Wer evidenzbasiert argumentiert, bezieht sich auf den aktuellen Wissensstand aus Studien, nicht nur auf persönliche Erfahrungen oder auf die Meinung von Experten. Das bloße Zitat einer Studie ist aber noch keine evidenzbasierte Argumentation. Studien bieten keine unumstößlichen Wahrheiten, sondern Ergebnisse statistischer Analysen. Jeder Studientyp hat dabei spezifische Stärken und Schwächen. Diese Artikelreihe konzentriert sich auf klinische Studien, die experimentell Wirksamkeit und Verträglichkeit von Arzneimitteln prüfen. In kurzen Beiträgen möchten wir Sie mit dem nötigen „Werkzeug" ausstatten, um klinische Studien zu Arzneimitteln kritisch zu lesen und sich Ihre eigene, evidenzbasierte Meinung zu bilden.

Klinische Studien zu Arzneimitteln – Wo ist der Haken?

Der Zahn der Zeit – Hazard Ratio und Kaplan-Meier-Kurven

Einhart, N.

Mathes, T.

Rückblick: Die Wirksamkeit von Arzneimitteln lässt sich am verlässlichsten durch randomisierte kontrollierte Studien beurteilen (Studientypen – ohne Kontrolle geht nichts; Randomisierung – der reine Zufall). Um Verzerrungen zu vermeiden, sollte die Datenanalyse auch Patienten mit Protokollverstößen berücksichtigen (Was machen wir mit den „Abtrünnigen"?) und fehlende Werte auf geeignete Weise ersetzen (Verloren, aber nicht unersetzlich). Viele patientenrelevante Endpunkte sind „dichotome Endpunkte", die entweder vorhanden oder nicht vorhanden sind (z. B. Herzinfarkt ja/nein). Dichotome Endpunkte werden in Studien zumeist als Risiko beschrieben. Dabei klingt die relative Risikoreduktion in der Regel sehr viel eindrucksvoller als die absolute Risikoreduktion (Wie groß ist der Nutzen?).

Auf ihrer letzten Urlaubsreise haben Sie von der wundersamen Kraft der Vitalonga-Knolle gehört: Wer sie täglich kaue, der werde seine Altersgenossen alle überleben. Zurück in Deutschland wollen Sie untersuchen, ob Extrakte der Knolle Vitalonga tatsächlich die Sterblichkeit betagter Patienten reduzieren. Dafür randomisieren Sie 20 Patienten des Geburtsjahrgangs 1940 1:1 zu Vitalonga-Extrakten bzw. Placebo und beobachten alle Patienten über 800 Tage.[1] Sie stellen fest, dass zu Studienende in beiden Armen noch ein Teilnehmer lebt. Die Wahrscheinlichkeit, bis Tag 800 zu überleben, betrug also in beiden Armen 10 % bzw. (wie Ihr Statistiker sagen würde) 0,1.[2] Entsprechend lag das relative Risiko[3], unter Vitalonga-Extrakt zu überleben, bei exakt 1,0 (0,1 geteilt durch 0,1) – die Vitalonga-Extrakte hatten keinerlei Einfluss auf die Wahrscheinlichkeit, nach 800 Tagen noch am Leben zu sein (zum relativen Risiko).

Sie greifen enttäuscht zum Hörer und erklären Ihrem Statistiker, dass er sich die Ergebnisse gar nicht erst anschauen müsse. Doch Ihr Statistiker findet zu Ihrer Verblüffung die Ergebnisse nicht so eindeutig wie Sie: Ihr Statistiker erklärt Ihnen, dass man bei dieser Studie eine Ereigniszeitanalyse durchführen müsse, denn von Interesse sei ja nicht nur, wie viele Patienten gestorben seien, sondern auch zu welchem Zeitpunkt der Studie – es sei durchaus denkbar, dass sich die Überlebenszeit innerhalb des Studienzeitraums zwischen den Armen unterschied. Dafür müsse er genau wissen, nach welcher Beobachtungsdauer welcher Patient verstorben sei. Sie schicken Ihrem Statistiker die gewünschte Liste (Tabelle 1) und erhalten am nächsten Tag eine treppenartige Grafik zurück (Abbildung 1) sowie die Aussage, die Hazard Ratio liege bei 0,470.

