A SAS Macro to Predict Event Dates

Kim Lea Weyer
PAREXEL International
Am Bahnhof Westend 15
14059 Berlin
Kim.Weyer@PAREXEL.com

Zusammenfassung

Ziel der vorliegenden Arbeit „A SAS Macro to Predict Event Dates“ war es, ein Makro in der Programmiersprache SAS vorzustellen, mit dem man das Datum einer Zwischen– oder Endauswertung einer ereignis-basierten Doppelblindstudie, vorhersagen kann. Die Methodik, welche bei der Implementierung des Makros verwendet wurde, besteht aus drei essenziellen Teilen: der Simulation von Rekrutierungsdaten, der Vorhersage von Ereignisdaten und der Vorhersage der Daten von Patienten, die vorzeitig aus der Studie ausscheiden werden.

Zuerst werden die mathematischen Grundlagen der bayesschen Inferenz erläutert, welche zur Simulation von Rekrutierungsdaten verwendet werden. Anschließend werden Aspekte der Überlebenszeitanalyse definiert. Diese werden zur Vorhersage der Daten von Ereignissen und Studienabbrechern verwendet. Danach wird das Makro vorgestellt, welches aus fünf Unter-Makros besteht. Anhand eines Beispiels wird die von dem Makro erzeugte Datei genauer betrachtet.

Die Arbeit wurde in Kooperation zwischen der Otto von Guericke Universität Magdeburg und der Firma Parexel International erstellt.

Schlüsselwörter: Event Prediction, Accelerated Failure Time Modell, Bayessche Inferenz, Ereignis-basierte multizentrische Doppelblindstudie

1 Vorwort

Die Arbeit wurde in Kooperation der Otto von Guericke Universität Magdeburg und dem CRO (engl. Clinical Research Organisation) Parexel International erstellt. Zur Programmierung wurde die Version SAS 9.3 verwendet.

2 Einleitung

In ereignis-basierten Doppelblindstudien werden oft neben der Endauswertung Zwischenauswertungen durchgeführt. Dabei spielt der Zeitpunkt beider Auswertungen eine wichtige Rolle. In ereignis-basierten klinischen Studien ist der Zeitpunkt einer Zwischenauswertung (Endauswertung) durch eine vorher definierte Anzahl von Ereignissen festgelegt, dem sogenannten „landmark event“ oder „milestone event“. Die Durchführung einer solchen Analyse (zwischen, als auch end) ist immer mit erheblichem Auf-

wand verbunden. Deshalb ist eine genaue Prognose des „landmark events“ für die Planung von großer Bedeutung.
Angenommen, es liegen Daten einer ereignis-basierten multizentrischen Doppelblindstudie vor, bei der die Rekrutierung noch nicht abgeschlossen ist. Im Folgendem wird eine Möglichkeit vorgestellt, mit der für eine solche Studie das Datum einer Zwischen- oder Endauswertung vorhergesagt werden kann. Dabei besteht die ausgewählte Methode aus drei essenziellen Teilen: der Simulation von Rekrutierungsdaten bis die geplante Anzahl von Patienten erreicht ist, der Simulation von Ereignisdaten und Daten von Studienabbrechern für Patienten, die weder ein Ereignis haben oder aus der Studie ausgetreten sind. Anschließend wird das Makro, mit seinen fünf Unter-Makros, vorgestellt.

3 Methodik

3.1 Simulation von Rekrutierungsdaten

In diesem Kapitel wird eine Methode zur Simulation von Rekrutierungsdaten vorgestellt. Dazu werden die Grundlagen der bayesschen Inferenz erläutert. Falls nicht anders beschrieben, folgen die Beschreibungen und Notationen denen von Gelman [4], Hartung [5] und Lessaffre [9].

Bayessche Inferenz
Die bayessche Inferenz ist eine Methode der statistischen Inferenz, welche auf dem Satz von Bayes basiert. Zunächst wird ein Experiment mit einer Beobachtung x betrachtet. Bei der bayessche Inferenz wird angenommen, dass diese Beobachtung x die Realisierung einer Zufallsvariable X ist. Weiterhin soll die Zufallsvariable X eine Verteilungsfunktion mit einer Dichtefunktion besitzen. Bevor das Experiment beginnt, wird vorausgesetzt, dass bereits eine Idee über den Parameter θ existiert. Diese Idee über den unbekannten Parameter θ kann durch eine Dichtefunktion $f(\theta)$ beschrieben werden, welche als 'A-priori Verteilung' bezeichnet wird. Nachdem das Experiment gestartet ist, sind Daten und damit Information verfügbar. Die gemeinsame Dichtefunktion für die Daten und dem unbekanntem Parameter θ ist gegeben durch

$$f(x,\theta) = f(x|\theta)f(\theta). \quad \textbf{(1)}$$

Dabei ist $f(x|\theta)$ der Likelihood der Zufallsvariablen X, gegeben den unbekannten Parameter θ. Die A-posteriori Verteilung kombiniert die Information der Daten und der A-priori Verteilung und ist gegeben durch

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int_{-\infty}^{\infty} f(x|\theta)f(\theta)d\theta}. \quad \textbf{(2)}$$

Alle weiteren Schlussfolgerungen über den Verlauf des Experimentes erfolgen nur durch die Betrachtung der A-posteriori Verteilung.

