

ベンチマーク ドース・テクニカルガイダンス

リスクアセスメントフォーラム

米国環境保護庁

Washington, DC 20460

Disclaimer (for translation) :

This was translated from the original, English language document entitled “Benchmark Dose Technical Guidance”, which is a product of the U.S. Environmental Protection Agency’s Risk Assessment Forum.

This document was translated by an outside party, and U.S. EPA makes no claims with respect to the translation.”

免責条項（訳文）：

本文書は米国環境保護庁のリスクアセスメントフォーラムにより作成された“ベンチマーク ドース・テクニカルガイダンス”と題された英語による文書を原文として翻訳された。この翻訳文は外部機関によって翻訳されているので、米国環境保護庁はこの翻訳について一切関与しない。

また、本訳文は原文の理解を手助けするために国立医薬品食品衛生研究所 総合評価研究室における研究行為の一環として翻訳されましたが、正確な記載は原文を参照ください。総合評価研究室は本訳文の使用に関しての一切の責任は負いかねます。

免責条項(原文)

本文書は米国環境保護庁のピアレビュー政策および管理上のレビュー政策に従って査読され、発表および出版が許可された。商標名または販売製品の記述については、その使用を支持したり推奨したりするものではない。

目次	
略語および頭字語リスト	6
著者、テクニカルパネル、およびスタッフ	7
概要	8
1. 緒言	10
1.1. 目的	10
1.2. 背景	12
1.3. ベンチマークドースに関連する文献の簡略レビュー	16
1.3.1. 用量反応評価におけるベンチマークモデル化の早期使用	16
1.3.2. ベンチマークドースの性質	16
1.3.3. BMD 算出へのアプローチ	17
1.3.4. 本ベンチマークドーステクニカルガイダンスの歴史的発展	19
2. ベンチマークドースのガイダンス	19
2.1. データの評価	19
2.1.1. 試験デザイン	19
2.1.2. データ報告の側面	20
2.1.3. モデル化される試験の選択	21
2.1.4. モデル化されるエンドポイントの選択	21
2.1.5. BMD 計算の最小データセット	21
2.1.6. BMD 算出に対するデータの結合	24
2.1.7. 用量測定 of 調整	25
2.2. ベンチマーク反応 (BMR) レベルの選択	26
2.2.1. 非連続 (二分) データ	26
2.2.2. 連続データ	27
2.3. データのモデル化	30
2.3.1. はじめに	30
2.3.2. モデル選択の背景	31
2.3.3. モデルの選択	32
2.3.3.1. エンドポイントのタイプ	32
2.3.3.2. 実験デザイン	33
2.3.3.3. 制約と共変量	34
2.3.4. モデルの適合	35
2.3.5. いかにモデルがデータをうまく説明しているかの評価	37
2.3.6. モデル適合性の改善	38
2.3.7. モデルの比較	39
2.3.8. BMDL 取得のための信頼限界値の計算	40
2.3.9. POD 算出に使用するモデルの選択	42
2.4. 推奨事項の報告	42
2.5. ディシジョンツリー	43
付録 A. 例	46

A.1 非連続データのモデル化.....	46
A.1.1. 適合モデルの選択（セクション 2.3.3）	46
A.1.2. 適合度の評価（セクション 2.3.5）	47
A.1.3. モデルの比較（セクション 2.3.7）	48
A.1.4. POD の基礎として使用するモデルの選択（セクション 2.3.9）	49
A.2. 非連続データ：用量群の除外（セクション 2.3.6 を参照）	50
A.3. 連続データ：よく適合したモデルを得る	52
A.4. がんバイオアッセイデータ：線形外挿法の POD を得るためにモデル化する	58
A.5. 発生毒性データ	61
A.6. ヒトのデータ	64
付録 B. 用語集	65
付録 C. 特定のベンチマークドースモデル.....	74
付録 D. ベンチマークドース技術ガイダンス文書の寄稿者と校閲者	77
参考文献	78

図のリスト

図 1 : BMD および BMDL を示した二分データに適合するモデル例。	15
図 2: [A: BMD のモデル化の実現可能性を決定するデータ評価ステップのフローチャート] [B: 図 2A に対応するデータセット A、B、C の図解]	24
図 3: 標準正規分布から平均の 1 SD シフトで得られる集団テール確率の差、1 SD の BMR 基準値の理論的根拠の図解	29
図 4: BMD のディシジョンツリー	44

略語および頭字語リスト

AIC : 赤池の情報量基準

AUC : 曲線下面積

BMC : ベンチマーク濃度

BMCL : ベンチマーク濃度下限値

BMD : ベンチマーク ドース

BMCL : ベンチマーク ドース 下限値

BMDS : BMD ソフトウェア

BMCL : ベンチマーク ドース 上限値

BMR : ベンチマーク反応

CI : 信頼区間

ED : 有効用量

GEE : 一般化推定方程式

HED : ヒト相当用量

IRIS : 総合リスク情報システム

BMCL : 有効量下限値

LOAEL : 最小毒性量

MLE : 最尤推定値

NOAEL : 無毒性量

PBPK : 生理学的なファーマコキネティクス

POD : point of departure (出発点)

RfC : 参照濃度

RfD : 参照用量

SD : 標準偏差

SE : 標準誤差

UCL : 信頼上限

UED : 有効量上限値

UF : 不確実係数

U.S. EPA : 米国環境保護庁

著者、テクニカルパネル、およびスタッフ

著者

本文書は、米国環境保護庁（U.S. EPA）リスクアセスメントフォーラムの後援の下で、テクニカルパネルによって作成された。本文書は、2000年12月7-8日に開催された公開ピアレビューワークショップの会合で、外部のピアレビューパネルおよび一般のメンバーから得たコメントさらに、その会合以後受け取ったコメントや方法の適用により得た経験についての考察を反映している。さらに、本文書は、環境保護庁内部の化学者のレビュー（2006-2008）および外部の非公式なレビュー（2009-2011）からの追加のコメントの考察も反映している。

テクニカルパネル

David Gaylor (Former Employee), U.S. Food and Drug Administration, Rockville, MD 20857

Jeff Gift, Office of Research and Development, U.S. EPA, Research Triangle Park, NC 27711

Karen Hogan (Lead), Office of Research and Development, U.S. EPA, Washington, DC 20460

Jennifer Jinot, Office of Research and Development, U.S. EPA, Washington, DC 20460

Carole Kimmel (Former Employee and Former Lead), Office of Research and Development, U.S. EPA, Washington, DC 20460

R. Woodrow Setzer (Former Lead), Office of Research and Development, U.S. EPA, Research Triangle Park, NC 27711

リスクアセスメントフォーラムスタッフ

Michael Broder, Office of the Science Advisor, U.S. EPA, Washington, DC 20460

Diane Henshel, Office of the Science Advisor, U.S. EPA, Washington, DC 20460

概要

米国環境保護庁（U.S. EPA）は、環境物質への曝露の結果と考えられる多数の健康への影響に対するリスクアセスメントを実施している。これらのアセスメントは、曝露と健康関連の結果との用量反応関係の解析を含むことが多い。用量反応評価は基本的に以下の 2 つのステップからなる。（1）point of departure（POD：出発点）の定義および（2）ヒト曝露に関連した POD の外挿ベンチマークドース（BMD）アプローチは、ベンチマークドース（すなわちデータの観察可能な範囲の最も低い値に近い特異的反応レベルに相当する用量レベル）を得るための用量反応のモデル化を含み、人体への非がん性作用に対し長年使用されている無毒性量（NOAEL）または最小毒性量（LOAEL）プロセスよりも多くの情報を包含し伝えるものである。このアプローチは、がんエンドポイントに対する POD を決定する方法に類似している（U.S. EPA 2005a）。EPA が、発がんとは非発がんのリスクアセスメントに対するアプローチの整合化に向かうとき、人体への作用を発がんとは非発がんに二分することは、作用機序の考慮および、注目する作用が低用量で直線または非直線関係の傾向にあるかどうかの検討にとって代われる。そのため、本文書の目的は、線形または非線形の用量反応外挿の有無にかかわらず、異なるタイプの人体への影響データへの POD の決定をはじめとする、様々な用途に対して BMD が誘導するアプローチの一貫性のある適用に関して、EPA と外部団体にガイダンスを提供することである。BMD の他の用途には、相対的強度（例、化学物質について）または相対的感受性（例、異なるサブ集団について）の比較が含まれる。BMD モデル化は、生態リスクアセスメントのような他の領域にも適用可能でもあるが、本文書では人体への影響に対する用量反応のモデル化に焦点を当てていることに注意して欲しい。

本ガイドラインでは、BMD、ベンチマーク濃度（BMC）およびその信頼性限界の計算、データの要件用量反応解析、および BMD または BMC の使用に固有の推奨事項の報告などについて検討している。本文書では、用語について以下の慣習を採用している。BMD は、全般的にベンチマークドースアプローチの意味で使用され、モデル結果の特徴付けをする特別な場合は、BMD と BMC は中央推定値を示す。BMDL または BMCL は、各々 BMD または BMC 上の片側 95%信頼区間の下限に対応する。これは Crump（1995）により導入された用語および、<http://epa.gov/NCEA/bmds/>で無料で入手できる EPA の BMD ソフトウェア（BMDS）で用いられている用語と一致する。本文書は、名前は類似しているが、EPA の BMDS ソフトウェアに特異的ではなく、ここでの推奨事項は他のソフトウェアパッケージおよび他の用量反応モデルにも適用できる。

上述したように、BMD アプローチは、用量反応評価で長年にわたり使用されてきたが限界が認められる NOAEL/LOAEL アプローチの代わりとして開発された。しかしながら、すべてのデータが BMD モデル化に補正可能ではないため（例、完全なデータが入手不能またはデータセットを適切に説明できるモデルの欠損のため）、NOAEL/LOAEL アプローチも継続する必要がある。

用量反応のモデル化に適したモデルの選択を優先するには、特定例に関連する生物学的プロセスと一貫性のあるモデルを使用することである。そのようなモデルは、生物学的プロセス（例、細胞成長のダイナミクス、飽和酵素プロセス）または考慮する反応の変数（例、反応時間）の明確な表現を含むことができる。生物学系モデルに欠けると用量反応のモデル化は大幅に曲線適合が必要となる。本文書は簡単な用量反応モデルを対象としている。

BMD アプローチの適用および結果の解釈は技術的に難しいので、BMD モデル化は、統計的手順およびこのタイプの解析における落とし穴について熟知した専門家と共同作業を実施することが推奨される。本文書では、一貫性のある BMD 適用を支える以下のような多数の問題点を検討する。

- 1) BMD 計算の基礎となる試験とエンドポイントの決定
- 2) ベンチマーク反応値の選択
- 3) BMD の算出に使用するモデルの選択
- 4) モデルの適合、モデル適合の評価およびモデルの比較
- 5) BMD (すなわち、BMDL) の信頼性限界の計算
- 6) BMD および BMDL 計算の提示に対する推奨事項の報告

BMD 計算の基礎となる試験とエンドポイントの決定

有害性の特徴付けと用量反応評価に使用するエンドポイントの選択後、モデル化と BMD 解析に関連する試験を評価できる。用量に伴い単調に段階的な反応を示す試験の一部では、BMD 解析の補正が可能であり、BMD 計算用の最小のデータセットは、選択したエンドポイントにおいて生物学的または統計学的に有意な用量関連傾向を示すことになる。BMD のより良い推定値を与えるために、BMR 付近レベルの用量を 1 つ以上有する試験が望ましい。用量レベルのすべてが、コントロール値に比べ異なる (すなわち、NOAEL が存在しない) 試験は、一般的に容易に BMD 解析に使用できる。

本ガイダンスは、通常直面するタイプのデータ、大部分は二分データ (非連続データ) と連続データ、の定義を提供し、さらにどの情報が反応をモデル化するために必要であるかを検討する。例えば、二分反応はおそらく、1 つの作用の有無にかかわらず報告され、一方連続データは実際の測定または対比 (コントロールとの相対変化) として報告される。連続データで個々のデータが入手できない場合、被験者数、反応変数の平均、および反応バラツキの大きさ (例、標準偏差 (SD)、標準誤差 (SE)、または分散) は各グループに必要である。用量反応評価で使用される傾向がある異なる試験から選択されたエンドポイントのすべては、特に異なる不確実係数 (UF) が異なる試験とエンドポイントに使用される場合は、モデル化されるべきである。リスク評価者は (一部のエンドポイントがモデル化できない場合)、結果として得た BMD および NOAEL/LOAEL を科学的判断、リスクアセスメントの原則、およびモデル化プロセスの結果を用いて、POD としての使用を評価する。本ガイドラインは、BMD のモデル化の技術的視点に限定している。

ベンチマーク反応 (BMR) 値の選択

BMD の計算は BMR の選択により直接決定される。BMR の選択では、データベースの統計学的かつ生物学的特徴、および得られた BMD/BMDL を使用する適用に対し判断を下す。用途が異なれば、異なる BMD 値が必要となる可能性がある。EPA は、反応レベルまたは BMD の選択を判断し、特定の適用 (例、参照用量または相対強度計数) に対する BMD のモデル化に使用するガイダンスは現在のところ存在せず、そのようなガイダンスは本文書の範囲を超えている。選択はケースバイケースでなされ、各 BMR 選択には透明性をもって判断がなされる必要がある。本ガイダンスは、非連続データと連続データにおける BMR の選択における一般的なアプローチを検討している。

非連続データでは、超過リスク 10%は、(化学物質間およびエンドポイント間の比較の基礎として用いられる) 標準的な報告に対する BMR であり、ハザードランク付けにすることが多い。その理由は 10%の反応は殆どのがんバイオアッセイおよび一部の非がんバイオアッセイにおける限界感度に近いからである。これは BMR の初期設定ではないことに注意すること。POD の決定では、低い (あるいは時々高い) BMR が統計学的/生物学的配慮に基づいて使用されることが多い。それでもなお、報告の目的では、超過リスク 10%に相当する BMD を常に提示されることが推奨される。

連続データの好ましいアプローチは、(専門家の判断または信頼性のあるガイダンス文書によって決定されるような) 作用が生物学的に有意になると考えられるエンドポイントの変更レベルに基づいて BMR を定義するこ

とである。個々のデータが使用できる場合、または（例えば、コントロール分布のパーセンタイルに基づいて）個々のレベルで有害性と判断される得る場合、そのデータは、カットオフ値に基づき二分され、BMR は上述の非連続的データのように設定される。代わって、どの反応レベルを有害と考えるかについて他に考えがない場合、コントロールの平均における 1 つの標準偏差（SD）と同程度の平均値の変化が使用できる。つまり、統計的/生物学的考慮により補償されるなら、コントロール SD の低い増分または高い増分が使用されることがあるかもしれない。コントロール SD は、以前のコントロールデータを含めて計算できるが、コントロール平均は考慮する処理群と同時に得たデータから算出すべきである。BMR を定義するどの方法に連続データセットを使用するかに関わらず、コントロール平均反応からのコントロール SD に対応する BMD は報告目的用に提示されることが推奨される。

BMD の算出に使用するモデルの選択

BMD 算出の数学的モデル化の最終目的は、データセットを説明する用量反応データに、特に観察可能な用量反応範囲の下端に、モデルを適合させることである。生物学系モデルに欠けると、用量反応のモデル化は大幅に曲線適合が必要となる。実際、これにはまずデータと実験デザインの特徴に基づき、さらに検討するためにモデルのファミリーを 1 つあるいは複数選択し、次いで確立された数少ない方法の 1 つによってそのモデルを適合させる。ガイダンス文書は、異なるタイプのデータへのモデル選択に関する情報を提供している。さらに、BMD と BMDL を得るにはどのモデルを使用するかを決定するために、モデル適合、適合度の測定、およびモデルの比較が検討される。ガイダンスでは、適合度の臨界値の算出に $\alpha=0.1$ を使用すること、およびモデル適合のグラフィック表示も同様に試験することを推奨している。BMD 算出用のモデルの比較およびモデルの選択には、赤池の情報量基準（AIC）が推奨される。

BMD の信頼性限界（すなわち、BMDL）の計算

本ガイダンスでは、信頼限界が得られる方法は、通常、データタイプと BMD がモデルから推定される方法に関係していることを認識しながら、BMD の信頼限界の算出を検討している。本文書では、特定のデータタイプ（例、非連続、クラスター化、連続）に固有の信頼限界を算出する詳細なアプローチを与えている。

BMD/BMDL 算出からの報告推奨事項

本ガイダンスには、BMD と BMDL の多数の報告推奨事項がリストされている。これらは、モデル化の試験およびエンドポイント、さらにこれらのエンドポイントを特徴付けする BMD と BMDL の選択を文書化するために重要である。

要約すると、本ガイダンスは、モデル化に適する試験とエンドポイントタイプの評価、BMR レベルの選択、モデルの適合と BMD の算出、モデル適合の判断、さらに BMDL の算出に使用される、ステップごとのプロセスを提供している。最後に、本文書には（EPA の BMD パッケージによる）BMD と BMDL 算出の例が載っている。

1 緒言

1.1. 目的

本文書の目的は、ベンチマークドースアプローチの適用について EPA と外部団体にガイダンスを提供することである。そのアプローチには、ベンチマークドース、すなわち特定反応レベル、または観察可能なデータ範囲の最低値に近いベンチマーク反応に対応する用量レベルを得るための用量反応のモデル化を含んでいる。そのた

め、これらのベンチマークドースは、人体への影響データの線形または非線形外挿に対する可能な **points of departure** (POD：出発点) として、また試験/化学物質/エンドポイントに対する用量と反応結果の比較の基礎として役立つ。本ガイダンスは、ベンチマークドースとベンチマーク濃度 (BMD と BMC)、およびその信頼限界の算出、データの要件、用量反応解析、および推奨事項の報告について検討している。本文書は、現在の知識と理解、さらに本アプローチを使って得た経験に基づくガイダンスを提供している。本文書は、その中で示されたアプローチの代替え、または追加など、いずれも新しいアプローチが得られるようになった時点で更新されるように作成されており、定量的リスクアセスメントを改善するリサーチを妨げるものではない。事実、EPA は、科学的な理解の向上、および用量反応のモデル化のよりメカニズム的なアプローチの開発を強く奨励している。

BMD 算出方法は特別なソフトウェアが必要なので、本文書のもう一つの目的は、好ましい算出アルゴリズムの情報を提供し、ユーザーがそのソフトウェアの情報を得て選択できるようにすることである。本文書は特定のソフトウェアパッケージの使用を支持しないが、EPA の BMDS パッケージのように方法の記載が十分なソフトウェアの使用が推奨される^{注1} (本ガイダンスでは、説明のため、EPA の BMDS パッケージを使った図式を提示している)。本ガイダンスは BMD と用量反応解析の算出に関する試験デザインの情報を与えることも予想されるが、これについては明確に言及されていない。

注1 BMDS の詳細については <http://epa.gov/NCEA/bmds/>を参照する。

本文書はガイダンスのみを目的としている。本文書は、米国連邦行政手続法 (Administrative Procedure Act) あるいはその他のいかなる法律下で実質的な「規則」を確立するわけではなく、EPA またはいずれの規制当局と結びつける効果は有していない。

本文書は BMD のモデル化の入門書ではない。BMD のモデル化には高い技術が必要であり、本ガイダンスは、人体への定量的リスクアセスメントの十分な経験を積んだ読者を対象とした技術文書である。解析を助けるソフトウェアを得ると、モデル化は見かけより易しくなるように見えるが、BMD アプローチの適用と結果の解釈は簡単でないことが多い。BMD のモデル化は、統計的手順およびこのタイプの解析における落とし穴について熟知した専門家と共同作業を実施することが推奨される。

本文書は使用可能な用量反応モデルの範囲または、それに関連する利点についても考慮していない。このガイダンスがなくとも、適切な方法を使用することにかわりはなく、EPA は使用可能なベストな科学的方法を使用することを強く奨励している。本文書の焦点は、用量反応のモデル化の基本原則および一貫性のある使用に当てている。必要に応じて、多変数解析、カテゴリカル回帰、時間-反応解析、分布解析、ブートストラップ法、モデル平均化、およびベイジアン法など、より特殊なトピック (しかしこれだけに限らない) が今後の本ガイダンスへの補遺または他のガイダンスで検討されることになる。

同様に、本文書は毒性学またはリスクアセスメントの入門書ではない。つまり、ここに記載されている手順は、毒性学者およびリスクアセスメントにおけるハザード特徴決定の問題点を指摘する人々など、専門家の判断に代わるものではない。試験の品質および観察された作用の毒性学的有意性などの問題点に対する専門家の評価や判断は、BMD 解析の使用とは別に必要であり、本文書の範囲を超えている。特に、本文書は何が生物学的有意性をもたらすのかについては取り上げていない。つまりこういった決断は、特定の適用状況において、また本決定を伝える他の入手可能な政府機関のガイドラインと併せてなされなければならない。そのため、どの程度の人体

への変化を有害と定義するのか、あるいは RfC、RfD、またはがん効力の算出に対するガイドラインの提供などはより一般的なリスクアセスメントの問題点でもあり、本文書の範囲を超えている。さらに本文書は、特定のエンドポイントや適用に対するベンチマーク反応（BMR）の選択、および他のリスクアセスメントの科学的政策問題点についてのガイダンスをも目的としていない。

最終的に、本文書の焦点は、実験動物試験の毒性学的データのモデル化に当てている。ヒトデータのモデル化の機会はさらに限定されており、ヒト試験は動物試験よりも標準化しにくく、さらにヒトデータのモデル化は、変数の補正など別の考慮事項が入ってくることが多い。そのため、通常ヒトデータのモデル化はより特異的な方法で実施される。ヒトデータに対する BMD のモデル化の例を示す、数件の引用文献について付録 A.6 を参照すること。

1.2. 背景

U.S. EPA は、環境物質への曝露の結果と考えられる多数の健康への影響についてリスクアセスメントを実施している。リスクアセスメントのプロセスは、米国学術研究会議（National Research Council : NRC）の概念（NRC 1983）に基づき、ハザード同定、用量反応評価、曝露評価、およびリスク特徴決定の数ステップを有する。有害性の特徴付けは使用可能なすべてのデータの完全な評価であり、人体への潜在的有害性を同定し特徴付ける。用量反応評価は、化学物質曝露と健康関連結果との関係の解析を含み、がんと他の毒性作用間のメカニズムの基礎には異なる受け止め方があるため、その評価は以前から人体への発がん作用と非発がん作用を非常に異なっており扱ってきた。しかし、毒性作用の基礎となる生物学の理解が進展するにつれ、発がん作用と非発がん作用の明確な差は減少した。本セクションでは、EPA の発がん作用と非発がん作用の用量反応評価アプローチおよび、より広範に適用可能な定量的方法の基礎についての概要を紹介する。

発がん作用と非発がん作用間のリスクキャラクター化の主な差異は、単独変異によるがん誘発の理論的な可能性、およびバックグラウンド反応への追加の可能性など多数の要因により、がんの超過リスクは低用量において線形であるとする予測であった（U.S. EPA 1986 ; Crump ら.1976）。非発がん作用は一方、十分な曝露レベル（閾値）後にのみ発生することが一般的に仮定された。発がん作用の用量反応評価の実際は通常、統計学的モデルを腫瘍発症データに適合させることであった。つまり、リスクの低用量への外挿アプローチは、時間とともに変化した。これについては他の文書に要約されている（U.S. EPA 1986, 2005a）。以前から、データのバラツキによる発がんリスク推定の不確実性は、通常 10^{-6} ~ 10^{-5} の非常に低リスクレベルで、曝露とリスクの関係のスロープに対する 95%上限を使用することにより扱われた。この不確実性は、現在、低用量外挿の中央推定用量に対する 95%信頼区間により扱われている（U.S. EPA 2005a）。

これに比べて、発がん以外の人体への影響についての用量反応解析の標準的な実施では、以前から適切な試験による最小毒性濃度（LOAEL）および無毒性濃度（NOAEL）に基づいて基準値を設定していた。LOAEL は、使用した化学物質で有害作用が検出された最低用量であり、NOAEL は、有害作用が検出されなかった最大用量である。NOAEL（あるいは、NOAEL がない場合は LOAEL）は使用可能なデータの限界と不確実性を考慮した「不確実係数：UF」を適用する場合は POD として役立ち、ヒトへの有害作用の大きなリスクがないと思われる曝露、つまり参照用量（RfD）または参照濃度（RfC ; U.S. EPA 2002a）に到達する。発がんの用量反応モデルとは異なり、観察される反応のバラツキは、（有意差検定を超えた）NOAEL/LOAEL アプローチでは指摘されない。

NOAELは時々このようなNOAELと真の閾値（すなわち、真の無作用レベル）の一致を推定するため、試験中の用量反応関係を説明する重要なポイントとして考慮される。しかし、一般的に作用の統計学的有意性に欠けて定義されたNOAELは、実際、限りある試験が固有の検出限界を有しているという事実による結果である^{注2}。

- ・NOAEL/LOAELは、試験で使用される用量の1つに限定されるので、用量の選択に高く依存している。
- ・NOAEL/LOAELは、サンプルサイズに強く依存する。投与群の反応を対照群の反応と区別するためのバイオアッセイの性能は、サンプルサイズが小さくなるほど低下する^{注3}。そのため化合物に対するNOAEL（およびNOAELに基づく場合はPOD）は、1用量群の動物数が少ない試験ほど高くなる傾向にある。
- ・より一般的には、NOAEL/LOAELアプローチは、用量選択、投与間隔、およびサンプルサイズなど試験デザインの特徴のため、実験結果のバラツキと不確実性を説明しない。
- ・NOAEL/LOAELは試験／化学物質／エンドポイントの比較で一貫性のある反応レベルに対応せず、またNOAELあるいはLOAELで観察された反応レベルはRfD/RfCの算定では考慮されない。
- ・用量反応曲線の形（例、BMDで傾きがどれだけ大きい小さいかは、推定閾値にPODがどれだけ近いかわかる程度を示す）のような実験から得られる他の用量反応情報は、考慮されない。
- ・NOAELが試験内に存在しない場合、LOAELをNOAELの算定に使用できない。その代わり、UFを最大10までLOAELに適用して、この限界を説明する。
- ・NOAELは通常閾値（無作用レベル）として解釈されてきたが、シミュレーション試験（例、LeisenringおよびRyan 1992、1群あたりの反復回数10、20、50回を含む試験デザイン）や発生毒性バイオアッセイデータの再解析（Gaylor 1992、Allenら1994a、1群約20匹の同腹子を含む試験）は、NOAELの基準に適合する用量で対照群を超える反応比は、試験デザインの範囲に対して、平均5～20%であって0%ではないことを実証した（詳細はセクション1.3.2を参照。）

1. ^{注2}分析化学による「検出限界」という記述は、毒性学試験の最小検出反応レベルを特徴付けるために時々使われてきた。統計学的パワーが含まれるか否か、もし含まれるならどの程度のパワーレベルが意図されるのかなど、この概念を一貫して適用する標準化された基準はない。一部の試験が他の試験よりよりパワフルであるという事実は、重要ではなく、試験の感度として定性的な意味を有している。

^{注3}二分データとして、例えば、1群6匹の動物を使用した試験では、観察された有害反応比率0%に対する95%信頼上限（UCL）は49%となる。すなわち、6匹の動物で観察されない反応に基づいて選択されたNOAELでの真の作用は0%よりかなり大きい可能性もある。10、20、および50匹群について、観察された有害反応比率が0%である95%UCLは、各々31%、17%、7%となり、適切なサンプルサイズの重要性を浮き彫りにしている。

NOAEL/LOAELアプローチの限界の一部を指摘する試みとして、Crump（1984）はBMDアプローチを代替方法として提唱した（詳細はセクション1.3を参照）。BMDのモデル化では、（バックグラウンド反応レベルに比べて）反応の大きさは通常高い用量では減少しないということ以外に、一般的に観察された用量反応関係の生物学的基礎について特定の仮定はしない。特に、推定無作用レベルとBMD間で固有な関係はない。十分なデータがある場合、BMDアプローチは、基準値（例、RfDまたはRfC）の算出用または低用量の線形外挿用の可能なPODとして、および／または、試験／化学物質／エンドポイントの一貫した比較に対する特定の反応レベルに対応する用量レベルとして役立てるために、BMDの算定に使用できる。

BMDアプローチは、腫瘍データおよび発がんプロセスで前駆体の重要なイベントと考えられる他の反応のモデル化に関して、2005年にEPAが発行した発がん性物質のリスクアセスメントガイドライン（U.S. EPA 2005a）

の勧告を実践するために使用できる。このガイドラインは、用量反応関係を測定する際の化学物質作用機序の理解を促進している。さらに、その用量反応の外挿手順は、化学物質の発がん性作用機序に関する有害性評価の結論に従っている。ガイドラインの用量反応評価には以下の2段階のプロセスがある。(1) 反応データが経験的観察範囲でモデル化される。つまり観察範囲のモデル化が、生物学的に基づく曲線適合モデルを使ってなされる。次いで、(2) 観察範囲未満への外挿が、十分なデータがあればモデル化によって、または初期設定手順（線形、非線形、または両方）によってなされる。初期設定外挿手順では、観察可能な範囲の最低値に近い POD がモデル化によって推定される。ガイドラインでは、POD は一般的に、データがモデル化を支持できる最低用量レベルの 95%信頼下限である。線形の初期設定値は、POD からのバックグラウンド反応レベルへの直線外挿であり、1 ユニット用量当たりのリスク推定値（上限）が示され、一方、非線形アプローチは、同定した POD に UF を適用し、低用量のリスク推定値ではなくがんの基準値（RfD または RfC に類似）を与える。

非発がん作用の基準値を算定する場合、実験動物からヒトへの外挿によって発生する不確実性を考慮し、入手可能なデータの他の限界値と同様にヒト内のバラツキを考慮するために、POD は下向きに補正される。Review of Reference Dose and Reference Concentration Processes（参照用量と参照濃度プロセスのレビュー）（U.S. EPA 2002a）では、基準値の誘導に関してより完全に検討している。NOAEL/LOAEL と BMD のアプローチの主な相違は POD をどのように決定するかであることに注意して欲しい。本文書では、EPA の説明（2002a）のように、非発がん作用に対する POD として、BMD の 95%下限（すなわち、BMDL）を適用することを推奨している。下限値を使用すると、与えられた試験に内在する実験的バラツキを考慮し、選択した BMR を超えないこと（BMR の検討についてはセクション 2.2 を参照）（実験内容の 95%信頼区間で）を確保することになる。95%限界値の使用は、発がんリスク推定値に從來から使用されているものと一致し、BMDL を POD として全般的に使用することは、EPA の発がんガイドラインでも記載されている（U.S. EPA 2005a）。これに反して、化学物質／エンドポイント／試験間で比較する場合は、中央推定値の使用が推奨される。EPA の発がんガイドライン（U.S. EPA 2005a）は、不確実性の指標を伝える役に立つように、関連する中央用量推定値および上限用量推定値の報告を推奨していることに注意して欲しい。

NOAEL/LOAEL アプローチの限界について本文書の始めに議論したように、BMD アプローチは NOAEL/LOAEL アプローチより好まれる。例えば、BMD（または BMDL）は、試験の全用量が有意に有害反応に関連している場合でさえ、推定可能である（すなわち、NOAEL が存在しない場合）。しかし、信頼性のある BMD が推定できず、NOAEL/LOAEL アプローチが正当化される可能性のある場合も時にはある。特に、利用可能なデータがモデル化に従わないことがある。例えば、曝露群すべてが最大反応を呈する場合などである。そのような場合、観察されたデータは、反応レベル全域にわたり非常にわずかな情報しか与えず、BMD モデルはその範囲内で信頼性のある推定値を提供できない（しかし、そのような場合、LOAEL からの情報も限定される）。用量反応のモデル化に従わないデータセットの別のサンプルに関する検討はセクション 2.1.5 も参照すること。そのような場合、NOAEL/LOAEL アプローチは、その限界およびデータセットの限界を認めながら、使用されることもなくはない。

