Studienauswertung per Knopfdruck. Code-Generierung direkt vom Statistical Analysis Plan - Ist das möglich?

Endri Endri
DataFocus GmbH
Lothringer Straße 23
D-50667 Köln
endri0501@yahoo.de

Zusammenfassung

CDISC SDTM ist der Datenstandard in der klinischen Forschung für die Einreichung bei der FDA. In der Erstellung von ADAM Datensätzen werden dabei bedeutsame Datenableitungen vorgenommen, um den Ansprüchen an die TLF's (Table, Listing, Figure) zu entsprechen.

Die FDA und die PhUSE haben die Arbeitsgruppe „Development of Standard Scripts for Analysis and Programming" gegründet ([1], [2]), deren Aufgabe darin besteht, mögliche Standardskripte für Datentransformationen und -analysen innerhalb klinischer Studien zu identifizieren.

Der Schritt zur Standardisierung ist nichts neues, doch sind wir in der Lage eine komplette Studienanalyse per Knopfdruck zu erzeugen? Die Antwort ist natürlich: „Nein, noch nicht". Der Grund dafür ist die Abhängigkeit vom Studiendesign, von der Datenstruktur sowie den Ableitungsregeln. Oder sei es nur eine kleine Fußnote, die in einer bestimmten Tabelle ausgegeben werden muss.

Dieses Paper beschreibt die Herausforderung der Zeit vor der Arbeitsgruppe der FDA / PhUSE und zeigt auf, wie Programmcodes möglicherweise direkt vom Statistical Analysis Plan (SAP) generiert werden, welche Schwierigkeiten dabei auftreten und wie diese möglicherweise bewältigt werden könnten. Die vorgeschlagene Lösung basiert auf der Logik und Technik, den SAP Text zu analysieren und in einen SAS Code zu übersetzen. Im besten Fall können mit diesem Ansatz gesamte Studien der klinischen Forschung vollständig automatisiert werden.

Schlüsselwörter: Code Generierung, SAP Analysierung, Makro, künstliche Intelligenz

1 Einleitung

Dieses Paper wurde mit mehrjähriger SAS Erfahrung im Bereich Klinische Forschung entwickelt. Die Idee dieses Papers ist es, die Möglichkeit einer automatisierten Umwandlung von Texten im Statistical Analysis Plan (SAP) direkt in SAS Programmen, die zur Auswertung von klinischen Forschungen benötigt werden, zu erörtern.

Folgende Überlegungspunkte werden zunächst im Vorfeld diskutiert:

- Datenstrukturen
  In der klinischen Forschung werden Daten meist in einer CDISC SDTM Struktur gespeichert.
  Clinical Data Interchange Standards Consortium (CDISC) entwickelt eine Reihe von offenen Standards für den Austausch von Daten aus klinischen Studien. Das Study Data Tabulation Model (SDTM) beschreibt die inhaltliche Struktur von Tabellen, in denen die in einzelnen Case Report Forms dokumentierten Fallberichte aus klinischen Studien zusammengefasst und bei der FDA eingereicht werden können.
  Die SDTM Strukturen sind zwar bereits sehr einheitlich, jedoch gibt es immer noch kleine Unterschiede in jedem Pharma-Unternehmen oder auch in Dienstleistungen wie Clinical Research Organisation (CROs) / Auftragsforschungsinstitut, sei es durch zusätzliche Variablen, Studien-spezifische Formate, Labels, usw.
  Aus den SDTM Datensätzen werden klinische Daten meist aber auch in Analysis Dataset Model (ADaM) Datensätze transformiert, sodass die Daten „analysis-ready“ sind. ADaM Datensätze sind leider noch sehr flexibel, weswegen es fast unmöglich ist eine einheitliche Struktur zu erreichen.
  Die Überlegung ist hierbei, ob es möglich wäre, klinische Daten mit SAS Programmen / Macros unabhängig von Datenstrukturen automatisch und korrekt zu erkennen, um eine Studie anhand des SAPs auswerten zu können.

- Statistical Analysis Plan (SAP)
  SAPs werden zwar immer einheitlicher zwischen den Studien innerhalb eines Projekts, aber sie werden niemals vollkommen gleich sein bzw. es werden immer auch Freitexte enthalten sein.
  In jeder Studie sind Unterschiede aufzufinden, wie zum Beispiel Ableitungsregeln / Definitionen für Baseline Werte, Missing Results, usw..
  Wie ist es nun möglich freie Texte im SAP zu analysieren?

