Zuordnung von Ereignisintervallen zu Bezugszeiträumen in größeren Datenbeständen

Thomas G. Grobe
ISEG
Lavesstr. 80
30159 Hannover
grobe@iseg.org

Zusammenfassung

Informationen zu Bezugszeiträumen (z.B. personenbezogen dokumentierte Berufstätigkeitsintervalle) sowie zu Ereignisintervallen (z.B. individuelle Krankschreibungsintervalle) werden üblicherweise primär in unterschiedlichen Datentabellen erfasst. Sollen Ereignishäufigkeiten oder Ereigniszeiten innerhalb von spezifischen Bezugszeiträumen ermittelt werden (z.B. Krankschreibungszeiten in bestimmten Berufsgruppen) ist eine Zuordnung von Informationen aus beiden Tabellen erforderlich. Diese Zuordnung kann insofern komplex sein, als dass es sich bei der Beziehung zwischen beiden Tabellen in der Regel um eine m:n Beziehung handelt (z.B. können mehrere Krankschreibungsintervalle in ein Berufstätigkeitsintervall fallen, es kann sich jedoch auch eine Krankschreibung über mehrere separat dokumentierte Tätigkeitsintervalle erstrecken). Zudem stehen primär keine fixen Werte zur Verknüpfung der Beobachtungen bereit, die Zuordnung setzt eine Überprüfung auf Überschneidungen von Wertebereichen (bzw. Zeitintervallen) voraus. Dargestellt werden im Rahmen des Beitrages drei unterschiedliche Programmroutinen zur Realisierung einer Zuordnung von Ereignis- und Bezugsintervallen in SAS, die innerhalb eines universell verwendbaren SAS-Makros formuliert wurden und als Aufarbeitungsoptionen ausgewählt werden können. Als Resultat erhält man in allen drei vorgestellten Varianten eine Datentabelle, die zumindest Ereignisintervalldaten einschließlich ausgewählter Merkmale zu jeweils relevanten Bezugszeiträumen beinhaltet. Die Vor- und Nachteile der einzelnen Varianten werden dargestellt und diskutiert.

Schlüsselworte: Datastep, Merging, SAS-Makro, LAG-Funktion, m:n Beziehung Ereignisintervalle, Bezugsintervalle, Zuordnung von Intervallen.

1 Hintergrund

Angaben zu personenbezogenen Ereignisintervallen und Bezugszeiträumen (z.B. individuelle Krankschreibungsintervalle und Berufstätigkeitsintervalle zu einer

großen Zahl von Versicherten) werden üblicherweise in Datenbanken in zwei unterschiedlichen Tabellen erfasst. Ein einzelnes Intervall wird dabei typischerweise in jeder der beiden Tabellen in einer Tabellenzeile mit einem Von- sowie einem Bis-Datum abgelegt, die den ersten und letzten Tag des Intervalls kennzeichnen[1]. Sollen Ereignishäufigkeiten oder Ereigniszeiten für spezifische Subpopulationen bzw. Bezugszeiträume angegeben werden, ist eine vorherige Zuordnung entsprechender Informationen erforderlich (z.B. für die Angabe von Krankschreibungszeiten in einer bestimmten Berufsgruppe). Eine solche Zuordnung ist insofern komplex, als dass sowohl innerhalb eines Bezugszeitraumes mehrere Ereignisintervalle anfallen können (z.B. mehrere Krankschreibungen in einem Berufsintervall) als auch ein Ereignisintervall sich über mehrere Bezugszeiträume erstrecken kann (z.B. bei fortdauernder Krankschreibung mit zwischenzeitlichem Wechsel des Berufsstatus). Es resultiert eine Viele-zu-Viele-Beziehung, bei der zudem eine gültige Zuordnung von Intervallen erst durch eine Überprüfung auf Überschneidungen von Wertebereichen ermittelt werden kann.

2 Zuordnungswege

Nachfolgend dargestellt werden insgesamt drei Möglichkeiten einer Zuordnung von Ereignis- und Bezugsintervallen in SAS-Datasteps. Als Resultat erhält man in allen drei Varianten eine Tabelle, die (zumindest) Angaben zu Ereignisdaten innerhalb von dokumentierten Bezugszeiträumen einschließlich optional ausgewählter Merkmale der zugeordneten Bezugszeiträume enthält. Sofern sich Ereignisintervalle über mehrere Bezugszeiträume erstrecken, werden die Ereigniszeiten aufgeteilt und zugehörige Informationen einer entsprechenden Zahl von Bezugszeiträumen zugeordnet.

Die dritte der hier vorgestellten Varianten übernimmt eine Reihe von weiteren Funktionen. Sie stellt wesentliche Informationen sowohl zu Bezugszeiträumen als auch zu Ereigniszeiträumen in einer einzigen Tabelle bereit. Hierzu werden sowohl Informationen zu Ereignisintervallen als auch zu Bezugsintervallen diskreten Zeitintervallen zugeordnet, die zuvor aus allen personenbezogen dokumentierten Datumsangaben in den beiden Ursprungstabellen abgeleitet wurden. Die resultierende Tabelle enthält neben ausgewählten Merkmalen zu Bezugs- und Ereignisintervallen ergänzend eine

[1] Hinweis: Das hier vorgestellte Makro geht grundsätzlich von der Verwendung ganzzahliger Datumswerte aus, die sich auch bei der Verwendung des üblichen SAS-Datumsformates ergeben. Dabei werden sowohl der Von-Datumswert als auch der Bis-Datumswert als Elemente des jeweils beschriebenen Zeitintervalls betrachtet. Die Intervalllänge berechnet sich dabei nach der Formel Intervalllänge = Bis-Datum - Von-Datum + 1.

