Adaptive Designs in der klinischen Forschung

Jan C. Schuller
EORTC, European Organisation for Research and Treatment of Cancer, Bruxelles, Belgique

Zusammenfassung

Durch die Entwicklungen in der Molekularbiologie hat sich eine Reihe neuer Aspekte der Krebstherapie ergeben. Einerseits sind viele neue potentiell wirksame Substanzen (z.B. die «Biologics») verfügbar geworden, andererseits zeigen neue Diagnoseverfahren (etwa durch Biomarker), dass viele Unterarten eines Krebses gezielte Therapien erfordern *(targeted therapies)*. Beim klinischen Studiendesign besteht hier in mancher Hinsicht ein Dilemma, denn einerseits geht man davon aus, mehr oder weniger homogene Patientenkollektive zu behandeln, andererseits besteht der Bedarf an der «personalisierten Medizin».
Bei den «klassischen» experimentellen Designs müssen *a priori* Annahmen über die Eigenschaften der zu testenden Behandlung gemacht werden. Bei vielen der ganz neuen Therapieoptionen liegen dafür allerdings nur unzureichende Informationen vor. Hier versprechen die adaptiven Designs Abhilfe. Sie sollen den Entwicklungsprozess von neuen Krebsbehandlungen beschleunigen und ebenso der Forderung nach der personalisierten Medizin gerecht werden.
Der vorliegende Artikel gibt eine Einführung in die Methodik der adaptiven Designs, zusammen mit einer kritischen Würdigung aus Sicht der Statistikabteilung der EORTC.

Signalmessung

Ein klinischer Versuch zielt darauf ab, ein Signal zu messen. Dies besteht in der Differenz zwischen der Wirksamkeit einer neuen Behandlung und der einer Kontrolle. Die Kontrolle kann ein Plazebo, eine Standardbehandlung oder eine historische Kontrolle sein. Ein wichtiger Teil der Aufgabe des Statistikers besteht darin, ein Studiendesign zu finden, dass mit hoher Wahrscheinlichkeit das gesuchte Signal messen kann (falls die neue Behandlung in Wahrheit tatsächlich besser ist als die Kontrolle). Das ist gemeint, wenn wir sagen, eine Studie sei «gepowert», um einen bestimmten Effekt zu messen. Es ist leicht einsehbar, dass man für schwache Signale präzisere Instrumente benötigt. In der Statistik erhält man ein präziseres Instrument, indem man einen höheren Stichprobenumfang wählt.
Die Anzahl an Patienten ist ein wesentlicher Kostenfaktor, und so wird diesem Aspekt besondere Beachtung geschenkt: Weder wollen wir zu wenige Patienten in eine Studie einschliessen, weil die Gefahr besteht, dass das gesuchte Signal nicht gemessen wird («underpowered»), noch wollen wir zu viele Patienten einschliessen, das wäre unethisch und teuer.

Die «klassische» Entscheidungsregel

Das klassische Vorgehen ist, die Hypothesen, Alpha und Beta (Typ-I- und Typ-II-Fehler) und die Stichprobengrösse vor Studienbeginn festzulegen. Wenn die Studie abgeschlossen ist und alle Daten gesammelt sind, liefert der statistische Analyseplan eine klare Entscheidungsregel, ob die neue Behandlung den gesuchten Effekt hat oder nicht. Der wesentliche Punkt hier: Die Entscheidung wird erst getroffen, wenn alle Patienten der Studie ausgewertet sind. Um den Fehler der ersten Art (fälschliches Verwerfen der Nullhypothese) unter Kontrolle zu halten, wird nicht «zwischendurch» geschaut.

