Automating the Process of Creating the Contents of Data Definition Document

David Wang, Sanofi-aventis, Malvern, PA
Gregory Ridge, Sanofi-aventis, Malvern, PA

ABSTRACT

Creating a data definition document is an important step for any newly created dataset. It can either be done before the development of any datasets or after the actual creation of a dataset, but in any case, it must be done before the validation of those databases. The 1999/2003 FDA electronic submission (eSub) guidance and the electronic Common Technical Document (eCTD) documents clearly specify that a document describing the content and structure of the included data should be provided within a submission. Normally three important Data Definition Documents are needed for a submission to FDA. The first one is for ODM (formally called CRT DD and commonly referred as "define.xml") to provide Case Report Tabulations Data Definitions in XML format. The second one is for SDTM to provide Study Data Tabulation Data Definitions in XML format, which has been mandated since year 2003. The third one is for ADS/ADaM to provide Analysis Data Definitions in PDF format. The process of creating those files is very tedious and time consuming. The creation process could be divided into two steps: the content creation (e.g. variables, variable's attributes and comments) and the format conversion (i.e. converted to XML or PDF). This paper will mainly focus on the automation of creating the contents of data definition document.

INTRODUCTION

Creating a data definition document is an important step for any newly created dataset. Given the fact they are the specifications for the variables defined in the database, some companies or teams even create them before the dataset is created. This document not only defines how the variable is created but also has other important information like origin, control terminology etc. The process of creating the document is very tedious and time consuming, and could be divided into two steps: 1) the content creation (e.g. variables, variable's attributes and comments) and 2) the format conversion (i.e. converted to XML or PDF with all needed hyperlinks). This paper will cover the first step only, which is to create and update the contents of data definition file in excel by two main macros. This paper will also show how the creation process helps facilitate the document review process during the study. The macro can produce both a short version and long version of the data definition document. The short version, which contains only key columns, is to facilitate the team review process. The long version could be used to load into an internal application that converts the excel file to the desired format (XML or PDF) with all necessary hyperlinks.

PROBLEM ANALYSIS

Define.pdf and Define.xml: Let's look at one of the final data definition documents for ADS in PDF and another one for SDTM in XML first. Figure 1 shows the format of a definition file for analysis data ADDM in PDF. Figure 2 shows the format of a definition file for SDTM (DM) in XML. In both of these documents some fields are easier to create than others and in many cases it is a very time consuming process which can be critical for submission milestones. The fields like Comments/Notes, Origin and Controlled Terms or Format have to be filled out manually in the past. Since version 3.1.2 of SDTM the field "Controlled Terms or Format" has become mandatory.

Due to the fact that CRF design does not always use the CDISC recommended controlled terms a mapping effort is typically needed. At times, even the Origin field is hard to determine. This paper will show you how we can use the existing source file(s) to make the creation process easier. At the same time, the approach introduced here can be utilized as a comparison against any internal standards for validation purpose.

Study: [redacted] Demographics addm.xpt				
Variable	**Label**	**Type**	**Codes**	**Comments**
STUDYID	Study Identifier	text		Set equal to "[redacted]".
USUBJID	Unique Subject Identifier	text		Subject identification unique across all trials and defined by protocol number (without character prefix) ($6) country ($3) center ($3) and subject ($3) with leading zeroes for each component where needed.
SUBJID	Subject Identifier for the Study	text		Subject identifier used within the study and defined as country ($3) center ($3) and subject ($3). Page 2