- SAS Programme als Ausgaben
  Zunächst - in Form eines kleineren Problems - sind die Header in jedem SAS Programm zu berücksichtigen, die je nach Unternehmen unterschiedlich sind. Darüber hinaus besteht die weitere Überlegung, ob SAS Programme eher mit Daten- oder PROC SQL-Schritten geschrieben werden sollen.
  Ein anderer wichtiger Aspekt ist die Nutzung von Standardmakros. In jedem Unternehmen wurden meistens Standardmakros entwickelt. Die Arbeit in jeder Abteilung / jedem Unternehmen ist an Standard Operating Procedures (SOP) ausgerichtet, welche besagen, dass für die Auswertung einer klinischen Forschung immer Standardmakros verwendet werden sollen, sofern dies möglich ist.

Natürlich gibt es noch viele andere Überlegungen zu diesem Paper, aber die genannten drei Überlegungen sollen als Basis dieser Entwicklung dienen.

2 Lösung

Im Folgenden werden die einzelnen Überlegungen im Detail diskutiert und eine bestmögliche Lösung für diese vorgeschlagen.

2.1 Datenstrukturen – Indexierung

Jeder kleine Unterschied in Datenstrukturen oder aber auch in Inhalten kann zu unterschiedlichen Ergebnissen führen.
Folgende Ansätze sollen dieses Problem beheben.

- SASHELP.VCOLUMN
  Mit der Hilfe von SASHELP.VCOLUMN können alle Variablen sowie deren Attribute gespeichert werden.
  SASHELP.VCOLUMN beinhalt folgende Informationen:
  - Datensatz-Name innerhalb der SAS Library
  - Variablen-Name von jedem Datensatz
  - Variablen-Länge, -Format und -Label

- FORMAT Katalog
  Studienspezifische Formate können mit Hilfe der Prozedur FORMAT in einer Datei gespeichert werden.

- Distinct-Werte
  Für eine bessere und komplette Erfassung von Daten können auch die „unique" Textvariablen gespeichert werden. Dies ist besonders wichtig für größere Datensätze, um später in weiteren Prozessen eine schnellere Suche von bestimmten Informationen zu ermöglichen.

Es gibt noch andere Methoden wie klinische Daten unabhängig von deren Strukturen korrekt erkannt werden können. Aber die bisher genannten Methoden sollen aufzeigen, dass eine Erkennung von Datensätzen sowie deren Inhalte kein großes Problem darstellen.

2.2 SAP-Text Analysierung

Im SAP wird festgelegt, wie eine klinische Studie ausgewertet werden soll. Der SAP beinhaltet Informationen über die Studie und definiert welche Tabellen, Figures und Listings erstellt werden sollen. In manchen Fällen beinhaltet der SAP auch Mocktables, die zum Layout der gewünschten Tabellen dienen.

In den meisten Fällen werden bereits zwar standardisierte Texte in SAP verwendet, jedoch ist SAP ein sehr offenes Feld, in dem heutzutage immer noch freie Texte verfasst werden.

Wie die Texte aus dem SAP automatisch analysiert werden können, wird im Folgenden im Detail diskutiert.

- Fuzzy-Wuzzy's Methode
  Der erste Ansatz war ein Versuch mit einer Wort-Vergleichsmethode, die Texte zu analysieren. Dieser Versuch jedoch misslang, da zumeist Abkürzungen im SAP verwendet werden oder sich Tippfehler im SAP befinden.

  Durch Meinungsaustausch und Diskussionen mit Freunden im Programmierumfeld entstand die Überlegung, SAS und Python miteinander zu verbinden, zumal. Python eine Programmierungssprache ist, die bereits eine Methode zum Zeichenkettenvergleich anbietet.
  Die Fuzzy-Wuzzy Methode im Python vergleicht Zeichenketten und drückt die Ähnlichkeit von Wörtern in Prozentwerten aus.

  Da die Fuzzy-Wuzzy Methode von Python für allgemeine Fälle entwickelt worden ist, führt sie leider nicht zu einem konkreten Vergleich innerhalb der klinischen Forschung. Beispielsweise werden „Normal" und „Abnormal" mit einer Ähnlichkeit von über 80% durch die Fuzzy-Wuzzy Methode im Python bewertet.

  Um eine bessere und flexiblere Berechnung der Fuzzy-Wuzzy Methode zu erhalten, wurde eine eigene Berechnung mit Hilfe von SAS-Makros entwickelt.
  Dabei wurde lediglich eine Zusammensetzung aus Schleifen und dem Buchstabenweisen Vergleich der Zeichenketten in die Berechnung eingebaut.

- Eigenwörterbuch
  Es wurde ein Eigenwörterbuch im System entwickelt, um zum einen Synonyme erkennen, zum anderen aber auch Studien-spezifische Definitionen übersetzen zu können. Beispielsweise haben „Subject" und „Patient" eine gleiche Bedeutung.

