

Written and researched by:
Dr. Kirk Mousley
and Karl Mousley

In the next issue of
EDC Today:

Edit Checks

About EDC Management:

EDC Management is the leader in Clinical and Data Management and Electronic Data Capture (EDC) consulting services for the biopharmaceutical industry. EDC Management publishes well-researched and timely information about Electronic Data Capture technologies and processes through EDC Today® and EDC In Depth. We do not sell or endorse any specific EDC software application or vendor. Improve process today; position for tomorrow.

EDC Management

P.O. Box 384
Conshohocken, PA 19428
484-530-0300 (voice)
610-567-0357 (fax)
info@edcmanagement.com
www.edcmanagement.com

Coding and Dictionaries

EDC Today is an independent publication on current information and issues in Electronic Clinical Systems (ECS) strategies and technologies for the Biotechnology and Pharmaceutical (Biopharma) industry. Each month we examine topics related to ECS theory, technology, practice, or implementation.

Recently the FDA rolled out its “Adverse Event Reporting System (AERS) ...a computerized information database designed to support the FDA’s post-marketing safety surveillance program for all approved drug and therapeutic biologic products.” A key component of this system is that “all reported adverse event terms are coded using a standardized international terminology, MedDRA (the Medical Dictionary for Regulatory Activities)”. At long last, the FDA has selected a dictionary as its preferred terminology dictionary.¹

In this issue, we discuss the reasons for coding clinical trials data that is captured in free-form text and how that coding is currently done. With the FDA’s selection of MedDRA, we suggest that the time might be near when a Biopharma can purchase an off-the-shelf autoencoding system for integration with their ECS. While there aren’t a lot of autoencoding systems yet available on the market, it seems prudent for a Biopharma to assess its autoencoding requirements and investigate the possible use of a commercially available system.

Introduction

Adverse Events or Adverse Experiences (AEs), prior or concomitant medications, and clinical procedures are often coded using one or more special dictionaries. AEs are usually “free-form” verbatim (textual) descriptions of what went wrong for a subject in a clinical trial. These descriptions might include numbers, technical medical terminology, foreign words, and even misspellings. In order for the AE to be used in an analysis it needs to be coded so that like AEs will be both grouped and counted together. An example of the need for categorization can be found on a drug label. Drug labels show percentages of trial subjects experiencing certain categories of AEs (e.g., dry mouth, headaches, and dizziness).

(continued on page 2)



Medications are also usually “free-form” text in nature, but the difficulty with medication naming is that a given medication might be known by a number of trade names, generic names, and compound names any or all of which might be used by the investigators participating in a clinical trial. An example of this is “aspirin”, the use of which by a subject might be recorded by the investigator as “Bayer”, “Excedrin”, “Aspirin”, “Acetylsalicylic Acid” or any other of a large number of acceptable names for “aspirin”. In order to perform analysis on medications, it is necessary to code them so that like medications can be grouped and counted together.

Another type of clinical trials data that are often collected as free-form text are clinical procedures, such as “Bypass Surgery”. And like AEs and medications collected in free-form text, clinical procedures often are known by a number of names and thus need to be coded in order for them to be properly analyzed. Unlike AEs and medications however, clinical procedures can sometimes be collected using a codelisted item. This is because the analysis of a specific clinical trial might only concern itself with a relatively small number of clinical procedures such as only those that are cardiac-related (e.g., a specific class of clinical procedures that might be considered “endpoints” for a study).

AEs and medications are rarely collected using codelisted items for a number of reasons. One reason is that the codelists would have to be very large and thus unwieldy for the person performing data entry to use. Another reason is that the codelist would always be incomplete, and updating a codelist to add new entries during the course of a study may not be desirable or even feasible within an EDC or CDMS application. But the most important reason might be that the use of a codelist, giving the investigator a fixed list of possible entries, could be construed as “biasing” the investigator as they would feel compelled to use an entry on the current list as opposed to describing an AE or medication in their own terms.

Over the years, as computer processing power grew, many Biopharmas attempted to convert the process of manually coding AEs, medications and clinical procedures to an automated process performed by the computer. This automation is known as autoencoding.