[1] Bei diesem Gedankenexperiment klammern wir statistische Fragen zur Errechnung der geeigneten Stichprobengröße ebenso aus wie die erforderliche Genehmigung Ihrer Studie durch die zuständige Bundesoberbehörde sowie das notwendige positive Votum der Ethik-Kommission in Ihrem Bundesland.

[2] Diese Berechnung des relativen Risikos berücksichtigt nicht die fehlenden Daten aufgrund von Studienabbrüchen. Das Studienergebnis wird stark davon beeinflusst, ob die „verlorenen" Studienteilnehmer als Überlebende oder Verstorbene gewertet werden (zum Umgang mit fehlenden Daten).

[3] Die Wahrscheinlichkeit für das Eintreten bestimmter Ereignisse wird in klinischen Studien häufig als „Risiko" beschrieben – unabhängig davon, ob diese Ereignisse für den Patienten negativ (Myokardinfarkt) oder positiv (Überleben) sind.

Tabelle 1: Vitalstatus und Follow-up Zeiten der Studienteilnehmer der Vitalonga-Studie

Vitalonga-Extrakte		Placebo	
Zeit (Tage)	Status	Zeit (Tage)	Status
353	tot	59	tot
365	tot	115	tot
377	zensiert (ausgeschieden)	156	tot
421	zensiert (ausgeschieden)	268	tot
464	tot	329	tot
475	tot	431	tot
563	tot	448	zensiert (ausgeschieden)
744	zensiert (ausgeschieden)	477	zensiert (ausgeschieden)
769	zensiert (ausgeschieden)	638	tot
800	zensiert (Studienende)	800	zensiert (Studienende)

