Modular automatisierte Datenbereinigung in einer großen Bevölkerungsstudie

André Werner
Stephan Maiwald
Kristin Henselin
Susanne Westphal
Jörg Henke
Dietrich Alte
Henry Völzke
Carsten Oliver Schmidt
Institut für Community Medicine, Abt. SHIP-KEF,
Universitätsmedizin Greifswald KdöR
Walther-Rathenau-Str. 48
17475 Greifswald
awerner@uni-greifswald.de

Zusammenfassung

In großen Bevölkerungsstudien werden in kurzen Zeitabständen große Datenmengen generiert, die verwaltet und qualitätsgesichert werden müssen. Voraussetzung für die Qualitätssicherung ist eine zeitnahe Datenbereinigung und –prüfung. Zur Vollautomatisierung der Datenbereinigung wurden mehrere interagierende SAS Module entwickelt, die eine tägliche Bereitstellung bereinigter Studiendaten ermöglichten. Der personelle Aufwand in der jetzigen Routine des Datenmanagements hat sich deutlich verringert.

Schlüsselwörter: SHIP, ORACLE, Datenmanagement, Automatisierung

1 Einführung

1.1 Hintergrund

Die Study of Health in Pomerania (SHIP) ist eine populationsbasierte Kohortenstudie [1], in welcher sehr große Datenmengen generiert werden. SHIP untersucht Risikofaktoren, subklinische Funktionsstörungen und bevölkerungsrelevante Erkrankungen in Nordostdeutschland [2]. Im Mitte März 2016 abgeschlossenen dritten Follow-Up SHIP-3 wurden ca. 1.700 Probanden untersucht. Je Proband wurden bis zu 6.000 Variablen erhoben. Zur Sicherstellung der Datenqualität mussten diese Daten zeitnah verwaltet sowie geprüft werden. Dies stellte erhebliche Anforderungen an die Infrastruktur des Datenmanagements sowie an die verwendeten Routinen [3]. Die Komplexität der Datenerfassung, -haltung und –aufbereitung sowie die Anforderungen an die Studienlogistik wurden dadurch erhöht, dass die einzelnen Untersuchungen an mehreren räumlich voneinander getrennten Orten und Kliniken in Greifswald stattfanden. Die Datenerfas-

sung erfolgte mit unterschiedlichen Hard- und Softwaretechnologien, so dass für den ETL-Schritt (extract transform load) diverse komplexe Schnittstellen berücksichtigt werden mussten.

1.2 Status Quo

In vorherigen SHIP-Untersuchungswellen wurde jeweils eine quartalsweise Datenaufbereitung mit der SAS Software umgesetzt. Hierfür gab es in der Regel ein SAS Skript je Untersuchungstabelle. Dies bedingte einen hohen Ressourcenaufwand von mehreren Personalwochen je Aufbereitungsintervall. Da hierbei mehrere Medizinische Dokumentare beteiligt waren, war es zudem schwierig, einen einheitlichen Programmierstandard sowie die gegenseitige Vertretbarkeit zu gewährleisten.

1.3 Zielstellung

Dieser hohe Ressourcenaufwand sollte im Rahmen von SHIP-3 deutlich reduziert werden. Die dafür zu entwickelnde Lösung sollte folgende Anforderungen erfüllen:

- Automatisierung & Standardisierung der Datenbereinigung,
- tägliche Bereitstellung bereinigter Studiendaten für die Qualitätssicherung,
- modularer Aufbau für die Adaptierbarkeit auf andere Studien.

Voraussetzung für eine Vollautomatisierung sind Rohdaten, welche im EAV-model (entity-attribute-value), also im „langen“ Datenformat vorliegen sowie Metadaten, die in einem Data Dictionary verwaltet werden.

2 Methodik

2.1 Prozessbeschreibung / Datenworkflow in SHIP-3

Die Datenerfassung in SHIP-3 erfolgte zum größten Teil über Webformulare. Die Daten wurden initial in einer ORACLE Datenbank (zukünftig PostgreSQL) abgelegt und verwaltet. Fremdformate wie z.B. Gerätedaten wurden zur zentralen Datenhaltung über Importroutinen ebenfalls täglich in die ORACLE Datenbank geschrieben. Der gesamte Prozess des Datenworkflows inkl. Datenbereinigung ist in folgender Abbildung dargestellt (siehe Abbildung 1 auf der nächsten Seite).

