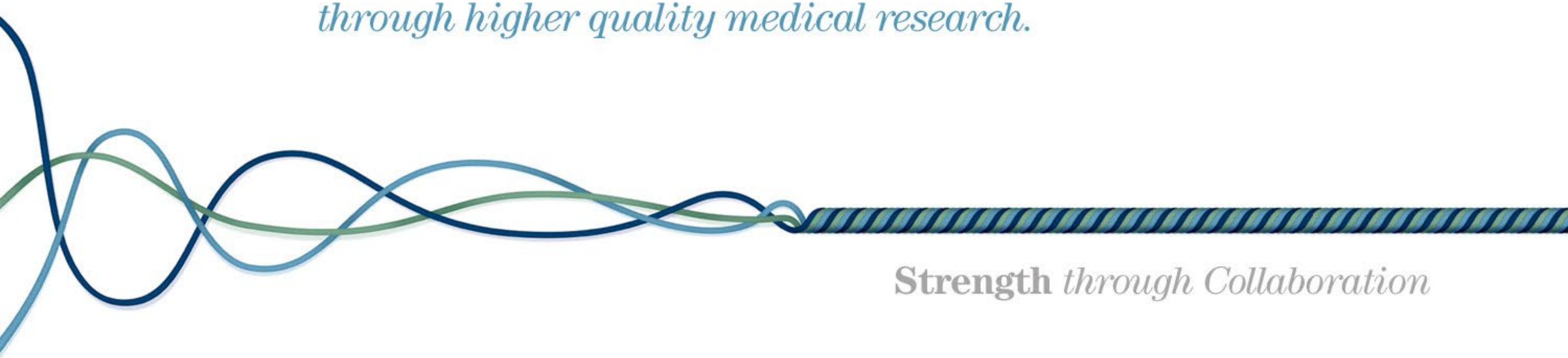
# CDISC®

CLINICAL DATA INTERCHANGE STANDARDS CONSORTIUM

*The CDISC vision is to inform patient care & safety through higher quality medical research.*



**Strength** *through Collaboration*

# An Approach to Combining Disparate Clinical Study Data across Multiple Sponsor’s Studies participating in Project Data Sphere®

Presented by Gene Lightfoot

**Strength** *through Collaboration*

# INTRODUCTION

- Project Data Sphere®
- The Challenge
- The Approach
  - Simplified Process Flow
  - Identify the data
  - Reviewing the Raw Data
  - Programming the Process
  - Reviewing and Data Quality
  - Basic Program Flow
- Documentation
- General Issues and Things to Ponder
- The Final Data Sets
- Conclusion

# PROJECT DATA SPHERE®

- An independent, not-for-profit initiative of the *CEO Roundtable on Cancer's Life Sciences Consortium* (LSC), operates the *Project Data Sphere* platform, a free digital library-laboratory that provides one place where the research community can broadly share, integrate and analyze historical, patient-level, comparator-arm data from academic and industry phase III cancer clinical trials.

- The *Project Data Sphere* platform is available to researchers affiliated with life science companies, hospitals and institutions, as well as independent researchers. Anyone interested in cancer research can apply to become an authorized user.

- A goal of the *Project Data Sphere* initiative is to spark innovation.

# PROJECT DATA SPHERE®

Some Project Data Sphere® metrics (December, 2016)

- 1,437 total users
- 51 countries
- 5,861 total downloads to date
- 40,500+ subjects
- Growing monthly

Tools are available to the registered users and the data can be downloaded and accessed locally.

# PROJECT DATA SPHERE®

| Publication | | Author | Pub. Date |
|---|---|---|---|
| 1 | Comparative Effectiveness of Mitoxantrone plus Prednisone versus Prednisone alone in Metastatic Castrate-resistant Prostate Cancer after Docetaxel Failure. | Angela Green, *et al.* | May 2015<br>***The Oncologist*** |
| 2 | Individual Patient Data Analysis of Randomized Clinical Trials: Impact of Black Race on Castration-resistant Prostate Cancer Outcomes. | Daniel Spratt, *et al.* | April 2016<br>***European Urology*** |
| 3 | A Patient-Level Data Meta-Analysis of Standard-of-Care Treatments from Eight Prostate Cancer Clinical Trials. | N. Geifman<br>A. Butte | May 2016<br>***Nature Scientific Data*** |
| 4 | Predicting Survival of Pancreatic Cancer Patients Treated with Gemcitabine Using Longitudinal Tumour Size Data. | Thierry Wendling, *et al.* | May 2016<br>***Cancer Chemotherapy and Pharmacology*** |
| 5 | "Threshold-crossing": A Useful Way to Establish the Counterfactual in Clinical Trials? | H-G Eichler, *et al.* | Oct. 2016<br>***Clinical Pharmacology & Therapeutics*** |
| 6 | Prediction of Overall Survival for Patients with Metastatic Castration-Resistant Prostate Cancer: Development of a Prognostic Model Through a Crowdsourced Challenge with Open Clinical Trial Data. | James Costello, *et al.* | Nov. 2016<br>***Lancet Oncology*** |
| 7 | Estimation of Tumour Regression and Growth Rates During Treatment in Patients with Advanced Prostate Cancer: A Retrospective Analysis. | Tito Fojo, *et al.* | Dec. 2016<br>***Lancet Oncology*** |
| 8 | Assessment of a Prognostic Model, PSA Metrics and Toxicities in Metastatic Castrate Resistant Prostate Cancer using Data from Project Data Sphere. | Anthony Joshua, *et al.* | Feb. 2017<br>***PLOS One*** |
| 9 | A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-resistant Prostate Cancer. | James Costello, *et al.* | *Under Review*<br>***JCO*** |