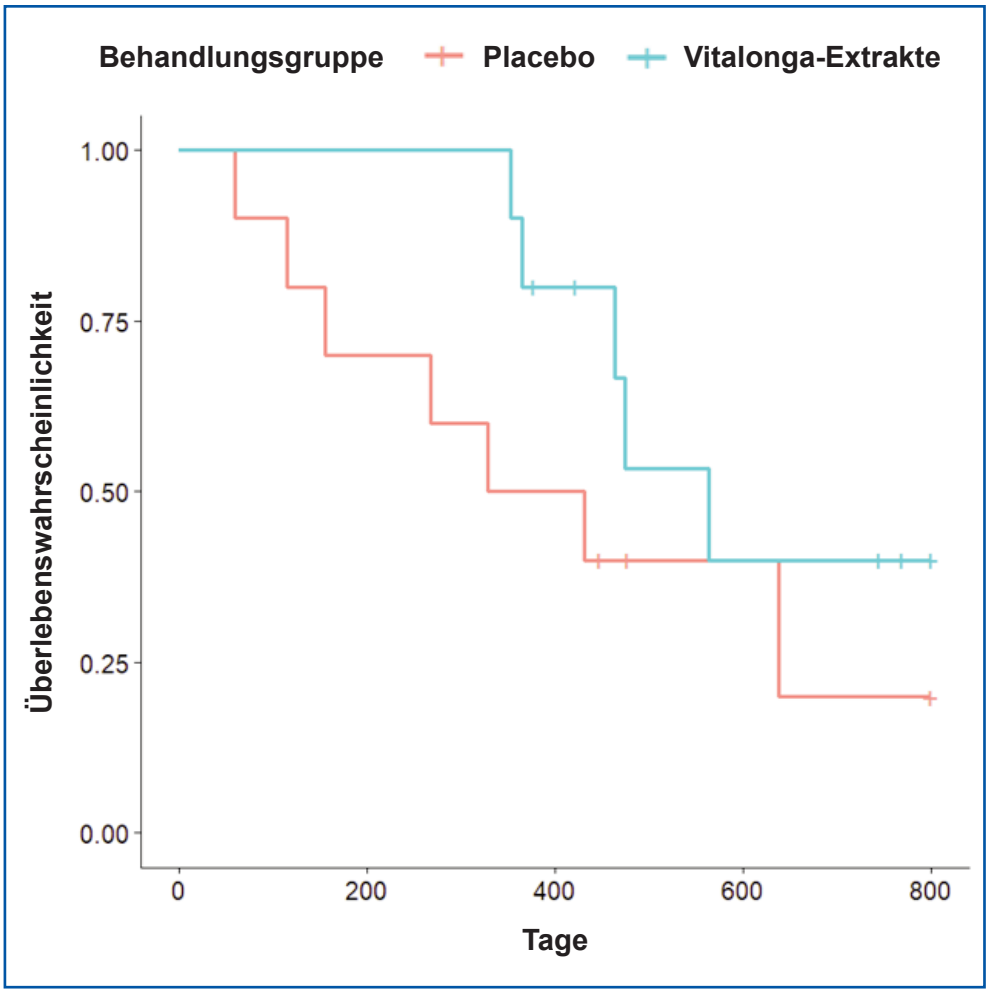

Abbildung 1: Kaplan-Meier-Kurve der Vitalonga-Studie

Die treppenartige Grafik stellt eine sogenannte **Kaplan-Meier-Kurve** dar. Die *x-Achse* repräsentiert die Beobachtungszeit vom Studienbeginn bis zum Studienende. Es ist wichtig, bei der Interpretation der Kurve auf die Einheit der Zeitachse zu achten (Tage, Wochen, Monate oder Jahre?). Auf der *y-Achse* wird der Anteil der Personen angegeben, bei denen das untersuchte Ereignis eingetreten ist. Die Kaplan-Meier-Kurve wird auch Überlebenskurve genannt, weil mit ihr häufig – wie in unserem Beispiel – das Ereignis „Überleben" für jeden Probanden in Abhängigkeit von der Zeit dargestellt wird. In diesem Fall beginnt die Kurve bei 1,0 bzw. 100 % (alle Teilnehmer leben zu Studienbeginn) und sinkt im zeitlichen Verlauf ab. Mit Kaplan-Meier-Kurven können auch andere Ereignisse dargestellt werden, z. B. das Auftreten von Herzinfarkten oder die Notwendigkeit von Rezidiv-Operationen. In diesem Fall beginnt die Kurve bei 0 % und steigt im Studienverlauf an. Jede Stufe in der Kurve bedeutet, dass zu diesem Zeitpunkt ein oder mehrere Ereignisse eingetreten sind. Die Höhe der Kurve ergibt sich aus dem Verhältnis zwischen Teilnehmerzahl und Anzahl der Ereignisse: Je größer die Teilnehmerzahl ist, desto kleiner ist die Stufe, die durch ein einzelnes Ereignis verursacht wird. In großen Studien sieht die Kaplan-Meier-Kurve deshalb tatsächlich wie eine (leicht gezackte) Kurve aus und nicht wie eine unregelmäßig gebaute Treppe.

