

# 食品安全委員会評価技術企画ワーキンググループ

## (第8回) 議事録

1. 日時 平成29年10月20日(金) 10:00～12:30
2. 場所 食品安全委員会中会議室(赤坂パークビル22階)
3. 議事
  - (1) 専門委員等紹介
  - (2) ワーキンググループの運営等について
  - (3) 座長の選出・座長代理の指名
  - (4) 統計学的に妥当なベンチマークドーズ(BMD)法の適用について
  - (5) その他
4. 出席者
  - (専門委員)  
赤堀専門委員、岡田専門委員、川村専門委員、小関専門委員、広瀬専門委員、山田専門委員
  - (専門参考人)  
小坂専門参考人、西浦専門参考人
  - (食品安全委員会)  
佐藤委員長、山添委員、吉田委員、山本委員
  - (事務局)  
川島事務局長、小平事務局次長、吉田評価第一課長、吉岡評価第二課長、池田評価情報分析官、橘評価技術企画推進室長、窪崎課長補佐、山原評価専門官、柳澤係長、石神技術参与、比留間技術参与
5. 配布資料
  - 資料1-1 食品安全委員会専門調査会運営規定
  - 資料1-2 食品安全委員会における調査審議方法等について
  - 資料1-3 「食品安全委員会における調査審議方法等について」に係る確認書について
  - 資料1-4 評価技術企画ワーキンググループの設置について(平成28年3月

29日食品安全委員会決定)

資料2 ベンチマークドーズ (BMD) 法で使用する数理モデルの意味合い的  
由来とモデル選択について

参考資料 評価技術企画ワーキンググループ (第7回) 配布資料

## 6. 議事内容

○吉田評価第一課長 おはようございます。それでは、定刻となりましたので、ただいまから第8回「評価技術企画ワーキンググループ」を開催いたします。

先生方におかれましては、お忙しい中、御出席をいただきまして、誠にありがとうございます。

10月1日付をもちまして、各専門調査会の専門委員の改選が行われました。本日は改選後の最初の会合に当たります。したがって、座長が選出されるまでの間、私、食品安全委員会事務局評価第一課長の吉田が議事を進行させていただきますので、どうぞよろしく願いいたします。

机上に配布いたしました資料の一番上の議事次第を御覧ください。本日でございますけれども、広瀬先生が若干遅れるとお聞きしておりますが、6名の専門委員の先生方が全員御出席いただく予定でございます。

また、専門参考人として、小坂専門参考人、西浦専門参考人にも御出席をいただいております。

初めに佐藤食品安全委員会委員長より挨拶いたします。

○佐藤委員長 おはようございます。この度は、先生方には専門委員への御就任を御快諾いただきまして、大変ありがとうございます。食品安全委員会の委員長としてお礼を申し上げます。実は再任の先生方がほとんどなので、こういう御挨拶を申し上げるのも何かという気もするのですが、任期の切りかえということもございまして、一言御挨拶をさせていただきたいと思っております。座って失礼します。

既に、安倍内閣総理大臣から平成29年10月1日付で食品安全委員会専門委員としての任命書がお手元に届いているかと思っております。専門委員の先生方が所属する専門調査会あるいはワーキンググループについては委員長が指名することになっており、先生方を評価技術企画ワーキンググループに所属する専門委員として指名をさせていただきました。

評価技術企画ワーキンググループは、食品安全にかかわるリスク評価を取り巻く状況の変化を踏まえ、新たな評価方法の潮流に対応する目的で、2016年4月に設置されました。これまで審議いただいた中では、近年のコンピュータ技術の発達に伴って海外のリスク評価機関等で導入が開始されている(Q)SAR及びRead acrossについて御議論をいただき、その経過を報告書にまとめていただきました。報告書で御提言いただきました取り組みに関しては、現在、食品健康影響評価への活用に向けてスピード感をもって推進しておるところであります。

さて、食品安全委員会はリスク評価機関としての独立性と中立性を確保しつつ、科学的な知見に基づき客観的で公正な立場から食品健康影響評価を行うということを掲げております。専門委員の先生方におかれましては、この大原則を御理解の上、それぞれ専門分野の科学的知見に基づき、会議の席で御意見を交わしていただけますようお願いしております。

通常、私どもが考える科学は、精密なデータをもとに正確な解答、真理を求めていくものです。一方、御承知のように、リスク評価は多数の領域の学問が力を合わせて判断をしていく科学、言わばレギュラトリーサイエンスの一部であると考えられております。(Q) SARやRead acrossについて議論していただいた先生方には、リスク評価において、ある時は限られたデータしかない場合でも、完璧さにこだわらず解答を出すことが求められることもあることを御理解いただいているものと思っております。また、新しい評価技術によって、限られた知見であっても科学的で客観性を持った結論を導けるものと思っております。加えて透明性も担保できるものと思っております。このワーキンググループにおいても透明性の高い御議論をしていただければ幸いです。

専門調査会やワーキンググループの審議については原則公開となっております。先生方のこれまでの研究から得た貴重な経験を活かした御発言によって、また、総合的な判断に至るまでの議論を聞くことにより、傍聴の方々にとっても情報共有に資するものと考えております。

食品のリスク評価は、国の内外を問わず強い関心が寄せられていると思います。リスク評価の方法論や技術を議論していただく専門委員としての任務は、食品の安全を支える重要かつ意義深いものと思っております。専門委員の先生方におかれましては、国民の期待に応えるべく、科学的にかつ迅速に所要の任務を遂行すべく、御尽力いただけますようお願い申し上げます。私からの挨拶とさせていただきます。どうもありがとうございます。

○吉田評価第一課長 どうもありがとうございました。

それでは、次に本日の机上配布をしております資料の確認をお願いいたします。

○窪崎課長補佐 本日の資料は議事次第、座席表、専門委員名簿のほか、

資料1-1 「食品安全委員会専門調査会運営規定」。

資料1-2 「食品安全委員会における調査審議方法等について」。

資料1-3 『食品安全委員会における調査審議方法等について』に係る確認書について」。

資料1-4 「評価技術企画ワーキンググループの設置について」。

資料2 「ベンチマークドーズ (BMD) 法で使用する数理モデルの意味合い的由来とモデル選択について」。これらは近日中にホームページに掲載されます。

参考資料として、第7回評価技術企画ワーキンググループにおいて、川村専門委員、広瀬専門委員に御発表いただきました資料をお配りしております。

さらに、本ワーキンググループの専門委員の先生方には、机上配布資料「今後の審議に

おける主な論点（案）」をお配りしております。

配布資料の不足等はありませんでしょうか。過不足等がございましたら、事務局までお申し出いただければと思います。

○吉田評価第一課長 よろしいでしょうか。

それでは、議事を進めさせていただきます。まず議事（１）としまして「専門委員等紹介」でございます。専門委員名簿も御参照を願えればと思います。今回改選がございましたのは小坂先生、川村先生、小関先生、広瀬先生の４名の先生方でございますが、専門委員若しくは専門参考人として引き続き審議に御参加いただくこととしております。今回改選がございませんでした先生方もあわせまして、本日御出席の先生方を改めて御紹介させていただきます。

