS, although many clinical data warehouses support alternative reporting tools such as Oracle BI Publisher. A common factor with reporting in clinical data warehouses is the ability to consolidate the output of multiple reporting programs into a single, collated output. The really clever part

of these consolidated reports is noticed when they are viewed in conjunction with the traceability features of a clinical data warehouse. Imagine that, at the last moment just before you are going to submit a report to the regulatory authorities, you discover a data point needs changing in one of the source tables for a report. A complex report for a regulatory submission may take many hours to run and the updated data point may be used in several places in the overall report.

Using the traceability inherent in a clinical data warehouse, it should be possible to reduce the time taken to run the overall report again by letting the system resolve the dependencies of data and programs and only run the programs that are directly related to the changed data point.

These are only a few examples of the tasks and business processes potentially affected by the adoption of a clinical data warehouse. Table 2 lists many other potential uses of a clinical data warehouse.

So there are many use cases for Clinical Warehousing. There is huge inherent value embedded within the warehouse. The challenge is building a flexible platform which can consume, aggregate, transform and enrich the data in such a way that its meaning remains true, and that can be accessed via a variety of tools to enable full data exploration. Realising the true benefit of clinical warehousing is often a multi-year cycle, and it's critical that any organisation embarking on a warehousing project has a clear definition of goals and benefits, and an understanding of dependencies and priorities, driving a phased implementation.

Building the right platform can be hard, but the potential rewards are huge, and a key differentiator in driving a clinical development program.

Iain Barnden, Director Clinical Data Warehousing, pharmaSOL (iain.barnden@pharmasol.de)
Jonathan Palmer, Senior Director, Clinical Warehousing, Oracle

Table 2

Potential uses of a Clinical Data Warehouse
• Ongoing medical review
• Data cleaning
• Data reconciliation
• Streamlined statistical analysis for submission
• Protocol design and trial simulation
• Responding to regulatory queries
• Safety monitoring and signal detection
• Cross-study analysis
• Single data storage for visualisation/ analysis tools

Jonathan Palmer, Senior Director of Product Strategy, Clinical Warehousing, Oracle Health Sciences
Jonathan Palmer has over 20 years of clinical experience. He began his career as a statistical programmer, then moved through the rapid growth of the contract research organisation (CRO) industry, ultimately being responsible for clinical solutions and support at Parexel. He has had numerous roles at Oracle and IBM, helping lifescience organisations realise benefits from implementing new clinical technologies. Jonathan is currently responsible for product strategy for Oracle Health Sciences Clinical Warehousing and Analytics solutions.

Iain Barnden, Director of Clinical Data Warehousing, pharmaSOL
Iain has worked in the pharmaceutical industry for over 20 years. He has experience with many industry-leading clinical database management, pharmacovigilance, clinical trial management, and data warehousing systems. He has taken leading roles in numerous implementation and strategic projects, and currently heads-up pharmaSOL's clinical data warehousing practice.