Sequentielle Methoden

Dies änderte sich im Jahr 1943, als der Mathematiker Abraham Wald, tätig an der Columbia University, dem *National Defense Research Committee* einen Report mit dem Titel «Sequential Analysis of Statistical Data: Theory» vorlegte. Walds Report betraf u.a. die Qualitätskontrolle von industriellen Prozessen: Hier kommen die Produkteinheiten, eine nach der anderen, am Ende des Prozesses heraus. Wir sind am Anteil defekter Einheiten interessiert. Gemäss der «klassischen» Entscheidungsregel müssten wir warten, bis eine vorher festgelegte Anzahl Einheiten produziert wurde. Anschliessend würden wir die Anzahl defekter Einheiten auszählen und unsere Statistik auf diese Zählung gründen.
Aber wie gross soll die Anzahl Einheiten sein, die man abwarten muss? Das hängt vom Anteil defekter Einheiten ab. Ist dieser gross, dann braucht man nur wenige Einheiten; ist er klein, dann benötigt man viele. Dies kann nachteilig sein, vor allem dann, wenn man keine Idee hat, welcher Anteil defekter Einheiten zu erwarten ist. Betrachten wir dazu das folgende extreme Beispiel: Angenommen, durch einen unbemerkten Fehler sei beinahe jede Einheit defekt. Allerdings wurde entschieden, dass die Qualitätskontrolle anhand einer Stichprobengrösse von N = 1000 durchgeführt wird. Bei ei-

Diese Arbeit wurde vom National Cancer Institute (NCI, Bethesda, Maryland, USA, Grant Nr. 5U10 CA011488-40) und vom Fonds Cancer (FOCA, Belgien) unterstützt.
Der Autor ist verantwortlich für den Inhalt dieser Arbeit. Sie gibt nicht notwendigerweise die offizielle Position des NCI wieder.

ner Tagesproduktion von 100 Einheiten vergingen zehn Tage, bis der Fehler bemerkt würde. Im vorliegenden Fall hätte aber auch ein Nichtstatistiker bereits nach vielleicht zehn oder zwanzig defekten Teilen sagen können, dass in der Fabrik etwas falsch läuft. Dies ist ein wichtiger neuer Aspekt der sequentiellen Analyse: Jede Einheit, die den Produktionsprozess verlässt, wird sofort begutachtet und bewertet und kann möglicherweise die Entscheidung bewirken.
Eine Variation dieser Idee bilden die *group sequential designs*, wie sie in der klinischen Forschung seit langem üblich sind. Hier werden im Verlauf der Studie eine oder mehrere Interimsanalysen durchgeführt. Falls eine neue Behandlung extrem wirksam oder unwirksam ist, dann kann man das mitunter bereits sehen, bevor das gesamte geplante Patientenkollektiv behandelt wurde. So kommen die Adaptationen des Versuchsdesigns ins Spiel: Nach einer Interimsanalyse der Daten kann, im Falle sehr hoher oder sehr niedriger Wirksamkeit der getesteten Behandlung, die Studie vorzeitig abgebrochen werden. Dieses Vorgehen ist ethisch und wirtschaftlich.

Bayesianische Methoden

Diese Methoden sind nach Rev. Thomas Bayes (ca. 1702–1761) benannt. Nach der Bayes-Regel wird die Wahrscheinlichkeit eines Ereignisses nicht nur von aktuellen experimentellen Ergebnissen bestimmt, sondern auch ganz wesentlich von der Vorerfahrung, die gemacht wurde (der *a-priori*-Wahrscheinlichkeit). Diese Methoden sind bei vielen Wissenschaftlern besonders beliebt. Der Grund ist wohl, dass diese Methoden so sehr dem Vorgehen der induktiven Naturwissenschaften, einschliesslich der Medizin, entsprechen. Auch Entscheidungen in unserem täglichen Leben sind bayesianisch geprägt: Wir sehen den Sonnenaufgang am ersten Morgen, am zweiten und an jedem anderen, und schliesslich erwarten wir, dass die Sonne auch morgen aufgeht. Der bayesianische Statistiker würde sagen: «Der Sonnenaufgang morgen früh hat eine hohe *a-priori*-Wahrscheinlichkeit».
Traditionell wird in der Medizin dem statistischen Testen von Hypothesen ein hoher Stellenwert eingeräumt. Man möchte «signifikante» Ergebnisse. Es ist aber lange bekannt, dass ein «p <0,05» für sich allein genommen wenig aussagt und zudem regelmässig falsch interpretiert wird. Die bayesianischen Methoden ermöglichen die Bewertung von Behandlungseffekten, ohne auf Signifikanzwerte angewiesen zu sein.

