

GRUNDLAGEN DER AUSWERTUNG UND OPTIMIERUNG VON HALTBARKEITSUNTERSUCHUNGEN

von Martin Holz, Donald Dill, Rudolph Frank und Theo Wember
aus einem Artikel erschienen in „Die Pharmazeutische Industrie“, Editio Cantor, September 2000

1. Zusammenfassung

Die behördliche Anforderung, die Haltbarkeitsdauer eines pharmazeutischen Produkts mit einer hinreichenden statistischen Sicherheit anzugeben und den Einfluss der relevanten Einflussfaktoren zuverlässig abzuschätzen, erfordert einen hohen zeitlichen, logistischen und analytischen Aufwand. Um dieses Ziel zu erreichen, erlaubt das Rahmenwerk der ICH Guideline für Stabilitätsprüfungen einen weiten Spielraum für die Versuchsplanung und Auswertung von reduzierten Haltbarkeitsstudien mit einem Matrixing und Bracketing Konzept. Die vorliegende Arbeit zeigt, wie durch individuell angepasste D - optimale Versuchspläne und multiple Regressionsrechnung die Präzision der Haltbarkeitsdauerschätzung deutlich erhöht und der experimentelle Aufwand gleichzeitig vermindert wird. Mit moderner statistischer Software ist die Erstellung und Analyse individueller optimaler Pläne auch dem Nichtmathematiker möglich.

Die grossen Vorteile der gemeinsamen Auswertung aller beteiligten Faktoren in einem generellen linearen Modell wird dargestellt gegenüber der separaten einfachen linearen Regression an die Zeitverläufe einzelner Chargen, Packungsarten oder Temperaturen.

Der FDA Empfehlung, bei der Entscheidung über das Zusammenfassen ('pooling') von Chargen, Packungen etc. generell ein Irrtumsniveau von $\alpha = 0.25$ zuzulassen, werden praxisgerechte Alternativen auf Basis von Trennschärfavorhersagen gegenübergestellt. Dabei wird die statistische Signifikanz eines Effekts von seiner praktischen Relevanz unterschieden. Das Verfahren stellt sicher, dass relevante Effekte auch als signifikant erkannt werden.

Die Problematik der Behandlung von nur drei Chargen als repräsentative Zufallsstichprobe und die besondere Fehlerstruktur analytischer Assays werden diskutiert.

In einem zweiten Teil in „Die Pharmazeutische Industrie“ wird ein praktisches Beispiel der Anwendung der vorgestellten Verfahren behandelt werden.

2. Einführung

Für ein neues pharmazeutisches Produkt muß gezeigt werden, daß am Ende der Haltbarkeitsdauer mit einer statistischen Wahrscheinlichkeit von 95 % ein bestimmter Wirkstoffgehalt im Mittel nicht unterschritten bzw. eine Grenze für Abbauprodukte nicht überschritten wird¹. Das Alterungsverhalten ist von entscheidender Bedeutung für die Festlegung von realistischen Spezifikationen für die Batch - Freigabe². Die Ziele der Planung und Auswertung von Stabilitätsstudien sind:

1. eine möglichst genaue Schätzung der Haltbarkeitsdauer und damit auch eine möglichst lange Haltbarkeitsdauer. Je mehr Messungen in die Schätzung eingehen, desto präziser wird diese.
2. die Schätzung der Effekte der Faktoren Verpackungstyp, Verschlusstyp, Dosierung, Charge, Temperatur, Feuchte, Lichtintensität, Wellenlänge usw. auf den Zerfallsprozess. Dabei sollen Effekte relevanten Ausmasses sicher erkannt werden.

Um beide Ziele mit möglichst geringem Aufwand sicher zu erreichen, wird eine synoptische Auswertung aller Daten aus einer für alle Effekte gemeinsam geplanten Studie mithilfe der multiplen Regressionsrechnung durchgeführt. Bei separater Auswertung der einzelnen Effekte mit einfacher linearer Regression ist die Schätzung der Effekte nicht zuverlässig möglich, da etwaige Wechselwirkungen zwischen den Faktoren nicht ohne weiteres erkannt werden können. Wichtige Effekte könnten übersehen werden, weil sie von einer Wechselwirkung maskiert sind. Ein Effekt kann als relevant erscheinen obgleich tatsächlich nur eine Wechselwirkung vorliegt. Ausserdem ist die getrennte Auswertung höchst unökonomisch, weil die Residualvarianz und damit auch die Präzision der Effektschätzung immer wieder aufs neue aus einer vergleichsweise kleinen Anzahl Messungen ermittelt wird, wodurch die Vertrauensgrenzen aufgeweitet und damit die Haltbarkeitsdauerschätzungen unnötig verkürzt werden.

3. Versuchspläne, Faktoren und Modelle

Der einfachste Versuchsplan besteht darin, bei jeder Kombination aus Charge, Packungstyp, Temperatur usw. eine komplette Zeitreihe aufzunehmen. Dieser vollfaktorielle Plan erfordert ein Maximum an Aufwand, garantiert dafür aber die Schätzung aller Effekte und Wechselwirkungen der beteiligten Faktoren. Durch gezieltes Weglassen von Messungen ('matrixing') um den Preis, einige Effekte nicht mehr eindeutig schätzen zu können, entstehen teilfaktorielle Versuchspläne. Durch moderne Software, z.B. RS/DISCOVER³, Cornerstone⁴ oder Statgraphics⁵ ist es möglich, solche Versuchspläne ohne grossen Aufwand zu erstellen und ihre Trennschärfereigenschaften zu dokumentieren.

Eine gute Versuchsplanung erfordert eine detaillierte Festlegung eines mathematischen Modells für alle am Alterungsprozess beteiligten Faktoren.

3.1 Faktoren

Um einen angemessenen Versuchsplan aufzustellen, müssen je nach Produkt alle oder einige der folgenden Faktoren berücksichtigt werden:

- Produktionslos („Batch“, "Charge", "Lot")
- Dosierung (z.B. 10mg, 50mg, 100mg, 250mg)
- Verpackung (HDPE-Flasche, Aluminium-Blister, Braunglas... und Verschlussarten)
- Temperatur (=Lagerungstemperatur z.B. 25, 30, 40 °C)
- Feuchte (rel. Luftfeuchtigkeit, je nach Klima im Vermarktungsgebiet ab 60 % bis 95%).
- Zeit nach der Herstellung (z.B. 0, 3, 6, 9, 12, 18, 24, 36 Monate),

Bei photosensiblen Produkten kommen noch "Lichtintensität" und "Wellenlänge" hinzu.

Wünschenswert ist eine Unabhängigkeit vom Batch, von der Verpackung und von der Temperatur und den Lichtbedingungen. Ideal ist eine Arznei, die überhaupt keinen Zerfall erkennen lässt (=Unabhängigkeit vom Faktor Zeit). Dies kann mit Hilfe der Regressionsanalyse eines geeigneten mathematischen Modells getestet werden.