In der Kaplan-Meier-Kurve der Vitalonga-Studie gibt es eine erste Stufe in der rot gefärbten Placebo-Kurve an Tag 59: Zu diesem Zeitpunkt ist der erste Teilnehmer unter Placebo verstorben. Sie können aus der Kurve ablesen, dass die Wahrscheinlichkeit, bis Tag 59 zu überleben, bei 90 % bzw. 0,9 (9 geteilt durch 10) lag. Nach jedem Ereignis beginnt ein neues Beobachtungsintervall, für das separat die Überlebenswahrscheinlichkeit berechnet wird. Die Beobachtungszeit ist dadurch in viele kleine Intervalle aufgeteilt, die jeweils von Ereigniszeitpunkt zu Ereigniszeitpunkt reichen. Das zweite Intervall reicht im Placeboarm von Tag 60 bis Tag 115. Ein Risiko, in diesem Intervall zu versterben, haben nur diejenigen Teilnehmer, die an Tag 60 noch leben, in unserem Beispiel also neun Personen. Diese Teilnehmer stehen, wie der Statistiker sagt, *„unter Risiko"*. Außerdem können offensichtlich nur diejenigen Patienten in die Analyse einbezogen werden, deren Vitalstatus zum Ende des Intervalls bekannt ist. Wer aus der Studie ausscheidet (z. B. weil er sein Einverständnis zur Teilnahme zurückzieht), wird als *„zensiert"* bezeichnet und verschwindet aus der Gruppe „unter Risiko". Jeder Patient geht somit nur mit seiner individuellen Beobachtungszeit in die Analyse ein. Zensierungen im Studienverlauf werden oftmals durch kleine senkrechte Striche visualisiert. In Tabelle 1 sehen Sie, dass auch alle Teilnehmer, bei denen am Studienende das untersuchte Ereignis nicht eingetreten ist, als „zensiert" bezeichnet werden – da am Studienende die Datenerhebung beendet wird, können wir logischerweise nicht wissen, was danach passierte.

In der Vitalonga-Studie wurden im Studienverlauf zwei Teilnehmer aus dem Placebo-Arm und vier Teilnehmer aus dem Vitalonga-Arm zensiert. Wir können keine Aussage darüber treffen, ob diese „verlorenen" Patienten zu Tag 800 noch leben oder nicht. Ihre Daten fließen deshalb nicht in die Berechnung des relativen Risikos ein – es können nur mittels Imputationsverfahren Annahmen darüber getroffen werden, ob die „verlorenen" Patienten verstorben sind oder überlebt haben. In die Ereigniszeitanalyse können wir dagegen alle bis zum Studienabbruch erhobenen Daten einbeziehen, zum Beispiel die Information, dass im Vitalonga-Arm zwei Patienten noch an Tag 743 bzw. 768 lebten. Ereigniszeitanalysen

Berechnung von Überlebenswahrscheinlichkeiten nach der Kaplan-Meier-Methode

Die Überlebenswahrscheinlichkeit wird immer dann neu berechnet, wenn ein Patient stirbt. Die Wahrscheinlichkeit, ein bestimmtes Intervall zu überleben, ist der Quotient aus der Anzahl der am Ende des Intervalls Lebenden und der Patientenzahl unter Risiko (d. h. der Patienten, die unmittelbar vor dem Ereignis noch leben und nicht zensiert sind).

Rechenbeispiele aus dem Vitalonga-Arm:

Das erste Intervall dauert vom Studienbeginn bis zum ersten Todesfall an Tag 353. Die Überlebenswahrscheinlichkeit bis Tag 353 beträgt 9/10 = 90 %.

Das zweite Intervall dauert von Tag 354 bis Tag 365. Die Zahl der Patienten unter Risiko hat sich durch den Todesfall an Tag 353 auf 9 verringert. Die Überlebenswahrscheinlichkeit von Tag 354 bis Tag 365 beträgt deshalb 8/9 = 89 %.

Das dritte Intervall dauert von Tag 366 bis Tag 464. An Tag 464 stehen nur noch 6 Patienten unter Risiko (von den ursprünglich 10 Studienteilnehmern sind zwei verstorben und zwei zensiert worden). Die Überlebenswahrscheinlichkeit von Tag 366 bis Tag 464 beträgt deshalb 5/6 = 83 %. Die Wahrscheinlichkeit, das dritte Intervall zu überleben, ist somit etwas niedriger als die Wahrscheinlichkeit, das erste oder zweite Intervall zu überleben. Entsprechend sehen Sie auch in der Kaplan-Meier-Kurve an Tag 464 eine etwas größere Stufe als an Tag 365. Ursache ist die in Folge der Zensierungen deutlich reduzierte Anzahl der Patienten unter Risiko.