表記法：文献では、BMD および BMDL の用語を様々な方法で使用してきた（Crump 1984, 1995）。中央推定値および信頼下限値、より一般的に定義された BMD を参照するために用量反応評価が頻繁に必要となる。この文書の以下で、ベンチマークドースを誘導するプロセスの技術的な詳細について説明する場合、BMD または BMC は、BMR を得ることが期待される用量または濃度の中央推定値を意味する。BMDL または BMCL は、中央推定値に対する片側信頼区間の下限を意味する。BMD はまた、全モデル化プロセスを参照するためにも使用され

る。低用量の外挿または RfD/RfC 設定に使用する POD は、BMDL または BMCL となる。本文書でのさらなる議論を簡単にするため、特に記載のない限り、一般的に経口または吸入値を意味するために BMD と BMDL を使用する。最後に、本文書では使用されていないが、BMD と BMDL の基本となる BMR 値を表記する下付字(例、5%超過リスクに対し、 BMD_{05} 、連続データの平均において、5%または1標準偏差 (SD) の変化には各々 BMD_{c05} または BMD_{1SD}) は、異なる BMR に基づく BMD/BMDL の定義および BMD/BMDL の識別に役立つと思われる。BMR を表示する明確な下付文字がない場合、各 BMD と BMDL に対応する BMR を明確に述べなければならない。

図解例：BMD アプローチを使って、実験データはデモル化され、BMD は観察可能な範囲で推定される。図 1 は、二分データに適合する BMD モデルの図であり、10%の超過リスクに対する BMD と BMDL を示している。上の曲線は、BMD の片側 95%信頼下限値に相当する。このデータセットの NOAEL は 50 ユニットであり、LOAEL は 100 ユニットであると思われる。NOAEL および LOAEL とは異なり、BMD および BMDL は実験用量の1つであることに留まらないため、BMDL は LOAEL または NOAEL のいずれよりも、特異的な BMR に基づきより一貫性のあるうまく定義された POD として使用されうる。与えられたモデルが真であると仮定すると、BMDL は試験デザインの特徴に起因する BMD 推定値の UF を明かにする。BMD アプローチは通常、試験の反応データすべてを使用し、用量反応曲線の形は BMD および BMDL の推定に不可欠である。

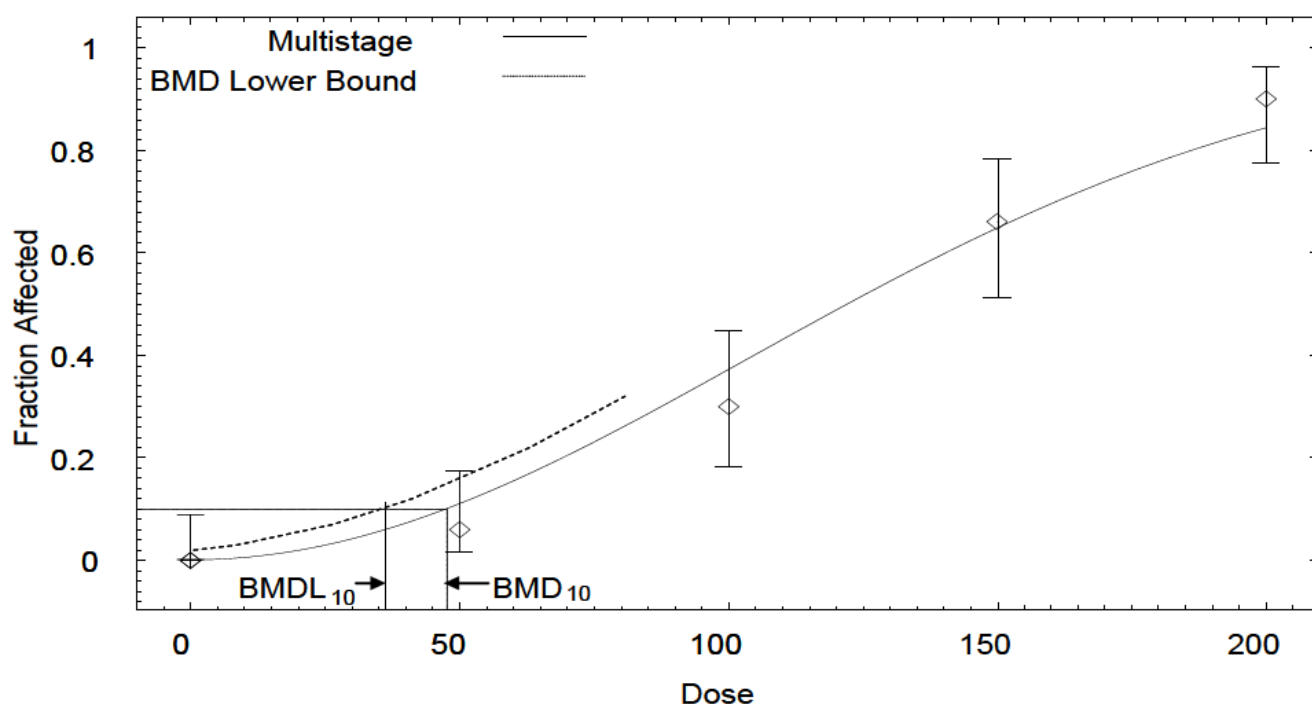


図 1：BMD および BMDL を示した二分データに適合するモデル例。

各群で影響を受けた動物の割合を菱形で示している。誤差バーは影響を受けた割合に対する 95%信頼区間を表している。この例で BMR の超過リスクは 10%である（または反応割合 0.1）。適合したモデルは実線で示され、この曲線で 10%超過リスクに対応する BMD は BMD_{10} と表記されている。 $BMDL_{10}$ と表記されている BMD_{10} の下限値は、点線の曲線から適合モデル曲線の左側に位置し、BMR 範囲の用量に対する推定下限値を示す。

BMD 算出手順は非常に一般的であるので用量反応評価で BMD アプローチが一貫性のある方法で利用できるように、本文書では以下の多数の問題点を少し詳しく検討している。

- 1) 試験と基本的な BMD 算出のためのエンドポイントの選択、およびデータセットの最低要件などのデータ評価 (セクション 2.1) ;
- 2) BMR 値の選択 (セクション 2.2) ;
- 3) BMD の算出に使用するモデルの選択 (セクション 2.3.3) ;
- 4) モデルの適合、モデル適合の評価およびモデルの比較 (セクション 2.3.4~2.3.7) ;
- 5) BMD の信頼限界の算出 (例、BMDL、セクション 2.3.8) ; および
- 6) BMD 算出のどの情報を報告するかの確認 (セクション 2.4)

本文書で検討している、非連続および連続データへの BMD の適用を図解した例を付録 A に示す。用語集は付録 B にあり、選択した BMD モデルに対する式は付録 C にある。

1.3. ベンチマークドースに関連する文献の簡略レビュー

これらの方法の適用について、最近、いくつかのレビューが Filipsson ら (2003)、Filipsson および Victorin (2003)、Gaylor ら (1998)、Parham および Portier (2005)、Sand (2005)、および Sand ら (2002、2008) によって提供されている。

1.3.1. 用量反応評価におけるベンチマークモデル化の早期使用

用量反応評価へのベンチマークドース様アプローチは新しくはない。Mantel および Bryan (1961) は、低用量発がんリスク評価の手順を提示した。この手順では、最小試験用量での過剰^{注4}腫瘍発生率の信頼上限、または 1%過剰腫瘍発生率を起こすと推定される用量、つまり基本的に BMD での過剰腫瘍発生率の信頼上限を計算した。プロビット対数用量モデルを仮定し、用量の 10 の倍数の減少につき、1 プロビットの低用量スロープを使用し、低用量での過剰がん発生率の推定値を導いた (「控えめな」値が目的)。Gaylor および Kodell (1980)、van Ryzin (1980)、および Farmer ら (1982) は、最小試験用量または過剰発生率 1%に相当する用量つまり再度 BMD で、有害作用のバックグラウンドを超えた過剰発生率の信頼上限から、過剰リスクゼロへの低用量線形外挿を提案し、サブ線形の用量反応曲線に低用量リスクの上限を与えた。Gaylor (1983) と Krewski ら (1984) は低用量リスクを制御する直線外挿と安全因子を比較している。Crump (1984) は、「ベンチマークドース」という用語を最初に導入した。

^{注4} 本文書で使用している「過剰発生率」と「過剰リスク」という用語は、コントロールまたはバックグラウンド反応を超えて増加した発生率またはリスクを広く意味する。これらの増加は、追加リスク、または過剰リスク (付録 B を参照) または相対リスクなど、複数の方法で表現されることがある。本文書では、リスクのこれらまたは他の定義を使用する多数の適用を支持し、一般的な記述では「過剰」を使用している。特定なリスク測定を意図する場合は常に、その旨が明確にされる。

1.3.2. ベンチマークドースの性質

多数の研究が、ベンチマークドースと NOAEL の比較を試みた。これらの多くは生殖発生毒性データを扱い、過剰確率に関して 10%以上の作用 (非連続データ) または従来の NOAEL における対照群平均からの変化 (連

続データ)を示した(例、Alexeeff ら 1993; Catalano ら 1993; Chen ら 1991; Krewski および Zhu 1994, 1995; Auton 1994; Crump 1995; Fowles ら 1999; Leisenring および Ryan 1992; Gaylor 1992)。Faustman ら (1994)、Allen ら (1994a,b)、および Kavlock ら (1995) による一連の論文では、BMD アプローチを発生毒性研究 (約 20 同腹子/群) の大規模データベースに適用した。簡単に言うと、これらの試験結果は、データを、影響を受けた胎児数/同腹子の割合 (入れ子型二分データ) で表す場合、NOAEL の 10%過剰反応確率は平均で、BMDL の 0.7 倍、さらに 5%過剰反応確率は平均で、BMDL とほぼ等しかった。データを非連続エンドポイント数として表す (すなわち、胚吸収または奇形を示す 1 用量群当たりの同腹子数) 場合、対照群値を超える反応の 10%確率での NOAEL は、BMDL よりも約 2~3 倍高く、5%過剰反応確率は BMDL よりも約 4~6 倍高かった。影響を受けた胎児/同腹子の割合としてデータを表すことは、発生毒性データの解析にはより正確な方法である。しかし、連続データの解析結果は、他の連続データによる BMD アプローチの使用に適合することもあり、またこのような場合の NOAEL は、サンプルサイズおよびバックグラウンド比に応じて、真の 10%過剰反応レベル以上である可能性を示唆している。

発生毒性試験での胎児体重の減少は、評価される様々なエンドポイント間で最小の NOAEL を示すことが多いので、これらの連続データへの BMD アプローチの適用も評価された (Kavlock ら 1995)。カットオフ値のバラツキが、対照値未満の体重減少の有害レベルを定義するために詳しく検討された。一部の例では、連続パワーモデルによって解析され、他の場合は、非連続データへ変換された。NOAEL との比較では、数個のカットオフ値が NOAEL と同様な BMDL 値を与えることが示された。これらの解析は、連続データの様々なエンドポイントから BMD が導き出される方法を示している。

Fowles ら (1999) は、急性吸入致死データを検討し、1%、5%および 10%過剰反応発生率に対応する NOAEL と BMDL を比較した。サンプルサイズは、1 群平均 10~20 匹であった。Allen ら (1994a, b) の試験結果の「非連続」部分と同様に、10%過剰発生率に基づく BMDL は NOAEL にほぼ相当した。しかし、これらの致死データに対する反应用量関係はかなり急激であったため、5%および 1%過剰発生率に対する BMDL は、10%過剰発生率の BMDL と非常に接近していた。結果として、1%過剰発生率に対する BMDL は、ログプロビットまたはワイブルモデルを使用することで、NOAEL よりも平均でわずかに約 1.6 倍または 3.6 倍小さかった。

NOAEL とのこれらの比較に加え、Kavlock ら (1996) によるシミュレーション試験では、発生毒性の 2 つのエンドポイント (奇形の発生数および胎児の体重減少) に対し、様々な見方の試験デザイン (投与群数、投与間隔、1 投与群のサンプルサイズ) に関連して BMDL を検討した。評価したデザインについて、ベストな結果 (つまり、信頼区間が狭い結果) は、2 つの用量レベルがバックグラウンドレベルを超える反応比率を有する時に得られ、そのうち 1 つは BMD に近かった。この試験では、サンプルサイズが 10~20 同腹子/用量群を超えると事実上利点はなかった。バックグラウンドレベルを超える反応率を有する 2 つの用量群とも、BMD に近い場合、満足する結果が得られたが、BMDL は低くなる傾向があった。バックグラウンドを超える反応比率を有する用量レベルが 1 つだけあり、BMD に近い場合、最尤推定値と BMDL について妥当な結果が得られたが、この場合、大きい用量群サイズで利点があった。反応比率が増加した群がたった一群であり、反応比率が BMR よりかなり高かった場合、最悪の結果が得られた。

1.3.3. BMD 算出へのアプローチ

がん以外の多数の人体への影響は、お互いに完全には独立していない複数のエンドポイントによって特徴付けられる。Lefkopoulou ら (1989)、Chene ら (1991)、Ryan (1992a, b)、Catalano ら (1993)、Zhu ら (1994)

Krewski および Zhu (1995)、および Fung ら (1998) は、発生毒性データを使ってこの問題点を検討し、殆どの場合、多項モデル化アプローチから誘導した BMDL は、いずれの個々のエンドポイントよりも低いことを示した。このアプローチは、他の人体への影響データに適用されなかったが、複数の関連結果が特定の人体への影響について考慮される場合は、留意する必要がある。

リスクアセスメントに対する連続的エンドポイントの用量反応のモデル化は、リスクを特徴付けに適した確率スケールが存在しないため、さらに難しい。課題は、連続スケール再解釈の際の影響にあり、非連続エンドポイントで実施したように、結果はリスクの観点から考えられ得る。1つのアプローチは、そのような連続エンドポイントを明確に二分し、続いてその二分したエンドポイントを、その他の任意の非連続エンドポイントのようにモデル化することである。別の論文で、Crump (1995) および Kodell ら (1995) が、Gaylor および Slikker (1990) によって最初に提案された方法に基づき、連続データに対する BMD を誘導するアプローチを詳しく述べた。このアプローチは、「ハイブリッド」法と呼ばれることが多く、連続データの分布を活用し、有害または少なくとも正常でないと考えられるレベルを超えるか、または超えない個別の発生率を推定し、さらに対照群より高い特定用量での反応の確率を与える。結果は、非連続データの解析から誘導したものと同一の用語で表現される。つまり、このアプローチは、連続データのモデル化に対する全パワーを保持し、一方で連続および非連続データに由来する BMD と BMDL の直接比較を可能にする結果を得ながら、暗にデータを二分する。Gaylor (1996) は、連続的エンドポイントに対して算出した BMD を、最初にデータを明確に二分した後算出したものと直接比較した。そして中程度のサンプルサイズであっても、データを明確に二分することでかなりの精度が失われることを見出した。West および Kodell (1999) は、連続データでそのような暗黙的方法を、明確に二分されたエンドポイントをモデル化した結果と比較した。West および Kodell は、1 群 10~20 匹のサンプルサイズについて、暗黙的アプローチは、明確に二分したデータのモデル化アプローチを行うよりも十分に良好な結果を与えることを見出した。そのため、可能な場合、おそらく Gaylor および Slikker (1990)、Crump (1995)、または Kodell ら (1995) によって記述されたハイブリッド法により、連続データのモデルから得た連続データに対する BMD と BMDL を誘導することは一般的により良好である。Crump (2002) は、連続データの BMD 計算における現在未解決の問題点について議論している。

BMD のモデル化の殆どのアプローチは、単一試験から 1 つまたは複数の反応をモデル化することに焦点を当てていた。カテゴリカル回帰のモデル化 (Dourson ら 1985 ; Hertzberg 1989 ; Hertzberg および Miller 1985 ; Guth ら 1997 ; Simpson ら 1996a, b) は、試験にまたがる複数エンドポイントの結果をより大規模なデータベースに基づく化合物毒性の全体的評価に使用できるようにする 1 つの方法である。この方法は今まで BMD 算出に広く使用されていなかったが、豊富なデータベースからより定量的に厳密に情報を組み合わせる方法として有望であることを示している。

BMD 算出へのベイジアン法は、確率分布による BMD 推定における不確実性 (ベイジアン用語では事後分布) を表し、より一般的に使用される頻度論的アプローチで使用される信頼限界と対比される (Hasselblad および Jarabek 1995)。ベイジアン法の広範な適用はまだ見出されていないが、有用な可能性のある特徴がいくつかある。ベイジアン法は異なるデータセットの結果を組み合わせやすくし、より確実な推定値およびその推定値の不確実性評価を与え、試験間のバラツキを考慮する。このタイプのアプローチは、単一試験に内在する不確実性のみを定量化するより広く使用されている方法よりも、改善点を与える可能性がある。

Gaylor ら (1998) は、BMD 算出用の統計的方法をレビューし、Murrell ら (1998) は、POD として BMD の信頼限界を使用する帰結のいくつかを検討し、連続的エンドポイントに BMR レベルを設定する方法を示唆した。

1.3.4. 本ベンチマークドーステクニカルガイダンスの歴史的発展

数件のワークショップおよびシンポジウムが開催され、BMD 方法論の適用に関して議論された (Kimmel ら 1989 ; California EPA 1994 ; Beck ら 1993 ; Barnes ら 1995 ; U.S. EPA 1996b)。全体的にみると、1995 年に EPA が共同スポンサーしたワークショップの参加者 (Barnes ら 1995) は、すべての非連続的非発がんエンドポイント、特に多数の研究がなされた発生毒性に対して、BMD 法を適用することを支持した。BMD 法のワークショップが開催された時点では、BMD 法の連続データへの適用に関して利用できる情報が少なく、多くの研究が促された。

これらのワークショップおよび議論は、2000 年に公表され科学的ピアレビューに公開された本文書の初期原稿の発展に情報を与えた。本最終文書に明示されたガイダンスおよび推奨選択肢の大部分は、2000 年の原稿、外部ピアレビュー委員会のコメント、および EPA リスクアセスメントの方法論の適用から得た経験に基づいている。

2 ベンチマークドースのガイダンス

このセクションでは完全な BMD 解析を実施するために提唱されたアプローチについて説明している。解析を進めるに当たっての理論的根拠および推奨される初期設定など、決定プロセスの形式で体系化されている。本ガイダンスでは、決定基準を通じて BMD 解析における制約をいくつか示唆し、可能なアプローチが 2 つ以上存在する場合、初期設定を提案している。

2.1. データの評価

有害性を特徴付けるプロセスの最初のステップは、特定の化合物または曝露状態に関連する有害物質を特定し特徴付けするために、得られる物質の毒性データを完全にレビューすることである。これには、すべての利用可能なデータから有害作用または有害作用の前兆を、さらに NOAEL または BMD の基本となる最も関連するエンドポイントを決定することが含まれる。有害性の特徴付けに対するエンドポイントデータのレビューに関するガイダンスは、発がん性、発生毒性、神経毒性、および他の人体への影響に焦点を当てた EPA の多数の出版物中に見られる (U.S.EPA 1991,1996a,1998,2005a)。これらのプロセスは、BMD または NOAEL 法のどちらを使用するかということと本質的には同じである。以後の項目での考察は、BMD 法を使用する際の試験デザインおよびデータの報告に関する、より重要な問題点の一部をまとめている。データ評価に関連し、本セクションで検討している意志決定ステップの一部を、図 2A のフローチャートにまとめている (セクション 2.1.5)。本ガイダンスでは、有害性を特徴付ける方法、特に毒性の決定およびエンドポイントの選択方法に関して、変更を加えていない。本ガイダンスは用量反応のモデル化に最も受け入れられるデータタイプと試験デザインを検討し、NOAEL/LOAEL が一部のデータベースで継続して使用される可能性を認めている。NOAEL/LOAEL 法に頼ることは、データセット本来の限界を解決しないが、データセットに限度があることを認識させる。

2.1.1. 試験デザイン

一般的に、用量群が多く用量に対し段階的で単調に反応する試験は、BMD 解析にはより有用である。対照群と異なる反応を示す用量群が 1 つだけの試験では BMD 解析を支持しないことがあるが、といても、反応上昇の 1 つが BMR に近ければ、結果として適切な BMD と BMDL が計算される可能性がある (Kavlock ら 1996)。反応がバックグラウンドと同一か、または最大反応レベルあるいはそれに近い結果だけを示す試験では、適切な BMD 解析にはならない (考察はセクション 2.1.5 を参照)。より良い BMD 推定値を得るために、1 つ以上の用

量が BMR レベルに近い試験があることが好ましい。

2.1.2. データ報告の側面

多くの場合、リスク評価者は重要な毒性学試験報告書のまとめを信頼するにちがいないが、それは BMD 法のデータ要件と比べて完成度は様々である。最適な状態は、個々の被験動物の情報を有することであるが、ピアレビューの文献では見込みはない。測定された作用、特に連続反応の変数に関する情報（群のレベルでの情報、例えば、平均および SD）にまとめることがより一般的であり、その情報のまとめが、適用する BMD 法にとって十分であるかを決定しなければならない。二分（非連続）データは、通常、個々のレベルで報告される（例、50 匹中 11 匹の動物が作用を示した）。時として、二分エンドポイントは、作用を示す動物数が示されず 1 群で観察されるとして報告される。こういったことは通常、報告されたエンドポイントの発生が報告の焦点から見て付属的なものである場合に起こる。二分データの BMD モデル化には、反応を示す動物数およびその群の全被験動物数の両方が必要である。

連続データは、対照群と曝露群に体重または酵素活性など、作用に関する測定値として報告される。反応は、数種の異なる方法、例えば、実際の測定値または対照からの相対的变化の対比として報告されるかもしれない。個々の動物データが入手できない場合に連続データをモデル化するには、被験動物数、反応変数の平均値、およびバラツキの指標（例、標準偏差（SD）、標準誤差（SE）、または分散）が各群に必要である。SD または SE が数値で報告されていないと BMD の計算ができない可能性がある。一部の例では、バラツキの指標が対照群にのみ提示されており、例えば、曝露群は対照群と同一であるというような仮定をたてることにより、この情報がモデル化に使用される可能性も少しある。しかし、この仮定は正しくない可能性があり、バラツキの情報が個々の群について得られている場合ほどにはデータのモデル化と信頼限界の計算の信頼性と精度は高くない。

カテゴリカルデータは、2 つ以上の定義されたカテゴリが、無作用カテゴリ以外に存在するデータである（カテゴリ内の反応は非連続である）。投与群での観察が、作用の重症度（例、軽度、中度、重度の組織的变化）で特徴付けられる場合、これらは順序付けられたカテゴリカルデータである（順序データとも呼ばれる）。カテゴリについて全投与群として報告するか（群レベル報告）、または各群の各カテゴリの動物数を報告することにより、結果は分類される可能性がある。例えば、上皮変性病変の報告では、曝露群が軽度の作用を示した（群レベル）、または曝露群で 7 匹の動物に軽度の作用が、3 匹は無作用であったことが提示される可能性も僅かにある。後者の場合、BMD は重度のカテゴリのデータを組み合わせた後、非連続モデルを使って計算できる（例、軽度を超える作用を示した全動物をモデル化する）。二分データは、1 つの作用カテゴリにおいて考えられる反応が（作用のある場合とない場合、など）2 変数である特別な事例とみることができる。モデル化の方法は、複数のカテゴリを有するカテゴリカルデータ（Dourson ら 1985 ; Hertzberg 1989 ; Hertzberg および Miller 1985）と、群レベルのカテゴリカルデータ（Guth ら 1997 ; Simpson ら 1996 a, b）について検討されてきた。これらのモデルは、異なる重症度レベルの作用の確率を推定することにより、BMD を誘導するためにも使用できる。

さらに、一般的にデータの評価に関しては、データ（反応と用量）は、可能な限り有効である必要がある。例えば、可能な元のデータ源を調査すべきであり、著者による意図的な用量群または被験動物の削除はすべて認識され、その根拠も理解される必要がある。対照条件の適切性も評価する必要がある。2 つのタイプの対照群が解析に使用可能であるなら、一般的に最も適切な方が選択される（例、溶媒対照）。

2.1.3. モデル化される試験の選択

毒性データのレビューが完全に終了後、リスク評価者は、扱われたヒト曝露状態、試験の質、報告の適切性、およびエンドポイントの関連性を基に、BMD 解析に用いる試験を選択する。BMD 解析用の試験を選択するプロセスは、BMD が算出できるように、モデル化が可能な試験を特定することを意図している。関連する試験すべてのモデル化を検討すべきである。ある事例では、選択プロセスにより、算出が適切であると同定される試験は1つであるかまたは非常に少ない。他の事例では、多くの BMD の算出を必要とする、多数の試験または多数のエンドポイントが報告された試験が存在する可能性がある。後者の事例では、標的器官または試験中の作用を代表するものとして、エンドポイントのサブセットを選択することが可能なこともある。この選択は、感度と重症度に基づいてなされるため、同一の標的器官では、試験間で比較するよりも単一試験内で比較する方が簡単であると考えられる。時として、数種のデータセットを組み合わせる選択肢もある（詳細な考察はセクション 2.1.6 を参照）。

2.1.4. モデル化されるエンドポイントの選択

一旦 BMD のモデル化の実現可能性について試験を評価したら、モデル化するエンドポイントの選択では用量反応関係に焦点を当てるべきである。通常、リスク評価者が曝露に関連していると判断した試験中のすべてのエンドポイントについてモデル化を検討すべきである。このことは、リスクアセスメント適用に最も感度が高いと考えられる作用、通常は最小の BMDL を有すると考えられるエンドポイントが解析から除外されないことを確認するのに役立つ。NOAEL/LOAEL に基づくエンドポイントの見かけの相対的感受性は、BMD のモデル化後、BMD または BMDL に基づく同じ相対的感受性とは一致しない可能性がある。そのため、エンドポイントの相対的感受性は必ずしも事前に判断できるとは限らない。例えば、エンドポイント間の（BMR での）傾きの相違は BMDL の相対的な値に影響する可能性がある。POD の決定に使用される可能性がある異なる試験で選択されたエンドポイントはすべて、特に異なる UF が別の試験とエンドポイントに使用される場合には、モデル化されるべきである。リスク評価者は BMDL を選択し、科学的判断とリスクアセスメントの原則、およびモデル化プロセスの結果を使って POD として役立つ。比較および他の目的で複数のエンドポイントについてリスク推定の誘導を遂行することが望ましい場合もあることに留意が必要である。

2.1.5. BMD 計算の最小データセット

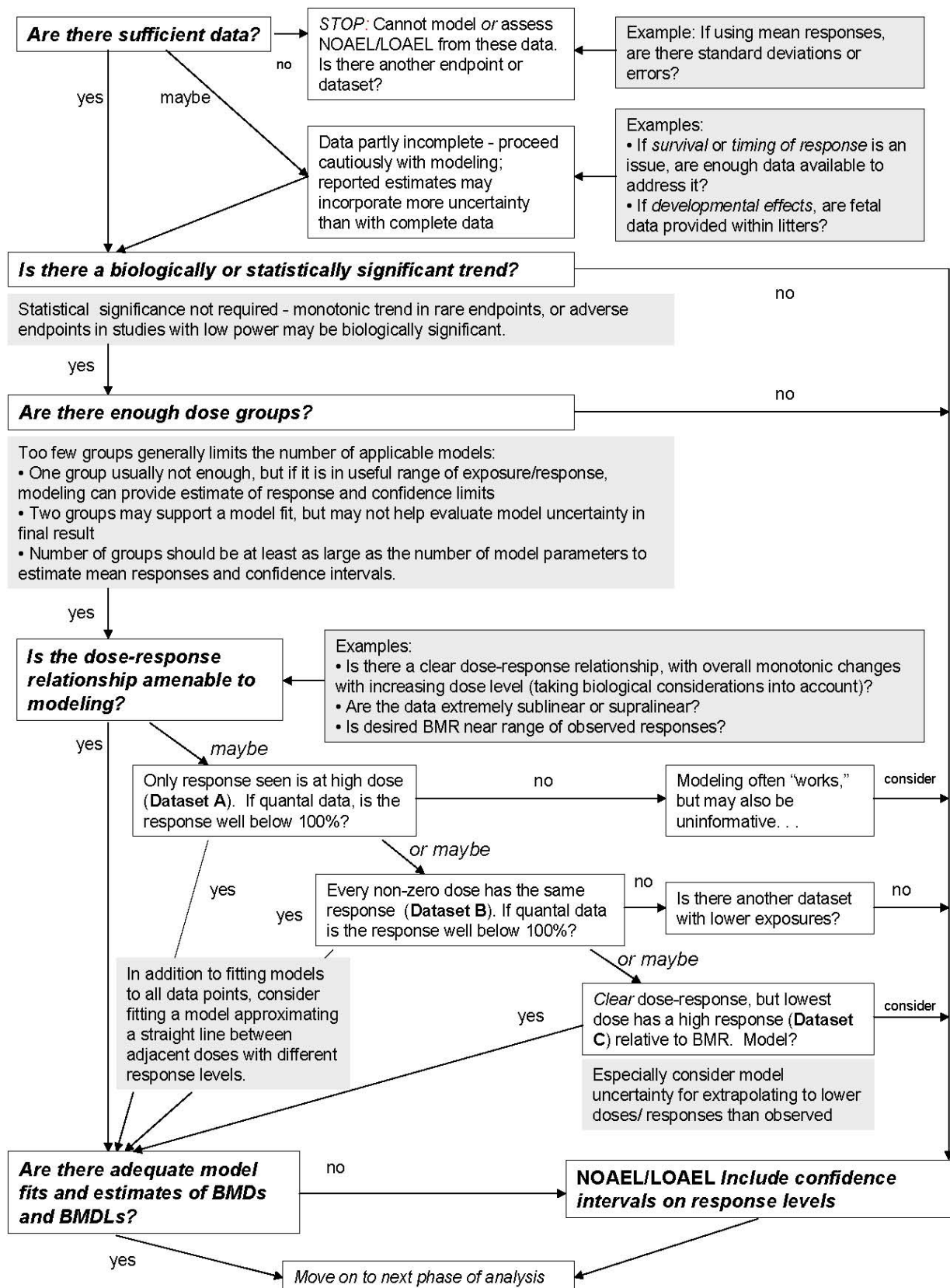
重要なエンドポイントを選択した後、データセットは BMD 解析の実現可能性について検討される。BMD のモデル化に推奨された最小データセットの基準は、図 2A および 2B に要約され、以下の基準を含む。

- ・選択したエンドポイントに、統計学的または生物学的に有意な用量に関連する傾向が少なくとも 1 つはなければならない^{注 5}。
- ・データセットは、対照群の最少値と観察された最大反応値との用量反応関係に関する情報を含んでいるべきである。理想的な状態は、BMR に近いデータポイントを有していることである。以下の例は、この最小データセットの基準を満たせない可能性のある事例を説明している。