# THE CHALLENGE

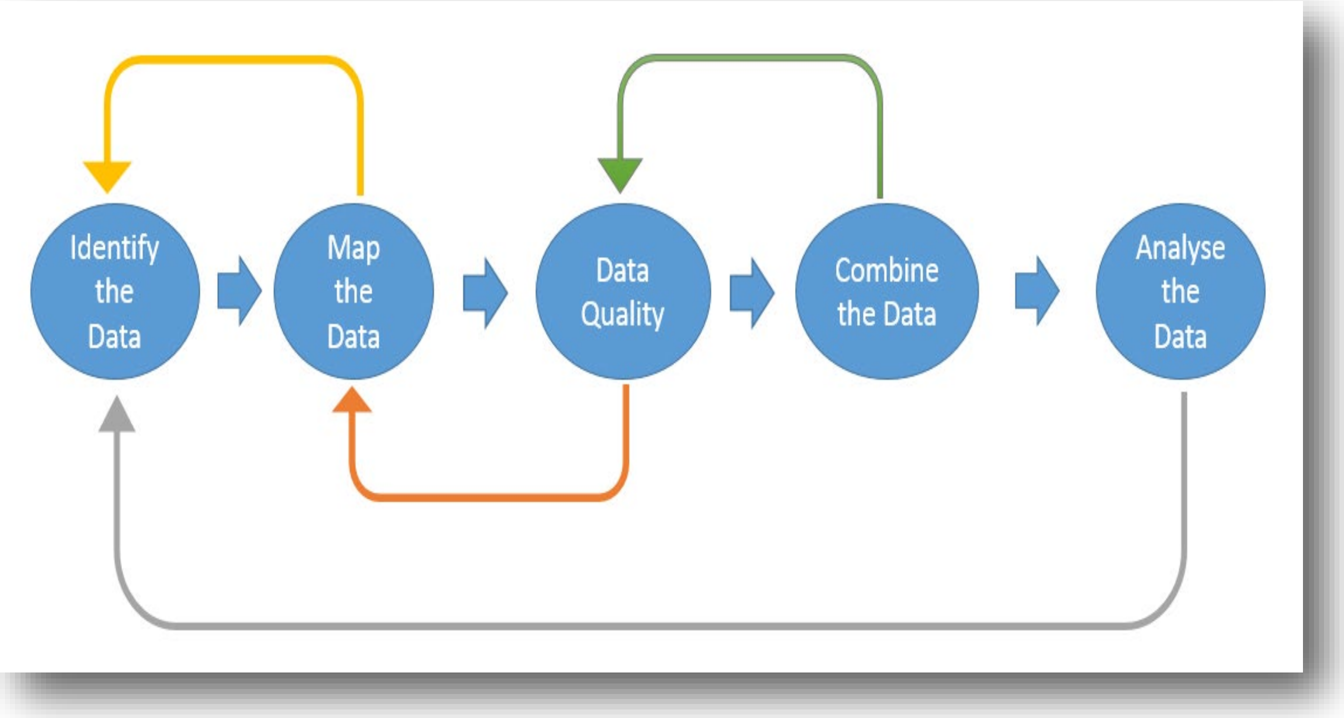
- Use available data provided for the prostate cancer studies to develop and implement a process to combine the data.

- The data comprised 12 separate studies spanning 20+ years from 7 different sponsors.  Standards represented were:
  - 1 ADaM
  - 5 SDTM
  - 6 Other

- Three data sets for analysis were identified; labs, adverse events, and demography.

- The task involved aggregating the data for each domain at the study level and then harmonizing the data for analysis across all 12 of the sponsor studies.

# THE APPROACH SIMPLIFIED PROCESS FLOW



After completing several studies across multiple sponsors, it became evident that a process had evolved that served well for this project.

# THE APPROACH: IDENTIFY THE DATA

Before the team started looking at the data, certain endpoints and populations were identified for the analysis. Of particular interest was the value for the Prostate Specific Antigen (PSA) used as a predictor for Prostate Cancer. This project was a single gender (male) population. It was decided to include all available labs, adverse events (AE), and demography data.

- Since SDTM is considered a global industry standard and recently conducted studies uploaded to Project Data Sphere® usually conformed to this model, it was decided to use SDTM as the standard.

- Disease expertise at this level would have made column selection and analysis much easier. Did not have access to this resource.

# THE APPROACH: IDENTIFY THE DATA

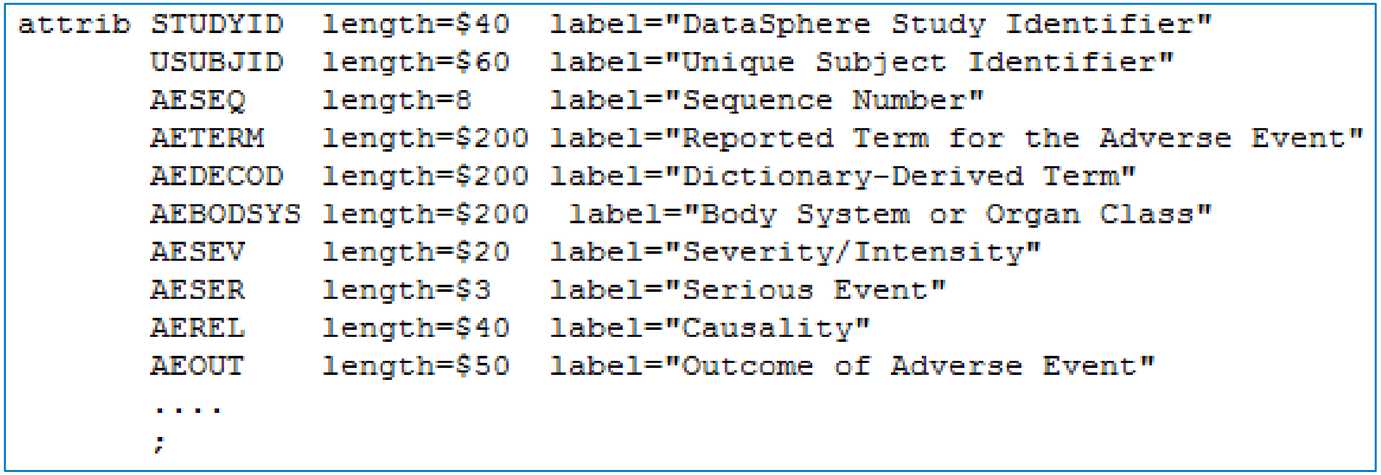
## Reviewing the Raw Data

- Undoubtedly the hardest aspect of this project.
- Supplied as SAS data sets
- Clinical data knowledge is invaluable here – not always obvious where the data is “hiding”. May require multiple data sets to build one domain.
- Data has been de-identified.
- Some of this data was 20+years old.
  - presenting some interesting aspects of data collection – long skinny (normalized) vs short fat (non-normalized) data sets.
  - Unusual data set names – made identifying contents less intuitive .
- All sponsors provided some combination of data dictionary documents, annotated CRFs,  a study protocol document, and SAS formats.

