Datenimport per SAS-Makro in papier-basierten EDC-Studien

Michael Nonnemacher
Institut für Medizinische Informatik, Biometrie und Epidemiologie - IMIBE
Universitätsklinikum Essen
Hufelandstrasse 55
D-45122 Essen
michael.nonnemacher@uk-essen.de

Dorothea Weiland
Institut für Medizinische Informatik, Biometrie und Epidemiologie - IMIBE
Universitätsklinikum Essen
Hufelandstrasse 55
D-45122 Essen
dorothea.weiland@uk-essen.de

Zusammenfassung

Bei der Datenerfassung in klinischen Studien gibt es neben der „klassischen“ papier-basierten Datenaufzeichnung (mit anschließender Erfassung z.B. im Datenzentrum) und den papierfreien EDC-Lösungen auch Mischformen. Eine solche Mischform ist die Aufzeichnung der Daten mit Hilfe eines elektronischen Stiftes, die auch als papier-basierte EDC bezeichnet wird. Dabei werden die Daten auf Spezialpapier durch einen Stift mit integrierter Kamera aufgezeichnet und nach der Übertragung durch ein web-basiertes Interface und einer Online-Verifizierung, die die Zweiteingabe ersetzen soll, im XML-Format zur Verfügung gestellt, wobei zu jeder Version einer jeweils einzelnen Seite zum einen je eine separate XML-Datei erzeugt wird und zum anderen ein elektronisches Abbild im JPG-Format mit integriertem Zeitstempel.

Das o.g. papier-basierte EDC-System wird am IMIBE zur Zeit in ersten Studien eingesetzt. Da die Daten nach erfolgter Verifikation direkt nach SAS importiert werden, haben wir zur Zusammenführung und Aufbereitung der XML-Dateien ein Paket von SAS-Makros geschrieben.

Die Charakteristika der XML-Dateien (Fehlen einer einheitlichen Dateistruktur, Übermittlung aller Werte im Textformat, durch Kommata getrennte Wertelisten für Items mit multiplen Antworten, Versionenkontrolle und Nachführung der Änderungen) bedingen die benötigten Makro-Funktionalitäten. Letztere umfassen das Einlesen, die Anlage einer einheitlichen Standardstruktur, die Ableitung von Zahlen- und Datumswerten aus den Rohwerten mit gleichzeitigen Datenvalidierungschecks, die Nachführung von Datenänderungen (besonders bei Items mit multiplen Antworten programmtechnisch komplex) die Führung eines Trackingsystems und Import-Logs sowie die Erzeugung der Standardstruktur der Analyse-Datasets.

Die Anwendung der SAS-Makros kann durch Personen mit relativ geringen SAS-Kenntnissen erfolgen. Eine Anpassung an die Besonderheiten einer gegebenen Studie ist leicht möglich. Nach unseren bisherigen Erfahrungen ist das hier angesprochene Hybridsystem für die Prüfzentren relativ leicht zu handhaben, bringt jedoch einen erhöhten Aufwand im Datenmanagement mit sich.

Schlüsselwörter: EDC, elektronischer Stift, Datenimport, Klinische Studien, XML.

1 Einleitung

In klinischen Studien gibt es verschiedene Formen der Aufzeichnung und Erfassung von Daten. Am längsten im Einsatz ist die „klassische“ papier-basierte Form, bei der die Daten im Prüfzentrum auf einer papier-basierten Case Report Form (CRF) aufgezeichnet und anschließend (z.B. im Datenzentrum) durch Dateneingabekräfte in eine Studiendatenbank übertragen werden. Zum anderen gibt es papierfreie Lösungen, bei denen die Daten mit Hilfe einer entsprechenden Software im Prüfzentrum direkt in die Studiendatenbank eingetragen werden (Electronic Data Capture, EDC). Daneben gibt es aber auch Mischformen, die die Aufzeichnung auf Papier mit einer automatisierten oder teilautomatisierten Erfassung der Daten (z.B. durch Einscannen der Papier-CRF) verbinden.

