

SAS macht Public-Use-Files public - Webbasierte Beantragung der Datennutzung und Datenübergabe per SAS-Macro im Greifswalder Forschungsverbund Community Medicine

Dietrich Alte
Institut für Community Medicine
Walther-Rathenau-Str. 48
Greifswald
Alte@uni-greifswald.de

André Werner
Institut für Community Medicine
Walther-Rathenau-Str. 48
Greifswald
Andre.Werner@uni-greifswald.de

Wolfgang Hoffmann
Institut für Community Medicine
Ellernholzstr. 1-2
Greifswald
Wolfgang.Hoffmann@uni-greifswald.de

Zusammenfassung

Im Rahmen des Forschungsverbundes Community Medicine der Uni Greifswald werden Daten der Study of Health in Pomerania (SHIP) an Datennutzer mit einem SAS-Macro übergeben. Wir beschreiben den Hintergrund, das Konzept und die Umsetzung des Macros, sowie die dadurch erreichten Vorteile, insbesondere der automatisierten Erstellung eines Protokolls.

Schlüsselwörter: Public Use, SAS-Makro, Data Dictionary, automatisches Protokoll

1 Einleitung / Hintergrund

1.1 Public Use Daten

In öffentlich geförderten Forschungsprojekten müssen die Studiendaten in der Regel nach einer gewissen Zeit der Fachöffentlichkeit in Form von Public-Use-Files zur Verfügung gestellt werden. Die Datennutzung muss dann von externen Datennutzern beantragt und ein Kooperationsvertrag mit der die Studie primär ausführenden Institution geschlossen werden. Die Beantragung der Daten, sowie deren Übergabe und Protokollierung kann bei größeren Projekten oder Forschungsverbänden und großer Nachfrage viel Zeit in Anspruch nehmen und erfordert eine angemessene Infrastruktur, um die Anfragen zeitnah bearbeiten zu können.

1.2 Forschungsverbund Community Medicine in Greifswald

Der Forschungsverbund Community Medicine (FVCM) an der Universität Greifswald [1] ist die zentrale Forschungsplattform der medizinischen Fakultät und führt mehrere Großprojekte durch wie die Study of Health in Pomerania (SHIP), und die Study of Neonates in Pomerania (SNiP) mit jeweils mehreren 1000 Probanden. Die Daten sind – da öffentlich gefördert – im Prinzip Public Use Daten und die Datennutzung ist zentral geregelt.

1.3 Problemstellung

Es gibt ein Antragsverfahren für interne/externe Nutzer und über die Anträge wird auf monatlichen Vorstandssitzungen (V-FVCM) zeitnah eine Entscheidung getroffen. Die Datenübergabe (jeweils nicht der ganze Datensatz, sondern nur die genehmigten Variablen) werden zeitnah übergeben. Dabei wird ein Protokoll geschrieben. Die Menge der Anträge und die Komplexität der Daten machte eine Teilautomatisierung nötig. Dazu wurden einige SAS-Macros entwickelt.

2 Material und Methoden

Wichtige Säulen des FVCM sind zwei große prospektive, epidemiologische Studien, die Study of Health in Pomerania (SHIP; [2]) und die Study of Neonates in Pomerania (SNiP; [3]), in die jeweils mehrere 1000 Probanden eingeschlossen wurden.

2.1 Study of Health in Pomerania (SHIP)

SHIP ist eine epidemiologische Längsschnittstudie in Vorpommern (MV, Abb. 1) mit n=4310 in der ersten (SHIP-0, 1997-2001), n=3300 in der zweiten (SHIP-1: 2002-2006) und geplanten n=2500-2700 Probanden in der dritten Erhebungswelle (SHIP-2: 2008-2012). Das Untersuchungsspektrum ist sehr breit und enthält u.a. medizinische Untersuchungen, Laboranalysen, zahnmedizinische Untersuchungen und ein Interview.

