

The Randomized Clinical Trial

Vorbereitungsmaterial

Laura Mittermeier

Seminar: Einblicke in verschiedene aktuelle Forschungsgebiete der Statistik

Wintersemester 2011/12

10.01.2012

1 Einleitung

Randomisierte klinische Studien sind das wichtigste Instrument in der medizinischen Forschung. Sie werden verwendet, um verschiedene Behandlungsalternativen gegenüber zu stellen. Dabei werden die Patienten zufällig einer Behandlungsmethode zugeordnet.

Durch die Randomisation wird sicher gestellt, dass die anderen möglichen Einflussgrößen, wie zum Beispiel Alter oder Geschlecht, gleichmäßig auf die einzelnen Gruppen aufgeteilt werden und somit durch das Gesetz der großen Zahlen der reine Behandlungseffekt untersucht wird. Die Zuteilung zu den Gruppen durch den behandelnden Arzt oder durch den Patienten selbst, könnte zu einer Verzerrung führen.

Problematisch dabei ist allerdings, dass die Methode der Randomisation einen zentralen Aspekt der Behandlung nicht beachtet, und zwar die Patient-Arzt-Beziehung. Ein Arzt muss immer die bestmögliche Behandlung für einen Patienten finden und dann diese anwenden. Da nun die Behandlung zufällig ausgesucht wird, kann es zu ethischen Problemen kommen (vgl. Abschnitt 3) (Harrington, 2000).

2 Studiendesign

Eine einfache Zufallsauswahl kann bei klinischen Studien zu unterschiedlich großen Gruppen oder zu einer Unausgeglichenheit der Einflussfaktoren führen. Daher wurden viele andere Randomisationsarten eingeführt.

Zelen (1974) schlug eine Randomisation von Blöcken vor. Dabei werden die Patienten, meist in der Reihenfolge, in der sie in die Studie aufgenommen werden, in gleichgroße Blöcke aufgeteilt. Dann werden innerhalb der Blöcke die verschiedenen Behandlungsmöglichkeiten zugeteilt. Man erhält zum Beispiel Blöcke mit je $2b$ Personen. Innerhalb jeden Blockes werden zufällig b Personen Behandlung A und b Personen Behandlung B zugeteilt. Jede der $\binom{2b}{b}$ Kombinationen ist gleich wahrscheinlich (Efron, 1971).

Um vor allem bei Studien mit kleinem Stichprobenumfang ausgeglichene Gruppen zu erhalten, hat Efron (1971) Biased-Coin-Designs entwickelt. Hierbei werden neue Studienteilnehmer mit einer bestimmten Wahrscheinlichkeit p den verschiedenen Gruppen zugeordnet. D gibt die Differenz der Gruppenößen an. Wenn ein neuer Patient in die Studie aufgenommen wird, erhält der Arzt einen Umschlag, in dem die Behandlungsgruppe angegeben ist, und zwar für die 3 unterschiedlichen Fälle $D = 0$, $D > 0$ und $D < 0$. Bei $D = 0$ ist die Wahrscheinlichkeit $p = 0.5$. Bei den anderen Fällen hat Efron (1971) gezeigt, dass die Wahrscheinlichkeit $p = \frac{2}{3}$ am besten geeignet ist.

Des weiteren kann man die Zuteilung auch anhand von Play-the-Winner-Designs vornehmen. Dabei wird ein Patient, wenn er in die Studie aufgenommen wird, der Behandlungsmethode zugeteilt, die

zu diesem Zeitpunkt die besseren Ergebnisse zeigt. Diese Methode mildert die ethischen Probleme, die bei klinischen Studien auftreten können, da der Arzt hier dem Patienten die im Moment beste Behandlung gibt.

3 Ethische Aspekte

3.1 Richtlinien

Die Frage, die sich nun stellt ist, unter welchen Bedingungen es moralisch vertretbar ist, eine Randomisierung vorzunehmen. Es wurden verschiedene Richtlinien eingeführt, mit dem Versuch, den Forschern und Ärzten Anhaltspunkte für die Durchführung solch klinischer Studien zu geben.