Eine solche Mischform ist die auch als „papier-basierte EDC“ bezeichnete Aufzeichnung der Daten mit einem elektronischen Stift. Dabei wird für die CRF ein spezielles Papier verwendet, das über ein aufgedrucktes Punkteraster eine eindeutige Zuordnung eines Eintrags auf eine bestimmte Stelle einer bestimmten Seite ermöglicht. Die auf den Papier-CRFs eingetragenen Angaben werden von einer in den Stift integrierten Kamera mitverfolgt. In der Literatur ist bisher v.a. der Einsatz solcher Systeme im Bereich der Patientenversorgung beschrieben. Cole et. al. vergleichen zwar verschiedene Dateneingabesysteme für klinische Studien, beziehen sich aber nur auf die Datenaufzeichnung und Verifikation, ohne sich zu weiteren Schritten der Verarbeitung mit dem elektronischen Stift gewonnener Daten zu äußern [1].

Am IMIBE wird zur Zeit das System dotforms® der Firma PharmaForms (http://www.pharmaforms.com) eingesetzt. Jeder Stift ist individuell einem Prüfarzt zugeordnet. Mit einer Dockingstation werden die Daten über ein Web-Interface auf einen Datenserver übertragen, wobei diese Übertragung nicht nach jeder erfassten Seite erfolgen muss (der Stift kann die Daten von bis zu 1000 Seiten zwischenspeichern). Dort werden programmtechnisch die Koordinaten den jeweiligen Variablen zugeordnet, die Werte für Checkbox-Felder werden ausgelesen, und für Zahlenfelder erfolgt eine Handschrifterkennung. Danach werden die Daten freigegeben für eine Online-Verifizierung, die die Zweiteingabe ersetzen soll und bei der in der von uns verwendeten Programmversion auch die Freitexte eingegeben werden müssen. Hierzu steht eine Oberfläche zur Verfügung, die die Werte so darstellt, wie sie vom System interpretiert wurden (z.B. werden alle Einträge, die als 0 erkannt wurden, in der Rubrik „0“ dargestellt) und zum Abgleich ein Abbild der jeweiligen CRF-Seite im GIF-Format anbietet. Nach erfolgreicher Verifikation werden die Daten zum Download bereitgestellt. Zu jeder einzelnen CRF-Seite für einen Patienten wird je eine separate XML-Datei erzeugt. Bei Datenänderungen gibt es pro Patient, CRF-Seite und Version eine weitere XML-Datei, die nur die Änderungen enthält. Darüber hinaus wird zu jeder Version einer Seite ein elektronisches Abbild im JPG-Format mit integriertem Zeitstempel bereitgestellt.

Das dotforms®-System wird am IMIBE zur Zeit in zwei Studien eingesetzt. Da die Daten nach erfolgter Verifikation nicht in einer Studiendatenbank zwischengespeichert,

sondern direkt nach SAS importiert werden, haben wir zur Zusammenführung und Aufbereitung der XML-Dateien ein Paket von SAS-Makros geschrieben, das wir hier vorstellen möchten. Darüber hinaus berichten wir über Erfahrungen im Produktivbetrieb.

2 Aufgabenstellung

Beim Einsatz eines papier-basierten EDC-Systems stellt sich die Frage, ob die Daten nochmals in einer Datenbank zwischengespeichert oder direkt nach SAS importiert werden sollen. Wir haben uns für den direkten Import nach SAS entschieden, woraus sich die Notwendigkeit ergibt, die in den einzelnen XML-Dateien abgespeicherten Daten zu importieren, sinnvoll zusammenzuführen und aufzubereiten und für die Auswertung geeignete Analyse-Datasets zu erzeugen. Diese Schritte werden über ein Paket von SAS-Makros abgearbeitet, wobei der Schwerpunkt der folgenden Darstellung auf den Makros zum Datenimport und zur Datenaufbereitung liegt.

3 Beschreibung der SAS-Makros

3.1 Input

Die XML-Dateien weisen einige Charakteristika auf, die für die Programmierung der Import-Makros relevant sind.