赤堀専門委員でございます。

岡田専門委員でございます。

川村専門委員でございます。

小関専門委員でございます。

広瀬専門委員でございます。

山田専門委員でございます。

また、本日は小坂専門参考人、西浦専門参考人にお越しいただいております。

食品安全委員会からは冒頭御挨拶をいただきました佐藤委員長を初め、本ワーキンググループの担当委員であります山添委員、吉田委員、山本委員が出席しております。

続きまして、事務局を御紹介いたします。

川島事務局長でございます。

小平事務局次長でございます。

吉岡評価第二課長でございます。

池田評価情報分析官でございます。

橋評価技術企画推進室長でございます。

窪崎課長補佐でございます。10月１日付で井上に代わりまして着任しております。

山原評価専門官でございます。

柳澤係長でございます。

後ろのほうですが、比留間技術参与でございます。

石神技術参与でございます。

以上、どうぞよろしくお願いたします。

続きまして、次の議事（２）に移らせていただきます。「ワーキンググループの運営等について」でございます。お手元の資料１－１、１－２、１－４を御用意いただければと思います。既に皆様は御存じの内容でございますので、要点のみ簡潔に説明をさせていただきます。

まず資料１－１「食品安全委員会専門調査会運営規定」でございます。これは調査会、

ワーキンググループの各種運営等に係る規定を定めたものでございます。

1 ページの裏側でございますが、ワーキンググループにつきましては第6条にその規定がございまして、委員会に諮ってワーキンググループを置くことができるとされているところでございますが、本ワーキンググループにつきましては資料1-4にありますとおり、平成28年3月29日の食品安全委員会決定に基づき設置されてございます。所掌事務でございますが、真ん中あたりの2にありますとおり、「WGは、食品健康影響評価の実施に用いる新たな技術の企画等に関する事項について調査審議を行う」という形になっているものでございます。

資料1-1にお戻りいただきまして、ワーキンググループの各種運営につきましては、第6条第3項におきまして、いわゆる調査会の運営規定、すなわち第2条～第5条までの規定を準用するという形になっております。具体的には、1-1の1枚目の表のほうに戻っていただきまして、例えば第2条の第3項におきましては、専門調査会をワーキンググループで読みかえますが、座長を置き、当該専門調査会に属する専門委員の互選により選出するでありますとか、あるいは第5項ですが、座長に事故がある時は当該ワーキンググループに属する専門委員のうちから座長があらかじめ指名する者がその職務を代理するといったような規定がございまして、さらに第4条ですが、座長はワーキンググループの会議を招集し、その議長となる。こういったような規定が定められているところでございます。

引き続きまして、資料1-2「食品安全委員会における調査審議方法等について」でございます。「1 基本的な考え方」にございまして、これはいわゆる中立公正な評価の確保の観点から、委員会等における調査審議などの方法を定めるものでございまして、いわゆる利益相反に関する規定でございます。具体的には2のところの内容が定められていますが、(1)にございまして、委員または専門委員が次に掲げる場合に該当する時は一部、当該委員等の有する科学的知見が委員会等の調査審議に不可欠であると認める場合を除き、調査審議等に参加させないものとするという形になってございます。具体的には①でございますが、調査審議等の対象となる企業申請品目の申請企業若しくはその関連企業または同業他社、特定企業と称しますが、特定企業から過去3年間の各年において新たに取得した金品等の金額が別表に掲げるような一定の金額を超えるような場合でありますとか、あるいは④でございますが、特定企業からの依頼により当該調査審議等の対象品目の申請資料等の作成に協力した場合、こういったような項目など6項目が定められているところでございます。

めくっていただきまして、裏側でございますが、このことを確認するため、(2)でございますけれども、委員等は別紙に掲げるような確認書を任命された日以降、初めて開催される委員会等に合わせて提出する。さらには(3)でございますが、先ほど申し上げた(1)に掲げる場合に該当するという事になった場合にも、その都度提出するという形になってございまして、(4)でございますけれども、それを委員会などの都度、当該確認書に記載された事実の確認を行うということになっているものでございます。

こういった内容になってございますので、御説明いたしました内容について再度御確認いただきまして、また御留意いただきながら専門委員をお務めいただければと存じます。運営等については以上でございます。特に御質問等はよろいでしょうか。

それでは、よろしければ、続きまして、議事（3）に移らせていただきます。議事（3）は「座長の選出・座長代理の指名」でございます。先ほど御説明いたしました食品安全委員会専門調査会運営規定第6条第3項において準用する第2条第3項の規定によりまして、ワーキンググループに座長を置き、当該ワーキンググループに属する専門委員の互選により選任することとされております。どなたか御推薦はございませんでしょうか。

岡田先生。

○岡田専門委員 このワーキンググループの発足以来、非常にうまくリードをしていただいていると思いますので、お隣にお座りの川村先生に引き続き座長をお願いしてはどうかと思います。先生には御苦労さまでございますけれども。

○吉田評価第一課長 ただいま川村専門委員を座長にという御推薦がございました。いかがでございましょう。御賛同される方は拍手をいただければと思います。

（拍手）

○吉田評価第一課長 どうもありがとうございました。それでは、御賛同いただきましたので、座長に川村先生が互選されました。今、座長の札を置かせていただいております。

それでは、川村座長から一言御挨拶をお願いいたします。

○川村座長 このたびは座長に御推薦いただきまして、ありがとうございます。微力ながら、このワーキンググループを取りまとめる係をさせていただきたいと思います。座って御挨拶をさせていただきます。

この食品安全委員会は科学に基づいて最適な規制のための指標値を出していくという、そういう使命を帯びている非常にアカデミックな機関だと思っておりますが、その中で評価の方法として今までは各専門調査会がそれぞれ最適と思う手法を用いて評価してまいりました。しかしながら、やはり共通なものも多々ありますし、そもそもその評価の技術のコンセプトと背景にあるものの考え方というものをきちんと打ち立てておくという必要もございます。そこで2年前くらいでしたか、この評価技術のワーキンググループを立ち上げるという話がありまして、組織され、実際に活動を始めて1年余りたったわけでありませけれども、その中でICTの技術を活用するなど、最近の科学のインフラとなる状況が大きく変化してまいりました。そういうことで評価の技術もまた進歩、進展させる必要もありますし、海外の評価機関も新しい手法を積極的に取り入れる方向にございます。

そこで私どもとしまして、専門調査会に横串を刺すといえますか、横断的に評価の技術について、きちんと見直し、また新しい必要な方法を提唱すると。そのためにきちんとした科学的な議論をするという機会も必要かと思っておりますので、この評価技術ワーキンググループが立ち上げられ、そして実際に活動を始めたわけでありませ。まだまだ緒についたばかりでよちよち歩きではありますけれども、第1弾の報告は出させていただきますの