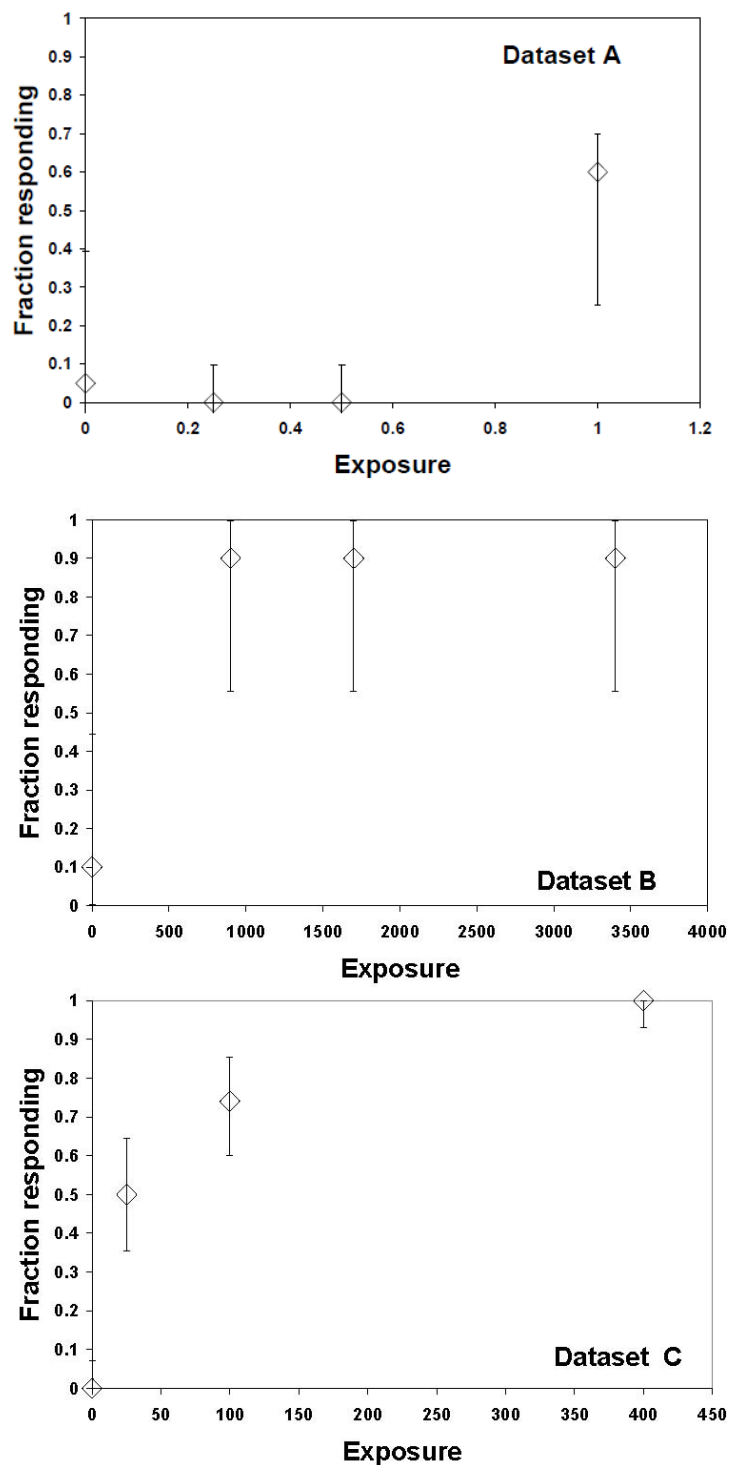
* 最大用量でしか反応しないデータセット（例、図 2B のデータセット A）は、最も低い値で BMD を一括するが、用量反応関係の形に関して情報を限定する可能性がある。そのような事例では、投与間隔と観察された反応レベルへの BMR の接近が、BMD 推定値の不確実性に影響を与えることになる。複数のモデルをデータセットに適合させると、この不確実性の大きさを評価するのに役立つ。モデル化の試行自体が、推定 BMD に関連する不確実性の程度に明察を与える可能性がある。

* バックグラウンドから最大までの反応の全範囲が最小用量未満のいずれかで起こり、すべての非対照用量が本質的に同一の反応レベルであるデータセット（例、図 2B のデータセット B）は、用量反応関係に関して情報が限定される。そのため、BMD は最初の用量にわずかに満たないか、桁数が少なくなる可能性がある。この場合、制限のないパワーパラメータを使ったワイブルのようなモデルを使用したくなるが（非連続データ、特に最大反応が 100% 未満の場合）、これでは低用量領域で異常に急勾配のモデルが得られる可能性がある（セクション 2.3.3.3 を参照）。そのような状態では残念なことに、データは低用量における用量反応関係について有益な情報を殆ど与えないのが現実である。理想的な解決法は、手持ちの試験から不足用量範囲のデータをさらに集めることである。

注 5 一部の事例では、生物学的有意性が同一の化学物質およびエンドポイントについて、他のデータから推論されることもある。



A. BMD のモデル化の実現可能性を決定するデータ評価ステップのフローチャート（データセット A、B、C は図 2B を参照）



B: 図 2A に対応するデータセット A、B、C の図解

図 2: [A: BMD のモデル化の実現可能性を決定するデータ評価ステップのフローチャート] [B: 図 2A に対応するデータセット A、B、C の図解]

2.1.6. BMD 算出に対するデータの結合

統計学的／生物学的に互換性のあるデータセットは、用量反応のモデル化前に組み合わせることができ、その結果、BMD 算出において統計学的／生物学的信頼性が増す。データセットを組み合わせる最も簡単な方法は、データすべてが同時に集計されたように取り扱うことである。複数のデータセットが用量反応の均一象を提示す

ることが妥当なら（例、2 つ以上のデータセットに共通な用量での反応が本質的に同一であり、統計学的に区別できない場合）、上述の方法は正当と認められるアプローチである。

Allen ら（1996）は、ホウ素関連の発生影響に関するデータを対数尤度の評価に基づいて BMD 解析に組み合わせた事例を示した。別の例は、1,3-ブタジエンの優性致死効果に対して EPA の評価で示されており、その中で同じ研究機関で実施した異なる 3 件の試験データが組み合わせられていた（U.S. EPA 2002b、セクション 10.3.3）。

ただし、データセット間にバラツキが存在することがあり、情報を適切に組み合わせるためにはより詳細にモデル化する必要がある。データセットをモデル化用に組み合わせる場合、統計学的に適切な方法と生物学的な判断を使用して正当化する必要があると言う以外に、このことに関する特定のガイドラインを提供するには、この状態での経験は今のところ殆ど实际的でも理論的でもない。試験間のバラツキを統計学的に調整する 1 つのテクニックは、カテゴリカル回帰解析（Simpson ら 1996a, b）であるが、この方法は注目する化学物質に関する多数の試験が必要である。単一エンドポイントで BMD を推定するために、データセットを組み合わせる際のバラツキを調整する例が、EPA による有機リン酸および n-メチルカルバミン酸系殺虫剤に関する累積リスクアセスメントに提示されている（U.S. EPA 2002c, 2005b）。

2.1.7. 用量測定の調整

実験動物に投与した用量を生体継続投与した場合のヒト相当用量（HED）に変換するために用量測定調整が用いられることが多い（例、U.S. EPA 1994, 2002a, 2011）。これらの調整を誘導または適用するためのガイダンスを提供することは本文書の範囲を超えているが、このセクションでは、用量反応のモデル化に先立ち、用量測定調整の検討が重要と考えられる一般的状況の一部について言及する。

種間外挿に関する仮定を喚起する前に用量反応関係を説明するため、実験動物の反応を用量（例、適用用量、内部用量測定）を使ってモデル化することが一般的に好まれる。調整が用量に比例する場合（例、継続曝露に対する一定の調整）、モデル化の前あるいは後に用量を調整するか否かは、最終的な結果には影響せず、利便性の問題である。

しかし、調整が用量に比例しない場合、用量反応モデル化前に用量測定調整をすることがより適切である可能性がある。これは、例えば、入手可能なデータが、体重の四分の三乗で種間補正するしかなく、体重が用量群間で明らかに異なる場合である。同様に、生理学に基づくファーマコキネティクス（PBPK）のモデル化は、用量と非線形なプロセスを反映することが多い。PBPK モデルに由来する用量測定が利用できる場合、複数の選択肢は考慮に値する可能性がある。BMD/BMDL を推定するのに、実験的曝露の用量／濃度を利用して非線形曲線を当てはめることができる。続いて、その BMD/BMDL は、実験動物の PBPK モデルを使って、ヒト相当値または妥当な用量測定レベル（例、肝臓の代謝物濃度の曲線下面積（AUC））に変換される。高度な超線形用量反応関係では、適用用量を使って十分に曲線を当てはめるのは難しいことがあるので、用量反応のモデル化に内部用量測定を使用することは有用である可能性がある。実験動物の PBPK モデルからの内部用量測定が使用される場合は、BMD と BMDL のヒト等価用量は、ヒト PBPK モデルから逆計算するかまたはその他の方法で推定される。内部用量測定に関する用量反応解析は、用量反応関係を簡単にして（例、代謝飽和による超線形曲線の直線化）、曲線を当てはめやすくする可能性があり、さらに PBPK 過程とファーマコダイナミクス過程の観察された用量反応関係への寄与を、対比して明らかにするのに役立つ可能性がある。

2.2. ベンチマーク反応（BMR）レベルの選択

BMR を選択することは、データベースの統計学的および生物学的特徴の判断と、得られた BMD/BMDL が使用される適用に対する判断を下すことである。EPA は、反応レベルまたは BMD の選択を判断し、殆どの適用を意図した BMD のモデル化（例、レファレンスドースまたは相対効力係数）に使用するガイダンスは現在のところ有しておらず、そのようなガイダンスは本文書の範囲を超えている。EPA による発がん性物質リスクアセスメントのガイダンス（U.S.EPA 2005a）では、発がんリスクの推定で BMR を取り上げている。本セクションでは、事例固有の問題とともに標準的報告として BMR を考慮するため、つまり化学物質またはエンドポイントを容易に比較するための一般的原則の一部について概要している。

通常、観察可能な範囲の下限に近い BMR を、BMD と BMDL を得るための基準として選択し、定量的推定値を観察範囲未満に誘導するための可能な POD として役立て、さらに化学物質、試験、またはエンドポイントに共通な反応レベルに対応する有効用量を比較するために使用する。異なる試験デザインでは、様々な用量で有害作用を観察するため異なる用量選択と異なる感度（すなわち、統計学的パワー）を有するので、観察の下限は、試験間の完全に異なる反応レベルに対応する可能性がある。反応が生物学的に有意と考えられれば、試験で対照群の反応と統計学的に有意に異なると検出される反応に、BMR を対応させる必要はない、ということを認識することは重要である。

一部のデータセットでは、観察値は選択した BMR をかなり超えた反応レベルに対応することがあり、観察可能な範囲未満で十分に外挿することはあまりに不確実で、選択した BMR に対し信頼性のある BMD/BMDL を推定できない（例、すべての投与群が最大反応に近い場合）。そのような事例では、BMD のモデル化は推奨されず^{注6}、その方法はデータの制限を解決することができないことを認識した上で、より多くのデータを得るか、または NOAEL/LOAEL 法を使うことが必要となる可能性がある（セクション 2.1.5 を参照）。

^{注6}モデル化されるデータセットの特徴が様々であり、作用機序など関連する情報の影響力が大きいため、詳細な決定プロセスは提供していない。BMD のモデル化に頼らない決定は、モデル化に熟知した統計学者または他の研究者、および考慮中のデータタイプまたは特定なデータベースに精通した科学者によって、事例毎になされるべきである。

以後のセクションでは、BMR の選択で使用する選択肢について述べている。非連続（二分）データについては、従来のアプローチはかなり正攻法である。一方、連続データについては、利用すべき前例となる方法は少ないが、いくつか妥当な選択肢が紹介されている。選択された各 BMR を支持する理論的根拠は与えられるべきである。一旦 BMR が選択され、用量反応データがモデル化されると、BMD は明確に決定される。

2.2.1. 非連続（二分）データ

上述したように、BMD/BMDL にはいくつかの応用例があり、BMR の選択では各々について決定が必要となる。二分データについて化学物質またはエンドポイント間で（例、化学物質の順位について）効力を比較する場合、10%超過リスクの反応レベルが、有効用量（すなわち、ED₁₀）としても知られる BMD を定義するのに一般的に使用されてきた。この反応レベルは多数の一般的な試験デザインにおいて、観察可能な範囲の下限に近いため、そのような比較に使用される。一般的に、化学物質／試験／エンドポイント間の比較は中央推定値に基づくことが推奨される。これは、基準値またはがん発生力推定値に対し POD の下限を使うことと対照的である。

しかし、POD の決定では、すべての化学物質またはエンドポイントに対して共通の反応レベルを使用することは必ずしも必要ではなく、観察可能な範囲未満の用量で定量的推定値を誘導する目的では、データセットの統計学的および生物学的特徴によって裏付けされるなら、10%超過リスク以外の反応レベルを使用することが望ましいこともある。さらに、疫学的データでは、10%超過リスクの反応比率は上向きの外挿に関与することが多いかもしれない。そういった事例では、より低いレベルを使用することが好ましく、1%超過リスクが **BMR** として使用されることが多い。**BMR** の選択において生物学的（例、作用機序などエンドポイントの性質）および統計学的（例、試験の感度）考慮に重きを置く場合の判断に対するガイダンスを提供することは、本文書の範囲を超えている。それでもなお、透明性を保つため、各 **BMD** の選択ではこれらの検討事項に対応しながら判断する必要がある。

そのため、比較の目的で 10%超過リスクに対応する **BMD**（および **BMDL**）を報告することは常に重要である一方、POD として使用される **BMD**（**BMDL**）は、10%未満（または 10%を超えることもある）の超過リスクの反応レベルに対応させることもある。標準化に向けて、概数として 1%、5%、または 10%が通常使用されてきた。

要約すると以下ようになる。

- ・化学物質またはエンドポイント間の比較目的には、10%の超過リスクが非連続データに対する標準報告レベルとして推奨される。反応レベル 10%は、殆どのがんバイオアッセイおよび比較可能なサイズの非がんバイオアッセイで感度の限界値であるか、またはそれに近いことから、習慣的に比較に使用されてきた。このレベルは、POD の作成または他の目的では初期設定の **BMR** ではないことに注意して欲しい。
- ・生物学的考察には、一部の作用タイプ（例、明確な作用）に 5%以下の **BMR** を使用するか、または基準値に対する POD の基本として 10%を超える **BMR**（例、初期の前駆作用）を使用することが必要となる可能性がある。
- ・（生物学的考察に基づき）10%未満の **BMR** が観察可能な範囲内に入ることもある。統計学的視点から、入れ子型試験デザインを持つ殆どの生殖および発生試験は 5%の **BMR** を容易に支持する。同様に、1%の **BMR** は通常、疫学試験の非連続なヒトデータに使用されてきた。他の事例では、観察可能な範囲未満でモデル化する場合、推定値の不確実度が増大することを頭に入れておく必要がある。そのような場合、**BMD** と **BMDL** は過剰な逸脱について比較することができる。さらに、モデルの不確実性はデータ範囲未満で増大する。

2.2.2. 連続データ

連続データでは、**BMR** の選択には色々な可能性がある。どの値が選択されるかに関わらず、対照群の平均値の 1 SD に等しい平均反応値の変化に対応する **BMD**（および **BMDL**）が、常に比較の目的で提示されることが推奨される。この値は、比較において標準化の基本として役に立ち、二分データの 10%超過リスクに対応する **BMD** のような物である。

理想的には、連続データの **BMR** は生物学的根拠を有する。例えば、連続エンドポイントの最小レベルの変化が何かという、合意が得られた科学的定義が生物学的に重要である。**BMR** 選択の生物学的根拠が存在する場合、モデル作成者は、連続モデルに適合させて **BMR** としてこの定義された変化レベルを使用するか、またはそのレベルの変化を基にデータを「二分して」非連続モデルに適合させるかという選択肢を持つ（以下のセクション 2.3.3.1 を参照）。後者のアプローチでは、情報が失われる結果となるが、使用可能な連続モデルがデータに不適

切である場合に有用となる可能性がある。例えば、成人体重の 10% の変化が生物学的に有意と考えられると、(10% の変化の BMR とする) 連続モデルは、平均 10% の体重変化に対応する BMD を与えることになる。データの二分化では、体重変化が 10% 以上である被験者の発生率として個々の体重変化のデータをまとめる。それでも続いて、超過リスク 10% と言われる、発生の重要な変化に関連する BMR を定義する必要がある。非連続モデルは、次いで、10% 以上の体重変化を有する被験者の超過リスク 10% に対応する BMD を推定するために使用される。しかし、この場合、BMD に関連する平均体重変化の情報は得られない。

生物学的有意性でカットポイントを確立するための十分な生物学的根拠がない場合、対照群の分布のパーセンタイルに基づきデータを二分化することが合理的かもしれない (例、Kavlock ら 1995)。

別のアプローチは、連続モデルを連続データに適合させるが、非連続データ (例、超過リスク) に使用される表現で BMD 示す「ハイブリッド」法である。データの二分化と同様に、2 つ以上のカットポイントが定義されなければならない。ハイブリッドモデルの実施に関する詳細情報は、セクション 2.3.3.1 を参照のこと。

連続データに対する BMR を選択するための説得力のある根拠が全くない場合、対照群の平均値からの 1 SD (統計的および生物学的検討によって妥当であるなら^{注7}、これ以下または以上のこともある) の変化の BMR を使用でき、連続データの比較用に標準化された報告レベルと同様に推奨される。対照群の SD は、以前のコントロールデータを含めて計算できるが、対照群の平均値は全般的に、考慮する処理群と同時に得たデータから算出すべきである (Crump 1995)。別々に推定可能な場合、測定誤差による SD とは別に、動物間の SD を使用すべきである (Gaylor および Slikker 2004)。しかし通常、測定誤差を含む全体の SD だけが利用可能である。Gaylor および Slikker (2004) によれば、全体の SD を使用すると BMD を高く推定することとなる。しかし、測定誤差の SD が動物間の SD の 1/3 未満であればバイアスは比較的小さく、この条件は、著者らによると殆どの実験デザインで達成可能である。

対照群の平均値の 1 SD のシフトは、正規分布効果に対する対照群の 1.4 パーセンタイル未満または 98.6 パーセンタイルを超えた被験動物の割合に対する 10% の超過リスクに相当する^{注8} (説明については図 3 を参照)。

^{注7} 例えば、入手可能なデータが試験デザインや作用機序などから、外挿を裏付けるような場合である。

^{注8} 1 SD の BMR は、一定の分散を有する正規分布では、対照群の分布の 1 または 99 パーセンタイルを超える集団の 10% が SD の平均 1.1 倍の変化に相当するという Crump (1995) の観察と一致している。

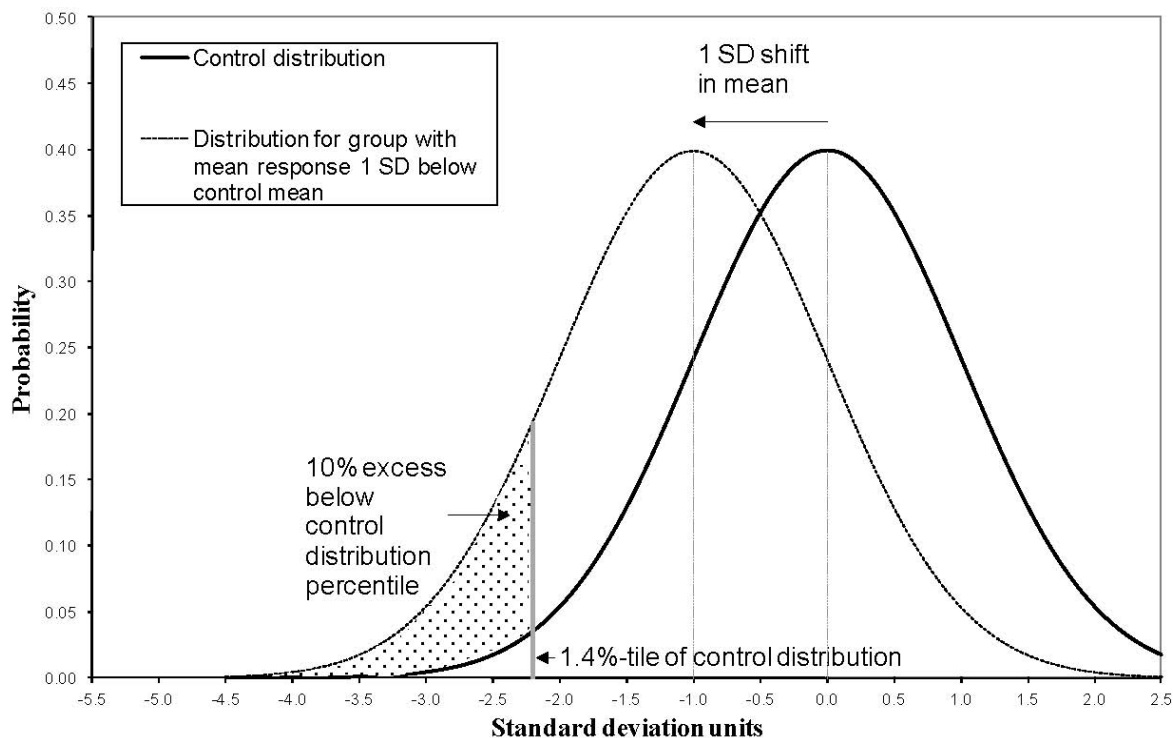


図 3:標準正規分布から平均の 1 SD シフトで得られる集団テール確率の差、1 SD の BMR 基準値の理論的根拠の図解

1 SD 変化は BMD の比較で推奨される BMR である一方、この値は、POD 決定の BMR として常に適当であるとは限らない。すなわち、対照群の平均値の 1 SD の変化は 1 群 10 匹以上の動物を使う殆どの試験で統計学的に有意であると思われ、対応する BMD は、もちろん測定された成績の生物学的有意性に依存し、一般的に LOAEL と解釈されるであろう。そのため、非連続データについて以前に検討したように、データの生物学的および統計学的特徴の判断がなされる必要がある。例えば、明確な作用では、より低い BMR が妥当である（例、0.5 SD）。

連続データで BMR を選択する際に生物学的根拠または確立された科学的合意（例、対照群の平均値からの体重の変化が 10%など）が無い場合、（異なるバラツキ度合の）異なるエンドポイントでパーセント変化が同一の場合、大幅な反応差に関連する可能性があるので、対照群の平均値のパーセント変化を BMR として使用することは推奨されない。

時として、BMR の定義の根拠として「ダイナミックレンジ」法が使用されてきた。これには、反応のバックグラウンドレベルから最大可能レベルの範囲で連続測定をスケーリングし（Murrell ら 1998）、合計反応にスケールされた反応の変化（例、1%または 10%）として BMR を選択することが含まれる。しかし、この方法は、達成可能な最大作用の信頼性のある推定値を必要とし（Gaylor および Aylward 2004）、高曝露量での毒性作用が低曝露量から質的に変化または相違するという状況では、あるパーセントのダイナミックレンジの毒性学的関連性は不確かである。さらに、BMR の他の定義との比較可能性は明確ではない。結果として、ダイナミックレンジ法は、現在では推奨されていない。

連続データの BMR 選択の全般的階層は、以下ようになる。上述で注意したように、選択した BMR に対す

る正当性は常に提供すべきである。報告の一貫性に関して、対照群の平均値の 1 SD シフトに相当する BMD は、どの BMR が POD に使用されても、BMD および BMDL と共に提示される必要がある。

要約すると以下ようになる。

- ・ **好ましい方法** 生物学的に有意であると一般的に考えられるエンドポイントに最小の変化がある場合、その変化量は BMR の定義に使用することができる。
- ・ 個別データが得られ、個々のどのレベルが合理的に有害であると考えられるか判断ができる場合、データは、ハイブリッドモデルにより暗黙的に二分化でき、またはそのカットオフ値に基づき明示的に二分化できる。さらに BMR は非連続データについて上述のように設定できる。明示的二分化では情報が欠損することがあるため、暗黙的二分化は明示的二分化よりも好ましいことに注意されたい。
- ・ どのレベルの反応が有害と考えられるかについてその他に一切見解がない場合、対照群の平均値からの 1 SD に等しい平均の変化（またはより重度の作用についてはこれより低い 0.5 SD など）が使用される必要がある。

2.3. データのモデル化

2.3.1. はじめに

BMD 算出における数学的モデル化の目的は、モデルをデータセットを説明する用量反応データに、特に観察可能な用量反応範囲の下端に、適合させることである。適合は、パラメータ推定値に関連する不確実性を定量化し、BMR を生み出す用量推定値に関連させることができる方法で実施しなければならない。実際、この手順はまずデータと実験デザインの特徴に基づき、さらに検討するためにモデルのファミリーを 1 つあるいは複数選択し、次いで数少ない確立した方法の 1 つによってそのモデルを適合させる。

その後、BMD と BMDL が BMR で計算される。このセクションはあまりにも短すぎて、モデル化の課題を紹介できない。さらに読むべき文献の一部として、Draper および Smith (1981, Chapter10)、Gallant (1987)、Bates および Watts (1988)、McCullagh および Nelder (1989)、Seber および Wild (1989)、Ross (1990)、Clayton および Hills (1993)、Davidian および Giltinan (1995)、Piegorisch および West (2005)、Nitcheva ら (2005)、および Wu ら (2006) などが挙げられる。

用量反応モデルは、用量、可能なら共変量、および定数セットの関数として表されるが、これらはパラメータと呼ばれ、得られたカーブの形を詳細に制御する。それらのモデルは、観察される用量値と共変量に対するモデルの予測値を、観察される反応に近くなるように補正するパラメータ値を見つけることにより、データベースに適合される。毒性学データの用量反応モデルは、数学用語では、通常「非線形」と呼ばれるタイプである。線形の数学モデルでは、モデルが予測する値は、パラメータの線形的組み合わせである。例えば、用量に対する反応 y の線形回帰において、推定値は、下式のような、 a と b の線形の組み合わせである。

$$a \times 1 + b \times \text{用量}$$

二次式または他の多項式も線形数学的モデルであることに注意されたい。この意味において、

$$y = a + b \times \text{用量} + c \times \text{用量}^2 + d \times \text{用量}^3$$

上記式は三次多項（三次元）式であるが、パラメータ a 、 b 、 c および d は、なおも線形の組み合わせである。これに対して、次のような非線形数学的モデル、例えばバックグラウンドを有する対数ロジスティックでは、下記式の通りであり、

$$p = P_0 + \frac{1 - P_0}{1 + e^{-[a + b \log(\text{用量})]}}$$

反応は、パラメータ（ここでは、 P_0 、 a および b ）の線形の組み合わせではない。パラメータが非線形であるモデルは通常より複雑な計算を必要とし、データへの適合がより難しく、また統計的推定はパラメータが線形であるモデルよりも近いことが多いため、区別することは重要である。

「線形」についてのこの定義は、用量と線形であるモデルを参照するために、用量反応モデル化でより特異的に使用されることが多い方法とは大きく異なっている。

最終的なモデル選択の基準は、様々なモデルがデータを説明するか否か、考慮した特定エンドポイントの慣習、および時には、同一の基本モデル様式を複数のデータセットに適合させる要求に基づくことになる。（うまく文書化された）特別な目的のモデル化ソフトウェアを使用することが好ましいので、EPA は、本文書で説明されている数種のモデルおよび初期設定プロセス（例、パラメータの制約）を含んだソフトウェアを開発した（<http://epa.gov/NCEA/bmds/>）。便宜上、本文書で述べたモデルの説明を付録 C に載せてある。

2.3.2. モデル選択の背景

適切なモデルの選択を優先するには、特定事例で作用すると理解される生物学的プロセスと一貫性のあるモデルを使用すること、および明らかに一貫性がないモデルを避けることである。用量反応モデルで取り上げうる特徴は、生体プロセスの明確な発現（例、Moolgavkar および Knudson (1981) と Chen および Farland (1991)）によって開発された 2 段階クローナル拡大モデル、およびミカエリス-メンテンモデルで特徴付けられる飽和プロセス）および考慮される反応の主な共変量（例、多段階ワイブルモデルでの反応時間、発生試験の入れ子型モデルでの前処理済み母体重）である。

生物学に基づくモデルがない場合、用量反応のモデル化は得られる多様な経験的モデル間で大幅な曲線適合が必要となる。モデルの選択を効率よく行うと思われる、推奨されるモデル階層は現在のところ存在せず、理由の一部は、用量反応パターンに影響するデータセットおよび試験デザインに多数の異なるタイプが存在するからである。よりフレキシブルなモデルが開発されると、エンドポイントのカテゴリに対する階層はよりふさわしいものになるだろう。一部のモデル階層が、好ましい慣例として確立されることも考えられる。例えば、がんバイオアッセイデータの発がん用量反応のモデル化に多段階モデルを選択することが、EPA の IRIS プログラムによる最新の慣例である（Gehlhaus ら、2011）。多段階モデル（実際、異なる段階の多項式モデルのファミリー）は、殆どのがんバイオアッセイデータに十分適応性があり、それを使用すると発がん用量反応解析に一貫性を与える。このセクションでは、解析されているデータに対するモデル構造の選択、モデルの適合、モデルの比較、および BMDL を POD として使用するように誘導するための信頼限界の計算について、基本的な統計学的背景とガイダンスの一部を提供する。

2.3.3. モデルの選択

データに適合するモデルグループの最初の選択は、注目するエンドポイント、およびデータを得るのに使用した実験デザインを表す測定法の特徴によって左右される。さらに、モデルの特定の制約またはそのパラメータ値を、定期的に観察する必要がある、モデル選択に影響を与える可能性がある。最終的に、同時に複数のエンドポイントをモデル化することが望ましいこともある。様々な物質に対して、候補となるエンドポイントと用量反応関係の形は多様であるため、BMD 算出用にセット数の少ないモデルを特定することができない。このため、用量反応評価用に最終モデルと BMDL を選択する場合、必然的に判断が必要となる。用量反応評価への BMD 方法論を使用した経験が蓄積すると、あるシナリオで特定される、十分に適応性があり、冗長ではないモデルの数が少なくなる可能性もでてくる。

2.3.3.1. エンドポイントのタイプ

注目するエンドポイントを表す測定変数の種類は、数学的モデルの選択に重要な検討事項である。一般的に、そのような変数は、肝重量または特定の肝酵素活性のように連続的か、あるいは肝の異常状態の有無のように二分的（非連続）である。しかし、生物学的データで見られるタイプもある。例えば、正常が 1 できわめて異常を 5 とする範囲の組織学的スコアのような順序カテゴリカルである。直面する可能性のあるすべての変数を考慮することは、本文書の範囲を超えており、詳細な検討は二分変数と連続変数に留める。

二分変数 二分変数データは、一般的に特定の用量または曝露レベルで、特定の状態を呈する個体の割合またはパーセントとして提示される。二分データのモデル化では正確な個数を使用することに注意が必要である^{注 9}。そのようなエンドポイント対し、我々は普通、ロジスティック、プロビット、ワイブル分布など、ゼロ用量を含むすべての可能な用量に対し予測がゼロと 1 の間にある確率密度モデルを選択する。

^{注 9}特に、必要に応じて、例えば poly-3 アプローチ（Bailer および Portier 1988）によって生存を補正した分母が使用される可能性があることに注意する。

連続変数 連続変数のデータは、平均値と SD（標準偏差）または SE（標準誤差）として提示されることが多いが、対照群または他の基準に対する % として提示されることもある。モデル化の観点から、そのようなデータの最も望ましい様式は個別による。二分変数の通常の条件と異なり、連続変数をまとめると、これらの変数の分布に関する条件を失う結果となる。さらに、解析で共変量の使用が目的である場合、個別のデータが必要とされる。

BMR の確立に使用される方法が連続データのモデル化へのアプローチを決定するであろう（考察はセクション 2.2.2 を参照）。アプローチの 2 つの大まかなカテゴリが以下のように提案されている。