Die *kumulative Wahrscheinlichkeit*, im Vitalonga-Arm von Tag 1 der Studie bis Tag 365 zu überleben, ist das Produkt aus den Teilwahrscheinlichkeiten beider Beobachtungsintervalle (0,9 x 0,89 = 0,8). Die kumulative Wahrscheinlichkeit, von Tag 1 bis Tag 464 zu überleben, liegt entsprechend bei 66 % (0,9 x 0,89 x 0,83 = 0,66). Durch Multiplikation der Teilwahrscheinlichkeiten lässt sich die kumulative Überlebenswahrscheinlichkeit für jedes beliebige Beobachtungsintervall berechnen.

Exkurs: medianes Überleben

Insbesondere in der Onkologie wird häufig das *mediane Überleben* angegeben. Das ist die Zeitspanne, nach der 50 % der Teilnehmer verstorben sind und 50 % noch leben. Das mediane Überleben wird von der Kaplan-Meier-Kurve abgelesen, indem eine horizontale Linie von der y-Achse bei 0,5 (50 %) bis zum Schnittpunkt mit der Kaplan-Meier-Kurve gezogen wird. Von hier aus wird eine vertikale Linie zur x-Achse gezogen. Der Schnittpunkt mit der x-Achse entspricht dem Median. In unserem Beispiel beträgt das mediane Überleben unter Placebo 380 Tage und unter Vitalonga-Extrakten 563 Tage.

Das mediane Überleben einer Studienpopulation ist robust gegenüber Ausreißern, d. h. der Wert verändert sich nicht durch einzelne Teilnehmer, die sehr früh oder sehr spät im Studienverlauf verstorben sind. Außerdem kann das mediane Überleben auch dann zuverlässig bestimmt werden, wenn es gegen Ende der Studienlaufzeit vermehrt zu Studienabbrüchen kommt: Auch wenn wir nicht wissen, ob Herr Schmidt (Studienabbruch an Tag 448) und Frau Müller (Studienabbruch an Tag 477) zu Studienende noch lebten, haben sie sicher an Tag 380 noch gelebt und fließen somit in die Analyse des medianen Überlebens unter Placebo ein.

Da das mediane Überleben nur einen isolierten Zeitpunkt beschreibt, sollte es immer im Gesamtkontext der Ereigniszeitanalyse bewertet werden: Ein höheres medianes Überleben in einem Studienarm ist keine Garantie dafür, dass auch die langfristige Prognose besser ist als im Vergleichsarm – es sagt nichts aus über den weiteren Verlauf der 50 % Studienteilnehmer, die bis zum Median überlebten.

erlauben somit eine vollständigere Nutzung inkompletter Daten als das relative Risiko. Da fehlende Werte über die Zensierung eingehen, müssen die fehlenden Werte nicht ersetzt werden. Trotzdem können fehlende Daten in Ereigniszeitanalysen zu relevanten Verzerrungen führen, wenn das Fehlen der Daten nicht zufällig ist, sondern beispielsweise durch die Verträglichkeit oder Wirksamkeit der Studienmedikation beeinflusst wurde. So vertragen z. B. multimorbide Patienten eine intensive Chemotherapie häufig schlechter als jüngere, gesündere Patienten. Dies kann dazu führen, dass multimorbide Patienten im Interventionsarm häufiger und früher die Studie abbrechen als im Kontrollarm. In die Ereigniszeitanalyse des Interventionsarm fließen dann überwiegend Daten von Patienten ein, die im Vergleich zur Kontrollgruppe jünger und gesünder sind und deshalb – unabhängig von der untersuchten Intervention – eine bessere Prognose haben als die Kontrollgruppe (zum Umgang mit fehlenden Daten).

