Paper 158-2007

Can Coding MedDRA and WHO Drug be as Easy as a Google Search?

Sy Truong, Meta-Xceed, Inc., Milpitas, CA
Na Li, Pharmacyclics, Sunnyvale CA

ABSTRACT

In a world where information can easily be accessed by applying a Google search, mapping unstructured clinicians term or verbatim terms from an adverse event or a drug name to a standard term no longer needs to be an arduous task. One of the challenges of working with coding medical terminology is combining several skills that are diverse from many different disciplines. The user needs to be clinically trained to understand and interpret the meaning of the adverse events or the drug names. A conceptual understanding of the normalized database and the multi-axial hierarchical structure is required to navigate the dictionary. The user must also be adept at adding to the source data and joining the case report form data with the proper fields of the dictionary tables to derive at the final mapped data. Expecting users to overcome these hurdles without a clear process or tools can lead to an error prone, laborious and painful process. This paper will address many of the issues confronted when coding terms by demonstrating tried and true methodologies and technologies to automate and make the process efficient and easy.

1. **Auto Coding** – Auto code with existing dictionaries against source data for efficient mapping.
2. **Google like Search –** Searching for terms in the dictionary or mapping decision history can be as simple as a Google Search.
3. **Optimize Coding Decision –** Intelligence search recommendations, review process, and managing split terms are some techniques used to optimize.
4. **Optimize Dictionary** – Loading source dictionary from MSSO (MedDRA) or UMC (WHO Drug) in an optimized fashion for performance.
5. **Managing Multiple Dictionaries** – Organize dictionary in a centralized and hierarchical manner to have consistent coding decisions.
6. **Build Knowledge Base –** Manual code only once and the term will be added to a knowledge base for future coding.
7. **Create new Mapped Data –** Techniques for creating mapped data sets with the use of email to make the process seamless.

It is essential that you have a consistent thesaurus dictionary when performing an analysis on clinical terminologies. This paper will show processes along with SAS based software solutions such as Sy/Map™ to allow clinical users to function optimally with data managers and clinical programmer analysts. Armed with the understanding of the process and the right tools, you can narrow the gap between the different disciplines required to perform mapping decisions in a manner that is as easy as applying a Google search.

INTRODUCTION

The coding of patient data is critical in the grouping, analysis, and reporting of data. Coding decisions directly impact submissions for New Drug Applications (NDAs), safety surveillance, and product labeling. The success of a submission to the FDA can be significantly impacted by the analysis of adverse events, medical history and concomitant medications. The analysis relies on the interpretation of what has been transcribed from the subject CRF (Case Report Form). The original clinical term is referred to as the clinician's term or verbatim term. This term needs to be re-interpreted or coded into a preferred term in order for it to be used during analysis. This is because different verbatim terms can have the same meaning such as in the example of the term "pain in head" or "headache". In this case, the two distinct verbatim terms are coded to one synonymous preferred term. The identical terms and the consistent classification of the term allow the analysis to draw valid statistical conclusions pertaining to the subject's experience. The coding process can therefore affect the statistical interpretation of the adverse events or medications in which the subject is taking during the clinical trial.

There is room for inconsistency or error since the process contains many factors that go into making a decision. The following considerations are evaluated in making the optimal interpretation of the true meaning of the clinician's term.

- **Clinical Accuracy** – The interpretation of the meaning of the original term has to be clinically correct. In addition to the original term, other supportive information in the CRFs (e.g. drug indication and strength) is also used to ensure the accuracy of the mapping decision. The person performing this task needs to be clinically trained to decipher the meaning of the verbatim term as it relates to the preferred terms.

