Paper S6-2007

Using the WHO Drug Dictionary for Reporting Clinical Trials

MWSUG 2007 Meeting

Thomas E Peppard, deCODE Genetics, Brighton, MI

ABSTRACT

This paper will introduce the application of the WHO Drug Dictionary to the analysis of clinical trial data. It will describe the structure of the dictionary, including PROC SQL example code to describe the relationships among the dictionary tables. Next the paper will provide example data summaries using different components of the tables. Finally, the author will discuss SAS® coding strategies for implementation, including how to manage dictionary updates.

INTRODUCTION

The WHO Drug Dictionary (WHO-DD) is administered and licensed by the World Health Organization's (WHO) Uppsala Monitoring Center (UMC). The UMC collaborates globally with regulators, researchers and other professionals from the health care and pharmaceutical industries in the practice of pharmacovigilence, which WHO defines as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems." [1]

A drug dictionary proves useful when tabulating medication usage because it classifies the same medication, often known by different names, into a single name. For example, Tylenol®, acetaminophen and paracetamol all refer to the same active ingredient, and WHO-DD uses the ingredient name paracetamol.

This paper will describe the structure of the dictionary, including PROC SQL example code to illustrate relationships among the dictionary tables. Next the paper will provide example data summaries using different components of the tables. Finally, I will discuss SAS® coding strategies for implementation.

In 2005 the UMC released the WHO-DD Enhanced, which follows the same structure as WHO-DD, but incorporates a more timely system for including newly launched pharmaceutical products. While this paper refers to the WHO-DD, the same lessons would apply to WHO-DD Enhanced. Examples in this paper follow "Format B" of the dictionary.

DICTIONARY STRUCTURE

The WHO-DD includes tables that describe the manufacturer of the pharmaceutical products and a published source (e.g., *Physician's Desk Reference*). These tables are omitted from the discussion below.

DD TABLE

The DD table contains the drug names that are used for coding source data records (e.g., case report form entries). Drug names can be generic or trade names, and many do refer to drug products that contain multiple active ingredients (e.g., Excedrin® contains aspirin and caffeine). There are different drug names for different salts or esters of the same active ingredient (e.g., morphine sulfate vs. morphine tartrate), but often one would want to collapse these together in order to tabulate on the active ingredient (morphine). For this purpose the WHO-DD provides what this paper refers to as the "preferred" drug name, which is typically a generic name omitting the salt/ester specification. Each drug name is identified by a unique combination of the Drug Record Number, Sequence Number 1 (Seq1) and Sequence Number 2 (Seq2), with the "preferred" drug name identified by Seq1=01 and Seq2=001. Different salt/ester formulations of the same drug are identified by different Seq1 values, and different names for the drug – whether trade names or generic names – are identified with different Seq2 values.

INGREDIENTS TABLES

Drug products are composed of one or more active ingredients, and the ingredients included in each drug product are listed in the ING table, by linking Drug Record Numbers with Chemical Abstract Service Registry Numbers (CAS Numbers). The names of the ingredients are listed in the BNA table, linked by the CAS Number. Non-"preferred" drug names often are not included in the ING table; therefore it is recommended that the DD and ING tables be joined on the Drug Record Number alone, after subsetting on Seq1=01 and Seq2=001.

ATC TABLES

With thousands of drug products on the market, there is an obvious need to group these into meaningful categories. The Anatomic Therapeutic Chemical (ATC) classification system does this, and it is part of WHO-DD. The ATC system originated in the early 1970s in Norway, and a search engine for it is available today in the public domain at the WHO website (http://www.whocc.no/atcddd). However, the linkage between Drug Record Numbers and ATC codes is only available in the WHO-DD.