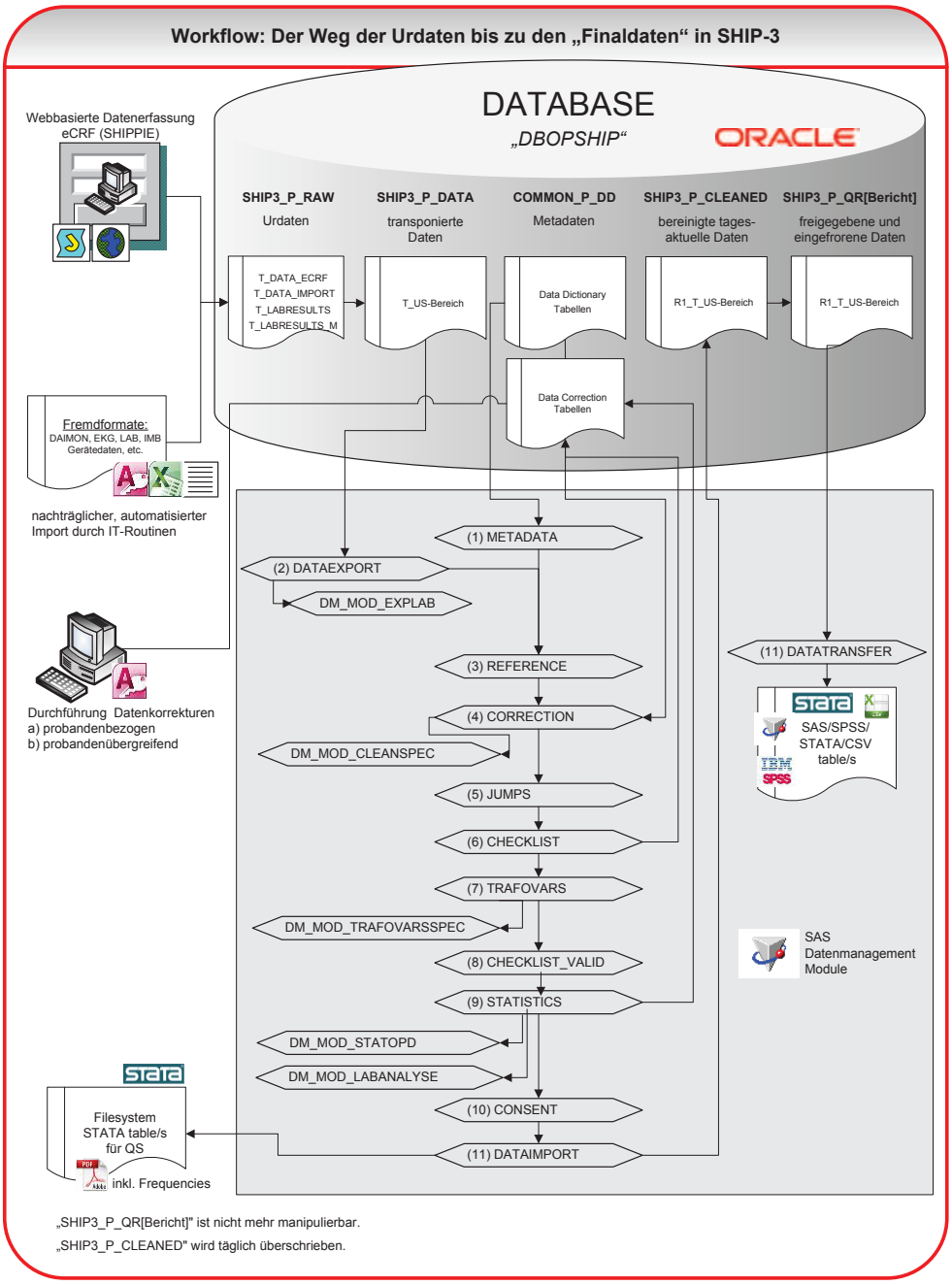

Abbildung 1: Datenworkflow

2.2 SAS Module im Überblick

Da die angestrebte Lösung vollautomatisch aber auch adaptierbar für andere Studien sein sollte, wurde ein modularer Aufbau in Form von SAS Modulen (Makros) angestrebt und umgesetzt. So wurden vorab Anforderungen für die diversen Module schriftlich definiert. Dies war Voraussetzung für die Interagierbarkeit dieser Module. Ein studienspezifisches Steuer-Makro triggert die Ausführung der einzelnen Module. Es werden u.a. folgende Parameter übergeben: Studie, Untersuchungsbereiche, Zeiträume. Dies bedingt eine flexible Nutzbarkeit z. B. für einen oder ausgewählte Untersuchungsbereiche verschiedener Studien. Insgesamt entstanden mehr als zehn SAS-Module (siehe Tabelle 1).

