# Frequentistische und Bayesianische Analyse von Überlebenszeiten in klinischen Studien – Wie wirkt sich die Nutzung von Vorwissen auf die Power aus?

**Joachim Gerß**

Institut für Medizinische Informatik und Biomathematik, Universität Münster

joachim.gerss@ukmuenster.de

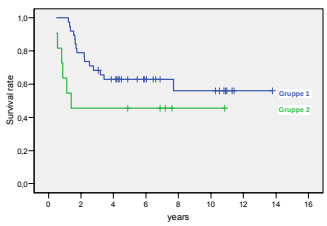
## Einleitung und Fragestellung

Klinische Studien stellen ein Anwendungsgebiet der Biometrie dar, in dem heute fast ausschließlich Methoden der frequentistischen Statistik eingesetzt werden. Die entsprechenden Gütekriterien haben sich als Standard etabliert und sind behördlich anerkannt. In den letzten Jahren werden jedoch in stärkerem Ausmaß die Vorteile alternativer Ansätze erkannt und diskutiert. So wurden sowohl von US-amerikanischer als auch von europäischer Seite Initiaven zur Effizienzsteigerung der medizinischen Forschung gestartet, die Critical Path Initiative der U.S. Food and Drug Administration (FDA) sowie die europäische Innovative Medicines Initiative. Im Rahmen dieser Initiativen werden Bayes-Verfahren als möglicher Ansatz angesehen, der zumindest ergänzend zu klassischen Verfahren eingesetzt werden kann. Dies wird hauptsächlich damit begründet, dass es im Rahmen des Bayes-Ansatzes möglich ist, zusätzliches Vorwissen in die Auswertung der Studiendaten einzubringen. Daraus ergibt sich die naheliegende Hoffnung, bei gleichem Aufwand die Aussagekraft der Ergebnisse zu steigern bzw. andererseits bei gleicher Aussagekraft den erforderlichen Aufwand zu reduzieren, etwa in Form einer geringeren Anzahl rekrutierter Patienten. Inwieweit lassen sich derartige intuitive Überlegungen mit Fakten belegen? Um dieser Frage nachzugehen, werden im Rahmen eines univariaten Cox-Regressionsmodells zur Analyse von Überlebenszeiten der frequentistische Ansatz sowie ein entsprechender Bayes-Ansatz vorgestellt und hinsichtlich klassischer Gütekriterien wie der Power miteinander verglichen.

## Material und Methoden

Als grundlegende Problemsituation wird der Vergleich zweier Patientengruppen hinsichtlich der Überlebens-zeit behandelt.

Der frequentistische Ansatz zur Datenauswertung im Rahmen eines univariaten Cox-Modells verläuft standardmäßig und liefert neben dem Ergebnis des Signifikanztests zum Vergleich beider Überlebens-kurven eine Intervallschätzung des Hazard Ratios HR=2.227 (95% KI 0.947-5.238, p=0.0990).



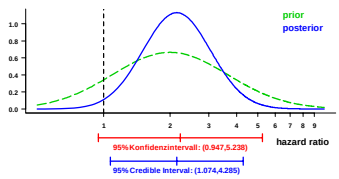

Um einen entsprechenden Bayes-Ansatz herzuleiten, nutzt man die Tatsache, dass der logarithmierte Schätzer des Hazard Ratios bei gegebener Anzahl eingetretener Ereignisse $n_e$ erwartungstreu und approximativ normalverteilt ist mit der Varianz $4/n_e$ [1,2]. Auf der Grundlage dieses Resultats lässt sich ein zweistufiges Bayes-Modell entwickeln, dass mit einer unterstellten A-priori-Verteilung des Hazard Ratios nach Hinzuziehung der empirischen Daten eine A-posteriori-Verteilung liefert. Aus der A-Posteriori-Verteilung ergibt sich direkt ein so genanntes Credible-Intervall, das ein Bayesianisches Gegenstück zum klassischen Konfidenzintervall darstellt.