Die klassische Theorie setzt die „clinical equipoise“ voraus. Diese besagt laut Peto und Baigent (1998), dass ein Patient nur dann in eine klinische Studie aufgenommen werden darf, wenn sich der Arzt wirklich unsicher ist, welche der zu untersuchenden Behandlungsmethoden die erfolgreichste für diesen Patienten ist.

Eine weitere Richtlinie als ethische Grundlage von randomisierten klinischen Studien ist die Deklaration von Helsinki des Weltärztebundes (World Medical Association). Diese beschäftigt sich mit allgemeinen Grundsätzen für jede Art von medizinischer Forschung und mit weiteren Grundsätzen für die medizinische Forschung in Verbindung mit ärztlicher Versorgung (WMA, 2008). Dort ist unter anderem festgehalten, dass alle Ergebnisse der Studien veröffentlicht werden müssen (auch negative) und dass die Teilnahme an Studien freiwillig ist.

3.2 Probleme

Wie oben bereits angesprochen, sind Ärzte dazu verpflichtet, bei einem Patienten die Behandlung anzuwenden, die ihnen für diesen am besten erscheint. Überlässt man nun die Wahl der Behandlungsmethode dem Zufall, kann es zu ethischen Konflikten bei den Ärzten kommen.

Außerdem ist die Beteiligung an klinischen Studien oft so gering, dass die festgestellten Behandlungseffekte nicht mit Sicherheit den unterschiedlichen Behandlungen zugeordnet werden können. Die beobachteten Effekte könnten auch zufällige Schwankungen sein, die entstehen, wenn bei einem zu kleinen Stichprobenumfang die anderen Einflussgrößen nicht gleichmäßig auf die Gruppen aufgeteilt wurden. Damit wird aber der stillschweigende Vertrag mit den Studienteilnehmern gebrochen, der beinhaltet, dass die Teilnahme an klinischen Studien dazu führt, eine wichtige medizinische Frage zu klären (Harrington, 2000).

4 Statistische Analyse

4.1 Metrische Zielvariablen

Bei metrischen Größen, wie zum Beispiel Blutdruck oder Cholesterin, kann man mit verschiedenen Tests den Behandlungsunterschied messen. Dabei berechnet man die mittlere Veränderung in den verschiedenen Behandlungsgruppen und testet diese auf Gleichheit. Unter der Normalverteilungsannahme kann man hierfür den Student-t-Test verwenden. Wenn diese nicht gegeben ist, kann man anstelle des t-Tests den Wilcoxon-Rangsummen-Test verwenden. Bei mehr als zwei Behandlungsgruppen, kann man die Mittelwerte mit Hilfe der Varianzanalyse vergleichen. Man kann den Behandlungseffekt auch mit Hilfe eine (multiplen) linearen Regression schätzen. (Schumacher und Schulgen/Kristiansen, 2000)

4.2 Binäre Zielvariablen

Bei der Analyse von binären Zielgrößen gibt es verschiedene Ansätze. Binäre Zielgrößen sind zum Beispiel Tod/Überleben oder Heilung/Nicht-Heilung. Wenn man nur eine einzige Einflussgröße betrachten will, also das Medikament, kann man den sogenannten Log-Rang-Test (Ziegler, Lange und Bender, 2007) durchführen. Dieser Test funktioniert wie der χ^2 -Anpassungstest. Voraussetzung dabei ist, dass die Survivorfunktionen $S_i(t)$ der Gruppen proportional zueinander sind, d.h. dass gilt:

$$S_2(t) = S_1(t)^c$$

Die Null-Hypothese dieses Tests ist $H_0 : c = 1$ gegen die Alternative $H_1 : c \neq 1$. Es wird die Abweichung der beobachteten Todesfälle von den unter H_0 zu erwartenden Todesfällen, wenn das Medikament keinen Einfluss hat, betrachtet. Die Teststatistik für den Log-Rang-Test beim Vergleich zweier Gruppen sieht dann folgendermaßen aus:

$$LR = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

wobei O_i die beobachtete Anzahl an Todesfällen in Gruppe i und E_i die erwartete Anzahl an Todesfällen in Gruppe i ist.