で、これから第2弾に踏み込みまして、またスパイラルのように最初に戻るところもあると思いますが、科学に基づいた評価方法をきちんと打ち立ててまいりたいと思います。

このワーキンググループは専門を異にする多様な委員で構成されておりますし、特に必要などころにつきましては専門参考人をお招きして深い議論をしてまいります。それぞれの持ち味を生かして積極的な議論をするというのがこのワーキンググループの一つの特性ですので、忌憚のない意見交換、率直な意見を皆さんで出し合って、いい検討をしてまいりたいと思います。これからはばらく難しい議論も続くかと思いますが、何とぞよろしくお願い申し上げます。

以上です。

○吉田評価第一課長 先生、どうもありがとうございました。

それでは、次に食品安全委員会専門調査会運営規定の第6条第3項において準用する第2条第5項におきまして、座長に事故がある時は当該ワーキンググループに属する専門委員のうち、座長があらかじめ指名する者がその職務を代理するという旨が規定されておりますので、座長代理の指名をお願いしたいと思います。また、これ以降の議事の進行については座長をお願いいたします。

○川村座長 課長さん、どうもありがとうございました。それでは、議事の進行を引き継がさせていただきます。

ただいま事務局から御説明があった座長代理の指名についてですが、私からは広瀬専門委員をお願いしたいと思います。いかがでしょうか。

○広瀬専門委員 私ではよければ、お引き受けしたいと思います。

○川村座長 信頼する片腕と言ったら語弊があるかもしれませんが、従前もこのワーキンググループ全体を支え、引っ張っていただいたので、引き続き座長代理をお願いしたいと思います。では、一言御挨拶をお願いします。

○広瀬専門委員 私ではよければというか、代理の代理が必要な気もしてしまいますけれども、川村先生の方針を引き受けて、IT技術が進んでいますけれども、やはり異なる分野の意見を用いるということも依然必要だと思いますので、そういう取りまとめのサポートをしていきたいです。よろしくをお願いいたします。

○川村座長 ありがとうございます。私の至らないところをしっかりと支えていただければありがたいと思います。

それでは、次の議事に入らせていただきます。資料1-2に基づきまして、必要となる専門委員の調査審議等への参加に関する事項について説明を行ってください。

○窪崎課長補佐 本日の議事について、専門委員の先生方から御提出いただきました確認書を確認したところ、平成15年10月2日委員会決定2の(1)に規定する調査審議等に参加しないこととなる事由に該当する専門委員はいらっしゃいません。

○川村座長 ありがとうございます。御提出いただいた確認書の記載に特に変更はございませんか。よろしいですね。

(「はい」と声あり)

○川村座長 ありがとうございます。

それでは、議事(4)に移らせていただきます。今日の主題でありますけれども、「統計学的に妥当なベンチマークドーズ(BMD)法の適用について」でございます。前回のワーキンググループでは、私と広瀬専門委員からベンチマークドーズ法の現状と課題について情報提供をいたしました。ベンチマークドーズを採用するか否かという問題はもちろんよく考えなければなりませんし、採用したとしても例えばモデルの選択ですとか、あるいはRestrictionの問題、BMRの設定、モデルそのものを統合するような話も含めまして、実際の取扱いに向けて合理性と実現性、そしてフィージビリティを含めたことを今後ワーキンググループで検討する必要性が明らかになってまいりました。

今回は特にBMD法を採用した後に生ずる課題について、統計学的にどのように議論を進めていけばよいかということで先生方と情報を共有したいと考えています。まずはBMDに使われている数理モデルそのものの統計学的背景の理解を深めまして、そして、次にモデルの選択やモデルの平均化について、どのように考えるべきか。そして、最後にBMRの設定やRestrictionをどう考えていくか順を追って考えてまいりたいと思います。

本日は専門参考人として北海道大学大学院の西浦先生をお呼びしております。先生には長時間お話をしていただくこととなりますが、どうぞよろしくお願いをいたします。進行の都合で、そのお話の中で2回ほど質疑応答を挟みまして、皆さんの理解をその都度そろえた上で先に進んでまいりたいと思います。それでは、西浦先生から資料に基づきまして、御発表をお願いしたいと思います。準備の間、西浦先生の御略歴を紹介いたします。

西浦先生は宮崎医科大学医学部医学科を御卒業された後、広島大学大学院におきまして、保健学の博士号を取得されておられます。英国、ドイツ、オランダ、香港で研究をなされ、平成25年に東京大学大学院医学系研究科の准教授として、そして、現在は北海道大学大学院医学系研究科の教授として御活躍でございます。博士研究員の時代にBMDソフトの一つでありますPROASTの改良に携わられておられまして、BMD法の統計学的な理論にも精通されておられます。また、英国、ドイツを初め、諸外国で研究もなされていて、それぞれの国や国際的な情勢にも明るくていらっしゃいます。御専門は理論疫学、生物統計学、感染症疫学でございます。厚生労働省のエイズ動向委員会の委員も務めておられます。

それでは、西浦先生、よろしくお願ひいたします。

○西浦専門参考人 御紹介をありがとうございました。初めまして、西浦と申します。

(PP)

今、私の簡単な略歴を御紹介いただいたのですが、オランダでPROASTの改良に携わったという御紹介はいただいているのですが、当時、ポスドクでユトレヒト大学の所属時にアルバイトで雇われて既に出来上がったモデルのコードを書いたり、バグ取りと言って、エラーを取り除く作業を短期的にしたことがある、というのが実際のところで、全くもって大した経験ではありません。ただし、今までずっと数理モデルを専門に研究に取り

組んできました。

私はこうやって数理モデルを専門に感染症疫学を名乗っているのですが、他の省庁で何らかの審議会だったり分科会だったりにお呼びいただいたときは「モデルの詳細は省いてください」と言われます。「モデルの意味は余り重視しません」という趣旨なのですが、その「結果がいかにか妥当で、それが国際誌にピアレビューされているかを基点に議論が進みます」と言われるのです。国の機関でこうやってお呼びいただいて、「モデルの意味を知りたい」というリクエストを受けるのは実は初めてで、本当にうれしく思っています。ずっとこの件ばかりをマニアに研究してきた自分としては、その話をいつもしたくてしようがなかったのですが、それをじっくりと学びたいということでリクエストを初めていただきました。わからないことがあれば、いつでもお尋ねいただいて、モデルの中身であれば私のできるレスポンスは差し上げますので、どうぞよろしくお願ひします。

(PP)

今日のコンテンツでカバーする部分がこちらに箇条書きしている部分になります。特に①と②が主体になると思います。BMD法でいろいろなモデルが使われますけれども、数理モデルを使うということが初めての方や、実験現場の方を対象に、いきなり降ってわいてきたというのがBMD法の衝撃的な出現だったのですが、今までNOAEL、LOAELだけでやっていた評価でも、数理モデルを用いざるを得なくなりつつあります。つまり、それぞれの数理モデルについて勉強する必要性が生じてきたのですけれども、一つ一つのモデルに一定のメカニスティックな由来があります。このモデルはこうこうこういう意味で使われていますよというのがあるので、その入り口の部分を皆さんと共有させていただきたいと思っています。その上でどういうふうにモデルが選択されていて、平均化がどのように行われているのかという点に関して、基本的概念の部分に集中して今日は共有させていただき、というのをゴールとして設定させていただいています。

(PP)