- **Global or Project Specific** – The coding decision of a term in one specific project can be used again on other projects. It is therefore important to keep a global perspective while making a decision. However, there are instances where a coding decision needs to be applied specifically to special circumstances of the project.
- **Patient History –** It is useful to look at the clinical history of the patient in order to understand what led up to the current situation. This allows the clinician to have a historic understanding and therefore make a more accurate interpretation of the terms. However, the decision cannot be subject specific since this needs to be applied to all subjects.
- **Dictionary Update** - Understanding the structure of the dictionary and keeping up with the changes to the dictionary is critical for the success of mapping terms.

There are many factors that affect the interpretation of a clinical term and therefore the process becomes very complex. Besides the decision process, there are other operational and logistical considerations. The original clinician term can contain multiple terms so it needs to be split into separate distinct terms. This will therefore be coded separately. There are different versions to the dictionaries so version control becomes very important. There are many team members involved in this effort so training and standard operating procedures need to be established in order for the team to work together consistently. This multi-faceted process is complex but once a process is established, everything can work together in harmony so that terms are coded systematically and accurately to produce efficient results.

HISTORY AND STRUCTURE OF STANDARD DICTIONARIES

DICTIONARY FOR ADVERSE EVENTS

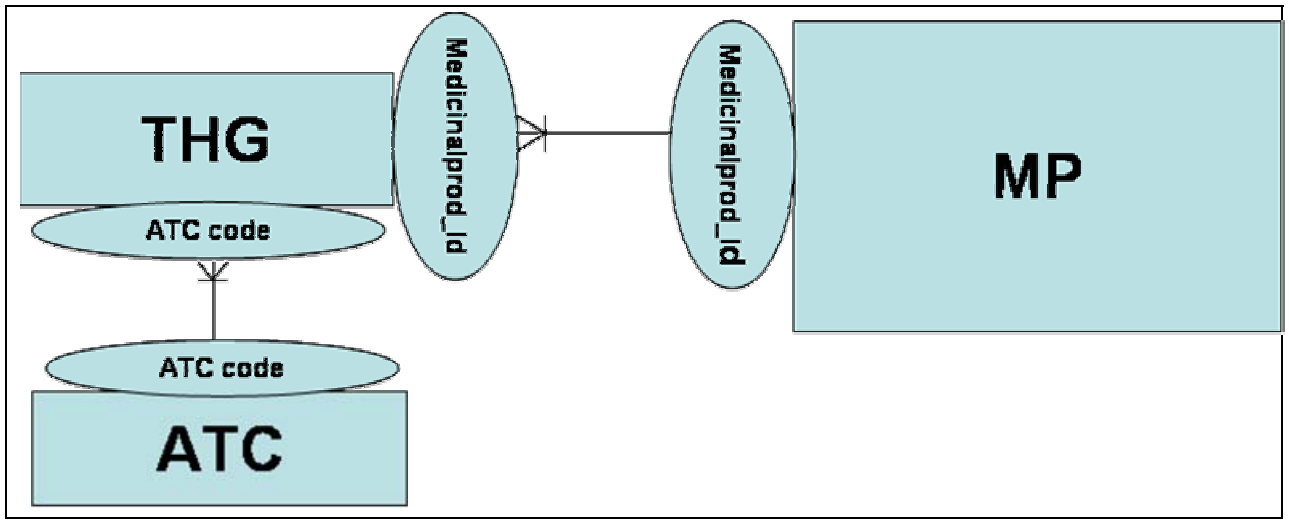
For statistical purpose and public health control, the need for an internationally accepted classification system was recognized at the First International Statistical Conference in Brussels in 1853. The standardization effort bore fruit in 1969 with the FDA's use of the COSTART (*Coding Symbols for a Thesaurus of Adverse Reaction Terms*) dictionary. For the next quarter century, adverse event coding was dominated by WHOART (*WHO Adverse Reaction Terminology,* required by EU) and COSTART (required by FDA). For regulation purposes, in 1994 the ICH (International Conference on Harmonization) began to standardize international medical terminology in all the phases of the regulatory process. In 1997 the first version of MedDRA (*Medical Dictionary for Regulatory Activities*) was born. The FDA recommends that adverse events are coded in MedDRA. This dictionary was also required by EU starting in 2003. MedDRA became commonly used to translate the Adverse Event verbatim terms into standard terms. It has a distinctive five level hierarchy structure ranging from System Organ Class (SOC), High Level Group Term (HLGP), High Level Term (HLT), Preferred Term (PT), to Low Level Term (LLT). The hierarchy allows the dictionary to accommodate both linear and multiaxial coding.