# THE APPROACH: PROGRAMMING THE PROCESS – (MAP THE DATA)

Programming approach

- Although data mapping solutions are available, it was decided to stick with traditional SAS programs to mimic how a solitary researcher might work.
- A global attribute program for each domain was created to manage the column metadata as the project progressed – column name, label, type, length, etc. This metadata was %included in each domain program.



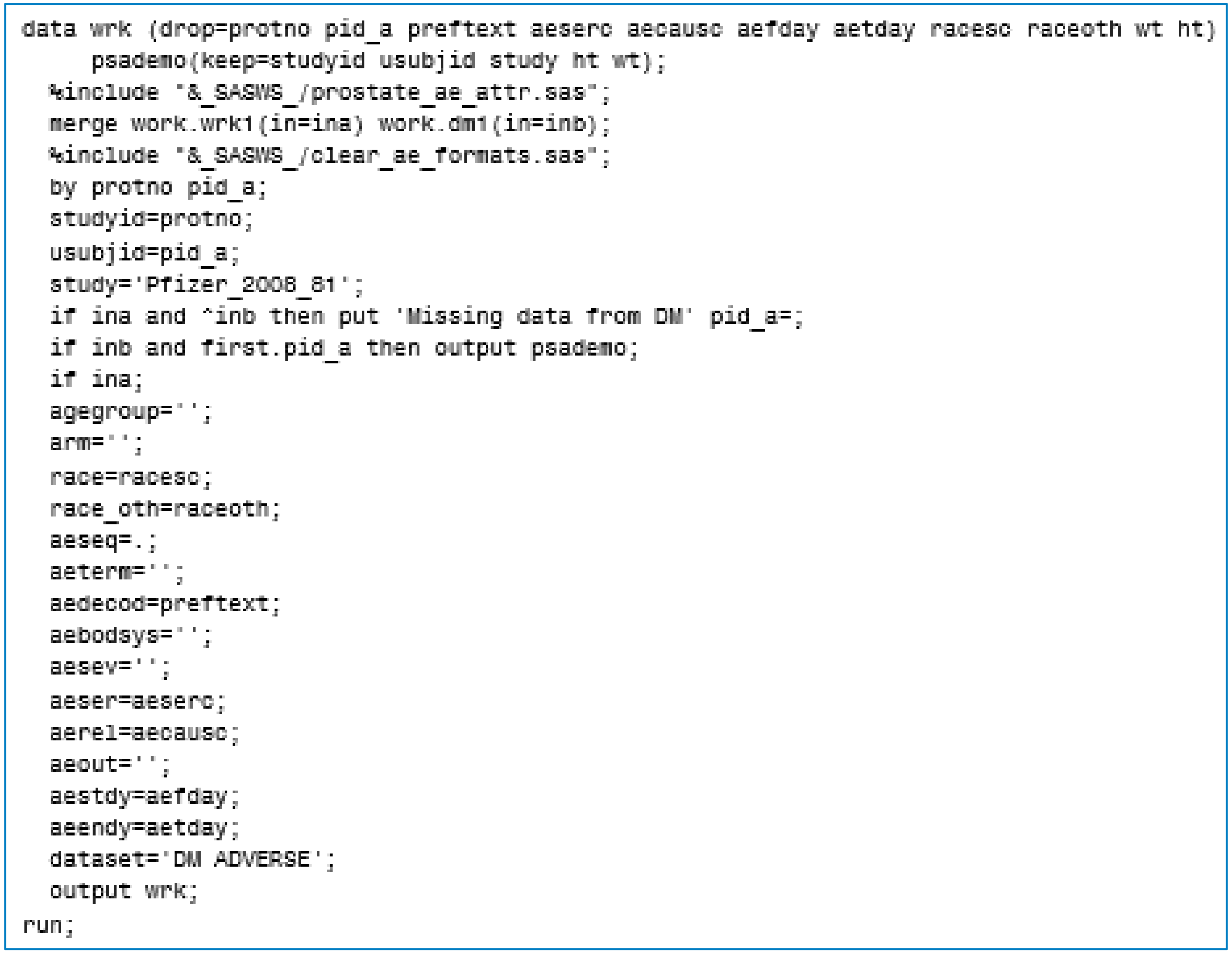
```
attrib STUDYID  length=$40  label="DataSphere Study Identifier"
       USUBJID  length=$60  label="Unique Subject Identifier"
       AESEQ    length=8    label="Sequence Number"
       AETERM   length=$200 label="Reported Term for the Adverse Event"
       AEDECOD  length=$200 label="Dictionary-Derived Term"
       AEBODSYS length=$200  label="Body System or Organ Class"
       AESEV    length=$20  label="Severity/Intensity"
       AESER    length=$3   label="Serious Event"
       AEREL    length=$40  label="Causality"
       AEOUT    length=$50  label="Outcome of Adverse Event"
       ....
       ;
```

# THE APPROACH: PROGRAMMING THE PROCESS – (MAP THE DATA)

## Map the Data

Mapping programs were written for each domain (DM, AE, etc.) within each study for each sponsor.

Don’t be alarmed - code reuse within sponsor and even within SDTM standards across sponsor resulted in program efficiencies.

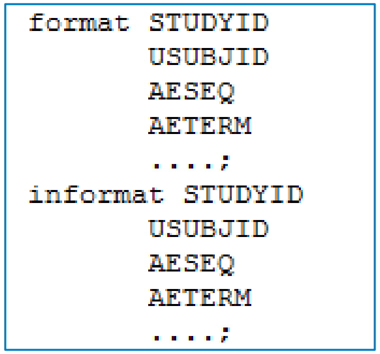


```
data wrk (drop=protno pid_a preftext aeserc aecausc aefday aetday racesc raceoth wt ht)
     psademo(keep=studyid usubjid study ht wt);
  %include "&_SASWS_/prostate_ae_attr.sas";
  merge work.wrk1(in=ina) work.dm1(in=inb);
  %include "&_SASWS_/clear_ae_formats.sas";
  by protno pid_a;
  studyid=protno;
  usubjid=pid_a;
  study='Pfizer_2008_81';
  if ina and ^inb then put 'Missing data from DM' pid_a=;
  if inb and first.pid_a then output psademo;
  if ina;
  agegroup='';
  arm='';
  race=racesc;
  race_oth=raceoth;
  aeseq=.;
  aeterm='';
  aedecod=preftext;
  aebodsys='';
  aesev='';
  aeser=aeserc;
  aerel=aecausc;
  aeout='';
  aestdy=aefday;
  aeendy=aetday;
  dataset='DM ADVERSE';
  output wrk;
run;
```