早速、中の話を始めさせていただきます。BMD法のモデルの成り立ちとその内容についてというところからご紹介します。

(PP)

BMD法というのは、NOAEL、LOAELとかの実験のドーズを決めて、反応が起こる最初のところとか、あるいは反応が起こらない最初のところを決めるのではなくて、観察データに数理モデルを適合することを活用するものです。そのために、数理モデルが多数登場するのは皆さんも御存じのとおりなのですけれども、そのときに使う関数をまず知っておいていただかないといけません。最初の入り口であえてこのスライドを事務局とも相談の上で用意をしています。

確率密度関数というものと累積分布関数という2つの種類の分布というのがあることを知っていただきたいのですが、私たちが普段の生活で分布と言っているものは、本当は確率密度関数に相当するものです。横軸がいつもドーズ、用量だと考えてやっている

のですけれども、それは他の現象だったら、他のものでも構わないのですが、何らかの入力があったときに反応の頻度を縦軸にプロットしているというものです。それがどれくらいの頻度で起こるかなというのが本来的な $f(x)$ のほうに相当します。それが累積でどれくらいまで起こったのかというのが本来的な意味での累積分布関数で、それがキャピタルFと言って、大文字のFを使っている関数です。

各Xという入力を与えられたときの反応の確率が累積で足されていくので、必ず1まで上がっていくような分布になっているというのが特徴です。数多くの累積分布関数の数式があって、それがドーズレスポンスの反応に当てはめられているというのが、まず入り口の話です。本来的には累積分布関数ではなくて確率密度的な話を扱うのですが、ドーズレスポンスに関しても数理をやっている人たちは横軸がドーズではなくて、ドーズの変化量、ドーズを少しずつ切った増え方と反応の出方の変化率を縦軸にとると確率密度みたいなのが出るのですけれども、今それは考えないようにしたいと思います。

(PP)

用量反応曲線にその累積分布関数を当てはめて検討していくのがドーズレスポンスカーブです。横軸は今この図にはログをとっているようなドーズを出していますけれども、そのままアブソリュートバリュー（絶対値）でも大丈夫ですが、特定のドーズが被験者に投与されて何らかの反応、発がん毒性など色々な生体反応とかがありますけれども、その頻度を縦軸で定量するということによって、こういうカーブが描かれるのですが、ある人は低いドーズで反応するのだけれども、他の人は極めて高いドーズで反応するので、反応する率がS字曲線を描くことが多いです。カーブがねじ曲がってしまっているから、S字曲線に対応する分布を適合する必要が生じています。本来的にはものすごく理想的な環境だと、ここまでのドーズだと反応が起こらなくて、ここからだとは必ず起こりますとかいうような反応があるとこんなカーブは要らないのですけれども、仕方がないから、観察パターンに対応するような反応曲線をサイエンスとして分析している、というのが用量反応曲線のベースにあるものです。

これから御一緒に見ていただくS字曲線というのは、ポピュラーだから使われているだけです。S字曲線の特定の関数が海外も含めて選ばれているのですけれども、今まで慣習的に使われているから選ばれているだけで、実際には直線でもいいわけです。閾値があって、そこから直線で上がるような折れ線のようなドーズレスポンスカーブでも本来的には構わないのですけれども、いろいろな恣意性などが混入するので、歴史的に今まで収められた数少ない分布が使われているというのが現状です。

(PP)

それと、事前に一つリクエストをいただいていたのですけれども、望ましい形というのはどうなのかという質問です。それが分かると余り苦労はしないのですが、ドーズレスポンスの反応のデータが得られたときに、どういう分布が一番よくて、どういう分布が一番よくないのかということなのですが、それは観察データのパターンによります。

確率密度で、どこで反応が起こりやすいのかというのを図にしたのは、ここで御覧いただいている図になります。左側だと右にスキュードだと言います。右に裾が長いというような言い方なのですけれども、右側の図が左側にスキュードだという言い方をしますが、ちょっと偏ったドーズで反応が起こりやすい場合の観察データに対応していると、こういうふうにゆがむような分布でなければならないということです。理想的には1点で反応が起こらないのだったら、真ん中にあるように左右対称な分布できれいに反応が起こってくれるというのが理想なのですけれども、現象によっては左右対称で起こるといふ蓋然性がないと、左右対称の分布をいつも無批判に当てはめることは危険です。

こういったときの分布の特徴というのをモーメントと呼びますけれども、その分布をたくさん微分していったときのことを言うのですが、ゆがみ方だけの話をさっきのSkewnessと言います。右に裾が長いとか左に裾が長いというのがSkewnessというもので、それが縦にどれだけ長いような分布になるのかというのがKurtosisと言うのですけれども、当てはまりがどれくらいいいのかを評価するために、特定の用量反応に関する観察データを蓄積して研究を積み重ね、これら特徴が少しずつ理解されてきました。

どんな分布が理想的かと言うと、もう最初に答えを書いていますけれども、箇条書きの2番目のところを御覧いただきたいのですが、パラメータ数が少ないけれども、フレキシブルでデータの当てはまりがよいようなモデルがよくて、さらにできるならドーズレスポンスとその現象に関してこれからお話ししますが、その対応があるといい、と考えられます。この委員会は私の理解では、評価技術を皆さんで検討されて、できれば正しいモデルの選択をするためのサイエンティフィックなベースを皆さんで議論されていると理解しているのですけれども、勝手な私たちの思いですが、ハイドーズの現象の観察データを利用してロードーズのところを何とか予測したいというのが、自分勝手な用量反応関係に対する期待です。ロードーズを含めて、現象が説明可能なようなモデルがあるとなおさらよい、ということまでが、私たちの理想としているものである、という事実が前提です。

(PP)

それで、それぞれの分布がどんな形かという話を踏まえて、BMD法の設定が行われるのですけれども、これからお話をする分布に関してどんな特徴を持っているのかという点について説明します。実は、この点は簡単に議論してしまふことができます。このスライドは引用のところに正直に書いていますけれども、ウィキペディアから入手可能です。日本語版のウィキペディアはパラメータの範囲とかを検討していることが少ないのですけれども、英語版のウィキペディアは累積分布関数も含めて、形状面に関して質的に捉えておかなければならないパターンは体系的・包括的にさらってくれています。

上の2つがロジスティックあるいは対数ロジスティック分布というもので、プロビットというのは正規分布と一緒になのですけれども、ロジスティックとプロビットは似ているので、対数プロビットははしょっています。下3つというのがQuantal-Linearと言われるものと、ガンマ分布とワイブル分布と呼ばれるそれぞれの分布ですけれども、簡単にこれを

描写してしまいうことができて、ロジスティックとプロビットというのは、左右対称の分布です。左右対称のときに、きれいな形になっているというのがこの図を見て分かりますか。トキシコロジーで言うとTD50ですかね。TD50のところを中心に右の形と左の形が折り返すと対称になっているというような、そういうのがロジスティックとかプロビットの分布の特徴です。