The ATC system is hierarchical and includes four levels of granularity. An example of this is listed below, with the four levels shown from most general (level 1) to most specific (level 4):

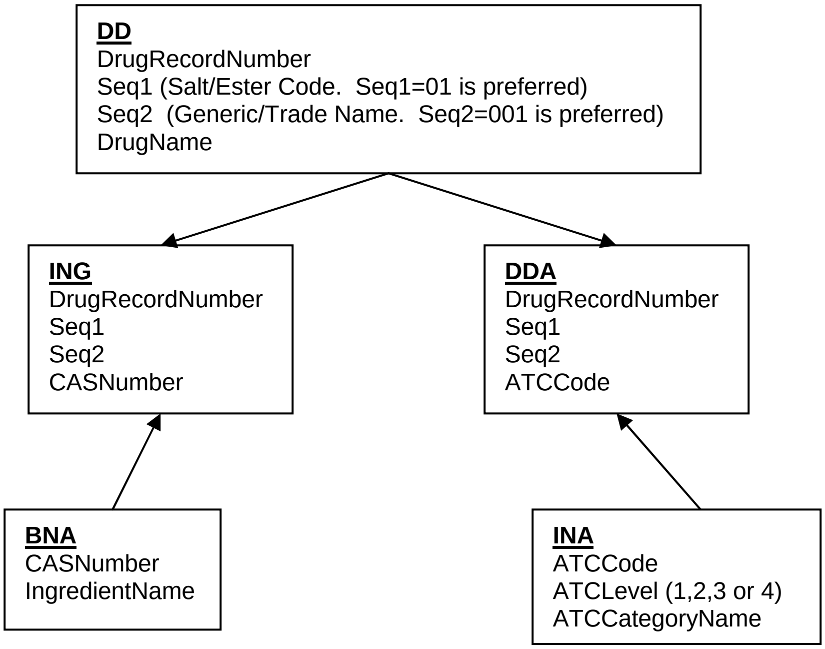
TABLE 1 – ATC LEVELS AND EXAMPLE

		ATC Example	
Level	**Type**	**Code**	**Category Name**
1	Anatomical main group	A	ALIMENTARY TRACT AND METABOLISM
2	Therapeutic subgroup	A02	DRUGS FOR ACID RELATED DISORDERS
3	Pharmacological subgroup	A02A	ANTACIDS
4	Chemical subgroup	A02AA	MAGNESIUM COMPOUNDS

The level 1 codes are always a single letter; the level 2 codes are always a two-digit number appended to the corresponding level 1 code; etc.

Each drug product in the DD table is associated with one or more ATC codes. (Some drugs operate on multiple anatomic systems, and thus are associated with multiple ATC codes). The ATC code(s) associated with each drug product are listed in the DDA table, by the highest ATC level for each association. (For example, a drug that is associated with chemical subgroup A02AA would be listed at only the A02AA level, not A02A, etc.). The names of the ATC categories are listed in the INA table by the ATC code. As with the association between drug names and ingredients, non-“preferred” drug names often are not included in the INA table; therefore it is recommended that the DD and INA tables be joined on the Drug Record Number alone, after subsetting on Seq1=01 and Seq2=001.

FIGURE 1 – FLOWCHART DESCRIBING STRUCTURE OF SELECTED WHO-DD TABLES



SAS® CODING EXAMPLES

The SAS® SQL Procedure is ideally suited for joining the WHO-DD tables into a format that is useful for statistical analysis. In the examples that follow, assume that SAS® data sets are available in a SAS® library named "MEDDICT", and that these data sets follow the WHO-DD table format shown in Figure 1.

CREATING A DATASET TO JOIN DRUG NAMES WITH INGREDIENTS

First consider the case of joining ingredients with "preferred" drug names. This is a one-to-many relationship (individual drug names are, in some cases, associated with multiple ingredients), but there are also drug names which are not associated with any ingredients. Examples of this include the 900000-series Drug Record Numbers, which are non-specific drug names such as "VITAMINS", "MINERALS" or "LAXATIVES".

