

## The Evolution Of Coding Technology

Richard Davies, Oracle, Reading, UK

### ABSTRACT

Clinical data coding was introduced to the life sciences industry over a decade ago to better ensure the safety of drugs in development, pushing forward technological innovations that enable open standards and global collaboration. Coding was once a tedious process wherein data managers worked from data listings and published encyclopedias to manually code medical terms. With technological innovation, coding has evolved into an automated process, bringing standardization and providing a centralized coding system across multiple global trials. Concurrently, the shift toward personalized healthcare combined with the growing global span of clinical trials, requires more insightful safety data to improve patient care. After providing key reasons for medical term coding, the paper details the evolution of coding technology. The speaker will discuss how the evolution of coding has supported the movement toward open standards and global collaboration to improve clinical coding processes.

### INTRODUCTION

Medical Coding has changed radically over the last 15 years, moving from an intensive, laborious, manual process to something that is highly automated and software driven. The technology that now supports medical coding has evolved over time, becoming increasingly sophisticated and also reacting to changes in the coding process. In this paper, we'll look at some of the milestones that have been reached during the last 15 years and finish on where technology has evolved to today, together with where it might evolve to in the future.

### THE EVOLUTION OF CODING TECHNOLOGY

#### WHY DO WE CODE ?

Initially lets look at the drivers for medical coding. Using some real CRF examples we can see how diverse the clinical data we collect can be. The diverse nature of the data being collected means that the same Adverse Event can be referred to in different ways, such as:

- ⤴ Headache
- ⤴ Head Pain
- ⤴ Pain in Head.

These may be exactly the same Adverse Event or they may be different. This drives the need to properly categorize the data so that we can clean it, understand it, identify interactions, report data and manage risk. Electronic dictionaries have been developed to help us code our clinical data and remove the ambiguities that can be introduced by those supplying the data to us. We also increasingly collect data in different languages, giving an additional problem when it comes to aggregating and categorizing our clinical data. Technology exists and can help us with these challenges. The technology is not new and so it has been through an evolution and maturation process. However, business requirements have also changed and continue to change in this area.

## **CODING REQUIREMENTS**

Having understood what drives the need to code clinical data we can look at some of the business requirements for technology that supports the coding process. Manually dealing with clinical data in an effort to consistently categorize it is supremely inefficient and leads to inconsistencies. Technology can support the process and over the last 15 years or so, software packages have been developed to remove the manual burden and increase to coding consistency. Ideally, what should coding technology be delivering to our business in order that the coding activities are as efficient as they can be ?. Here are some examples:

- ⤴ Timely
- ⤴ Consistent
- ⤴ Controlled
- ⤴ Manageable
- ⤴ Automated
- ⤴ Not a burden to our business

Coding data in a timely fashion is important because the coding process can raise issues that need to be clarified to the Investigator. Such issues can be significant, even leading to the withdrawal of a subject from the clinical trial or worse, impacting the safety of the subject. Hence being able to code data quickly is imperative. The coding process is designed to remove the inconsistencies and lead us to well categorized clinical data. This is critical to help us understand our clinical data, including the safety profile. This is not only important for the sponsors benefit, but also that of the regulators. Having a controlled coding environment is important because for a number of reasons, but especially the maintenance of the electronic dictionaries. We don't want the wrong version of dictionary version being used to code our data, and we don't want the dictionary modified or enhanced without control, so that the coding process will become inconsistent. The process should be automated so that as few people as possible should be needed to manage the coding process and as a whole, coding should not be painful in terms of cost, time or quality. Coding should not be a rate-limiting step in the clinical development process.

## **A JOURNEY THROUGH CODING (1993 - PRESENT)**

The business requirement to code clinical data existed long before technology was available to help support the process. When I started my industry career in 1993, coding was very much a manual, tedious and time-consuming process. Some of my colleagues in data management coded data by working with listings of data from our Clinical Data Management System and used paper reference manuals to code data manually. The time consuming nature of this process was clear but in the absence of any software solution, there was no other option than to use in-house skills to create a 'coding solution' using statistical programs or other CDMS functionality to create a rudimentary auto-encoding tool.

The gap in the industry was soon identified by some of the software providers and by 1996, off-the-shelf packages were appearing as integrated functionality in Clinical Data Management or Drug Safety systems. The capabilities were simplistic, but at least they introduced automation. Users could take electronic versions of the standard categorization dictionaries, and load the contents into the tables of their auto-encoding software. The tables extended to allowing synonyms to be defined by the end users, as well as defining 'noise' words that the auto-encoder would look for and remove from the source data before comparing the remainder against the dictionary to find a match. In this way, 'Head Pain' could be used as a synonym for 'Headache' and the auto-encoder would now happily begin to code our data consistently without manual intervention. However, the tools were specific to the applications they were built within and hence to the users of those systems. Data was being coded in silos and that led to reconciliation overheads. The concept of a Central Coding department was also evolving but no technology existed to support this type of user.