3.2 Modelle

Zur Ermittlung der Effekte der Faktoren kann ein einfaches lineares Modell angenommen werden⁶:

$$\begin{aligned}
 Y(t) = & \alpha_0 + \alpha_i^{\text{Batch}} + \alpha_j^{\text{Pack}} + \alpha_k^{\text{Temp}} + \\
 & \alpha_{ij}^{\text{Batch-Pack}} + \alpha_{ik}^{\text{Batch-Temp}} + \alpha_{jk}^{\text{Pack-Temp}} + \alpha_{ijk}^{\text{Batch-Pack-Temp}} + \\
 & \beta_0^{\text{Batch}} \cdot t + \beta_j^{\text{Pack}} \cdot t + \beta_k^{\text{Temp}} \cdot t + \\
 & \beta_0 \cdot t + \beta_{ij}^{\text{Batch-Pack}} \cdot t + \beta_{ik}^{\text{Batch-Temp}} \cdot t + \beta_{jk}^{\text{Pack-Temp}} \cdot t + \beta_{ijk}^{\text{Batch-Pack-Temp}} \cdot t + \varepsilon
 \end{aligned}$$

Formel 1

$Y(t)$:	Gehalt als Funktion der Zeit (oder Log(Gehalt))
α_0 :	Globaler Mittelwert, entspricht dem mittleren Y-Achsenabschnitt beim 'common slope' model)
β_0 :	Globale Steigung =Zerfallsrate z.B. in %/Jahr. Haupteffekt des Faktors Zeit, entspricht der mittleren Steigung beim 'common slope' model
α_i^{Batch}	Haupteffekt des Faktors Batch = Y-Achsenabschnittsabweichung von Batch i von α_0)
β_i^{Batch}	Wechselwirkung des Faktors Batch mit der Zeit = Steigungsabweichung des Batch i von der globalen Steigung β_0
β_j^{Pack} :	Wechselwirkung des Faktors Verpackung mit der Zeit = Steigungsabweichung der Verpackung j von der globalen Steigung β_0
β_k^{Temp}	Wechselwirkung des Faktors Temperatur mit der Zeit = Steigungsabweichung von der globalen Steigung β_0 bei Temperatur k .
$\beta_{ij}^{\text{Batch-Pack}}$	Wechselwirkung der Faktoren Batch, Pack und Zeit = Steigungsabweichung der Packung j von der Steigung der Charge i .
$\beta_{ik}^{\text{Batch-Temp}}$	Wechselwirkung der Faktoren Batch, Temp und Zeit = Steigungsabweichung der Charge i von der Steigung bei Temperatur k.
$\beta_{jk}^{\text{Pack-Temp}}$	Wechselwirkung der Faktoren Pack, Temp und Zeit = Steigungsabweichung der Packung j von der Steigung bei Temperatur k.

$\beta_{ijk}^{\text{Batch-PackTemp}}$	Wechselwirkung der Faktoren Batch, Pack, Temp und Zeit = Steigungsabweichung der Charge i von der Steigung der Packung j bei Temperatur k.
ε	Residuum, sollte normalverteilt sein mit der "intermediate precision" - Standardabweichung des Assays.

Diese doch recht unhandliche Funktion lässt sich wegen einiger Besonderheiten bei Haltbarkeitsstudien in der Regel erheblich vereinfachen. So ist bei gegebener Gehaltshomogenität innerhalb der Chargen nicht zu erwarten, dass unterschiedliche Verpackungen unterschiedliche Achsenabschnitte aufweisen. Dies gilt ebenso für die Temperaturen, weil die Messung zum Zeitpunkt $t=0$ vor Aufteilung auf die Verpackungen und Klimaschränke stattfindet. Daher können die Haupteffekte dieser Faktoren α_i^{Pack} und α_k^{Temp} und ihre Wechselwirkung $\alpha_{ij}^{\text{Batch-Pack}}$, $\alpha_{ik}^{\text{Batch-Temp}}$, $\alpha_{jk}^{\text{Pack-Temp}}$ und $\alpha_{ijk}^{\text{Batch-Pack-Temp}}$ aus dem Modell entfernt werden. Die Verpackungen bzw. Temperaturen können sich bzgl. der Geradensteigung unterscheiden, nicht jedoch hinsichtlich des Y - Achsenabschnitts. Formel 2 zeigt das so reduzierte Modell.

$$Y(t) = \alpha_0 + \alpha_i^{\text{Batch}} + \beta_i^{\text{Batch}} \cdot t + \beta_j^{\text{Pack}} \cdot t + \beta_k^{\text{Temp}} \cdot t + \beta_0 \cdot t + \beta_{ij}^{\text{Batch-Pack}} \cdot t + \beta_{ik}^{\text{Batch-Temp}} \cdot t + \beta_{jk}^{\text{Pack-Temp}} \cdot t + \beta_{ijk}^{\text{Batch-Pack-Temp}} \cdot t + \varepsilon$$

Formel 2

Bei einem guten Produkt sollte man erwarten, dass der Effekt der Temperatur auf die Zerfallsrate nicht von der Charge abhängt. Und dass dieser Effekt obendrein innerhalb jeder Charge auch noch vom Packungstyp abhängen soll, ist noch unwahrscheinlicher. Zumindest die Wechselwirkung $\beta_{ijk}^{\text{Batch-Pack-Temp}}$ und vielleicht auch $\beta_{ik}^{\text{Batch-Temp}}$ und $\beta_{ij}^{\text{Batch-Pack}}$ sollten daher sehr klein sein. Hingegen kann der Temperatureffekt auf die Steigung sehr wohl vom Packungstyp abhängen, so dass $\beta_{jk}^{\text{Pack-Temp}}$ zur Sicherheit im Modell verbleiben sollte. Damit ergibt sich schliesslich das reduzierte Modell in Formel 3.

$$Y(t) = \alpha_0 + \alpha_i^{\text{Batch}} + \beta_i^{\text{Batch}} \cdot t + \beta_j^{\text{Pack}} \cdot t + \beta_k^{\text{Temp}} \cdot t + \beta_0 \cdot t + \beta_{jk}^{\text{Pack-Temp}} \cdot t + \varepsilon$$

Formel 3

Eine explizite Angabe von Formeln für die Berechnung der Koeffizienten des Modells ist mit akzeptablem Aufwand nur für vollständige balancierte Versuchspläne möglich. Schon wenn eine Messung zu einem einzigen Zeitpunkt ausfällt, ist die Balance des Versuchsplans gestört und die Angabe einfacher Formeln wird unmöglich, weil die Schätzungen der Koeffizienten nun voneinander abhängen (korreliert sind). Die Auswertung sollte daher mithilfe von Statistik-Software durchgeführt werden, die multiple Regression für generelle lineare Modelle (GLM) beherrscht. Wenn allerdings der zugrundeliegende Versuchsplan ungeeignet ist, versagt auch die avancierteste Software.

Die frühzeitige Festlegung auf ein bestimmtes mathematisches Modells ist von enormer Wichtigkeit, weil eine optimale Versuchsplanung immer nur für ein gegebenes Modell erfolgen kann. So kann es z.B. passieren, dass ein für das Modell in Formel 3 optimierter Plan nicht mehr imstande ist, Auskunft über die Wechselwirkung $\beta_{ijk}^{\text{Batch-Pack-Temp}}$ zu geben. Ob die Reduktion von Formel 2 zu Formel 3 vor der Studiendurchführung statthaft ist, muss von Fall zu Fall mit den Behörden bei der Studienplanung verhandelt werden. Eine weitere Reduktion des Modells in Formel 3 ist vor der Studiendurchführung auf keinen Fall sinnvoll. Erst wenn die Messungen vorliegen, kann durch statistische Tests der Koeffizienten entschieden werden, ob weitere Terme aus dem Modell entfallen dürfen, weil sie kleiner als eine relevante Grenze sind. Das Herausnehmen einzelner Terme des Modells bedeutet das Zusammenfassen der Stufen des betreffenden Faktors. Wenn der statistische Test erlaubt, z.B. β_j^{Pack} gleich Null zu setzen, so bedeutet dies, dass die Regression über alle Packungsarten gemittelt erfolgt. Die Messwerte der unterschiedlichen Verpackungen werden dann als einfache Messwiederholungen aufgefasst und kommen dadurch der Präzision der Schätzung der im Modell verbleibenden Faktoren zugute. Daher ist die Prüfung der Modellkoeffizienten von herausragender Bedeutung.

3.3 Test der Einflüsse der Faktoren

Wie immer beim statistischen Beurteilen, so gibt es auch bei der Beurteilung der Modellkoeffizienten zwei Risiken:

1. Das Risiko α einen Effekt zu folgern, wo in Wahrheit kein Effekt vorliegt (also der Koeffizient in Wahrheit Null ist)
2. Das Risiko β , einen tatsächlich vorhandenen relevanten Effekt zu übersehen, also einen Modellterm zu entfernen, der in Wahrheit einen relevanten Wert übersteigt.