The example code below creates a dataset named DRUG_ING that contains each drug name and its associated ingredients (if any), along with the Drug Record Number and CAS Number.

```
proc sql;
  create table meddict.drug_ing as select distinct
    a.drecno
   ,a.drugname
   ,b.cas_num
   ,(select c.ingrdnt
      from meddict.bna as c
      where b.cas_num=c.cas_num) as ingrdnt
   from
    meddict.dd  as a left join
    meddict.ing as b
   on
    a.drecno=b.drecno
   where
    a.seq1=01 and a.seq2=001 and
    b.seq1<=01 and b.seq2<=001
   order by
    drecno, cas_num
  ;
quit;
```

Key points in the "Ingredients" code shown above:

- Use left joins so that Drug Record Numbers without ingredients are included.
- Subset on Seq1=01 and Seq2=001 to select only the "preferred" drug names; but
- Use Seq1<=01 and Seq2<=001 on the ING dataset to allow for drug names that have no associated ingredients. (Otherwise these are lost if equality is used).
- Select distinct to drop any duplicates.

CREATING A DATASET TO JOIN DRUG NAMES WITH ATC CODES

The two ATC tables are related to the DD table in a manner that is very similar to the relationship between the DD table and the two ingredients tables. The DDA table associates drug names with ATC codes according to a many-to-one relationship, and each ATC Code is associated with its ATC Category Name in the INA table.

There are four levels of ATC codes, but many drug names are associated with only the first, second or third level ATC code. As a solution to those cases, the example code shown below assigns "Not Specified" when a drug name is not associated with an ATC code at a particular level.

The example code below creates a dataset named DRUG_ATC that contains each drug name and its associated ATC codes. Only the first two levels of ATC codes are included.

First, PROC SQL subqueries are avoided by creating a format that maps from the ATC Code to the ATC Category Name (ATC_TEXT). In the format, *other* (i.e., any value not included in the left-hand side of the format) is mapped to "Not Specified".

```
proc sort data=meddict.ina nodupkey out=fmt(keep=atc_code atc_text);
```

```
  by atc_code atc_text;
run;
data fmt;
  set fmt(rename=(atc_code=start atc_text=label)) end=eof;
  fmtname='$ATC';
  type='C';
  output;
  * Add OTHER --> Not Specified;
  if eof then do;
    hlo='O';
    start=' ';
    label='Not Specified';
    output;
  end;
run;
proc format cntlin=fmt;
run;
```

Next, join Drug Names with their ATC codes, and use the $ATC format to populate the ATC Category Names. Drug names that are missing ATC codes are assigned an ATC code of "_", which will sort last, and an ATC Category Name of "Not Specified".

```
proc sql;
  create table meddict.drug_atc as select distinct
    a.drecno
   ,a.drugname
   ,case
     when(compress(b.atc_code)^="") then substr(b.atc_code,1,1)
     else '_'
    end                             as ATC_CD1
   ,put(calculated atc_cd1,$atc.)   as ATC1
   ,case
     when(length(b.atc_code)>=3) then substr(b.atc_code,1,3)
     else '_'
    end                             as ATC_CD2
   ,put(calculated atc_cd2,$atc.)   as ATC2
   from
    meddict.dd  as a left join
    meddict.dda as b
   on
    a.drecno=b.drecno
   where
    a.seq1=01 and a.seq2=001 and
    b.seq1<=01 and b.seq2<=001
   order by
     drecno, atc_cd1, atc_cd2
  ;
quit;
```

SUGGESTIONS FOR SUMMARIZING MEDICATION USAGE

ATC codes are an excellent tool for grouping drug names in a summary, but levels 3 and 4 tend to be so granular that there are very few medications in each category. Thus in the example below only levels 1 and 2 are used.