# THE APPROACH: COMBINING THE DATA SETS – (COMBINE THE DATA)

## Code to Remove Data Formats and Informats

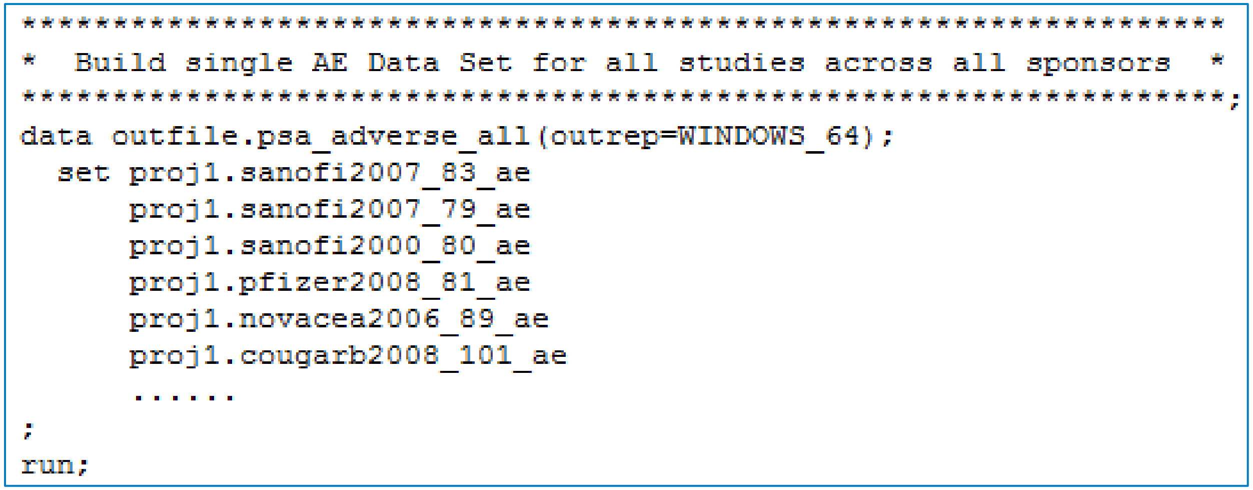
- To reduce notes and any warnings in the SAS log – any SAS informats/formats were removed from the raw input data sets.
- Used %include to use this code



```
format  STUDYID
        USUBJID
        AESEQ
        AETERM
        ....;
informat  STUDYID
        USUBJID
        AESEQ
        AETERM
        ....;
```

## Programs to Combine the Data Sets

- Simple data step procedure with multiple sets



```
********************************************************************
*  Build single AE Data Set for all studies across all sponsors  *
********************************************************************;
data outfile.psa_adverse_all(outrep=WINDOWS_64);
  set proj1.sanofi2007_83_ae
      proj1.sanofi2007_79_ae
      proj1.sanofi2000_80_ae
      proj1.pfizer2008_81_ae
      proj1.novacea2006_89_ae
      proj1.cougarb2008_101_ae
      ......
;
run;
```

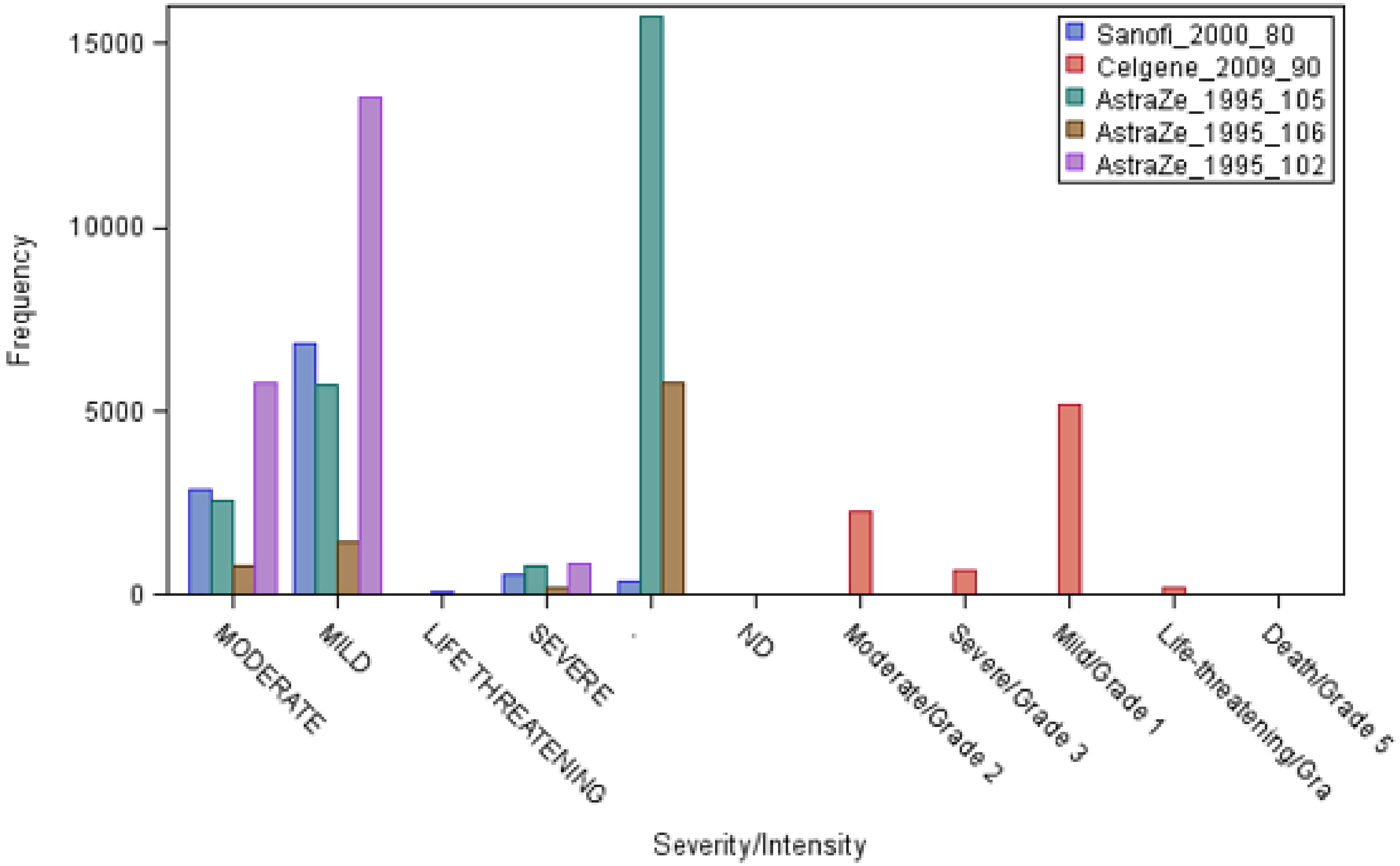
# THE APPROACH: REVIEWING AND DATA QUALITY – (DATA QUALITY)