Figure 1 A definition file for ADS in PDF

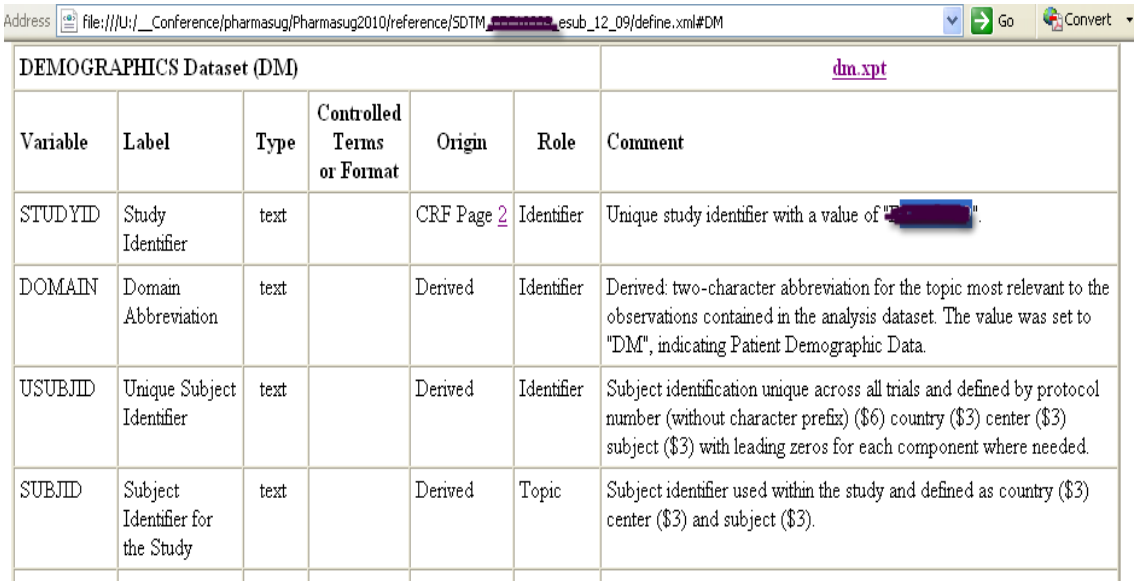


Figure 2 A definition file for STDM in XML

Output files: In order to produce the above two types of tables in the data definition document we need to have all above column information available. In our case we will create a CSV file, which contains at least the following variables: Variable, Label, Type, Code/Format, Controlled Terms, Origin, Role and Comments. Among them Variable, Label, Type, Code/Format are from the SAS contents of the datasets. The rest of variables would have to have been created manually in the past or will come from the sources other than SAS dataset in our case. The macro we introduced here could create two versions of output files with the same key columns. The long/detailed one is used to load into our internal application only. The short/less detailed version is used for the team members to review the key columns easily (see Figure 7).

The challenge is how to utilize the existing documents (standard or other similar study data definition document) to create the data definition document for your study electronically. This is especially useful for the case where the datasets were created without formal dataset specification documents. For the scenarios where dataset specifications already exist, we could use the approach to enhance or improve them during the analysis of the study. The process will be discussed in detail in the following sections.

PREREQUISITES

A well-defined Master Data Definition File (MDDF) is needed. MDDF could be a company standard ADS and/or SDTM or ones downloaded from CDISC website (http://www.cdisc.org/content1039). For our case we use a company standard for ADS as MDDF.

PROCESS AND INPUTS

Figure 3 shows the process of creating the data definition files electronically. There are two main SAS macros used in the process: **%CONTENT** and **%DEFINE.**

%CONTENT is used to create the Study Specific Data Contents File (SSDCF) based on the existing datasets.

%DEFINE is the macro utilized to join between different input datasets. In the very beginning when there are no existing data definition files, the input files for this macro would be SSDCF and MDDF. Once the first Draft Data Definition File (DDDF) is created a user can have an option to create either the next version of DDDF based on its predecessor or the MDDF until the final file is produced. The final data definition could be a long or short version or both as needed.