Ziel behördlicher Kontrolle ist, sicherzustellen, dass β für einen relevanten Effekt ein bestimmtes Mass nicht übersteigt. Für den Hersteller ist hingegen das Risiko erster Art α wichtig, denn er möchte vermeiden, dass ein in Wahrheit stabiles Produkt versehentlich als instabil deklariert wird. Daher wird α auch Produzentenrisiko und β Konsumentenrisiko genannt. $1 - \beta$ wird auch Trennschärfe (Power) genannt. Dies ist die Wahrscheinlichkeit einen in Wahrheit relevanten Effekt auch als signifikant zu beurteilen.

Um eine bestimmte Trennschärfe zu erreichen, muss je nach Präzision des Assays eine Mindestanzahl von Messungen in die Auswertungen einbezogen werden. Denn alle statistischen Tests werden gegen die Residualvarianz gemacht. Und diese wird um so zuverlässiger geschätzt, je mehr Messungen in ihre Schätzung einbezogen werden.

Ebenso bewirkt eine Erhöhung des Irrtumsniveaus α ein Absinken von β , denn wenn man eher zulässt, dass etwas Gleiches als unterschiedlich gesehen wird, so wird man erst recht tatsächlich Unterschiedliches als unterschiedlich erkennen. Dieser Umstand bewog FDA dazu, zur Sicherstellung einer hohen Trennschärfe ein α -Niveau von 25 % zu empfehlen^{7,8,9,10}. Wir lehnen dieses Verfahren ab, weil diese undifferenzierte Pauschalmaßnahme keine Rücksicht auf den jeweils konkret vorliegenden Versuchsplan nimmt. Moderne Software ermöglicht die Darstellung der Trennschärfe für jeden der Modellterme eines noch so komplizierten Modells, so dass die Einhaltung einer Mindesttrennschärfe auf jeden Fall kontrollierbar ist. Damit ist das pauschale Heraufsetzen des Irrtumsniveaus obsolet.

Eine absurde Konsequenz aus dem hohen α -Wert für ist, dass im Falle vieler einbezogener Faktoren durch das mehrfache simultane Testen nahezu sicher ist, dass irgendeiner dieser Tests einen signifikanten Befund ergibt, auch wenn alle Faktoreinflüsse in Wahrheit Null sind. Daher muss $\alpha = 0.25$ als nicht praxisgerecht und nicht mehr zeitgemäss abgelehnt werden.

Die Trennschärfe ist nicht nur eine Funktion der Anzahl Messungen N , der Präzision des Assays und des Irrtumsniveaus α , sondern auch der Grösse des wahren Koeffizienten: Grosse Effekte sieht man eher als kleine. Zur Trennschärferechnung muss man also festlegen, was als relevanter Effekt anzusehen ist. Diese Relevanzgrenze, deren Überschreitung man mit einer gegebenen Trennschärfe testen möchte, muss vor Studienbeginn für jeden Faktor mit der Aufsichtsbehörde ausgehandelt werden. Hierbei stellen die sog. Gütekurven eine grosse Hilfe dar (siehe **Fehler! Verweisquelle konnte nicht gefunden werden.**). Sie zeigen für einen gegebenen Versuchsplan und ein gegebenes α -Niveau, mit welcher Wahrscheinlichkeit ein wahrer Effekt von dem statistischen Test als signifikant erachtet werden wird. Aus der ersten Kurve in **Fehler! Verweisquelle konnte nicht gefunden werden.** kann z.B. gefolgert werden, dass bei Verwendung des vollfaktoriellen Versuchsplans ein Achsenabschnittsunterschied zwischen zwei Chargen von ca. 0.5 mg mit einer Wahrscheinlichkeit von mindestens 80 % auch als signifikant auf dem 5% Niveau entdeckt wird.

Für jeden der Modellterme aus Formel 2 oder Formel 3 sollte vor Studienbeginn eine solche Gütefunktion vorgelegt werden, um zu zeigen, dass die Trennschärfe für die Entdeckung relevanter Effekte ausreicht. Wenn die Trennschärfe als ausreichend erachtet wurde, so kann bei der späteren Auswertung in gewohnter Weise auf dem $\alpha = 5\%$ Niveau getestet werden. Es werden dann alle Koeffizienten, die zu $p(H_0) > 0.05$ führen, aus der Modellgleichung gestrichen, solange bis nur noch signifikante Terme im Modell übrigbleiben. In unserer Praxis wurde dieses Verfahren von FDA bereits als wertvoll anerkannt.

Eine Alternative, die ohne die Berechnung von Gütekurven auskommt, ist die Durchführung einseitiger T-Tests an den Relevanzlimiten. Dieses aus der Bioäquivalenzstatistik stammende Verfahren¹¹ ist für den Praktiker sehr einfach durchführbar, indem geprüft wird, ob die 95% - einseitigen Vertrauensbereiche der Koeffizienten innerhalb der Limiten liegen. Weil bei diesem Verfahren lediglich die Hypothesen umgekehrt werden (die Nullhypothese ist die der Überschreitung der Limite, die zu 'beweisende' Alternative ist die der Äquivalenz), ist zwar das Konsumentenrisiko unter Kontrolle, aber das Produzentenrisiko ist unkontrolliert. Dieses Verfahren ist besonders effektiv für Situationen mit kleiner Residualvarianz. Dies ist bei Haltbarkeitsstudien aber selten der Fall.

Welcher Test auch immer zur Anwendung kommt, so ist die Abfolge der Tests der einzelnen Modellterme immer von den Termen höherer Ordnung zu denen niedrigerer¹² (sog. Modellhierarchie). So macht die Betrachtung der Terme für den Effekt der Chargen auf den Achsenabschnitt α_i erst dann Sinn, wenn zuvor sichergestellt wurde, dass eine gemeinsame Steigung für alle Chargen angemessen ist, dass also alle $\beta_i = 0$ sind.

3.4 Modelle mit zufälligen Effekten („Random Effects Models“)

Die Produktionslose (Batches) werden beim bisherigen Ansatz als fester Effekt („fixed effect“) im Modell aufgenommen. Streng genommen darf man dann nur Aussagen über die drei Batches machen, die in die Studie einbezogen sind. Dagegen ist die Forderung der Behörden nach drei Losen sicher durch die Vorstellung begründet, damit eine repräsentative Zufallsstichprobe aus dem späteren Produktionsgeschehen zu erhalten. Daher findet man in der Literatur Ansätze, die die Batches als Zufallseffekte berücksichtigen^{13,14,15}. Bei der Varianzkomponentenschätzung wird getestet, ob die Varianz zwischen den Chargen signifikant grösser als Null ist.

Dieser Ansatz wird hier aus folgenden Gründen nicht weiter verfolgt:

- Bei einem Stichprobenumfang von nur 3 Produktionslosen ist eine Schätzung der „Batch zu Batch“ Streuungskomponente mit einem sehr breiten Vertrauensintervall behaftet. Eine Varianzschätzung erfordert vernünftigerweise 20 und mehr Einheiten.
- Die algorithmische Bewältigung im Rahmen eines Modells mit sonst vielen festen Effekten ist nicht trivial und führt zu iterativen Methoden (restricted maximum likelihood: REML).
- wenn die Batches nicht rein zufällig hergestellt sind, sondern Grenzmuster darstellen (Batch 1: untere Grenze, Batch 2: Sollwert, Batch 3: obere Grenze) ist das Modell mit festen Effekten angemessen.

4. Versuchsplanoptimierung

4.1 Vollfaktorieller Versuchsplan

Der sicherste wenn auch aufwendigste Versuchsplan ist ein vollfaktorieller Plan, also ein Plan, bei dem jede Stufe jedes Faktors mit jeder Stufe aller anderen Faktoren kombiniert wird. In Tabelle 1 ist ein solcher Plan für 3 Chargen (Batches), 2 Klimagruppen (25 °C, 30 °C) und 2 Verpackungsarten (A, B) angegeben, wobei der Übersichtlichkeit wegen nur bis 12 Monate geplant wurde.