Background

In the past, Biopharmas had to develop Adverse Event and Medication autoencoding software for their own use. There were at least four reasons for developing and using proprietary solutions:

1. There was no universally accepted or regulatory mandated terminology dictionary. Each Biopharma selected a dictionary that it preferred to use or developed a custom dictionary. The FDA accepted a wide variety of coding and coding schemes.
2. Even if a dictionary was available from a standards body (such as COSTART and WHOART), many Biopharma would, to some extent, modify or customize their copy of these dictionaries for a given protocol or therapeutic area, believing that such modifications improved the “accuracy of the coding” for the given protocol or therapeutic area.

(continued on page 3)



3. Many Biopharma that perform autoencoding developed unique strategies for coding terms in an attempt to reduce the number of autoencoding failures. These strategies are perceived as business advantages so they remain proprietary in nature and are not shared in any development of an autoencoding solution available for licensed use by others.
4. Each Biopharma that performed autoencoding developed a unique process for resolving verbatim terms that did not automatically encode resulting in coding failures. The resolution process had to be tightly integrated with their coding strategy. For instance, if they used an “ignore word” list, this list might be updated in order to resolve an encoding failure.

Coding - Automatic or Manual?

To start with, one really needs to consider whether the coding of AE, medication names, and/or clinical procedure names should even be automated. Certainly the manual process can be drawn out and involved, and dictionary lookups and coding decision making can be tedious. Moreover, some clinical trials might capture hundreds or thousands of terms to be encoded. However, automating a process that involves specialized human knowledge is a difficult process at best. While a medically trained individual can look at a verbatim term and sometimes immediately decide how to code it, it may be difficult for a computer to be programmed to do the same. One source of difficulty is that the verbatim term may contain regional spelling or even misspellings. Some automated coding strategies employ a set of replacement words. For example, one might want to replace “centre” with “center” to allow it to code. Clearly, such a list of replacement words could be extensive.

How extensive an autoencoding algorithm is needed or desired will depend in part on the number of verbatim terms that need to be coded. If there are not a large number of terms, it may be more cost effective to perform manual coding. If there aren't huge numbers of terms to be encoded, a simple direct match algorithm might be sufficient. Only in the case where there are large numbers of terms to be encoded (and more are expected over a fair time span), it might be cost effective to develop (or employ) a more sophisticated autoencoding algorithm.

Pitfalls to Autoencoding

Once an autoencoding system is implemented, the need for medical specialists to perform coding is not ended. Even in the case when a verbatim term successfully autoencodes, there is a need for someone to verify accuracy of the encoding. Worse yet, no matter how sophisticated the autoencoding algorithm, there will always be verbatim terms that cannot be autoencoded, so there will always be a need to manually resolve these coding failures. This may mean revising the verbatim terminology, updating one or more of the autoencoding supporting dictionaries (e.g., “replacement words”, “ignore words” and perhaps others) or even adding an entry into the main dictionary, which for AEs is now usually the MedDRA dictionary.

(continued on page 4)



In any event, there is a need to develop, implement, and document a branch in the overall data processing workflow loop. In many Biopharmas, a whole group of medical personnel familiar with the dictionary and autoencoding algorithm, often known as “the dictionary group”, are tasked with maintaining the electronic dictionaries, reviewing autoencoded terms for accuracy, and resolving encoding failures. Obviously, encoding terminology remains a necessary but costly proposition.

Multiple Dictionaries

Many Biopharmas have different versions of dictionaries per study agent (drug) classification, protocol or study, or even per therapeutic area. Also, these biopharmas may even have multiple versions of MedDRA, WHO Drug, WHOART, ICD-9, or a self customized or created dictionary. Biopharmas often justify protocol or therapeutic-based versions of dictionaries by stating that specific dictionaries allow more accurate coding. EDC Management is not convinced using protocol or therapeutic-based versions of dictionaries are a good idea. Certainly one has to be concerned about introducing bias in the resultant coding.

Any autoencoding system should run independently of whatever dictionary is used, within certain limits. Dictionaries supporting medication coding are different than dictionaries supporting AE coding (and often the autoencoding algorithms are different between medications and adverse events). However, the autoencoding system should be able to use different dictionaries for adverse events interchangeably, especially if it is offered to other users on a licensed basis.

Hit Rate Concerns

When discussing autoencoding systems, the term “hit rate” is defined as the percentage of verbatim terms that successfully autoencode. The hit rate should be as high as possible, that is, it should be as close to 100% as possible, without causing errors by incorrectly coding. The more elaborate and sophisticated the coding strategy, the more likely that the computerized interpretation of the original verbatim term will be changed to something that is not intended and no longer reflects its original meaning. Clinically valuable information in the verbatim term is maybe disregarded by autoencoding systems that are too aggressive in attempting to raise the hit rate associated with their use. It is preferable to be safe and not code a term (and have an expert manually code the term) rather than code to an incorrect term.