By 1998 the initial coding solutions were beginning to evolve to meet new user roles and more sophisticated business requirements. Central Coding groups had been established in many organizations to handle the challenges of coding data from different systems consistently, whilst managing the coding technology and the dictionaries that they utilized. Interfaces were now designed for this 'Central' user, giving access to data across all the clinical trials being run, rather than working on a trial by trial basis as had been done up to this point. Data was now being coded consistently across all the trials being run, not just within a single trial. The coding technology also included an electronic query facility that allowed Central Coding users the same ability to create queries on clinical data as their Data Manager colleagues, streamlining the creation, management and resolution of coding-related data queries. New electronic dictionaries had also emerged. Rather than just being simple lists of Terms, dictionaries now had a multilayer hierarchical structure that included more useful information for categorizing the clinical data. The end users had also been given some control over the auto-encoding algorithms their software was using to find matches against the dictionary. As well as being able to manipulate how the auto-encoder was looking for matches, they were able to introduce more sophisticated data processing steps such as word-stemming. In this way, the auto-encoder was able

## PhUSE 2011

to narrow words down to their stem, so 'includes', 'included' and 'including' would be stemmed down to 'include' for instance, helping to remove the tense of the verbatim text before matching it against the dictionary.

By 2006 there was new technology and new challenges to be met. With the adoption of the internet and the possibilities it brought, coding technology was evolving with thin-client web interfaces. This helped organizations implement a single tool that was globally accessible to all relevant stakeholders and lessened the IT overheads of managing the technology internally. The internet also brought the capability that technology could cross organizational and business boundaries. For instance, a Pharmaceutical company could allow a Contract Research Organization (CRO) to leverage its coding technology so the CRO users were using the same environment and learning that what already existed in the Pharma organizations system. It also facilitated review and approval processes across organizations, allowing Pharma companies to review CRO coding activities within a single tool, real-time. Coding technology had also expanded to make full use of the hierarchical information now embedded within the dictionaries, allowing coding or reporting of data on user-definable levels. More information was now being processed by the coding technology and was also available to the interactive users. Coding drug information can be challenging when different formulations of indications exist. Hence, just using the verbatim drug name from the clinical data isn't always enough to identify the product taken by the subject. In such situations, knowing the indication or the route of administration can additionally help the interactive coder when faced with multiple coding results to choose from, and such additional contextual information can be passed to and used by the coding technology. Some of the electronic dictionaries used for coding purposes also went through regular update cycles, meaning that any coding technology had to enable multiple versions of dictionaries to be implemented at any one time and potentially used concurrently. There were significant challenges regarding when and how a new version of a dictionary was used with clinical data, especially considering the update frequency of the dictionaries and the long duration of many clinical trials.

By 2008, coding technology was helping users with the complexities of up-versioning dictionaries and adopting new dictionary versions with their clinical data. Firstly, coding technology was able to compare two versions of the same dictionary, identifying changes to the published dictionary structures as well as taking into account any synonyms that users had added to the coding configurations. Once these differences were understood, the coding technology could look back across the data coded with the old dictionary version and spot any coded data that would require attention if the new dictionary version was applied, providing users with an actionable Impact Analysis report that could be used for recoding the impacted data. Additional enhancements came with the support for Japanese data and dictionaries. Japanese language can be written in different character sets and hence eCRF data can be supplied in different character representations. Coding Technology could normalize the eCRF data into a single Japanese character set before searching against the dictionary.

### The Future

Before concluding and looking towards the future of coding technologies, we can return to our Coding Requirements and see how well the Coding Technology of today has addressed those needs. Starting with data being coded in a timely manner, the fact that we can nowadays collect our clinical data from the sites via the internet, and have it coded potentially seconds after submission from the site, shows that the initial timely coding of data is easily achievable. This gives us the ability to review and react to the coded data and make timely decisions around it. Consistency however is still a challenge. This largely comes from organizations still using multiple coding technologies to code data in different places within their business. A prime example is the coding of adverse event information from the clinical databases and the coding of the same adverse event information in the safety database. A lot of energy is still put into the reconciliation of data between these two databases to ensure consistent coding between them. Controlling any single coding technology is also achievable with functionality such as Impact Analysis to even help control the adoption of different dictionary versions. Configurable rights and roles with these systems ensure that only privileged users are allowed to perform the more sensitive operations and maintenance. Automation has been achieved and was so early on the life cycle of coding technologies. So much so that many organizations quote in excess of 90% hit rates for their coding technologies. As far as the process being manageable is concerned, partial success can be claimed. Individual technologies provide excellent management of their respective coding processes. However, the different technologies in play, leading to data reconciliation overheads, mean there are management overheads to absorb as well and so there is still work to do. Coding still adds a burden to the business that has yet to be addressed because there is no single technology solution that can address the coding needs of clinical data across all lines of business. There is still work to be done as we look to the future and develop technologies further.

### CONCLUSION

So with work still to do, where should coding technology evolve to in the future ? Some organizations now look at coding not from a data management perspective or from a drug safety perspective, they simply see it as a fundamental need that is required for their data. Data must be coded consistently, with a single technology platform, whenever it is appropriate, either during capture, safety management or even further downstream when the data is stored in a clinical warehouse. They should also be able to up-version the coding to new versions of dictionaries whenever they like and regardless of source. Reconciliation should be minimized and automated. One way to look at this is to have a single piece of coding technology, available to all appropriate stakeholders, implementing consistent coding definitions to clinical data, regardless of original source of data or current storage location. Whether coding technology will evolve in this way we shall see, but evolve it must do, as there remain challenges to be addressed.

### CONTACT INFORMATION

(In case a reader wants to get in touch with you, please put your contact information at the end of the paper.)

Your comments and questions are valued and encouraged. Contact the author at:

Richard Davies  
Oracle, Health Sciences Global Business Unit  
Oracle Parkway,  
Thames Valley Park,  
Reading, RG6 1RA, UK  
Work Phone: +44 1189249732  
Fax: +44 1189249732  
Email: rich.davies@oracle.com  
Web: www.oracle.com

Brand and product names are trademarks of their respective companies.