Tabelle 1 Vollfaktorieller Versuchsplan für 3 Chargen, 2 Temperaturen und 2 Verpackungen

Zeit in Monaten	Batch 1				Batch 2				Batch 3				
	25 °C		30 °C		25 °C		30 °C		25 °C		30 °C		
	A	B	A	B	A	B	A	B	A	B	A	B	
0	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X	X
6	X	X	X	X	X	X	X	X	X	X	X	X	X
9	X	X	X	X	X	X	X	X	X	X	X	X	X
12	X	X	X	X	X	X	X	X	X	X	X	X	X
18	X	X	X	X	X	X	X	X	X	X	X	X	X
24	X	X	X	X	X	X	X	X	X	X	X	X	X
36	X	X	X	X	X	X	X	X	X	X	X	X	X

(X gibt an, dass eine Messung durchgeführt wird)

Dieser Versuchsplan erfordert insgesamt 60 Einzelbestimmungen (5 Zeitpunkte mal 12 mögliche Varianten). Mit einem solchen Plan können alle Modellterme des vollen Modells mit allen möglichen Wechselwirkungen nach Formel 1 geschätzt werden. Wie oben gezeigt, ist die universelle Schätzbarkeit aller möglichen Modellterme in vielen Fällen gar nicht erforderlich. Aus diesem Grund kann um den Preis der Schätzbarkeit einiger Koeffizienten auf einen Teil der Messungen verzichtet werden.

4.2 Sukzessiver D-optimaler Versuchsplan

Zur Einsparung von Versuchen bei Stabilitätsstudien wurden klassische teilfaktorielle Versuchspläne vorgeschlagen¹⁶. Diese Pläne haben gute statistische Eigenschaften, sind jedoch nicht sehr flexibel an gegebene Forschungssituationen und Modelle anpassbar. Inzwischen erlauben elaborierte Softwarealgorithmen jedoch massgeschneiderte Lösungen, sogenannte D-optimale Versuchspläne. Der Algorithmus sucht aus der Menge aller theoretisch möglichen Experimente iterativ diejenigen heraus, die zu einer maximalen Präzision der Koeffizientenschätzer führen.

Ein direkter Einsatz dieser Versuchspläne für Haltbarkeitsstudien ist nicht ohne weiteres möglich, da der Software der zeitlich aufeinander aufbauende Verlauf der Modelle mit Zwischenanalysen zu den ver-

schiedenen Untersuchungszeitpunkten nicht nahegebracht werden kann. Dieses Handicap lässt sich jedoch sehr einfach durch sukzessiv aufeinander aufbauende Versuchspläne lösen, wie an einem Plan für das Modell nach Formel 3 gezeigt werden soll.

Generiert man für die zwei Stufen 0 und 3 Monate des Faktors Zeit ein D-optimales Design für 3 Chargen, die Verpackungen A und B und die Temperaturen 25 °C und 30 °C, so könnte der Plan aus Tabelle 2 a entstehen. In diesem Beispiel wurde angestrebt, die Anzahl der Messungen zu halbieren. Der Algorithmus muss also 12 Messungen den 24 Planzellen so zuordnen, dass alle Effekte schätzbar und die Fehler in den Koeffizientenschätzern minimal werden. Nur bei bestimmten Anzahlen von Faktoren und ihrer Stufen ergeben sich dabei Pläne, die auch optisch symmetrisch wirken (sog. klassische Pläne). Im allgemeinen sollte man sich aber der numerischen Plansuche bedienen, da nur diese einen freien sukzessiven Aufbau erlaubt.

Tabelle 2 D-optimaler Versuchsplan

a) Teilplan von 0 bis 3 Monate

Zeit in Monaten	Batch 1				Batch 2				Batch 3				
	25 °C		30 °C		25 °C		30 °C		25 °C		30 °C		
	A	B	A	B	A	B	A	B	A	B	A	B	
0	X		X			X		X	X				X
3	X	X						X	X		X	X	

b) Der D - optimale Versuchsplan für 3 Chargen, 2 Temperaturen und 2 Verpackungen bis 12 Monate

Zeit in Monaten	Batch 1				Batch 2				Batch 3				
	25 °C		30 °C		25 °C		30 °C		25 °C		30 °C		
	A	B	A	B	A	B	A	B	A	B	A	B	
0	X		X			X		X	X				X
3	X	X						X	X		X	X	
6		X						X				X	
9				X	X					X			
12	X	X				X	X				X	X	

(X gibt an, dass eine Messung durchgeführt wird)

Dieses Design wurde nun sukzessiv für die fortschreitende Untersuchungszeit erweitert, indem das jeweils vorhergehende Design dem Algorithmus vorgegeben wurde (als sog. 'inclusions'). Die Tabelle 2 b zeigt das Design nach 12 Monaten, wobei jeweils 3, 3 und 6 Beobachtungen hinzugefügt worden sind. Insgesamt ergeben sich 24 Messungen, also eine Einsparung von 36 der sonst erforderlichen 60 Messungen. Dennoch sind alle Modellterme aus Formel 3 schätzbar. Es gibt bisweilen Schwierigkeiten mit FDA bei Plänen, die beim ersten und beim letzten Zeitpunkt nicht alle verbleibenden Faktorenkombinationen abdecken. Da diese Einschränkung statistisch nicht begründet ist, wird sie hier nicht berücksichtigt.

4.3 Qualitätskriterien für Versuchspläne

Die Pläne müssen verhindern, daß relevante Effekte aufgrund der Reststreuung oder der Vermengung von Effekten nicht erkannt werden. Bei der Planung von Experimenten sollte man daher stets sicherstellen, dass für die Schätzung der Restvarianz genügend Freiheitsgrade übrig bleiben. In unserem Beispiel verbraucht der Faktor Charge und seine Wechselwirkungen zwei Freiheitsgrade während alle anderen Effekte nur einen Freiheitsgrad beanspruchen. Daher verbleiben bei der Auswertung nach 12 Monaten $24 - 3 \cdot 2 - 5 \cdot 1 = 13$ Freiheitsgrade. Dies sollte für eine solide Schätzung ausreichen. Beim Zeitpunkt 3 Monate liegen nur 12 Messungen vor, so dass nur ein Freiheitsgrad für die Restvarianzschätzung verbleibt. Daher ist das Testen der Koeffizienten schon nach 3 Monaten noch sehr unsicher.

Neben der ausreichenden Anzahl Freiheitsgrade muss ein gutes Design gewährleisten, dass die Faktoren untereinander unkorreliert sein und wie schon erwähnt sollte die Trennschärfe des Plans ausreichend sein, relevante Effekte auch zu erkennen.

Die Unkorreliertheit kann sehr einfach durch Betrachtung der Korrelationsmatrix der Modellterme geprüft werden.

Tabelle 3 zeigt die aus diesem Design resultierende prozentuale Korrelation der Koeffizientenschätzer untereinander in Form einer Matrix.

Tabelle 3 Korrelationsmatrix der Modellkoeffizienten bei der Auswertung nach 12 Monaten

	α_0	α_{i1}	α_{i2}	β_0	β_{i1}	β_{i2}	β_j	β_k	β_{jk}
α_0	100	1	0	9	4	-3	-3	3	14
α_{i1}	1	100	-50	0	1	-12	32	19	10
α_{i2}	0	-50	100	0	1	13	-7	-24	6
β_0	9	0	0	100	1	-1	-2	2	3
β_{i1}	4	1	1	1	100	-46	-24	-7	25
β_{i2}	-3	-12	13	-1	-46	100	-7	-30	-19
β_j	-3	32	-7	-2	-24	-7	100	14	1
β_k	3	19	-24	2	-7	-30	14	100	0
β_{jk}	14	10	6	3	25	-19	1	0	100

Es besteht keine Korrelation grösser als 50 %. Die Indices 1 und 2 beim Faktor Charge deuten an, daß bei allen Effekten, bei denen dieser Faktor vorkommt, 2 Freiheitsgrade verbraucht werden, weil dieser Effekt kategorial ist und auf 3 Stufen eingestellt wird.