## Data Quality

- Our most important concern was the quality of the mapped data. Did we assign the proper column during the mapping process.
- An additional programmer was tasked to review the data and confirm correct observations counts and correct patient populations.
- Constantly ran frequencies against the raw data and the harmonized data to verify output, paying particular attention to the remapped columns.
- Any outliers or any data that was questioned by this programmer was reviewed and, if found to be incorrect, the appropriate changes were made to the mapping code.
- No original source data was ever modified.

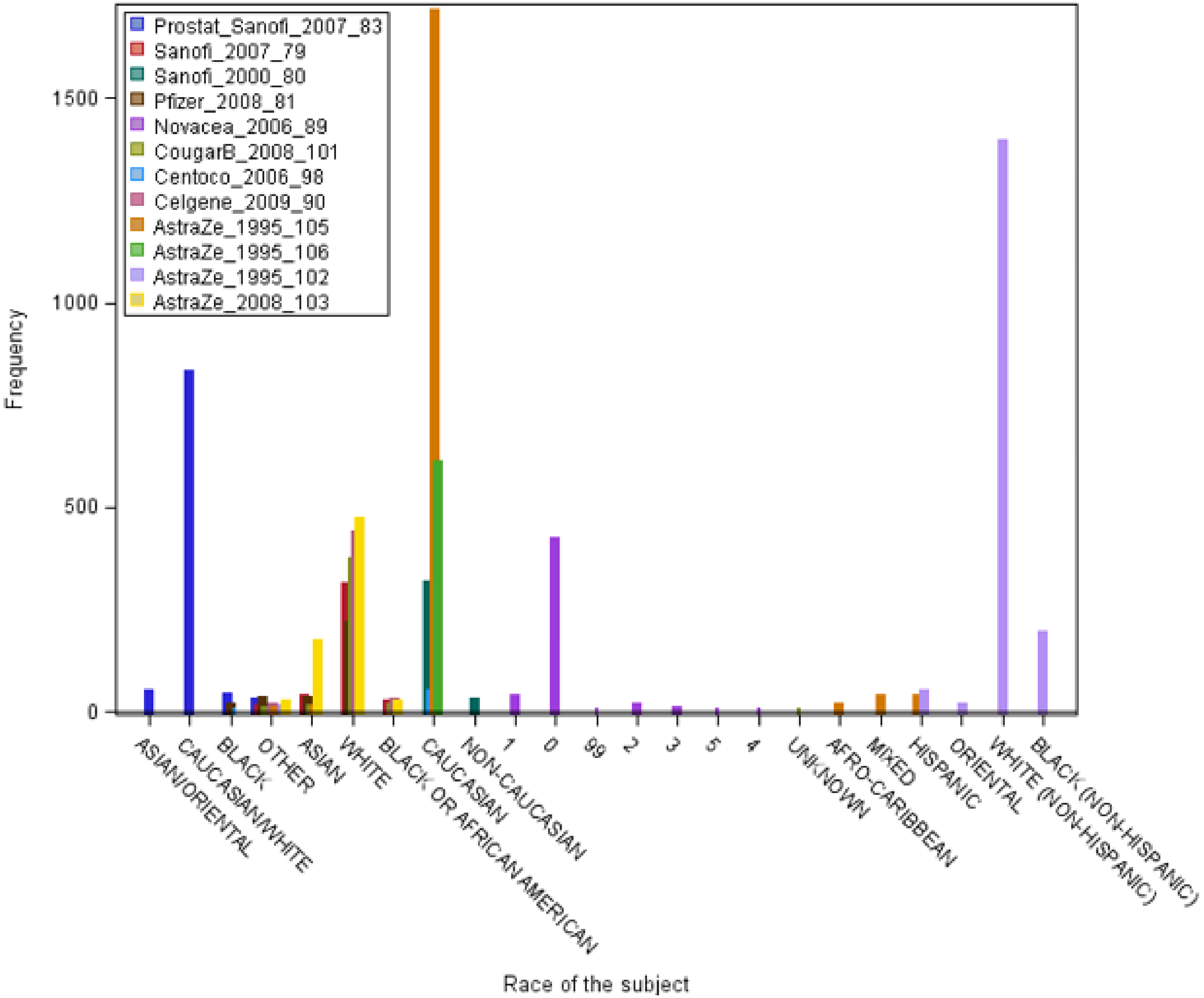
# THE APPROACH: REVIEWING AND DATA QUALITY – (DATA QUALITY)