Reihe von Zähl- und Indikatorvariablen, die unter anderem auch eine Überprüfung der Zuordnung sowie eine einfache Quantifizierung von zeitlichen Überlappungen von Intervallen innerhalb beider Ausgangstabellen sowie von Ereigniszeiten ohne Zuordnung zu Bezugsintervallen erlauben. Die vorgestellten drei Zuordnungsvarianten, die als auswählbare Ablaufvarianten eines übergeordneten Makros realisiert sind, wurden im Hinblick auf ihre Performance bei großen Datensätzen in einem spezifischen Anwendungsfall unter SAS 9.1 auf einem PC-System unter Windows 2000 überprüft.

Die nachfolgenden beiden Tabellen zeigen kurze Datenbeispiele zu Ausgangsdaten, zu denen später die Ergebnistabellen als Resultate der drei Aufarbeitungsvarianten vorgestellt werden sollen. Neben den obligaten Angaben Personenkennzeichen (im Beispiel die Variable ID), Von-Datum und Bis-Datum beinhaltet in den Beispielen die Tabelle zu Bezugsintervallen drei weitere Merkmale (De_V1-De_V3), die Tabelle zu Ereignisintervallen zwei Merkmale bzw. Variablen (Num_V1, Num_V2).

Tabelle 1: Beispieldaten zu Bezugsintervallen

ID	De_From	De_To	De_V1	De_V2	De_V3
1	01/01/2000	10/01/2000	Otto	100	d1
2	01/01/2000	20/08/2000	Hans	110	d2
2	25/08/2000	31/01/2003	Hans	410	d3
4	01/01/2000	31/12/2004	Gudrun	110	d4
5	01/08/2000	31/07/2003	Oskar	110	d5
5	01/08/2003	31/12/2003	Oskar	500	d6
5	01/02/2004	31/05/2004	Oskar	100	d7
6	01/09/2000	31/12/2000	Maria	100	d8
6	01/10/2000	31/12/2000	Maria	500	d9

Tabelle 2: Beispieldaten zu Ereignisintervallen

ID	Num_From	Num_To	Num_V1	Num_V2
1	01/01/2000	10/01/2000	F10	n1
3	04/03/2001	05/03/2001	D12	n2
4	01/01/2001	10/01/2001	H04	n3
4	08/05/2001	18/05/2001	L11	n4
4	14/12/2004	14/01/2005	A10	n5
5	01/07/2003	30/06/2004	M54	n6
6	01/08/2000	15/10/2000	E14	n7

2.1 Variante 1

Eine vom Grundgedanken her sehr einfache Möglichkeit der Zuordnung von Ereignis- und Bezugsintervallen wird in Variante 1 realisiert: Jedes personenbezogen vorhandene Bezugsintervall wird im Rahmen einer Programmschleife einzeln selektiert. Die einzelnen Bezugsintervalle werden anschließend mit allen personenbezogen vorhandenen Ereignisintervallen (in beliebiger Zahl) über ein Merging in einem SAS Datastep verknüpft. In eine Ergebnisteildatei werden in dem Datastep schließlich nur die Ereignisintervalle geschrieben, die zeitliche Überschneidungen zu dem jeweiligen Bezugsintervall aufweisen. Durch die Auswahl jeweils nur eines einzelnen Bezugsintervalls besteht in jedem Bearbeitungsschritt eine 1:n Beziehung, die beim Merging (über eine personenbezogene ID-Variable) im Datastep keine Probleme bereitet. Abschließend werden alle Teildateien in einer Ergebnisdatei mit zuordnungsfähigen Ereignisintervallen zusammengefasst.

Die Laufzeit eines entsprechenden Programms hängt maßgeblich von der Zahl der erforderlichen Programmschleifendurchläufe und damit von der Zahl der personenbezogen vorhandenen Bezugsintervalle ab. Die Zahl der Bezugsintervalle variiert in den hier beispielhaft aufgeführten Daten zur Berufstätigkeit allerdings extrem. Obwohl zu einem wesentlichen Teil der Versicherten lediglich ein Bezugsintervall erfasst ist, können in einzelnen Fällen mehr als 1000 entsprechende Intervalle dokumentiert sein. Vor diesem Hintergrund erscheint eine Aufteilung der Daten entsprechend der Anzahl der personenbezogen dokumentierten Bezugsintervalle nahezu obligat.
Die Bestimmung der Anzahl von personenbezogenen Bezugsintervallen und die anschließende Aufteilung sowohl der Bezugs- als auch der Ereignisdaten in mehrere Teildatensätze in Abhängigkeit von der Bezugsintervallzahl bildet daher einen ersten, wesentlichen Programmschritt. Hierfür können beim Aufruf des Makros (in allen drei Varianten) unterschiedliche Modalitäten gewählt werden (vgl. Tabelle zu Makro-Parametern im Anhang A). So führt der Aufruf des Makros mit dem Parameter "V 2 5 100" beispielsweise dazu, dass durch die Kennung "V" die nachfolgenden, durch Leerzeichen getrennten Zahlen als einfache Liste von Werten interpretiert werden, die im Programmablauf jeweils als obere Grenzwerte der personenbezogenen Bezugsintervallzahl in den einzelnen Teildateien verwendet werden.

Einen Sonderfall bilden Datensätze zu Personen mit nur einem Bezugsintervall[2]. Diese Datensätze werden grundsätzlich gesondert betrachtet, da in entsprechenden Fällen per se eine 1:n Beziehung besteht. Für entsprechende Datensätze genügt grundsätzlich ein einfaches Merging des einen Bezugsintervalls mit allen dokumentierten Ereignisintervallen, um den Ereignisintervallen Informationen zum Bezugszeitraum zuzuordnen.