2.2 Umfang und Komplexität der Daten

Durch den längsschnittlichen Aufbau der Studie (SHIP-0/1/2 ...) und das Vorliegen mehrerer Untersuchungsbereiche sind auch mehrere Qualitätssicherungs-Verantwortliche an der Datenaufbereitung beteiligt und legen die Daten in Tabellen ab, die insgesamt >5000 Variablen haben. Ein Data Dictionary (DD) wird in MS-Access gepflegt und hat folgende Hierarchie (je 1 Tabelle mit Schlüssel):

1. Untersuchungsbereich (Bsp: Labor, Zahnmed. Unters., ...)
2. Gruppe/Bereich (Bsp: Blut, Urin, ...)
3. Untergruppe/Frage (Bsp: Blutbild, Gerinnung, ...)
4. Variable (Bsp: Glukose, MCV, GGT, ...)
5. Codierung/Labels (Bsp: 0=nein, 1=ja, ...)

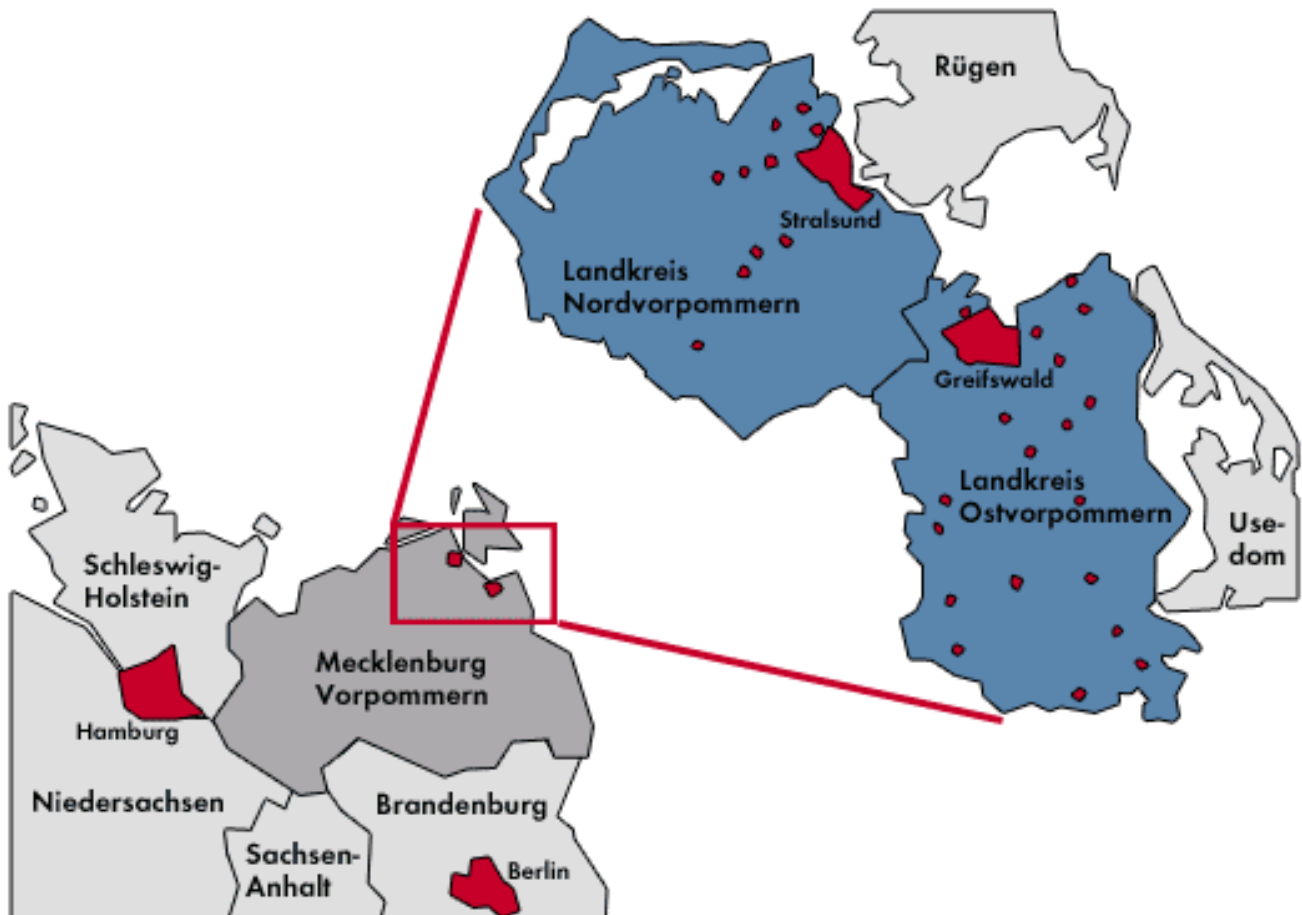


Abbildung 1: SHIP Studienregion

2.3 Traditionelle Datenübergabe

Vor der Überarbeitung des Prozesses machten Antragsteller unsystematische Angaben zu den gewünschten Daten und ein langwieriges Zusammenstellen der Variablenlisten war erforderlich, um mit einem SAS-Skript zur Datenauswahl die entsprechende Datei bereitzustellen. Die Protokollierung erfolgte von Hand (Word-Dokument) und die Gesamtdauer einer Datenübergabe betrug oft mehrere Stunden. Die Anzahl der Anträge stieg mit der Zeit auf >50/Jahr und somit wurde eine weitgehende Automatisierung inkl. automatischer Protokollierung (Word-Dokument) angestrebt.

2.4 Konzept: Datenbeantragung per Webformular

Der Lösungsansatz sah die Beantragung mit einem Webformular und einer dann folgenden Datenauswahl mit einem SAS-Macro vor. Die Datennutzung (sowohl fakultäts- bzw. universitätsintern als auch –extern) ist zentral geregelt und wird über eine Transferstelle abgewickelt [4]. Jede Datennutzung muss beim FVCM beantragt werden und wird dann folgendermaßen prozessiert:

- Antrag + Datenauswahl per Webformular (HTML / PHP / MySQL)
- Speicherung in MySQL
- Übertragung des Antrags aus MySQL in MS-Access

- Erstellung eines Word-Serienbrief (Daten aus MDB) und PDF-Umwandlung
- Versand des Antrags per Email an FVCM-Vorstand
- Diskussion und Entscheidung im FVCM
- Information über Genehmigung und Vertrag
- Datenübergabe an Datennutzer

Jährlich werden mehrere Dutzend Datennutzungsanträge per Webformular [5] (s. Abb. 2) gestellt. Dabei müssen Angaben zum Inhalt (Fragestellung, Hypothesen etc.) und Umfang (Art der Daten, Variablen) gemacht werden. Nach Entscheidung durch den FVCM-Vorstand und Abschluss des Nutzungsvertrages werden die Daten per SAS-Macro übergeben. Dieses übernimmt die per Webformular aus einem Data Dictionary ausgewählten Variablennamen und wählt aus den aktuellen, in einer Versionierungsdatei verwalteten SAS-Dateien die entsprechenden Daten aus. Ein standardisiertes Protokoll wird in Form einer MS-Word-Datei generiert, welches die übergebenen Daten beschreibt. Aktuell können Daten als SAS- oder SPSS-Dateien übergeben werden. Der Transfer in SPSS erfolgt über ein SAS Macro, welches SPSS startet und die Daten in SPSS einliest, bearbeitet und speichert.

Ernst-Moritz-Arndt-Universität Greifswald
Transferstelle für Daten- und Biomaterialienmanagement

Medizinische Fakultät Klinikum Lageplan English

Transferstelle für Daten- und Biomaterialienmanagement

Schritt 2 - Auswahl von Daten aus SHIP-1

Bitte klicken Sie auf die Untersuchungsteile und wählen Sie ihre Variablen aus...