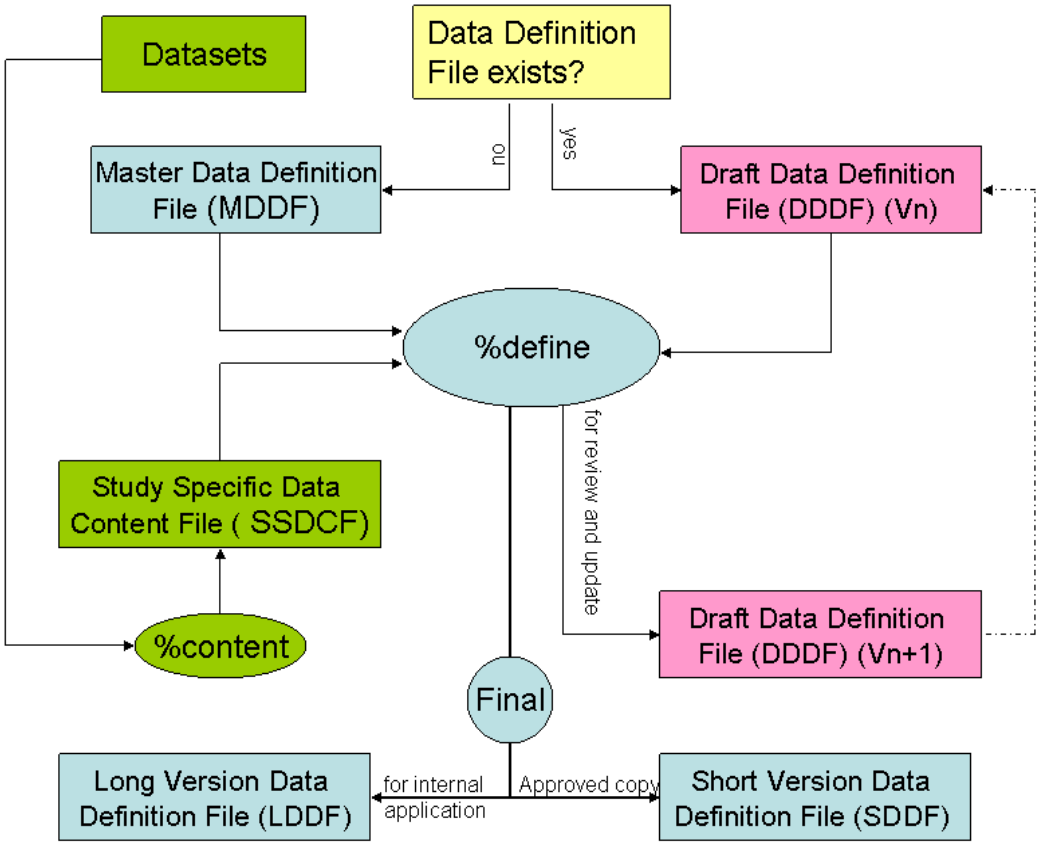


Figure 3 Process of creation of definition file content

MASTER DATA DEFINITION FILE (MDDF)

The Master Data Definition File (MDDF) could be a company standard ADS or SDTM or ones downloaded from CDISC website. Figure 4 shows a sample of an ADS standard definition file used by our company. We utilize it as an MDDF in our application.

A	B	C	F	G	H	I	J	K	L	M	N
ADS name	Order of the variable	Variable name	Source variable name	Format to be applied to SRC_VAR to obtain VARIABLE	Source type	Variable label	Variable type	Variable length	Controlled terminology or format	Origin	Role
ADS	POSITION	VARIABLE	SRC_VAR	SRC_FMT	SRC_TYPE	LABEL	TYPE	TGT_LEN	CNTLTER	ORIGIN	ROLE
$8	8	$8	$15	$50	$10	$40	$10	8	$50	$50	$200
ADDM	38	DOMAIN				Domain	Text	2	DM	Assigned	Identifier
ADDM	39	RFSTDTC				Subject Reference Start Date/Time	Datetime	19	ISO 8601	Derived	Record Qualifier
ADDM	40	RFSTDX				Subject	Text	10	YYYY-MM-	Derived	Timing
ADDM	41	RFSTTX				Subject	Text	5	HH:MM.	Derived	Timing

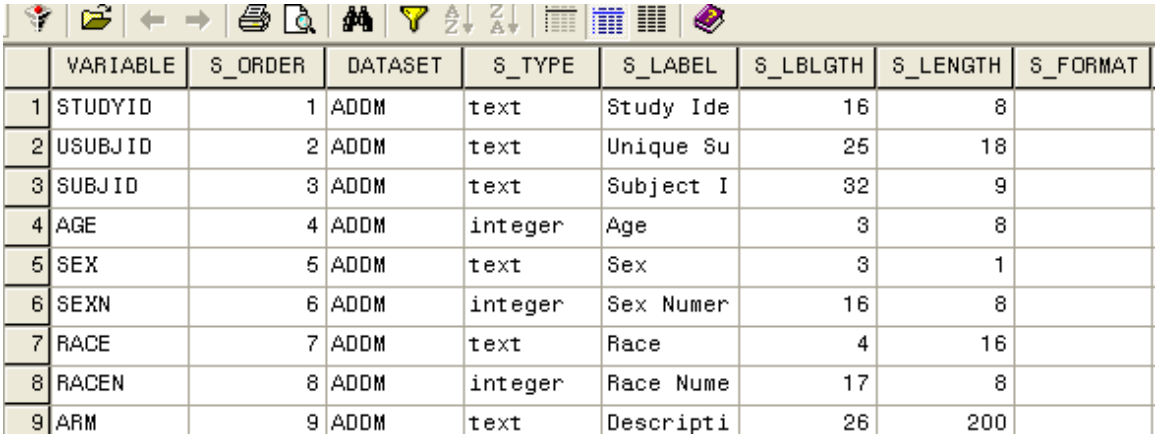
A	O	P	Q	R	S	T
ADS name	Notes	Req/Exp/Perm	SDS or SUPPQUAL	Key variable number	Page number	Note number
ADS	CDISCDSC	CORE	TARGET	KEY	PAGE	NOTE
$8	$255	$4	$8	8	$200	$200
ADDM	Two-character abbreviation for the domain.	Req	SDS			
ADDM	Reference Start Date/time for the subject in ISO 8601 character format. Is the basis for all relative day calculations and analysis windows. Usually equivalent to date/time of first intake of drug but this may be defined as screen date/time,	Exp	SDS DM			
ADDM		Exp				
ADDM		Perm				