Die höchste Zahl der personenbezogen vorhandenen Bezugsintervalle wird im Makro automatisiert bestimmt und im Regelfall als maximales Limit zur Aufteilung der Ursprungsdateien und zur Bestimmung der maximal erforderlichen Programmschleifendurchläufe verwendet. Existieren in ausgewerteten Daten beispielsweise personenbezogen maximal 512 Bezugsintervalle, werden, in Abhängigkeit von der Zahl der Bezugsintervalle, nach dem zuvor genannten Aufruf mit "V 2 5 100" folgende Teildateien (jeweils getrennt für Bezugs- und Ereignisintervalle) gebildet:
Teildateien zu Personen mit genau einem Bezugsintervall,
zu Personen mit zwei Bezugsintervallen,
zu Personen mit 3 bis 5 Bezugsintervallen,
zu Personen mit 6 bis 100 Bezugsintervallen,
zu Personen mit 101 bis 512 Bezugsintervallen.

Die Zuordnung jeweils aller Ereignisintervalle zu einzelnen Bezugsintervallen wird über ein Sub-Makro mit Programmschleifen realisiert, welche jeweils einzelne, zuvor fortlaufend (unter Nutzung von PROC RANK) durchnummerierte Bezugsintervalle mit allen personenbezogen verfügbaren Ereignisintervallen verknüpfen bzw. mergen. In den Teildatensätzen zu Personen mit 3 bis maximal 5 Bezugsintervallen wird die Programmschleife also beispielsweise 5 mal durchlaufen, um nacheinander die personenbezogen ggf. vorhandenen Bezugsintervalle Nr.1 bis maximal Nr.5 mit jeweils allen Ereignisintervallen zu mergen. In Ergebnisdateien werden dabei nur jeweils die Beobachtungen übernommen, bei denen Überschneidungen zwischen Bezugs- und Ereigniszeiträumen festgestellt wurden.

Je Programmschleifendurchlauf wird eine Subdatei (der Teildatei) erzeugt (jeweils zu Bezugsintervall 1, 2, 3, ...u.s.w.). Die Dateinamen zu diesen Subdateien werden in einer einzigen Makro-Variablen (&M) gespeichert, welche in einem SET-Statement in einem abschließenden Datastep zur Zusammenführung aller Subdateien zu einer Ergebnis(teil)datei verwendet wird.

[2] In Variante 3: Personen mit insgesamt einem Intervall, vgl. spätere Abschnitte.

Da die Zeichenkettenlänge von Makro-Variablen begrenzt ist, lassen sich auf diese Art maximal etwa 5600 Subdateien in einem Schritt zusammenführen. Vor diesem Hintergrund ist die hier beschriebene Variante 1 des Makros in seiner Anwendung automatisch auf personenbezogene Datensätze mit maximal 5600 Bezugsintervallen limitiert. Finden sich in den verarbeiteten Daten einzelne Personen mit mehr Bezugsintervallen, werden die Daten zu diesen Personen grundsätzlich in Variante 1 nicht verarbeitet.

Auszug SAS-Kode zu Variante 1:

```
%LET M=;        *MACRO-Variable to keep names of sub-datasets;

%DO Z=1 %TO &MAXINT.;

     data r&Z. (drop=obs_nor bnn_id);

     *merge Denominator interval No. &Z. and Numerator interv.;
     merge
       d&MAXINT. (drop=obs_no bnn_id where=(obs_nor=&Z.))
       n&MAXINT. (where=(bnn_id ge &Z.));
     by &ID_VAR.;

     *Estimation of overlapping;
     if &DBas_TO. ne . and &DNum_TO. ne .
          and &DBas_FROM. ne . and &DNum_FROM. ne . then
          &Int_in.= min(&DBas_TO.,&DNum_TO.) -
                    max(&DBas_FROM.,&DNum_FROM.)+1;

     *write to r&Z. if intervals are overlapping;
     if &Int_in. ge 1 then output r&Z.;
     run;

     %LET M=&M. r&Z.;
%END;
```

Ausdrücklich hingewiesen sei an dieser Stelle auf die Tatsache, dass im Rahmen des Makroablaufs ggf. sehr viele temporäre Datendateien generiert werden können, die zum Abschluss des Gesamtdurchlaufs wieder gelöscht werden. Durch Überschneidungen bei der Dateibenennung können daher möglicherweise Dateien, die vom Nut-

zer zuvor selbst im WORK-Verzeichnis erstellt wurden, überschrieben werden, weshalb das WORK-Verzeichnis vor dem Makro-Aufruf sicherheitshalber keine später vom Anwender benötigten Dateien enthalten sollte.

2.2 Variante 2

Die Variante 2 zur Zuordnung von Bezugsintervallmerkmalen zu Ereignisintervallen basiert vorrangig auf einer extensiven Nutzung von LAG-Funktionen innerhalb einer Sub-Makro-Schleife. LAG-Funktionen ermöglichen in SAS eine komfortable Zugriffsmöglichkeit auf Variablen-Werte aus vorausgehenden Tabellenzeilen, wobei über Makro-Schleifen eine quasi beliebige Anzahl von vorausgehenden Tabellenzeilen schrittweise angesprochen werden kann (vgl. Kode-Beispiel). Um über LAG-Funktionen die erwarteten (und einfach interpretierbare) Resultate zu erhalten, sollten allerdings nach praktischen Erfahrungen zwei Regeln eingehalten werden:
Eine LAG-Funktion sollte innerhalb eines Datasteps immer unbedingt, also nicht innerhalb einer IF-Bedingung berechnet werden, die unterschiedliche Behandlungen einzelner Beobachtungen zur Folge hat.
Gleichzeitig sollte eine LAG-Funktion ausschließlich auf Variablen zugreifen, die ihrerseits innerhalb des selben Datasteps nach dem Aufruf der LAG-Funktion unverändert bleiben.