Simulationsmethode

Bagiella und Heitjan stellten 2001 in ihrem Paper „Predicting analysis times in randomized clinical trials“ [2] eine Methode, welche auf der bayesschen Inferenz basiert, zur Simulation von Rekrutierungsdaten vor. Diese Methode basiert auf einem Poisson-Modell. Es wird angenommen, dass die Zufallszahlen x_i, i =1,…,n Poisson-verteilt sind,

$$x_i \sim Poi(y_i\theta)\,,$$

mit y_i als positive, erklärende Variable von x und θ als unbekannten Parameter. Die konjugierte A-priori Verteilung von θ ist die Gammaverteilung mit den Parametern α und β, Gamma(α, β). Die A-posteriori Verteilung von θ unter der Bedingung $\boldsymbol{x} = (x_1, \dots, x_n)$ ist geben durch

$$\theta \mid \boldsymbol{x} \sim Gamma\left(\alpha + \sum_{i=1}^{n} x_i\,, \beta + \sum_{i=1}^{n} y_i\right). \qquad \textbf{(3)}$$

Bagilla und Heitjan verwenden in ihrem Paper [2] eine zu (3) ähnliche A-posteriori Verteilung für eine Poisson-verteilte Rekrutierungsrate θ. Dabei werden die Parameter folgendermaßen interpretiert. Bei der A-priori Verteilung wird der Parameter α als die Anzahl der rekrutierten Patienten in Zeitintervall β angesehen. Weiterhin ist im Likelihood (Information aus den beobachteten Daten) die Gesamtanzahl der Patienten N, welche im Zeitintervall t_0 rekrutiert wurden, gegeben. Daraus resultiert folgende A-posteriori Verteilung:

$$Gamma(\alpha + \mathrm{N}, \beta + t_o). \qquad \textbf{(4)}$$

Durch das Ziehen von Zufallszahlen aus der A-posteriori Verteilung können Rekrutierungsdaten simuliert werden.

Umsetzung in SAS

In der Programmiersprache SAS können Gamma-verteilte Zufallszahlen mit Hilfe der SAS-Funktion `RAND('GAMMA',<alpha>)`, mit Skalenparameter $\lambda = 1$ simuliert werden. Um Zufallszahlen der Gammaverteilung mit Skalenparameter $\lambda \neq 1$ zu generieren, muss eine andere Darstellungsform der Gammaverteilung gewählt werden

$$Gamma(\alpha, \lambda) = Gamma(\alpha, 1) * \lambda, \quad mit\ \lambda = 1\ /\ \beta > 0\ . \qquad \textbf{(5)}$$

Die Simulation eines Rekrutierungsdatum erfolgt in folgenden Schritten. Zuerst wird eine Gamma-verteilte Zufallszahl `xx` generiert. Anhand dieser wird die Wartezeit `xx_3` bis zur nächsten Rekrutierung berechnet. Das neue Rekrutierungsdatum ergibt sich dann durch das Addieren der Wartezeit `xx_3` zum Letzen vorhandenen Rekrutierungsdatum

`startDt`. Für die Simulation von n Rekrutierungsdaten kann folgende SAS-Schleife in einem DATA-Step verwendet werden:

```
DO j=1 TO n;
  xx   = RAND('GAMMA',posteriorA)/posteriorB;
  xx_2 = RAND('GAMMA',1);
  xx_3 = xx_2/xx;
  startDt = startDt+xx_3;
  OUTPUT;
END;
```

3.2 Simulation der Daten von Ereignissen und Studienabbrechern

In diesem Kapitel werden die Survival Funktion, die Hazard Funktion und das Accelerated Failure Time Modell eingeführt. Falls nicht anderes angegeben, folgen die Beschreibung und Notationen denen von Collett [3], Kahle [6] und Lawless [7].

Die Survival Funktion und die Hazard Funktion
Die Survival Funktion, auch Überlebensfunktion genannt, beschreibt die Wahrscheinlichkeit, dass ein Individuum ein bestimmtes Ereignis überlebt. Sei T eine Lebenszeit, dann ist die Survival Funktion gegeben durch

$$S(t) = 1 - F(t), \qquad t \geq 0, \tag{6}$$

mit F(t) als Verteilungsfunktion.