Für die Beurteilung der Trennschärfe des Versuchsplans wurden mit der RS/DISCOVER Versuchsplanungssoftware³ die Gütekurven für einen Test auf dem $\alpha = 5\%$ Niveau zu jedem Auswertungszeitpunkt für alle Modellkoeffizienten berechnet. Dabei wurde eine Reststandardabweichung von 1 mg angenommen. In der Praxis wird man hier zunächst die Standardabweichung aus der Validierung der 'intermediate precision' - der analytischen Methode annehmen.

und Fehler! Verweisquelle konnte nicht gefunden werden. zeigen exemplarisch die Verläufe für die Detektion von Chargenunterschieden bezüglich Achsenabschnitt bzw. Steigung. Als Vergleich ist die Gütefunktion des vollfaktoriellen Plans nach Tabelle 1 eingezeichnet.

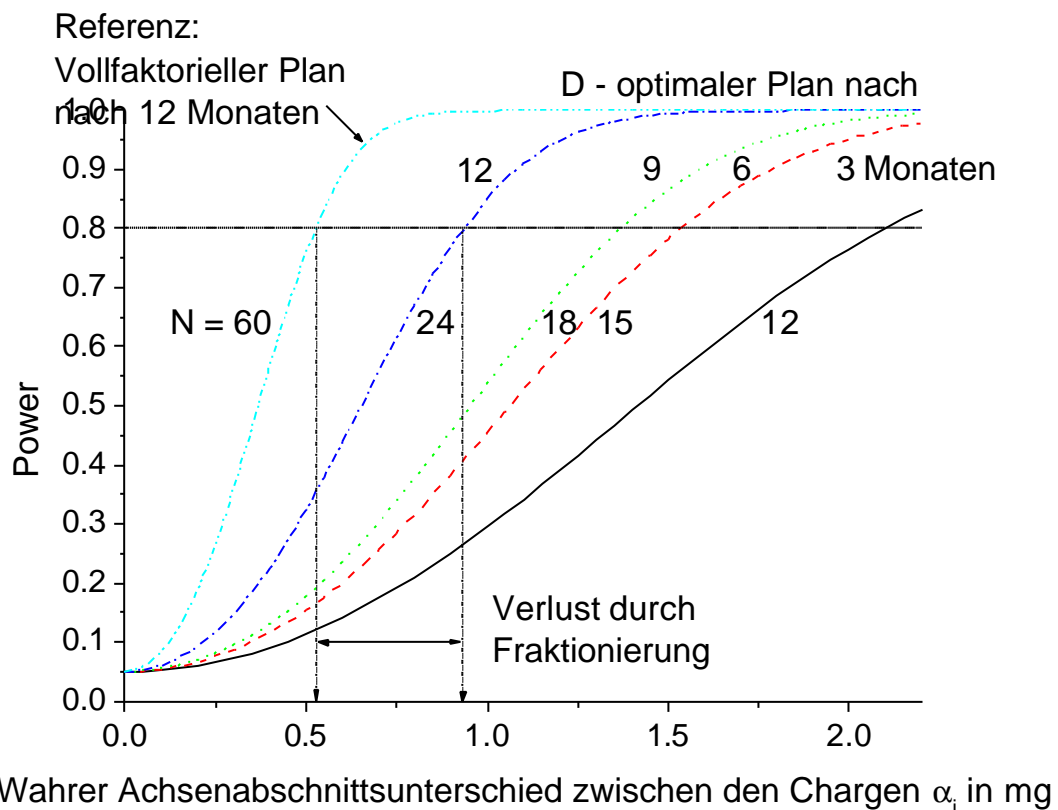


Abbildung 1 Die Gütefunktion für den Achsenabschnittsunterschied zwischen den Chargen bei Auswertung zu unterschiedlichen Zeitpunkten

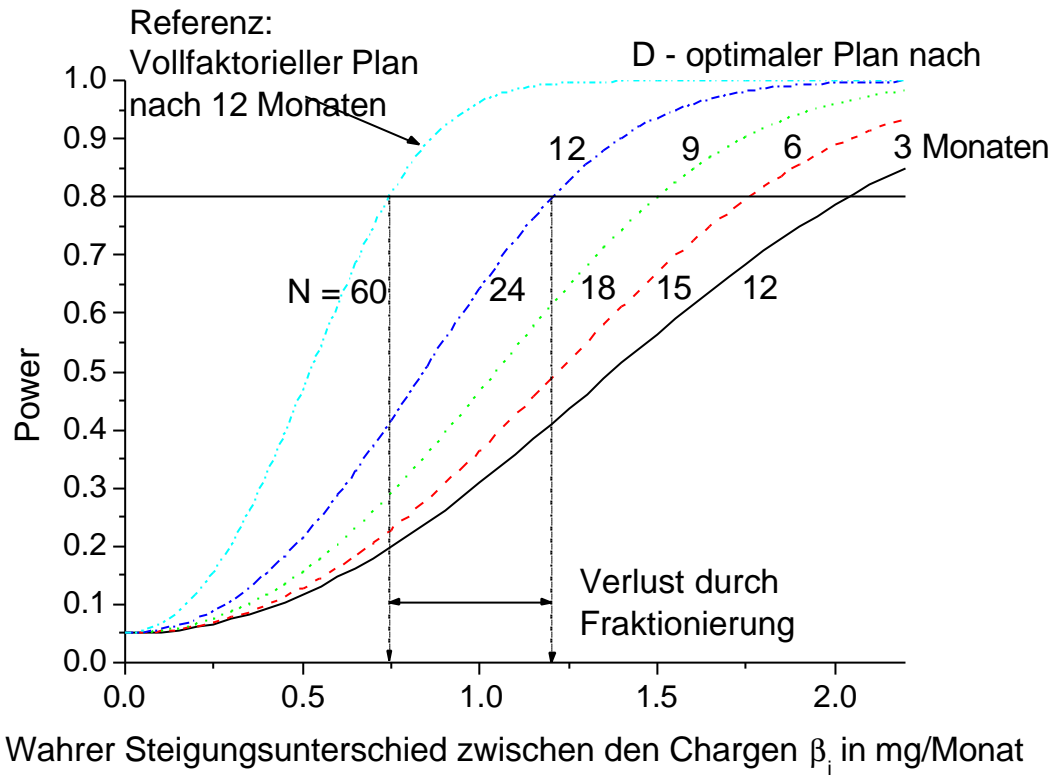


Abbildung 2 Die Gütefunktion für den Steigungsunterschied zwischen den Chargen bei Auswertung zu unterschiedlichen Zeitpunkten