- 1) BMR が連続エンドポイントの変化レベルとして定義される場合（通常、平均反応の特定変化として、可能なら対照群の平均の割合として、または非投与群の測定値の SD の割合として）、連続モデルが使用できる。典型的な連続モデルには多項式モデル、パワーモデル、および Hill モデルがある。
- 2) データが二分化され、BMR が連続エンドポイントの変化が特定レベルよりも大きい被験動物の割合として定義される場合、得られる変数は二分変数としてモデル化できる。しかし、二分化では情報が失われ、一般的には回避すべきであることを喚起したい（セクション 2.2.2）。

別の方法は、Gaylor および Slikker (1990)、Kodell ら (1995)、および Crump (1995) により説明されたようなハイブリッド法を使用することであり、そこでは、連続モデルを連続データに適合させ、データの分布を推定することで、影響を受けた割合から BMD を計算する。この方法にを使うと、有害レベルを有する被験者の確率（リスク）を用量の関数として直接推定できる。ハイブリッド法は以下に概要するように、4 つのステップを使用する。

最初のステップでは、対照群の個々の連続測定に対する確率分布が特徴付けられる。この分布はおおよそ対数正規分布（すなわち、生物学的測定値の対数が正規分布する）であると考えられることが多い。殆どの生体作用は負の値を前提としないので、対数正規分布はこの条件を満たす。高値が有害である場合、分布の高いパーセンタイル値（例、99 パーセンタイル）が正常レベルのカットオフ値として選択される可能性があり、値が高いほど有害と考えられる。逆に、低値が有害である場合、小さいパーセンタイル値（例、1 パーセンタイル）がカットオフとして選択され、値が低いほど有害と考えられる。

2 番目のステップでは、用量反応モデルを、平均値が用量関数としてどのように変化するかを確認するためのデータに適合させる。3 番目のステップでは、平均値について被験動物間のバラツキが計算される。これは平均値の SD によって簡単に表現できることが多い。平均値に比例すること、すなわち変動係数が一定であることは、生物学的測定 SD では一般的である。これもまた、対数正規分布の特性である。しかしながら、変動係数が用量とともに変化することもあり、データの解析はより複雑となる。この事例では、累乗した平均値に比例してバラツキをモデル化することが有用な場合が多い。この方法でバラツキをモデル化すると、変動係数が一定となり、バラツキが平均値の二乗に比例し、変動係数が比例定数の平方根となる状態に対応できる。（付録 A の例 A.3 で、標準的連続モデル、つまり Hill モデルのバラツキのモデル化について説明する。）

2 番目のステップの用量反応モデルから推定した平均値、および 2 番目のステップで推定した平均値の変動値から、4 番目のステップでは、個体が最初のステップで確認された有害範囲内にある確率を、いかなる用量についても推定することが可能となる。一方、BMD（および BMDL）は特定の BMR を推定できる。

上述 3 つの各テクニックの使用を説明した例は、EPA による 1,3-ブタジエンの健康アセスメントにおける胎児体重解析に見出すことができる（U.S. EPA2002b、セクション 10.3.2）。この解析では、連続パワーモデルは、連続エンドポイントとして平均胎児／同腹児の平均をモデル化するのに使用された。対数ロジスティックモデルは、データの二分化後、同腹児の関係を考慮して使用され、ハイブリッドモデルが適用された。

2.3.3.2. 実験デザイン

モデルの選択に影響する実験デザインの特徴には、使用する投与群総数と実験対象のクラスタ化の可能性が含まれる。投与群数は推定されうるパラメータ数に影響を与える。つまり用量反応曲線の全体の形に影響するパラメータ数は通常、投与群数を超えることはできない。

被験動物のクラスタ化は、実際にはモデル形態自体の選択よりも、モデルの適合方法の問題である。クラスタ化が起こる最も共通する状況は、発生毒性の実験においてであり、そこでは、物質が母動物に投与され、同腹児内の各児について有害作用が検討される。（付録 A、例 A.5 参照）。

クラスタ化の別の例は、個々の被験対象が複数の所見をもたらすデザインに関してである（反復測定）。これ

は、例えば、各被験動物が投与物質およびコントロールの両方を投与されるか（ヒトを対象とする試験に多い）、または各被験動物が投与後、複数回観察される（例、神経毒性試験）場合に起きうる。これらの例すべてにおける問題は、互いに独立して個々の観察を行うことが出来ないことにある。モデル適合で使用される殆どの方法は、データが独立しているという仮定に依存しており、より複雑な依存関係を呈するデータセットでは特別な適合方法を使用する必要がある（例として Ryan 1992a,b ; Davidian および Giltinan 1995 を参照のこと）。詳細は本文書の範囲を超えている。

2.3.3.3. 制約と共変量

用量反応のモデル化では、モデル作成者は、数値的に可能なパラメータ値の設定を制約するような、選択肢を考慮しなければならない可能性があり、これは通常結果の生物学的妥当性を強化することが目的である。

二分データ用モデルでの明らかな制約についてはすでに考察した。つまり、確率が 1 以下の正の数に制限される。生物学的現実性はモデルに他の明確な制約を課す。例えば、殆どの生物学的測定値は正であり、そのため、その予測値は、少なくとも適用範囲内で、その制約に適合するようにモデルを選択する必要がある。

他の選択は、用量反応パターンの生物学的妥当性で対処する必要がある。多数の毒性作用では、用量に伴う作用の単調な増加が予想される。つまり、高用量では、低用量以上の作用が得られる。そのため、既存の実践では殆ど、単調なモデルに制約されており、例えば、多段階モデルの適合では、パラメータは負にならないように制約される。一部の状況では、単純ではない関係が見られることがあり、通常、質的に変化した生体メカニズムまたは高用量データに伴う観察限界がある場合である（セクション 2.3.6. を参照）。

他に課題となるのは、一部のパラメータ値について急勾配で超線形となるモデルについてである。推定パラメータであるパワーまで用量を増やしたモデル（ワイブルモデルなど）では、1 未満のパワーパラメータ値について、低用量で用量反応曲線の傾きは非常に急になる。これは評価者にとって難しい課題となりうる。一方、超線形の反応パターンを示すことは、観察された範囲のデータでまれではなく（例、ミカエリス・メンテンの関係式の形）、そのため 1 未満のパワーパラメータを排除しても、データに最適に適合せず、または観察範囲の反応に対する不確実性の適切な評価ができない。原則として、BMD のモデル化で非常に低用量まで外挿することは一般的に求められないので、元のモデルに近い制約されない一部のモデルで見られる急な傾きは、それ自体基本的な問題とはならない。一方、場合によっては、計算された BMD および BMDL はパワーパラメータが 1 未満の場合に非常に低くなりうる。これは、このデータは用量反応曲線の下限を制約しないという事実を反映している。例えば、本文書の始めの方で述べたように、特に問題があるのは、非コントロール用量すべてが選択した BMR を超える反応レベルであり、反応が平坦または浅いというようなデータセットについてである。この状況は非連続データで起こる場合、特に最大反応が 100% 未満なら、パワーパラメータに制約がないワイブルのようなモデルを使用したくなる。その理由は、そのようなモデルは 100% 未満でプラトーに達し、さらに殆どのモデル化プログラムはこの特性を有する非連続データに対する他のモデルを含まないからである。しかし、そのような制約のないモデルを使用すると、反応のすべての変化が起こっているこのような低用量範囲では、データが用量反応曲線を制約しないため、非常に不明確な BMD が得られる可能性がある。理論的には、BMD を強制的に、最小 BMDL 値と最小投与量の間の値にするような他のモデルが見出される可能性もありうる。よって、ここで算出された BMD は、選択したモデルだけに依存しており、適合度は可能性の選択には役立たない。そのような状態では残念なことに、データは用量反応関係について有益な情報を殆ど与えないというのが現実である。理想的な解決法は、手作業で、試験に入っていない用量範囲について別なデータを集めることである。

一般的に、モデル作成者は、制約パワーパラメータは 1 以上であると考えるべきである（これは **BMDs** ソフトウェアの初期設定値である^{注 10}；例 **A.1** を参照）。しかし、観察されるデータが超線形に現れた場合、非制約モデルまたは漸近項を含むモデル（例、**Hill** モデル）は、それらのモデルが妥当な **BMD** 値と **BMDL** 値を裏付けることができるか否かの調査を必要とする。裏付けのない場合、他のモデル形式を **POD** に考慮するべきである。時としてモデル化から有用な結果を得られず、**NOAEL/LOAEL** アプローチが考慮されるかもしれない。しかしデータのギャップおよびそのアプローチの固有の限界を理解しておくべきである。

注 10 そのようなパラメータが制約され、推定中にその制約値 (=1) に設定された場合、その事実を報告すべきである。こうなった場合、信頼区間の名目上の範囲が正確でなくなり（漸近的）、さらに真（未知）のパラメータが <1 であると、意図する範囲よりもかなり狭くなる可能性があり、これについても報告すべきである。

非連続モデルでは、バックグラウンドパラメータが、モデル化される結果が曝露せずとも起こり得る確率を定量化する。バックグラウンドパラメータ値をゼロに固定することで、推定されるパラメータ数を減らしたいという気持ちが起るかも知れない。しかし、曝露なしでは結果が起こらないであろうことが明白な場合のみ、バックグラウンド値をゼロに固定することが適切である。

モデルでいわゆる「閾値」という用語を含めることは、全般的に **BMD** 解析では推奨されない。そのようなパラメータは生物学的閾値の推定値ではないが、用語の混乱により間違えやすい。さらに、殆どのデータセットはこのパラメータがなくとも、それに関連する自由度のロスは起こらずに適切に適合できるからである。しかし、まれに、反応の増加がかなり急激なため、用量反応モデルの化には一般的に利用できるモデルと共に閾値パラメータを含めることが必要となることがあり、そのような事例では、パラメータを含むことが容認される。

用量反応モデルを適合する場合、個別に共変量を含めることが望ましいこともある。例えば、同腹子のサイズを、発生毒性試験で実験動物データのモデル化の共変量として含めることが多い。別の例は、免疫学的データをモデル化する場合であり、この時、結果に影響を与えることが期待され、曝露と相関する可能性がある共変量（例、年齢、出産回数）が含まれる。共変量が反応に影響する場合、それをモデルに含めると、共変量をモデルに含めなければ最後に残差分散となるバラツキを明らかにすることにより、全体の推定値の精度が改善されることがある。用量に相関し（因果関係なし）結果に影響を与えるいずれの変数も共変量として考慮すべきである¹¹。

¹¹ 特定の体重範囲の場合のように、共変量が試験集団のサブセットを定義することに注意されたい。従って、異なるレベルの共変量に対して明確に異なる **BMR** が必要となる可能性がある。

2.3.4. モデルの適合

適合プロセスの目標は、得られた適合モデルが可能な限りこれらのデータを説明するような、すべてのモデルパラメータ値を見出すことである。つまりこれは「パラメータ推定」とも言われる。これを達成する 1 つの方法は、関数全体の最小値（または同様に全体の最大値）に対応するパラメータ値が、望ましいモデル予測を与えるという性質を持つように、全パラメータと全データの関数（目的関数）を、同定することである。

目的関数を構築して最適化する最も一般的な方法は、最大尤度、非線形最小二乗、および一般化推定方程式（Generalized estimating equations: **GEE**）などである。目的関数の選択は、適合モデルのエンドポイントと

データのバラツキの特徴によって大部分が決定される。そのため 1 つの方法だけがデータタイプに適合することが時々ある。これらの方法を一般的なデータタイプで使用する場合のいくつかの限界によって、さらに以下で説明する。

- ・二分データ—そのような状況の例は、個別に独立して処理された動物（すなわち、同腹児にクラスタ化しない）が単一反応の存在に対してスコアされる場合である。ここで、反応する動物数が、用量関数として表される反応の確率により、二項分布に従うと仮定することは妥当である。

- ・連続変数、特にいくつかの観察値の平均は、正規（ガウス）分布または対数正規分布になることが多い。変数が一定のバラツキで正規分布する場合、二乗の合計を最小にすることは、尤度を最大にすることと同等である。このことは、以下に考察されているように、なぜ最小二乗法が連続変数に使用されることが多いかの一部を説明している。

- ・発生毒性データでは、妊娠母動物が実験単位であり、統計学的方法によって、同腹児が同様に反応する傾向を説明できるはずである。有害結果を示す動物数の分布は、同腹児間の独立性の欠落に対応するため、だいたい β 2 項分布であると解釈されることが多い（同腹効果：例、Chen および Kodell 1989 ; Williams 1975）。この方法の 1 つの欠点は、同腹効果が間違ってモデル化されるか否かについてロバスト性に欠けることである（Kupper ら 1986 ; Williams 1988）。別な解析は、擬似尤度、あるいはより一般的な、一般化推定方程式に基づき、以下に検討されている。単純なデータ変換による簡単な方法が Rao と Scott (1992) および Krewski と Zhu (1995) によって説明され、GEE または最大尤度法のように効率が高いことが示された（Fung ら 1998）。最大尤度法は、データ分布に適切な仮定を行うことができる場合、目的関数を引き出す一般的な方法である。最大尤度法によって導き出された推定値は、漸近的正規性（ある種の規則性のある状況）のような良好な統計的性質を持っているので、最大尤度は、その仮定が真の値にかなり近い時、推定の好ましい形であることが多い。

非線形最小二乗法は、目的関数が観察データ値とモデル予測値との差を二乗した合計であり、観察値が独立して取る場合の連続変数を扱う一般的な方法である。この方法の基本となる仮定は、投与群平均付近の個別観察結果のバラツキが用量全域で一定であるというものである。この仮定に反する場合（一般的に、連続変数のバラツキが平均の関数として変化する場合で、平均の二乗に比例することが多く、一定の変動係数を与える）、方法が修正され（一般化非線形最小二乗法 ; Davidian および Giltinan 1995）、適合平均の関数としてバラツキがモデル化されることがある。この方法は、適合するデータが少なくともほぼ正常に分布していると推定される場合、特に関連性がある。

パラメータ推定の第三番目のアプローチは、関連偽似尤度法 (McCullagh および Nelder 1989) と GEE 法 (例、Zeger および Liang 1986 ; Liang および Zeger 1986) であり、それらは、平均、バラツキ、およびデータの相関構造が特定されるものについてのみ必要とされる。GEE 法は、繰り返し解を必要とする最大尤度の推定手順に似ており、漸近的正規分布であるパラメータ推定値と標準誤差の推定値、さらにパラメータ推定値の相関を与える。GEE は一般的に最も広範なデータタイプに適用可能である一方、よく知られておらず、その使用は今まで主に、同腹児データなどで独立性に欠ける様式を扱うことであった（例、Ryan 1992a）。これらの方法は、臨床試験および反復神経行動検査で起こるような、多数の反復測定デザインのいずれにも有用であろう。

一旦適切な目的関数が同定されると、モデル適合に「ベスト」なパラメータの決定においてより実際的な問題は、実際の適合プロセスをどうやって開始するかである。通常、ソフトウェアの常法はパラメータ値に対する最初の「推定」で開始される。次いで、この推定は、反復して更新され、（通常）1 つの値に集中する一連の推定値

を生じる。多数のモデルの殆どのデータセットで、開始パラメータ値の適切ないずれのセットからも正しい推定値に収束する。しかし、一部のモデルとデータセットでは、モデルが収束する前に、初期値として複数の推定値が必要なこともある。時には、適合手順が、異なる初期推定値から異なる推定値に収束することもある。これらの推定値セットが1つだけであることが「ベスト」である。パラメータの非線形モデルを適合させる場合、万の場合に備え、異なる初期値を使う試みは常に良い慣習である。この状況では専門的な判断が有用な可能性がある。(付録A、例A.3.を参照。)

2.3.5. いかにモデルがデータをうまく説明しているかの評価

適合モデルを選択する重要な基準は、モデルが特にBMRの領域内でデータを十分に説明していることである。殆どの適合法が全体的な適合度の指標、通常はp値を与える。これらの指標は、モデルによって予測される投与群の平均値が実際の投与群の平均値と異なる度合、つまり投与群の平均値にどの程度バラツキが予想されるかに関連させて、定量化する。p値が小さいことは、もしデータが実際にモデルからサンプリングされているならば、少なくとも極限において、統計的に値が適合度に達している可能性が低く結果としてモデルがデータにあまり適合しないことを示している。BMDのモデル化は通常、一組のモデルが関与して曲線適合が行われ、データがBMDの算出で適切にモデル化されることが重要であるので、適合度の臨界値^{注12}を算出するために、より慣習的な値の $\alpha=0.05$ または 0.01 に代わって^{注13}、 $\alpha=0.1$ を使用することが推奨される。この推奨値の例外は、特別なモデルを選択する先験的な理由が存在する時であり、その場合、より慣習的な値として $\alpha=0.05$ または 0.01 が考慮される可能性がある。p値は異なるモデルが正しいという仮定の下に推定されるので、1つのモデルを別のモデルと比較することはできない。つまり、実験結果に一貫したこれらのモデルの同定のみ可能である。同腹児サイズなど、モデルに他の共変量がある場合も考え方は同一であるが、算出がより複雑になる。この場合、用量と他の共変量の範囲は別のセルに分割され、各セルに入った観察値数はモデルによる予測値と比較される。

注12 χ^2 適合度検定では、臨界値が適切な自由度では χ^2 分布の $1-\alpha$ パーセンタイルである。この目的用に選択した第1種の過誤（偽陽性）の限界確率である α 値未満のp値に相当する、大きな χ^2 の値は棄却する。

注13 一部の事例では、使用可能な殆どのモデルは、適合度p値単独、すなわちp値が0.1未満に基づく適切ではないようにみえる可能性があることに注意されたい。(データ範囲の最も関連性のある領域のエンドポイント、視覚的適合、および残差に対するバラツキの特性を含む)他の基準を考慮すると、これらの適合度が低い一部のモデルが満たされる可能性がある。これらの事例では専門的な判断が有用である。

モデルがデータポイントから決して遠く離れていない（適合度検定のp値が小さすぎない）が、常に投与群平均値の片側またはもう一方の側にあるということは起こりうる。反応の広範囲な反応が存在することもあり得、モデルは高用量反応をうまく予測するが、低用量反応は予測できないかもしれない。そのような場合、適合度検定は、有意でないかもしれないが、適合は注意を払って扱う必要がある。そのような状況を検出する1つの方法は、残差の表またはプロット、モデルにより予測した反応の実際のデータからの偏差の測定によって行う。残差が推定されたバラツキ（SE）によりスケーリングされる場合、そのようなスケーリングされた、または標準化された残差が絶対値2を超える場合は、モデル適合性をさらに試験する必要がある。

適合からのこれらの偏差の形を見つける別の方法は、グラフ表示である。プロットは常に適合度検定を補足する。データポイントを含むプロットが、信頼限界など、それらのポイントの分散測定値を含む場合は非常に役立つ。

2.3.6. モデル適合性の改善

時には、使用可能なモデルすべてが特定データセットに妥当な適合性を与えないこともある。例えば、用量反応関係が単調である可能性が少ない場合、またはバックグラウンド反応のみを与える一部の低用量投与後、突然反応が上昇する場合、標準的な試験デザインに典型的なモデルは観察データと共に使用できない。このセクションでは、適合性の改善に使用するための検討事項、注意点および方法の一部を提供している。

使用可能なモデルのすべてがデータへの適切な適合性を与えない場合は常に、モデル作成者は初期試験で評価できなかったデータの質または実験の問題点をまず考慮すべきである（例、日和見性感染、投与過誤；セクション 2.1 を参照）。時には、データの調整（例、用量の対数変換または無関係な死の調整）が必要になることがある。一部のプラトーに達したか、単調でない反応パターンは、高い曝露量でより高頻度で見られる他の反応への進行またはその反応によるマスキングという点からの方がより良く理解でき、反応のより広範な定義を考える必要があることを示唆している。より複雑なモデル（例、反応時間を説明するモデル）を使用すると、使用可能なデータにより裏付けされることがある。または、簡単に入手可能なモデルによる用量反応関係をより容易に得られる、適切な用量単位を与える可能性がある、関連する PBPK データまたはモデル（例、最大の毒性物質、または他の複雑な薬物動態の代謝系または送達系の飽和状態を扱う）が存在すると思われる。

時として適合性の欠損は、モデル適合性プロセスの側面、例えば、非線形適合手順が「ベスト」な推定値に実際到達したか否か、またはすべての不均一なバラツキの影響が適切に考慮されてきたか否かによることもある。セクション 2.3.4 で詳細が記述されているように、パラメータが非線形であるモデルに適合させる場合、万一、推定プロセスがあまり代表的ではないパラメータ推定値セットに収束される場合に備えて異なる初期値を試みることは、常に良い慣習である（付録 A、例 A.3. も参照）。

不均一なバラツキは、BMR として使用する標準偏差の推定値など、連続モデルの適合に悪影響を及ぼす可能性がある。1つの方法は、付録 A の例 A.3 で示すように、累乗に増加した平均値に比例してバラツキをモデル化することである。この方法でバラツキをモデル化すると、変動係数が一定となり、バラツキが平均値の二乗に比例し、変動係数が比例定数の平方根となる状態に対応できる。加重最小二乗法、および一般化非線形最小二乗法などの他の方法を検討することができる。（セクション 2.3.4. を参照）

適合性の欠損が継続する場合、一つの選択肢は、用量反応関係を適切に説明するより、柔軟性のある経験的モデルを探すことである。この方法の見かけの利点は、データすべてを解析に組み入れることができる可能性があることである。この方法の危険性は、用量範囲の特定部分にデータを適合させようと試みると、より直接的に注目している用量範囲で用量反応曲線をゆがめる可能性があることである。多数の状況では、BMD が試験の最小用量に近く、そのためモデル作成者は BMD 領域でモデルの適合度を評価することができる（セクション 2.3.5. を参照）。

用量反応パターンには実現可能な生物学的説明があることもあるが（例、生存率または体重増加が、高用量群と低用量群とで顕著に異なる場合）、生体メカニズムに対応するモデルが利用できないことも多く、またそのようなモデルに適切に適合させる十分な追加データが得られない可能性もある。毒性物質への生物学的反応の機序に基づく理解がないと、BMR より極度に大きい反応を与える曝露データは実際には、BMR 領域内の反応の形について多くの情報を与えない可能性がある。しかし、そのような曝露は、最大用量が最大反応を示す場合のように、BMD 領域で適合したモデルの形に強く影響する可能性がある。

適用が低用量外挿に関与している場合、使用可能なモデルが適切な適合性を与えない時に考慮するアプローチは（適合度検定で $p < 0.10$ であるような、ある客観的基準に従って）、最大用量のデータを削除し、残るデータにモデルを再度適合させることである^{注 14}。しかしながら、これはすでに適切である適合を改善するために行うべきではない。適切な適合が得られるまで十分な投与群が存在すれば、最大用量でデータを削除するプロセスは繰り返すことができる。このプロセスを毒性学的試験データに適応させることは、通常これらの実験では投与群数が少ないため大きく限定される。実際には、自由度の欠損および適切なモデルの多様性の同時損失が起こる。連続エンドポイントに対しモデルを適合させる場合の高用量（および中用量）が下がると、結果としてバラツキのモデル化に対する情報が欠損する可能性があることにも注意されたい^{注 15}。

注 14 漸近項を推定するモデル（例、Hill モデル）の使用時に高用量データを低下させる場合は注意が必要である。例 A.3 を参照のこと。

注 15 バラツキモデルが適切にデータを説明しない場合、より柔軟性の高いモデルを検討する必要がある。この状況は、異常値の存在を示唆している可能性もある。

用量群を除外することは、注意深く取り組み、実行し、さらに詳細に提示することが必要である（セクション 2.4 も参照）。用量群を外すには、常に明確な判断を行う必要がある。このアプローチを含むモデル化の報告では、どの群を除外したかを明確に示し、結果がモデル化されたデータ範囲に限定されていることを言及するべきである。モデル作成者は、除外した用量を包含した結果と除いた結果を提示し、データ除外の影響を考察する必要がある。

付録 A の例 A.2 はある種のがんバイオアッセイデータに適用するように、高用量データを除いた図を示している。連続エンドポイントに対するこのアプローチの例を、付録 A の例 A.3 に示す。

2.3.7. モデルの比較

いくつかのモデルは与えられたデータセットに十分適合することが多い。現在、モデルの平均化方法が検討されており、その方法では、データが各モデルに示唆した裏付けを検討した適切に適合するモデルを収集し、そこから最終的に推定値を統合できる。これらの方法は、ベイジアンまたは他の方法により、各モデルから推定値を加重することによりリスク推定値の合成を可能にする（例、Kang ら 2000 ; Bailer ら 2005 ; Wheeler および Bailer 2007, 2008）。モデルの平均化は食品中の遺伝毒性発がん物質の症例研究で使用されてきた（Benford ら 2010）。しかし、このような方法は、モデル不確実性のリスク推定値への影響を説明するのに役立つ一方、適用は簡単ではなく、相違する結果がでる可能性があり、よって明確な EPA ガイダンスが必要となる。この時点で、リスクモデル作成者は、よく適合する実行可能なモデルを選択することが勧められる。モデル適合性の比較に使用するために、以下のガイダンスが提供されている。

適切に適合するモデルセットは、本質的にはお互い関連していないか（例えば、ロジスティックモデルとプロビットモデルは適合する二分データで同様に対応することが多い）、または、パラメータがある初期設定値で固定されているという点では異なるが、同じファミリーメンバーという点で互いに関連している可能性がある。例えば、対数ロジスティック、非ゼロバックグラウンドを持つ対数ロジスティック、閾値と非ゼロバックグラウンドを持つ対数ロジスティックすべてを、同じファミリーメンバーのモデルと考えることができる。適合度の統計値は異なるモデルの比較用にはデザインされていない。特に、1 つのモデルに対する適合度 p 値が高い場合は、よ

り低い p 値を有する別のモデルよりもうまく適合することを必ずしも意味しない。そのため BMD の算出に使用するモデルを選択する別の方法を追求する必要がある（詳細はセクション 2.3.5 を参照）。

用量反応モデルファミリー内で追加パラメータが導入されると、一般的に適合は改善するようにみえる。尤度比検定は、追加パラメータの推定により得た適合の改善が正当であるかを評価するために使用できる。しかし、このような検定は、統計的モデルを異なるファミリーと比較するためには適用できない（すなわち、対数正規分布 vs 正規分布）。一部の統計データとりわけ、赤池の情報量基準 (AIC) (Akaike 1973 ; Linhart および Zucchini 1986 ; Stone 1998 ; AIC は $-2L+2p$ 、ここで L は推定パラメータ p に対する最大尤度推定値 [MLE] における対数尤度) は、同様な適合方法でモデルを異なるファミリーと比較するために使用できる（例、最小二乗法または二項最大尤度法）。そのような方法は正確ではないが、モデルの選択に有用なガイダンスを与えうる。

同様のエンドポイントで他のデータセットが存在する場合、外部の検討事項が適用できる。仮に、データすべてが同じモデル様式で適合し、モデルがデータセットすべてを説明すると認められると予測した場合、試験全体の BMD 算出結果を比較することが可能になることもある。別の検討事項は、特殊なデータを適合させる従来のアプローチの有無である。これらの検討事項のいずれも、適合しないモデルを使用することの正当性として理解すべきではない。最後に、可能なら、よりパラメータの少ないモデルを使用することが好ましいと考えられることも多い。

付録 A に、モデル適合およびモデル比較の問題を探求する多数の例（例 A.1-A.5）を示す。

2.3.8. BMDL 取得のための信頼限界値の計算

得られた反応に関連する用量（すなわち BMD）の信頼限界値は、殆どの統計学的ソフトウェアパッケージによって与えられない。そのような結果を得るには、統計的問題を正しく組み立て、特別なプログラムを作成する必要がある。そのようなプログラムは、正しい結果を生み出すことを立証するために検定される必要がある。そのため、EPA の BMDS パッケージのような十分に裏付けされた方法論があるソフトウェアが使用され、特別に作成されたプログラムは十分に裏付けされていることが望ましい。このセクションは、BMD に対する信頼区間を推定するための、好ましい計算アルゴリズムについて概説する。

信頼区間は、サンプリングおよび／または測定誤差の両方または一方による、パラメータ推定値の不確実性を表している。「信頼」の定量化は、解析中のデータを生み出した実験を無限に反復するという概念的実験を実施することに由来する。信頼区間に関連する「信頼」または「範囲」は、推定されているパラメータ、例えば BMD を含むこれらの反復区間の一部である。この概念的実験の結果は、一般的に信頼限界を算出するアルゴリズムに変換され、統計学的理論を使用し既定の範囲レベルから区間を計算する。信頼水準の選択は、データ収集の費用と必要なデータの精度と、どちらをとるかである。0.05 は有意性検定における従来のカットオフレベルであると同時に（必ずしもすべてのデータに適しているわけではない）、95%は殆どの限界値に使いやすい選択であり、本ガイダンスで推奨する初期設定値である。信頼区間の両端は信頼限界と呼ばれる。信頼限界値は、特定のモデルファミリー内でデータと一致するこれらの値をまとめるが、モデル化された動物データと関心のあるヒト集団データとの一致を説明あるいは仮定しない。稀だが重要な例外として、計算された信頼区間は、その区間の実際の範囲が通常、望ましいレベルから少しずれているという意味で、概算となる。

信頼区間 (CI) は両側にあり得るが、対応するパラメータ値の境界を両側で示すか、または対応するパラメー

タ値の境界を片側だけで示すことができる。両側区間は、通常一般的な科学的用途にみられ、推定値の全不確実性を特徴付ける必要がある場合に適切である。片側区間も、不確実性を特徴付けるが、焦点は特定の方向に当てられる。例えば、片側区間を使用して、**BMD** の真の値が指定値以上であることの確認に役立てる。片側信頼区間の信頼限界値を算出する 1 つの方法は、他方の限界が無限大または負の無限大となる両側区間の片側限界として算出するものである。例えば、1 つのパラメータに対する片側 95%信頼区間が、そのパラメータの両側 90%信頼区間と限界値を共有するかもしれないし、2 番目の限界値としてプラスまたはマイナス無限大（恐らく、負の値をとらない **BMD** のようなパラメータではゼロ）を有することもある。

BMR を超えない高い信頼性（例、95%）を確保する用量（**BMDL**）を得るために、低い信頼限界値が **BMD** に設定される。このプロセスにより、より正確な **BMD** 推定値を与える、より良い実験デザインと手順が得られ、その結果、信頼区間が狭まり、そのため **BMDL** が高くなる。**BMDL** または **BMCL** を計算する方法と例の一部が Gaylor ら（1998）によって示されている。

信頼区間を得る方法は、通常 **BMD** がモデルから推定される方法に関連している。最大尤度法によってパラメータを推定する場合、信頼区間は尤度比（プロフィール尤度）の漸近的分布に基づくか、最大尤度推定値の漸近的分布に基づく可能性がある。漸近理論を使用する必要があるという仮定が弱くなり始める時（例、サンプルサイズの減少）、両方の方法が問題を与える可能性があるとはいえ、良好な適用範囲の挙動を与える傾向のために、最大尤度法によって推定したパラメータの信頼区間を、最大比率の漸近的分布に基づいて算出することは通常好ましい（Crump と Howe 1985）。

尤度比の分布に基づいてモデルパラメータに対する片側 $100 \times (1 - \alpha)\%$ 信頼区間を算出するには、まずそのモデルのパラメータすべての最大尤度推定値を算出する。次に、その信頼区間が算出されているモデルパラメータ（ μ と呼ぶ）を他のパラメータと分ける。続いて、尤度を最大にするために他のパラメータを補正する場合、対数尤度が正確に $\chi^2_{(1, 1-2\alpha)/2}$ 分だけ最大尤度推定値の対数尤度から減少するような、 μ 値を見つける。ここで、 $\chi^2_{(1, 1-2\alpha)/2}$ は自由度 1 で上限確率 2α に相当する χ^2 分布の変位値を表す（例えば、Crump および Howe 1985、Venzon および Moolgavkar 1988 を参照）。注目する値がモデルパラメータとして表現できない場合、同様なしかしより複雑な方法が使用される。