Die Hazard Funktion ist eine nicht-negative Funktion und wird auch als Ausfallrate bezeichnet. Sie beschreibt den Grenzwert der Wahrscheinlichkeit, dass ein Ereignis in einem Zeitintervall $(t, t + t_0]$ eintritt, unter der Bedingung, dass kein Ereignis vor dem Zeitpunkt t eingetreten ist. Weiterhin wird diese bedingte Wahrscheinlichkeit durch die Zeit t_0 dividiert. Für eine stetige Lebenszeit T ist die Hazard Funktion gegeben durch

$$h(t) = \lim_{t_0 \to 0} \frac{P(t < T \leq t + t_0 | T > t)}{t_0} = -\frac{\partial}{\partial t} \log S(t), \qquad t \geq 0. \tag{7}$$

Das Accelerated Failure Time Modell
In klinischen Studien gibt es oft Kovariablen oder erklärende Variablen, die einen Einfluss auf die Lebenszeit der Patienten haben. Dies können z.B. Behandlungsgruppen, Art der Behandlung, Alter usw. sein. Regressionsmodelle wie das Accelerated Failure Time (AFT) Modell können dazu in Betracht gezogen werden.

Bei dem AFT Modell wird angenommen, dass die erklärende Variable einen Einfluss auf die Lebenszeit hat, sie beschleunigt oder verlangsamt diese. Sei $p \in \mathbb{N}$ und $T \geq 0$ eine Lebenszeit. Weiterhin sei $\boldsymbol{X} = (X_1, \ldots, X_p)'$ ein $p \times 1$ Vektor der erklärenden Vari-

ablen und $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)'$ ein $p \times 1$ Vektor des Regressionskoeffizienten. Dann ist das AFT Modell gegeben durch

$$S(t|\boldsymbol{X}) = S_0\left(\frac{t}{\exp(\boldsymbol{\gamma}'\boldsymbol{X})}\right), \qquad t \geq 0. \tag{8}$$

Dabei ist S_0 die Basis Survival Funktion mit Kovariablenvektor $\boldsymbol{X} = \boldsymbol{0}$ und $\exp(\boldsymbol{\gamma}'\boldsymbol{X})$ der Beschleunigungsfaktor. Die zugehörige Hazard Funktion ist gegeben durch

$$h(t|\boldsymbol{X}) = \exp(-\boldsymbol{\gamma}'\boldsymbol{X})h_0\left(\frac{t}{\exp(\boldsymbol{\gamma}'\boldsymbol{X})}\right), \qquad t \geq 0. \tag{9}$$

Für eine Lebenszeit $T > 0$ kann das AFT Modell für die logarithmierte Lebenszeit $Y = log(T)$ durch das folgende logarithmierte lineare Modell beschrieben werden:

$$Y = log(T) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X} + bZ. \tag{10}$$

Dabei ist $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ ein $p \times 1$ Vektor der unbekannten Koeffizienten, β_0 der Achsenabschnitt, b der Skalenparameter und Z eine Zufallsvariable mit Verteilungsfunktion.

Um die AFT Modelle anzupassen, wird häufig die Maximum-Likelihood-Methode verwendet. Da die Beschreibung der Maximum-Likelihood-Modelle den Rahmen dieser Publikation sprengen würde, wird sie nicht erläutert. Eine detaillierte Beschreibung der Maximum-Likelihood-Methode für AFT Modelle kann in Lawless [7] auf den Seiten 292 bis 294 entnommen werden. Der Newton-Raphson Algorithmus, welcher bei der Maximum-Likelihood-Methode zur Maximierung der Informationsmatrix in SAS verwendet wird, wird in den Büchern „Statistical Methods for Survival Data Analysis“ von Lee [8] auf den Seiten 428 bis 432 und „Survival Analysis Using SAS: A Practical Guide, Second Edition.“ von Allison [1] auf den Seiten 90 bis 93 beschrieben.

Simulationsmethode

Angenommen, die Daten der Ereignisse oder Studienausscheidern folgen einer bestimmten Verteilung und die Parameter der Verteilung wurden durch die Maximum-Likelihood-Methode geschätzt. Für Patienten, die in der Zukunft rekrutiert werden, kann die Zeit bis zu einem Ereignis oder Studienabbruch durch das Generieren von Zufallszahlen aus der Verteilung simuliert werden. Für Patienten, die bereits in der Studie rekrutiert wurden, muss die Zeit, in der sie sich bereits in der Studie befinden, berücksichtig werden. Dazu kann die bedingte Survival Funktion betrachtet werden

$$S(t + t_0|t_0) = \frac{S(t + t_0)}{S(t_0)}. \tag{11}$$

Um t zu simulieren, wird angenommen, dass die bedingte Wahrscheinlichkeit gleich einer gleichverteilten Zufallsvariable $U, U \sim Uniform(0,1)$, mit Realisierung u ist

$$\frac{S(t+t_0)}{S(t_0)} = u. \quad \textbf{(12)}$$

Durch die Umformung der Gleichung (12) und der Verwendung der Quantil Funktion kann t folgendermaßen dargestellt werden:

$$t = F^{-1}(1 - [S(t_0) * u]) - t_0. \quad \textbf{(13)}$$

Durch das Generieren von Zufallszahlen aus der Gleichverteilung und der Verwendung der Formel (13) können Daten für Ereignisse und Studienabbrechern, gegeben dass diese in der Studie noch kein Ereignis bzw. Studienabbruch hatten, simuliert werden.