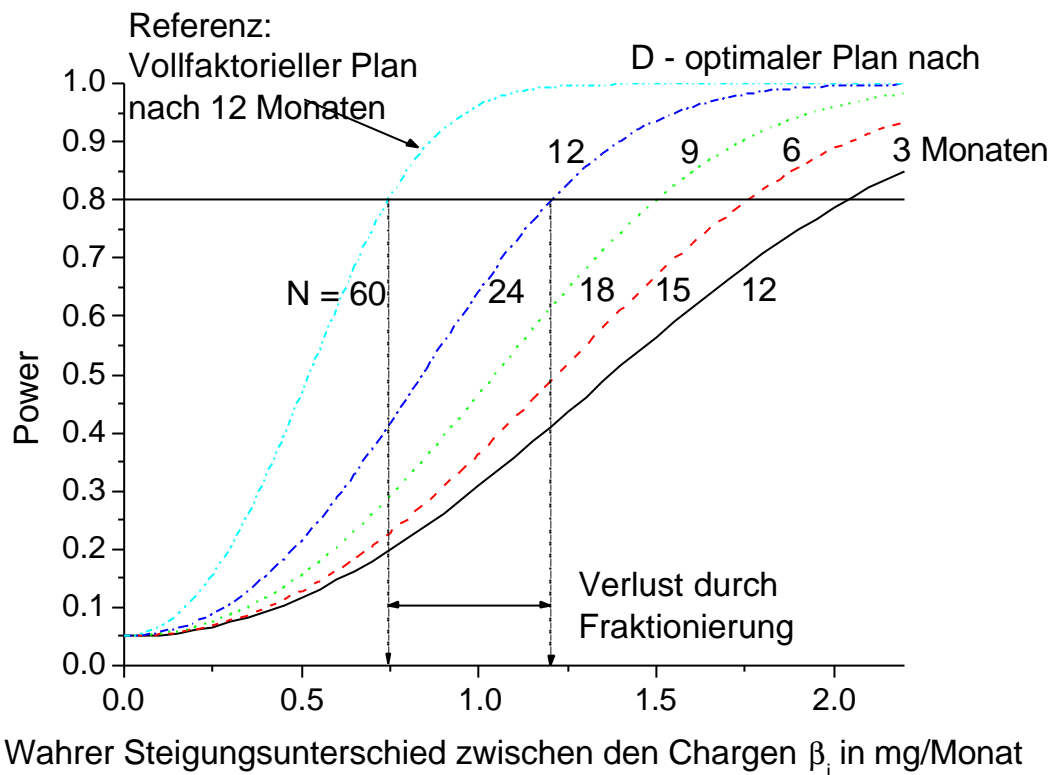


Abbildung 2 zeigt, dass bei der Auswertung nach 3 Monaten ein Chargenunterschied der Steigung vom Mittelwert von etwa 2 mg/Monat mit einer Sicherheit von 80 % auch als signifikant auf dem 5 % Niveau erkannt werden. Nach 12 Monaten verbessert sich dieser Wert auf 1.2 mg/Monat und beim vollfaktoriellen Plan ergäbe sich eine Empfindlichkeit von etwa 0.75 mg/Monat. Diesem Empfindlichkeitsverlust von 60 % durch die Fraktionierung des Versuchsplans steht der 60 - prozentige Gewinn durch die Einsparung von Messungen gegenüber.

Für das Aushandeln der konkreten Versuchspläne besteht noch ein grosser Spielraum, weil es noch keine verbindlichen Standards für die geforderten Relevanzgrenzen bzw. Trennschärfen gibt. Mit Sicherheit sind Trennschärfen unter 80 % unakzeptabel. Eine Trennschärfe von 50 % bedeutet, dass relevante Effekte nur rein zufällig entdeckt werden würden. Die Entscheidung, ab welcher Grösse ein Effekt als relevant anzusehen ist, ist ausschliesslich auf Basis chemisch, pharmazeutisch, toxikologischer Kriterien zu fällen. Die Statistik liefert lediglich die Auskunft darüber, mit welcher Sicherheit diese Relevanzlimiten eingehalten werden können. Diese klare Trennung ist ein Schritt zu mehr Arzneimittelsicherheit.

5. Die Schätzung der Haltbarkeitsdauer

Die Haltbarkeitsdauer ist derjenige Zeitraum, in dem der Gehalt (die Verunreinigung) im Mittel mit 95 % iger statistischer Sicherheit oberhalb (unterhalb) der festgelegten Limite liegt. Daher kann die HD aus dem Zeitpunkt des Schnittpunkts des 95 % Vertrauensbandes für den Fit (CLF = confidence limit for the fit) mit dem einzuhaltenden Grenzgehalt (GG) geschätzt werden.

Abbildung 3 zeigt die Situation für den einfachsten Fall $Y(T) = \alpha_0 + \beta_0 \cdot T$. Für kompliziertere Modelle wie z.B. Formel 3 gilt sinngemäss das gleiche, nur dass dann statt einer Geraden eine Fläche bzw. Hyperfläche vorliegt.

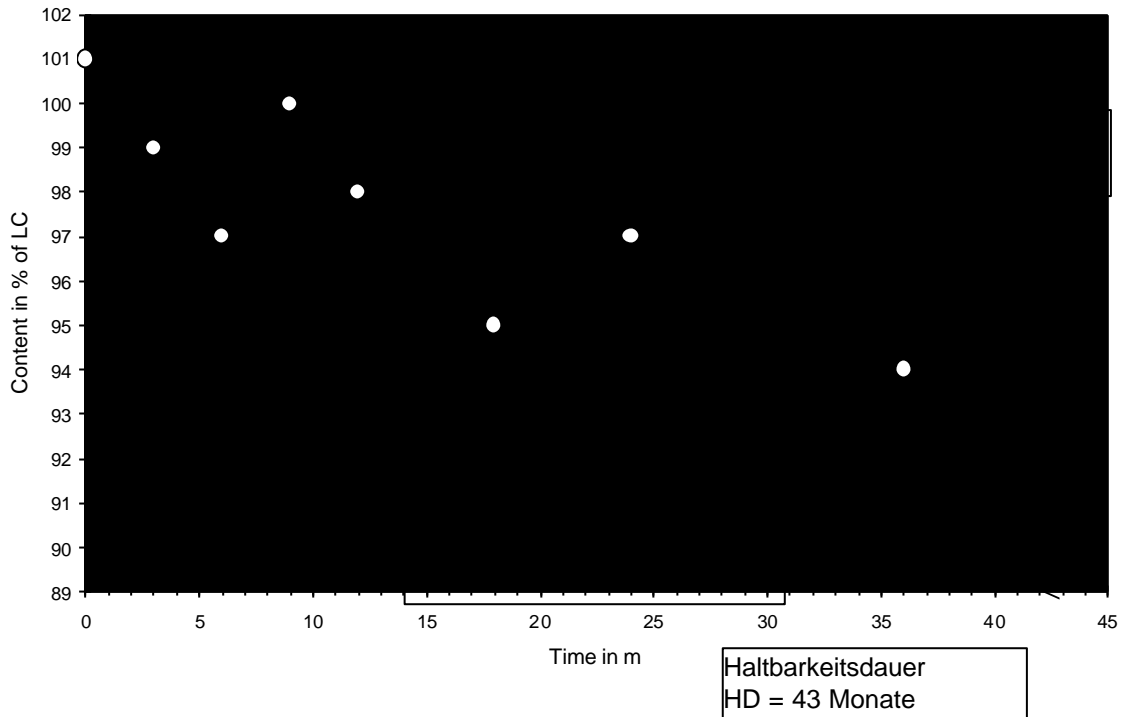


Abbildung 3 Die Berechnung der Haltbarkeitsdauer

Die Interpretation des Vertrauensbereichs für den Fit CLF ist:

Wenn die gefundene Fitgrade $\bar{Y}(T)$ der wahren Zerfallsfunktion $Y(t)$ entspricht und die gesamte Haltbarkeitsstudie 100 mal wiederholt wird, so wird voraussichtlich nur in 5 Fällen aufgrund der Meßstreuung die sich ergebende Regressionsgerade auch mal unterhalb des Vertrauensbands für die Fitgerade liegen. Das Vertrauensband für die Fitgerade lautet für Gehaltsbestimmungen:

$$\begin{aligned}
 CLF(T) &= \bar{Y}(T) - t_{1-\alpha, dF_{Residual}} \cdot \sqrt{\frac{SS_{Residual}}{N} + \frac{SS_{Residual}}{\sum_{j=1}^N (t_j - \bar{t})^2} \cdot (T - \bar{t})^2} \\
 &= \bar{Y}(T) - t_{1-\alpha, dF_{Residual}} \cdot \sqrt{Var_{Residual} + Var(\bar{b}) \cdot (T - \bar{t})^2}
 \end{aligned}$$

Formel 4

Hierin sind:

- $\bar{Y}(T)$: Fitfunktion zum Zeitpunkt T, $\bar{Y}(T) = \bar{a} + \bar{b} \cdot T$
- N: Anzahl aller Datenpunkte
- $t_{1-\alpha, dF_{Residual}}$: Student's t- Faktor für 1- α prozentige statistische Sicherheit und $dF_{Residual}$ Freiheitsgrade. Im Falle einfacher linearer Regression gilt $dF_{Residual} = N-2$, allg. gilt $dF_{Residual} = N-k$, wobei k die Anzahl der Fitparameter ist.
- $SS_{Residual}$: Summe der Residuen = Mass für die Streuung der Daten um die Fitgerade herum.
- $Var_{Residual}$: Restvarianz
- $Var(\bar{b})$: Varianz der Steigungsschätzung
- \bar{t} : Mittelwert aller Zeitpunkte

Die Halbarkeitsdauer HD lässt sich nun durch die Gleichsetzung $GG = CLF(HD)$ und Auflösen nach HD ausrechnen. Dies führt auf eine in HD quadratische Gleichung.¹⁷

Aus den beiden Varianzbeiträgen in der Wurzel von Formel 4 wird ersichtlich, wo statistische Versuchspannung eingreift:

1. Durch eine grosse Anzahl von einbezogenen Messungen N werden der Student-Faktor und die Restvarianz der Regression klein, so dass sich eine schlankere Taille der Funktion CLF im Schwerpunkt der Daten ergibt. \Rightarrow die Haltbarkeitsdauerschätzung steigt.

2. Durch kluge Wahl der Messzeitpunkte wird die Summe der quadrierten Zeitabweichungen im Nenner des rechten Terms gross und damit die Varianz der Steigungsschätzung klein. \Rightarrow die Aufweitung der CLF - Kurve zu den Enden hin wird geringer und die Haltbarkeitsdauerschätzung steigt.

Wegen des Arguments No. 1 ist es sinnvoll, immer alle möglichen Faktoren in die Auswertung einzubeziehen. Denn ganz unabhängig davon, ob man nun Chargen oder Packungstypen zusammenfassen (poolen) darf, wird die Restvarianz des Fits immer auf Basis aller gemachten Messungen geschätzt. Wenn man z.B. für jede Charge in jeder Packungsart bei jeder Temperatur eine separate lineare Regression machen würde, so ergäbe sich jeweils ein N von 5 bzw. $dF_{\text{Residual}} = 5-2 = 3$. Der Student-Faktor ist bei dieser enorm kleinen Fallzahl $t_{95\%,3} = 2.353$. Bei simultaner Auswertung z.B. durch Regression nach Formel 2 ergibt sich $N = 60$, $dF_{\text{Residual}} = 47$ und $t_{95\%,47} = 1.678$ beim vollfaktoriellen Design bzw. $N=24$, $dF_{\text{Residual}} = 11$ und $t_{95\%,11} = 1.796$ bei obigem D - optimalen Plan. Damit vermindert

sich die Breite des Vertrauensbereichs auf $\frac{1.796}{2.353} \cdot \sqrt{\frac{5}{24}} = 34.8\%$ bzw. $\frac{1.678}{2.353} \cdot \sqrt{\frac{5}{60}} = 20.6\%$

des Wertes bei separater Auswertung. Das bedeutet, dass in diesem Beispiel allein durch die Anwendung angemessener Auswertemethoden eine bis zu fünffach höhere Präzision der Haltbarkeitsdauerschätzung möglich ist. Die Haltbarkeitsdauerschätzung verlängert sich dadurch beträchtlich.

Im Lichte des Arguments No.2 erscheint die derzeitige Praxis, regelmässig bei 0, 3, 6, 9, 12, 18, 24 und 36 Monaten zu messen, als wenig förderlich für eine optimale Schätzung der Steigungskoeffizienten. Theoretisch ist für Linearfunktionen ein Versuchsplan optimal, der nur bei $t=0$ und bei $t=36$ Monate je die Hälfte der Messungen plant. Je näher eine Messung dem Mittel aller Zeitpunkte kommt, desto weniger trägt diese Messung zur Verbesserung der Steigungsschätzung und damit der Haltbarkeitsdauer bei. Eine tendentielle Bündelung der Messungen zu Beginn und am Ende bzw. zu den wichtigen Zwischenauswertungen sollte daher unbedingt angestrebt werden. Im oben vorgestellten D - optimalen Versuchsplan könnte man z.B. auf die 6 - Monatsmessungen verzichten und die freigewordenen Kapazitäten besser auf 0 und 12 Monate aufteilen. Ein strikt optimaler Plan, also nur noch bei 0 und 36 Monaten zu messen, muss jedoch als nicht praxismässig verworfen werden, weil das Risiko von Fehlmessungen nie mehr korrigiert werden kann. Ausserdem lassen sich so nie Abweichungen von der Linearität testen. Ein Minimum von 3 Messzeitpunkten ist daher zwingend.

6. Lohnen sich Messwiederholungen ?

In manchen Situationen ist es sinnvoll, an einigen Messzeitpunkten (vornehmlich denen an den beiden Enden der Zeitskala) Mehrfachmessungen durchzuführen. Für einfach zu messende technisch-physikalische Charakteristika wie z.B. der Bruchhärte, Wasseraufnahme oder dem pH kann man durch wiederholte Messungen zu einem Zeitpunkt die Präzision der Koeffizientenschätzungen erhöhen, weil ja das N und die Präzision der Steigungsschätzung erhöht werden.

Bei aufwendigeren chemischen Analysen gilt dies jedoch nur eingeschränkt, weil die Ergebnisse wiederholter Assays nur die Wiederholbarkeitsstreuung ('repeatability precision') und eventuelle Produktinhomogenität abdecken, die in der Regel deutlich kleiner ist als die langfristige 'intermediate precision' - Standardabweichung. Aber die Letztere ist es, die durch die Residualvarianz um die Fitfunktion herum geschätzt wird. 'Replicate Assays' unter Wiederholungsbedingungen würden diese Streuung in unredlicher Weise unterschätzen. Dies kann aber schnell entdeckt werden, wenn im $Y(T)$ - Plot die Daten an den einzelnen Zeitpunkten jeweils eng gruppenweise beieinanderliegen, diese Gruppen jedoch stark um die Regressionsgrade herum streuen. Aus diesem Grunde sind wiederholte Messungen nur bei solchen Methoden sinnvoll, bei denen die Wiederholstandardabweichung kaum kleiner ist als die der langfristigen 'intermediate precision'. In der Regel erlauben die Daten der Haltbarkeitsstudie erstmals eine wirklich gute Schätzung der 'intermediate precision' - Standardabweichung, denn in den seltensten Fällen wird während der Methodvalidierung die Systemstreuung der analytischen Methode wirklich langfristig erfasst.

7. Diskussion

Die klar umrissenen Ziele einer Haltbarkeitsstudie, Erhalt einer möglichst präzisen Schätzung der Haltbarkeitsdauer und zuverlässige Ermittlung der Effekte der beteiligten Faktoren, lassen sich durch die vorgestellten Versuchsplanungs- und Regressionsmethoden mit minimalem Aufwand optimal erreichen. Welches statistische Modell dabei *ab initio* verwendet werden muss, ist eine Frage der Verhandlung über den konkreten Einzelfall. Insbesondere wenn über die Homogenität innerhalb der Chargen und Verpackungen (content uniformity) noch keine verlässliche Information vorhanden ist, sollte man bei der

Versuchsplanung eher zum Modell nach Formel 1 tendieren. Auch gibt es bei einigen Zulassungsbehörden Vorbehalte gegenüber einigen teilfaktoriellen Plänen. So ist von FDA zu lesen "Environmental factors should not be matrixed..."¹. Dies darf nicht unwidersprochen bleiben, weil keinerlei Begründung geliefert wurde. Die Entscheidung über das Herausnehmen bestimmter Faktorenkombinationen aus dem vollfaktoriellen Plan sollte ausschliesslich durch deren Relevanz für das Testen wichtiger Effekte begründet sein. Wenn ein wichtiger Effekt auch nach "Ausdünnen" des Plans immer noch mit einer bestimmten Trennschärfe getestet werden kann, so spricht nichts gegen dieses Vorgehen. Es ist Sache der Behörden festzulegen, was wichtige Effekte sind und mit welcher Trennschärfe diese geprüft werden müssen. Es liegt dann allein an der Industrie, geeignete Pläne zu entwerfen, die diese beiden Anforderungen angemessen erfüllen und das Produzentenrisiko nicht zu gross werden lassen. Ein darüber hinausgehender willkürlicher Behördeneingriff in die Versuchsplanung ist nicht akzeptabel.