Tabelle 1: SAS Module (Auszug)

SAS Modul	Beschreibung
METADATEN	Bereitstellung von Metadaten für die Datenbeschreibung, wie Variablen- und Formatlabels
TRANSPOSE	Aufteilung "lange Tabelle" in transponierte untersuchungsbezogene Tabellen
DATAEXPORT	Auslesen der Studiendaten aus ORACLE in SAS
REFERENCE	Abgleich mit einer Referenzdatei zur Vollständigkeitskontrolle
JUMPS	Setzen von erlaubten Sprüngen, mit Bezug zum Data Dictionary
CHECKLIST	Detektion und Weiterleitung via Email von unplausiblen oder fehlenden Werten an Untersucher, Qualitätsverantwortliche oder Medizinische Dokumentare
CORRECTION	Auslesen der Korrekturfälle sowie Datenkorrekturen
TRAFOVARS	Berechnung transformierter Variablen für die QS
STATISTICS	Erstellung deskriptiver Statistiken für metrische und kategoriale Variablen
DATAIMPORT	Ablage bereinigter Daten in die ORACLE DB und ins Filesystem zur weiteren Verarbeitung
DATATRANSFER	Übergabe bereinigter sowie qualitätsgesicherter Studiendaten an interessierte Wissenschaftler

2.3 SAS Module im Detail

- **METADATEN**
  Die Metadaten werden zentral in einem eigenen Datenbankschema verwaltet und beinhalten eine komplexe Struktur mit mehr als 30 Data Dictionary (DD) Tabellen. Das SAS Modul zieht relevante Informationen für die Variablen- und Wertebeschreibung aus diesen DD-Tabellen und legt diese zentral auf dem Server ab (METADATENTABELLE).
- **TRANSPOSE**
  Rohdaten werden im EAV-model (entity-attribute-value), also im „langen" Datenformat erhoben und müssen transponiert sowie in untersuchungsbezogene Tabellen aufgeteilt werden. Diese hier entstehenden Tabellen werden in einem separaten DB-Schema abgelegt.
- **DATAEXPORT**
  Hier werden die transponierten Tabellen mit Hilfe des SAS Moduls aus der ORACLE DB in SAS überführt. Es erfolgt weiterhin das dynamische Labeln von Variablen- und Werteausprägungen anhand der METADATENTABELLE.
- **REFERENCE**
  Dieses Modul sorgt für den Abgleich der Ergebnisdaten mit der Referenzdatei. Hier erfolgt ein Check, das Vorliegen von Ergebnisdaten mit den gegebenen Einverständnissen korrespondiert. Diskrepanzen gehen anschließend in die Prüfschleife.