Was bedeutet nun die eingangs durch unseren Statistiker berechnete **Hazard Ratio (HR)** von 0,470? Das *Hazard* ist die Rate für die Geschwindigkeit, mit der ein Ereignis (z. B. Tod) zu einem genau definierten Zeitpunkt eintritt. Das Hazard ist eine theoretische Größe, die sich in klinischen Studien nicht vollständig berechnen lässt – hierfür wären unendlich viele Datenpunkte erforderlich. Mithilfe mathematischer Modelle (z. B. dem Cox-Proportional-Hazards-Modell) lässt sich jedoch die Hazard Ratio schätzen, d. h. das Verhältnis der Hazards zweier Gruppen über die beobachtete Zeitspanne hinweg. Voraussetzung dafür ist, dass die Hazards der verglichenen Gruppen über den Beobachtungszeitraum in einem festen Verhältnis zueinanderstehen (sogenannte „Proportional-Hazards-Annahme“). Stark abweichende Kurvenverläufe oder sich schneidende Kurven können darauf hindeuten, dass die Proportional-Hazards-Annahme verletzt ist. Solche Verläufe entstehen vor allem dann, wenn sich die Effekte der Therapie über die Zeit (z. B. unmittelbar postoperativ erhöhtes Sterblichkeitsrisiko) verändern, wodurch die berechnete Hazard Ratio ungenau oder sogar irreführend wird. Auch im Beispiel der Vitalonga-Studie schneiden sich die Kurven von Interventions- und Kontrollgruppe am Studienende. Dies deutet darauf hin, dass im zeitlichen Verlauf das Sterblichkeitsrisiko unter Vitalonga-Extrakten wieder zunimmt. Denkbar wären z. B. unter einer Langzeittherapie auftretende unerwünschte Ereignisse wie Leber- oder Nierenschäden.

Unsere Beispielstudie zu Vitalonga-Extrakten illustriert, dass sich das relative Risiko und die Hazard Ratio wesentlich unterscheiden können: In der Vitalonga-Studie lag das relative Risiko bei 1,0, die Hazard Ratio dagegen bei 0,470. Woran liegt das? Das relative Risiko vergleicht bei zwei Gruppen, wie hoch das Risiko für das untersuchte Ereignis war – unabhängig davon, wie lange die Beobachtungszeit in den Gruppen war. Die Hazard Ratio setzt dagegen die beobachteten Ereignisse und die Zeitspanne der Beobachtung in ein Verhältnis zueinander. Auch im Alltag beziehen wir viele Größen auf eine bestimmte Zeitspanne. Die Aussage Ihres Partners „Ich habe schon dreimal die Spülmaschine ausgeräumt“ können Sie zum Beispiel nur korrekt einordnen, wenn klar ist, ob Ihr Partner damit die Hausarbeit des heutigen Tages, der letzten Woche oder des letzten Jahres meint. Schlichtweg zu kontern, dass auch Sie bereits dreimal die Spülmaschine ausgeräumt haben, macht offensichtlich nur Sinn, wenn Sie sich auf die gleiche Zeitspanne wie Ihr Partner beziehen. In ähnlicher Weise lässt sich das Risiko für das Auftreten bestimmter Ereignisse nur sinnvoll vergleichen, wenn die Beobachtungszeit in den Studienarmen ähnlich war.