Direct Match/Match a Synonym

The direct match of a verbatim term with a dictionary entry or a synonym table might be the safest approach for autoencoding. This approach minimizes the introduction of coding errors. Some Biopharmas are skeptical about the efficiency and accuracy of autoencoding algorithms and actually prefer to only do direct match autoencoding and manually code all non-direct matches. It is hard to argue with this approach unless the Biopharma adds a large number of similar dictionary entries in an attempt to raise the hit rate. Having such records in the dictionary means that the Biopharma must maintain a custom dictionary, something the FDA seems to be slowly indicating is not a desirable practice.

(continued on page 5)



Coding Strategies and Natural Language Processing (NLP)

In order to improve the hit rate of their autoencoding systems and to reduce the number of terms that need encoding failure resolution or manual coding, many Biopharmas have designed sophisticated systems that go beyond a direct match or a “match to a synonym” process that involves applying one or more coding strategies (rules) and possibly even Natural Language Processing. Among the various strategies that can be used by autoencoding systems are removing words that don’t add meaning to the verbatim term such as the word “the,” replacing alternately spelled or misspelled words with a correction, ordering key words (i.e., the words leftover after “meaningless words” are removed from the term) and applying “sounds like” functions to some words. There are many other strategies, all of which are aimed at reducing the investigator-supplied terminology with a phrase that has the fewest possible words needed to retain the term’s original clinical meaning, albeit sometimes the words will be in a strange order!

While Natural Language Processing has made strides in the computer science world and is familiar to many in the form of the Microsoft Office suite member’s (e.g., Word and Excel) “Clippit” help animation, it is likely to be too probability-oriented to be used in coding strategies. EDC Management is not aware of a Biopharma or CDMS Vendor using NLP in an autoencoding system, but this technology will almost certainly be used in the future.

No known autoencoding system currently takes into consideration the clinical context. Clinical context is often useful when coding terms since references to laboratory test results, physical exams, medical history, EKGs, and so on, may be the deciding factor in how a term gets coded. Even with manual coding systems, the absence of clinical context may make it difficult for the medical expert to properly select an appropriate code. A report encapsulating the clinical data that most often supplies clinical context for terms that are to be encoded is usually a most useful thing.

Autoencoding Failure Resolution Process

With the FDA’s selection of MedDRA, perhaps now is a good time for a Biopharma to consider purchasing an off-the-shelf autoencoding system for integration with their ECS. Although there are not many autoencoding systems currently available for purchase, it seems sensible for a Biopharma to evaluate its requirements and the advantages of a commercially available system.

Formulating an efficient and manageable process (workflow) might be the most important part of creating a successful encoding system (be it automatic or manual) and here is where most Biopharma’s solutions diverge from being uniform and this will be the area in which a commercially available product will face one of its greatest challenges.

Sometimes the difference in how an autoencoding failure is resolved is due to the size of the Biopharma’s organization and its infrastructure. Some Biopharmas have complete standalone groups dedicated to maintaining the various corporate dictionaries. Some even go so far as to have separate groups dedicated to each therapeutic area in which the biopharma is conducting clinical trials. The smaller ones may not have a formal organization dedicated to dictionary maintenance. The autoencoding system will need to be flexible and support segmented access roles of user that span a number of groups.

(continued on page 6)



Building or Buying an Autoencoding System

Even if most Biopharmas would agree that a direct matching or matching to a synonym table is the safest and most effective way to autoencode terms, then the ability to process verbatim terms that do not match, and constitute a coding failure, will be of significant importance. However, a lot of Biopharmas think a more sophisticated autoencoding algorithm is necessary. So any commercially available autoencoding system will need to be implemented taking either the simplicity of a direct match, single dictionary, single dictionary maintainer approach, or the much more costly, sophisticated, but highly user tailorable, autoencoding algorithm, multi-dictionary, multi-dictionary maintainer approach.

From these observations, one can view proposed autoencoding solutions in terms of how well they support multiple dictionaries, how well they support your Biopharma's specific workflow process relating to coding failures, and how well they support unique strategies for coding.