- **CHECKLIST**
  Hier erfolgen die Detektion von unplausiblen oder fehlenden Werten und die direkte Weiterleitung per Email an Untersucher, Qualitätsverantwortliche oder Medizinische Dokumentare. Die Prüffälle werden parallel in entsprechenden data correction Tabellen in die ORACLE DB geschrieben (Backend). Mit Hilfe eines MS ACCESS Frontends können die Prüfungen bearbeitet sowie Datenkorrekturen vorgenommen werden. Hierbei wird zwischen probandenbezogenen und probandenübergreifenden Korrekturen differenziert.
- **CORRECTION**
  In diesem Modul erfolgt das Auslesen der Korrekturfälle aus den data correction Tabellen (CHECKLIST). Diese Datenkorrekturen werden durch datengetriebene dynamische SQL Update-Statements umgesetzt.
- **JUMPS**
  Dieses Modul ist zuständig für das Setzen erlaubter Sprünge anhand der Informationen aus der METADATENTABELLE. Fällt hierbei auf, dass ein erlaubter Sprung in eine Variable gesetzt werden soll, in der sich jedoch bereits ein Wert befindet, wird dies als Prüffall markiert und die Prüfschleife geschickt.
- **TRAVOVARS**
  Oftmals müssen die erhobenen Daten erst transformiert werden, um für statistische Berechnungen oder zu Qualitätssicherungszwecken verwendbar zu sein. Mit diesem Modul erfolgt die Berechnung solcher transformierter Variablen für die Qualitätssicherung und wissenschaftliche Analysen, z.B. BMI nach WHO aus Größe und Gewicht, metabolisches Syndrom (Blutdruck, Laborparameter, Diabetes etc.).
- **STATISTICS**
  Das Modul erzeugt deskriptive Statistiken für metrische sowie Häufigkeitsübersichten für kategoriale Variablen. Die Reports werden täglich erzeugt und im Filesystem als PDF-File abgelegt. Die Qualitätsverantwortlichen können die Reports u.a. für einen initialen Datencheck nutzen.
- **DATAIMPORT**
  Im letzten Schritt des Datenbereinigungsprozederes werden die aufbereiteten Studiendaten in einem separatem DB-Schema abgelegt. Dieses beinhaltet die finale Datenhaltung (persistance layer) und ist Grundlage für Datentransfers. Parallel werden die aufbereiteten Daten im SAS & STATA File im Filesystem für die interne Qualitätssicherung abgelegt.
- **DATATRANSFER**
  Dieser Baustein ist verantwortlich für die Übergabe bereinigter sowie qualitätsgesicherter Studiendaten an Wissenschaftler. Die Übergabe erfolgt z.B. in den Formaten SAS, STATA, SPSS und CSV. Voraussetzung für die Datenübergabe ist ein durch den Vorstand des Forschungsverbundes Community Medicine – Gutachtergremium genehmigter Antrag auf SHIP Datennutzung.

3 Ergebnisse

3.1 Leistungsfähigkeit & Effizienz

Die Datenbereinigung in SHIP-3 lief täglich vollautomatisch im Hintergrund (Laufzeit ca. 3 Stunden). Somit konnte täglich auf aufbereitete Studiendaten zugegriffen werden, was der Datenqualität zu Gute kommt. So konnte die Qualitätssicherung zeitnah intervenieren, sofern sich Auffälligkeiten in den Daten zeigten. Weiterhin wurden Prüffälle zeitnah kontrolliert und korrigiert.

3.2 Flexibilität & Adaption für andere Studien

Eine erste Implementierung in einer kleinen Partnerstudie verlief positiv. Derzeit erfolgt die Implementierung in einer weiteren großen Kohorte (SHIP-TREND-1).

3.3 Aufwand, Kosten

Die Entwicklung der SAS-Module verursachte einen Programmieraufwand von mehreren Personen-Monaten und erforderte eine hohe und regelmäßige Kommunikationsbereitschaft zwischen Programmiererinnen und Projektleitung. Die jetzt voll automatisierte Datenbereinigung benötigt dafür nur noch geringe personelle Ressourcen. Zuvor waren es mehrere Personalwochen je Aufbereitungsintervall. Langfristig ist durch die Automatisierung eine deutliche Aufwandsreduktion für die Datenbereinigung zu verzeichnen.

4 Schlussfolgerungen

Das modular entwickelte Datenbereinigungsprozedere konnte im Zusammenspiel von ORACLE und SAS wie geplant umgesetzt und in die Routine überführt werden. Durch den modularen Aufbau ist eine Implementierung in anderen Studien mit überschaubarem Aufwand möglich. Die wesentliche Voraussetzung dafür ist das Vorhandensein einer ähnlichen Datenstruktur wie in SHIP oder eine entsprechende Schnittstellendefinition. Weitere Optimierungen sind derzeit in Arbeit.

Literatur

[1] Lüdemann J. (2000): Methoden zur Qualitätssicherung im medizinischen Untersuchungsbereich epidemiologischer Feldstudien: Die „Study of Health in Pomerania“ (SHIP). Das Gesundheitswesen, 2000 Apr;62(4):234-43.

[2] Völzke et al. (2010): Cohort Profile: The Study of Health in Pomerania (SHIP). International Journal of Epidemiology, 2011 Apr;40(2):294-307.

[3] Nonnemacher M., Nasseh D., Stausberg J (2014): Datenqualität in der medizinischen Forschung: Leitlinie zum Adaptiven Datenmanagement in Kohortenstudien und Registern. TMF, Berlin. 2014 (ISBN: 3954661217)