Umsetzung in SAS

In SAS können die Parameter eines AFT Modells mit Hilfe der Prozedur `PROC LIFEREG` geschätzt werden. Im folgenden Beispiel ist der Aufruf der Prozedur für die Schätzung der Parameter für das AFT Modell der Ereignisdaten gezeigt. Dabei beschreibt im Modell die Variable `aval` die Zeit bis zum Ergebnis bzw. Zensur der Patienten und die Indikatorvariable `event` beschreibt ob die Patienten bereits ein Ergebnis erlebt haben oder nicht (1=Ereignis, 0=Zensur). Durch die Option `DIST=` wird die zu verwendende Verteilung angegeben.

```
PROC LIFEREG DATA=<Datensatz>;
MODEL aval*event(0) = / DIST = <Verteilung>;
RUN;
```

Mit den geschätzten Parametern kann ein Datum eines Ereignisses vorhergesagt werden. Dazu werden Formeln (12) und (13) implementiert. Zuerst kann eine gleichverteilte Zufallszahl u mit Hilfe der Funktion `RAND('UNIFORM')` generiert werden. Die Umsetzung der Formel (13) erfolgt in drei Schritten. Zunächst wird das Innere der Quantil Funktion `x` $= 1 - [S(t_0) * u]$ durch die Verwendung der Survival Funktion `SDF()` berechnet. Im zweiten Schritt wird eine obere Grenze für die berechnete Zahl `x` eingebaut. Durch die Funktion `QUANTILE()` kann das Quantil der Verteilung unter der Verwendung von `x` und den geschätzten Parametern bestimmt werden. Für diesen Schritt wurde die obere Grenze von `x` benötigt, da bei der Funktion `QUANTILE()` nur Zahlen mit endlichen Dezimalzahlen verwendet werden können. Zuletzt wird die Zeit $t + t_0$ bis zum Ereignis auf die Studienzeit addiert, um ein simuliertes Ereignisdatum zu erhalten.

```
DATA <Name>;
  SET <dataset>;
  u = 1-RAND('UNIFORM');
```

```
  x = 1-(SDF("<Verteilung>", t_0, <Parameter>)*u);
  IF x GE 0.99999999999 THEN x = 0.99999999999;
  t = QUANTILE("<Verteilung>", x, <Parameter>);
  eventTimes = startDt + t;
RUN;
```

Die Simulation der Daten für Studienabbrechern erfolgt nach dem gleichen Prinzip.

4 Die SAS Makros

In Kapitel 2 wurde die Methodik gezeigt, mit der Rekrutierungsdaten, Ereignisdaten und Daten von Studienabbrechern simuliert werden können. Basierend auf dieser kann folgender Algorithmus zur Vorhersage vom „landmark event“ verwendet werden:

1. Definiere die Anzahl an Simulationen und führe die Schritte eins bis sechs für jede Simulation aus.
2. Falls der Rekrutierungsprozess noch nicht abgeschlossen ist, simuliere Rekrutierungsdaten für neue Patienten in der Zukunft, bis die geplante Anzahl an Studienteilnehmern erreicht ist.
3. Ordne jedem Patienten einen Status 0, 1 oder 2 zu. Die Status sind folgendermaßen definiert:
   0: Der Patient hat ein Ereignis oder ist aus der Studie ausgeschieden.
   1: Der Patient hat weder ein Ereignis noch ist er aus der Studie ausgeschieden.
   2: Der Patient wurde in der Zukunft rekrutiert.
4. Simuliere für jeden Patienten mit dem Status 1 oder 2 ein Ereignisdatum.
5. Füge Daten für Studienaussteiger hinzu:
   - Simuliere für jeden Patient mit Status 1 oder 2 ein Ausstiegsdatum. Falls das simulierte Ausstiegsdatum vor dem simulierten Ereignisdatum liegt, definiere den Patienten als Studienabbrecher.

   Oder
   - Definiere zufällig eine festgelegte Anzahl an Patienten als Aussteiger.
6. Ordne die Ereignisdaten in aufsteigender Reihenfolge.
7. Entnehme jeder Simulation das Datum des „landmark event“. Verwende den Median dieser Daten als vorhergesagtes Datum.

Dieser Algorithmus wurde in fünf Makros implementiert. Anhang A zeigt das Strukturdiagramm dieser Makros. Dabei sind die Pflichteingaben mit einem Stern gekennzeichnet.