対数ロジスティックとか対数プロビットと言われるのは、対数正規分布に似ているのですけれども、右に裾が長い、ライトスキュードとさっき言っていたような分布です。低いドーズのほうでピークが早めに来ていて、でも、右側で少しずつだらだらと反応する率が上がっていきますよと言われるようなものです。ロジスティックと対数ロジスティックの違い、というのはそれぞれ現象論的に当てはまりやすい単純なモデルだから用いられていて、用量反応曲線では当てはまりのよさが違うだけ、と理解しています。それと一線を画すのが下の3つです。

Quantal-Linearというのは指数分布と一緒になのですが、あるいはシングルヒットモデルと呼ばれるものですが、一つのランダムに起こる現象がドーズの関数だったときにドーズをふやしていくと、どんな形になるかなということなのですが、上に凸な分布にならざるを得ないことが知られています。S字ではないというものです。そんなランダムなイベントが何回も起こるように仕掛けたものというのがガンマ分布です。

Quantal-Linear、あるいは指数分布、シングルヒットというのは1回のランダムな何らかのDNAの損傷が起こるような現象をモデル化するのに対応しており、何回も同じようなランダムなDNAの損傷が重なって現象が起こるときはガンマみたいなのがよくて、それと指数分布が数学的には何回も折り畳まれたものがガンマ分布と言われます。そうするとS字曲線の形がキャプチャーされるということです。この辺の詳しい解釈はこれからまた話していきます。さらにワイブルみたいに現象自体に記憶があるものだと、S字曲線の形状よりキャプチャーしやすいことが知られています。

これら分布と、プラスアルファでマルチステージというものもありますけれども、基本的なスタンダードディストリビューションと呼ばれるものが多いのですが、それらを使ってドーズレスポンスの曲線をモデル化するというのがBMD法でやられていることです。

(PP)

こちらは広瀬先生の前回のスライドを使わせていただきましたけれども、BMD法で用いられるモデルと数式がどんなもので、不明なパラメータが幾つありますということを表にしたものです。さらに、コンストレイントと言ってパラメータの制約に関して右端の列に書いています。今日はここに8つのモデルがあるのですが、8つのモデルに関してのメカニズムがどういう解釈ができて、それがどうしてドーズレスポンスで使えてというあたりの入り口の部分までを御紹介できればと思っています。

ちょっと戻りますが、今見ていただいたのですが、パラメータの数というのがプロビットのところは消えていますけれども、2ですね。2、2、あとは3、3、3、3、

3で最後のところだけ4というのが続いているのがお気づきになると思います。ロジスティックとプロビットのモデル以外の下の6つのモデルを見ていただきたいのですが、そこで $\gamma$ というパラメータがあります。これはロジスティックだったり、ワイブルだったりガンマだったり、本来の累積分布関数を書くときには使わないパラメータです。だから、本来的にはワイブルとかガンマというのは2パラメータの分布なのですが、ここにいつも $\gamma$ というパラメータがついているというのが最初に気づいていただきたいのです。

(PP)

この点が、最初に気づいていただきたいBMDモデルの仕掛けです。ロジスティックとプロビットに関しては左裾と言って、ドーズが限りなくゼロに近いときに、ベースラインで何もエクスポージャがなかったときにどれくらいリスクで反応が起こっているかという話をするときにゼロの値をとりません。他方、ロジスティックとプロビット以外の分布は全部ゼロをとります。もともと数理的にゼロから始まる関数なのですが、ロジスティックはゼロドーズでの反応がゼロに近い数値しか取りませんし、プロビットに関してはドーズを理論上のマイナス無限遠まで減らさないと反応ゼロにはならないからベースラインを設定できないのです。そのため、2つの分布ではこのままにしているのですが、それ以外の6つの分布というのは左裾が反応ゼロになる分布です。

だから、BMD法でその観察データにフィットするときに何もエクスポージャがなかったときの発がんのリスクとかをモデル化するためにさっきの $\gamma$ を使っています。左テールのリスク最小値というのが $\gamma$ です。その $\gamma$ は、本当はモデル的には推定はするけれども、捨ててしまって、残った2つのパラメータでBMD10だとか、いわゆる特定のパーセントイルポイントを使ってBMDLとかの議論をすることになります。 $\gamma$ という分布は本来的にはそれぞれの分布には備わっていないパラメータなのですが、観察データにそのまま適用できるようにしたいから、BMDモデルには備わっているということです。本来的なパラメータ数はこの表に書いているとおりになります。指数分布だったら1個しか本当はないですし、ワイブルだったら2個、ガンマ分布も2個なのですが、BMD法に入れるときにはそれぞれプラスでパラメータ1個分を付加して推定しているということです。

実はこれは今後こういう分布を改善する上でも意外に重要になります。 $\gamma$ の情報は実はハイドーズしかないような観察データにモデルを適用したときは実は余り十分に識別できなくて、 $\gamma$ の信頼区間が大きくなって推定されることもあります。 $\gamma$ の信頼区間はBMD10に直接影響しないのではないかと思われるかもしれませんが、他のパラメータの信頼区間に影響を与えるので、実はBMDLとかも $\gamma$ の識別性が低い場合には信頼区間が $\gamma$ のせいで大きくなってしまいうことがあるといのが結構意外な余り知られていない問題です。BMDLとは、BMD10の信頼区間から決まるのですが、分位の信頼区間をモデルにとるので、それを決めるときにはパラメータの信頼区間というのがすごく重要になるのです。意外にこの $\gamma$ に気を遣っていない一方で、モデル内に入ってしまったことを理由に、BMDLがやたらと低い値になってしまうこともあるかも知れないので、そのような設定がされて

いるのだなというのを理解した上で、次のそれぞれの本来のモデルの意味を見ていかないといけないというところに気づかないといけません。あえてこれは最初にお話をさせていただきました。

(PP)

それで、今日お話しさせていただく話の統計モデルというのは、定量的リスクアセスメントで使用するモデルのうちでも「外挿モデル」と呼ばれ、今日お話をさせていただくBMD法で使うものは合計で8つです。その他にもたくさんモデルはあるのですが、慣習で使われているものがそれら8つです。国立衛研の先生方がたくさんいらっしゃる中でこの話を私がするというのもおこがましいのですが、外挿モデルは確率分布に基づくものと発病メカニズムに基づくものに大別できて、きれいに分けられるものではないです。どちらにも分類できるようなものもあります。

確率分布と言って、もともとの数学的関数に基づくものの典型例がプロビットとかロジスティックです。発病のメカニズムに基づくものの典型例がマルチステージです。ワイブルはどちらとも捉えられる。あるいは用量詳細で例えばばく露時間、タイムドーズというようなモデルもありますけれども、そういったものを考慮したモデルもありますし、より生物学的なメカニズムを微分方程式で書いたようなモデルだってあります。あるいはPB-PKモデルとあるのですが、体内で吸収をされて、どのように分布して代謝されるまでどうなるかというのを直接にモデル化した上で、ドーズレスポンスを利用したリスク分析に役立っているというような研究も、リスク評価研究で実施されたことがあると記憶しています。

(PP)

