Conversion de données vers SDTM

Groupe des utilisateurs francophones des standards CDISC

10 février 2011

Thierry Lambert

AdClin

Spécialisée en traitement des données (après le data management, avant l'analyse statistique)

- Logiciels de reporting
  - TPF: listings et rapports sous SAS
  - Report Builder: gestion des tables de résultats sous Word
  - Navigator: appréhension et exploration rapide des données
- Récupération d'historique de données cliniques
  - En entrée, formats divers et variés
  - En sortie, une base exhaustive et analysable

Projet

- Conversion de 10 études d'oncologie phases 1/2 au format SDTM
- Entrée
  - Documentation CDISC, NCI Controlled Terminology
  - Standards du Sponsor
  - Protocoles, SAP, CRF, define.xls/doc/pdf
  - Raw Data, Analysis Data
- Sortie
  - Fichiers SDTM 3.1.2 au format transport SAS v5 (*.xpt)
  - CRF annoté correspondant
  - « Reviewer's Guide »
  - define.xml
- Contrôle
  - OpenCDISC, WebSDM

Entrées générales

- Documentation CDISC (ordre chronologique)
  - Case Report Tabulation Data Definition Specification (define.xml) 1.0.0 (2005, 45 pages)
  - Study Data Tabulation Model 1.2 final (2008, 35 pages)
  - SDTM Implementation Guide 3.1.2 final (2008, 298 pages)
  - XML Schema Validation for Define.xml 1.0 (2009, 12 pages)
  - SDTM Metadata Submission Guidelines 1.0 draft (2009, 23 pages)
- Standards Sponsor
  - reprise et extension du SDTM-IG 3.1.2, 277 pages
  - conseils pratiques pour contourner les situations de non-validation des sorties

Entrées par étude

- Raw Data: format proche du CRF
- CRF annotés vers ces Raw Data
- Analysis data: format assez proche de SDTM 3.1.1 + extensions
- Define.xls au format « classique » SDTMIG:
  - datasets: nom, label, clé (théorique), structure
  - variables: nom, label, type/length, controlled terms or format, origin, role, comments, core

Sorties

- Fichiers SDTM 3.1.2 au format transport SAS v5 (*.xpt)
  - doivent suivre les règles décrites dans SDTMIG
- CRF annoté correspondant
  - pages uniques seulement
  - double table des matières
  - système de couleurs de fond par domaine sur chaque page
- « Reviewer's Guide »
  - documentation de tout ce qui ne valide pas
  - autres commentaires
- define.xml
  - description de tout ce que précède, et des liens entre eux
  - dans la limite de ce qui est dicible actuellement

Validation

- OpenCDISC
  - gratuit
  - sur-ensemble déclaré de WebSDM
- WebSDM
  - payant (euphémisme)
- Comparaison sur une étude
  - Les problèmes de sévérité « high » détectés sont les mêmes
  - Les autres se recouvrent à 90%

Méthode de conversion

- Entrer la documentation dans une base de données
  - domaines & classes de domaines
  - variables associées
  - listes de codes
- Charger la définition de chaque étude en entrée
  - datasets
  - variables
  - description CRF
- Etablir les « maps »
  - des données en entrée vers le SDTM
  - des modules uniques du CRF vers les données en sortie

Méthode de conversion

- Valider les « maps »
  - saisie/validation des transcodages
- Génération automatique du CRF annoté
  - positionnement manuel des annotations => mise à jour des maps et/ou de la description du CRF (« not entered in database »)
- Dérivations/corrections systématiques
  - exemple: une variable « expected » qui n'est pas là => la générer à manquant + commentaire « not collected »
  - régler les problèmes « classiques » de validation comme « start > end » ou « 24:00 illegal »
- Génération du define + validation de l'ensemble

Difficultés

- Les documents se contredisent
  - external codelist « ISO8601 » for datetime variables
  - --TEST est le label pour --TESTCD, donc pas plus de 40 caractères, mais IETEST fait 200 caractères
- La validation valide quoi ?
  - Qu'est-ce qu'une violation « grave » ?
  - Si on documente la violation, tout va bien ?
  - Certaines règles ne sont pas validables automatiquement.
  - L'exemple donné par SDTM-MSG ne valide pas
    - define.xml: 25 erreurs + 34 avertissements
    - data: 22 erreurs, 437 avertissements, 4 infos (n'a pas pu valider QS)
- La standardisation n'est souvent qu'apparente (technique)
  - datasets SUPP-- pour cacher le nombre variable de colonnes
  - fichiers de « findings » sur-verticalisés + exigence des value-level metadata