特定なデータタイプに特異的な信頼区間算出への他の方法も使用可能である。例えば非連続データに対する別の方法は、標準的な統計理論（具体的にはデルタ法、例、Gart ら 1986）を適用し、推定した **BMD** のバラツキを概算することである。このバラツキの推定値は次に、**BMD** のより低い信頼限界を構築するための基礎として使用できる。用量の対数は正の **BMDL** を確保するために使用できる。クラスタ化データ、例として生殖および発生作用では、このデータタイプに対しセクション 2.3.4 で説明したモデル化方法が適切な信頼区間算出方法に直接つながる。複数の結果では、生殖プロセスの多数の異なる段階で影響を示す、発生および生殖毒性データで見られるように、用量反応モデルの開発と信頼限界の算出に対処する多数の方法が利用可能である（Chen ら 1991 ; Ryan ら 1991 ; Catalano および Ryan 1992 ; Ryan 1992 b ; Catalano ら 1993 ; Zhu ら 1994 ; Krewski および Zhu 1994, 1995）。

よって、**BMDL** は以下のような手順で決定される。

1) エンドポイントの選択

- 2) BMR の同定（コントロールと比べた反応の変化の既定レベル）
- 3) 適切な推定手順による、データを十分に適合させるモデルの確立
- 4) アプリケーションに応じた、片側または両側信頼限界および信頼水準（例、95%）の特定
- 5) モデルおよび BMD に対する同一推定手順を使用した、選択した BMR での信頼限界値の計算

2.3.9 POD 算出に使用するモデルの選択

以下のアプローチは、特定データセットの POD となるような BMDL を算出するために使用するモデルを選択するのに推奨される。すでに述べたように、これらの決定の一部は、統計的手順およびこのタイプの解析における潜在的な落とし穴について熟知した専門家が行うか、このような専門家と共同作業することにより、最も良好に実施される。

- 1) $\alpha=0.1$ の値を使って適合度を評価し、臨界値（または $\alpha=0.05$ または一連のモデルに適合するより特定の 1 つのモデルを使用する理由がある場合 $\alpha=0.01$ ；セクション 2.3.5 を参照）を決定する。
- 2) モデルの残差とグラフおよびデータを調査しながら、用量反応関係の関連する低用量部分を明らかに適切に説明しないモデルをさらに破棄する（セクション 2.3.5. を参照）。
- 3) 残りのモデルが、推奨される適切性に対する初期設定の統計学的基準に適合し、視覚的にデータに適合したとき、それらのいずれのモデルも BMDL の決定に使用できる可能性がある。BMDL を選択する残りの基準は必然的に幾分任意となり、初期設定値として示される。
- 4) 残りのモデルによる BMDL 推定値が、個々のモデルの特定の影響を反映せず、十分に近ければ（評価の必要性を前提とする）、最小の AIC を有するモデルを POD に対する BMDL の算出に使用することができる。この基準は、客観的で再現性のある方法で単一 BMDL 値への到達に役立てることを目的としている。2 つ以上のモデルが最小の AIC を共有する場合、最小の AIC と共に BMDL の単純平均または幾何的平均値が使用されることもある。これは、適切な適合モデルの全セットの加重を含む「モデルの平均化」と同一ではない（セクション 2.3.7. を参照）。さらに、そのような平均は（平均 BMD で）95% 下限値ではないという事実も含め、欠点がある。つまり、平均は単に考慮中の特定 BMDL の平均である（すなわち、平均は個々の推定値の統計学的性質を失っている）。
- 5) 残りのモデルによる BMDL 推定値が十分に近くない場合、その推定値がある程度モデルに依存していると推定できる。専門家による統計学的判断は、結果の一部またはすべてを信頼するにはモデルの不確実性が大きすぎるか否かを、その時点で判断することに役立つ可能性がある。結果の範囲が妥当であると判断される場合、結果の中から選択する残りの明確な生物学的または統計学的根拠は存在せず、妥当で控えめの推定値として最小の BMDL が選択される可能性がある。別の解析と考察には、追加モデルの考慮、使用するモデルのパラメータ値の調査、または同一パターンが BMDL について存在するか否かを決定するための BMD の評価が含まれるかもしれない。判断手順を検討する機会は常に与えなければならない。
- 6) 一部の事例では、モデル化の試みが有用な結果を与えないこともある。有用な結果を与えず、生物学的に最も関連のある影響が、適切だがモデル化の影響を受けないと考えられる試験に由来する場合、NOAEL（または LOAEL）が POD として使用される可能性があるかもしれない。持ち上がったモデル化の問題点は、関連データの全限界が代わりの NOAEL/LOAEL 法の結果に及ぼす影響と共に、評価時に検討すべきである。

2.4. 推奨事項の報告

セクション 2 を通じて検討したように、選択したアプローチと値を裏付けるための選択肢について、正当な理由を徹底的に提示すべきである。BMD または BMDL のいかなる算出に対しても、以下の要素を含めることが推

奨される。

- 1) BMD を算出するために選択された試験
 - a) 試験選択の論理的根拠
 - b) エンドポイント（作用）選択の論理的根拠
 - c) 使用した用量反応データのリスト
- 2) 各事例で選択した用量反応モデル
 - a) 論理的根拠
 - b) 推定手順（例、最大尤度、最小二乗法、一般化推定方程式）
 - c) モデルパラメータの推定値
 - d) 適合度（例、カイ二乗統計量）、対数尤度、および AIC
 - e) 標準化された残差（観察値－予測反応）／標準誤差
- 3) 各事例での BMR の選択
 - a) 論理的根拠
 - b) 連続データの場合の使用手順
- 4) 各事例での BMD の算出
- 5) 各事例での BMD の信頼下限（すなわち、BMDL）の算出
 - a) 信頼限界手順（例、尤度プロフィール、デルタ法、ブートストラップ）
 - b) BMDL 値
- 6) 各事例の図
 - a) データポイントおよび誤差（SD）バーと適合した用量反応曲線のプロット
 - b) 適合した曲線に対する信頼限界値のプロット（任意に含める場合は、演算に使用した方法を、順を追って説明する）
 - c) BMD と BMDL の同定
- 7) 標準化 BMR に対する BMD および BMDL（比較用）
 - a) 二分データでは、超過リスク 0.10 に対する BMD と BMDL
 - b) 連続データでは、対照群の平均値から 1 SD に等しい平均反応値の変化に対応する BMD および BMDL
- 8) 推定の適用と実行可能性によって、BMDU（BMD の信頼上限）

2.5. ディシジョンツリー

図 4 に示すディシジョンツリーは、データ初期評価の終了後、BMD/BMDL 計算における全般的なステップ進行の概要を示している（セクション 2.1 および図 2A 参照）。別に実施した BMD の計算が、最終の定量的リスク推定値に対する妥当な候補である各エンドポイント／試験の組み合わせを裏付ける。エンドポイントまたは試験全体を通じて NOAEL または LOAEL を比較する場合と異なり、可能な BMD の相対値は、モデル化が完了するまですぐには明白にならない。ディシジョンツリーのステップの一部の詳細を以下で検討する。

各エンドポイント／試験の組み合わせの候補に対して以下を行う。

- 1) データのタイプ（すなわち、非連続または連続）、試験デザインの感度、毒性エンドポイント、および連続データならば特定レベルのエンドポイント変化に対する難度判断に基づき BMR を選択する（セクション 2.2）。
- 2) データのタイプに特定のモデル構造（すなわち、非連続または連続、BMR をどのように設定するかによる）

と試験デザイン（例、入れ子構造、セクション 2.3.3）特異的なモデル構造を使用して、用量反応データをモデル化する。がんバイオアッセイデータのモデル化では、代替モデルが優れている可能性のある事例に特異的な状況（例、腫瘍発生までの期間のモデル、生物学的モデル）を除いて、特定の初期設定アルゴリズムが一般的に使われる。実験動物データの他のタイプでは、様々なモデルによる曲線適合を試みることができる。ヒトデータは症例に特異的な方法でモデル化され、死因に競合するような、共変量を捕らえる必要がある（付録 A、例 A.6 を参照）。

3) モデルの適合性を評価する（セクション 2.3.4～2.3.7）。p 値 0.1 で棄却されないモデルを保持する（事前にモデルが選択されている場合は除く、セクション 2.3.5 を参照）。残差を分析してデータとモデルをプロットする。つまり、モデルが特に BMR 領域内で、適切にデータを説明していることをチェックする。時に、なんらかの方法でデータを変換するか、または良好な適合性を得るためにさらに統計学的评价を実施することが必要な場合もある（セクション 2.3.6.を参照）。

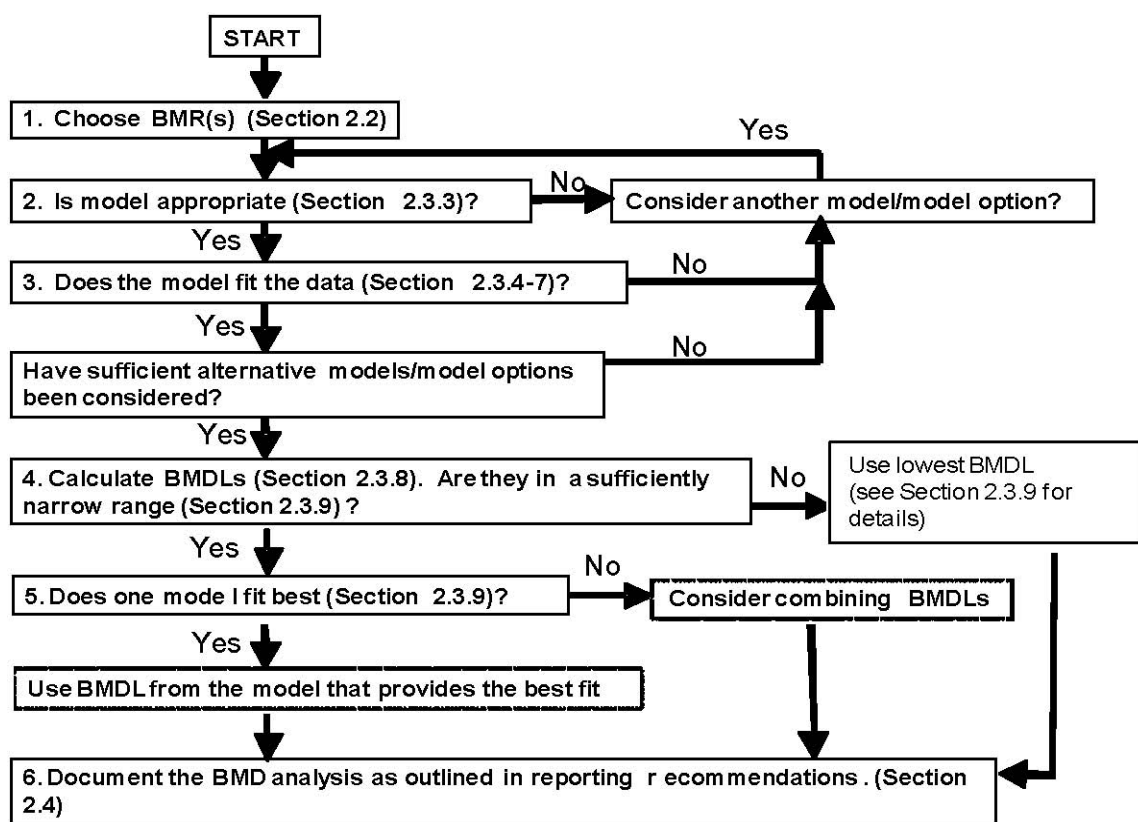


図 4: BMD のディシジョンツリー

- 4) データに十分に適合するモデルを使って、BMD 候補値の 95%信頼下限（すなわち、BMDL）を算出する（セクション 2.3.8 を参照）。
- 5) データに適切に適合するモデルから選択する（セクション 2.3.9）。これらの残りのモデルによる BMDL 値が十分に近い場合（評価の必要性があれば）、最小の AIC を有するモデルが BMDL を与えるために選択される可能性がある。BMDL 値が十分に近くない場合、モデル依存性がある程度想定され、科学的指針による判断をする必要もある。
- 6) 推奨事項の報告に関するセクション 2.4 で概要を示したように BMD 解析を文書化する。

付録 A. 例

以下の例は、単一のデータセットと単一のエンドポイントについて BMD と BMDL を算出する重要な特徴を説明するために選択した。以下の例に示す計算は EPA の BMDS パッケージを用いて作成したが、プリントアウトは示していない。すでに述べたとおり、本文書では特定のソフトウェアパッケージの使用方法を超える包括的問題を取り上げている。BMR の選択、どのエンドポイントとデータセットをモデル化するか、どのモデルを検討するか、などの他の決定事項は以下の例の範囲を超えている。

A.1 非連続データのモデル化

この例では、セクション 2.3 で考察した、様々なモデルを当てはめ、適合度を評価し、BMDL を選択して POD に利用するプロセスについて説明している。この例では、臨界データセットと BMR 値がすでに選択されていると仮定する。

Table A.1.1. 非連続データ

Dose	Number Affected	Fraction Affected	Number of Animals
0	1	0.02	50
8	6	0.12	50
21	15	0.31	49
60	20	0.44	45

片側 95%信頼区間を用いて超過リスク 0.1 について BMD と BMDL を計算する。超過リスク 0.1 (=BMR) に対応するように BMD を定義した場合、 $P(\text{BMD})$ が BMD で影響を受けた動物の割合であり、 $P(0)$ が対照群の割合であれば、BMR は以下のように定義される。

$$BMR = \frac{P(\text{BMD}) - P(0)}{1 - P(0)}$$

これは以下のように置き換えられる。

$$P(\text{BMD}) = P(0) + [1 - P(0)]BMR$$

BMR=0.10 について検討しているため、これは $0.02 + (0.98 \times 0.1) = 0.118$ の反応に相当する。検討した動物の 31% が最小非対照用量で影響を受けた。そのため、BMD での予測反応は観察された最低反応よりも大幅に低い。このため、モデルの選択が BMD の計算に何らかの影響を及ぼすことを知っておく必要がある。

A.1.1. 適合モデルの選択（セクション 2.3.3）

この例では、特定の 1 つのモデルを適用する（例えば、機構的な）理由はないため、表 A.1.2 に示すとおり、多数のモデルをデータに適合させている。モデルが異なると曲線の形状および低用量での挙動は様々となるため、

どの形態のモデルがデータを十分説明すると予想されるか不確実であることを考えるとこれは筋が通っている。このモデルは、生物学的プロセスと一致すると広義に考えられる制約を用いて適合した（多段階モデルの係数 ≥ 0 、ガンマおよびワイブルモデルの出力係数 ≥ 1 、対数ロジスティックモデルの勾配係数 ≥ 1 。セクション 2.3.3.3 および付録 C を参照）。推定時、ガンマおよびワイブルモデルの出力係数の設定は 1（制限値）とし、対数ロジスティックモデルの勾配パラメータの設定は 1（制限値）とし、高次多段階パラメータ β_2 および β_3 の設定は 0（標準的制限値の下限）とした。結果として、すべての適合モデルで 2 つのパラメータの推定に自由度 2 が必要であった。

ワイブルモデルであり、指数がそれぞれ正確に 1 または 2 と指定された非連続直線モデルまたは非連続二次モデルは適合させなかった。この場合、指数パラメータを明記する根拠はなかった。より少数のパラメータを要求してもパラメータ値の指定は正当化されず、指定したパラメータには十分な科学的根拠があり、通常はそのような根拠はないことに注意して欲しい。また、一次多段階モデル（非連続線形モデルにも相当）の利用は、高次多段階係数をゼロに指定することと同じであることにも注意して欲しい。

A.1.2. 適合度の評価（セクション 2.3.5）

表 A.1.2 は、AIC の昇順に並び替えたモデルの適合結果を示している。[AIC は $-2 \times (\text{LL} - p)$ であり、LL は MLE の対数尤度、 p は推定パラメータ数であることを思い出していただきたい。他はすべて等しく、AIC 値は低い方が好ましい。]

5 つのモデルで χ^2 値が推奨カットオフ p 値の 0.1 を超える。2 つのモデルで $p < 0.10$ であり、（同時にではないが）少なくとも 1 つ多少大きなスケールされた残差があり、少なくとも 1 つのデータポイント（この場合は中用量）に適合していないことを示している。

Table A.1.2. モデル適合度の統計値

モデル	χ^2	χ^2 の p -値	AIC	スケールされた残差			
				用量 0	8	21	60
対数ロジスティック ^a	0.96	0.618	173.6	-0.049	-0.151	0.772	-0.584
ガンマ ^b	2.29	0.318	174.9	-0.304	0.182	1.186	-0.872
一次多段階 ^c	2.29	0.318	174.9	-0.304	0.182	1.186	-0.872
ワイブル ^b	2.29	0.318	174.9	-0.304	0.182	1.186	-0.872
対数プロビット	0.58	0.445	175.3	-0.035	-0.383	0.600	-0.273
プロビット	7.82	0.020	181.1	-1.688	0.012	2.125	-0.672
ロジスティック	8.39	0.015	181.9	-1.801	-0.066	2.183	-0.613

^a 勾配パラメータを ≥ 1 に制限。

^b 出力パラメータを ≥ 1 に制限。

^c 負以外のパラメータの標準的な制限を考慮し、これらのデータに適合する唯一の多段階モデル。

A.1.3. モデルの比較（セクション 2.3.7）

最小 AIC のモデルは対数ロジスティックモデルである。このモデルについては、予測曲線は各データポイントの信頼限界内に十分入るため、スケーリングされた残差 [すなわち、(実測値-予測値) / SE は小さい (± 2 単位以内。セクション 2.3.5 および 2.5 を参照)] と目視検査でこのモデルの選択が裏付けられる (表 A.1.2 および 図 A.1.1)。

次に、表 A.1.2 の次の 3 つのモデル、つまり多段階、ガンマ、ワイブルは、(スケーリングされた残差が同じであることから示されるとおり) すべて正確に同じ適合度と BMD 予測値が得られることに注意して欲しい。実際、これらのデータについては、事実同じモデルであり、多段階パラメータは正値に制限されるため、 β_2 と β_3 はゼロに設定され、出力パラメータは制限値の 1 に設定された。そのため、3 つのモデルはすべて $0.0310 + [1 - 0.0310] \times [1 - \exp(-0.110 \text{ 用量 } 1)]$ と推定された。AIC 174.9 は対数ロジスティックモデルの AIC 173.6 よりもごくわずかに悪かった。図 A.1.2 は、ワイブルに適合させたものを示し、3 点すべてが適合していることを示している。すべての用量の適合は対数ロジスティックよりもわずかに悪く、グラフを詳細に検討すると明らかである。リスク評価の応用では、通常最も関心がある部分は低用量範囲の適合であることを思い出していただきたい (セクション 2.3.5)。

図 A.1.3 に示した対数プロビットモデルもデータに十分適合し、スケーリングされた残差が小さく、予測される曲線が十分各データポイントの信頼限界内にある。このモデルでは、すでに考察したモデルよりも低用量での適合度がわずかに低い。最後に、表の最後の 2 つのモデルでは $p < 0.10$ であり、これらはさらなる検討から除外できる。

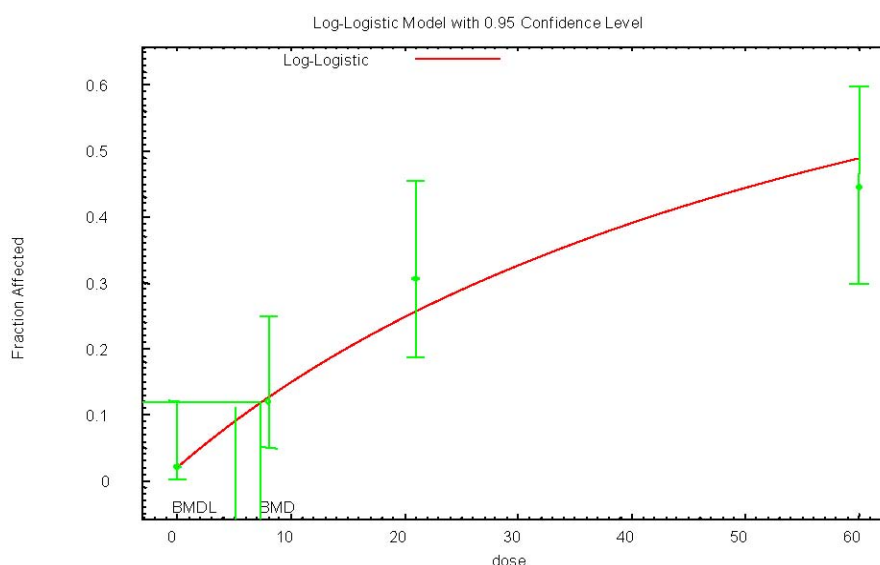


図 A.1.1. 対数ロジスティックモデルの適合。誤差バーは個々の反応平均値の 95%信頼限界を示す。

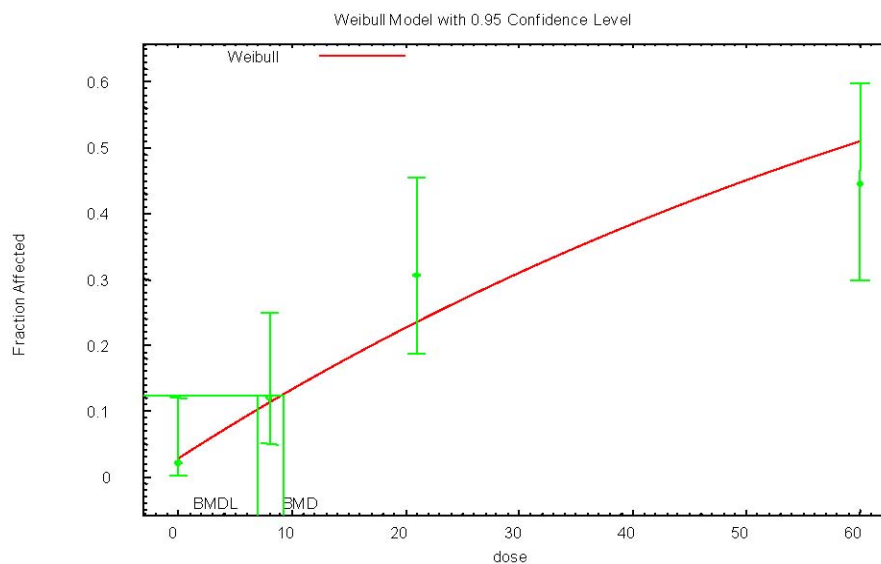


図 A.1.2. ワイブルモデルの適合誤差バーは個々の反応平均値の 95%信頼限界を示す。

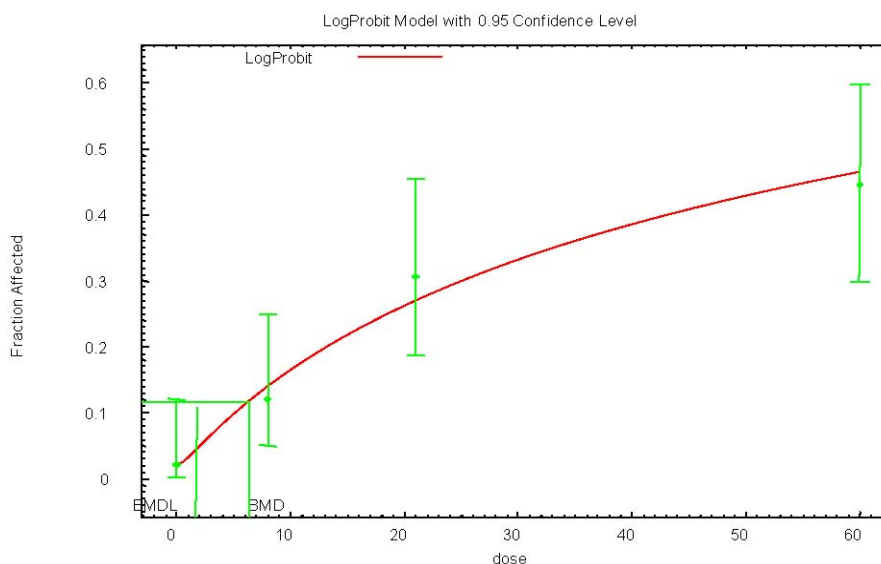


図 A.1.3. 対数プロビットモデルの適合誤差バーは個々の反応平均値の 95%信頼限界を示す。

A.1.4. POD の基礎として使用するモデルの選択（セクション 2.3.9）

我々の適合度評価では、表 A.1.3 に示した 3 つのモデルが残った。

Table A.1.3. 適合度評価後に許容されたモデル

モデル	χ^2	p -値	AIC	BMD	BMDL
対数ロジスティック	0.96	0.618	173.6	7.3	5.2
多段階、ガンマ、ワイブル	2.29	0.318	174.9	9.2	6.9
対数プロビット	0.58	0.445	175.3	6.4	1.7

3つの許容できるモデルのうち、BMD と BMDL の基礎として利用すべきモデルはどれか。この場合、BMDL は約 4 倍の範囲内で 1.7～5.2 である。適用の必要性によっては、BMDL は十分に近いとは考えられない。リスクアセスメントの目的で、例えば、この範囲は BMDL がもっとも低いモデルが妥当で控えめな推定として望ましいと考えるのに十分な広さである。

A.2. 非連続データ：用量群の除外（セクション 2.3.6 を参照）

セクション 2.3.6 で考察したとおり、用量反応評価において、特に目標の反応領域で十分にモデルを適合させるため、用量群を除外することを検討可能な場合がある。BMR が最低用量に近いまたはそれ未満の場合、最高用量のデータを除く理論的根拠は、最高用量のデータが目標の低用量領域、すなわち BMR で付近の情報が最も少ないと考えられることである。これは、二分および連続データの両方に当てはまる。以下の例では、二分データセットを使用し、低用量と高用量モデルの適合度を比較するため、スケーリングされた残差の評価を含め、用量群を除外するか否かを検討する際に従うべき、いくつかの基本的な原理を説明する。次の例では、より大きなデータセットを使用し、連続データの平均および分散情報のモデル化に関するこのような検討事項を示している。

以下のデータセットは、長期的バイオアッセイの 1 用量群あたりの典型的な被験者数を利用しているが、通常の用量群の数よりも多い長期的がんバイオアッセイから得られる腫瘍反応データの一例である。

Table A.2.1. 非連続データ

用量 (ppm)	影響を受けた動物数	影響を受けた割合	試験動物数
0	0	0.00	50
50	1	0.02	50
100	10	0.20	50
150	35	0.7	50
250	30	0.75	40

データの初期検査で分かるとおり、影響を受けた割合に関する反応は最高用量でプラトーに達しているように見える。この所見には生物学的理由があると考えられ（例えば、主要酵素が飽和となった）、または目標のエンドポイントはより重大な効果または他の原因による初期の死亡率により最高用量でマスクすることができ（例えば、腫瘍発生とは関係のない急性毒性）、目標エンドポイントのリスクで有効となる数が減少する。前者の例では、1つの選択肢で漸近線の項を含むモデルの使用を検討することになり、100%の反応レベルに達する前にプラトーとなる可能性のある反応を考慮に入れる。しかし、このタイプのモデルが利用できない場合、または（例えば、最高用量群で死亡率が高いために）1つの用量で報告された反応が疑わしい理由がある場合、用量群の除外は許容できる代替アプローチと考えられる。

これまでの例において、我々は臨界データセットと BMR はすでに選択されたものと仮定している。この例の目的でも、腫瘍発生までの時間のモデル適合を容易にする、個々の動物に関する適切な生物学的モデル、または死亡時間データが利用できないことが確定した後、モデル、つまり多段階モデルを推測的に選択することを想定している。まず、我々はデータをすべて適合させる試みを行う。セクション 2.3.7 で考察し、これまでの例で

説明したモデルを使用し、様々な多段階モデルの選択肢を評価した後、表 A.2.1.のデータに最も適合するモデルとして、3 次多段階モデルを選択した。このモデルの適合および適合度の p 値のプロットを図 A.2.1 に示す。

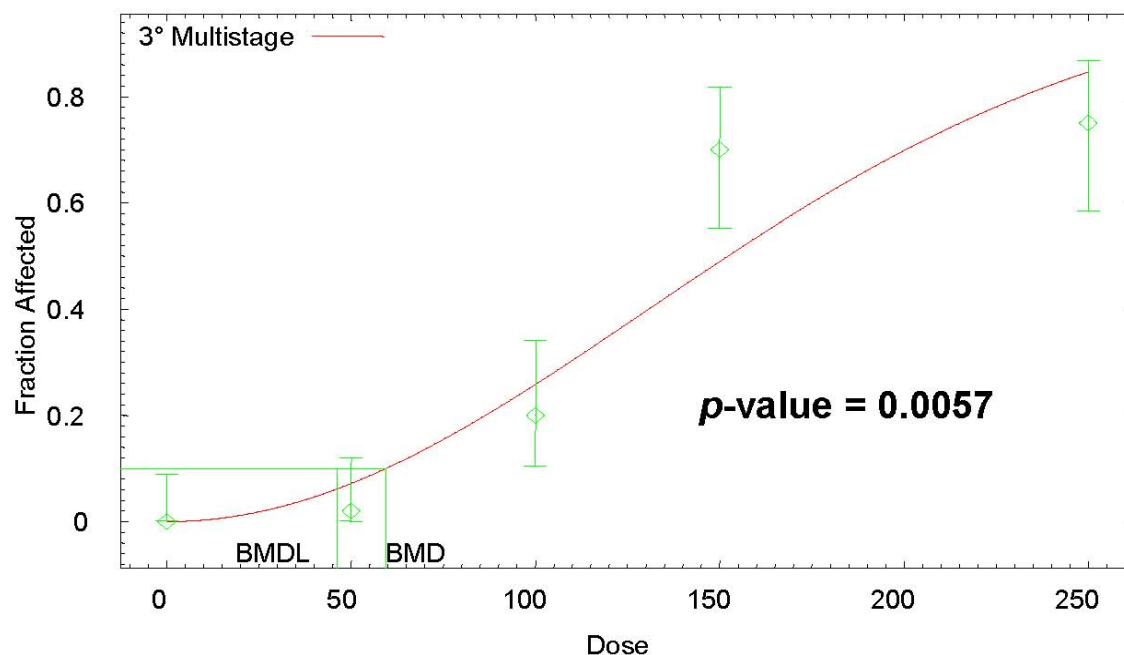


図 A.2.1. 最高用量群を含めた 3 次多段階モデルの適合誤差バーは 95%信頼限界を示している。

このモデルの適合は、セクション 2.3.5 で説明した適合度の基準を満たしていない（すなわち、 p 値は、推測的に選択したモデルの慣習的な p 値の 0.05 以下である）。リスク評価者は、この時点で推測的に選択したモデルを適合する際にいくつかの選択肢があり、通常は以下の順である。(1) 用量を除外する生物学的根拠がある場合（例えば、最高用量群で死亡率が高いなど）、用量を除外し、選択したモデルを再度適合させる、(2) 別のモデルを試す、(3) 別の同等なデータセットを選択する、または (4) 用量群の一つを除外し、選択したモデルを再度適合させる。最高用量群は、このように、通常、BMR がデータの下端にある場合に除外される用量群である（セクション 2.3.6）。しかし、実験誤差などの他の要因が最高用量群以外の用量群除外の妥当性を示すこともある。この例では、最高用量群を除く理由は、高い死亡率が低用量での用量反応の適合とは関連していないというものであり、多段階モデルを残りの 4 群に再度適合させている。最高用量なしでは、3 次多段階モデルが表 A.2.1 のデータに最もよく適合すると判断された（図 A.2.2）。