Formal schreibt sich das Normal-Normal-Modell wie folgt. Dabei bezeichnet $\theta=\ln(HR)$ das logarithmierte Hazard Ratio und $\hat{\theta}$ die entsprechende empirische Schätzung auf der Basis eines angepassten Cox-Modells.

$$\begin{cases} \text{Datenmodell: } \hat{\theta}|\theta \overset{asy}{\sim} N\left(\theta,\frac{4}{n_e}\right), \text{ mit } n_e = \text{Anzahl eingetretener Ereignisse} \\ \text{A-priori-Verteilung: } \theta \sim N(\mu_0,\sigma_0^2) \end{cases}$$

Die Kombination der A-priori-Verteilung mit den realisierten Daten liefert als A-posteriori-Verteilung ebenfalls eine Normalverteilung mit

$$\theta|\hat{\theta} \sim N\left(\frac{\frac{1}{\sigma_0^2}\mu_0+\frac{n_e}{4}\hat{\theta}}{\frac{1}{\sigma_0^2}+\frac{n_e}{4}}, \frac{1}{\frac{1}{\sigma_0^2}+\frac{n_e}{4}}\right)$$




**Power des frequentistischen und des Bayes-Ansatzes**

Gegeben sei ein einseitiges Testproblem, mit dem die Frage geklärt werden soll, ob eine zweier Patientengruppen über eine signifikant ungünstigere Prognose verfügt als die Vergleichsgruppe.

$H_0$: HR≤1 gegen $H_1$: HR>1, d.h. $H_0$: θ:=ln(HR)≤0 gegen $H_1$: θ>0

Die approximative Power des frequentistischen Ansatzes ergibt sich direkt aus dem obigen Resultat einer approximativen Normalverteilung des logarithmierten HR-Schätzers $\hat{\theta}|\theta \overset{asy}{\sim} N\left(\theta,\frac{4}{n_e}\right)$.

$$\text{Power}_{klassisch} = P\left(\frac{\hat{\theta}}{\sqrt{4/n_e}} > u_{1-\alpha}|\theta\right) = 1-\Phi\left(u_{1-\alpha}-\frac{\theta}{\sqrt{4/n_e}}\right) = \Phi\left(\frac{1}{2}\cdot\sqrt{n_e}\cdot\theta-u_{1-\alpha}\right)$$

Im Rahmen des Bayes-Ansatzes ergibt sich ein Gegenstück zum Signifikanztest wie folgt. Man sieht den „Bayesianischen Test" als signifikant an, falls 95% der A-Posteriori-Wahrscheinlichkeitsmasse des Hazard-Ratios jenseits des neutralen Wertes 1 liegt [3]. Formal ist das der Fall, wenn gilt:

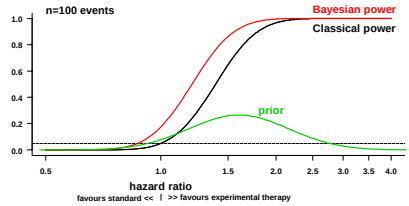
$$\begin{aligned} & P\left(\theta>0|\hat{\theta}\right) > 1-\alpha \\ \Leftrightarrow\ & P\left(\theta<0|\hat{\theta}\right) < \alpha \\ \Leftrightarrow\ & \Phi\left(\frac{-\frac{1}{\sigma_0^2}\cdot\mu_0-\frac{n_e}{4}\cdot\hat{\theta}}{\sqrt{\frac{1}{\sigma_0^2}+\frac{n_e}{4}}}\right) < \alpha \\ \Leftrightarrow\ & \hat{\theta} > \frac{1}{n_e}\cdot\left(-\frac{4}{\sigma_0^2}\cdot\mu_0-2u_\alpha\cdot\sqrt{\frac{4}{\sigma_0^2}+n_e}\right) \end{aligned}$$

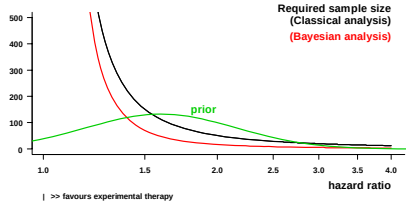
Die „Power" des Bayes-Ansatzes entspricht dann der Wahrscheinlichkeit dieses Ereignisses bei gegebenem θ. Sie kann berechnet werden, indem man das obige Resultat $\hat{\theta}|\theta \overset{asy}{\sim} N\left(\theta,\frac{4}{n_e}\right)$ nutzt:

$$\text{Power}_{Bayes} = P\left[\hat{\theta} > \frac{1}{n_e}\cdot\left(-\frac{4}{\sigma_0^2}\cdot\mu_0-2u_\alpha\cdot\sqrt{\frac{4}{\sigma_0^2}+n_e}\right)\middle|\theta\right] = \Phi\left(\frac{1}{2}\cdot\sqrt{n_e}\cdot\theta+\frac{2\cdot\mu_0}{\sigma_0^2\cdot\sqrt{n_e}}+u_\alpha\cdot\sqrt{\frac{4}{n_e\cdot\sigma_0^2}+1}\right)$$