Wie oben bereits beschrieben wird für jede einzelne Version einer einzelnen CRF-Seite eines gegebenen Patienten eine separate XML-Datei erzeugt, die im Fall der ersten Version alle in dieser Version ausgefüllten Items enthält und im Fall einer Änderungsversion nur die jeweils geänderten Items. Dies führt dazu, dass für ein und dieselbe CRF-Seite sowohl die Struktur der XML-Dateien variieren kann als auch die Anzahl möglicher XML-Dateien zu dieser Seite beliebig groß ist je nach Häufigkeit der Änderungen. Darüber hinaus ist eine Versionskontrolle notwendig, damit die Änderungen richtig zugeordnet werden können.

Jedes Item auf einer CRF-Seite besitzt ein eigenes Tag, allerdings enthält die jeweilige XML-Datei nur die Tags für die auf dieser Seite für diesen Patienten ausgefüllten Items. Mehrere Werte zu einem Item werden durch Kommata getrennt (z.B. bei Checkboxen mit multiplen Antworten zu einer Frage). Alle Werte werden im Textformat übermittelt (bei Zahlenfeldern ist dies der vom System erkannte handschriftlich eingetragene Wert, bei Checkboxen ist jeder Antwortoption ein fester Wert zugeordnet). Werden in Checkboxen oder Zahlenfeldern Daten geändert, so werden alter und neuer Wert durch Komma getrennt ausgegeben und der alte Wert mit einem Stern '*' markiert (Löschmarkierung). Bei Änderungen in Freitextfeldern wird der Text ausgegeben, der über die Verifikations-Eingabemaske eingetragen wurden (d.h., die verifizierende Person ist darauf hinzuweisen, dass bei Änderungen in Freitextfeldern der alte Text, so weit benötigt, nochmals abgeschrieben werden muss).
Die folgenden Beispiele illustrieren die Besonderheiten der XML-Dateien.

Beispiel 1: Erstversion einer CRF-Seite

```
<?xml version="1.0" encoding="windows-1252"?>
<DotForm>
	<P1234_75_4711_3_1>
		<eingabe_t>01</eingabe_t>
		<eingabe_m>09</eingabe_m>
		<eingabe_j>2006</eingabe_j>
		<ber_abschluss>0,2,4</ber_abschluss>
		<ber_txt>Arzt</ber_txt>
		<anz_autage>4</ anz_autage>
		<patnr>372</patnr>
		<zentrum>75</zentrum>
	</P1234_75_4711_3_1>
</DotForm>
```

Beispiel 2: Änderungsversion einer CRF-Seite

```
<?xml version="1.0" encoding="windows-1252"?>
<DotForm>
	<P1234_75_4711_3_2>
		<ber_abschluss>0*,1,2*</ber_abschluss>
		<ber_txt>Zahnarzt</ber_txt>
	</P1234_75_4711_3_2>
</DotForm>
```

Das DotForm-Tag ist für alle Dateien einheitlich. Der zweite (eigentlich erste) Tag-Level enthält die auch im Dateinamen vorhandenen Metadaten zu Studie, Zentrum, Nummer des CRF-Satzes (diese kann auch als Screening-Nummer für den Patienten verwendet werden), Seitennummer und Versionsnummer (1 bei Erstversionen, für Änderungsversionen wird dieser Index hochgezählt). Der unterste Tag-Level enthält die Tags für die einzelnen Items.

3.2 Funktionalität

Das Makropaket deckt mehrere Funktionen ab. Hierzu gehören der Import und die Aufbereitung der Erstversion und aller Änderungsversionen einer für einen gegebenen Patienten ausgefüllten bestimmten CRF-Seite, die Zusammenführung dieser Daten in Seiten-Datasets (ein Dataset für jede Seite im CRF mit je einer Observation pro Patient, die den jeweils neuesten Stand der Änderungen enthält) und die Erzeugung einer Standardstruktur für die Analyse-Datasets. Letztere wird im Folgenden nicht näher beschrieben.

Für den Import und die Aufbereitung der Erstversion einer Seite werden die XML-Dateien mit der XML-Engine von SAS eingelesen. Danach erfolgt die Erzeugung einer Standardstruktur, die alle benötigten Variablen enthält (sowohl zur Speicherung der Rohwerte als auch der daraus abgeleiteten Werte), damit unabhängig von der Anzahl der tatsächlich ausgefüllten Items sichergestellt ist, dass alle auf einer CRF-Seite möglichen Items auch im SAS-Dataset angelegt sind.