Das Grundprinzip der Variante 2 ist einfach. Alle Bezugs- und Ereignisintervalle werden in einer Datei zusammengeführt, wobei personenbezogen erfasste Bezugsintervalle grundsätzlich in separaten Tabellenzeilen vor den potentiell zugehörigen Ereignisintervallen stehen. Dies wird dadurch gewährleistet, dass die anordnungsrelevante Variable bei Bezugsintervallen aus dem VON-Datum und bei Ereignisintervallen aus dem BIS-Datum (unter Addition eines konstanten Wertes von 0,1) des jeweilig dokumentierten Intervalls gebildet wird. Anschließend werden in einer Sub-Makro-Schleife über LAG-Funktionen potentiell relevante Informationen aus den voranstehenden Datensätzen gelesen. Bei jeder personenbezogen festgestellten Überschneidung von Ereignis- und Bezugsintervallen werden entsprechende Datensätze in eine Ergebnisdatei herausgeschrieben (vgl. Auszug zum SAS-Kode).

Wie in Variante 1 hängt auch in Variante 2 die Zahl der Durchläufe der Submakro-Schleife von der Anzahl der personenbezogen dokumentierten Intervalle ab, weshalb sich eine Aufteilung in Teildateien in der Regel empfiehlt und entsprechend den Optionen in Variante 1 auch durchgeführt werden kann. In Variante 2 ist allerdings nicht die Zahl der Bezugsintervalle, sondern die Summe der Anzahl von Ereignis-

und Bezugsintervallen relevant (sofern von beiden Intervallarten mindestens eines personenbezogen existiert - andernfalls ist eine Zuordnung ausgeschlossen), da diese beim Submakroablauf in einer Datei zunächst separat untereinander stehen. Die Aufteilung in Teildateien in Variante 2 erfolgt also analog zu Variante 1, richtet sich jedoch abweichend nach der Gesamtzahl der personenbezogen dokumentierten Ereignis- und Bezugsintervalle.

Auszug SAS-Kode Variante 2:

```
data m&MAXINT. (keep=&ID_VAR. &DBas_FROM. &DBas_TO. &BAS_VAR.
          &DNum_FROM. &DNum_TO. &Num_VAR. &Start_Ind. &Int_in.);

merge d&MAXINT. n&MAXINT.;          *MERGING ;
by &ID_VAR. mdate;        *date variable used to order observations;

%DO Z=0 %TO &MAXINT.;

     lid=lag&Z.(&ID_VAR.); &DBas_FROM.=lag&Z.(bf);
     &DBas_TO.=lag&Z.(bt);

     *handling of optional denominator variables;

     %IF &CBAS_VAR. gt 0 %THEN %DO ZZ=1 %TO &CBAS_VAR.;
          &&BV&ZZ.=lag&Z.(BV&ZZ.);
     %END;

     *if prev.denom.intervals have same ID
      and date values not missing;

     if lid=&ID_VAR. and &DBas_FROM. ne . and &DBas_TO. ne .
          and &DNum_FROM. ne . and &DNum_TO. ne . then do;

          &Int_in.=min(&DBas_TO.,&DNum_TO.)-
               max(&DBas_FROM.,&DNum_FROM.)+1;

          if &Int_in. ge 1 then output m&MAXINT.;
     end;
%END;
run;
```

Sowohl Variante 1 als auch Variante 2 stellen verhältnismäßig einfache Aufarbeitungswege dar. Der Hauptteil des zugehörigen SAS Programmkodes ist zur automati-

sierten Aufteilung der Ausgangsdateien in Teildateien und zur Vorbereitung der entsprechend angepassten Aufrufe der Sub-Makros erforderlich, welche dann die eigentliche Zuordnung übernehmen.
Variante 1 ist in der hier besprochenen Makroumsetzung in der Anwendung auf Datensätze beschränkt, die *personenbezogen* maximal 5600 Bezugsintervalle umfassen. Daten zu Personen mit mehr als 5600 Bezugsintervallen werden nicht verarbeitet. Entsprechende grundsätzliche Beschränkungen bestehen bei der Variante 2 nicht. Die in Variante 2 maßgeblich verwendete LAG-Funktion arbeitet auch bei einer sehr großen Intervallzahl noch zuverlässig, ihre Verwendung dürfte daher an sich nicht zu praxisrelevanten Einschränkungen führen.

Tabelle 3: Ergebnisdatei zu Beispieldaten Variante 1 und 2

ID	De_From	De_To	De_V1	De_V2	De_V3	Num_From	Num_To	Num_V1	Num_V2	A_days
1	01/01/00	10/01/00	Otto	100	d1	01/01/00	10/01/00	F10	n1	10
4	01/01/00	31/12/04	Gudrun	110	d4	01/01/01	10/01/01	H04	n3	10
4	01/01/00	31/12/04	Gudrun	110	d4	08/05/01	18/05/01	L11	n4	11
4	01/01/00	31/12/04	Gudrun	110	d4	14/12/04	14/01/05	A10	n5	18
5	01/08/00	31/07/03	Oskar	110	d5	01/07/03	30/06/04	M54	n6	31
5	01/08/03	31/12/03	Oskar	500	d6	01/07/03	30/06/04	M54	n6	153
5	01/02/04	31/05/04	Oskar	100	d7	01/07/03	30/06/04	M54	n6	121
6	01/09/00	31/12/00	Maria	100	d8	01/08/00	15/10/00	E14	n7	45
6	01/10/00	31/12/00	Maria	500	d9	01/08/00	15/10/00	E14	n7	15