Um E_i zu berechnen benötigt man die Anzahl r_{ij} der Personen aus Gruppe i, die zum Zeitpunkt j unter Risiko stehen. Bei insgesamt $r_{.j}$ Patienten unter Risiko zum Zeitpunkt j ergibt sich E_i unter der Nullhypothese durch Aufsummieren über alle Zeitpunkte j

$$E_i = \sum_j r_{ij} \frac{d_j}{r_{.j}}$$

wobei d_j die Anzahl der Patienten ist, die zum Zeitpunkt j ausgeschieden sind.

Wenn man diesen Test auf mehrere Behandlungsgruppen erweitern will, kann man einfach weitere Summanden hinzufügen. Bei k unterschiedlichen Behandlungsgruppen erhält man dann diese Teststatistik:

$$LR = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

LR ist asymptotisch χ^2 -verteilt mit $(k-1)$ -Freiheitsgraden.

Wenn man weitere Einflussgrößen in das Modell mit aufnehmen will, bietet sich das Proportional-Hazard-Modell von Cox an (Cox und Oakes, 1984 und Fahrmeir, Hamerle und Tutz, 1996). Für jedes Individuum i beobachtet man p verschiedene Einflussgrößen $\mathbf{x}'_i = (x_{1i}, \dots, x_{pi})$. Dieses Modell basiert auf der Hazardrate $\lambda(t|\mathbf{x})$, die in der Ereigniszeitanalyse verwendet wird.

$$\lambda(t) = \lim_{\substack{\Delta t \rightarrow 0 \\ \Delta t > 0}} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

wobei T der Zeitpunkt des Zustandswechsels ist und $[t, t + \Delta t)$ das betrachtete Intervall. Dabei gibt der Zähler die bedingte Wahrscheinlichkeit dafür an, dass das interessierende Ereignis T im Intervall $[t, t + \Delta t)$ eintritt, gegeben dieses Intervall wird erreicht.

Bei der Cox-Regression wird folgende Struktur der Hazardrate angenommen:

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}'\beta)$$

Dabei ist $\lambda_0(t)$ eine nicht genauer spezifizierte Grundhazardrate, die keinen Einschränkungen unterliegt, aber bei allen Untersuchungseinheiten gleich ist. Der lineare Prädiktor $\mathbf{x}'\beta$ enthält keine Konstante. Beim Proportional-Hazard-Modell wird die Proportionalität der Hazardraten angenommen, d.h. der Quotient der Hazardraten zweier Untersuchungseinheiten ist nicht von t abhängig (Fahrmeir, Hamerle und Tutz, 1996).

$$\frac{\lambda(t|\mathbf{x}_1)}{\lambda(t|\mathbf{x}_2)} = \exp((\mathbf{x}_1 - \mathbf{x}_2)'\beta)$$

Der Parameter β kann durch die bedingte Likelihood geschätzt werden (Cox, 1972).

Man nimmt an, dass die Grundhazardrate $\lambda_0(t)$ beliebig ist. Daher kann man in den Zeitintervallen, in denen keine Person ausscheidet, keine Information über β erhalten, da hier die Grundhazardrate $\lambda_0(t) = 0$ sein könnte. Daher wird nur die Menge $\{t_{(i)}\}$ aller Zeitpunkte $t_{(i)}$, zu denen Personen ausscheiden, betrachtet.

Die Wahrscheinlichkeit für das Individuum i zum Zeitpunkt $t_{(i)}$ auszuscheiden, gegeben der Menge der Personen unter Risiko $R(t_{(i)})$, ist:

$$P(T_i = t_{(i)} | R(t_{(i)})) = \frac{\exp(\mathbf{x}'_i \beta)}{\sum_{l \in R(t_{(i)})} \exp(\mathbf{x}'_l \beta)}$$

Daraus ergibt sich die gemeinsame log-Likelihood

$$l(\beta) = \sum_{i=1}^k \mathbf{x}'_i \beta - \sum_{i=1}^k \log\left(\sum_{l \in R(t_{(i)})} \exp(\mathbf{x}'_l \beta)\right)$$