DICTIONARY FOR CONCOMITANT MEDICINE

In the pharmaceutical industry, two types of dictionaries are commonly used. One type of dictionary is used to standardize the adverse event terms while the other type of dictionary is used to standardize the drug terms. WHO-DD (*WHO Drug Dictionary*) is commonly used for the standardization of drug names. Drug information has been entered into the dictionary since 1968. WHO-DD is the world's most comprehensive dictionary for medicinal product information. It is used by pharmaceutical companies, clinical trial organizations and drug regulatory authorities for identifying drug names, active ingredient(s), therapeutic usage, and frequently used drugs in most major drug markets in the world. The organization that maintains this dictionary is named UMC (*Uppsala Monitoring Centre*). UMC organizes drug information by classifying them into codes according to the Anatomical-Therapeutic-Chemical classification (ATC). This allows for grouping of drugs in different ways for useful analysis. The information is also classified according to the main therapeutic use of the main active ingredient on the basic principle of only one ATC code for each pharmaceutical formulation.

LOADING EXTERNAL DICTIONARIES

Usually dictionaries are provided in modules (files) stored in ASCII format. This provides flexibility for putting together dictionaries based on specific needs of different organizations. The modules usually can be downloaded through the internet with license. For example, in our case, we construct an in-house SAS based WHO-DD by downloading these modules and importing these ASCII data into a SAS based database. The construction diagram for the WHO-DD type C is shown below.



Medicinal Product (MP) module provides the main part of the WHO-DD. The generic drug name is provided in this module. Therapeutic Group (THG) module provides a bridge to link the MP module with the ATC module. The common field called ATC code from THG and ATC is first used to join these two modules together so that the ATC classification information such as ATC level and ATC text can be obtained. And then the common field of Medicinalprod_Id from both MP and THG is used to bring the MP information in to form our dictionary database. There are other modules which are provided by WHO-DD, such as Pharmaceutical Product (PP), Ingredient (ING), and Substance (SUN). The main modules include MP, PP, THG, and ING. Other additional modules are available as subsidiary files. Depending on the company's pharmaceutical interest, these modules can be used accordingly.

Once we decide which modules are needed for the dictionary, these modules (files) need to be converted into SAS data set. Using SAS data steps with INFILE statement we can easily accomplish this task. The sample below shows the SAS code which converts the MP module into a SAS data set giving the name as MPpart.

```
filename mp "\\sample\path\mp.txt";

*** Get the data from medicinal product (MP)***;
data MPpart;
   infile mp missover pad;

   label med_id="Medicinal Product ID" drgnum="Drug record number"
         seqnum1="Sequence number 1" seqnum2="Sequence number 2"
         generic="Generic" drugname="Drug name"  sourcode = "Source code";

   input @1 med_id 6. @7 drgnum 6. @13 seqnum1 2.
         @15 seqnum2 3. @18 generic $1.
         @19 drugname $80. @144 sourcode $3. ;
run;
```

After the conversion of all these required modules into SAS data set, the dictionary can be established by following a structure such as the above diagram. The converted SAS data set (modules) can be joined by the prescribed key fields from the diagram. The join of the modules might be a one to one relationship or it can be a one to many or even a many to many relationship. It is therefore important to gain a good understanding of the structure in order to perform the join correctly. In the following example, a one to many relationship join is performed:

```
data THG_ATC;
   merge thg atc;
by atccode;
run;

*** med_id is the rename of Medicinal Product ID ***;
data MPTHGATC;
   merge THG_ATC mp;
   by med_id;
run;
```

MAPING METHODOLOGIES

After the SAS based customized dictionary is established, the mapping can be preformed. There is a series of steps that needs to be performed in order to have your verbatim adverse events or drug names coded to the synonymous preferred terms. This section will describe the methodologies used to effectively manage thesaurus dictionaries and code the verbatim terms.