Figure 4 Master Data Definition File

The main columns we need from this file are: Variable Name, Controlled Term, Origin, Role, Notes, Req/Exp/Perm, and SDS/SUPPQUAL. Among them **Variable Name** is a join key.

STUDY SPECIFIC DATA CONTENTS FILE (SSDCF)

The Study Specific Data Contents File (SSDCF) is produced by **%CONTENT**. Figure 5 shows the format of this file. As you can see, all columns but S_TYPE can be produced by a Proc Contents directly. The S_TYPE is supposed to have one of five data types (text, integer, time, data or float) in the data definition file instead of Char or Num. Other features of %CONTENT will be discussed in the SAS code section.



	VARIABLE	S_ORDER	DATASET	S_TYPE	S_LABEL	S_LBLGTH	S_LENGTH	S_FORMAT
1	STUDYID	1	ADDM	text	Study Ide	16	8	
2	USUBJID	2	ADDM	text	Unique Su	25	18	
3	SUBJID	3	ADDM	text	Subject I	32	9	
4	AGE	4	ADDM	integer	Age	3	8	
5	SEX	5	ADDM	text	Sex	3	1	
6	SEXN	6	ADDM	integer	Sex Numer	16	8	
7	RACE	7	ADDM	text	Race	4	16	
8	RACEN	8	ADDM	integer	Race Nume	17	8	
9	ARM	9	ADDM	text	Descripti	26	200	

Figure 5 Study specific data contents file

DRAFT DATA DEFINITION FILE

The Draft Data Definition File (DDDF) is considered as both input and output of macro **%define**. It was created by jointing Master Data Definition File or previous DDDF and Study Specific Data Contents File (SSDCF).

Figure 6 shows the intermediate one. The column VARIABLE is the key for any joins or merges between or among files. The columns starting with S_ are from SSDCF created by %CONTENT. The columns starting with M_ are from MDDF. Other columns are copied from the corresponding M_ columns (from MDDF) first and are either kept or modified (in yellow) based on your study specifications. For the non-standard variables (in green) that do not exist in MDDF we have to develop the column Notes, Origin and SDS_SUPPQUAL etc. from scratch. The positive aspect is that we only have to write it once and it can then be utilized many times. Those columns without prefixes are the ones that need to be kept for the final data definition document (see the OUTPUT section).

Once DDDF is updated it should be used for next run to carry forward those updated columns until a final one is created and approved. The %define can be run as much as needed until the datasets and their definition documents are finalized and approved.