Figure 4: Adverse Event Severity



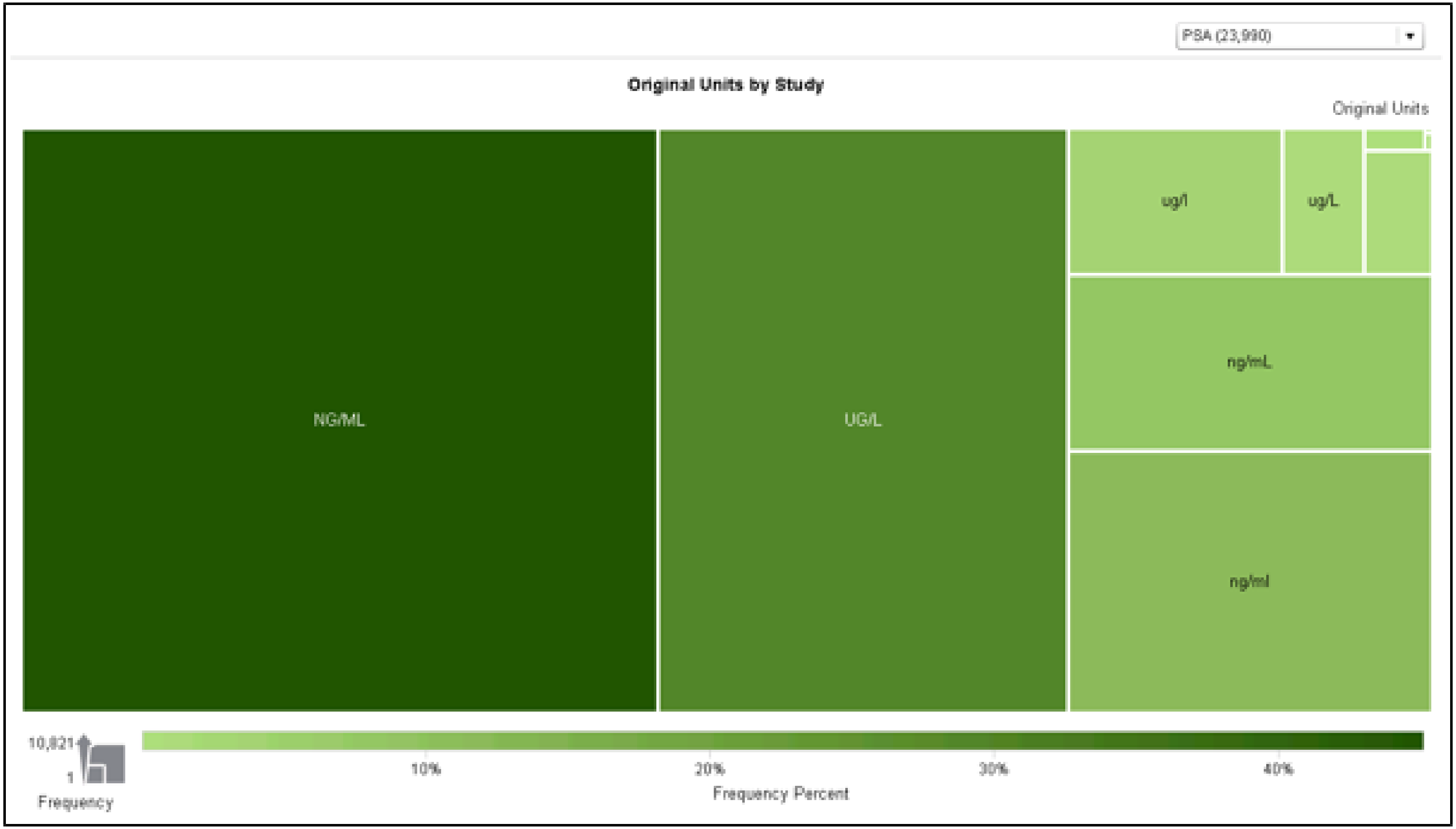
# THE APPROACH: REVIEWING AND DATA QUALITY – (DATA QUALITY)

*Figure 5: Race Group Names Bar Graph*



# THE APPROACH: REVIEWING AND DATA QUALITY – (DATA QUALITY)

*Figure 8: Original Units by Study for PSA Tree Map*



In the upper right corner are four blocks with missing values. Their values from high to low are: missing, MCG/L, UG/l, and NG/DL.

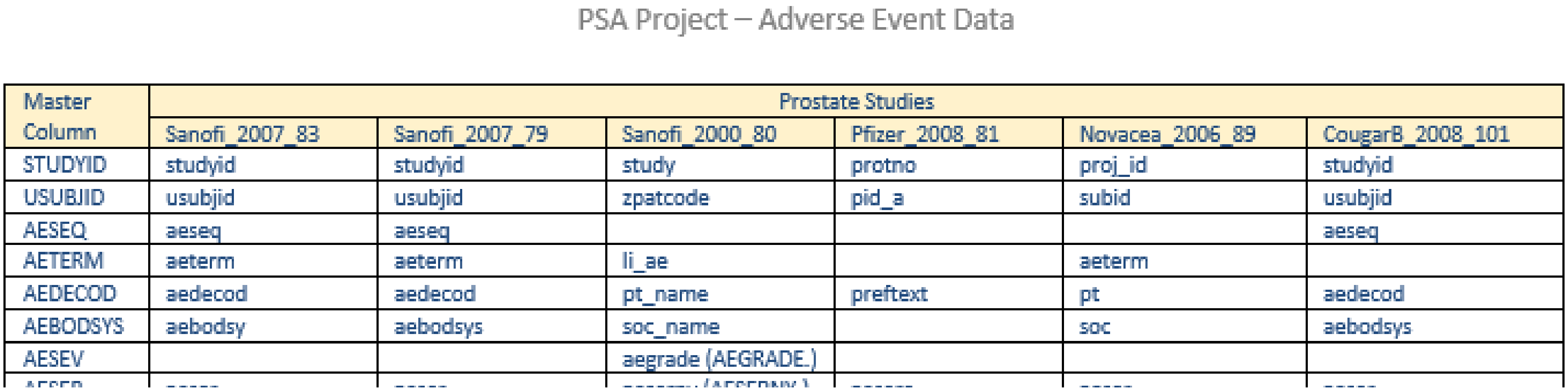
# THE APPROACH: BASIC PROGRAM FLOW

Programming Flow

1. Review the data and identify needed tables and columns.
2. Create a “global” metadata file for each domain. For this project it was the SAS attrib statement used for each domain and across each study.
3. Create mapping programs for each study – should be able to re-use code within sponsor.
4. Create data quality process flow to check the data for correct metadata, patient counts, and any “outliers”.
5. Create code to combine data across studies – simple SET statement.
6. [Optional] Create one process that submits all the code created in items 2-5.