Das Makro *EventPrediktion* organisiert die Kommunikation zwischen den anderen Makros in folgender Reihenfolge – *EventPrediction_Enrollment*, *EventPredicion_Event-Dates*, *EventPredicion_ETDates* und *EventPrediction_Output*. Diese Reihenfolge ist von großer Bedeutung, da das Ergebnis jedes Makros als Eingabeparameter für das darauffolgende Makro verwendet wird. Jedoch ist es auch möglich, die Makros separat aufzurufen.

4.1 Beschreibung der Makros

EventPrediction_Enrollment
EventPrediction_Enrollment erweitert für jede Simulation einen eingegebenen Datensatz mit neuen Rekrutierungsdaten, bis die geplante Anzahl von Patienten erreicht ist. Der Anwender hat unter anderem die Möglichkeit, die Anzahl der Simulationen zu definieren. Per Voreinstellung werden 1000 Simulationen durchgeführt.

EventPrediction_EventDates
EventPrediction_EventDates simuliert Ereignisdaten. Dazu benötigt das Makro den Datensatz, der in *EventPrediction_Enrollment* erstellt wurde, und mindestens ein „Vorhersagemodell". Dieses muss vom Anwender definiert werden. Ein Vorhersagemodell ist ein Modell, das aus einem Modell für Ereignisse und einem Modell für Studienabbrechern besteht. Das Ereignismodell gibt eine Verteilung an, die zur Simulation von Ereignisdaten verwendet werden soll. Entweder bestimmt der Anwender die Form- und Skalenparameter der Verteilung selbst oder die Parameter werden durch das Makro geschätzt. Die von den Makros unterstützenden Verteilungen für den Simulationsprozess sind die Exponentialverteilung, Logistische Verteilung, Logarithmische Normalverteilung und die Weibullverteilung.

EventPrediction_ETDates
Basierend auf den Datensätzen, die in *EventPrediction_EventDates* erstellt wurden, ersetzt das Makro *EventPrediction_ETDates* einen Teil der Ereignisdaten durch Daten von Studienabbrechern. Dieser Austauschprozess kann auf zwei verschiedene Arten stattfinden, wie in Schritt 5 des oben beschriebenen Algorithmus erläutert wurde. Für die erste Möglichkeit muss der Anwender eine Verteilung für den Simulationsprozess angeben. Dabei ist die Auswahl an Verteilungen die gleiche wie für das Ereignismodell. Jedoch können die Parameter nicht vom Anwender definiert werden und müssen vom Makro geschätzt werden. Bei der zweiten Möglichkeit wird vom Anwender eine Anzahl von gewünschten zukünftigen Studienabbrechern angegeben. Das Makro wählt nach dem Zufallsprinzip Patienten aus und definiert diese als Studienausscheider. Falls keine Studienabbrecher einbezogen werden sollen, kann die Zahl 0 angegeben werden.

Der Anwender kann ein Vorhersagemodell spezifizieren, indem er ein Ereignismodell mit einem Modell Studienabbrecher angibt. Dabei ist es möglich, mehrere Vorhersagemodelle in einem Programmlauf der Makros zu definieren. Das Makro führt für jedes Vorhersagemodell eine separate Simulation durch.

EventPrediction_Output
EventPrediction_Output bestimmt das Datum eines oder mehrerer „landmark events", wie in Schritt 7 des Algorithmus beschrieben und gibt eine rtf-Datei aus. Damit das Makro die „landmark events" bestimmen kann, muss der Anwender folgende Eingaben tätigen. Zunächst muss die geplante Anzahl von Ereignissen für die Finalanalyse angegeben werden. Weiterhin müssen die „landmark events" für die das Datum vorhergesagt

werden soll (z.B. mehrere Zwischen- und/oder die finale Analyse), eingegeben werden. Dazu werden entweder die jeweilige Anzahl der Ereignisse oder die Prozentzahl angegeben.

4.2 Die Ausgabe eines Beispiels

Input

Die dargestellte Ausgabe eines Beispiels wurde durch das Ausführen des Makros *EventPrediction* erstellt. Dazu wurden beim Anwenden folgende erforderlichen Eingaben getätigt. Zuerst wurde ein simulierter Datensatz eingelesen, welcher vier Variablen besaß. Die erste Variable enthielt die Rekrutierungsdaten und die zweite die Studienzeit der Patienten. Die dritte und vierte Variable waren zwei Indikatorfunktionen. In der ersten war dargestellt, ob ein Patient ein Ereignis hatte oder nicht und die zweite, ob der Patient aus der Studie ausgeschieden war oder nicht. Von 511 Patienten aus dem simulierten Datensatz hatten 251 ein Ereignis und 30 hatten die Studie abgebrochen.