  In diesem Wörterbuch werden Schlüsselwörter in zwei Typen unterteilt:
  - System-Schlüsselwörter.
    Zu System-Schlüsselwörtern gehören zum einen allgemeine SAS Prozeduren, Funktionen und Syntax, zum anderen Definitionen von Standardmakros der Abteilung / des Unternehmens, so dass alle Standardmakros auch in diesem System eingebettet werden.
    Andere Schlüsselwörter sowie Synonyme, welche sehr oft im SAP verwendet werden, wie zum Beispiel: „only", „at least", können als Schlüsselwörter definiert werden.

  - Inhalts-Schlüsselwörter.
    Zu Inhalts-Schlüsselwörtern gehören alle anderen, die nicht System-Schlüsselwörter sind, wie zum Beispiel Studien-spezifische Inhalte in den klinischen Daten.

Das Eigenwörterbuch ist eine intelligente Methode, um Definitionen sowie Schlüsselwörter schneller festlegen zu können. Je mehr Studien das System bereits ausgewertet hat, desto intelligenter und schneller wird das ganze System, da in ihm eine Memory-, Erinnerungsmethode eingebaut ist.
Diese Memory-, Erinnerungsmethode speichert zum einen alle Wörter, zum anderen auch die Definitionen sowie deren Lösungsmethode in seinem System.

Mit dieser Eigenwörterbuch-Methode kann gleichzeitig auch die eigene Fuzzy-Wuzzy Berechnung zum großen Teil eingespart werden.

2.3 SAS Programme – Code-Generierung

Die Code-Generierungs-Methode wurde bereits auf der PhUSE Conference 2011 in Brighton [3] vorgetragen. Dabei wurde im Detail diskutiert, wie ADaM Datensätze mit Hilfe von Microsoft Excel programmiert werden können.
Die gewünschten ADaM Datensätze werden in einer Excel-Datei definiert. Anschließend wird diese Datei von SAS Makros eingelesen und dann in SAS Programme transformiert.

Dieses PhUSE Paper ist sehr hilfreich in der klinischen Forschung, da zum einen in der klinischen Forschung alles fehlerfrei programmiert und zum anderen im Detail dokumentiert werden muss.

Im Rahmen der Systemmakro-Entwicklung wurden jedoch noch mehr Funktionen und eine höhere Flexibilität eingebaut. Folgende wichtige Punkte wurden in der Code-Generierung berücksichtigt und in SAS Programmen ausgegeben.

- Einrücken
- Groß/Klein-Buchstaben
- Genügend Kommentare
- Template Technik für Standardmakros, Header, usw..

Eine nützliche und doch überaus einfache Technik innerhalb der Code-Generierung ist die Template Technik. Damit kann jedes beliebiges Layout (wie z.B. Header, Kommentare) in SAS Programmen ausgegeben werden.

In der klinischen Forschung werden oft ähnliche Tabellen jedoch mit unterschiedlichen Einschränkungen angefragt, wie z.B. demographische Tabellen, aber mit drei verschiedenen Populationen (Safety Population, Per Protocol Population, Efficacy Population). Es wäre einfacher in der Validierung, wenn diese Arten von Tabellen mit einer Makro-Programmierung erstellt werden könnte. Dieser Aspekt wurde mit der Template Technik abgedeckt.

3 Beispiele

3.1 Tabelle Programmierung: ". . . (MALE ONLY)"

Als einfaches Beispiel möchten wir versuchen die Informationen aus der angefragten Tabelle zu extrahieren, wie z.B. „Male only" Population-Begrenzung.
Wie bereits beschrieben, soll das System automatisch die vorgegebenen Daten „verstehen" und die wichtigsten Informationen (Metadaten) extrahieren, wie z.B. Datenstrukturen, studienspezifische Formate usw.

Die vorgegebenen Daten sind wie folgt:

```
* Study specify format;
PROC FORMAT LIB = WORK;
  VALUE _sex
     1 = 'M'
     2 = 'F';
QUIT;
* Sample unknown data structure;
DATA random;
  ATTRIB subj FORMAT = $20.
  gender FORMAT = _sex.;
  subj = '1'; gender = 1; OUTPUT;
  subj = '2'; gender = 2; OUTPUT;
RUN;
```

Mit allen Informationen aus den gegebenen Daten und der angefragten Tabelle kann folgende Lösung als Resultat der Automatisierung-Prozesse betrachtet werden:

- Eigenwörterbuch
  Innerhalb des Eigenwörterbuch wurde beispielsweise „only" als System-Schlüsselwort definiert: „IF …… EQ …….. "
- „Male" ist kein System-Schlüsselwort und deshalb wird mit der internen „string-compare method" (Fuzzy Wuzzy's Methode) weiter analysiert.
- Als Ergebnis erhält man: IF gender EQ 1;

3.2 Population Definition

Eine „safety population" ist z.B. wie folgt in der SAP definiert:
". . . if he/she is randomized to a treatment group and has taken at least one unit of the study medication and has post-treatment safety data available"

In diesem Beispiel werden die CDISC SDTM Datensätze verwendet. Somit ist die Lösung wie folgt;

- "randomized" ist in DS dokumentiert.