Autoencoding solutions must support changes to verbatim terms, dictionaries, and coding software or strategies. Thus, an ideal autoencoding solution requires the tracking and maintenance of one or more status flags during the coding process.

For example, if a verbatim term is modified, the corresponding code may need to be modified. One potential solution for dealing with this situation may be to delete the previous code. Then on the next autoencoding pass, all verbatim terms with missing codes are autoencoded. This will permit the modified verbatim term to be re-coded.

Changes or updates in dictionaries, for example going from COSTART to MedDRA or even going from MedDRA version 7.0 to version 7.1, need to be handled with care. During the life of shorter duration studies, it may not be sensible to change or update a terminology dictionary. Ongoing studies of long duration means changing or updating the dictionary will be inevitable. A process must be developed to locate and determine what previously coded records, if any, will need to be re-autoencoded to reflect the new version of the dictionary.

Changes to autoencoding strategy should probably be minimized, at least while a particular study is ongoing. However, coding tools may not be capable of supporting multiple coding strategies on a protocol-by-protocol basis. Therefore, a full impact assessment of the autoencoding strategy change should be performed, and the new strategy validated before implementing any changes.

Availability of Autoencoding and Dictionary from Vendors

The move towards standardizing dictionaries (specifically MedDRA) will help improve the availability of autoencoding products. However, issues remain that have prevented widespread adoption of autoencoding solutions in the past. One such issue is the perceived need of Biopharmas to create a custom dictionary view per study or protocol. It is not immediately evident how a commercially available autoencoding system will handle Biopharma dictionary personalization and dictionary modifications. Obviously, the more sophisticated autoencoding systems will need to be enormously flexible.

(continued on page 7)



The extent that Biopharmas continue this practice will make it harder for a general encoding product to be successful. A Biopharma's physicians (medical monitors) will always have their own view on how something should be coded, which may make using a commercially available autoencoding system cumbersome and difficult.

Conclusions

With the FDA's selection of the MedDRA dictionary as the preferred dictionary for verbatim term encoding, it may be possible for Biopharmas to buy off-the-shelf AE and medication name autoencoding systems that work well in their EDC or CDMS environments for a considerably lower cost than that of systems developed in-house. There are some significant hurdles for both the autoencoding system vendor and the Biopharmas that will need to be surmounted. Both the autoencoding failure resolution process and the autoencoding algorithm will need to be flexible and able to be customized to meet the perceived needs of the Biopharma users.

Obviously, there will be no way of getting around the need for important clinical trials data that is captured as free-form text. Computer software may eventually reach the point where it can successfully determine how to codify terminology captured in this cumbersome format in a large majority of the time. As with many human endeavors, the computer remains a powerful tool to achieve timesavings and realize efficiency improvements, but it remains a tool that requires some intelligent forethought as to how it best be applied to a business need.

Readers are encouraged to share their recommendations for how vendors can improve their products. We would also like to encourage feedback and suggestions to this issue, and welcome suggestions of topics for future issues.

References

¹ <http://www.fda.gov/cder/aers/default.htm>



Who's behind the research?

Our lead researcher, Kirk Mousley, PhD received BS and MS degrees in Electrical Engineering from MIT and a PhD in Computer Science from Lehigh University. He has been the President of Mousley Consulting, Inc. since its founding in 1993 and has directed the company's efforts in the areas of clinical database design, data editing/cleaning, document management, and submissions.

Karl Mousley received his BS in Mechanical Engineering from Rose-Hulman Institute of Technology and a MS in Computer Science from Villanova University. He has been a senior member of the technical staff at Mousley Consulting, Inc. since 1993. Among his significant accomplishments are the investigation, evaluation, and implementation of new computer technologies for clinical data management systems and developing strategic plans for integrating these technologies into current systems. He has extensive experience preparing Standard Operating Procedures (SOPs).



EDC Today and EDC In Depth

EDC Management publishes well-researched, timely information about EDC technologies and processes.

EDC Today is a free electronic technical bulletin.

Each month we examine topic areas related to Electronic Clinical Systems (ECS) theory, technology, practice, or implementation.

Each *EDC In Depth* research report comes with an executive summary and may be purchased individually for \$395 or as a group of related reports for \$975. Available via downloadable electronic version or paper version sent via mail.

To subscribe to ***EDC Today*** or purchase a specific ***EDC In Depth*** research report:

Order online at
www.edcmanagement.com

Email us at
info@edcmanagement.com

Call us at
1-484-530-0300