Wir haben uns dazu entschieden, die Rohwerte in eigenen Variablen mitzuführen (diese Variablen sind durch das Präfix „R_“ gekennzeichnet), da die XML-Datei wie oben beschrieben nur Textwerte enthält und es bei der Ableitung von Zahlen- und Datumswerten zu Fehlern kommen kann. Aus dem gleichen Grund werden gleichzeitig mit der Ableitung erste Datenvalidierungschecks durchgeführt. Wenn bei der Ableitung aus den Rohwerten Probleme auftreten, die dazu führen, dass die abgeleitete Variable nicht oder nicht korrekt erzeugt wird oder aufgrund der Datenvalidierungschecks andere Fehler entdeckt werden, wird dies in einer RTF-Datei vermerkt, die als Fehler-Log dient.

Die eingelesenen Daten werden – mit je einer Observation pro Patient – an das entsprechende Seiten-Dataset angehängt (beim Import der ersten Observation für ein gegebenes Dataset wird dies neu angelegt).
Zum Schluss wird ein Eintrag in das Import-Log geschrieben. Dies ist ein SAS-Dataset, das für jeden importierten Datensatz eine Angabe zu Patientennummer, CRF-Seite, Seitenversion und Importdatum enthält (siehe Tabelle 1).

Tabelle 1: Beispiel für einen Auszug aus dem Import-Log

Pat.nr.	CRF 1 Version	CRF 1 Datum	CRF 2 Version	CRF 2 Datum	...
4711	1	01/02/2007	1	01/02/2007	...
	.	.	2	03/02/2007	...
	.	.	3	04/02/2007	...
4712	1	17/01/2007	1	24/01/2007	...
4715	1	01/02/2007	.	.	...
	2	02/02/2007	.	.	...
.....					

Bei Datenänderungen werden ebenfalls zuerst die XML-Dateien mit der XML-Engine von SAS importiert. Danach werden zunächst die Rohwerte aktualisiert. Dies ist besonders bei Checkboxen mit Mehrfachnennungen programmtechnisch komplex, da die XML-Datei für ein solches Item eine Liste mit durch Komma getrennten Werten enthält, wobei die Anzahl der Werte variabel ist. Bei Änderungen wird diese Liste neu übermittelt, wobei alle mit einem Sternchen gekennzeichneten Werte gelöscht und alle nicht mit einem Sternchen gekennzeichneten Werte hinzugefügt werden müssen. Da die

Anzahl der Werte in der Liste variieren kann, muss die geänderte Liste jedes mal mit der aus der Vorversion stammenden Liste abgeglichen werden. In Beispiel 1 und Beispiel 2 ist dies anhand der Variable „ber_abschluss“ nachvollziehbar. Hier wurde in der ersten Version eine Werteliste mit „0,2,4“ übermittelt und die erste Änderungsversion lautet „0*,1,2*“. Dies bedeutet, dass die Werte 0 und 2 zu löschen und der Wert 1 zu ergänzen ist, so dass die korrigierte Version der Liste „1,4“ lautet.
Die Ableitung von Zahlen- und Datumswerten aus den Rohwerten sowie die Durchführung der oben beschriebenen Datenvalidierungschecks erfolgen analog zur Vorgehensweise beim Import einer Erstversion.

Zum Schluss wird auch hier ein Eintrag in das Import-Log geschrieben (siehe das Beispiel in Tabelle 1). Zusätzlich wird ein Eintrag in ein separates Tracking-Dataset geschrieben. Das ist ein SAS-Dataset, das für jede Änderung den Namen der Studie, den Name der XML-Datei, die CRF-Seite, die Versionsnummer der CRF-Seite, das Datum der Änderung, die Patienten-Nummer, das jeweilige Item, die Änderung sowie den alten und neuen Rohwert enthält (siehe Tabelle 2).