Ziel war es, 700 Patienten zu rekrutieren. Weiterhin sollten Vorhersagen für 350 Ereignisse (Zwischenanalyse) und für 500 Ereignisse (finale Analyse) erfolgen. Der 22APR2016 wurde als Datum des Datenextraktes festgelegt und 10000 Simulationen für die Vorhersage ausgewählt. Zuletzt mussten die Vorhersagemodelle definiert werden. Diese sind in Tabelle 1 dargestellt. Die erste Spalte zeigt die Ereignismodelle und die zweite Spalte die Modelle für Studienabbrecher. Für die ersten beiden Vorhersagemodelle wurden die Parameter des Ereignismodells angegeben, da sie nicht durch das Makro geschätzt werden sollen.

Tabelle 1: Vorhersagemodelle eines Beispiels

Prediction Model		Event Model	
Event Model	Early Termination Model	Shape Parameter	Scale Parameter
Exponentail	Weibull		400
Weibull	40	0.8	380
Lognormal	Exponentail		
Exponentail	Exponentail		
Weibull	Lognormal		

Ausgabe

Mit dem Aufruf des Makros *EventPrediction* mit den oben beschriebenen Eingaben wurde für 350 und 500 Ereignisse jeweils ein Datum vorhergesagt. Ausgegeben wurde ein Datensatz und eine rtf-Datei mit vier Tabellen und zwei Abbildungen. Die Tabellen und die Grafiken der rtf-Datei werden im Folgenden einzeln betrachtet.

Tabelle 2 zeigt Information über die Eingaben an. In den Spalten eins bis fünf werden das erste Datum der Rekrutierung, das Datum des Datenextraktes, die Anzahl der zu verwendenden Simulationen, die geplante Anzahl der Patienten und die geplante Anzahl an Ereignissen für die finale Analyse angezeigt. Die Hauptmerkmale des Eingabedatensatzes werden in den Spalten sechs bis acht dargestellt. Dies sind die Anzahl der rekrutierten Patenten und wie viele davon ein Ereignis hatten und wie viele aus der Studie ausgeschieden waren.

Tabelle 2: Input des Anwenders

					User Input Data Set		
First Date	Cut Off Date	#Simulations	Target #Subjects	Target #Events	#Subjects	#Events	#Early Terminations
25-JAN-2014	22-APR-2016	10000	700	500	511	251	30

Tabelle 3 beschreibt die Parameter der Ereignismodelle. Jedem Vorhersagemodell wird eine ID zugeordnet. Die zweite Spalte zeigt die Verteilung der Ereignismodelle und in der dritten Spalte werden die ausgewählten Modelle für Studienabbrecher angezeigt. Zu sehen ist, dass die Modelle mit ID 1 oder 2 in Spalte vier ein „No“ zugeordnet wurde. Dies bedeutet, dass die Parameter der Ereignismodelle nicht vom Programm geschätzt wurden, sondern vom Anwender beim Aufruf des Makros eingegeben wurden. Die Parameter der anderen Vorhersagemodelle wurden vom Makro geschätzt.

Tabelle 3: Parameter der Ereignismodelle

ID	Event Model	Early Termination Model	Parameter Estimation	Shape Parameter	Scale Parameter
1	Exponential	Weibull	No	.	400
2	Weibull	40	No	0.8	380
3	Exponential	Exponential	Yes	.	377.58566
4	Lognormal	Exponential	Yes	5.5273031	1.5484958
5	Weibull	Lognormal	Yes	0.9421106	386.14514

Tabelle 4 stellt die Parameter und Statistiken der Modelle der Studienabbrecher dar. Die Spalten eins bis drei zeigen die ID, das Ereignismodell und das Modell der Studienabbrecher. Diese sind die gleichen wie in Tabelle 3 dargestellt. In der vierten Spalte wird mit „Yes“ bzw. „No“ angegeben, ob die Parameter des Studienabbrechermodells geschätzt wurden oder nicht. Bei einer Schätzung der Parameter, werden diese in den Spalten sechs und sieben angezeigt. Wie beschrieben, wird ein Patient als Studienabbrecher definiert, falls das Datum des Studienabbruchs vor dem Datum des Ereignisses liegt. Da es sich um simulierte Daten handelt, ist die Anzahl der Studienabbrüche in jeder Simulation unterschiedlich. Aus diesem Grund wird in den Spalten acht bis zwölf das Minimum, das 25%-Quantil, das 50%-Quantil, das 75%-Quantil und das Maximum der Studienabbrüche aus allen Simulationen dargestellt. Falls ein Vorhersagemodell eine Zahl als Modell der Studienabbrecher erhält (z.B. Modell mit ID 2) dann zeigt Spalte

sieben die Gesamtzahl der Studienabbrüche an (Anzahl der Studienabbrüche im Eingabedatensatz plus die hinzugefügten). Weiterhin wird Tabelle 4 nicht erstellt, falls alle Modelle der Studienabbrecher die Zahl 0 besitzen.