今日お話をするのはその8つで、一気に深められるところの限界くらいの数です。8つに関して、より詳しく知りたければ、ぜひお話をいただければと思うのですが、今日は表層的な部分しか御紹介できませんが、より詳しくは最も詳しく数理的記述が与えられている文献というのがこちらです。Johnson & Kotzというのがもともと書いていたもので、今はBalakrishnanという人が加わってリビジョンをしているのです。『CONTINUOUS UNIVARIATE DISTRIBUTIONS』の本でボリューム1のところほとんど最初のもので載っているのですが、ガンマ分布とかでも一般化すると、ガンマ分布は、本当は2パラメータではなくて5パラメータのものがあるのですが、マニアな話をし出すといっぱいあって、その由来とか数理的な特性に関して一気に書いているようなものです。楽しいのですが、ドーズの話をするときには余りこれ以上言うと皆さんが離れていくので、この程度に留めて書籍のみご紹介します。

(PP)

個々の分布がどんな分布なのかというのをご紹介させていただきますけれども、そのために分かりやすい順番に変えさせていただいています。あえてBMD法のマニュアルとかに載っている分布の順ではなくて、ワイブル分布やガンマ分布というわかりやすい順番で今日は話させていただきます。

最初はワイブルです。技術者だったワイブルさんが考えた分布です。1939年にできた分布で、歴史としてはとても新しい分布です。故障の分布として名高いことが知られています。ここから眠くなる可能性があるので、どうしても押さえていただきたいものは、下線を引いていますので、そこを押さえていただくとテークホームするものがあります。

ワイブルと言うと故障の分布です。発端の話を書かせていただきましたけれども、昔、産業革命が終わって、これから軍需拡大という話になったときに、鎖の強さを皆さんが評価したかったのです。鎖の材料の破壊強度をデータとして分析することをヨーロッパで行っていたのですが、正規分布に破壊強度のデータがうまくフィットしないことが明らかになりました。局所的な強度の定義というのは1個の鎖の輪の強さで見たら正規分布をしているのに、うまく鎖がいつ、どれくらいの強度で引っ張ったら壊れるのかという現象が正規分布しないということが発端で、技術者の間でどうしてなのか、という疑問が沸き起こりました。理由として最もサウンドだったのがワイブルの考えの原点なのですが、破壊が起こるのは鎖の全部が壊れるのではなくて、鎖の最も弱い一つというのを見つけて発生しますという事実が鍵でした。最も弱いところの1カ所がやられると、ばんと鎖は外れてしまうので、数理的にもそのメカニズムに集中しましょうということです。1つの鎖の強度は構成物であるリング一つ一つの強度の平均値とか分散でもともと正規分布するルールに従わない。そうではなくて、最も弱いリングがどこかにあって、その特性で故障の確率決まりますよということで、ワイブルがその分布をモデル化したのですが、ここでちょっとだけ数式を見て頑張ってください。

(PP)

一個一個のリングがn個できている鎖があったとします。それを両側から引っ張るのです。加重xをかけてリングの1つが切れるときの確率が「 $1 - \exp(-\phi(x))$ 」みたいな形でモデル化できますということだったのですが、鎖のどこかが加重xで切断される確率を計算するとうまくいくのではないかとことをワイブルは見つけたのです。「 $1 - \exp(-\phi(x))$ 」というのがモデル化のミソで、「 $-\phi(x)$ 」の中身は正規分布です。鎖がどこか1カ所で壊れるので、それは確率的に起こっていて、「 $1 - (1 - G(x))^n$ 」とありますけれども、n個のリングがあるので、n回こういうトライアルをやっているというモデル化をしたら、この現象はうまく説明できますよということが言われていて、すると、その数式を変更すると、もともとの式に似た感じになります。エクスポネンシャルをn乗とるとエクスポネンシャルの中にnが入れるという、高校数学で学ぶルールを使っています。「 $\phi(x)$ 」の部分を経験的にうまく書きかえると、今日使われているワイブル分布になりますよということを証明したのが1939年の研究です。

(PP)

なので、ワイブル分布というのは一つのランダムイベントが起こるのですが、全体にランダムに何か加重があって、鎖一つ一つが独立な事象として負荷がかかった時にどこか1個が壊れるような確率と言われます。n個のリングがあるというような鎖があった

ときに、何回も繰り返せるわけです。そのうちのどこかでイベントがあるということで起こるものなので、そういう壊れるという過程に記憶があることをモデル化できるというのが特徴です。

ここがまあまあ重要になるのですけれども、例えば一段階低いドーズでは何も起こらなかったけれども、一段階ドーズふやすと反応が起こりやすいというときに、前のドーズで反応が起こらなかったという事実を加味した上で次のイベントの発生確率が決まるような場合のモデル化が可能です。普通のランダムイベントだと、単に新しい確率の掛け算になっている場合があります。そういうモデルが他にもあるのですけれども、そうではなくて、プロセスに記憶がある場合に向いているということです。言いかえるとわかりやすいのですけれども、古い車ほど壊れやすいことは皆さん御存じのとおりですね。古い車というのは今まで壊れなかったという、きのうまで私の車は壊れなかったというプロセスを覚えた上で、では、今日はどうなのという、その壊れる確率はどんどん上がっていくわけです。その事実をモデル化できるのがワイブルということです。プロセスに記憶があるということです。

(PP)

一方でガンマ分布というのは、ワイブル分布に少し対比すると分かりやすいです。指数分布というのがこの後も出てきますけれども、単なるランダムな1つのイベントが記憶とか何も関係なくて、1回のランダムイベントに関して「いつ起こるか」という話を確率分布にすると指数分布になるのですけれども、そういった分布で済む話を繰り返したらどうなるのかというのを数理統計学者が考えた結果がガンマ分布です。発端が分かりにくいのですけれども、分かりやすい例題をあえてつくりました。

例えば私がどこかのバーでバーテンダーをしてアルバイトをしています。平均で年間3個グラスを割ってしまいます。ある日そのお店がグラスを1ダース(12個)買ったのですけれども、そのときに私が1ダース全部を割ってしまうまでにかかる時間はどんな分布に従うでしょうということです。単純な普通の分布とかを考えない割り算の計算だったら、4年ですね。そうではなくて、統計の人はその分布を考えるために一つ一つを割るというイベントをモデル化するのは、1つのコップが割れるイベントはランダムであり、コップを割るというイベントはきのうまでの話と関係ないですね。車の故障と比べていただきたいのですけれども、単に手が滑って割れるというイベントなので、メモリーがありません。そうするとイベント発生までの時間は1回のコップが割れるプロセスですので指数分布に従います。指数分布で2個が割れるときはランダムに次に1個が割れるのと2個が割れるのとの時間の和になります。同じようなリスクで起こっているとすると、指数分布の2つが和をとったものになります。統計学的にはそれを畳み込みと呼びます。2つの分布の変数を足して積分をただけのものです。式で言うと指数分布の代わりにこんな形のものになります。問題はn個が割れるまではどうなるのかということ、n回畳み込めばいいということなのですけれども、畳み込んでいるとガンマ分布が出てくるというのが、このガ

ンマ分布が生み出された背景です。

(PP)