A	B	C	D	E	F	G	H	I	J	K
VARIABLE	S_LABEL	S_TYPE	Notes	M_NOTES	Origin	M_ORIGIN	SDS_or_SUPPQUAL	M_SDS_OR_SUPPQUAL	Req_Exp_Perm	Controlled_terminology_or_format
DOMAIN	Domain Abbreviation	text	Two-character abbreviation for the domain most relevant to the observation. Set to "DM"	Two-character abbreviation for the domain most relevant to the observation. Set to "DM"	Derived	Derived	SDS	SDS	Req	DM
BRTHDT	Date of Birth	date	The date of birth in yymmdd10 SAS format.	The date of birth in yymmdd10 SAS format.	Derived	CRF	ADS	ADS	Perm	YYMMDD10.
ARM	Description of Planned Arm	text	Treatment group name with a value "ALVOCIDIB" as it is a single arm trial. For screen failures equal to 'Screen Failure'.	Treatment group name from randomization schedule. This means the description of each unique treatment arm as defined in the protocol. Includes description, dose	Derived	Derived	SDS	SDS	Req	
ARMN	Planned Arm Number	integer	Numeric version of ARM with a value equal to 1. For screen failures set equal to missing.	Numeric version of ARM used for sorting. 0 should be used for placebo and the dose levels for a dose-ranging study. Screen failures should be missing.	Derived	Derived	ADS	ADS	Perm	
PCOHORT	Primary Cohort	text	Indicate if the patients who have received at least 2 cycles of study treatment and have baseline and at least one post-baseline valid response assessment. Set "Y" if the answer is YES otherwise "N"		Derived		SUPPDM		Perm	Y or N

Figure 6 Intermediate draft data definition file

OUTPUT

There are two final output files to produce: 1) a short version of Data Definition File (SDDF) for team to utilize in study development for review and approval, and 2) a long version of Final Data Definition File (LDDF) including all necessary variables for an internal company application. SDDF is designed to provide the users/reviewers with an excel file with less columns (e.g. Comments, Origin, SDS_or_SUPPQUAL Controlled_term). LDDF is utilized to load into an internal department application, which can convert CSV file into submission ready data definition file for SDTM in XML and/or ADS in PDF.

Figure 7 shows the short version of definition file. It is customized from the intermediate data definition file at the end of study. The macro %define has a great flexibility. Users can choose: 1) what columns need to keep in the short version in any run; 2) what columns need to be overwritten by the previous draft/intermediate data definition file.

	A	B	C	D	E	F	G	H	I	J
1	VARIABLE	S_LABEL	S_TYPE	Notes*	Origin*	SDS_or_SUPPQUAL*	Controlled_terminology_or_format*	Req_Exp_Perm*	S_ORDER	STD_FLAG
2	DOMAIN	Domain Abbreviation	text	Two-character abbreviation for the domain most relevant to the observation. Set to "DM"	Derived	SDS	DM	Req	30	Y
3	SITEID	Study Site Identifier	text	Unique identifier for the study site within a submission. Concatenation of country number and center number. Defined as the first six charactors of SUBJID in character format.	Derived	SDS		Req	31	Y
4	INVNAM	Investigator Name	text	Name of the investigator for a site. Concatenation of the first name variable and the last name variable with a blank between the two.	Derived	SDS		Perm	32	Y
5	BRTHDT	Date of Birth	date	The numeric date of birth in yymmdd10 format.	Derived	ADS	YYMMDD10.	Perm	33	Y
				Character representation of BRTHDT in						

Figure 7 Short version of data definition file

The long version of Data Definition File (LDDF) is not shown here.

SAS CODE

Only critical code and tips are shown here.

1. %CONTENT

The SAS code contains three parts: 1) a Proc Contents call to generate the SAS dataset with all attributes; 2) data step manipulation to modify the data type (from two types to five types); and 3) data step and Proc Report to produce a data consistency checking report. Before using the Study Specific Data Contents File (SSDCF) as an input of **%DEFINE** we want to make sure that the variable attributes in the file are compliant with submission requirements (e.g. the length of variable and variable labels etc.) and the same variable is used consistently across the dataset. The code in the macro is very basic and not shown here.

2. %DEFINE

It has the following main parts.

1) Define location of input and output

```
libname std_lib "&path..standard.";          * MDDF location *;
libname stu_lib "&path..data_m.";             * SSDCF location *;
```

```
libname upd_lib "&path..draft.";             * DDDF location *;
%let def_path = &path..esub.;               * DDDF and FDDCF in CSV location *;
```

2) Identify variables to keep or update

```
* Variables for short version *;

%let part_var =%str(variable, s_label, s_type, notes, m_notes, origin, m_origin,
req_exp_perm, sds_or_suppqual, m_sds_or_suppqual, Controlled_terminology_or_format,
role, s_length, s_order, std_flag) ;

* Variables for long version *;

%let full_var =%str( s_order, variable, source_library, source_dataset,
source_variable_name, format_to_be_applied_to_src_var, source_type, s_label,
s_type, s_length, controlled_terminology_or_format, origin, role, notes,
req_exp_perm, sds_or_suppqual, key_variable_number );

* Variables for updates in next run *;

%let updvars=notes origin role req_exp_perm sds_or_suppqual
Controlled_terminology_or_format;
```