Definition und Beispiele für adaptive Designs

Ein adaptives Design erlaubt Änderungen des Studienablaufs und/oder der statistischen Parameter nach Studienbeginn, ohne die Valididät und Integrität der Studie zu verletzen. Absicht ist es, klinische Versuche effizienter, schneller und flexibler zu gestalten.
Im Folgenden werden die wichtigsten Beispiele für adaptives Studiendesign vorgestellt. Das zuerst besprochene *group-sequential*-Design ist schon lange in der Praxis angekommen und nimmt eine gewisse Sonderstellung ein, auch weil die möglichen Adaptationen (fast immer Studienabbruch wegen Unwirksamkeit oder Sicherheitsbedenken) sehr begrenzt sind. Dem gegenüber stehen die neueren «flexiblen» adaptiven Designs, welche die verschiedensten Anpassungen vorsehen, z.B. Änderung des Stichprobenumfangs, adaptive Randomisierung, nahtloses («seamless») Phase-II/III-Design, Verwerfung von Studienarmen, Änderung des Endpunktes sowie Wechsel von «Überlegenheit» nach «Nicht-Unterlegenheit» zwischen Behandlungsarmen.

Group sequential designs* und *Early stopping

Hier kann die Studie aufgrund von Zwischenergebnissen vorzeitig abgebrochen werden (entweder weil die Behandlung viel wirksamer oder unwirksamer ist als vorausgesehen). Dies ist die gängigste Form des adaptiven Designs und findet bei Phase-II- und -III-Studien häufige Anwendung.

Phase-II/III-Design *(Seamless phase II/III design)*

Dieses Design kombiniert die Zielsetzungen von Phase II (feststellen, ob überhaupt eine Antitumor-Wirkung erzielt wird) und Phase III (Vergleich der neuen Behandlung mit einer Kontrolle) in einer einzigen Studie. Nehmen wir an, es solle herausgefunden werden, welche von drei neuen Behandlungen die beste sei. Ein mögliches Vorgehen ist, alle drei Behandlungen zunächst im Rahmen einer Phase-II-Studie zu testen, d.h. mit geringer Patientenanzahl und entsprechend geringer Power gegen eine (ggf. historische) Kontrollgruppe. Diejenige der drei neuen Behandlungen, die am besten abschneidet (und eine Verbesserung gegenüber der Kontrolle erwarten lässt), würde beibehalten und in einer ausreichend «gepowerten» Phase-III-Studie gegen die Kontrolle getestet. Zwischen Phase II und Phase III würde eine gewisse Zeit vergehen (Monate bis Jahre), die zur Analyse der Phase-II-Daten und zur Planung der Phase-III-Studie benötigt wird.
Bei einem kombinierten Phase-II/III-Design wird die Patientenrekrutierung zwischen den beiden Phasen nicht unterbrochen. Stattdessen führt man Interimsanalysen durch und verwirft jedes Mal den Behandlungsarm mit der niedrigsten Wirkung. Die Patienten des Phase-II-Teils würden so Eingang in die Endanalyse des Phase-III-Teils finden. Theoretisch wird so eine geringere Anzahl an Patienten benötigt, und die Phase-III-Ergebnisse lägen schneller vor als beim Verfahren mit getrennten Studien.
Die Wirklichkeit lässt diese Vorteile manchmal in weniger strahlendem Licht erscheinen. Ein Grund dafür ist der sogenannte «Overrun»: Während der Interimsanalyse werden weitere Patienten rekrutiert, mitunter so schnell, dass die Entscheidung für eine Adaptation zu spät kommt, weil bereits alle Patienten eingeschlossen sind.
Aus Phase-II-Studien wird im Allgemeinen sehr viel über eine neue Behandlungsmöglichkeit gelernt, und diese Erkenntnisse haben grossen Einfluss auf die Planung und Durchführung von Phase III. Das Phase-II/III-Design ist hier im Nachteil, denn es wird als Einheit

geplant, und Erkenntnisse aus Phase II können nur bedingt Eingang in Phase III finden.