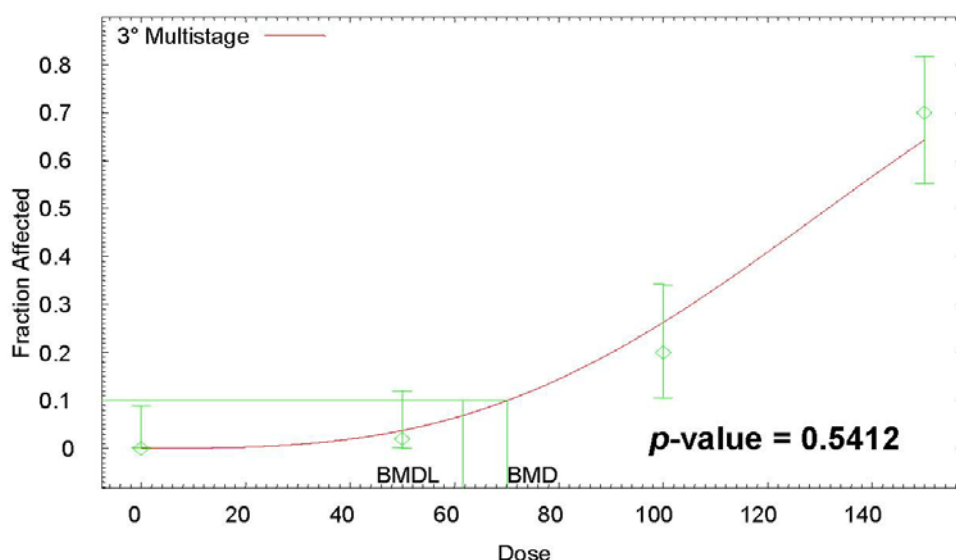


図 A.2.2. 最高用量群を含めない 3 次多段階モデルの適合誤差バーは 95%信頼限界を示している。

表 A.2.2 は、3 次多段階モデルの適合について、すべての用量群での BMD、BMDL、p 値、スケーリングされた残差を示している。5 つの用量群すべてで用量反応データを十分に適合させることができない (p 値<0.05)、最高用量群を除外した場合に用量反応データへの適合が許容できる (p 値>0.05)、プロットの目視検査 (図 A.2.1 対図 A.2.2)、BMR 付近のスケーリングされた残差の比較 (すなわち、推定 BMD 値に最も近い用量群について、-1.421 対-0.650) など、この分析のいくつかの側面は、最高用量群の除外を支持している。一般に、BMD 付近の用量群でスケーリングされた残差が低くなるモデルが好ましい。(セクション 2.3.5 および 2.5 を参照。) 最高用量群を除いたモデルの適合は、モデルの範囲内の用量で用量反応関係の特徴付けるために適しているに過ぎない (セクション 2.3.6)。

Table A.2.2 最高用量群有無による 3 次多段階モデルの比較

Fit statistics	3 rd -Order with high dose	3 rd -Order without high dose
p-Value	0.0057	0.5412
BMD	59.4	70.1
BMDL	46.2	61.6
Scaled residual at 0 ppm	0.000	0.000
Scaled residual at 50 ppm	-1.421	-0.650
Scaled residual at 100 ppm	-0.939	-1.014
Scaled residual at 150 ppm	2.981	0.839
Scaled residual at 200 ppm	-1.667	--

A.3. 連続データ：よく適合したモデルを得る

この例は、連続モデルの適合に関するいくつかの検討事項を説明しており、分散のモデル化など、(パラメータが) 非線形モデリングソフトウェアを使用する場合に必要な注意点、用量群の除外など、BMD および BMDL を計算するために十分なモデル適合度を得るために必要な一部のデータ操作などについて説明している。さらに、本文書では 2 つ以上の技術ポイントについて説明する。最初に、パラメータが非線形であるモデルの収束は MLE が達成されたことを保証しない。MLE を得るためには、常識的な認識と再適合が必要になることもある。第二に、一度 MLE が達成されると、モデルは十分に適合しないこともあり、モデルの適合を改善するために他の措

置が必要になることもある^{注16}。

注16 この例の挙動の一部は、BMDS の Hill モデル 2000 年 4 月 3 日版でその初期値をどのように選択するかに依存する。他のソフトウェアは、BMDS の Hill モデルの後期版であっても、これらのデータで異なる挙動を示す可能性がある。これはソフトウェアの「バグ」を示しているのではなく、むしろデータベースによっては、尤度関数に複数の「極大」がある可能性がある。(BMDS 同様) 最適化に純粋に個別の方法を利用しているソフトウェアは、極大でトラップされる可能性があり、尤度関数の真の全体の最大値に収束させるため、代わりの初期パラメータ値を用いて実験することが必要と考えられる。ソフトウェアパッケージは、最適化用に開始時のパラメータ値を選択するために使用するアルゴリズムが異なるため、極大値が異なる可能性がある。

表 A.3.1 のデータは、投与後のラットの生化学的反応を示している。本文書では比較の目的で推奨しているため、この例では、反応の平均が 1 つの対照 SD で示される用量として BMD を計算する。図 A.3.1 に見られるとおり、用量反応データはプラトーを示している。そのため、Hill モデル (BMDS で利用可能) を当てはめることは妥当であるように思われる。付録 C の指数モデルなど、他のモデルも検討することができるが、この例の目的では Hill モデルを使用する。

Table A.3.1. 連続データ

用量	動物数/群	平均値	SD
0.	8	100.	30.4
0.3	8	98.24	49.8
1.	8	111.34	59.9
3.	8	172.16	58.4
10.	8	357.48	167.5
30.	8	1695.03	260.9
100.	8	1576.11	169.7
300.	8	1896.22	141.7

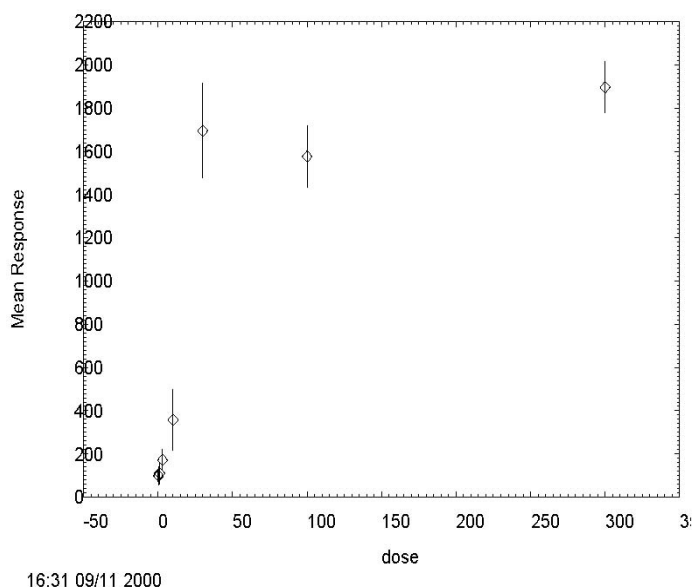


図 A.3.1. 例のデータの平均と 95%信頼区間

このデータセットの分散は平均が増加するにつれて増加する傾向があり、生化学的データでは分散が平均の二乗に（おおよそ）比例することが多いという所見に適合する。この例では、分散が平均の検出力に比例するものとしてモデル化されているデータにモデルを当てはめている。つまり、我々のモデルは以下の通りであり、

$$\mu(d)=\gamma+VX^n/(k^n+d^n)$$

$$\sigma^2(d)=\alpha(\mu(d))^\rho$$

式中、 d は用量、 $\mu(d)$ は反応の平均、 $\sigma^2(d)$ は用量 d の観測値の分散を表す。残りのパラメータのうち、 γ はバックグラウンドの反応（または切片）であり、 V は最大反応、 k は半分の反応が起こった用量、 α は比例定数、 n および ρ はモデル化過程で決定される指数部である。

このモデルのパラメータの概算はデータのグラフから読み取ることができ、アルゴリズムを当てはめる際の確認に有用である。例のデータに Hill モデルを当てはめる際、 γ （切片）は反応のほぼバックグラウンドレベルであるため、約 100 反応単位と予測し、 V はバックグラウンドレベルでの最高用量のおよその増分であるため、約 1,600 反応単位とし、 k は 10～30 用量単位の範囲とする。さらに、経験から、 n は比較的小さく、例えば 1～10 であり、上述のとおり、そのようなデータでは分散が平均の二乗に比例することが多いため、 ρ は 1～2 の間に入るはずである。

BMDs の Hill モデルの 2000 年 4 月 3 日版を用いると、アルゴリズムを適合することで明らかに解に収束する。この解から推定されるパラメータは以下の通りである。

Table A.3.2. Hill モデルのパラメータ予測値と標準誤差

変数	予測値	SE
α	4381.57	2211.67
ρ	0.266572	0.0668979
intercept	105.045	22.8759
V	1634.05	51.087
n	4.76591	1.62145
k	14.256	1.80324

対数尤度=-345.786

ρ の値が 1～2 の範囲であると予測していたのに対し、0.27 であったことを除き、すべての予測値は予測される範囲に入ったと言える。

このモデルを当てはめた結果の予測値を表 A.3.3 に示す。このモデルは高用量の平均値と SD をうまく予測しているが、低用量の SD は 2～4 倍過剰に推定される。過剰に推定された SD によって 1 SD の BMR が決定されるため、 BMD_{1SD} はデータから示されるよりもやや高く予測され、 $BMDL_{1SD}$ も誤って特定される可能性がある。低用量での適合を改善するため、高用量を除くことを考えることもできるが、まずは全体のデータセットで別の

点の説明を続ける。

Table A.3.3. Hill モデルによる予測平均と標準誤差

用量	データ数	観測値 平均	観測値 SD	予測値 平均	予測値 SD	スケール後 残差
0	8	100	30.4	105	123	-0.115
0.3	8	98.2	49.8	105	123	-0.156
1	8	111	59.9	105	123	0.138
3	8	172	58.4	106	123	1.518
10	8	357	168	360	145	-0.059
30	8	1700	261	1690	178	0.028
100	8	1580	170	1580	179	-2.570
300	8	1900	142	1740	179	2.483

これがこのモデルでできる最善の結果であるが、アルゴリズムの適合が尤度表面の極大値に至るようには見えず、おそらく、パラメータのいくつかで初期値が改善されるならば、より良い推定値が得られたであろう。平均のモデルはデータを十分説明していると考えられるため、平均のモデルのパラメータについて、初期値として古い推定値を選択し、分散関数のパラメータを推定する新しい初期値を得ることで、モデルを再度適合することができる。これらの新しい推定値は、平均の実測値の対数に対して分散の実測値の対数（つまり、SD の二乗）を回帰させることから来る。つまり、下記の式であり、

$$\text{Log}(\text{var}) = \log(\alpha) + \rho \log(\text{mean}),$$

式中、対数は自然対数を指す。この回帰のパラメータの推定値は $\rho=1.0$ 、 $\log(\alpha)=3.166$ であるため、 α の推定値は $e^{3.166}=23.7$ である。これらの新しい値から開始し、最終的な推定値は表 A.3.4 に示し、新しい予測値は表 A.3.5 にある通りである。BMD_{1SD} は 7.3467、BMDL_{1SD} は 5.96733 である（表には示されていない）。

Table A.3.4. Hill モデルの新しいパラメータ予測値と標準誤差

変数	予測値	SE
α	24.8892	24.5755
ρ	1.04671	0.162142
intercept	117.097	10.798
V	1629.2	64.9209
n	4.18855	1.33386
k	14.8385	1.86453

Table A.3.5. 最終 Hill モデルによる予測平均と標準誤差

用量	データ数	観測値 平均	観測値 SD	推定値 平均	推定値 SD	スケール後 残差
0	8	100	30.4	105	123	-0.115
0.3	8	98.2	49.8	105	123	-0.156
1	8	111	59.9	105	123	0.138
3	8	172	58.4	106	123	1.518
10	8	357	168	360	145	-0.059
30	8	1700	261	1690	178	0.028
100	8	1580	170	1580	179	-2.570
300	8	1900	142	1740	179	2.483

この適合の対数尤度は-333.2であり（表 A.3.6）、これまでの適合からかなり改善している。さらに、平均の推定値が実測値とよく一致しているだけでなく、SD の推定値は実測値にかなり近い。しかし、適合は改善しているが、この例の以下の BMD5 結果の抜粋が示している通り（表 A.3.7、A.3.8）、分散のモデル（以下の検定 3 の結果）、平均のモデル（以下の検定 4 の結果）ともデータには適合していない。

Table A.3.6. Likelihoods of Interest

モデル	対数（尤度）	自由度	AIC
A1	-343.706	9	705.4
A2	-317.77	16	667.5
A3	-324.533	10	669.1
Fitted	-333.127	6	678.3
R	-458.043	2	920.1

Table A.3.7. Explanation of Tests

試験 1	Does response and/or variances differ among dose levels? (A2 vs. R)
試験 2	Are variances homogeneous? (A1 vs. A2)
試験 3	Are variances adequately modeled? (A2 vs. A3)
試験 4	Does the model for the mean fit? (A3 vs. fitted)

Table A.3.8. Tests of Interest

試験	-2×対数（尤度比）	試験ごとの自由度	p 値
試験 1	280.547	14	<.0001
試験 2	51.8732	7	<.0001
試験 3	13.5263	6	0.0354
試験 4	17.1876	4	0.001777

何が起きているのだろうか。上記の適合値の表（表 A.3.5）で、特にスケーリングされた残差の欄で、この

モデルが用量「3」の反応を大幅に過小予測しており（スケーリングされた残差>2）、2つの最高用量の反応をいずれの側からも見逃している。さらに、このモデルは2つの最高用量でSDを過大予測している（おそらくこれが、分散のモデルが受け入れられない理由である）。しかし、低用量での過小予測は、適合したモデルから分かる限り、BMDの領域内にあるため、最も重要である。

何ができるだろうか。3つの最高用量30、100、300は、BMDからかなり離れている。これらの用量を除く場合、モデルでは十分に説明できない反応を示す用量を排除するが、BMDからは離れているため、BMDの実際の位置についてはあまり多くの情報が排除されない。さらに、プラトーの反応が低下した場合、他の単調な用量反応モデルをデータに適合させることができる。また、Hillモデルには下式で1次多項式を検討し、

$$\mu(d) = \beta_0 + \beta_1 d$$

パワーモデルは以下の通りである。

$$\mu(d) = \beta_0 + \beta_1 d^\gamma$$

一次多項式モデルは、より高次の項を検討した後の結果であり、高次の項ではモデルのデータ適合力は有意に増大しなかった。

これにより、最大反応Vの情報が実際に含まれていないデータセットが減少するため、Hillモデルの推定値Vは疑わしいことに注意が必要である（上述の表で報告されたモデルの推定値は過剰に大きく、143289、巨大なSEは 5.8×10^8 である）。しかし、BMRは漸近線が関与する領域から離れているため、これは低用量のBMDの計算に影響せず、このBMDは減少したデータセットについて対照の平均を超える1SDのBMRに対応する。

Table A.3 9. Final Model Comparison

モデル	適合度の p 値	AIC	BMD	BMDL
多項式モデル	0.98	375.5	1.46	1.11
Power モデル	0.95	377.4	1.66	1.11
Hill モデル	0.76	379.4	1.70	1.14

ここで報告した結果概要（表 A.3.9）とグラフおよび残差（ここには示していない）のより詳細な検討によれば、3つのモデルすべては減少したデータによく適合している、しかし多項式モデルのAICは他の2つよりもやや優れているため、これがBMDとBMDLを計算するために選択したモデルである。つまり、1SDの変化に基づくBMDとBMDLは1.46および1.11である。

この例では3つのポイントについて説明しており、いずれも連続データのモデル化に特異的なものではない。
 (1) 戻したパラメータが実際にMLEであると保証できるソフトウェアパッケージはないため、モデルをデータに適合する場合には判断力を使うことが重要であり、許容できる答えを得るため、分析者は試行錯誤することになる可能性がある（例えば、モデルパラメータについて異なる初期値を検討するなど）、
 (2) 我々は、BMR/BMD

の領域でデータを十分説明できるモデルが必要であり、他に利用できるモデルがなければ、モデルに試す用量範囲を賢明に狭めることになるかもしれない、(3) 必要な BMR の BMD を計算するため、科学的判断を行う必要があるかもしれない（例えば、適切なモデルを確認し、分散の不均一性を特徴付け、対照群の SD を十分に推定するなど）。どのような科学的およびリスク分析学的判断が望ましい答えを決定付けるかは、ソフトウェアの実行能力に従属すべきではない。

A.4. がんバイオアッセイデータ：線形外挿法の POD を得るためにモデル化する

この例では、用量反応のモデル化で典型的に制約されるモデルとして、多段階モデルを利用している（つまり、モデルの係数 ≥ 0 、例 A.1 を参照）。この多段階モデルは、十分なデータがない場合に標準的なバイオアッセイデータに EPA で長年使用されているモデルであり^{注17}、より生物学的なモデルをサポートする。EPA の 2005 年のがんガイドライン（U.S.EPA 2005a）では、がんバイオアッセイデータの定量的なリスク推定が、観察された範囲のデータをモデル化することで、典型的に計算され、10%超過リスクの BMR の BMDL が推定され、これらの値は一般に標準的ながんバイオアッセイデータで観察される範囲の下端付近になる。

^{注17} EPA では、（モデル係数を負以外に制限するだけの）BMDS 多段階モデルが利用できるようになるまで、（モデル係数を負以外に、低用量領域の上限を線形に制限する）線形多段階モデルを使用した。102 のデータセットを利用しこれらの 2 つのモデル形態を比較することで、BMD10 と BMDL10 は事実上同一になることが示された（Subramaniam ら 2006）。

この BMDL は線形の低用量外挿では POD の役割を果たし、がん発生日推定（すなわち、単位リスクまたは傾斜因子の推定）が得られる。線形外挿では、BMR の選択は通常、がん発生日推定に実質的には影響しないことに注意が必要である。しかし、EPA（2005a）で考察されている通り、発がん物質の作用機序から非線形的なアプローチが必要な場合、POD で選択した BMR 値は最終的な基準値に大きく影響する可能性がある。非線形的なアプローチを利用する場合、BMR の選択では、モデル化される前駆作用の生物学的性質（例えば、重症度）と用量反応関係の統計学的性質を検討する。（セクション 2.2 を参照。）

この例では、経口投与による発がんリスクの定量的推定について、米国 EPA のジブロクロロメタンに関する健康および環境への影響報告書（U.S.EPA 1988）で提示された用量反応データを利用する。要約情報は EPA の IRIS ウェブサイト（<http://www.epa.gov/iris/subst>）で利用できる。B6C3F1 マウスを用い、胃管栄養法により投与したがんバイオアッセイにおける腫瘍エンドポイントは、肝細胞の腺腫または、がんとした。研究選択とエンドポイント選択の理論的根拠は、BMD 算出に関する包括的な記載の重要な要素であるが、定量的な本例の範囲を超えている。

Table A.4.1. 用量反応データ^a

投与量 (mg/kg/day)	ヒト相当量 (mg/kg-day)	腫瘍発生率
0	0	6/50
50	2.83	10/49
100	5.67	19/50

^a NTP (1988). Toxicology and carcinogenesis studies of chlorodibromomethane (CAS No. 124-48-1) in F344/N rats and B6C3F1 mice (gavage studies). TR-282.

Available from http://ntp.niehs.nih.gov/ntp/htdocs/LT_rpts/tr282.pdf.

観察された範囲の下端に近かったため、BMR は 10%の超過リスクが選択された。EPA のがんガイドライン (U.S.EPA 2005a) では、BMR の選択は外挿法と独立させるべきであると強調しているが、10%の超過リスクは、POD からの線形外挿を利用する場合、標準的ながんバイオアッセイデータでは典型的な BMR である。片側の BMDL を 95%信頼水準について計算した。EPA のがんガイドラインでも、不確実性の指標を伝えるため、BMD または BMDU の上限を報告することが推奨されている。したがって、95%片側 BMDU についても推定した。上限、下限値を合わせると、90%両側信頼区間が得られる。

モデルの適合

まず、2 次 (すなわち $n-1$) 多段階モデルをデータに適合した。モデルの形態は以下の通りである。

$$P(\text{用量}) = \text{バックグラウンド} + (1 - \text{バックグラウンド}) \times [1 - \exp(-\beta_1 \times \text{用量} - \beta_2 \times \text{用量}^2)].$$

このモデルは、3 つの実測値すべてと正確に適合しているため (図 A.4.1、表 A.4.2)、 χ^2 適合度の p 値は定義されず、スケールされた残差はすべてゼロである。AIC は 158.7 であった。BMD および下限と上限 (尤度プロファイルから推定) の推定は以下の通りである。

BMD (ED_{10}) = 2.91 mg/kg-day

BMDL (LED_{10} ; 95%片側信頼限界) = 1.25 mg/kg-day

BMDU (UED_{10} ; 95%片側信頼限界) = 4.59 mg/kg-day

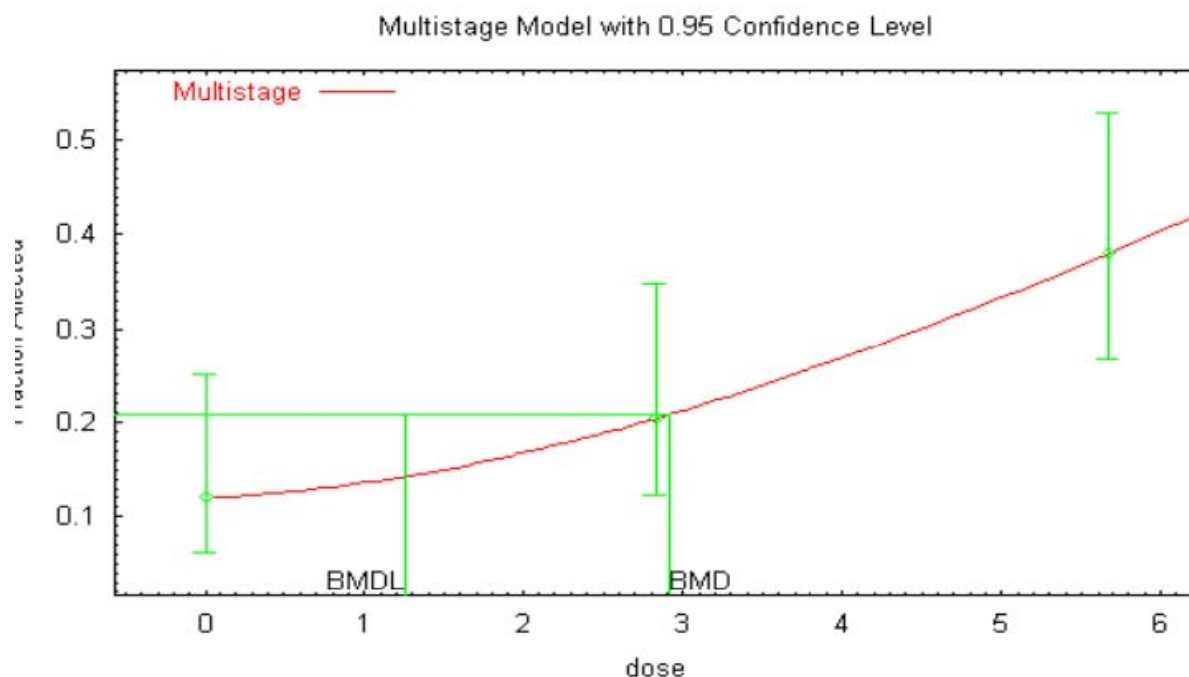


図 A.4.1. 適合した 2 次多段階モデル、およびデータの平均と標準誤差

Table A.4.2. 2 次多段階モデルのパラメータ予測値と標準誤差

パラメータ	最大尤度の予測値 (MLEs)	SE
バックグラウンド	0.12	0.132665
beta1	0.00930036	0.141898
beta2	0.00925286	0.0246904

次に、1 次多段階モデルをデータに適合し、より簡素なモデルでも十分に適合するかを観察した。モデルの形態は以下の通りである。

$$P(\text{用量}) = \text{バックグラウンド} + (1 - \text{バックグラウンド}) \times [1 - \exp(-\text{beta1} \times \text{用量}^1)]$$

1 次多段階モデルもデータは十分適合し (表 A.4.3 および A.4.4、図 A.4.2 を参照)、 χ^2 適合度の p 値は 0.4494 で、スケーリングされた残差は表 A.4.4 に示す通り、異常に大きくはない。AIC は 157.3 であった。BMD、BMDL、BMDU の推定は以下の通りである。

BMD (ED₁₀)=1.88 mg/kg/day

BMDL (LED₁₀ ; 95%片側信頼限界)=1.20 mg/kg-day

BMDU (UED₁₀ ; 95%片側信頼限界)=4.59 mg/kg-day

Table A.4.3. 1 次多段階モデルのパラメータ予測値と標準誤差

パラメータ	MLE	SE
バックグラウンド	0.111488	0.120556
beta1	0.0559807	0.0391492

Table A.4.4. 適合度

用量	Estimated probability	Expected number responding	Observed number responding	Group size	スケール後 残差
0.0000	8	100	30.4	105	-0.115
2.8300	8	98.2	49.8	105	-0.156
5.6700	8	1900	142	1740	2.483

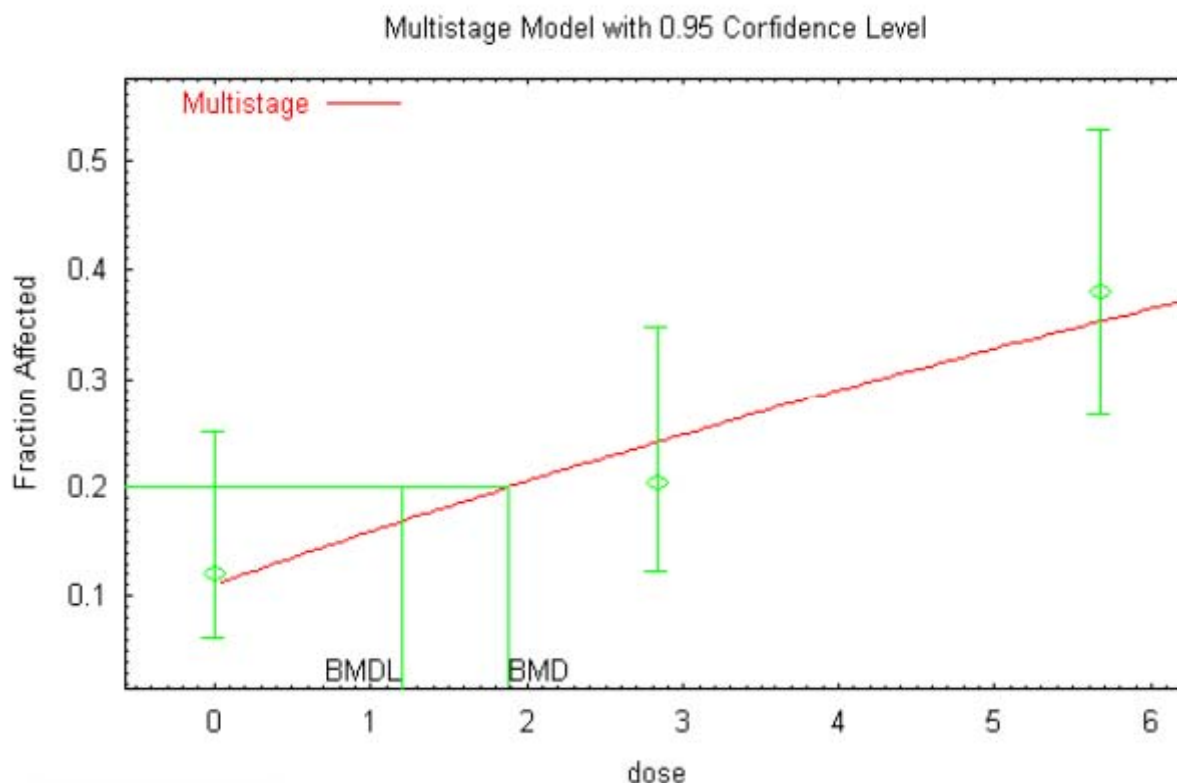


図 A.4.2. 適合した 1 次多段階モデル、およびデータの平均と標準誤差

モデルの比較

AIC は 1 次モデルでは低く、これが好ましいモデルであることを示唆している。多段階モデルは統計的に比較可能な k 次モデルの一種であるため、尤度比検定を利用し、追加パラメータを推定することで得られた適合の改善が妥当であるか否かを評価することもできる。この場合、2 次モデルの対数尤度は -76.3439 、1 次モデルの対数尤度は -76.6361 であった。従って、対数尤度の絶対差の 2 倍は 3.84 未満、つまり、自由度 1 のカイ二乗（すなわち 2.1）であり、1 次多段階モデルは 2 次モデルと有意な差はないことを示唆している。

POD の計算に使用するモデルの選択

ベンチマークドースのガイダンスを推奨するとすれば、一般にはより簡潔な 1 次モデルの方が好ましいだろう。この問題の最終判断は、エンドポイントを限定したガイダンスに従うことができる。従って、この例では、1 次モデルの BMDL (1.20 mg/kg-day) を POD として使用できるだろう。

A.5. 発生毒性データ

一般に、齧歯類における発生毒性研究のデータは、入れ子型モデルで最もよくモデル化される。これらのモデルは、同腹児内の関係、つまり同腹児が用量群の他の同腹児と比べて互いに同様に反応する傾向を説明している。（用量によって変化する）この関係が推定されていない場合、分散の推定、また従ってベンチマーク反応および用量の信頼限界は、全体的に誤って定められる。さらに、これらのモデルには、母動物の初期体重など、同腹児に特異的な共変量の条件が含まれることが多く、これが目的の結果と関連し（ただし、用量とは関連しない）、反応パターンの解明に役立つ可能性がある。この例では、適切なモデルの適合を選択する上でのこれらのパラメ

ータの評価が強調されている。

この例では、妊娠 6～15 日のマウスにエチレングリコールジエチルエーテルを経口投与した場合の発生毒性について、George ら（1992）が報告した用量反応データを使用している。（表 A.5.1 を参照。）本ガイドラインの他の例同様、この例では、1 つの用量反応パターンにモデル、ここではロジスティックな入れ子型モデルを当てはめて説明している。このモデルでは、入れ子型データの様々な用量反応形態を適合している。研究選択とエンドポイント選択の理論的根拠は、BMD 算出に関する包括的な記載の重要な要素であるが、定量的な本例の範囲を超えている。

モデル化した結果は、非連続エンドポイントである骨格奇形の発生率であった。同腹児数は、最高用量を除き投与量増加との関連性を示さなかったが、同腹児特異的な共変量と考えられた。BMR の 10%超過リスクは本例の目的のみで仮定している。

ロジスティックな入れ子型モデルでは、用量群の平均反応（示していない）と視覚的にかなりよく適合していることが証明され、適合度の p 値は 0.45 であった。このモデルの適合を承認する前に、同腹児数と、同腹児内の相関関係の重要性を評価した。反応率の予測における同腹児数の影響を評価する係数はかなりゼロに近く（それぞれ 0.0013 と -0.1507）、この場合、同腹児数は重要ではないことが示唆されたため、モデルは同腹児数なしで再度適合した。得られた適合では p 値が 0.184 となり、BMD の評価を裏付けるには十分であった。その AIC 450.6 も最初の適合の AIC 452.5 よりわずかに低く、モデルのパラメータが少ないことを反映している。