どんなことに使われるかと言うと、ランダムなイベントが起こって、いつまで待てば次のイベントが起こるかという回数が多いようなイベントに関してモデル化をするのに向いていて、自然現象でよく活用されます。

(PP)

今回はトキシコロジーの話なので、次のスライドでもうちょっと分かりやすい話を用意していますけれども、洪水がポアソン過程に従うというのが知られています。洪水が1回起こるとするのはグラスを1回滑らせて割ってしまうのに似ているようなランダムなイベントだと知られているのですが、それが何回起こりますかとかいうような分布に役立ちますし、あるいは数理的な特性があって、ポアソン密度のランダムな変動に利用できる、という特性があります。ここは実践的には余り重要ではないかも知れませんが、2、3、4というのは数学的な理由ですけれども、ガンマ分布がいろいろなモデルで利用されるというのには、実はポアソン分布とガンマ分布を結合させるとネガティブバイノミアルと言って負の二項分布になるとか、数理的な特性でいろいろ便利なことが知られていて使われています。

(PP)

よりリスク評価という立場で見ると、ガンマ分布に関しては別名がついていて、マルチヒットモデルと呼ばれています。これは1981年の『Environmental Health Perspectives』という雑誌に載っていた最初のマルチヒットのフォーマルな定式化の部分に相当するのですけれども、単にガンマ分布の累積分布を書いているだけなのですが、ドーズレスポンスの現象で、対象としているイベントはどんなものかと言うと、観察イベントが起こるまでにヒットが何回も起こらないといけないようなものですよということですが、ヒットというのは具体的に言うと例えばDNAの損傷とかです。ランダムにDNAの損傷が何回か起こったら特定の発がんというイベントが起こる。ドーズが  $d$  で与えられると、ガンマ分布の  $\theta$  というパラメータを使って  $\theta d$  回のヒットですとかいうような回数が計算できますし、 $k$  回のヒットが起こる確率は  $k-1$  の階乗ですとか、数理的にも色々な便利な特性があるということが知られていて、マルチヒットという形で使われたのは実は新しくて、1980年代からです。理論的な起源は1953年のNordingという人が『Communications in Statistics』に遡りますが、スライドの引用文献を読んでいただくとマルチヒットのヒストリーが書いてあります。

Nordingが記述的に発がん年齢をもとにがんというのはエージというのがドーズだとしたときにヒットを繰り返すことによって起こっていることが明らかで、ガンマ分布にがん患者の年齢分布が従いやすいということを想起しました。実際の適用事例もまあまあ合致していて、スライドでは無理やり書きましたけれども、エチルニトロソ尿素と消化管がんの発がんのリスクというあたりとかにマルチヒットモデルが用いられて、他のモデルより

も適合がまあまあいいという事例があります。

(PP)

ワイブル分布とガンマ分布に関してはそういう形で覚えやすいものです。ワイブルは故障の分布でメモリーがある。ガンマ分布というのは何回もランダムなことを繰り返すようなモデル化ができるというようなモデルで、メモリーのない、そういうヒットを繰り返すようなモデルですということなのです。

ロジスティックに関しては、実はメカニズムのある分布としては余り使われていません。これは何かと言うと、ロジスティック関数というものに由来しているもので、平たく言うとS字曲線のパイオニア的な分布です。こういう関数で与えられますというのがここで書いているのですけれども、これは何に由来するかというと、1844年にさかのぼるのですが、フェルフルストさんが人口でロジスティックの法則に従って人口がふえるというのをモデル化したのが最初です。

ここで言うPというのは人口サイズです。Pというのは無限にふえていくのではなくて、一つのキャリングキャパシティーと言って大文字のKなのですけれども、Kよりも大きくならないようにできているというのを微分方程式でモデル化可能です。すると、その解が今日使われるようなロジスティック曲線になっていきます。何でこういうのが出ているかと言うと、昔、人口論で人口爆発というのが1800年代に物すごく議論されることがあって、マルサスのころからもそうなのですけれども、指数関数的に今から人口がふえていって食料問題が絶対に出てくるという話が19世紀から20世紀初頭くらいまで、物すごく人口学者の間で議論されました。そうではないのだぞというような対抗する理論がそのときに同時に発展をしたのですけれども、そのうちの最も有望なものの一つがロジスティックだったということです。

生物群集で餌がずっとないのでキャリングキャパシティーというKを限界にしながら人口がふえるというのが、実際に例えば北海道のエゾシカなどがそうです。個体群と言ってシカの頭数がある一定のところまで止まる。上げ止まりをするというのがモデル化できる分布で、それによってS字曲線になることがわかったのですけれども、ロジスティックがドーズレスポンスのS字曲線に向いているかと言うと、その裏付けは実はそんなにたくさんあるわけではないです。

(PP)

ロジスティック分布の数理的特性は左右対称という話はしましたけれども、正規分布に似ています。左右対称で正規分布よりも多少Kurtosisが高くて、縦に長いような分布でもモデル化ができるということで、個体群生態学でさっきエゾシカの話と言及しましたけれども、成長曲線と解釈できるようなこともあったので今まで信用されてきたのですが、実はリスクアセスメントの分野だと「形状が使いやすいから」用いられているだけだと理解しています。

本来的には頻繁に出てこないのですけれども、デフォルトで1個設定されると、その慣

習がずっと続きます。なぜ頻繁に出てこないかという、数学的に他の統計分布と比べると、ドーズがゼロになっても現象がゼロになるとか、そういうところが制御できませんよね。リスクゼロから開始しなかったり、単なる当てはめでその中のメカニズムがロジスティックのルールで説明できなかったり、もともと正規分布からそんなに離れていないとかいうことがあって、微細なフィットのよさがプロビットと比べて、ちょっと優れていた、という事例が何例かだけあって、慣習的に使われるようになったというものです。なので、ロジスティックに関しては左右対称であるということと正規分布よりも高さが高いと見ていただければと思います。

(PP)

対数ロジスティックに関してです。数学が苦手な方の中では、ロジスティックの解説が終わると、その上に対数がついた時点で頭の中のシャッターがおりていくのですけれども少しだけ頑張りましょう。馴れ初めを頑張ってお書いていますが、正規分布と対数正規分布の関係に似ていて、ドーズのLogがロジスティック分布に従うというだけのことです。ドーズのLogがロジスティックに従うのだったら、本当のドーズというものは対数ロジスティックに従いますよということが数理的に明らかに証明できます。この関係は、正規分布と対数正規分布の関係にとっても似ています。だから、対数ロジスティック分布は対数正規分布にとっても似ていて、右に裾が長い分布です。分布自体は左に偏っているわけです。

分布は経済学におけるFiskモデルと呼ばれているものと全く一緒です。富の分布をモデル化しているのですけれども、横軸をドーズではなくて人の富だと思って分布すると中産階級くらいまでの人が高いようなところに入って、ライトテールのところにたくさんお金持ちがいる。あるいは気象学で降水量の分布にもこのモデルが頻繁に用いられます。

(PP)