STEP 1: MANAGE DICTIONARIES

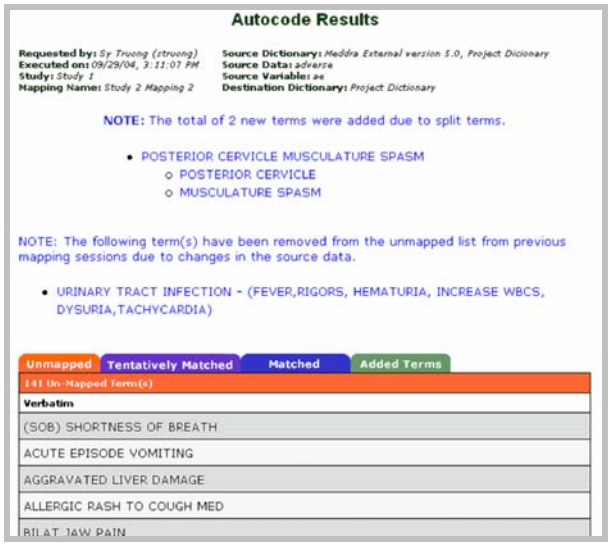
Before individual terms can be coded, thesaurus dictionaries need to be organized and managed. You would need to first identify and classify the types of dictionaries. The types of classifications for dictionaries include:

- **Dictionary Class** - Example classifications of dictionaries include WHO Drug, MedDRA or Costart. This describes the content of the terms pertaining to drug or adverse event names and how it is related to an "external" dictionary. The word "external" in this case means a level of dictionary which will be described in more detail in the next section.
- **Level** – A dictionary can be globally applied to all studies or it can be used specifically for a certain study. It can therefore be distinguished as either "global" or "project specific". These two types of dictionary levels pertain to terms managed internal to your organization. This is different from the dictionaries that are managed by an "external" organization such as MSSO who manages MedDRA. These external dictionaries are updated regularly by an external vendor and are managed differently from internal project specific or global level dictionaries.
- **Attribute –** Administrative attributes or metadata that describe the dictionaries are also essential in managing dictionaries. This includes values such as a unique name, physical location where associated dictionary files are located, and data sets name that stores the dictionary.

The classification information mentioned above needs to be managed in a way which allow users to register new dictionaries in the event that a new version of the dictionary is made available. Modifications to existing information are also necessary in the event that the underlying metadata is changed. The deletion of an existing dictionary can also be applied. Note that this does not necessarily mean that you are deleting the underlying data sets which store the content of the dictionary, but rather just removing the registered information or metadata. External dictionaries are managed by the vendor but internal dictionaries such as global or project specific need the capabilities of having old terms retired. This means that when a specific coding decision based on a term in the dictionary is no longer valid, it can be removed from the dictionary by a designated administrator. These are some of the tasks that are necessary in managing thesaurus dictionaries in order to optimize the performance of coding terms.

STEP 2: AUTOCODING

Once all the dictionaries are managed, you can merge the verbatim terms from the source data, which usually comes from the Case Report Form, against the dictionaries to determine if there are any matches. This is referred to as auto coding since it automates the comparison between source data and dictionaries. It will then produce reports detailing those terms that matched along with those that did not match. An example report below displays the results of an auto coding process by showing how some terms from the source data matched and some terms did not match. The results can be easily navigated through a series of colored tabs. The list of unmatched terms will determine how much manual coding you will have to do.