Beide bislang besprochenen Varianten führen zu einer Ergebnisdatei, die lediglich Informationen zu Ereignisintervallen enthält, die mindestens einem Bezugsintervall vollständig oder partiell zugeordnet werden können[3]. So ist in der oben dargestellten Ergebnisdatei beispielsweise das Ereignis zur ID 3 nicht aufgeführt (vgl. Tabelle 2).
In beiden Varianten 1 und 2 werden *alle* Informationen zu Ereignisintervallen mit zeitlicher Überlappung zu mindestens einem personenbezogenen Bezugsintervall in der Ergebnisdatei erfasst. Existieren dabei auch *innerhalb* der ursprünglichen Ereignis- oder Bezugsdaten personenbezogene Intervalle mit zeitlichen Überlappungen (also beispielsweise mehrere separat dokumentierte personenbezogene Angaben zur Berufstätigkeit für denselben Zeitraum), werden ungeachtet dieser nicht-diskreten Dokumentation von Zeitintervallen alle Kombinationen vorhandener Ereignisintervalle mit zeitlichen Überschneidungen zu Bezugsintervallen in der Ergebnisdatei erfasst. Also werden beispielsweise bei überlappenden Bezugsintervallen ggf. dieselben

[3] Eine Modifikation zur Bereitstellung von Ereignisintervallen ohne Bezugsdatei ließe sich in Variante 2 relativ einfach realisieren, ist jedoch als Option im Makro-Aufruf bislang nicht vorgesehen.

Ereignisintervalle auch mehrfach in die Ergebnisdatei geschrieben (vgl. dargestellte Resultate zu ID 6 (Maria). Soll dies vermieden werden, ist vor der Verwendung der Varianten 1 oder 2 eine vorangehende Aufarbeitung der Ereignis- und Bezugsdaten zur Sicherstellung personenbezogen diskret dokumentierter Zeitintervalle erforderlich. Möglichkeiten für eine derartige Aufarbeitung wurden vom Autor auf der KSFE vor zwei Jahren vorgestellt. Eine entsprechende Aufarbeitung wird zumindest partiell jedoch auch von der nachfolgend beschriebenen Variante 3 des hier präsentierten Makros geleistet.

2.3 Variante 3

Variante 3 ist im Vergleich zu den Varianten 1 und 2 aufwändiger programmiert und verursacht deutlich längere Laufzeiten. Dafür erhält man in der Ergebnisdatei Informationen sowohl zu allen zeitlich diskret abgrenzbaren Ereignisintervallen als auch zu allen diskret abgrenzbaren Bezugszeiträumen. Bestehen innerhalb der Ereignis- oder Bezugsdaten auf personenbezogener Ebene zeitliche Überlappungen, lassen sich diese auf Basis der Ergebnisdatei leicht quantifizieren, die primären Ergebnisse beruhen jedoch auf der Berücksichtigung von personenbezogen diskreten Zeitintervallen. Zusätzlich können aus der Ergebnisdatei zur Kontrolle des Programmablaufs auch bestimmte Parameter der Ursprungsdaten (Anzahl der Beobachtungen, ursprünglich und inklusive Überlappungen dokumentierte Intervalldauer) ermittelt werden.

Ein wesentliches Element der Variante 3 ist die personenbezogene Ermittlung aller diskret abgrenzbaren Zeitintervalle, die sich bei einer Zusammenführung von Ereignis- und Bezugsintervallen ergibt. Dazu werden zunächst alle vorhandenen Von- und Bis-Datumswerte personenbezogen aus den Ursprungsdaten ermittelt. Aus den Original-Von-Werten werden durch Subtraktion des Wertes 1 ggf. zu ergänzende Bis-Datumswerte für neue Zwischenintervalle gebildet, aus Original-Bis-Werten durch Addition von 1 werden ggf. zu ergänzende Von-Datumswerte zu Zwischenintervallen abgeleitet.
Es resultiert im Makro-Ablauf so eine Datei, die personenbezogen alle diskret abgrenzbaren Zeitintervalle zwischen dem kleinsten ursprünglichen Von-Datum bis zum größten ursprünglichen Bis-Datum aus einer der beiden Ursprungsdateien umfasst (einschließlich Intervall-Datensätzen zu ggf. vorhandenen dokumentationslosen Zeitintervallen).
Mit dieser Datei werden in einem nächsten Schritt sowohl Ereignis- als auch Bezugsdaten über ein Merging verknüpft (nach personenbezogener ID-Variable sowie Von-Datum; unter Beibehaltung der Bis-Datumswerte aus beiden Ursprungsdateien).

Auszug SAS-Kode Variante 3:

```
data m&MAXINT. (drop=........);
merge
m&MAXINT.       /*data set with all discrete time intervals*/
                /*denominator data*/
d&MAXINT. (rename=(&DBas_FROM.=date_from &DBas_TO.=bto)...)
                /*numerator data*/
n&MAXINT. (rename=(&DNum_FROM.=date_from &DNum_TO.=nto)...);
by &ID_VAR. date_from;

*weights to count intervals with identical from-date;
value_intb=1; value_intn=1;
if nf gt 1 and nf gt bf then do;
     value_intb=max(1,bf)/nf;
end;
if bf gt 1 and bf gt nf then do;
     value_intn=max(1,nf)/bf;
end;
bas_int=0; if bto ne . then bas_int=value_intb;
Num_int=0; if nto ne . then Num_int=value_intn;

%DO Z=1 %TO &MAXINT.*2-2;          *<<<< SUBMACRO LOOP;
     lid=lag&Z.(&ID_VAR.); lb_to=lag&Z.(bto); ln_to=lag&Z.(nto);
     lbval=lag&Z.(value_intb); lnval=lag&Z.(value_intn);

     if lid=&ID_VAR. then do;      *if previous ID still the same;
          *if prev.denom.interval overlapping (always total);
          if lb_to ge date_from then do;
          *count overlapping denom.intervals;
               bas_int=bas_int+lbval;
          end;
          *if prev.num.interval overlapping (always total);
          if ln_to ge date_from then do;
               *count overlapping num.intervals;
               Num_int=Num_int+lnval;
               end;
          end;
%END;
if bas_int gt 0 and Num_int gt 0 then &Int_in.=date_int;
obs_no=_N_;
run;
```

