Aufarbeitung von überlappenden zu eindeutig abgegrenzten Zeitintervallen unter Beibehaltung der in Primärdaten enthaltenen Informationen

Thomas G. Grobe
Institut für Sozialmedizin, Epidemiologie
und Gesundheitssystemforschung
ISEG
30159 Hannover, Lavesstr. 80
grobe@iseg.org

Zusammenfassung

Für viele epidemiologische Auswertungen ist eine personenbezogene Bestimmung der Bezugszeiträume erforderlich. Will man Daten von Krankenkassen für epidemiologische Auswertungen nutzen, ist man unter Umständen mit einer Erfassung von Versicherungsverhältnissen in den Primärdaten der Krankenkassen konfrontiert, die regulär personenbezogen sowohl zeitliche Diskontinuitäten der Versicherungsintervalle als auch personenbezogen multiple überlappend erfasste Zeitintervalle beinhalten kann (z.B. bei gleichzeitiger Beschäftigung von Versicherten bei mehreren Arbeitgebern). Eine Summation der ausgewiesenen Versicherungszeiten führt beim Vorliegen mehrerer Versicherungsverhältnisse zu einem Zeitpunkt nicht zur korrekten Bestimmung der Bezugszeiten. Gleichfalls ist eine eindeutige Zuordnung von Ereignissen oder Ereignisintervallen zu den ggf. mehrfach für einen Zeitpunkt ausgewiesenen Bezugsintervallen nicht möglich.

Im Rahmen des Beitrags wird ein SAS-Makro vorgestellt, welches die Aufarbeitung von Daten zu Zeitintervallen mit beliebigen zeitlichen Überlappungen erlaubt. Es resultieren nach Programmausführung Datensätze, die personenbezogen eindeutig zeitlich abgegrenzt sind, wobei optional alle Informationen der Originaldaten erhalten bleiben. Zusätzlich lassen sich Informationen bei einem Fehlen in aktuellen Beobachtungen aus zeitlich vorausgehenden Beobachtungen fortschreiben. Als wesentliches Element wurden bei der Umsetzung in der SAS-Syntax Lag-Funktionen genutzt, die über Programmschleifen innerhalb des Makros modifiziert werden. Das vorgestellte SAS-Makro wurde erfolgreich bei der Aufarbeitung großer Datensätze mit mehreren Millionen Beobachtungen eingesetzt.

Keywords: Datenaufarbeitung, Zeitintervalle, SAS-Makro.

1 Hintergrund

Für viele epidemiologische Auswertungen ist eine Bestimmung der Bezugszeiträume erforderlich. Will man Daten von Krankenkassen für entsprechende Auswertungen nutzen, ist man u.U. mit einer Dokumentation von

Versicherungsverhältnissen konfrontiert, die regulär personenbezogen zeitliche Diskontinuitäten der Versicherung, aber auch multiple überlappende Zeitintervalle aufweisen kann (z.B. durch eine zwischenzeitliche Beschäftigung bei mehreren Arbeitgebern). Eine Summation der ausgewiesenen Versicherungszeiten führt beim Vorliegen mehrerer Versicherungsverhältnisse zu einem Zeitpunkt nicht zur korrekten Bestimmung der Bezugszeiten, gleichfalls ist eine eindeutige Zuordnung von Ereignissen zu Bezugszeiten nicht möglich.

2 Aufarbeitungsziele

Im Rahmen des Beitrags wird ein SAS-Makro vorgestellt, welches die Aufarbeitung von (in einzelnen Beobachtungen mit Von- und Bis-Datum erfassten) Zeitintervallen bei beliebiger zeitlicher Überlappung zu Beobachtungen erlaubt, die personenbezogen eindeutig zeitlich abgegrenzt sind. Dabei können optional alle Informationen der Originaldaten erhalten bleiben. Exemplarisch sind in den nachfolgenden beiden Tabellen Dokumentationen von Zeitintervallen bei einem Individuum vor und nach einer Aufarbeitung dargestellt. Das beschriebene Makro wurde erfolgreich auch bei der Aufarbeitung großer Datensätze mit mehreren Millionen Beobachtungen eingesetzt (1).

Tabelle 1: Beispieldaten vor Aufarbeitung

ID	Von_Datum	Bis_Datum	Arbeitgeber	monatl. Einkommen
007	01.01.1991	31.05.1991	A	1200
007	01.08.1991	31.08.1991	B	800
007	01.09.1991	31.12.1995	C	1000
007	01.01.1994	31.12.1998	D	200
007	01.06.1995	31.12.1999	E	800

Tabelle 2: Beispieldaten nach Aufarbeitung

ID	Von_Datum	Bis_Datum	Arbeitgeber	monatl. Einkommen
007	01.01.1991	31.05.1991	A	1200
007	01.06.1991	31.07.1991		.
007	01.08.1991	31.08.1991	B	800
007	01.09.1991	31.12.1993	C	1000
007	01.01.1994	31.05.1995	D,C	1200
007	01.06.1995	31.12.1995	E,D,C	2000
007	01.01.1996	31.12.1998	E,D	1000
007	01.01.1999	31.12.1999	E	800

3 Programmschritte

Die Aufarbeitung erfolgt im Wesentlichen in vier Schritten.