Der Besonderheit vieler pharmazeutischer Assays, bei langfristiger Anwendung stärkere Streuung aufzuweisen als bei kurzfristiger Wiederholung wurde durch Einführung eines Modells mit hierarchischer Fehlerstruktur Rechnung getragen¹⁸. Dieser Ansatz erfordert auf jeden Fall Messwiederholungen an den einzelnen Zeitpunkten, aus denen jedoch auch nur das geschätzt werden kann, was doch aus der Methodvalidierung bereits hinlänglich bekannt ist, nämlich die kurzfristige Wiederholpräzision und evtl. die Produktinhomogenität. Diese ist jedoch für das Testen der langfristigen Effekte in Haltbarkeitsstudien unbrauchbar. Vielmehr sollten die Analytiker alle erdenkliche Mühe darauf verwenden, etwaige Ursachen für eine erhöhte Langzeitvarianz zu ergründen und abzustellen. Die routinemässige Ko-Analyse von Proben mit eingewogenen Aliquots des Wirkstoffs bzw. seiner Abbauprodukte ('control samples') scheint beim Assay pharmazeutischer Formulierungen nicht immer hinzureichen, die langfristigen Abweichungen zu korrigieren. Hier scheinen bisher unerfasste Varianzursachen beim Aufschluss der pharmazeutischen Form eine Rolle zu spielen. Daher ist es in diesen Fällen eventuell angezeigt, tiefgefrorenes Material der Formulierung mitzuanalysieren. Statt die Ergebnisse der Analysen bei den anderen Temperaturen auf die Gehalte im Tiefkühlsample zu normieren und erst danach mit GLM auszuwerten, ist es probater, die Ergebnisse der als stabil angenommenen Begleitprobe unter Einschluss eines weiteren Faktors im linearen Modell aufzunehmen. Neben der Reduktion der Residualvarianz und damit der Verlängerung der Haltbarkeitsdauerschätzung stellt dieses Vorgehen sicher, dass eine etwaige relevante Instabilität des Tiefkühlsamples auch erkannt werden kann. Eine andere Möglichkeit, bisher ungefaßte analytische Unrichtigkeit kompensieren zu können, besteht darin, Placebo-Material bei den entsprechenden Temperaturen mit einzulagern und über das Spiken mit Wirkstoff eine matrixkontrollierte Auswertung zu erhalten. Praktische Erfahrungen hierzu liegen uns zur Zeit noch nicht vor.

Haltbarkeitsstudien sind ihrer Natur nach multivariat, denn es werden viele korrelierte Zielgrössen als Funktion der Zeit und der anderen Faktoren erfasst. Doch die Anwendung multivariater statistischer Methoden für diesen Zweck ist noch sehr wenig erforscht¹⁹. Ihr Nutzen läge in einer Erhöhung der Trennschärfe bei der Erkennung der simultanen Effekte der Einflussfaktoren. So könnten die Messgrössen z.B. einer Hauptkomponenten- oder Faktorenanalyse unterzogen werden, um zu ermitteln, welche Messgrösse die meiste Information über den Degradationsprozess repräsentiert. Zur Ermittlung der Haltbarkeitsdauer sollte allerdings nach wie vor die "worst case" - Messgrösse genommen werden und nicht etwa nur die erste Hauptkomponente. Die multivariaten Verfahren sind an bestimmte Voraussetzungen über die Kovarianzstruktur der Messungen gebunden, die bei der Diversität der verwendeten Assays und physikalischen Methoden nicht unbedingt erfüllt sind.

Die vorgestellten Methoden der Versuchsplanung und Auswertung sind sicher, denn das Konsumentenrisiko β ist in jedem Fall unter Kontrolle und es besteht volle Transparenz über die Eigenschaften des Versuchsplans.

Diese Verfahren sind besonders wirtschaftlich, denn es werden nur diejenigen Messungen durchgeführt, die für die Schätzung der relevanten Effekte wirklich notwendig sind.

Die Verfahren sind auch für den Nichtmathematiker durch den Einsatz moderner Software handhabbar.

In einem zweiten Teil in „Die Pharmazeutische Industrie“ zu diesem Artikel werden wir die Anwendung der Verfahren an einem Beispiel im Detail demonstrieren.

8. Literatur

- 1 ICH Sept. 1994. Q1A Stability Testing of New Drug Substances and Products + Annex: Reduced Stability Testing Plan - Bracketing & Matrixing
- 2 P. Wessels, F. Erni, K. Krummen and M. Holz. 1997. Statistical Evaluation of Stability Data of Pharmaceutical Products for Specification Setting. *Drug Dev Ind Pharm.*23(5):427-39
- 3 Brooks Automation GmbH, Freisinger Str. 32, D-85737 Ismaning (<http://www.brooks.com>)
- 4 Brooks Automation GmbH, Freisinger Str. 32, D-85737 Ismaning (<http://www.brooks.com>)
- 5 Statistical Graphics Corp. 2115 East Jefferson Street, Rockville, Maryland 20852 USA (<http://www.manugistics.com>)
- 6 Chen JJ Ahn H and Y Tsong 1997. Shelf-Life Estimation for Multifactor Stability Studies. *Drug Information journal* Vol 31:573-87
- 7 FDA(Feb. 1987) Guideline for Submitting Documentation for the Stability of Human Drugs and Biologics. Rockville, MD, Dpt. of Health and Human Service
- 8 ICH 1993. ICH Harmonized Tripartite Guideline "Stability testing of New Drug Substance and Products
- 9 Fairweather WR Lin TYD and R Kelly 1995. Regulatory, Design, and Analysis Aspects of Complex Stability Studies. *J Pharm Sci* Vol.84(11):1322-6
- 10 TD Lin 1994. Applicability of Matrix and Bracket Approach to Stability Study Design. In: Proceedings of the Biopharmaceutical Section of the American Statistical Association, Alexandria, VA, pp. 142-7
- 11 Shuirmann DJ 1987. A Comparison of the two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. *J Pharmacokinet Bioavailab* Vol. 15(6): 657-80
- 12 Ahn H Chen JJ and TYD Lin 1995. Classification of Batches in a Stability Study. In: Proceedings of the Biopharmaceutical Section of the American Statistical Association, Alexandria, VA: 256-61
- 13 Shao J and SC Chow 1994. Statistical Inference in Stability Analysis. *Biometrics* Vol.50(3):753-63
- 14 Chows SC and SG Wang 1991. On the Estimation of Variance Components in Stability Analysis. *Commun. Statist. Theory Methods* Vol.23:289-303
- 15 Ho Ch, JP Liu and SC Chow 1992. On the Analysis of Stability Data. In: Proceedings of the Biopharmaceutical Section of the American Statistical Association, Alexandria, VA, pp. 198-203
- 16 Nordbrock E 1992. Statistical Comparison of Stability Study Designs. *J Biopharm Statist* Vol.2:91-113
- 17 Hartmann V, Krummen K, Schnabel G and H Bethke. Techniques for Stability Testing and Shelf-life Predictions. *Pharm Ind* 1982, Vol. 44(1): 71-9
- 18 TE Norwood 1986. Statistical Analysis of Pharmaceutical Stability Data. *Drug Dev Ind Pharm* Vol.12(4):553-60
- 19 Kowalski K Beno M Bergstrom C and H Gaud 1987. The Application of Multiresponse Estimation to Drug Stability Studies. *Drug Dev Ind Pharm* Vol.13(15):2823-38