# DOCUMENTATION

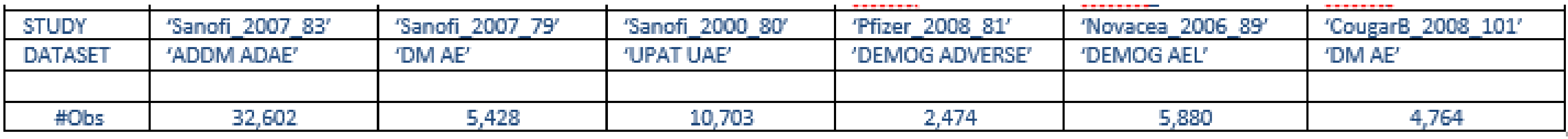
## Data Matrix Document



PSA Project – Adverse Event Data

| Master Column | Prostate Studies | | | | | |
|---|---|---|---|---|---|---|
| | Sanofi_2007_83 | Sanofi_2007_79 | Sanofi_2000_80 | Pfizer_2008_81 | Novacea_2006_89 | CougarB_2008_101 |
| STUDYID | studyid | studyid | study | protno | proj_id | studyid |
| USUBJID | usubjid | usubjid | zpatcode | pid_a | subid | usubjid |
| AESEQ | aeseq | aeseq | | | | aeseq |
| AETERM | aeterm | aeterm | li_ae | | aeterm | |
| AEDECOD | aedecod | aedecod | pt_name | preftext | pt | aedecod |
| AEBODSYS | aebodsy | aebodsys | soc_name | | soc | aebodsys |
| AESEV | | | aegrade (AEGRADE.) | | | |

The data matrix document was dynamic during the development process. The end result is a document that can be provided to the researcher tracing the harmonized data back to the original source columns and source data sets and providing a quick overview of the data.

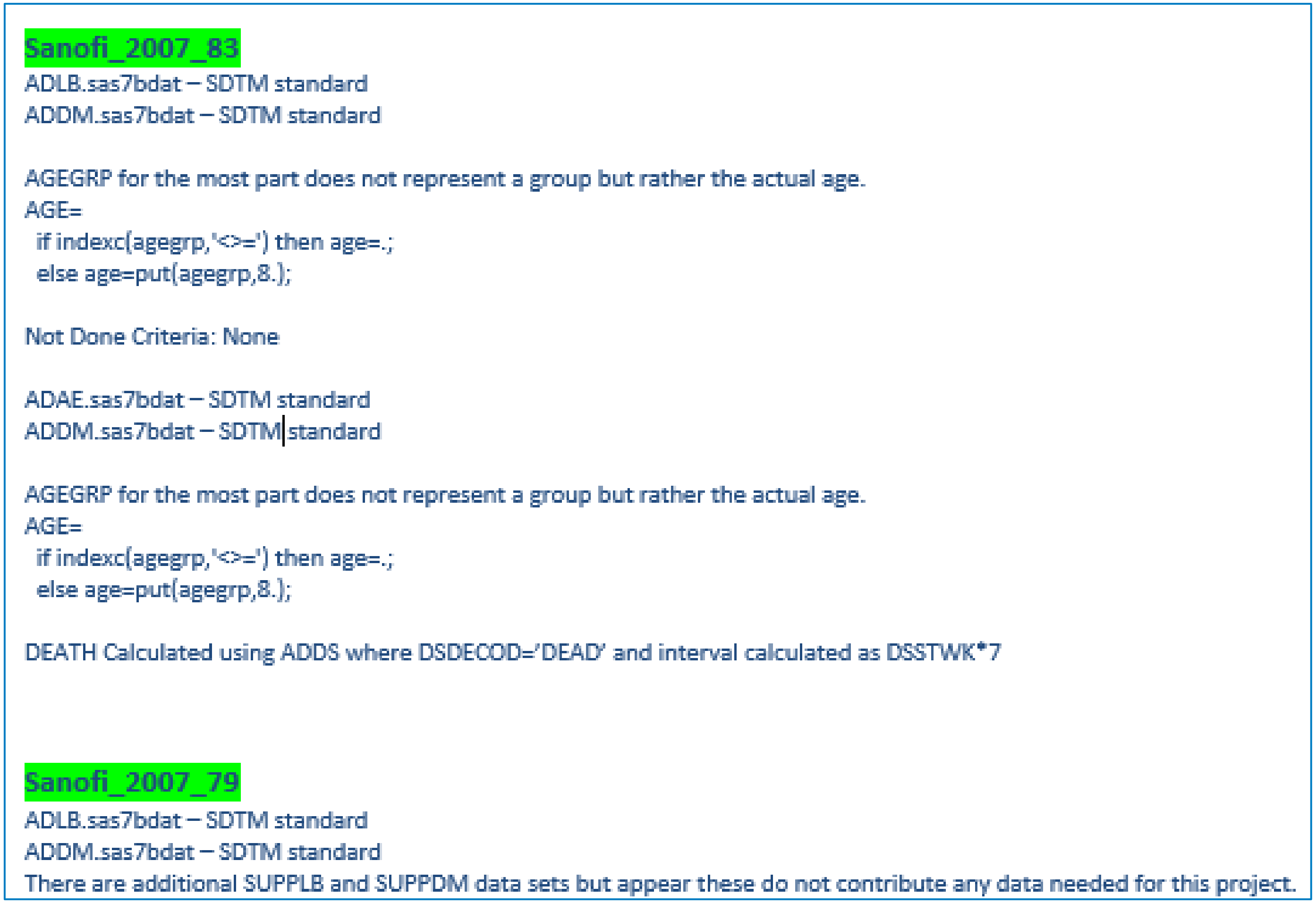


| | | | | | | |
|---|---|---|---|---|---|---|
| STUDY | 'Sanofi_2007_83' | 'Sanofi_2007_79' | 'Sanofi_2000_80' | 'Pfizer_2008_81' | 'Novacea_2006_89' | 'CougarB_2008_101' |
| DATASET | 'ADDM ADAE' | 'DM AE' | 'UPAT UAE' | 'DEMOG ADVERSE' | 'DEMOG AEL' | 'DM AE' |
| | | | | | | |
| #Obs | 32,602 | 5,428 | 10,703 | 2,474 | 5,880 | 4,764 |

# DOCUMENTATION

Data Traceability Document

This was dynamic also and recorded observations and notes about the data. It also contains any decisions that were made during mapping that might affect the harmonized data.