Sind *innerhalb* beider Ursprungsdateien keine überlappenden Zeitintervalle mit identischem Von-Datum enthalten, besteht bei dieser Zusammenführung eine 1:1 Beziehung zwischen der Datei mit diskreten Intervallen sowie den beiden Ursprungsdateien. Nur bei Überlappungen innerhalb einzelner Ursprungsdateien besteht beim Merging mit diskreten Zeitintervallen eine 1:n Beziehung, die jedoch gleichfalls keine Probleme bereitet.
Ähnlich wie in Variante 2 wird über eine Sub-Makroschleife unter Verwendung von LAG-Funktionen überprüft, ob vorausgehend beginnende Intervalle (nach Bis-Datum gemäß der Ursprungsdatei) ein einzelnes diskretes Teilintervall überschneiden (hier ggf. grundsätzlich vollständig), um relevante Informationen zu den jeweiligen diskreten Zeitintervallen verfügbar zu machen. Ggf. vorhandene Mehrfachdokumentationen zu einzelnen Zeitintervallen werden über Zählvariablen erfasst. Von den optional auswählbaren Merkmalen aus den Ausgangsdaten werden in der Ergebnisdatei allerdings nur die Merkmalausprägungen beibehalten, die zum jeweiligen Intervall im Ursprungsintervall mit dem frühesten Beginn dokumentiert sind (bei überlappenden Intervallen mit identischem Von-Datum werden, bedingt durch die spezifisch gewählte Sortierung der Datensätze, Werte aus dem Intervall mit der jeweils längeren Dauer übernommen).

Sofern 1:n Beziehungen beim Merging der Ausgangsdateien mit den diskreten Zeitintervallen bestehen, sind die diskreten Zeitintervalle in der Ergebnisdatei nach dem Merging zunächst ggf. mehrfach vorhanden, was in diesen Fällen in einem abschließenden Bearbeitungsschritt korrigiert wird. Die Endfassung der Ergebnisdatei nach Aufarbeitung durch die Makro-Variante 3 ist in Tabelle 4 dargestellt.
Die erste Spalte enthält die personenbezogene Kennung (*ID*).
Spalte 2 und 3 beinhalten die Grenzen des jeweiligen Zeitintervalls, welches immer diskret ist (*date_from, date_to*), die folgende Spalte gibt immer die Länge des Intervals an (*date_int*).
Die folgenden drei Spalten beinhalten Merkmale zu Bezugsintervallen, die zur Übernahme in die Ergebnisdatei (optional) ausgewählt wurden (hier: *De_V1, De_V2, De_V3*). Die Merkmalausprägungen werden nur für diskrete Intervalle angegeben, für die Informationen in den Ausgangsdaten zum entsprechenden Zeitraum ausgewiesen sind.
Analog beinhalten die beiden nachfolgenden Spalten Informationen zu Ereignisintervallen (hier: *Num_V1, Num_V2*).
Die vom Makro generierten Variablen *Bas_int* und *Num_Int* sind Zählvariablen, die angeben, wie viele der ursprünglich dokumentierten Bezugs- bzw. Ereignisintervalle die jeweiligen diskreten Zeitabschnitte überlappen.

Die vom Makro gleichfalls generierten Variablen *bf*, *bt*, *nf*, und *nt* zählen, wie viele ursprünglich dokumentierte Ereignis- bzw. Bezugsintervalle im jeweiligen Zeitintervall beginnen (*bf*, *nf*) oder enden (*bt*, *nt*).
Die Varable *A_days* gibt analog zur Ergebnisausgabe der Auswertungsvarianten 1 und 2 schließlich die erfassten Ereigniszeiten innerhalb von Bezugsintervallzeiträumen an. Sie entspricht in der Variante 3 der Dauer der diskreten Intervallzeiten bei genau den Beobachtungen bzw. Zeitintervallen, zu denen sowohl Bezugs- als auch Ereignisintervalle dokumentiert sind (also die Bedingung *bas_int >0 and Num_int >0* erfüllt ist).

Tabelle 4: Ergebnisdatei zu Beispieldaten Variante 3

ID	date_from	date_to	date_int	De_V1	De_V2	De_V3	Num_V1	Num_V2	Bas_int	Num_int	bf	bt	nf	nt	A_days
1	01/01/00	10/01/00	10	Otto	100	d1	F10	n1	1	1	1	1	1	1	10
2	01/01/00	20/08/00	233	Hans	110	d2			1	0	1	1	.	.	.
2	21/08/00	24/08/00	4						0	0	.	.	.	.	.
2	25/08/00	31/01/03	890	Hans	410	d3			1	0	1	1	.	.	.
3	04/03/01	05/03/01	2				D12	n2	0	1	.	.	1	1	.
4	01/01/00	31/12/00	366	Gudrun	110	d4			1	0	1	.	.	.	.
4	01/01/01	10/01/01	10	Gudrun	110	d4	H04	n3	1	1	.	.	1	1	10
4	11/01/01	07/05/01	117	Gudrun	110	d4			1	0	.	.	.	.	.
4	08/05/01	18/05/01	11	Gudrun	110	d4	L11	n4	1	1	.	.	1	1	11
4	19/05/01	13/12/04	1305	Gudrun	110	d4			1	0	.	.	.	.	.
4	14/12/04	31/12/04	18	Gudrun	110	d4	A10	n5	1	1	.	1	1	.	18
4	01/01/05	14/01/05	14				A10	n5	0	1	.	.	.	1	.
5	01/08/00	30/06/03	1064	Oskar	110	d5			1	0	1	.	.	.	.
5	01/07/03	31/07/03	31	Oskar	110	d5	M54	n6	1	1	.	1	1	.	31
5	01/08/03	31/12/03	153	Oskar	500	d6	M54	n6	1	1	1	1	.	.	153
5	01/01/04	31/01/04	31				M54	n6	0	1	.	.	.	.	.
5	01/02/04	31/05/04	121	Oskar	100	d7	M54	n6	1	1	1	1	.	.	121
5	01/06/04	30/06/04	30				M54	n6	0	1	.	.	.	1	.
6	01/08/00	31/08/00	31				E14	n7	0	1	.	.	1	.	.
6	01/09/00	30/09/00	30	Maria	100	d8	E14	n7	1	1	1	.	.	.	30
6	01/10/00	15/10/00	15	Maria	100	d8	E14	n7	2	1	1	.	.	1	15
6	16/10/00	31/12/00	77	Maria	100	d8			2	0	.	2	.	.	.