In der Vitalonga-Studie war die Zahl an Todesfällen in beiden Armen gleich (9 von 10 Patienten), entsprechend lag das relative Risiko bei 1,0 (zum relativen Risiko). Die Beobachtungszeit unterschied sich zwischen den Armen jedoch deutlich: Die durchschnittliche Beobachtungszeit betrug 533 Tage im Interventionsarm und 372 Tage im Kontrollarm. Die Hazard Ratio von 0,470 spiegelt wider, dass im Interventionsarm die Patienten später verstarben als im Kontrollarm bzw. – anders ausgedrückt – dass pro Zeiteinheit weniger Todesfälle auftraten als im Kontrollarm. Qualitativ kann das Hazard Ratio interpretiert werden wie das relative Risiko: Bei einer Hazard Ratio unter 1 tritt das Ereignis in der Interventionsgruppe *pro Zeiteinheit* seltener auf als in der Kontrollgruppe. Wenn es sich um einen negativen Endpunkt handelt, bedeutet eine Hazard Ratio < 1 (genau wie beim relativen Risiko) einen Vorteil für die Interventionsgruppe. In der Medizin ist es in der Regel nicht nur wichtig, ob ein Ereignis auftritt, sondern auch wie schnell es auftritt: Ein verlängertes Überleben oder das spätere Auftreten unerwünschter Ereignisse (z. B. Implantatversagen) beeinflussen maßgeblich die Wahl der Therapie.[3] Bei unterschiedlichen Beobachtungszeiten sollte deshalb in der Regel die Hazard Ratio anstelle des relativen Risikos herangezogen werden, um den Therapieeffekt zu beurteilen.

[3] Dabei muss selbstverständlich immer kritisch geprüft werden, ob das Ausmaß des Effekts tatsächlich patientenrelevant ist und der Nutzen die Risiken überwiegt. Wie beim relativen Risiko müssen auch hier die absoluten Effekte beurteilt werden.

Die Kaplan-Meier-Kurve einer Interventionsstudie: vier wichtige Fragen

1. Schauen Sie sich die Beschriftung der y-Achse an. Reicht die y-Achse von 0 % bis 100 % (bzw. 0 bis 1)? Falls nein, dann wird nur ein vergrößerter Auszug dargestellt – durch diesen Kniff sieht der Effekt größer aus als er tatsächlich ist.
2. Ab wann trennen sich die beiden Kurven, d.h. wie lange dauert es durchschnittlich, bis ein Patient von der Therapie profitiert? Dies kann insbesondere bei Patienten, deren Prognose durch andere Erkrankungen eingeschränkt ist, eine wichtige Information sein.
3. Kreuzen sich die beiden Kurven im zeitlichen Verlauf oder zeigen stark abweichende Verläufe? Dies spricht dafür, dass die positiven Effekte einer Intervention im Verlauf nachlassen oder Nebenwirkungen die positiven Effekte überlagern.
4. Wie viele Patienten flossen in die Analyse ein? Schlussfolgerungen, die sich nur auf wenige Patienten stützen, sind mit großen Unsicherheiten behaftet. Dies betrifft regelmäßig das rechtsseitige Ende der Kurve, da gegen Studienende oftmals viele Patienten zensiert wurden bzw. das untersuchte Ereignis bereits bei ihnen eingetreten ist.

Zusammenfassung

Die Kaplan-Meier-Kurve zeigt den zeitlichen Verlauf von Überlebenswahrscheinlichkeiten (oder anderen Ereignissen) und ermöglicht einen visuellen Vergleich von unterschiedlichen Studienarmen. Die Hazard Ratio fasst bei einer Ereigniszeitanalyse den Unterschied zwischen zwei Gruppen in einer einzigen Kennzahl zusammen. Bei unterschiedlichen Beobachtungszeiten ist die Hazard Ratio präziser als das relative Risiko, da die Hazard Ratio berücksichtigt, nach welcher Beobachtungszeit die Ereignisse aufgetreten sind.

Interessenkonflikte
Die Autorin und der Autor geben an, keine Interessenkonflikte zu haben.

Dr. med. Natascha Einhart
Bundesärztekammer, Berlin

Prof. Dr. Tim Mathes
Ressortleiter Gesundheitsökonomie am Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Köln