The auto coding can be performed against multiple dictionaries before a match is found. For example, it will start by merging the source data against the external dictionary such as MedDRA. If no match is found for a particular term, it would then continue to merge against the global dictionary and then the project specific dictionary in order to find the best match. The order in which the merge is performed is established in a "session" before the auto code is performed. Some of the information captured in a session includes:

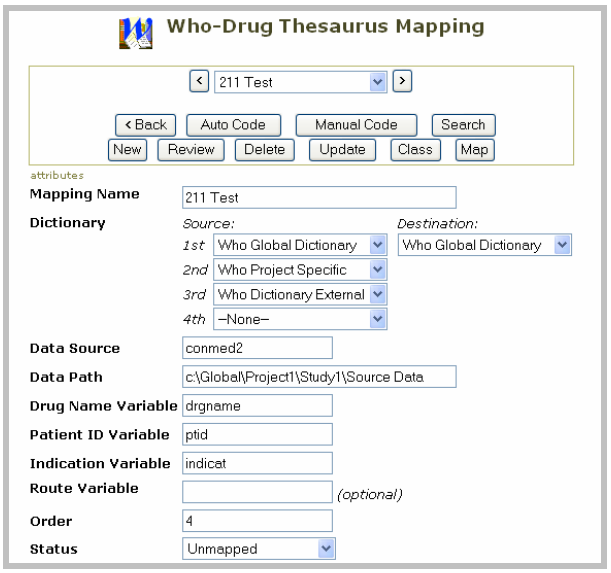
- **Name –** This is a unique name that distinguishes one session from another in the same study.
- **Source Dictionaries –** This lists all the dictionaries that will be used to merge the source data against the dictionaries in order to find a match. Note that there can be multiple dictionaries at different levels. The order is significant in that if a match is found in a dictionary listed first, it will not continue to search through the subsequent dictionaries to find a match.
- **Destination Dictionaries –** This is a dictionary that is populated with terms made by the user. It is a knowledge base that you build so that the next time an auto code is performed, this growing knowledge base stored in the destination dictionary would have more matches.
- **Data Attributes –** This contains information such as where the data is located and what column name in the specific data set which corresponds to terms such as drug name or verbatim adverse events.

After you have prepared the loading of the external dictionaries, you need to check how many verbatim terms from your source data match with the dictionary. If you are starting out for the first time, most likely, the matching results are small. There are strategies for increasing the matches so it will save you time from performing manual coding. These include:

Auto Coding Strategies	Description
Internal Dictionaries	Creating an internal dictionary is complementary to an external dictionary. This will keep track of any matches you may perform manually so that future auto code will use this knowledge base to match.
Order Internal Dictionaries	You may establish multiple internal dictionaries. This can distinguish coding at the global or project specific level. It will increase the accuracy of your matching for the project that you are working on. You would then order the priority in your auto code process where it may try to match up project specific coding and then if it does not find a match, it goes to the global dictionary.
Email Results	To facilitate the efficiency of the auto code process, you can have this emailed to the entire team with the results for review.
Scheduling	If you know that your source data will be updated on a regular basis, you can schedule this process so that it will perform this auto code processing upon these updates.

There is a series of steps that can be applied to perform an auto code. Some of these processing procedures are applied during setup while others can be done repeatedly upon each auto code request.

SETUP MAPPING SESSION – Before performing an auto code, you would need to determine where the source data is located and which dictionary to apply the auto code to. This information is referred to as a "Mapping Session".