3) Join between SSDCF, and MDDF and DDDF

```
%do i=1 %to &nb_pgm.;
%if %sysfunc(exist(stu_lib.&stu_pfix.&&pgm_&i.)) and
%sysfunc(exist(std_lib.&std_pfix.&&pgm_&i.)) %then %do;
        proc sql noprint;
             create table &&pgm_&i. as select
                a.*,
                b.*,
                b.notes as m_notes,
                b.origin as m_origin,
                b.sds_or_suppqual as m_sds_or_suppqual
                from stu_lib.&stu_pfix.&&pgm_&i. a left join
                std_lib.&std_pfix.&&pgm_&i.  b
                on a.variable = b.variable_name
                order by a.s_order;
        quit;
    %end;

    %if %sysfunc(exist(stu_lib.&stu_pfix.&&pgm_&i.)) and not
        %sysfunc(exist(std_lib.&std_pfix.&&pgm_&i.)) %then %do;
        proc sql noprint;
             create table &&pgm_&i. as select
                a.*,
                b.*,
                b.notes as m_notes,
                b.origin as m_origin,
                b.sds_or_suppqual as m_sds_or_suppqual
                from stu_lib.&stu_pfix.&&pgm_&i. a left join
                std_lib.&std_pfix.adall  b
                on a.variable = b.variable_name
                order by a.s_order;

        quit;
    %end;

    data &&pgm_&i.;
         set &&pgm_&i.;
         by s_order;
         length std_flag $1.;
         if _n_>=&com_n.;
         s_codes='';
         if notes='' and origin='' and role='' then std_flag='n';
         else std_flag='y';
```

```
    run;

%if %upcase(&updated.)=y and %sysfunc(exist(upd_lib.&upd_pfix.&&pgm_&i.)) %then
%do;
    proc sort data=&&pgm_&i. out=&&pgm_&i.(drop=&updvars.);
        by variable;
    run;
    proc sort data=upd_lib.&upd_pfix.&&pgm_&i. out=&upd_pfix.&&pgm_&i.
                                                    (keep=variable &updvars. );
        by variable;
    run;

    data &&pgm_&i.;
        merge &&pgm_&i.(in=a) &upd_pfix.&&pgm_&i.;
        by variable;
        if a;
    run;
%end;

** generate the editable define.doc for editing and updating **;
proc sql;
  create table &edt_pfix.&&pgm_&i. as
  select &part_var.
  from &&pgm_&i.
  order by s_order;
quit;

proc export data= &edt_pfix.&&pgm_&i.
            outfile= "&def_path./&edt_pfix.&&pgm_&i...csv"
            dbms=csv replace;
run;

%end;
```

Code to generate the full CSV file for internal application was omitted.

CONCLUSION

Today's regulatory environment mandates the submission of define.xml and define.pdf as part of submission package. The process of creating these files is typically very tedious and time consuming. This paper introduces a way to utilize the existing standard documents to make the process much easier. More than 50% of variables/rows in the data definition table do not need anything changed if the standard documents (e.g. company ADS or SDTM) are well-created and the study is well-designed. About 40% of variables/rows in the data definition table need some modification. Above 10% of the variables are unique to the study and need significant inputs. In the mean time the macro **%DEFINE** could be used to check if all required and expected variables (by SDTM) are included. Most importantly, %DEFINE is very flexible when creating a customized data definition file for reviewers in terms of choosing columns and rows. For example we can choose only a few important columns like **Notes, Origin** and **SDS_SUPPQUAL** and only the rows/variables that need to be modified (in yellow in Figure 6) and write from scratch (in green in Figure 6). Ultimately, automating the process of creating the data definition file in a flexible way will increase both the production speed and quality of work.

CONTACT INFORMATION

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

David Wang
Sanofi-aventis
9 Great Valley PKWY
Malvern, PA 19355
David.wang@ sanofi-aventis.com or david_2_wang@yahoo.com

Gregory Ridge
Sanofi-aventis

9 Great Valley PKWY
Malvern, PA 19355
Gregory.Ridge@ sanofi-aventis.com