Tabelle 2: Beispiel für einen Auszug aus der Tracking-Datei

Pat.nr.	Item	Änderung	Rohwert(alt)	Rohwert(neu)	...
4711	WIRKUNG	3		3	...
	GESCHLECHT	1,2*	2	1	...
	BER_TXT	Zahnarzt	Arzt	Zahnarzt	...
	BER_ABSCHLUSS	0*,1,2*	0,2,4	1,4	...
4712	AUS_J	2006		2006	...
	AUS_M	2		2	...
	AUS_T	15		15	...
.....					

Sowohl der Import einzelner Dateien als auch der automatische Import aller in einem Verzeichnis abgelegten XML-Dateien sind möglich. Da das EDC-System eine große Zahl von Einzeldateien erzeugt (so führen z.B. 100 Patienten mit jeweils 40 ausgefüllten CRF-Seiten schon ohne Berücksichtigung von Datenänderungen zu 4.000 XML-Dateien), ist der automatische Import aus einem Verzeichnis als Standard vorgesehen.

4 Diskussion

Der Aufwand für Entwicklung, Test und Validierung der SAS-Makros betrug ungefähr 15 Personentage. Für jede weitere Studie wird dieser Aufwand geringer ausfallen. Eine endgültige Aufwandsabschätzung ist jedoch noch nicht möglich, da der Einsatz des Systems zur Zeit nur in zwei Studien erfolgt.

Eine Anpassung der Makros an die Besonderheiten einer gegebenen Studie ist leicht möglich, da nur an einigen wenigen Stellen im Makro-Code Meta-Informationen über die Details der CRF-Struktur (Anzahl der CRF-Seiten, Bezeichnung der Items, Art der Items (Freitextfeld, Zahlenfeld, Datumsangabe, Checkbox), maximale Länge und / oder spezielle Formatierung von Itemwerten) geändert werden müssen. Die Bezeichnung der Items kann dabei aus einer Metadatei entnommen werden, die von der Firma PharmaForms schon vor dem CRF-Druck und zusammen mit dem annotierten CRF zur Verfügung gestellt wurde. Eine weiter mögliche Entwicklungsstufe wäre die automatische Anpassung der SAS-Makros an die jeweilige CRF-Struktur, in dem die Metadaten zur CRF-Struktur in einem SAS-Dataset gespeichert werden, der von den Makros direkt eingelesen werden kann.

Die große Anzahl der vom System erzeugten Einzeldateien (sowohl XML als auch JPG) führt zu erhöhtem Aufwand für die Prozesse im Datenmanagement. Für den Import und die Aufbereitung erscheint das Problem der vielen Dateien beherrschbar durch die Einhaltung strikter Vorgaben zur strukturierten Dateiablage. Auch hier ergibt sich eine mögliche Weiterentwicklungsstufe für die SAS-Makros, bei der diese strukturierte Ablage stärker automatisiert und damit unterstützt wird.

Die Anwendung der SAS-Makros kann durch Personen mit relativ geringen SAS-Kenntnissen erfolgen, ist aber in jedem Fall durch eine ausreichende Benutzeranleitung zu unterstützen. Zu Beginn sollte eine entsprechende Schulung und Einarbeitung der Anwender erfolgen, die nicht nur den Umgang mit den SAS-Makros umfasst, sondern auch die Regeln für die strukturierte Ablage.

Nach unseren bisherigen Erfahrungen stellt das hier angesprochene Hybridsystem ein für die Prüfzentren relativ niedrigschwelliges Erfassungssystem dar (Gebrauch von Papier und Stift wie gewohnt), und die Daten stehen relativ schnell für das Datenmanagement und das Monitoring zur Verfügung. Als nachteilig würden wir den hohen Aufwand bezeichnen, der sich aus dem Handling der vielen Einzeldateien und der doch relativ komplexen Programmierung ergibt.

Literatur

[1] Cole E, Pisano ED, Clary GJ, Zeng D, Koomen M, Kuzmiak CM, Seo BK, Lee Y, Pavic D. A comparative study of mobile electronic data entry systems for clinical trials data collection. Int J Med Inform. 2006 Oct-Nov;75(10-11):722-9.