次に、同腹児内の相関関係（係数 ϕ_1 - ϕ_5 ）をゼロに設定することで、同腹児内の相関関係を評価した。この適合は成功せず、適合度 p 値は 0、AIC は 570.4 であった（上記の 450.6 と比較）。従って、同腹児内の相関関係は、このデータセットで観察された変動を説明するために重要である。結果として、同腹児内の相関関係を取り入れているが、同腹児特異的な共変量は取り入れていないモデルを選択した。用量群別の適合したモデルおよび平均反応は図 A.5.1 に示す。

Table A.5.1. Dose-Response Data for Skeletal Malformations Resulting From Ethylene Glycol Diethyl Ether Administered Orally to Mice on Days 6–15 of Gestation, George ⁵. (1992)

Dose	Litter-Specific Covariate	Litter Size	Number Affected	Dose	Litter-Specific Covariate	Litter Size	Number Affected	Dose	Litter-Specific Covariate	Litter Size	Number Affected
0	6	6	0	150	3	3	0	1000	3	3	3
0	8	8	0	150	10	10	0	1000	3	3	3
0	8	8	0	150	10	10	1	1000	3	3	3
0	9	9	0	150	11	11	0	1000	3	3	3
0	9	9	0	150	11	11	4	1000	3	3	3
0	10	10	0	150	11	11	5	1000	9	9	8
0	10	10	0	150	12	12	0	1000	9	9	9
0	11	11	0	150	12	12	0	1000	9	9	9
0	11	11	0	150	12	12	0	1000	10	10	5
0	11	11	0	150	12	12	0	1000	10	10	7
0	11	11	0	150	12	12	0	1000	10	10	8
0	12	12	0	150	12	12	1	1000	10	10	10
0	11	11	0	150	13	13	0	1000	11	11	5
0	11	11	0	150	13	13	0	1000	11	11	11
0	11	11	0	150	13	13	0	1000	11	11	11
0	11	11	0	150	13	13	0	1000	12	12	7
0	14	14	0	150	13	13	1	1000	12	12	11
0	14	14	0	150	14	14	0	1000	12	12	12
0	14	14	4	150	14	14	0	1000	13	13	8
0	15	15	0	150	15	15	1	1000	13	13	13
0	15	15	0	150	18	18	0	1000	14	14	13
0	15	15	0								
50	2	2	0	500	6	6	0				
50	5	5	0	500	8	8	0				
50	9	9	0	500	9	9	6				
50	9	9	0	500	10	10	0				
50	9	9	0	500	10	10	0				
50	10	10	0	500	10	10	2				
50	10	10	0	500	11	11	0				
50	11	11	0	500	11	11	0				
50	12	12	0	500	11	11	1				
50	12	12	0	500	11	11	2				
50	12	12	0	500	11	11	3				
50	12	12	0	500	11	11	4				
50	13	13	0	500	11	11	7				
50	13	13	0	500	12	12	0				
50	13	13	0	500	12	12	0				
50	13	13	0	500	12	12	0				
50	13	13	0	500	12	12	1				
50	14	14	0	500	12	12	1				
50	15	15	0	500	12	12	4				
				500	13	13	0				
				500	13	13	6				
				500	15	15	0				

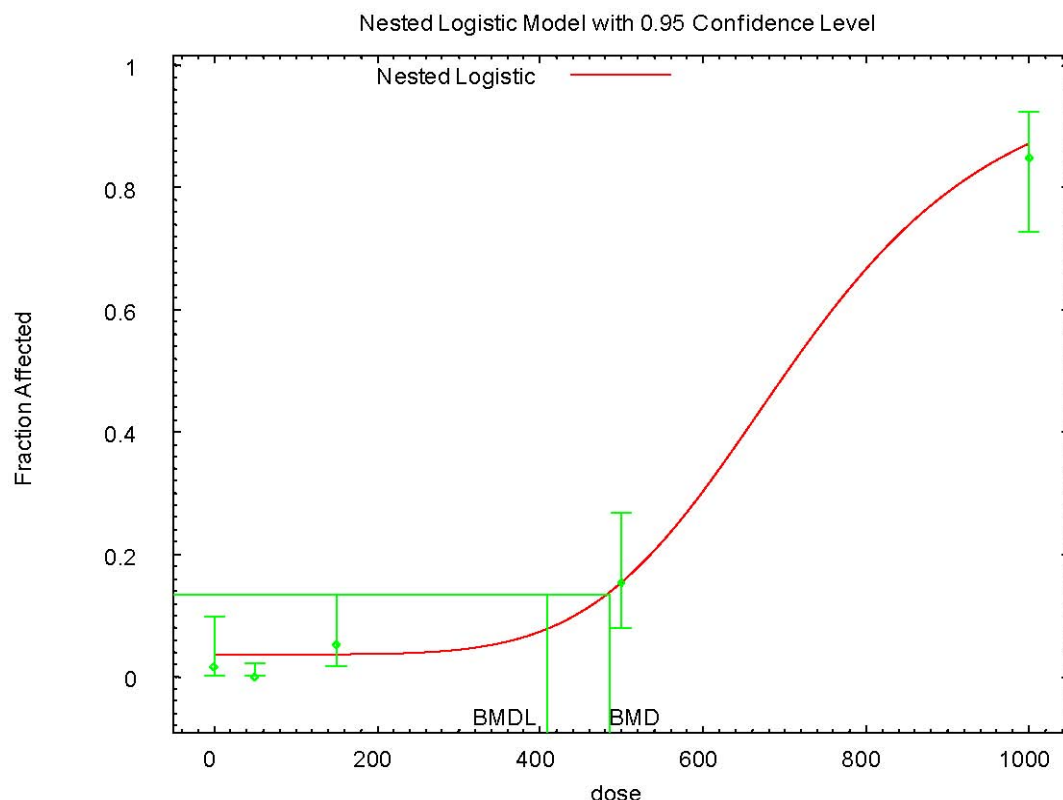


図 A.5.1. 骨格奇形がみられる児の割合と適合した入れ子型ロジスティックモデル

A.6. ヒトのデータ

ヒトの毒物学的データをモデル化する機会は限られ、ヒトの研究は実験動物の研究ほど標準化されていない。そのため、ヒトデータのモデル化は症例特異的に行われる。さらに、ヒトデータのモデル化には、共変量の調整が関与することが多い。加えて、一般に高齢であることと関連した影響（例えば、ほとんどのがんおよび循環器疾患）については、生命表の分析が行われることが多い（例えば、EPA による 1,3-ブタジエンに関する評価 (U.S.EPA 2002b) セクション 10.1 のがんのモデル化）。ヒトデータのベンチマークドースモデルの他の例については、以下の参考文献がある。EPA の IRIS データベースで紹介された例は、職業的に曝露された作業員の二硫化炭素で誘導される末梢神経系機能不全についてである (U.S.EPA 1995b)。IRIS の別の例は、メチル水銀に曝露したヒト幼児の発育神経異常についてである (U.S.EPA 1995c)。異なるヒトデータベースから得られたメチル水銀関連の発生神経学的影響に関するベンチマークドースモデルのより最近の例は、Budtz-Jorgensen ら (2000) と van Wijngaarden ら (2006) が報告している。ポリ塩化ビフェニルへのヒト曝露の発生神経学的影響のベンチマークドースモデルの例は、Jacobson らが示している (2002)。ヒトデータのベンチマークドースモデルの別の最近の例は、Suwazono ら (2006) によるカドミウム誘導腎毒性のモデル化である。

ヒトの毒性学的データは実験動物の毒性学的データに関する報告と同様に報告されることもあり（例えば、亜慢性作用のグループ分けしたデータで、共変量がなく、曝露の平均は用量グループごとに示される）、この場合、本ガイドライン文書が適応する（例えば、米国 EPA のベンゼンに関する評価 (U.S.EPA 2002d) について述べたセクション 5.1.2 において、リンパ球絶対数に関するヒトデータのモデル化を参照）。

付録 B. 用語集

以下の定義は本文書を分かりやすくするために示している。他の EPA 文書との不一致は意図的でない。
<http://www.epa.gov/iris> および <http://sis.nlm.nih.gov/enviro/enviropubs.html> の用語集も参照のこと。これらのリンクは、本文書を完成した時点で正しいと確認されたものであることに注意して欲しい。

追加リスク(Additional Risk)：追加リスクは、危険（本文書では、化学物質の特定用量または濃度）に曝露した曝露群と危険に曝露していない非曝露群のリスク（または反応確率）の差である。バイオアッセイおよびその用量反応分析の文脈では、追加リスクとは有害反応の確率がバックグラウンドの確率を超える増分であり、 $P(d)-P(0)$ と計算され、式中、 $P(d)$ は用量 d における反応リスクの確率、 $P(0)$ は用量ゼロでの反応確率（つまり、バックグラウンドのリスク）である。超過リスク（Extra Risk）も参照のこと。

赤池の情報量基準（Akaike Information Criteria: AIC）：特定のモデル同士を比較するために利用することのできる用量反応モデルの情報喪失の指標。AIC は $-2 \times (LL - p)$ と定義され、式中、LL はそのデータに与えられたモデルの対数尤度、 p はモデルに含められる推定パラメータの数である。特定モデルのなかでは、AIC がもっとも低いモデルが「最適」である。

漸近検定(Asymptotic Test)：サンプルサイズが無制限に大きくなるにつれて、検定統計値の分布が既知の分布に収束する統計学的検定。そのため、仮説を検証するための近似値として制限された分布を使用することができる。

ベイズアン(Bayesian)：これまでのデータ収集を基に、（集団の平均またはモデルのパラメータなどの）パラメータに確率または分布を割り当て、ベイズ理論を当てはめ、さらなる実験データを得た後、確率および分布を修正する統計学的方法に関する。

ベンチマーク濃度（Menchmark Concentration: BMC）：吸入した場合に、この作用のバックグラウンド反応率に対して有害作用の反応率が所定の変化を示す物質濃度。この所定の変化は「ベンチマーク反応」または BMR と呼ばれる。

ベンチマークドース（Benchmark Dose: BMD）法：摂取した場合に、この作用のバックグラウンドの反応率に対して有害作用の反応率が所定の変化を示す物質濃度。この所定の変化は「ベンチマーク反応」または BMR と呼ばれる。

ベンチマーク反応（Benchmark Response: BMR）：この作用のバックグラウンド反応率に対する有害作用反応率の所定の変化。BMR は BMD と BMDL を導く基礎である。

ベータ 2 項分布(Beta-Binomial Distribution)：クラスタ化またはネスト化された値、例えば、同腹仔の児の指標を示すために使われることがある統計分布。クラスタの事象の平均割合がベータ分布によって説明され、クラスタの事象数は 2 項分布によって説明される。

2 項分布(Binomial Distribution)：事象が起こる確率が同じ、独立した試験 n 件のサンプルから、 $0, 1, 2, \dots, n$ 件の事象が観察される確率の統計分布。

BMCL : ベンチマーク濃度 (BMC) の片側信頼限界の下限。

BMDL : ベンチマークドース (BMD) の片側信頼限界の下限。

BMDU : ベンチマークドース (BMD) の片側信頼限界の上限。

ブートストラップ(Bootstrap) : 実測値を何度も再サンプリングして置き換える統計手法 (ノンパラメトリックブートストラップ)。パラメトリックブートストラップでは、実測値から推定される確率分布を用い、新しいサンプルを作成する。例えば、(十分定義された集団の) 20 人の身長ランダムサンプルを基に、5,000 回データを置き換えて再サンプリングし、標準偏差と平均を毎回計算してもよい。得られた何らかの目的の量 (例えば、標準偏差または平均) の分布を用い、信頼限界を計算する、または計算が複雑な状況または推定または検定統計値の特定の分布が仮定できない場合に統計学的検定を行う。

がん発生力 (発がんスロープファクター) (Cancer Potency (Cancer Slope Factor)) : 単位用量あたりの基質に対する生涯曝露によりがん発生リスクが上昇した場合の増分を表す値。がん発生力は用量単位の逆数である単位で表される。ある用量での生涯発がんリスクを定量するため、特定の用量で乗じることができる。実際には、予想される値よりも推定される上限値に基づいていると考えられる。

カテゴリカルデータ(Categorical Data) : 自然順序関係のない情報 (名義的、例えば、修理屋、仕立屋、またはスパイ) や、または自然な順序 (序数的、例えば、軽度、中等度、重度) をカテゴリとして記録されたデータ。

中心推定(Central Estimate) : データセットの平均または中央値の推定。

カイ二乗適合度検定(Chi-square Goodness-of Fit Test) : 観察された総数を、2 つ以上のカテゴリに分類した独立した所見の予想数と比較するために用いる統計学的仮説検定。合計の総数は固定値と仮定する (多方向分類では、限界総数が固定されていることがある)。用量反応モデル化の文脈では、この検定は、各用量で観察された反応率が選択されたモデルの対応する予想 (または推定) 反応率と有意に異なるか否かを決定するために用いることが多い。観察された反応率が予測値から大きく逸れていると、カイ二乗値が大きくなり、選択したモデルが適合していないことを示す。例えば、用量と関連した結果の確率モデルを用い、4 つの用量レベル (カテゴリ) それぞれで反応する 50 匹の動物から動物数を予測し、予想数 (分数の場合もある) を出してもよい。カイ二乗統計値は、予想数から観察される総数の偏差の二乗の (4 用量の) 合計として計算し、それぞれを予想数で割る。モデルが正しければ、統計値の正確な分布は多項式になる。(妥当な大きさの反応確率の) 大きなサンプルでは、分布はカイ二乗分布に近づく。小さなサンプルでは、代わりに Fisher の直接確率検定を用いることができる。

クラスタ化されたデータ(Clustered Data) : 群またはクラスタで発生した個々のデータで回収される測定値。例えば、生殖・発生試験の同腹仔など。その群のメンバーで共通の状況 (母体環境、遺伝形質、飼育条件) は実験の結果に共通の影響を及ぼし、群のメンバー間で反応がより類似する可能性がある。そのような実験結果を説明する反応の統計モデルは、群内および群間の反応の変動を調節する。

コンケーブ(Concave)-コンベックス(Convex)を参照。

信頼区間 (両側) (Confidence Interval (Two-Sided)) : (上限と下限から成る) 統計的に導き出された区間であり、同じ集団から繰り返しサンプリングし、各サンプルからパラメータの不偏推定値を計算する場合、何らかの推定パラメータの真の値の境界となる特定の確率を有する。1つのサンプルの特定の信頼区間は、真のパラメータ値を含む場合と含まない場合がある。この区間には、特定の信頼値で推定パラメータの真の値が含まれると予想され、例えば、そのような区間の 95% が推定パラメータの真の値を含むと予想される。

信頼区間 (片側) (Confidence Interval (One-Sided)) : 上限または下限いずれかを含むが、その両方は含まない信頼区間。例えば、超過リスクが 10% 増加する用量 (BMD₁₀) の片側信頼区間の上限値は、上限 (BMDU) を 12.5 mg/kg-day として報告される。区間のもう一方の端は数学的または自然 (例えば、用量ゼロ) 下限である。片側信頼区間の下限値は下限 (BMDL)、例えば 2.67 mg/kg-day として報告され、この区間は数学的 (無限大) または自然上限まで広がると理解される。BMD の信頼限界を報告する場合、信頼水準と BMR をいずれも報告することが重要である。

信頼限界 (Confidential Limit) : 信頼区間の下限および/または上限 (信頼区間: Confidence interval を参照)。

条件付き用量反応モデル (Constrained Dose-Response Model) : モデルの 1 つ以上の推定が、例えばゼロに等しいか、それ以上である特定の範囲に制限されるモデル。

連続データ (Continuous Data) : データがカテゴリで記録されるカテゴリカルデータに対し (カテゴリカルデータ (Categorical Data) を参照)、臓器重量や酵素濃度など、連続的に測定されるデータ。

収束 (Convergence) : パラメータの推定の場合、サンプルサイズが増加するか、計算回数が増加するにつれて単一の値に近づくこと。

コンベックス (Convex) : 2 つの値につながっている関数の線 (弦) が、その関数値以上に位置する場合、関数は (その領域の一部の区間で) 凸面である。従って、関数 x が増加している場合、 x が増加すると勾配は増加する。例えば、スキーのジャンプまたはスケートボードの斜面はコンベックスであるが卵の上部はコンケーブである。「サブ線形 (sublinear)」は用量反応解析に固有のコンベックスと同義語であるが、「超線形 (supralinear)」はコンケーブに対応する同義語である。これらの用語の概念は、用量反応曲線は切片 (必ずしもゼロとは限らない) から曲線上のいずれかの注目点 (例えば BMD や他の POD) に引いた直線の下または上にあるというものである。

相関 2 項分布 (Correlated Binomial Distribution) : クラスタ化されたデータに代表される統計分布。クラスタの個々の要素、例えば同腹仔は、それぞれ作用を示す確率が同じである。

共変量 (Covariate) : 例えば、年齢、体重、多型など、目的の作用に影響する可能性のある、用量以外の独立変数。

収束確率 (Coverage Probability) : (理論的確率に対して) 母集団パラメータが一定の信頼区間で限界が決まる、実際の確率 (信頼区間 (Confidence Interval) を参照)。

立方 (モデルの立方項) (Cubic (Cubic Term in a Model)) : 3 乗 (X^3) となるモデルの項 (例えば、用量)。モデ

ルの最高次項とも呼ばれる（例えば、 $a+bX+cX^2+dX^3$ は、立方モデル、方程式、または式と呼ぶことができる）。

自由度(Degrees of Freedom)：用量反応モデルの適合では、「自由度」はデータポイント数－データから推定されるモデルのパラメータ数である。

デルタモデル(Delta Method)：テイラー級数の短縮により、ランダム変数の積率項についてランダム変数の関数の中心積率（分散など）を近似する方法。

二分(Dichotomize)：物、データ、事象を 2 つのグループに分けるまたは分類するプロセス。例えば、体重が特定の値を超えているかどうかで、50 匹の動物を 2 つのグループに分類することができる。

二分データ(Dichotomous Data)：ある作用が考えられる 2 つの結果のうち 1 つのみ、例えば、死亡または生存、腫瘍ありまたは腫瘍なしに分類される、カテゴリカルデータの一種。

バラツキ(Dispersion)：中心（平均または中央値）の値を囲む、ある量の変動の一般項。

用量反応モデル(Dose-Response Model)：ある用量の作用の指標を定量的に関連付ける（予測する）数学的関係（関数）。

用量反応傾向(Dose-Response Trend)：用量が増加すると反応の発生数または重症度が増加または低下する、生物学的反応と用量との定性的関係。

EC_P：対照反応と比較した、有害作用の P%の増加に対応する濃度。大気中濃度に基づく吸入曝露に用いられることが多い。

ED_P：対照反応と比較した、有害作用の P%の増加に対応する用量。投与量に基づく経口曝露に用いられることが多い。

推定値(Estimate)：典型的には、既知の母集団パラメータを表すことを意図したサンプル値。通常、サンプルデータに適用される推定量（Estimator: q.v）に基づく。

推定量(Estimator)：サンプルデータの未知の母集団パラメータの推定を導き出すために用いる手順、式、または値。例えば、サンプルの平均は母集団の平均の推定量である。数理統計学の本質的要素は、望ましい性質（例えば、確度および精度）を有する推定量を見出すことである。

過剰リスク(Excess Risk)：例えば、追加または超過リスクなど、比較群に対する有害事象が生じるリスクの増分。

超過リスク(Extra Risk)：同じ作用のバックグラウンド発生数に対して調整した有害作用リスクの比例増加の指標。つまり、用量（d）についてバックグラウンドよりも増加したリスクをバックグラウンドリスクに反応しなかった集団の割合で割った比率。超過リスクは $[P(d)-P(0)]/[1-P(0)]$ として計算される。追加リスク(Additional Risk)も参照のこと。

明確な作用(Frank Effect)：明白な、または臨床的に明らかな毒性作用。

ガンマ分布(Gamma Distribution)：様々な機能的形状、例えば平坦、尖鋭化している、非対称的などを説明することのできる、ゼロ以上の作用に限定された単峰型統計分布（用量など、何らかの指標の関数としての反応者の相対的割合）。

ガウス（正規）分布(Gaussian (Normal) Distribution)：平均を中心とし、標準偏差で測定される広がりまたはバラツキを有する、単峰型、対称的、釣鐘型の分布。

一般化推定方程式 (Generalized Estimating Equation: GEE)：モデルのパラメータを推定するために使用される統計学的手法。（最尤推定のように）分布を完全に指定する手法に対し、分布の最初の 2 つの積率を指定することのみが必要。

適合度統計量(Goofness-of Fit Statistic)：予想または仮定値で観察されたデータの偏差を測定する統計量。一部の適合度統計量は、統計的仮説検定に使用することができ、十分適合しないためにモデルを棄却する（または棄却できない）。

危険性の特定(Hazard Identification)：ヒトに起こる可能性がある作用の定性的記述など、化学的危険に曝露した結果である有害作用の特定。

Hill の式(Hill Equation)：酵素反応速度に用いられることが多い、用量反応関数。指数的に上昇する用量（d）の関数として、漸近線（最大値）に単調に近づく。関数は、 $F(d) = \gamma + \nu d^n / [k^n + d^n]$ である。

ハイブリッドモデル(Hybrid Model)：対照（曝露されていないヒトまたは動物）の極端な値に基づき連続データの異常値を確定し、用量の関数として異常反応レベルのリスクを推定するモデル。

発生数(Incidence)：一定の対象数または特定の集団について、特定の期間に発生した新しい症状の数。その期間中に感受性があった（まだ罹患しておらず、生存している）症状で、新たな発生率または出現を説明する。比率としても表される（例えば、104 週あたり 1 匹、1 年あたり 100,000 人など）累積発生数は、特定の期間にある疾病（例えば、がんまたは疾患）を示す特定集団の割合（例えば、特定の用量群で 2 年間に肝臓がんを示す被験動物の数）である。

非依存性(Independence)：いずれかの確率が他方の事象が発生するか否かに関わらず同じであれば、2 つの事象は依存していない。実験または観察研究では、これは、1 匹の動物または個体における結果が別の動物または個体に発生した同じ結果の確率に影響しないことを意味する。

切片項(Intercept Term)：用量反応モデルにおいて、用量ゼロまたはゼロの効果に対応する用量の推定値である。

最小二乗法(Least Squares)：観察されたデータポイントのモデルに基づく推定値からの偏差の二乗の合計を最小限にすることで、モデルのパラメータを推定する統計手法。すなわち、推定される残差分散を最小限にする。

尤度(Likelihood)：一定のデータセットを観察する確率（尤度関数で表される）に比例する数。特定の確率モデルと仮説に基づいたパラメータは正しいものと仮定する。これは、観察されたデータ、モデル、モデルパラメータの一部の特定値を条件とすることに注意して欲しい。最大尤度の方法によって、尤度関数の値を最大限とするこれらのパラメータ値が選択される。

尤度比検定(Likelihood Ratio Test)：一般モデルに基づくデータの最大尤度の、別のより限定的なモデルの最大尤度に対する比率に基づく統計的仮説検定。例えば、1 次および 2 次多段階モデルの最大尤度比を用い、多段階モデルの 2 次係数がゼロであるという仮説を検定することがあるかもしれない。 $-2 \log (L1/L2)$ の量は χ^2 の変量として漸近的に分布する（自由度は L1 と L2 について推定されたパラメータ数の差に等しい）。

線形用量反応モデル(Likehood Ratio Test)：反応の変化が用量の固定変化量に比例する数学的関係。例えば、反応 $=a+b \times$ 用量。これは、パラメータを線形に組み合わせた、より一般的な線形数学的モデルとは区別される。

最小作用量 (Lowest Observed Adverse Effect Level: LOAEL)：曝露集団とその適切な対照群の間で有害作用の頻度または重症度が生物学的に有意に増加する最小曝露量。

極大解(Local Maximum Solution)：母数空間の局所領域での関数の最大値の数学的解。許容される母数空間全体の最大値であることもあれば、そうでないこともある。BMD モデルで解が出る数値アルゴリズム（例えば、パラメータの最大尤度推定）は、特に非線型モデルでは、必ずしも全体の最大値が分かるとは限らず、そのパラメータの異なる初期値を用いて最大値決定プロセスを再開することで得られた解を検証するようなアドバイスが取られる。

ロジスティックモデル(Logistic Model)：特定の形のシグモイド（S 字型）関数。独立変数、例えば用量（d）に特異的な特徴を有する固体の割合に関する。関数は、 $P(d)=1/[1+\exp\{-\alpha - \beta \times \text{用量}\}]$

対数変換(Log Transformation)：データの対数を取るプロセス。対数変換は連続反応データに適用され、(a) 生データが対数正規分布の場合は、変換された反応が正規仮説を満たすようにするか、(b) 分散がすべての用量群でほぼ同じとなるような変換された反応を得ることもある（分散は均一と仮定）。

最尤推定 (Maximum Likelihood Estimate: MLE)：(サンプリング誤差に関する特定のモデルでの) 母集団パラメータの推定。尤度関数を最大化することで分かり、サンプルの所見が得られる可能性がもっとも高い。

ミカエリス・メンテン式(Michaelis-Menten Equation)：酵素反応速度の説明に用いられることが多い式。用量ゼロで勾配が最大となり、用量が増加すると漸近的に最大値に近づく。

曝露マージン (Margin of Exposure: MOE)：予想されるヒトの用量に対し、特定の作用、例えば、ベンチマークドースを生じる用量比。代わりに、LED10 または他の出発点を実際または投影された、目的の環境曝露で割ることもある。

単調な用量反応(Monotonic Dose-Response)：用量が増加（または低下）しても低下（または増加）しない用量

反応曲線。

多項分類(Multinomial Classification)：動物または被験者を2つ以上のカテゴリに分類すること。例えば、生殖試験では、「胎児は死亡」、「生存していて正常」、または「生存しているが異常」に分類することができる。

無毒性量 (No Observed Adverse Effect Level: NOAEL)：曝露集団とその適切な対照との間に有害作用の頻度または重症度に、生物学的に有意な増加がない最高曝露量。この用量レベルで生じる作用もあるが、有害である、または有害作用を引き起こすものとは考えられない。

非線形用量反応モデル(Nonlinear Dose-Response Model)：反応の変化が用量の固定変化量に比例しているため、単純に表すことのできない数学的関係またはモデル。非線形用量反応モデルの例は、(1) 反応= $a+b \times \text{用量}$ 、および (2) 反応= $a+b \times \log\{\text{用量}\}$ である。これは、パラメータを非線形に組み合わせた、より一般的な非線形数学的モデルとは区別される点に注意して欲しい。

目的関数(Objective Function)：最大または最小となるべき関数（例えば、最尤推定の尤度）。

順序データ(Ordinal Data)：順序または階級を付けることのできるデータ。

P 値(P-Value)：仮説を検定する場合、第1種の過誤（偽陽性）の確率。つまり、帰無仮説が真である場合は、帰無仮説を棄却する確率。

パラメータ(Parameter)：システムの測定または定量可能な特徴（例えば、用量反応関係、確率分布）。モデル化では、パラメータは通常未知で、推定量(Estimators)を用いてサンプルの測定に基づき推定する必要がある。

パーセンタイル値(Percentile)：昇順に並べられた n 個の測定値セットの k パーセンタイル値は、それ未満の $k\%$ およびそれを超える $(100-k)\%$ の測定値を有する値である。母集団パラメータとして、 k パーセンタイル値は、累積確率 $(k/100)$ 、 $k=1, 2, \dots, 98, 99$ に対応する確率分布関数の範囲の値 x である。

出発点 (Point of Departure : POD)：低用量の外挿で出発点を示す用量反応点。POD は NOAEL/LOAEL とすることもできるが、理想的には、実験データの BMD モデル化によって確定し、一般に選択し、推定される低レベルの反応に対応する（例えば、非連続作用では 1～10%の発生数）。作用機序および他の利用できるデータによっては、低用量リスクの推定に POD 未満で何らかの外挿を利用するか、一連の不確実計数で POD を割り、参照用量 (RfD) とすることができる。

多項式 (1 変量) (Polynomial (in one variable))：変量に係数、例えば $a+bx^2+cx^3$ を乗じた累乗の合計から成る数学的関数。最も大きな累乗が（一変量）多項式の次数となる。

確率(Probability)：特定の結果または事象発生の可能性。確率は 0～1 の値をとり、0 は事象が発生しないことを示し、1 は事象が常に発生することを示す。

確率分布(Probability Distribution)：事象の考えられるすべての結果の相対確率の（分布の形での）統計的記述。

プロビット関数(Probit Function)：用量の関数としての作用の相対確率が正規分布で説明されるとの仮定に由来する関数。用量の関数としての累積確率はシグモイド形である。プロビット用量反応関数は $P(d)=\Phi(\alpha+\beta\times\text{用量})$ であり、 Φ は累積標準正規または誤差関数である。

プロファイル尤度(Pfile likelihood)：(1) プロファイル尤度法は尤度の漸近分布を利用し、パラメータまたはパラメータの関数に関する仮説を検証し、その信頼区間を作成する。(2) 尤度プロファイルは、パラメータの固定値に対する尤度関数の最大値のプロットである。

二次項(Quadratic Term)：二次となった数学的関数の変量。

非連続データ(Quantal Data)：特定の腫瘍タイプの有無、またはホルモンレベルが正常か異常かなど、全か無かの作用を示すデータ。二分データ(Dichotomous Data)を参照。

変位値(Quantile)：確率分布関数の範囲内の特定パーセンタイル値。例えば、自由度 1、累積確率 0.95（すなわち、 $\Pr\{x\leq X\}\geq 0.95$ ）の χ^2 分布の変位量は 3.84（四捨五入）である。

擬似尤度(Quasi-Likelihood)：完全な定義がなく、一般には平均と分散を含む表現のみに基づく尤度関数。

回帰分析(Regression Analysis)：従属変量（生物作用）を独立変量（用量、曝露期間、年齢など）に定量的に関連付ける数学的関数（回帰方程式）を推定する統計手法。

反復測定(Repeated Measures)：複数の時点で同一被験者について測定される生物学的エンドポイント（例えば、異なる年齢での体重など）。

残差分散(Residual Variance)：独立変量（用量、曝露期間、年齢など）による分散を説明した後に残った実験的測定の分散（分散(Variance)を参照）。

残差(Residuals)：通常は回帰分析の文脈における実測値と推定値の数値差。スケーリングされた残差を参照。

参照濃度 (Reference Concentration: RfC) または参照用量 (Reference Dose: RfD)：(感受性の高いサブグループを含む) ヒト集団が曝露する可能性があり、生涯で悪影響が出る特記すべきリスクはないと考えられる、(不確実性がおそらく一桁分に及ぶ) 物質の濃度または用量の推定値。

リスク(Risk)：動物またはヒトが特定の曝露条件で特定の有害作用を示す確率。通常、0～1 のスケールで表す。

リスクの特徴決定(Risk Characterization)：危険性、曝露用量反応に関する情報を統合し、同定された有害作用がヒトで発生する尤度を推定する、リスクアセスメントプロセスの最終段階。

スケーリングされた残差(Scaled Residuals)：本文書において、スケーリングされた残差は標準誤差 (SE) で割って標準化した残差、つまり、(反応の実作用値－予想値) / SE である。

2次(Second Degree) : 2次つまり二乗項を含む数学的関数。

形状パラメータ(Shape Parameter) : 関数の曲率を示す、用量反応関数の用量の指数部。

有意性 (統計的有意性) (Significance(Statistical Significance)) : P 値(P value)を参照。

サブ線形 (sublinear)、超線形 (supralinear)、コンベックス(Convex)、コンケーブ(Concave) : コンベックス(Convex)の項目を参照。