Tabelle 4: Parameter und Statistiken der Modelle der Studienabbrecher

						Statistics for Early Termination					
						Fix Number	Estimated #Subjects Early Terminated				
ID	Event Model	Early Termination Model	Parameter Estimation	Shape Parameter	Scale Parameter	Total #Early Terminations	Min	Quantil 25	Quantil 50	Quantil 75	Max
1	Exponential	Weibull	Yes	0.8514224	4899.3006	.	49	66	70	74	94
2	Weibull	40	No	.	.	70	.	.	.	.	.
3	Exponential	Exponential	Yes	.	3159.1333	.	54	70	75	79	101
4	Lognormal	Exponential	Yes	.	3159.1333	.	81	105	110	116	144
5	Weibull	Lognormal	Yes	9.2946844	2.6582848	.	45	59	63	67	88

In Tabelle 5 sind die vorhergesagten „landmark events“ dargestellt. Die Spalten eins, zwei und vier zeigen die IDs, die Ereignis- und Studienabbrechermodelle. Angaben ob die Parameter dieser Modelle geschätzt wurden, wird jeweils in den Spalten drei und fünf angegeben. Für jedes „landmark event“, welches vorhergesagt werden soll, werden zwei weitere Spalten generiert. Die erste dieser beiden Spalten gibt die Anzahl der Ereignisse an und die zweite Spalte das vorhergesagte Datum. Weiterhin wird über den beiden Spalten der Prozentsatz angegeben, welcher den Anteil der Ereignisse an der Anzahl der Ereignisse der finalen Analyse angibt. In unserem Beispiel sollten die Daten für 350 und 500 Ereignisse vorhergesagt werden. Somit wurden vier Spalten generiert.

Tabelle 5: Daten von vorhergesagten „landmark events“

	Event		Early Termination		70 Percent		100 Percent	
ID	Model	Parameter Estimation	Model	Parameter Estimation	#Events	Predicted Date	#Events	Predicted Date
1	Exponential	No	Weibull	Yes	350	11-OCT-2016	500	05-AUG-2017
2	Weibull	No	40	No	350	23-OCT-2016	500	18-OCT-2017
3	Exponential	Yes	Exponential	Yes	350	03-OCT-2016	500	17-JUL-2017
4	Lognormal	Yes	Exponential	Yes	350	28-OCT-2016	500	17-JAN-2018
5	Weibull	Yes	Lognormal	Yes	350	06-OCT-2016	500	30-JUL-2017

In Abbildung 1 sind zwei Grafiken unter einander dargestellt. Die obere Grafik stellt für jedes Vorhersagemodell die originalen und vorhergesagten Daten aller Ereignisse dar. Auf der x-Achse wird die Anzahl der Ereignisse dargestellt und auf der y-Achse das zugehörige Datum. Für das „landmark event“ der Zwischenanalyse ist eine vertikale Linie eingezeichnet. Die untere Grafik ist ein Auszug der oberen Grafik. Es werden nur die Ereignisse angezeigt, für die das Datum vorhergesagt wurde.