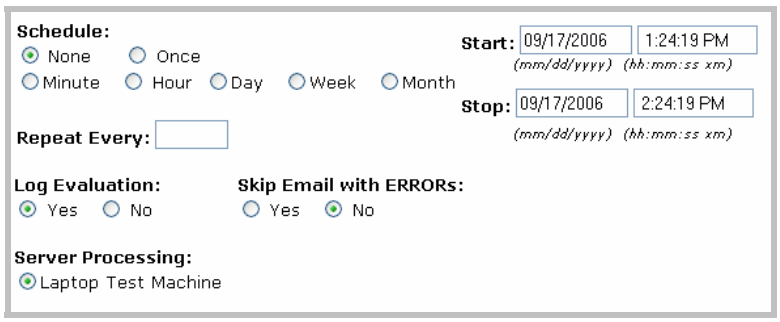
In this example, the source dictionary can be the same as the destination dictionary. This is an example of the knowledge base that you are building. The more terms that you add to the destination dictionary, the more "knowledge" will be applied to the merge for the next merge. This means the more terms you build in your dictionary, the smarter the auto code process will be.

The process for WHO Drug and MedDRA is very similar in this case. The key information includes:

1. Mapping Name – This is a label that is used as a unique identifier that you can refer to. It can contain a similar name to the study or some variation.
2. Source Dictionaries – This is a list of dictionaries used during the matching. The order is critical since this is the order in which it will try to find a match.
3. Source Data- This includes all the attributes pertaining to the source variables. It will specify the "verbatim" variable. In the WHO Drug case, it will be the variable pertaining to the drug name while the MedDRA case is referring to the adverse event verbatim variable. The other distinction is that the WHO Drug can also have an indication and route variable. If these variables are available, the matching decision will be applied by these variables.

PREPARE EMAIL – The results of the auto code are delivered to the requestor among other users within the team. You can therefore decide whom to send the results to in very much the same way an email is composed.

AUTOCODING OPTIONS – You can optimize and automate the auto coding session by having the process scheduled to be repeated at specified times. You can also have it skip sending out emails to the team if it encounters errors but it will always send the results to the requestor. For optimal performance, the process can be performed on a separate machine that is not busy.



In the above processing, only SETUP MAPPING SESSION procedure has to be done once. Once that has been established for your study, you can re-apply the subsequent procedures without having to repeat the initial steps. It is recommended that you incorporate processing procedures similar to these into your training and SOP so that your team will perform auto coding in a consistent manner.

The auto coding process will utilize all the information established in the mapping session in order to perform the merges to find matches. This process can be performed on demand by the user or it can be scheduled so that if the source data is periodically updated, an auto code will generate the most up to date report informing you of all the matches. The report can also be emailed to the user upon completion since these tasks may be performed asynchronously to the user's request.

STEP 3: MANUAL CODING

The auto coding process will attempt to match the verbatim terms against all the specified dictionaries that have been established for that session. However, a subject can record unique adverse event terms or have drug names that are unexpected and therefore will not match. In this case, manual coding needs to be performed in order for the user to evaluate the term and find a match against one of the preferred terms in the dictionary. The basic steps in manual coding include:

1. **Searching –** Similar to a Google search, the user can type text criteria to search through all specified dictionaries to find a match for the unmatched terms. There can be optimization algorithms which capture texts that are similar in the English language so even if it is not an exact match, it can show up in the result. This is similar to when Google suggests a "Did you mean:" followed by the text that is close to the text which the user typed.
2. **Entering Notes –** The user can enter a short message documenting how this decision was made. This adds context and meaning to the coding decision so that if it were to be reviewed, the reviewer would understand the thought process that went into this decision.
3. **Coding Type –** A manual coding decision can be of type final or tentative. If the user is confident about the decision, it can be assigned to be final. In this case, the verbatim term matching up with the preferred term will go directly to the destination dictionary to build a larger knowledge base. However, if the user is unsure of the coding decision, it can be applied as a tentative coding type. In this scenario, all the information collected for this decision is placed into a holding area for another reviewer to confirm before it will be placed into the destination dictionary.