Anpassung der Stichprobengrösse
(Sample size re-estimation)
Aufgrund von Interimsergebnissen kann man die Stichprobengrösse anpassen, um die gewünschte Power für den Nachweis eines erwarteten Behandlungseffekts zu erreichen. Als Faustregel gilt, dass man die Stichprobe aufstocken kann, wenn die Interimsanalyse eine «conditional power» zwischen 30 und 90% erreicht (dies ist die Wahrscheinlichkeit, einen tatsächlich vorhandenen Effekt auch nachweisen zu können. Im Gegensatz zur ursprünglich festgelegten Power, z.B. 80%, ist diese von den Interimsergebnissen abhängig). Ist sie kleiner als 30%, besteht wenig Hoffnung, dass die Studie ein positives Ergebnis haben wird, selbst mit vergrösserter Stichprobe. Ist sie grösser als 90%, dann wird der Versuch wahrscheinlich auch ohne Anpassung einen Effekt nachweisen können.
Dieses Design kann die anfänglichen Kosten einer Studie niedrig halten. Wenn dann aber die Stichprobe erweitert wird, muss das Budget ebenfalls aufgestockt werden. Dies ist manchmal problematisch, weil die Entscheidung zur Anpassung der Stichprobengrösse gewöhnlich von einem unabhängigen Monitoring-Komitee getroffen wird *(IDMC – Independent Data Monitoring Committee)*, während über das Budget von anderer Stelle (z.B. einem Industrieunternehmen) verfügt wird.
Eine entscheidende Grösse des Studiendesigns ist der klinisch relevante Effekt. Die Anpassung der Stichprobengrösse aufgrund von Interimsdaten bedeutet, dass ebenfalls angepasst wird, wie gross ein Behandlungseffekt sein muss, um klinisch relevant zu sein. Dies ist konzeptionell problematisch.

Adaptive Randomisierung
Normalerweise wird nach einem festgelegten Verhältnis randomisiert (z.B. 1:1 oder 1:2). Bei der adaptiven Randomisierung wird mit grösserer Wahrscheinlichkeit in einen wirksamen Behandlungsarm randomisiert. Diese Wahrscheinlichkeit wird durch Zwischenanalysen laufend angepasst (Bayesianische Methode) und findet besonders beim *Biomarker adaptive*-Design Anwendung. Die Methode ist problematisch, weil sie an sich dem Grundgedanken der Randomisierung widerspricht und ein deterministisches Element bei der Zuteilung der Studienbehandlungen einfügt *(operational bias)*.

***Pick the winner*-Design**
Dies ist typischerweise eine Phase-II-Studie mit mehreren Behandlungsarmen und einer oder mehreren Interimsanalysen (wie im besprochenen Beispiel zum Phase-II/III-Design). Aufgrund der Zwischenergebnisse werden wirksame Arme beibehalten und unwirksame nicht fortgeführt.

***Biomarker adaptive*-Design**
Modifikationen der Studiendurchführung werden aufgrund des Ansprechens verschiedener Biomarker durchgeführt, die mit der Krankheit assoziiert sind. Dieses Design kann dabei helfen, die richtige Patientenpopulation für eine neue Therapie zu finden und somit einen Beitrag zur «personalisierten Medizin» und den «targeted therapies» liefern. Dabei ist es wichtig zu beachten: Die Assoziation eines Biomarkers mit einem klinischen Endpunkt bedeutet noch nicht, dass ein prädiktives Modell vorliegt.

Aufstockung der Rekrutierung
(Ramping-up of patient accrual)
Wegen des «Overruns» (s.o.) wird ein adaptives Design seine Vorteile besser ausspielen, wenn die Rekrutierung langsam ist, d.h., sie kann auch künstlich niedrig gehalten werden. Wenn dann die Adaptierung des Designs stattgefunden hat, kann die Rekrutierung erhöht werden.

Adaptive Dosisermittlung *(Adaptive dose finding)*
In einer Phase-I-Studie wird die maximal tolerierbare Dosis einer neuen Substanz gesucht. Anstatt die zu testenden Dosen vor Studienbeginn festzulegen, können sie bei diesem Design aufgrund der Zwischenergebnisse angepasst werden (z.B. bei der *continual re-assessment method, CRM*).

Änderung des primären Endpunkts
Die Änderung des primären Endpunkts ist eine extreme Form der Adaptation. Dies ist nur in ganz wenigen Ausnahmefällen gerechtfertigt. In den meisten Fällen ist eine Änderung des primären Endpunkts einer Studie aufgrund von Interimsergebnissen sehr kritisch zu sehen.