Teil I.: Zunächst werden die Daten nach der Anzahl der Versicherten-bezogen ausgewiesenen Zeitintervalle aufgeteilt. Bei den de facto aufgearbeiteten Kassendaten und somit auch im hier dargestellten Programm wurde dabei eine Einteilung in vier Gruppen vorgenommen (Gruppe mit personenbezogen genau einem Intervall, Gruppen mit 2-10, 11-100 sowie mit 101 bis maximal 1000 dokumentierten Intervallen). Ziel dieser Aufteilung ist es, die Zahl der erforderlichen Durchläufe von Programmschleifen für einen überwiegenden Teil der Beobachtungen möglichst gering zu halten. Die Zahl der Durchläufe richtet sich dabei nach der maximal in einer Gruppe für einzelne Versicherte erfassten Anzahl von Beobachtungen. Die Daten zu drei der vier Gruppen mit personenbezogen mehr als einer Beobachtung werden nachfolgend unter Verwendung desselben Makros getrennt aufgearbeitet und erst zum Abschluss des Programms wieder zu einer Datei zusammengefügt. Daten zu Personen mit lediglich einem dokumentierten Zeitintervall bedürfen selbstredend keiner spezifischen Aufarbeitung und können daher weitgehend unverändert übernommen werden.

Teil II.: Sind personenbezogen mehrere Zeitintervalle erfasst, werden im zweiten Programmschritt Von- und Bis-Datumsgrenzen zu allen Zeiträumen ermittelt, die einen abgrenzbaren Zustand aufweisen. Bei Versicherungslücken resultieren dabei gänzlich neu beschriebene Zeitintervalle. Aus zwei partiell überlappenden Zeitintervallen resultieren bei dem Vorgehen beispielsweise drei diskrete Zeiträume. Eine im Programmablauf generierte vollständige Datei zu Zeitintervallen enthält in den einzelnen Beobachtungen neben der ein-eindeutigen Identifikationsnummer des Versicherten Angaben zu allen abgrenzbaren Zeitintervallen mit einem Von- und Bis-Datum. Zwangsläufig enthalten sind in den Datensätzen alle Von-Datumsangaben sowie alle Bis-Datumsangaben der Originaldaten. Hinzu kommen ergänzte Datumsangaben. Die Datei ist nach ID und Von-Datum sortiert, das Bis-Datum erhält in dieser Arbeitsdatei eine vom Original abweichende Bezeichnung.

Teil III.: Die Arbeitsdatei mit den Zeitintervallen lässt sich nun mit der nach ID, Von- und absteigend nach Bis-Datum sortierten Originaldatei nach den Merkmalen ID und Von-Datum zusammenfügen bzw. mergen (zu jedem Von-Datum der Originaldatei existiert eine Beobachtung in der Arbeitsdatei mit Zeitintervallen, zu jedem Von-Datum der Arbeitsdatei existieren keine, eine oder mehrere Beobachtungen in der Original-Datei). Durch die gewählte Sortierung stehen in der resultierenden Datei bzw. Tabelle alle relevanten Originaldaten oberhalb oder in der Zeile der Beobachtung, die das kleinste,

diskret abgrenzbare, Zeitintervall darstellt, welches zuvor in der Arbeitsdatei mit den Zeitintervallen eindeutig ausgewiesen wurde.

Teil IV.: In einem vierten Schritt werden unter Verwendung von Lag-Funktionen Informationen aus den in der Datentabelle vorausgehend stehenden Beobachtungen in die jeweils aktuell bearbeitete Beobachtung übertragen, sofern sie für das Zeitintervall relevant sind (vorausgehende Beobachtung mit gleicher ID und Bis-Datum >= aktuelles Von-Datum). Grundsätzlich sind dabei unterschiedliche Arten der Aggregation von Daten bei zeitlich überlappend vorliegenden Informationen möglich (einfache Verkettungen alphanumerischer Variablen, bedingte Verkettung, falls Information nicht bereits vorhanden, bei nur begrenzt vorkommenden Überlappungen die Bildung neuer Variablen....). Bei Bedarf lassen sich Informationen, die in der aktuellen Beobachtung fehlen – sofern inhaltlich gerechtfertigt – auch aus vorausgehenden Beobachtungen ergänzen (z.B. Ergänzung der Merkmale Geburtstag und Geschlecht für versicherungsfreie Intervalle).

Nach der eigentlichen Aufarbeitung werden die separat aufgearbeiteten Daten der vier Gruppen schließlich wieder zu einer Datei zusammengefügt. Details sowie verwendete Syntax der Aufarbeitung sind dem auf der Tagungs-CD enthaltenen SAS-Programm zu entnehmen. Das Programm kann bei Bedarf auch vom Autor per Mail direkt angefordert werden.

4 Resümee

Im Rahmen der Analyse größerer Datenbestände entfällt auf die auswertungsorientierte Aufbereitung der Daten oftmals ein erheblicher Anteil des Gesamtarbeitsaufwandes. Die SAS-Programmsyntax bietet für entsprechende Aufarbeitungen vielfältige Möglichkeiten, von denen einige wenige im Rahmen des vorgestellten Programms aufgezeigt werden konnten. Der Autor dieser Zeilen hofft, dem Leser damit Anregungen für eigene Programmierungen bei der Datenaufbereitung geben zu können, die im Arbeitsalltag eine nicht zu vernachlässigende Rolle spielt.

Literatur

1. Grobe, T.G., Dörning, H., Schwartz, F.W. (2002). GEK-Gesundheitsreport 2002, Asgard-Verlag, Hippe. ISBN 3-537-44022-7