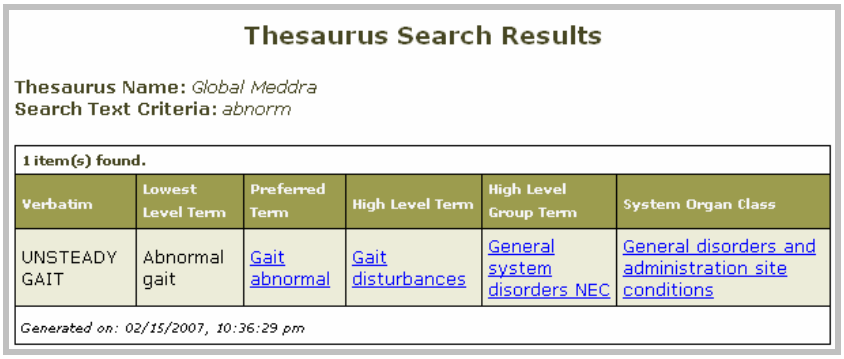
Associated information relating to a drug coding decision such as in the example of WHO Drug can be different as compared to MedDRA. In WHO Drug for example, there are other considerations such as indication of the drug, route and other drug information such as International Generic Name (INN), US Generic Name (USAN) and associated ATC Code. On the other hand, in the manual coding process for MedDRA, the user would need to evaluate the hierarchical information of the term including SOC, HGLT, HLT, PT and LLT. The content and information stored in each dictionary is different. It will therefore affect the decision process during manual coding. In general, however, the basic steps of searching for a matching term and then making a final decision are the same for all manual coding processes.

The coding decision made for each term is intended to be used for other situations that contain the same term. That is to say that a particular adverse event for example which is matched up with a preferred term can later be used for a different subject in a different study. The goal is to make the coding decision work universally for all situations. There are instances, however, that by just looking at a single term does not give the user enough information to make that determination. A common report can therefore be generated, referred to as a "patient report", which will list the history of that particular patient containing that term. This historic view will show all the verbatim terms in chronological order along with the others for that patient. The user will then be able to see that verbatim adverse event or series of drug use in the context of the patient's history. This context will give more information and help the user understand the meaning of that term. Once the user grasps this understanding, the decision being made still needs to be applied universally. It is still useful to see a complete report of an individual patient to understand the meaning before deciding on a universal coding decision.

Another special situation that arises during manual coding is the splitting of the verbatim terms. There are instances where the verbatim adverse event or drug name is actually multiple adverse events or multiple drug names. In this case, the terms needs to be split into separate instances and coded separately. Commonly, this term is sent back to the operational database which originally captured the CRF data. However, this process can be time consuming and prohibits the ability to perform coding efficiently. It is therefore useful to have the ability to split the verbatim terms and then code each individual term within the manual coding step. The original term is kept for audit historical records. However, the manual coding decision will then be made upon the multiple new split terms.

During manual coding, you can search through the dictionaries to find the matching preferred terms. The search can be applied in a similar way to how the Google search engine is applied. This is reliant upon key words entered by the user used with a matching algorithm to similar words in the English language. Using this searching methodology is therefore the easiest and most powerful way of getting the information needed to apply your coding decision.

In this example search, the key word “abnorm” is used as search criteria. This key word can appear anywhere inside the term. The word does not have to be exactly the same but have a close match. This elevates misspelling while still capturing accurate results. The goal is to apply a search that best matches the meaning of the user’s intent.



Thesaurus Search Results

Thesaurus Name: *Global Meddra*
Search Text Criteria: *abnorm*

1 item(s) found.