閾値用量(Threshold Dose) : この値未満では特定の生物学的作用が発生しない用量。

不確実性(Uncertainty) : 不確実性は、定性的または定量的に何が真実であるかについて、正確な知識がないことと定義することができる (NRC、1994)。不確実性は、一般にさらに研究を重ねることで減少する可能性のある変動性 (Variability: q.v.) とは異なる。

不確実係数(Uncertainty Factor) : RfC または RfD を導き出すため、NOAEL、LOAEL、またはベンチマークドースの調整に用いられる数値(3 または 10 の倍数であることが多い)。不確実係数は実験動物からヒトへの外挿、感受性の高いサブグループを含む個人間バラツキ、LOAEL から NOAEL への外挿、亜慢性曝露の結果から慢性曝露への外挿、データベース不足を説明するため、必要に応じて使用される。

制限なしモデル(Unconstrained Model) : 母数空間に、したがってパラメータの推定に、制限がかからないモデル。

上側確率(Upper-Tail Probability) : 変量が特定の値を超える確率。

変動性(Variability) : 生物学的感受性または反応、および曝露の特徴 (呼吸数、食事摂取量など) における固有の観察可能な多様性 (個体間であることが多い)。これらの差は、さらに研究を重ねることでより理解される可能性があるが、一般に減少する。

分散(Variance) : 変動性の統計学的指標。標準偏差の二乗。

加重最小二乗推定(Weighted Least Squares Estimate) : 実測値の二乗の合計ー推定値を関数で加重した合計値を最小限にすることで得られるパラメータの推定値。この関数は観察値の分散の逆数であることが多い。

付録 C. 特定のベンチマークドースモデル

本文書で述べた一部のモデルの説明を以下に示す。追加情報は、例えば Filipsson ら（2003）に見ることができる。

非連続モデル

ここでは、 $0 \leq P(X) \leq 1$ が用量 $X > 0$ で二分結果となる確率である。以下に示すパラメータの制限により用量反応関数は増加するものと仮定している。

ガンマモデル

$$P(X) = \gamma + (1 - \gamma) [\Gamma(\alpha)^{-1} \{0 \int^{\beta x} t^{\alpha-1} e^{-t} dt\}], \alpha \geq 0, \beta > 0, 0 \leq \gamma < 1$$

- γ は「バックグラウンド」
- α は「指数」。通常は無限大の勾配が初期値に近づかないようにするため、 $\alpha \geq 1$ に制限している。
- β は「勾配」

ロジスティックモデル

$$P(X) = F\{-(\alpha + \beta X)\}, 0 \leq \gamma < 1, -\infty < \alpha < +\infty, \beta > 0$$

$$= F\{-(X + (-\alpha)(1/\beta)) / (1/\beta)\}$$

$$\text{式中、} F\{-(\alpha + \beta X)\} = [1 + \exp\{-(\alpha + \beta X)\}]^{-1}$$

- α は「切片」
- β は「勾配」

対数ロジスティックモデル

$$P(X; \gamma, \beta) = \gamma + (1 - \gamma) F\{-(\alpha + \beta \ln X)\}, 0 \leq \gamma < 1, -\infty < \alpha < +\infty, \beta > 0$$

$$= \gamma + (1 - \gamma) F\{-(\ln X - (-\alpha)(1/\beta)) / (1/\beta)\},$$

$$\text{式中、} F\{-(\alpha + \beta \ln X)\} = [1 + \exp\{-(\alpha + \beta \ln X)\}]^{-1}$$

- γ は「バックグラウンド」
- α は「切片」
- β は「勾配」。通常は無限大の勾配が初期値に近づかないようにするため、 $\beta \geq 1$ に制限している。

多段階モデル

$$P(X) = \gamma + (1 - \gamma) [1 - \exp\{-\sum \beta_j X\}], j=1, \dots, k, 0 \leq \gamma < 1$$

- γ は「バックグラウンド」
- $\beta_1, \dots, \beta_k$ は「勾配」。通常は単調な曲線とするため、 $\beta_j \geq 0$ に制限している。

プロビットモデル

$$P(X) = P(X; \gamma, \beta) = \Phi\{\alpha + \beta X\}, 0 \leq \gamma < 1, -\infty < \alpha < +\infty, \beta > 0$$

$$= \Phi\{[X + (-\alpha)(1/\beta)] / (1/\beta)\}$$

- α は「切片」
- β は「勾配」

対数プロビットモデル

$$P(X)=\gamma+(1-\gamma)\Phi\{\alpha+\beta\ln X\}, 0\leq\gamma<1, -\infty<\alpha<+\infty, \beta>0$$

$$=\gamma+(1-\gamma)\Phi\{[\ln X-(\alpha)(1/\beta)]/(1/\beta)\}$$

- γ は「バックグラウンド」
- α は「切片」
- β は「勾配」

ワイブルモデル

$$P(X)=\gamma+(1-\gamma)[1-\exp\{-\beta x^\alpha\}], \alpha\geq 0, 0\leq\gamma<1, \beta>0$$

- γ は「バックグラウンド」
- β は「勾配」
- α は「指数」。通常は無限大の勾配が初期値に近づかないようにするため、 $\alpha\geq 1$ に制限している。

二分 Hill モデル

$$P(X)=v[1+g \exp\{-(a+b \log(X))\}]/[1+\exp\{-(a+b \log(X))\}]$$

- $0\leq g<1, 0<v\leq 1, b\geq 0$
- v はモデルにより予測される反応の最大確率
- g に $v(v\times g)$ を掛けると、反応の確率のバックグラウンド推定値
- b は「勾配」

入れ子型対数ロジスティックモデル $P(X)=\alpha+\theta 1r_{ij}+[1-\alpha-\theta 1r_{ij}]/[1+\exp\{\beta+\theta 2r_{ij}-\gamma \log(X)\}]$, 用量>0 の場合
 $=\alpha+\theta 1r_{ij}$, 用量=0 の場合

oith 用量群の j^{th} 同腹仔については、 r_{ij} は同腹仔特異的な変量

- $\alpha\geq 0, \beta>0, \gamma\geq 0$, およびすべての r_{ij} で $\alpha+\theta 1r_{ij}\geq 0$
- α は「バックグラウンド」
- β は「勾配」
- γ は「指数」。通常は無限大の勾配が初期値に近づかないようにするため、 $\gamma\geq 1$ に制限している。

連続モデル

ここでは、 $\mu(X)$ は用量 $X>0$ での平均反応である。用量群の分散を（平均の指数関数として）モデル化し、一定であると仮定することができる。平均反応をモデル化する

優良アプローチを以下に示す。

多項式連続モデル

$$\mu(X)=\gamma+\sum \beta_j X^j, j=1, \dots, n$$

- 曲線を単調にするため、 β_j は通常 $\beta_j\geq 0$ （反応が増加している場合）または $\beta_j\leq 0$ （反応データが減少している場合）に限定する。

指数連続モデル

$$\mu(X)=\gamma+\beta X^\alpha, \alpha>0, \beta>0$$

- γ は「バックグラウンド」
- β は「勾配」

- α は「指数」。通常は無限大の勾配が初期値に近づかないようにするため、 $\alpha \geq 1$ に制限している。

Hill の連続モデル

$$\mu(X) = \gamma + vX^n / (k^n + X^n)$$

- γ は「バックグラウンド」
- k は「勾配」
- v は漸近線
- n は「指数」。通常は無限大の勾配が初期値に近づかないようにするため、 $n \geq 1$ に制限している。

指数関数連続モデル、入れ子型モデルの例：

$$\text{モデル 2 : } \mu(X) = \gamma \exp\{\text{sign } k X\}$$

$$\text{モデル 3 : } \mu(X) = \gamma \exp\{\text{sign}(k X)^d\}$$

$$\text{モデル 4 : } \mu(X) = \gamma (c - (c-1)\exp\{-1 k X\})$$

$$\text{モデル 5 : } \mu(X) = \gamma (c - (c-1)\exp\{-1(k X)^d\})$$

- γ は「バックグラウンド」
- b は「勾配」
- 「sign」は変化の方向を示す。+1 は反応が増加、-1 は反応が減少している。
- c は漸近線のパラメータであり（モデル 3 および 5 のみ）、データが減少している場合は $0 < c < 1$
- d は「指数」。通常は $d > 1$ に制限する（モデル 3 および 5 のみ）

付録 D. ベンチマークドース技術ガイダンス文書の寄稿者と校閲者

寄稿者

John Fox, Office of Research and Development, U.S. EPA, Washington, DC

Allan Marcus, Office of Research and Development, U.S. EPA, Research Triangle Park, NC

David Svendsgaard, Office of Research and Development, U.S. EPA, Research Triangle Park, NC

Paul White, Office of Research and Development, U.S. EPA, Washington, DC

Diane Henshel, Office of the Science Advisor, U.S. EPA, Washington, DC

外部専門校閲者

George Alexeef, Office of Environmental Health Hazard Assessment, California Environmental Protection Agency, Oakland, CA 94612

Kevin Brand, Department of Epidemiology and Statistical Medicine, University of Ottawa, Ottawa ON, K1N 5C8 Canada

Paul Catalano, Department of Biostatistical Science, Harvard School of Public Health, Boston, MA 02115

Harvey Clewell, ICF Consulting, Ruston, LA

George Daston, Miami Valley Laboratories, Proctor and Gamble, Cincinnati, OH

Elaine Faustman, Department of Environmental Health, University of Washington, Seattle, WA

Clay Frederick, Toxicology Department, Rohm and Haas Company, Spring House, PA

Lynne Haber, Toxicology Excellence for Risk Assessment, Cincinnati, OH

Dale Hattis (Workshop Chair), Center for Technology, Environment and Development, Clark University, Worcester, MA

Colin Park, Consultant, Midland, MI

Lorenz Rhomberg, Gradient Corporation, Cambridge, MA

Robert Sielken, Jr., JSC Sielken, Bryan, TX

William Slikker, Jr., U.S. Food and Drug Administration, Rockville, MD

R. Webster West, Department of Statistics, University of South Carolina, Columbia, SC

Yiliang Zhu, Department of Epidemiology & Biostatistics, University of South Florida, Tampa, FL

参考文献

注： 以下に示すウェブリンクは、本文書を完成した時点で正確なものであった。

- Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov, BN; Csaki, F; eds. Proceedings of the second international symposium on information theory. Budapest, Hungary: Akademiai Kiado, pp. 267-281.
- Alexeeff, GV; Lewis, DC; Ragle, NL. (1993) Estimation of potential health effects from acute exposure to hydrogen fluoride using a 'benchmark dose' approach. *Risk Anal* 13(1):63-69.
- Allen, BC; Kavlock, RJ; Kimmel, CA; Faustman, EM. (1994a) Dose-response assessment for developmental toxicity: II. Comparison of generic benchmark dose estimates with NOAELs. *Fundam Appl Toxicol* 23:487-495.
- Allen, BC; Kavlock, RJ; Kimmel, CA; Faustman, EM. (1994b) Dose-response assessment for development toxicity: III. Statistical models. *Fundam Appl Toxicol* 23:496-509.
- Allen, BC; Strong, PL; Price, CJ; Hubbard, SA; Daston, GP. (1996) Benchmark dose analysis of developmental toxicity in rats exposed to boric acid. *Fundam Appl Toxicol* 32:194-204.
- Auton, TR. (1994) Calculation of benchmark doses from teratology data. *Regul Toxicol Pharmacol* 19:152-167. Bailer, AJ; Noble, RB; Wheeler, MW. (2005) Model uncertainty and risk estimation for experimental studies of quantal responses. *Risk Anal* 25:291-299.
- Bailer, AJ; Portier, CJ. (1988) Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. *Biometrics* 44::417-431.
- Barnes, DG; Daston, GP; Evans, JS; Jarabek, AM; Kavlock, RJ; Kimmel, CA; Park, C; Spitzer, HL. (1995) Benchmark dose workshop: criteria for use of a benchmark dose to estimate a reference dose. *Regul Toxicol Pharmacol* 21:296-306.
- Bates, DM; Watts, DG. (1988) Nonlinear regression analysis and its applications. New York, NY: Wiley.
- Beck, BD; Conolly, RB; Dourson, ML; Guth, D; Hattis, D; Kimmel, C; Lewis, SC. (1993) Symposium overview: improvements in quantitative noncancer risk assessment. *Fundam Appl Toxicol* 20:1-14.
- Benford, D; Bolger, PM; Carthew, P; Coulet, M; DiNovi, M; Leblanc, J-C; Renwick, AG; Setzer, W; Schlatter, J; Smith, B; Slob, W; Williams, G; Wildemann, T. (2010) Application of the Margin of Exposure (MOE) approach to substances in food that are genotoxic and carcinogenic. *Food Chem Toxicol* 48:S2-S24.
- Budtz-Jorgensen, E; Grandjean, P; Keiding, N; White, RF; Weihe, P. (2000) Benchmark dose calculations of methylmercury-associated neurobehavioural deficits. *Toxicol Lett* 112-113:193-199.
- California EPA (California Office of Environmental Health Hazard Assessment). (1994) Safety assessment for non- cancer endpoints: the benchmark dose and other possible approaches. Summary report.
- Catalano, PJ; Ryan, LM. (1992) Bivariate latent variable models for clustered discrete and continuous outcomes. *J Am Stat Assoc* 87:651-658.
- Catalano, PJ; Scharfstein, DO; Ryan, LM; Kimmel, CA; Kimmel, GL. (1993) Statistical model for fetal death fetal weight and

malformation in developmental toxicity studies. *Teratology* 47:281-290.

Chen, C; Farland, W. (1991) Incorporating cell proliferation in quantitative cancer risk assessment: approaches, issues, and uncertainties. In: Butterworth, B; Slaga, T; Farland, W; et al., eds. *Chemical induced cell proliferation: implications for risk assessment*. New York, NY: Wiley-Liss, pp. 481–499.

Chen, JJ; Kodell, RL. (1989). Quantitative risk assessment for teratologic effects. *J Am Stat Assoc* 84:966-971.

Chen, JJ; Kodell, RL; Howe, RB; Gaylor, DW. (1991) Analysis of trinomial responses from reproductive and developmental toxicity experiments. *Biometrics* 47:1049-1058.

Clayton, D; Hills, M. (1993) *Statistical models in epidemiology*. Oxford, UK: Oxford University Press. Crump, KS.

(1984) A new method for determining allowable daily intakes. *Fundam Appl Toxicol* 4:854-871. Crump, KS. (1995)

Calculation of benchmark doses from continuous data. *Risk Anal* 15:79-89.

Crump, KS. (2002). Critical Issues in benchmark dose calculations from continuous data. *Crit Rev Toxicol* 32:133- 153.

Crump, KS; Hoel, DG; Langley, CH; Peto, R. (1976) Fundamental carcinogenic processes and their implications for low dose risk assessment. *Cancer Res* 36:2973-2979.

Crump, KS; Howe, R. (1985). A review of methods for calculating statistical confidence limits in low-dose extrapolation. In: Clayton, DB; Krewski, D; Munro, I; eds. *Toxicological Risk Assessment*. Boca Raton, FL: CRC Press, Inc.

Davidian, M; Giltinan, DM. (1995) *Nonlinear models for repeated measurement data*. London, UK: Chapman and Hall.

Dourson, ML; Hertzberg, RC; Hartung, R; Blackburn, K. (1985) Novel methods for the estimation of acceptable daily intake. *Toxicol Ind Health* 1:23-41.

Draper, N; Smith, H. (1981) *Applied regression analysis*. 2nd ed., chap 10. New York, NY: Wiley.

Farmer, JH; Kodell, RL; Gaylor, DW. (1982) Estimation and extrapolation of tumor probabilities from a mouse bioassay with survival/sacrifice components. *Risk Anal* 2(1):27-34.

Faustman, EM; Allen, BC; Kavlock, RJ; Kimmel, CA. (1994) Dose-response assessment for developmental toxicity: I. Characterization of data base and determination of NOAELs. *Fundam Appl Toxicol* 23:478-486.

Filipsson, AF; Sand, S; Nilsson, J; Victorin, K. (2003) The benchmark dose method—a review of available models, and recommendations for application in health risk assessment. *Crit Rev Toxicol* 33:505-542.

Filipsson, AF; Victorin, K. (2003) Comparison of available benchmark dose softwares and models using trichloroethylene as a model substance. *Regul Toxicol Pharmacol* 37:343-355.

Fowles, JR; Alexeeff, GV; Dodge, D. (1999) The use of benchmark dose methodology with acute inhalation lethality data. *Regul Toxicol Pharmacol* 29:262-278.

Fung, KY; Marro, L; Krewski, D. (1998) A comparison of methods for estimating the benchmark dose based on overdispersed data from developmental toxicity studies. *Risk Anal* 18:329-342.

Gallant, AR. (1987) Nonlinear statistical models. New York, NY: Wiley.

Gart, JJ; Krewski, D; Lee, PN; Tarone, RE; Wahrendorf, J. (1986) Statistical methods in cancer research. Vol. 3. The Design and Analysis of Long-Term Animal Experiments. Lyon, France: International Agency for Research on Cancer.

Gaylor, DW. (1983) The use of safety factors for controlling risk. *J Toxicol Environ Health* 11:329-336.

Gaylor, DW. (1992) Incidence of developmental defects at the no observed adverse effect level (NOAEL). *Regul Toxicol Pharmacol* 15:151-16

Gaylor, DW. (1996) Quantalization of continuous data for benchmark dose estimation. *Regul Toxicol Pharmacol* 24:246-250.

Gaylor, DW; Aylward, LL. (2004) An evaluation of benchmark dose methodology for non-cancer continuous-data health effects in animals due to exposures to dioxin (TCDD). *Regul Toxicol Pharmacol* 40:9-17.

Gaylor, DW; Kodell, RL. (1980) Linear interpolation algorithm for low dose risk assessment of toxic substances. *J Environ Pathol Toxicol* 4:305-312.

Gaylor, DW; Ryan, L; Krewski, D; Zhu, Y. (1998) Procedures for calculating benchmark doses for health risk assessment. *Regul Toxicol Pharmacol* 28:150-164.

Gaylor, D; Slikker, W, Jr. (1990) Risk assessment for neurotoxic effects. *Neurotoxicology* 11:211-218. Gaylor, D;

Slikker, W, Jr. (2004) Role of the standard deviation in the estimation of benchmark doses with continuous data. *Risk Anal* 24:1683-1687.

Gehlhaus, MW; Gift, JS; Hogan, KA; Kopylev, L; Schlosser, PM; Kadry, AR. (2011) Approaches to cancer assessment in EPA's Integrated Risk Information System. *Toxicol Appl Pharmacol* 254:170-180.

George, JD; Price, CJ; Marr, MC; Kimmel, CA; Schwetz, BA; Morrissey, RE. (1992) The developmental toxicity of ethylene glycol diethyl ether in mice and rabbits. *Fundam Appl Toxicol* 19:15-25.

Guth, DJ; Carroll, RJ; Simpson, DG; Zhou, H. (1997) Categorical regression analysis of acute exposure to tetrachloroethylene. *Risk Anal* 17:321-332.

Hasselblad, V; Jarabek, AM. (1995) Dose-response analysis of toxic chemicals. In: Berry, DA; Stangl, DK; eds. Bayesian biostatistics. New York, NY: Marcel Dekker, Inc.

Hertzberg, RC. (1989) Fitting a model to categorical response data with application to species extrapolation of toxicity. *Health Physics* 57:405-409.

Hertzberg, RC; Miller, M. (1985) A statistical model for species extrapolation using categorical response data. *Toxicol Ind Health* 1:43-57.

Jacobson, JL; Janisse, J; Banerjee, M; Jester, J; Jacobson, SW; Ager, JW. (2002) A benchmark dose analysis of prenatal exposure to polychlorinated biphenyls. *Environ Health Perspect* 110:393-398.

Kang, SH; Kodell, RL; Chen, JJ. (2000) Incorporating model uncertainties along with data uncertainties in microbial risk

assessment. *Regul Toxicol Pharmacol* 32:68–72.

Kavlock, RJ; Allen, BC; Kimmel, CA; Faustman, EM. (1995) Dose-response assessment for developmental toxicity: IV. Benchmark doses for fetal weight changes. *Fundam Appl Toxicol* 26:211-222.

Kavlock, RJ; Schmid, JE; Setzer, RW, Jr. (1996) A simulation study of the influence of study design on the estimation of benchmark doses for developmental toxicity. *Risk Anal* 16:391-403.

Kimmel, CA; Gaylor, DW. (1988) Issues in qualitative and quantitative risk analysis for developmental toxicity. *Risk Anal* 8:15-20.

Kimmel, CA; Wellington, DG; Farland, W; Rose, P; Manson, JM; Chernoff, N; Young, JF; Selevan, SG; Kaplan, N; Chen, C; Chitlik, LD; Siegel-Scott, CL; Valaoras, G; Wells, S. (1989) Overview of a workshop on quantitative models for developmental toxicity risk assessment. *Environ Health Perspect* 79:209-215.

Kodell, RL; Chen, JJ; Gaylor, DW. (1995) Neurotoxicity modeling for risk assessment. *Regul Toxicol Pharmacol* 22:24-

Krewski, D; Brown, C; Murdoch, D. (1984) Determining "safe" levels of exposure: safety factors or mathematical models? *Fundam Appl Toxicol* 4:S383-S394.

Krewski, D; Zhu, Y. (1994). Applications of multinomial dose-response models in developmental toxicity risk assessment. *Risk Anal* 14:613-627.

Krewski, D; Zhu, Y. (1995) A simple data transformation for estimating benchmark doses in developmental toxicity experiments. *Risk Anal* 15:29-39.

Kupper, LL; Portier, C; Hogan, MD; Yamamoto, E. (1986) The impact of litter effects on dose-response modeling in teratology. *Biometrics* 42:85-98.

Lefkopoulou, M; Moore, D; Ryan, L. (1989) The analysis of multiple binary outcomes: application to rodent teratology experiments. *J Am Stat Assoc* 84:810-815.

Leisenring, W; Ryan, L. (1992) Statistical properties at the NOAEL. *Regul Toxicol Pharmacol* 15:161-171. Liang, KY;

Zeger, SL. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73:13-22. Linhart, H; Zucchini, W.

(1986) *Model Selection*. New York, NY: Wiley.

McCullagh, P; Nelder, JA. (1989) *Generalized linear models*. 2nd ed. London, UK: Chapman and Hall. Mantel, N; Bryan,

WR. (1961) Safety testing of carcinogenic agents. *J Natl Cancer Inst* 27:455-470.

Moolgavkar, SH; Knudson, AG. (1981) Mutation and cancer: a model for human carcinogenesis. *J Natl Cancer Inst* 66:1037–1052.

Murrell, JA; Portier, CJ; Morris, RW. (1998) Characterizing dose-response: I: Critical assessment of the benchmark dose concept. *Risk Anal* 18:13-26.

- NRC (National Research Council). (1983) Risk assessment in the federal government: managing the process. Washington, DC: National Academy Press.
- NRC. (1994) Science and judgment in risk assessment. Washington, DC: National Academy Press.
- Nitcheva, DK; Piegorsch, WW; West, RW; Kodell, RL. (2005) Multiplicity-adjusted inferences in risk assessment: benchmark analysis with quantal response data. *Biometrics* 61(1):277-286.
- Parham, F; Portier, C. (2005) Chapter 14: Benchmark dose approach. In: Edler, L; Kitsos, CP; eds. Recent advances in quantitative methods in cancer and human health risk assessment. Chichester, UK: John Wiley & Sons, Ltd; pp. 239-254.
- Piegorsch, WW; West, RW. (2005) Benchmark analysis: shopping with proper confidence. *Risk Anal* 25(4):913- 920.
- Rao, JNK; Scott, AJ. (1992) A simple method for the analysis of clustered binary data. *Biometrics* 48:577-585. Ross, GJS. (1990) Nonlinear estimation. New York, NY: Springer-Verlag.
- Ryan, LM. (1992a) The use of generalized estimating equations for risk assessment in developmental toxicity. *Risk Anal* 12:439-447.
- Ryan, L. (1992b) Quantitative risk assessment for developmental toxicity. *Biometrics* 48:163-174.
- Ryan, LM; Catalano, PJ; Kimmel, C; Kimmel, G. (1991) Relationship between fetal weight and malformation in developmental toxicity studies. *Teratology* 44:215-223.
- Sand, S. (2005) Dose-response modeling: Evaluation, application, and development of procedures for benchmark dose analysis in health risk assessment of chemical substances [Thesis]. Karolinska Institute, Stockholm, Sweden. Available online at: <http://publications.ki.se/jspui/bitstream/10616/39163/1/thesis.pdf>.
- Sand, S; Filipsson, AF; Victorin, K. (2002) Evaluation of the benchmark dose method for dichotomous data: model dependence and model selection. *Regul Toxicol Pharmacol* 36:184-197.
- Sand S; Victorin, K; Filipsson AF. (2008) The Current State of Knowledge on the Use of the Benchmark Dose Concept in Risk Assessment. *J Appl Toxicol* 28:405-421.
- Seber, GAF; Wild, CJ. (1989) Nonlinear regression. New York, NY: Wiley.
- Simpson, DG; Carroll, RJ; Zhou, H; Guth, DJ. (1996a) Interval censoring and marginal analysis in ordinal regression. *J Agr Biol Environ Stat* 1:354-376.
- Simpson, DG; Carroll, RJ; Zhou, H; Guth, DJ. (1996b) Weighted logistic regression and robust analysis of diverse toxicology data. *Commun Statist Meth* 25:2615-2632.
- Stone, M. (1998) Akaike's criteria. In: Armitage, P; Colton, T; eds. *Encyclopedia of Biostatistics*. New York, NY: Wiley.
- Subramaniam, RP; White, P; Coglianò, VJ. (2006) Comparison of cancer slope factors using different statistical approaches. *Risk Anal* 25(3):825-830.
- Suwazono Y; Sand S; Vahter M; Filipsson AF; Skerfving S; Lidfeldt, J; Akesson, A . (2006) Benchmark dose for

cadmium-induced renal effects in humans. Environ Health Perspect 114:1072-1076.

U.S. EPA (Environmental Protection Agency). (1986) Guidelines for carcinogen risk assessment. Federal Register 51(185):33992–34003. Available online at <http://www.epa.gov/raf/pubalpha.htm>.

U.S. EPA. (1988) Health and environmental effects document for dibromochloromethane. Environmental Criteria and Assessment Office, Office of Health and Environmental Assessment, Cincinnati, OH; ECAO-CIN-GO40. Available from the National Technical Information Service, Springfield, VA.

U.S. EPA. (1991) Guidelines for developmental toxicity risk assessment. Federal Register 56(234):63798-63826. Available online at <http://www.epa.gov/raf/pubalpha.htm>.

U.S. EPA. (1994) Methods for derivation of inhalation reference concentrations and application of inhalation dosimetry. Environmental Criteria and Assessment Office, Office of Health and Environmental Assessment, Cincinnati, OH; EPA/600/8-90/066F. Available online at <http://www.epa.gov/iris/backgrd.html>.

U.S. EPA. (1995a) Use of the benchmark dose approach in health risk assessment. Risk Assessment Forum, Washington, DC; EPA/630/R–94/007. Available online at <http://www.epa.gov/raf/pubalpha.htm>.

U.S. EPA. (1995b). Integrated risk information system (IRIS): Online substance file for carbon disulfide. Available at <http://cfpub.epa.gov/ncea/iris/index.cfm?fuseaction=iris.showSubstanceList>.

U.S. EPA. (1995c). Integrated risk information system (IRIS): Online substance file for methylmercury). Available online at <http://cfpub.epa.gov/ncea/iris/index.cfm?fuseaction=iris.showSubstanceList>.

U.S. EPA. (1996a) Guidelines for reproductive toxicity risk assessment. Federal Register 61(212):56274-56322. Available online at <http://www.epa.gov/raf/pubalpha.htm>.

U.S. EPA. (1996b) Report on the benchmark dose peer consultation workshop. Risk Assessment Forum, Washington, DC; EPA/630/R-96/011. Available online at <http://www.epa.gov/raf/pubworkshop-rpts.htm>.

U.S. EPA. (1998) Guidelines for neurotoxicity risk assessment. Federal Register 63(93):26926-26954. Available online at <http://www.epa.gov/raf/pubalpha.htm>.

U.S. EPA. (2002a) A review of the reference dose and reference concentration processes. Risk Assessment Forum, Washington, DC; EPA/630/P-02/002F. Available online at <http://www.epa.gov/raf/pubalpha.htm>.

U.S. EPA. (2002b) Health assessment of 1,3-butadiene. National Center for Environmental Assessment, Washington, DC; EPA/600/P-98/001F. Available online at <http://cfpub.epa.gov/ncea/iris/index.cfm?fuseaction=iris.showSubstanceList>.

U.S. EPA. (2002c) Organophosphate pesticides: revised cumulative risk assessment. Office of Pesticide Programs, Washington, DC. Available online at <http://www.epa.gov/pesticides/cumulative/rra-op/>.

U.S. EPA. (2002d) Toxicological review of benzene (noncancer effects). National Center for Environmental Assessment, Washington, DC; EPA/635/R-02/001F. Available online at: <http://cfpub.epa.gov/ncea/iris/index.cfm?fuseaction=iris.showSubstanceList>.

- U.S. EPA. (2005a) Guidelines for carcinogen risk assessment. Federal Register 70(66)177650-18717. Available online at: <http://www.epa.gov/raf/pubalpha.htm>.
- U.S. EPA. (2005b) Estimation of cumulative risk from N-methyl carbamate pesticides: preliminary assessment. Health Effects Division, Office of Pesticides Programs, Washington, DC. Available online at <http://www.epa.gov/scipoly/sap/meetings/2005/august/preliminarynmc.pdf>.
- U.S. EPA. (2011) Recommended use of body weight^{3/4} as the default method in derivation of the oral reference dose. Federal Register 76(38) 10591-10592.
Available online at: <http://www.epa.gov/raf/publications/interspecies-extrapolation.htm>
- Van Ryzin, J. (1980) Quantitative risk assessment. *J Occup Med* 22(5):321—326.
- Van Wijngaarden, E; Beck, C; Shamlaye, CF; Cernichiari, E; Davidson, PW; Myers, JM; Clarkson, TW. (2006) Benchmark concentrations for methyl mercury obtained from the 9-year follow-up of the Seychelles Child Development Study. *Neurotoxicology* 27:702-709.
- Venzon, DJ; Moolgavkar, SH. (1988) A method for computing profile-likelihood-based confidence intervals. *Appl Stat* 37:87-94.
- West, RW; Kodell, RL. (1999) A comparison of methods of benchmark-dose estimation for continuous response data. *Risk Anal* 19:453-459.
- Wheeler, WM; Bailer, AJ. (2007) Properties of model-averaged BMDLs: a study of model averaging in dichotomous response risk estimation. *Risk Anal* 27:659-670.
- Wheeler, WM; Bailer, AJ. (2008) Model averaging software for dichotomous dose-response risk estimation. *J Stat Software* 26(5):1-15.
- Williams, DA. (1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* 31:949-952.
- Williams, DA. (1988) Estimation Bias Using the Beta-Binomial Distribution in Teratology. *Biometrics* 44:305-309. Wu, Y;
- Piegorsch, WW; West, RW; Tang, D; Petkewich, MO; Pan, W. (2006) Multiplicity-adjusted inferences in risk assessment: benchmark analysis with continuous response data. *Environ Ecol Stat* 13:125-141.
- Zeger, SL; Liang, KY. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42:121- 130.
- Zhu, Y; Krewski, D; Ross, WH. (1994) Dose-response models for correlated multinomial data from developmental toxicity studies. *Appl Stat* 43:583-598.