**Sanofi_2007_83**

ADLB.sas7bdat – SDTM standard
ADDM.sas7bdat – SDTM standard

AGEGRP for the most part does not represent a group but rather the actual age.

```
AGE=
  if indexc(agegrp,'<>=') then age=.;
  else age=put(agegrp,8.);
```

Not Done Criteria: None

ADAE.sas7bdat – SDTM standard
ADDM.sas7bdat – SDTM standard

AGEGRP for the most part does not represent a group but rather the actual age.

```
AGE=
  if indexc(agegrp,'<>=') then age=.;
  else age=put(agegrp,8.);
```

DEATH Calculated using ADDS where DSDECOD='DEAD' and interval calculated as DSSTWK*7

**Sanofi_2007_79**

ADLB.sas7bdat – SDTM standard
ADDM.sas7bdat – SDTM standard
There are additional SUPPLB and SUPPDM data sets but appear these do not contribute any data needed for this project.

# GENERAL ISSUES AND THINGS TO PONDER

## Not All Data is Created Equal

- Mixture of character and numeric
- Normalized versus non-normalized
- Some studies were more robust (contained more data)

## Some Studies May Not Fit the Analysis

- May not find what you are looking for in the data – a key column may be missing (ie AEREL)

## To Compute or Not to Compute?

- May need to make a decision to compute relative day, age, gender??

## Age and Age Groups

- If age was not available it was usually reported in an age group – across sponsor this age group was not consistent (ie 40 – 55, 45-55, 50 – 65, etc..)

## Race

- A variety of race types seen here, mostly with the legacy data.

# GENERAL ISSUES AND THINGS TO PONDER

## Categorical Data

- Use of provided data dictionaries and SAS formats
- Cannot always make assumptions

## External Terminology/Dictionary

- Found a combination of COSTART and MedDRA dictionaries
- Made no effort to upgrade to MedDRA

## Dates versus Date Intervals

- Dates were rare in the data no doubt due to de-identification
- Relied on duration – But how is it calculated?? (event-start) or (event-start)+1
- Duration unit – days vs weeks

## Unique Subject Identifiers

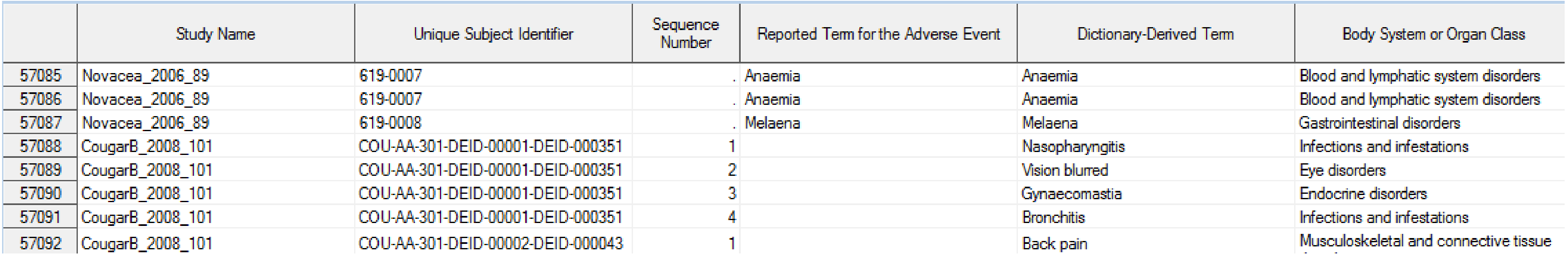
- Some studies simply gave a unique identifier starting with 1 to N number of subjects

## Can the Data be too De-identified?

- In some cases yes, lack of dates, age

# THE FINAL DATA SETS

AE Domain consisted of 127,067 observations



| | Study Name | Unique Subject Identifier | Sequence Number | Reported Term for the Adverse Event | Dictionary-Derived Term | Body System or Organ Class |
|---|---|---|---|---|---|---|
| 57085 | Novacea_2006_89 | 619-0007 | . | Anaemia | Anaemia | Blood and lymphatic system disorders |
| 57086 | Novacea_2006_89 | 619-0007 | . | Anaemia | Anaemia | Blood and lymphatic system disorders |
| 57087 | Novacea_2006_89 | 619-0008 | . | Melaena | Melaena | Gastrointestinal disorders |
| 57088 | CougarB_2008_101 | COU-AA-301-DEID-00001-DEID-000351 | 1 | | Nasopharyngitis | Infections and infestations |
| 57089 | CougarB_2008_101 | COU-AA-301-DEID-00001-DEID-000351 | 2 | | Vision blurred | Eye disorders |
| 57090 | CougarB_2008_101 | COU-AA-301-DEID-00001-DEID-000351 | 3 | | Gynaecomastia | Endocrine disorders |
| 57091 | CougarB_2008_101 | COU-AA-301-DEID-00001-DEID-000351 | 4 | | Bronchitis | Infections and infestations |
| 57092 | CougarB_2008_101 | COU-AA-301-DEID-00002-DEID-000043 | 1 | | Back pain | Musculoskeletal and connective tissue |

DM Domain 8,116 subjects

LB Domain 1,170,346 observations

# CONCLUSION

- This was a great project since it covered various aspects of data that a user would expect from 20+ years of research.

- Data conforming to the SDTM models obviously were the easiest to combine. The legacy data, as expected, required more work but in the end conformed nicely.

- Disease experts/researchers and clinical data programmers clearly benefit any project of this nature

- Effective analysis tools provide excellent data quality review.

# CONCLUSION

- Data harmonization requires careful analysis and understanding of the underlying clinical data especially when legacy data exists without any associated clinical data standard. Document, document, document.
- Choose a target standard such as SDTM when working with legacy data.
- Regard data harmonization as a continuous and valuable learning experience as processes for data harmonization will surely evolve with time.

As a result of this work, currently working on a more robust process to harmonize incoming data for Project Data Sphere®. A questionnaire/checklist was created for sponsors to provide certain information felt necessary to help get researchers started.

# FURTHER INFORMATION

**Project Data Sphere®**

https://www.projectdatasphere.org/projectdatasphere/html/about

**Author Contact information**

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

Gene Lightfoot
SAS Institute Inc.
SAS Campus Drive Q2372
Cary, North Carolina 27513 USA
+1 (919) 677-8000
gene.lightfoot@sas.com

- www.sas.com