Verbatim	Lowest Level Term	Preferred Term	High Level Term	High Level Group Term	System Organ Class
UNSTEADY GAIT	Abnormal gait	Gait abnormal	Gait disturbances	General system disorders NEC	General disorders and administration site conditions

Generated on: 02/15/2007, 10:36:29 pm

STEP 4: **MAPEX**

All previous steps have been applied in preparation for the final step of creating a data set that contains all mapped terms. This is the step that takes all the decisions made during manual coding and then merges all the dictionaries to derive a final data set that contains the preferred terms along with other supporting information. This step is referred to as “mapex” because it will create a new data set with the same name as the old one, but with an additional letter “x” added to it. For example, if the data was named AE, the new created mapped data would be AEX. The tasks involved in this step include:

1. **Merging –** This will merge the source data against all the dictionaries defined in the session. It is similar to the manual coding search in that it will merge the source data against the dictionaries to acquire the matching preferred terms. Once all the merges are complete, it will create the new data set with the preferred terms along with the verbatim terms in one data set.
2. **Supporting Variables –** In the MedDRA case, it will include the hierarchy name and coded values such as SOC, HLT and LLT. On the other hand, for WHO Drug, it will contain all variables from the dictionary such as INN and USAN. If the term has been split, it will contain the original term and the term that has been split.
3. **New Data –** It will then create a new data set with a new name which is the same name as the old data set with the letter “x” added to the name. To help distinguish MedDRA coding, it will include “mx” to the data set.

4. **Optional Tasks –** The new data needs to be delivered to the intended user. The new data set is created in the same location as the source data set. Optionally, it can also be compressed into a file using ZIP and emailed to a specified recipient. An additional password protection can also be added for security upon this compressed file. This can also be scheduled so the mapex process can be performed on a periodic basis upon data updates.

STEP 5: **REPORTING**

The previous steps 1 through 4 are usually performed in that specified order. This last step can be performed at any moment. The user needs to perform a review of the mapping process through reports. Some of the common reporting needs include:

- **Tentative Mapped Terms –** If the user has made a decision during manual coding but is not sure of that decision, it is placed into a holding area. Another user can then review this in this report and confirm if the decision is correct or not. Once confirmed, it can be placed into the final destination dictionary.
- **Unmapped Terms –** During the process, there will be terms that are completed and mapped while there may be those that do not yet have a match. These unmapped terms are displayed in this report. It represents a list of terms that the user must manually code. An optional report can also be included to display the patient ID along with the unmapped terms.
- **Mapping History -** All decisions made including tentative or final are recorded in an audit history. This can be reviewed in the mapping history report.

There are other recommended steps used to optimize the methodologies to process the coding of terms from source data to the standardized synonyms stored in thesaurus dictionaries. The five steps described above capture the main tasks required to efficiently derive at the data set which contains mapped terms.

CONCLUSION

This paper presents the clinical significance of coding decisions and how it affects your analysis. This has an impact on your submission resulting in the failure or success of a FDA filing. A historical demonstration of the structure of the dictionary is described to provide context on how the information is organized. The external source dictionary data is delivered as ASCII raw files which are then loaded into a structured database for optimal management. Each dictionary is managed to keep track of version control along with its hierarchical levels. The user would then apply an auto code to find matches and mismatched terms. For those terms that did not match, a manual coding process is performed with the use of a powerful yet easy to use search engine. These decisions will then be populated into a knowledge base so future coding decisions can be applied with these same decisions. Upon completion of all these decisions, a map between the source data and the dictionaries will result in a final data set that contains consistent preferred terms used for meaningful analysis. During the entire process, reports can be generated to give you status information such as a list of terms left to be coded. The methodologies presented in this paper demonstrate the complexity of the process. However, the use of logical steps along with the powerful searching approaches can simplify this into steps that are as simple as performing a Google search.

CONTACT INFORMATION

Sy Truong is President of MXI (Meta-Xceed, Inc.) They may be contacted at:

Sy Truong
MXI, Meta-Xceed, Inc.
1751 McCarthy Blvd.
Milpitas, CA 95035
(408) 955-9333
sy.truong@meta-x.com

If you have any comments and questions regarding this paper, please contact:

Na Li
Pharmacyclics, Inc.
995 East Arques Avenue
Sunnyvale, CA 94089
Office: (408) 990-7293
E-mail: nli@pcyc.com