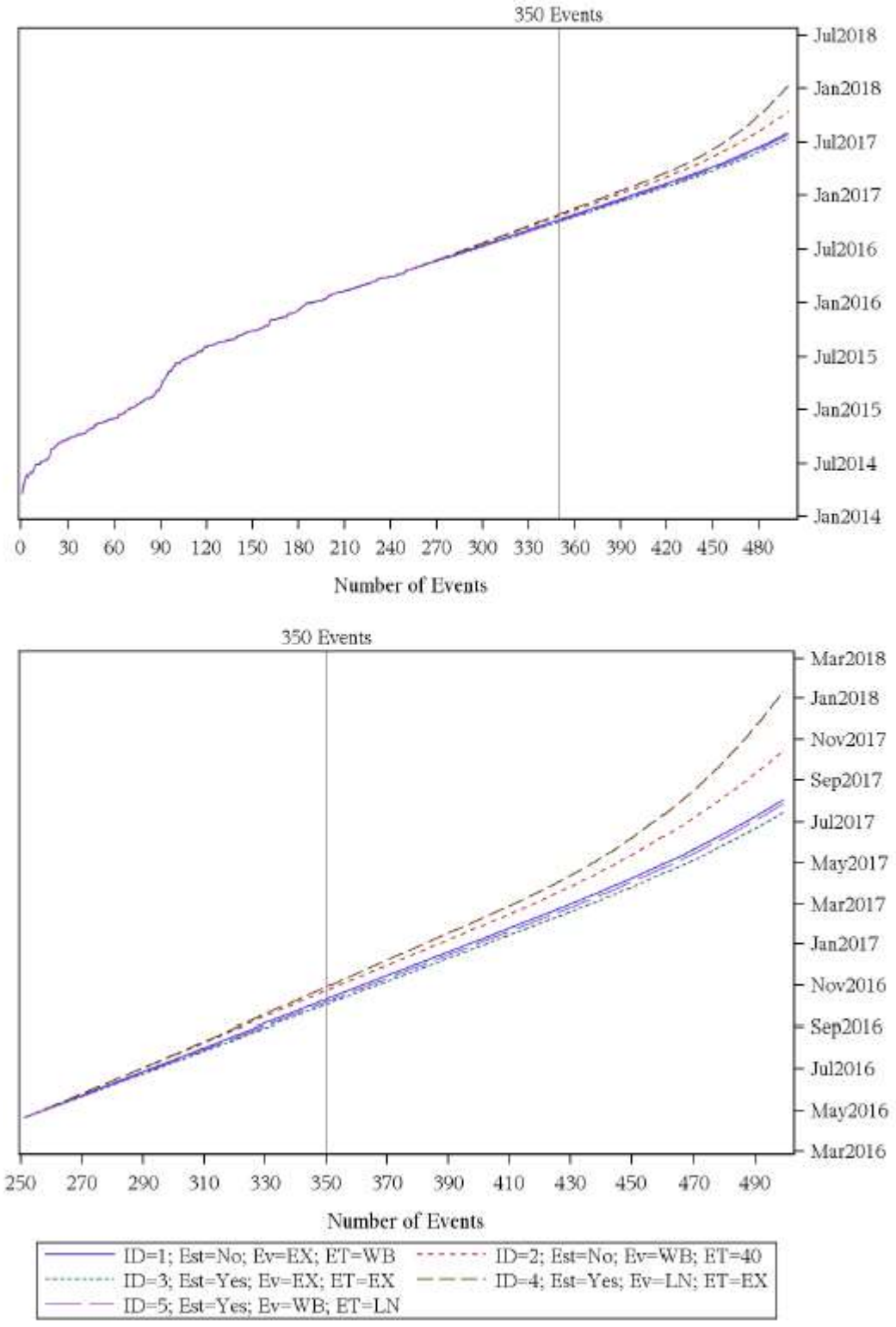

Abbildung 8: Grafische Darstellung der originalen Daten und der vorhergesagten Daten aller Ereignisse.

Die obere Grafik zeigt die Originaldaten und die vorhergesagten Daten aller Ereignisse für alle Vorhersagemodelle an. Die untere Grafik ist ein Auszug aus der oberen Grafik. Est: Parameterschätzung, Ev: Ereignismodell, ET: Studienabbrechermodell, EX: Exponentialverteilung, LL: Logistische Verteilung, LN: Logarithmische Normalverteilung, WB: Weibullverteilung

5 Zusammenfassung

Im Rahmen dieser Publikation wurden fünf Makros, welche in der Programmiersprache SAS implementiert wurden, vorgestellt. Mit diesen kann das Datum eines „landmark event“ vorhersagt werden. Die von den Makros verwendete Methode besteht aus drei Teilen: der Simulation von Rekrutierungsdaten, der Vorhersage der Daten von Ereignissen und von Studienabbrechern.

Zuerst wurde die Idee der bayesschen Inferenz behandelt. Anschließend wurde die Methode zur Simulation von Rekrutierungsdaten kurz erläutert. Danach wurde verkürzt das AFT Modell betrachtet und die Methode zur Vorhersage von Ereignissen und Studienabbrechern vorgestellt. Zum Schluss wurden die Makros eingeführt und die Ausgabe dieser präsentiert.

Literatur

[1] Allison, P. D. (2010). Survival Analysis Using SAS: A Practical Guide, Second Edition. SAS Publishing, 2nd edition.

[2] Bagiella, E. and Heitjan, D. F. (2001). Predicting analysis times in randomized clinical trials. Statistics in Medicine, 20(14):2055–2063.

[3] Collett, D. (2003). Modelling Survival Data in Medical Research, Second Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

[4] Gelman, A. (1995). Bayesian data analysis. Chapman & Hall, London New York.

[5] Hartung, J., Elpelt, B., and Klösener, K. (2009). Statistik: Lehr- und Handbuch der angewandten Statistik ; [mit zahlreichen durchgerechneten Beispielen]. Oldenbourg.

[6] Kahle, W. (2013). Zuverlässigkeitsanalyse und Qualitätssicherung. Oldenbourg Verlag, München Germany.

[7] Lawless, J. F. (2002). Statistical Models and Methods for Lifetime Data. Wiley-Interscience.

[8] Lee, E. T. (2003). Statistical Methods for Survival Data Analysis. Wiley-Interscience.

[9] Lesaffre, E. (2012). Bayesian biostatistics. Wiley, Chichester, West Sussex.

Anhang A Strukturdiagramm