Fazit

Die wichtigsten Techniken gegen die Verzerrung von Versuchsergebnissen («bias») sind Randomisierung und Verblindung. Adaptive Designs machen Zugeständnisse an beide Aspekte: Wenn die Ergebnisse der Interimsanalysen bekannt werden, ist die Integrität der Studie gefährdet. Bei adaptiven Designs besteht also immer die Gefahr einer schwer zu kontrollierenden Verzerrung der Resultate. Diese Gefahren sind auch den regulatorischen Behörden bekannt: Besonders die europäische Behörde EMA nennt eine Reihe strikter Anforderungen für adaptive Designs, damit z.B. eine Zulassungsstudie auch tatsächlich als konfirmatorisch angesehen werden kann. Im Vordergrund steht dabei die Kontrolle des Fehlers erster Art (irrtümliche Annahme, die neue Behandlung sei wirksam, wenn sie in Wahrheit wirkungslos ist). Manche Adaptationen, z.B. die Änderung des primären Endpunktes, sind vor der EMA nur bei sehr begründeten Ausnahmen zu rechtfertigen.
Es ist augenfällig: Adaptive Methoden benötigen eine viel aufwendigere Logistik, Planung und damit mehr Zeit als «klassische» Methoden. Oft steht dieser Aufwand im Missverhältnis zum erzielten Gewinn.
Letztlich ist auch zu bedenken, dass eine Studie mit adaptivem Design nicht empfindlicher ist für Behandlungseffekte als ein klassisches Design. Somit sind adaptive Designs keine «Wundermittel». Auch für sie gilt, dass für kleinere Effekte grössere Stichproben benötigt werden.

Diese Aspekte sollen sorgfältig abgewogen werden, und es ist von Fall zu Fall immer zu überlegen, ob ein «klassisches» Design nicht ähnlich gute Dienste leistet wie ein adaptives Design.

Verdankung
Der Autor dankt einem anonymen Reviewer für wertvolle Hinweise zu einer früheren Version dieser Arbeit.

Korrespondenz:
Jan C. Schuller
Senior Biostatistician
EORTC, European Organisation for Research
and Treatment of Cancer
Avenue E. Mounierlaan, 83/11
Bruxelles 1200 Brussel
Belgique
jan.schuller@eortc.be

Empfohlene Literatur

– Guidance for Industry. Adaptive Design Clinical Trials for Drugs and Biologics. Draft Guidance. U.S. Department of Health and Human Services, Food and Drug Administration. February 2010.
– Emerson Scott S. Issues in the use of adaptive clinical trial designs. Statist. Med. 2006;25:3270–96.
– Reflection paper on methodological issues in confirmatory trials planned with adaptive Design; Committee for medicinal products for human use. EMA, Doc. Ref. CHMP/EWP/2459/02EMA, October 2007.

Assoziation? Eine 37-jährige Frau kommt mit der Angabe von Fieber (39,2 °C) seit 3 Tagen, Kopfweh, Nausea und Verwirrung. Die Glasgow Coma Scale beträgt 11. Es besteht eine Nackensteife, fehlende Extensorreflexe. Die Lumbalpunktion ergibt einen Druck von 48 cm, 394 Leukozyten, Protein 1,21 g/l, Glukose 0,6 mmol/l. Ein MRI des Gehirns zeigt multiple Läsionen mit gesteigerter Intensität und ringförmiger Gadoliniumansammlung. Liquorkulturen ergeben Streptokokken. Was ist los? *(Auflösung siehe rechte Spalte)*

Auflösung: *Eine* ***Meningitis****? Ein* ***Hirnabszess****? Beides trifft zu, aber welcher Ursache? Die nachgewiesenen Streptokokken gehören in die Lancefield-Gruppe C (Streptococcus equi, Subspecies zooepidemicus). Es stellt sich heraus, dass die Patientin eine enthusiastische Reiterin ist und kürzlich von ihrem Pferd gebissen wurde. Meningitiden werden meist von den Streptokokken der Subspecies zooepidemicus ausgelöst und sind Penicillin-empfindlich. Nach zwei Jahren ist die Patientin aber bei schwerer Amnesie noch immer unselbständig. (Lancet. 2010;376:1194.)*