Daraus ergibt sich direkt mit $\xi, \eta = 1, \dots, p$ als Index für die Kovariablen die Scorefunktion für β_ξ

$$S_\xi(\beta) = \frac{\partial l(\beta)}{\partial \beta_\xi} = \sum_{i=1}^k (x_{(\xi i)} - A_{(\xi i)}(\beta)), \text{ mit } A_{(\xi i)}(\beta) = \frac{\sum_{l \in R(t_{(i)})} x_{\xi l} \exp(\mathbf{x}'_l \beta)}{\sum_{l \in R(t_{(i)})} \exp(\mathbf{x}'_l \beta)} \quad (1)$$

Die beobachtete Fisher-Information für β_ξ und β_η ergibt sich dann durch die negative Ableitung der Scorefunktion

$$I_{\xi\eta}(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta_\xi \partial \beta_\eta} = \sum_{i=1}^k C_{\xi\eta i}(\beta) \quad (2)$$

wobei

$$C_{\xi\eta i}(\beta) = \frac{\sum_l x_{\xi l} x_{\eta l} \exp(\mathbf{x}'_l \beta)}{\sum_l \exp(\mathbf{x}'_l \beta)} - A_{(\xi i)}(\beta) A_{(\eta i)}(\beta)$$

Den Maximum-Likelihood-Schätzer für β erhält man durch iterative Verfahren, wie das Fisher-Scoring, unter Verwendung von (1) und (2). Der Startwert $\hat{\beta}^{(0)}$ kann beim Fisher-Scoring beliebig festgelegt werden. Die einzelnen Iterationen mit $k = 1, 2, \dots$ erhält man durch

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + I^{-1}(\hat{\beta}^{(k)}) S(\hat{\beta}^{(k)})$$

wobei $I^{-1}(\hat{\beta}^{(k)})$ die inverse Fisher-Information und $S(\hat{\beta}^{(k)})$ die Scorefunktion ist. Als Abbruchkriterium des Fisher-Scoring verwendet man mit $\varepsilon > 0$

$$\frac{\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\|}{\|\hat{\beta}^{(k)}\|} \leq \varepsilon$$

Wenn dieses Abbruchkriterium erreicht ist, wird keine weitere Iteration mehr durchgeführt und man erhält als Schätzer $\hat{\beta}^{(k)}$. Für diesen Schätzer gilt $\hat{\beta} \stackrel{a}{\sim} N(\beta, I^{-1}(\beta))$, wobei $I^{-1}(\beta)$ wieder die inverse Fisher-Information von β ist.

Ein Signifikanztest mit $H_0 : \beta = 0$ hat die Teststatistik $S(\mathbf{0})^T I(\mathbf{0})^{-1} S(\mathbf{0}) \stackrel{H_0}{\sim} \chi_p^2$.

Hier kann man den Parameter kausal interpretieren, da durch die Randomisierung nur noch der Behandlungseffekt gemessen wird. Bei $\beta = 0$ hat die Behandlung keinen Einfluss auf das Überleben. Wenn $\beta > 0$ hat die Behandlung einen negativen Einfluss und bei $\beta < 0$ einen positiven Einfluss auf das Überleben.

4.3 Non-Compliance

Bei der statistischen Analyse ist aber vor allem interessant wie man solche Studien auswertet, bei denen Non-Compliance vorliegt. Non-Compliance bedeutet, dass nicht alle Studienteilnehmer die Behandlung erhalten, der sie zugeteilt wurden. So kann es sein, dass ein Studienteilnehmer seine Medikamente nicht wie verordnet jeden Tag einnimmt oder zusätzlich andere Medikamente einnimmt. Wenn die Gründe sich nicht an die vorgegebene Behandlung zu halten, mit den Ergebnissen korreliert sind, wird die Randomisierung rückgängig gemacht und man kann nicht mehr davon ausgehen, dass die festgestellten Gruppenunterschiede auf die unterschiedliche Behandlung zurück gehen.

Um solche Probleme bei der Analyse miteinzubeziehen, gibt es verschiedene Methoden.

Wenn man herausfinden kann, welcher Patient welche Behandlung erhalten hat, kann man den Behandlungseffekt mit einem zeit-abhängigen Cox-Modell schätzen. Dies ist eine Erweiterung des oben genannten Modells, zu den zeit-unabhängigen Kovariablen x_i , hat man hier die Behandlung $D(t)$ in Abhängigkeit der Zeit gegeben.