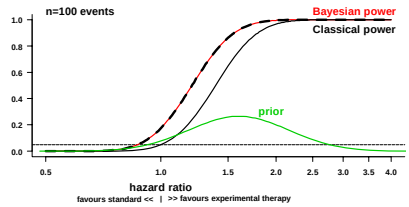
## Ergebnisse der Poweranalyse

Wie zu erwarten zeigt sich beim Vergleich des frequentistischen Signifikanztests mit dessen Bayesianischer Variante, dass die zusätzliche Nutzung von Vorwissen zu einer Steigerung der Power führt. Übertragen auf eine Aussage zur notwendigen Fallzahl kann dies unter realistischen Annahmen bedeuten, dass zur Gewährleistung einer bestimmten Power nur etwa halb so viele Patienten in eine Studie eingeschlossen werden müssen.




Auf der anderen Seite zeigen die Powerfunktionen beider Ansätze jedoch, dass der Powergewinn des Bayes-Ansatzes auf Kosten einer vergrößerten Wahrscheinlichkeit eines Fehlers 1. Art geschieht. Dies kann natürlich gerade in konfirmativen Auswertungen nicht akzeptiert werden, in denen die Kontrolle des Fehlers 1. Art als elementares Gütekriterium gilt. Aber auch in dem Fall, dass man die Anhebung des Fehlers 1. Art akzeptieren würde (etwa in explorativen Auswertungen), muss der obige Powergewinn in jedem Fall dementsprechend relativiert werden. Dies kann geschehen, indem man das Signifikanzniveau des klassischen Tests anpasst, so dass es dem vergrößerten Fehler 1. Art des Bayes-Tests entspricht. Dabei zeigt sich ein erstaunliches Resultat, dass nämlich in diesem Fall beide Powerfunktionen identisch sind. In der Schlussfolgerung gilt damit, dass der Bayes-Ansatz keinerlei echten Gewinn im Vergleich zum klassischen Ansatz mit sich bringt. Zwar resultiert aus der Nutzung von Vorwissen eine gesteigerte Power, den gleichen Powergewinn kann allerdings der klassische Ansatz leisten, wenn nur das Signifikanzniveau entsprechend angepasst wird.




## Diskussion

Bei dem obigen Vergleich des frequentistischen mit dem Bayes-Ansatz ergibt sich trotz Nutzung von zusätzlicher Information keine echte Überlegenheit des Bayes-Ansatzes. Bei diesem Resultat ist allerdings zu berücksichtigen, welche Art Kriterien als Vergleichsmaßstab verwendet wurden. Der Fehler 1. Art und die Power sind klassische Konzepte. Der frequentistische Signifikanztest ist unmittelbar auf die Optimierung dieser Konzepte zugeschnitten, womit ein relativ gutes Abschneiden natürlich zu erwarten ist. Das schlechtere Abschneiden des Bayes-Ansatzes ist möglicherweise dadurch zu erklären, dass der Bayes-Ansatz eben nicht auf der Grundlage klassischer, sondern alternativer Gütekriterien hergeleitet wird. Unter diesem Blickwinkel ist es intuitiv nachvollziehbar und nicht verwunderlich, dass der Bayes-Ansatz seine Vorteile einbüßt, wenn er mit den „falschen" Gütekriterien beurteilt wird. Sollte der Vergleich beider Verfahren also mit anderen als den klassischen Gütekriterien erfolgen? Dem ist entgegenzuhalten, dass die klassischen Kriterien sich in der Vergangenheit in der wissenschaftlichen Forschung bewährt haben und etabliert sind. Diese Tatsache rechtfertigt es, von einem neuartigen Verfahren tatsächlich zu fordern, dass es sich nach klassischen Kriterien an den Standardverfahren messen muss und ggf. gegenüber diesen durchsetzt. So bleibt als Fazit die Feststellung, dass das vorgestellte Bayes-Verfahren zumindest in konfirmativen Auswertungen nicht als vorteilhaft gegenüber dem frequentistischen Ansatz anzusehen ist.

## Literatur

1 Schoenfeld DA: The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 1981; 68: 316-9.

2 Schoenfeld DA: Sample-size formula for the proportional-hazards regression model. Biometrics 1983; 39: 499-503.

3 Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. New York: Wiley; 2004.