Limites des validateurs

- NCI: la RACE est codée par la liste *extensible* suivante:
  - AMERICAN INDIAN OR ALASKA NATIVE
  - ASIAN
  - BLACK OR AFRICAN AMERICAN
  - NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER
  - WHITE
- Autre liste, qui validera puisque la liste est extensible:
  - CAUCASIAN
  - BLACK
  - ASIAN/ORIENTAL
  - OTHER
- Mais il est précisé sur le site de la FDA que la liste est extensible en *subdivisant*, mais doit être *regroupable* dans les catégories de la liste de base.
- Donc il faut documenter une exception dans le reviewer's guide.

Validations

- Define.xml soit suivre son DTD
  - minimum syndical
- Il ne doit pas y avoir de contradiction entre les données et define.xml
  - assez facile
- Pas de message de sévérité « high »
  - empêchent (paraît-il) le chargement dans Janus
  - un exemple: start > end
- Documentez les exceptions de sévérité « medium » et « low » comme connues

Recommandations (conversion)

- ne faites pas entrer les données avec un chausse-pied dans SDTM, définissez plutôt des domaines adaptés (X, Y et Z)
- saisissez les standards sous forme structurée (y compris les informations cachées dans les commentaires), pour pouvoir les lire par programme
- décrivez formellement le CRF et le map de vos données vers SDTM, puis générez automatiquement les fichiers finaux, le CRF annoté et le define.xml
- interfacez-vous avec NCI, ne le prenez pas comme base (mélange terminologie et codelist)

Recommandations (en amont)

- ne suivez pas le SDTM
- n'écrivez pas de define
- documentez
  - extensivement
  - de façon lisible par programme

Ne suivez pas SDTM

- La FDA a déjà demandé un format XML de transfert: les limitations SAS v5 vont disparaître
- Define.xml vient d'ODM, et est sémantiquement inadapté: ne le prenez pas comme base
- La verticalisation aboutit à une dégradation de la qualité des données cliniques, en particulier des méta-données
  - disparition des formats: liste des choix possibles non maintenue (le define n'est pas couplé aux données, le format de la colonne « controlled term or format » n'est pas... contrôlé)
  - verticalisation à outrance: les attributs des variables disparaissent
    - où indiquez-vous que « HEIGHT » est une longueur, alors que « WEIGHT » est une masse, et donc que « kg » n'est pas une unité qui s'applique à HEIGHT ? (impossible de le dire avec define.xml)
    - où est-il indiqué que ces variables sont numériques (VSORRES est caractère) ?
    - si vous appliquez la terminologie standard pour LBTESTCD, où documentez-vous quelles autres variables doivent être connues pour fixer la valeur de LBSTRESU quand LBTESTCD = « GLUC » ?
    - pouvez-vous écrire un programme qui lit les indications précédentes et vérifie que les données suivent les règles documentées ?

N’écrivez pas de define

- Nom, type SAS, ordre dans le fichier, label v5: maintenus dans SAS, donc duplication = source d’erreur
- « Controlled Terms, Codelist or Format » : joyeux mélange, inexploitable par programme
  - aggravé par la structure verticale des « findings »
- Colonne « Origin »
  - « CRF » ou « Derived » insuffisants, souvent mélangés du fait de la verticalisation
- « Comments » : pour qui?
  - le programmeur qui passe des raw data aux analysis datasets
  - le statisticien qui analyse les données
  - le reviewer qui essaie de comprendre les données
- Où entrez-vous un commentaire relatif à plusieurs variables ?

Documentez

- Vos données et méta-données
  - noms longs
  - structure naturelle (horizontale)
  - custom domains (X, Y, Z)
  - types détaillés (date, time, datetime, duration, string, float, coded…)
  - tables auxiliaires (formats, scheduled visits, etc.)
  - commentaires internes vs. externes
  - links vers programmes, pas de copier/coller de code
- Leur « map » sur le standard
  - Générer la soumission devrait idéalement être 100% automatique (y compris CRF annotés)
  - Evolution du standard = évolution du générateur