- Archiv
- Zahnmedizinische Untersuchung
- Zahnmedizinisches Interview
- Persönliches Interview
- Blut- und Urinuntersuchungen
- Medizinische Untersuchungen
 - Medizinische Untersuchungen ohne Ultraschall
 - Allgemeine Information
 - alle keine
 - Blutdruck und Herzfrequenz
 - Blutdruck: Armumfang
 - Blutdruck: Beginn der Messung
 - Blutdruck: Besonderheiten
 - Blutdruck: Ende der Messung
 - Blutdruck: Geräte-ID
 - Blutdruck: Herzfrequenz 1
 - Blutdruck: Herzfrequenz 2
 - Blutdruck: Herzfrequenz 3
 - Blutdruck: Manschettengröße
 - Blutdruck: Welche Besonderheiten
 - Blutdruck: Diastolischer Blutdruck 1
 - Blutdruck: Diastolischer Blutdruck 2
 - Blutdruck: Diastolischer Blutdruck 3

Abbildung 2: Webformular - Variablenauswahl

3 Ergebnisse

Der Entwicklungsprozess erforderte eine Erweiterung der IT-Struktur (Abb. 3). Insbesondere musste das DD vervollständigt und eine Versionskontrolle für die SAS-Tabellen eingeführt werden.