3 Exemplarische Ergebnisse zur Performance

Allgemeingültige Aussagen zur Performance der Makrovarianten lassen sich nur sehr eingeschränkt formulieren. Der Ablauf wird wesentlich durch den Datenumfang, die Häufigkeitsverteilung der Zahl personenbezogen vorhandener Datensätze sowie letztendlich auch durch eine mehr oder minder effektive Wahl der Option zur Bildung von Teildatensätzen bestimmt.
Eine hohe Zahl von Programmschleifendurchläufen sollte möglichst durch eine entsprechende Aufteilung auf Teildateien beschränkt bleiben (mit personenbezogen hoher Zahl von dokumentierten Intervallen), in denen diese auch erforderlich sind.

Zwangsläufig sehr exemplarische Ergebnisse wurden bei Testdurchläufen mit realen Daten auf einem aktuellen PC unter Windows 2000 und SAS in der Version 9.1 ermittelt. Diese Daten umfassten 2,15 Mio. Ereignisintervalle (Arbeitsunfähigkeiten, Dateigröße 68 MB) und 4,19 Mio. Bezugsintervalle (Erwerbstätigkeitsintervalle, Dateigröße 200 MB) aus einem Kalenderjahr zu insgesamt 2,7 Mio. Personen. Personenbezogen waren maximal 445 Intervalle dokumentiert. Nach der Aufarbeitung resultierten in den Varianten 1 und 2 (ohne Korrektur für überlappend dokumentierte Intervalle) Ergebnisdateien mit insgesamt 2,2 Mio. Beobachtungen (138 MB). Die Ergebnisdatei zu diskreten Zeitintervallen nach Variante 3 (inklusive Informationen zu ausschließlichen Bezugsintervallen und nicht dokumentierten Zeiträumen) umfasste 8,1 Mio. Beobachtungen (951 MB). Für die Bereitstellung der Ergebnisdateien (bei leicht variabler Bearbeitungsdauer in mehreren gleichartigen Durchläufen, Aufteilungsoption "E 2", vgl. Tabelle im Anhang zu Makro-Parameteroptionen) benötigte:
Variante 1 etwa 90 Sekunden,
Variante 2 etwa 80 Sekunden und
Variante 3 etwa 600 Sekunden.

Variante 2 verursachte in den meisten durchgeführten Testdurchläufen mit praxisnahen Daten in der hier beschriebenen Art die geringsten Laufzeiten und dürfte daher der Variante 1 regelmäßig vorzuziehen sein.
Variante 3 liefert eine Reihe zusätzlicher Informationen, die im hier beschriebenen Anwendungsfall in einer akzeptablen Bearbeitungszeit bereitgestellt werden. Die Ergebnisdatei erreicht strukturbedingt eine Größe, welche die Größe der beiden Ausgangsdateien um ein mehrfaches überschreiten kann, was insbesondere bei extrem umfangreichen Datenbeständen zu beachten ist.

4 Diskussion - Resümee

Der Beitrag stellt drei Zuordnungsmöglichkeiten von personenbezogen erfassten Ereignis- und Bezugsintervallen vor. Ziel der Zuordnung sind immer Datentabellen, die nachfolgend ihrerseits als Basis für multiple Auswertungen zu Ereignissen bzw. Ereigniszeiten innerhalb von unterschiedlich definierbaren Subgruppen von Bezugsintervallen bzw. Bezugspopulationen genutzt werden können.
Alle drei Zuordnungsvarianten wurden als Auswahloptionen innerhalb eines universell einsetzbaren SAS-Makros formuliert. Erhebliche Teile des SAS-Kodes resultieren aus dem Bestreben, das Makro mit weiteren Aufrufoptionen für unterschiedliche Ausgangsdaten auszustatten. Der zugehörige Programmkode ist im Original-Makro-Kode zumindest ansatzweise kommentiert, konnte im Rahmen dieses Beitrages jedoch nur auszugsweise erläutert werden. Obwohl das Makro an diversen eigens konstruierten Datensätzen und Beispieldaten aus der Praxis getestet wurde, sollten die Ergebnisse auf Basis des hier präsentierten Makros nicht ohne eigene Überprüfung übernommen werden - der Entwicklungsstatus des Makros ist in jedem Fall zunächst als "experimentell" anzusehen.