- “treatment group” Informationen sind in der Variable ARMCD im DM Datensatz zu finden.
- “at least” ist als System-Schlüsselwort als „HAVING MIN (##var##) >= ##“ definiert.
- “study medication” Informationen sind im EX Datensatz zu finden.
- “post-treatment safety data” Informationen sind in z.B. in VS zu finden.

3.3 Mocktables / SAS generator

Das Ergebnis der SAS Code-Erzeugung ist wie folgt:

```
/************************************************************
* Program name : ADEG_script.sas
* (This script is automatic generated by %ADaM_Gen
* Author       : Endri
* Date created : 29.07.2011
* Study        : (Study number)
*                (Study title)
* Purpose      : ADEG - Analysis Dataset for Electrocardiogram
* Template     :
* Inputs       :
* Outputs      :
* Program completed : Yes/No
* Updated by : (Name) - (Date):
*              (Modification and Reason)
************************************************************/

/* Analysis Value # Analysis Value - Converted from EGSTRESC - SDTM.EG */
DATA adeg_010_eg;
  SET sdtm.eg;
  aval = egstresn;
  aval2 = %_help_c2n(var = egstresc);
RUN;

/* Baseline Value - SDTM.EG */
DATA adeg_020_base;
  SET sdtm.eg;
  base = egstresn;
  WHERE EGBLFL = 'Y';
RUN;

/* Join all */
%auto_join( intab  =   adeg_010_eg
                     $ adeg_020_base # base
          , outtab = adeg_30_all);

/* Change from Baseline # Change Group - DERIVED */
DATA adeg_40_chg;
  SET adeg_30_all;
  chg = AVAL - BASE;
  IF (CHG/BASE) < 10 THEN chg1g = 1;
```

Abbildung 1: Code Generator [3]

4 Diskussionen

Diese automatisierte Lösung kann in folgenden Gebieten angewendet werden:

- Einfache Zuweisung
  Z.B. Zuweisung der lokalen Laborparameternamen in CDISC Controlled Terminology Codelist
- Mapping / Derivation
  Z.B. SDTM Mapping, ADaM Mapping
- Query / Data Validation
- Programmierung von Tabelle und Listing

Die Ideen, um das Ziel einer vollständig automatischen Analyse von SAP Texten und deren Umwandlung in SAS Programme zu erreichen, beinhaltet einen sehr komplexen Algorithmus.

Das System wurde bereits in zahlreichen Situationen getestet, wie z.B. ADSL Mapping, Baseline Definition, Tabelle/Listing Programmierung usw. Ein erstes Testergebnis von dem ganzen System ist die Produktion von SAS Programmen für die Auswertung einer Studie mit 100 Tabellen/Listings innerhalb wenigen Stunden!

Das System ist noch in der Entwicklungsphase, wurde unter Betreuung sowie Inputs / Kritiken von Herrn P. Jähnig (PJ Statistics) entwickelt und wird demnächst durch freundliche Unterstützung der Firma ICRC Weyer getestet.

Natürlich gibt es noch viele weitere Aspekte, die während der Auswertung einer klinischen Forschung berücksichtigt werden müssen. Dieses Paper gibt nur eine kleinen Überblick darüber, wie SAS Programme direkt aus einem Statistical Analysis Plan automatisch generiert werden können.

Das Ziel ist es, eine vollständige Programmierung für die Auswertung einer klinischen Studie innerhalb kürzester Zeit zu ermöglichen, um etwas Zeit in der Validierung sowie Daten-Validierung zu gewinnen.

Literatur

[1] PhUSE Working Group: http://www.phuse.eu/CSSWorking-Groups5.aspx

[2] PhUSE Blog (Just press the button):
http://www.phuse.eu/blog/just-press-the-button

[3] E. Endri: PhUSE Conference 2011, Brighton (DH07: Much ADaM about Nothing - a PROC Away in a Day):
http://www.phusewiki.org/docs/2011%20Papers/DH07%20paper.pdf