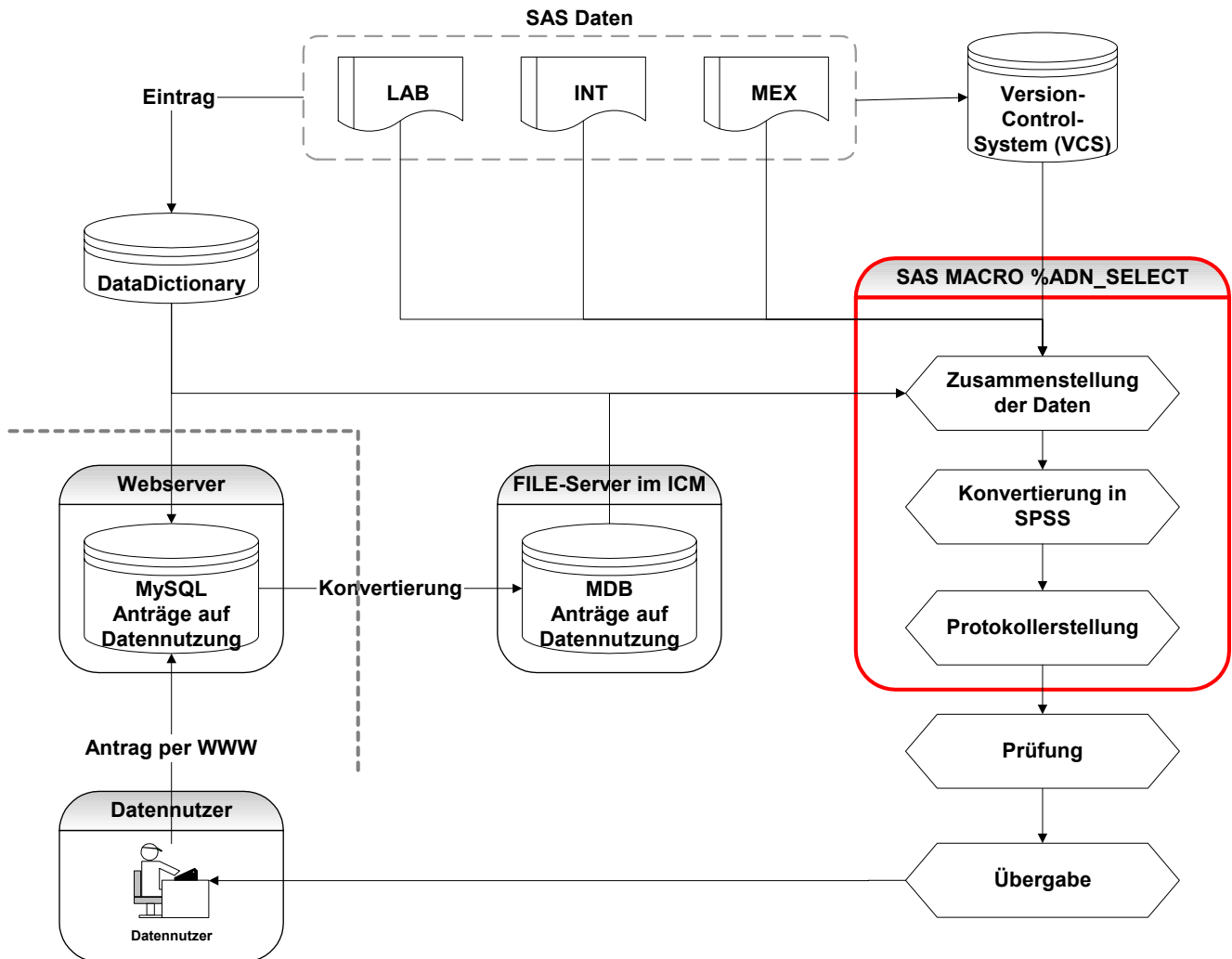


Abbildung 3: Ausschnitt der IT-Struktur

3.1 Haupt-Macro

Es wurde ein Hauptmacro entwickelt, mit folgendem Header und Optionen:

```
%MACRO ADN_SELECT (
/* ----- Antragsbezogene Optionen ----- */
Antragsnummer      =, /* Geschäftszeichen zum Zugriff auf MDB der
                        Anträge, zB SHIP/2007/15 (mit M/D) */
ExportPfad         =, /* Pfad auf Server zur Ausgabe der Dateien*/
Format             =, /* SAS oder SPSS */
```

```
BASISDATA      =1, /* Basisdaten (age, sex...) hinzufügen(1=ja)*/
Studie        =0 1, /* Aus welcher SHIP Studie?: SHIP-0+SHIP-1 */
DROP_EX_SHORT  =, /* Ausschluss von Untersuchungs-Teilen */
protokoll      =1, /* 1: Protokoll, 3: nur Prot., keine Daten,
                    sonst: kein Protokoll, nur Daten */

/* ----- Allgemeine Optionen -----*/
VCS            =, /* Dateienliste / Versions-Control-System*/
ADN_DB         =, /* Pfad für Anträge auf Datennutzung */
dd0            =, /* Pfad + MDB des DataDictionary SHIP-0 */
dd1            =, /* Pfad + MDB des DataDictionary SHIP-1 */
transfer_MDB   =, /* Pfad+Datei: Transfer-MDB */
transfer_TAB   = /* Tabelle in MDB für Missingtransfer */)
```

3.2 Macro-Struktur

Das Makro hat knapp 500 Zeilen und enthält folgende Bausteine:

- Initialisierung: Libnames, Pfade etc.
- Einlesen der Antragsdaten
 - von Projekt/Person-Daten aus der MDB (s. Abb. 4)
 - der beantragten Variablen
- Prüfung / Zuordnung
 - Variablen zuordnen zu DD und zu Dateien
 - Anzeigen der nicht zugeordneten Variablen
- Extraktion der Daten
 - Erstelle Gesamtdatei der ausgewählten Untersuchungs-Teile
 - keep „nur ausgewählte Variablen“
 - Zusammenfassen der Formatdateien
- ggf. Transfer in SPSS
- Erstellung des Protokolls als Word-Datei.

3.3 Hilfsmacros

Zur Unterstützung des Hauptmacros wurden einige Hilfsmacros geschrieben bzw. vorhandene genutzt mit folgenden Funktionen:

- %ADN_CHECK:** Übernahme / Prüfung der beantragten Variablen
- %GETFMTS:** Extrahiert benötigte Formate aus SAS-Formate-Datei
- %SASToSPSS:** Transferiert SAS-Daten in SPSS-Datei:
 - Umkodierung von SAS-Missings in Zahlen
 - SAS schreibt SPSS Skript (*.sps) zum Import der Daten
 - SAS schreibt SPSS Produktionsjob (*.spp)
 - SAS startet SPSS via spp (darin Aufruf der *.sps Datei):

3.4 Automatisches Protokoll

Da die Protokollierung einen beträchtlichen Anteil der Zeit für die Datenübergabe erforderte, wurde in das Hauptmacro eine vollautomatische Protokollierung eingebaut. Folgende Technik wird eingesetzt: SAS "spricht" mit MS-Word via DDE und WordBasic. Es werden dann vorbereitete Word-Dateien und per DDE übertragenen Textbausteinen zusammengefügt:

MS-WORD:	Standard-Titelseite + Infoblatt sind vordefiniert
Proc Contents:	Angaben der Variablenlisten als RTF per ODS
Data _null_:	Öffnet Titelseite (vorher vordefiniert als *.doc)
Data _null_:	Eintrag Datennutzer + Geschäftszeichen etc.
Data _null_:	Anhängen: Infoblatt (vordefiniert *.doc)
Data _null_:	Anhängen: Variablenliste (*.rtf)

Die Kommunikation mit MS-Word gestaltet sich mit der (wenn auch etwas veralteten Technik) DDE und WordBasic relativ einfach und sieht für die Übernahme der Antragsdaten ins Titelblatt etwa so aus:

```
filename sas2word dde 'winword|system';
data _null_;
  file sas2word;
  put '[FileCloseAll(1)]'; * offene Dateien speichern/schließen;
  * Öffne Vorlage mit Layout;
  put '[FileOpen.Name="\Pfad\Vorlage_Protokoll.DOC]';
  set _projekt_info end=done;
  if _n_=1 then put '[EndOfDocument]';
  put Label ":"; put '[Insert Chr$(9)]' Inhalt;
  put '[InsertPara]' '[InsertPara]';
  if done then do;
    put 'Übergabedatum: '; put '[InsertPara]' '[InsertPara]';
    put '_____';
    put 'SHIP / Verantwortlicher für Datenübermittlung: ';
    put "_____";
    put 'Projekt / Empfänger: ';
    put "_____";
    put "Anlagen:" '[InsertPara]';
    put "- Hinweise" '[InsertPara]';
    put "- Übersicht der Quell-Dateien ..." '[InsertPara]';
  end;
run;
```

3.5 Resultat

Vor der Fertigstellung der Web-basierten Beantragung und Übergabe von Daten nahm die Bearbeitung eines einzelnen Antrags mehrere Stunden in Anspruch, u.a. weil der

Umfang der Daten oft nicht hinreichend beschrieben wurde und dies im Detail bei den Antragstellern nachgefragt werden musste. Durch die zwingende Datenauswahl per Webformular und den Einsatz mehrerer SAS-Macros zur Datenübergabe ab April 2007 dauert die Übergabe jetzt normalerweise nicht mehr als ein halbe Stunde und ist weitgehend standardisiert. Die automatische Protokollerstellung sorgt für vollständige und wohl strukturierte Übergabeprotokolle. Aktuell werden nur die Daten aus SHIP mit diesen Macros übergeben.

4 Diskussion / Schlussfolgerungen

Durch systematische Nutzung von Metadaten der Public-Use-Files (Data Dictionary, Dateiversionsverwaltung) kann eine steigende Anzahl von Datennutzungsanträgen in kurzer Zeit vom Antrag bis zur Datenübergabe bearbeitet werden. Selbst geschriebene SAS-Macros, die Schnittstellen zum Betriebssystem und zu Office-Produkten flexibel nutzen, unterstützen diese Arbeit. Voraussetzungen dafür waren die Erstellung eines DD, einer Versionskontrolle (VCS) und eines Webformulars. Geholfen haben die SAS Flexibilität allgemein und besonders beim Datenaustausch mit allen Datenformaten. Der Programmieraufwand für die Macros beträgt etwa 100h.

Das Konzept kann für die Übergabe der Daten weiterer Studien oder Forschungsverbände erweitert werden. Die Macros können bei Interesse beim Erstautor angefordert werden.

5 Ausblick

Die Einbindung weiterer Daten (z.B. Medikamente - mehrere je Proband) und die Einbindung weiterer Studien sind für die nächste Zeit geplant. Das Webformular wird in einer neuen Version mit verbesserter Funktion umgesetzt. Weiterhin müssen noch Werkzeuge zum Update des DD ins Web entwickelt werden. Die Steuerdaten (DD+VCS) sollen von MS-Access auf eine Oracle-DB umziehen. Beim Datenexport ist eine Erweiterung auf andere Datenformate (stata, R, etc.) geplant.

Literatur

- [1] www.community-medicine.de
- [2] ship.community-medicine.de/
- [3] www.uni-greifswald.de/~neo_cm/
- [4] www.medizin.uni-greifswald.de/icm/transferstelle/
- [5] www.medizin.uni-greifswald.de/icm/transferstelle/dd_service/data_use_intro.php