Die Zuordnungsvarianten 1 und 2 ermöglichen (zumindest nach den Erfahrungen des Autors mit spezifischen größeren Datensätzen) eine relativ schnelle Zuordnung von Informationen aus Bezugsintervallen zu Ereignisintervallen, sofern diese überhaupt in die dokumentierten Bezugszeiträume fallen.
Liegen in den verarbeiteten Daten sowohl Bezugs- als auch Ereignisintervalle bereits nachweislich als personenbezogen diskret dokumentierte Intervalle vor (personenbezogen jeweils nur maximal ein dokumentiertes Bezugs- und/oder Ereignisintervall mit Gültigkeit für einen bestimmten Zeitpunkt) oder ist z.B. die mehrfache Zuordnung einzelner Ereignisintervalle zu mehreren überlappend dokumentierten Bezugszeiträumen beabsichtigt und werden zudem Informationen zu Ereignisintervallen ohne Zuordnung zu Bezugszeiträumen nicht benötigt, empfiehlt sich der Einsatz der Zuordnungsvarianten 1 und 2. Bei identischen Ergebnisdateien beider Varianten (abgesehen von der Sortierung der Intervalle) dürfte bei entsprechenden Auswertungen die nicht im Hinblick auf die Intervallzahl limitierte Variante 2 auch aufgrund der zumeist etwas günstigeren Laufzeiten vorzuziehen sein.

Bei der hier dargestellten Zuordnungsvariante 3 werden - verbunden mit einer merklich längeren Laufzeit - in der Ergebnisdatei Informationen sowohl zu Bezugsintervallen als auch zu Ereignisintervallen *für alle dokumentierten Bezugs- oder Ereigniszeitpunkte* bereitgestellt. Die Informationen werden dabei, unabhängig von der Art

der Dokumentation in den Ausgangsdaten, grundsätzlich und ausschließlich in Bezug auf diskret abgegrenzte Zeitintervalle angegeben. Durch automatisch gebildete Indikator- und Zählvariablen ergeben sich eine Reihe von Informationen, die potenziell zur Kontrolle der Struktur der Ausgangsdaten sowie des Makro-Ablaufs genutzt werden können (z.B. über eine nachfolgend leicht durchführbare Bestimmung des Anteils der ursprünglich überlappend dokumentierten Zeiträume oder des Anteils nicht zugeordneter Intervalle). Sind entsprechend umfangreiche Informationen oder die Gewährleistung einer Auswertungsdatei mit diskret aufgearbeiteten Zeitintervallen erwünscht, empfiehlt sich die Zuordnungsvariante 3.

Insbesondere die Ergebnisse der Zuordnungsvariante 3 dürften sich auch zur Kontrolle eigener Aufarbeitungsroutinen in großen Datenbeständen eignen. Sollten sich dabei Hinweise auf mögliche Fehlfunktionen des hier präsentierten Makros ergeben, würde sich der Autor über entsprechende Rückmeldungen im Sinne einer Qualitätssicherung freuen.

Anhang A: Parameter zum Makro-Aufruf

Macro Parameter	**What to provide, Comments** Macro **%INTZU100(** parameter1, 2, ...**)**; Version 1.10
ID_VAR	name of ID variable (has to be used both in numerator/denominator file, obligate nonmissing values)
Bas_FILE	name of denominator file (file must be sorted by ID_VAR! - obligate)
Num_FILE	name of numerator file (file must be sorted by ID_VAR! - obligate)
Res_FILE	name of final result file (obligate, use libname prefix to create a permanent result file)
DBas_FROM	name of FROM date variable: denominator interval (obligate nonmissing numerical integer)
DBas_TO	name of TO date variable: denominator interval (obligate nonmissing numerical integer)
DNum_FROM	name of FROM date variable: numerator interval (obligate nonmissing numerical integer)
DNum_TO	name of TO date variable: numerator interval (obligate nonmissing numerical integer)
Bas_VAR	facult. variable list: Denominator variables separated by blanks to be included in result file
Num_VAR	facult. variable list: numerator variables separated by blanks to be included in result file
Start_Con	facult. additional condition to count DNum_FROM as a start point if within

Macro Parameter	**What to provide, Comments** Macro **%INTZU100(** parameter1, 2, ...**)**; Version 1.10
	denom. interval, has to start with "and ....."
Start_Ind	facult. name of the indicator variable to indicate start of interval within denom. interval (value will be 1 if From-date is start point (only if within denom. interval in v.1 and v.2) - if blank: variable Start_Ind will not be estimated)
Int_in	name of new variable holding estimated numerator time within denominator interval in result file (obligate)
NOPRINT	insert word "noprint" to reduce output, or leave blank to see full output
NUM_ONLY	state "Y" to run only part of the Macro to estimate number of denom.intervals, "B" to use estimations done before, otherwise (default) all estimations will be done
SEPSTEP	How to divide data in separate files according to the estimated number of intervals per ID, always have to be one of the three letters S, V or E and at least one number: ***S xxx*** use S to indicate that constant intervals (given by xxx) should be used to devide (e.g.: "S 10" will divide in intervals like 1, 11, 21, 31....., upper limit) ***V xx xx***... use V to indicate that the following numbers should be used to divide files according to no. of intervals per ID (e.g.: "V 2 5 100" will divide in intervals 1, 2, 5, 100, upper limit) ***E xx*** use E to indicate to use the following number as a base to calculate an exponential row (e.g.: "E 2" will divide in intervals 1, 2, 4, 8, 16, ..., upper limit) ==== additional option (optionale use): ========== ***... L xx*** use L followed by a number to indicate an upper limit of intervals per ID to be read - all IDs with more intervals will be ignored! - CAVE: different estimations are used in different variants (see below)!
VARIANT	to select variant (1,2 or 3) - see text of the publication for comments
FORMDAT	to define Format of Date-Variables (variant 3 only)
CHECK	if "Y" or "YNP" then comparing time sum numerator file vs. result file is done (***only partly implemented jet!!!***) if "NP" or "YNP" then a NEW calculation time protocol data set will be created (previous will be deleted!!!)
You can use the additionally provided small program **Intzu_test_data.sas** to generate test data files with numerator and denominator intervals!	