TABLE 2 – EXAMPLE SUMMARIZING BY ATC LEVELS 1 AND 2

Anatomic Group	**Treatment A**
Therapeutic Subgoup	**(N=xxx)**
Preferred Drug Name	**n pct**
ALIMENTARY TRACT AND METABOLISM	
STOMATOLOGICAL PREPARATIONS	
ACETYLSALICYLIC ACID	xx xx%
DOXYCYCLINE	xx xx%
IRON	xx xx%
METRONIDAZOLE	xx xx%
POTASSIUM	xx xx%
DRUGS FOR ACID RELATED DISORDERS	
ALMINOX	xx xx%
ESOMEPRAZOLE	xx xx%
OMEPRAZOLE	xx xx%
RABEPRAZOLE	xx xx%
RANITIDINE	xx xx%
ANTIDIARR., INTEST. ANTIINFL./ANTIINFECT. AGENTS	
BETAMETHASONE	xx xx%
BUDESONIDE	xx xx%
IMODIUM	xx xx%
MESALAZINE	xx xx%

At times it is desirable to tabulate by active ingredients, which are always generic names, rather than by "preferred" drug name, which is sometimes a trade name. (Often this is the case for a combination agent). When this is desired, one can sort the ingredients' names in alphabetic order. Unfortunately, there is no direct relationship between ATC codes and active ingredients. Because, in some cases, the same active ingredient is used in very different ways in two different drugs, an attempt to relate ingredients back to ATC codes through the drug names could lead to some strange results. Consider the examples shown below.

TABLE 3 – EXAMPLE OF THE SAME ACTIVE INGREDIENT USED IN DISPARATE DRUG PRODUCTS

Active Ingredient	**Preferred Drug Name**	**ATC Codes (Level 1 → Level 2)**	**What is it?**
Paracetamol	Tylenol®	Nervous System → Analgesics	
Paracetamol	Sudafed®	Respiratory System → Cold and Flu Preparations	
Timolol	Betacentyl®	Cardiovascular → Beta Blocking Agents	Oral antihypertensive
Timolol	Xalcom®	Sensory Organs → Ophtamological	Eye drops for glaucoma

IMPLEMENTATION STRATEGIES

WHO-DD is now available as two different data formats: B Format and C Format. The examples shown above are from the B Format. The C Format includes a more specific definition related to the drug name, the "medicinal product ID". The C Format uses this additional level of detail to classify the drugs into ATC codes a bit differently than the B Format.[2]

When analyzing concomitant medication usage data from clinical trials, it can be desirable to store the data in three different data structures:

- As captured from the source documents, with one row in the dataset for each entry in the source documents.

- Joined with ATC codes. Since there is often more than one ATC code corresponding to each drug name, this join increases the number of rows in the dataset.
- Joined with generic ingredients. Since many drugs include multiple active ingredients, this join also increases the number of rows in the dataset.

The redundant data in the three data structures described above suggests that constructing these as SAS® PROC SQL views is a more efficient alternative to permanent data sets.

CONCLUSION

The WHO-DD dictionary is used to code medications, classify these into ATC (anatomic, therapeutic and chemical) categories, and identify the active ingredients associated with each medication.

For the purposes of analyzing clinical trial data, it is useful to use only the "preferred" drug names (those with Seq1=01 and Seq1=001).

One can group drug names by ATC code, but drug ingredients cannot be grouped in this manner because a single ingredient can be used in disparate types of drugs.

Given the conflicting desires to: preserve the number of medication records; expand the number of records corresponding to the ATC code(s) associated with each drug name; and expand the number of records corresponding to the active ingredient(s) associated with each drug name; it makes sense to create SAS® views to provide these different data structures.

REFERENCES

1. World Health Organization Uppsala Monitoring Committee. "Welcome Page". http://www.who-umc.org/DynPage.aspx
2. World Health Organization Uppsala Monitoring Committee. WHO Drug Dictionary Sample Getting Started document. http://www.umc-products.com/DynPage.aspx?id=2844. Download zip file from "WHO Dictionary Samples" link.

ACKNOWLEDGMENTS

I wish to acknowledge Kjartan Guðmundsson, formerly of deCODE Genetics and Sigurður D. Sigfússon of Encode for their assistance in loading and configuring my employer's license of the WHO-DD dictionary.

CONTACT INFORMATION

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

Thomas E. Peppard
deCODE Genetics
1032 Karl Greimel Dr, Suite 11
Brighton, MI 48116
Work Phone: 810-522-1909
E-mail: tpeppard@decode.com