Die Hazardrate $\lambda(t|H(t))$ ist gegeben durch

$$\lambda(t|H(t)) = \lambda_0(t) \exp(\beta D(t))$$

mit $H(t) = \{D(u); 0 \leq u \leq t\}$, also der Behandlungsgeschichte bis zum Zeitpunkt t , die angibt, zu welchen Zeitpunkten der Studienteilnehmer die Medikamente eingenommen hat. $D(t)$ ist der Behandlungsindikator, also $D(t) = 1$, wenn das Individuum zum Zeitpunkt t seine Medikamente genommen hat. Der Parameter β entspricht dem Behandlungseffekt, $\beta = 0$ bedeutet daher, dass kein Behandlungseffekt vorliegt. Den Schätzer von β erhält man ähnlich wie oben.

Wenn die Non-Compliance nicht zufällig ist, ist bei der Cox-Regression allerdings keine kausale Interpretation von β möglich, da die Studienteilnehmer, die sich nicht an die vorgegebene Behandlung halten andere Eigenschaften haben könnten, als die Studienteilnehmer, die sich an die Behandlung halten. Diese Eigenschaften könnten sich auch auf den Verlauf der Krankheit auswirken. Wenn sich alle Studienteilnehmer an die vorgegebene Behandlung halten, wird durch die Randomisierung eine kausale Interpretation möglich, was bei Vorliegen von nicht-zufälliger Non-Compliance nicht der Fall ist.

Als weitere Möglichkeit bietet sich bei Non-Compliance die Intent-to-treat-Analyse an. Dabei wird der Behandlungseffekt aufgrund der Behandlungsgruppe geschätzt und nicht auf der tatsächlich erhaltenen Behandlung.

Dabei erhält man folgende Hazardrate:

$$\lambda(t|R) = \lambda_0(t)\exp(\beta R)$$

wobei R der Indikator für die zugeteilte Behandlungsgruppe ist.

Auch hier kann man β wie oben schätzen. Wenn man bei der Schätzung $\beta = 0$ erhält, liegt wiederum kein Behandlungseffekt vor. Bei der Intent-to-treat-Analyse ist durch die Randomisierung die Vergleichbarkeit der Gruppen gewährleistet. Allerdings misst β hier nicht den wahren biologischen Effekt in der Grundgesamtheit. Der Behandlungseffekt in der Studie wird kleiner, je mehr Patienten sich nicht an die zugeteilte Behandlung halten. Wenn sich nun das Medikament als erfolgreich herausgestellt hat und deswegen zugelassen wird, gibt es in der Grundgesamtheit mehr Menschen, die sich an die Behandlung halten. Der Effekt des Medikaments ist nun nicht mehr derselbe wie in der Studie (Robins and Tsiatis, 1991).

Literatur

- Cox, D. R. (1972). Regression models and life-tables, *Journal of the Royal statistical Society. Series B (Methodological)* **34**: 187-220.
- Cox, D. R. und Oakes, D. (1984). *Analysis of Survival Data*, Chapman and Hall.
- Efron, B. (1971). Forcing a Sequential Experiment to be Balanced, *Biometrika* **58**(3): 403-417.
- Fahrmeir, L., Hamerle, A. und Tutz, G. (1996). *Multivariate statistische Verfahren*, Walter de Gruyter.
- Harrington, D. P. (2000). The randomized clinical trial, *Journal of the American Statistical Association* **95**(449): 312-315.
- Peto, R. und Baigent, C. (1998). Trials: The next 50 years, *British Medical Journal* **317**: 1170-1171.
- Robins, J. M. und Tsiatis, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models, *Communications in Statistics A - Theory and Methods* **20**: 2609-2631.
- Schumacher, M. und Schulgen/Kristiansen, G. (2000). *Methodik klinischer Studien*, Springer-Verlag.
- WMA (2008). Declaration of Helsinki. **URL**: <http://www.wma.net/en/30publications/10policies/b3/>
- Zelen, M. (1974). The randomization and stratification of patients to clinical trails, *Journal of Chronic Diseases* **27**: 365-375.
- Ziegler, A., Lange, S. und Bender, R. (2007). Überlebenszeitanalyse: Der Log-Rang-Test, *Deutsche medizinische Wochenschrift* **S01**: e39-e41.