対数ロジスティックモデルというものがリスク解析でどのくらい使われているかがここに書いていることなのですけれども、実を言うと対象とする興味のあるイベントがたくさんの変数がランダムに影響し合っただけで起こっているものを平均と分散だけでモデル化するのだったら、対数正規分布のほうが意味合いは分かりやすく、より自然です。対数正規分布というのは対数プロビット分布のことを意味するのです。対数正規分布のほうが自然な分布ですけれども、データが生まれてくる背景に実はランダムにドーズに反応したイベントを見ているのではなくて、関心のあるイベントが多数の変数が1つ、2つの要因が絡み合っている、相関しているとか、1つの変数がかなりドミナントに発生を決定するというような実践的なデータ生成過程に対応する場合は対数正規分布よりも狭い分布になりがちだから、対数ロジスティックのほうがいいですよということが頻繁に言われます。この点は難しくて分かりにくいので、次の図を用意しました。

タイムドーズイベントで、具体的にこれは『Scientific Reports』に今年出版されたもので、この図の右側が、対数ロジスティック分布がドーズレスポンスで使われているものです。

(PP)

これは実は単なるドーズレスポンスの研究ではなくて、そのドーズレスポンスと図のタイトルのところに「Logistic-accelerated failure time location-scale mixture regression model.」と書いてあるのですけれども、大変複雑です。何かと言うと、対数ロジスティックのモデルというだけではなくて、対数ロジスティックの変数が別の次元で、さらに左2つにあるような別の分布がタイムドーズで言うタイムのほうの情報があったりとかいうのに使われていて、それに混ぜたりとかしないといけないときにパラメータ間の従属性と言うのですけれども、それを処理しないといけないことがあるのですが、そういったときにも用いられるので、対数ロジスティックがフレキシブルで使いやすいということで、パラメータが複雑に絡み合ってもいいような現象をモデル化するのに向いています。

(PP)

プロビットモデルですけれども、2分変量とあって、イベント発生の有無とあって、皆さんがQuantalレスポンスと言っているものと一緒です。説明変数を入力情報として利用した場合に正規分布で与えられるということで、これは正規分布に従うようなモデル化をするというものです。プロビットのほうがロジスティックよりもリスク解析に使われているヒストリーが長くて、1930年代に薬物処方量と回復率の関係というものが研究されて発展したことに遡れます。今現在では実験研究のデータをもとにモデル化するときロジスティックと対比されるシンプルな分布です。Excelとかでも実装されていますので実験医学専門の皆さんは御存じだと思います。

(PP)

プロビットとロジスティックは、2つとも、もともと他の分布と比べると慣習的に使われているのですけれども、特徴でここに書いていますが、ロジスティックのほうが右のテールが厚いという言い方をしますけれども、100%に近づくときにゆっくりとドーズを高くしていても、ロジスティックのほうがゆっくりと100%に近づいていくのですけれども、プロビットというのは素直な分布過ぎるので、100%にすぐ近づいてしまいます。ハイドーズでハイレスポンスのものを何らかの形で扱うような機会があると、ロジスティックのほうが実は現実に寄っているかもしれないということと、数理的な分解をしたときにプロビットはただの正規分布なので、数理的な解釈はロジスティックのほうが比較的容易と考えられるときが多いです。

(PP)

一方でプロビットはそういうので解釈に乏しいのですけれども、対数プロビットモデルは相当リスク解析寄りに発展してきたモデルであることが知られています。対数正規分布のことです。個々の動物はこれ以下だという反応が起こらないという用量を持つはずだというような発想をもとに、閾値より上だったらイベントが起こるという発想をもとに出て

きたものです。これはどういうことかと言うと、閾値ドーズであって、それに達するまでの時間あるいは閾値レベルに達するまでのドーズを見ると、その対数が正規分布に従うので、それを最終的にドーズと反応の関係に従って対数正規分布ですねというのがわかってきたということです。

(PP)

ED50の計算を治療でやったりとかするときに対数プロビットがたくさん使われるので、やたらと盛んに使われています。なぜこれが便利なのかと言うと、観察しているイベントが発生するという頻度はLog (dose) が平均Log (dose) からどれくらい離れているかをはかることでイベントの発生頻度を正規分布を基に決めてしまうことができるためです。下線を引いている部分が大事ですけれども、発がんまでの時間の分布が対数正規分布に従うとき、用量反応関係は対数プロビットになることも知られていて、発がん性などを見るに当たっては、対数正規分布は必ずチェックしておかないといけないものなので頻用されているということです。

(PP)

Quantal-Linearモデルみたいに英語で書かれると、統計の苦手な人はより窓だけではなくて雨戸も閉めるくらい遠くに行ってしまうのですけれども、Quantalというのはどういうことかと言うと、2値データということです。イベントがあるかないかというような、イベントが起こるかどうかという意味なのですけれども、Quantal-Linearというのは実を言うと指数分布です。ただの指数分布なのですけれども、2値のイベントの発生確率をモデルするためには指数分布が使えるということです。

実は、詳しいBMDとかのモデル化をするとQuantal quadraticと言って、quadraticというのは2乗の式、二次式のこと、より難しい2乗のモデルもあります。指数分布はエクスポネンシャルの中は $\beta x$ とXは1乗でしかないのですけれども、Quantal quadraticというのはXが2乗のような場合にドーズが1個ふえるごとにそのドーズ分の2乗、反応が起こりやすくなるようなモデルです。より反応が起こりやすいようなモデルも使えます。ただ、Quantal quadraticはデータにフィットしやすいから使われているだけで、実は指数分布ほどリッチなヒストリーはありません。ここでは、指数分布について知っていただきたいと思えます。

(PP)

指数分布というのは、興味があるイベントというのが明らかに確実にランダムなプロセスに従うというとき、例えば地震の発生というのがポアソン過程であるとする、今日地震が起こるかどうかの確率は指数分布に従います。前の大規模地震の発生から次の地震が起こるまでのときに、前の地震の発生日付から何時間たっているから、地震の発生確率はどれくらいであるというのが経時的に計算をすることができるのですけれども、そのときは指数分布を使っていきます。

(PP)

リスクアセスメントでの意義ですけれども、短い時間に1回のみ起こるか否かという、すごくシンプルなことに興味があって、単位時間当たりのイベント発生ハザードが一定であると想定できる場合に指数分布が有用です。だから、車の故障と異なり、コップを1回割るか否かというだけの分布が指数分布であるということです。数学的には1パラメータしかないのでもっとも簡単なのですけれども、リスクアセスメントの場で言うと指数分布の別名はワンヒットモデルと現象をイメージしつつ呼称が与えられているのに注意をしないといけません。

これは発がん物質が遺伝子に衝突して損傷を受けた遺伝子ががん細胞に転化することをモデル化するためのモデルです。毒性を持つ物質がランダムに分布していて、その摂取量もポアソン分布に従って、分布はそんなにスキュードではないと。毒物量当たりのイベント発生の確率が一定のとき、一つの細胞とか、あるいは分子に発がん物質がヒットする確率が捉えられます。1ヒットで1イベントが起こるような、そういうランダムの一意の関係がある時には、指数分布でよいというような背景の研究が今までされてきています。

(PP)

ただ、発がんはマルチステージで起こるものです。細胞がイニシエータによってイニシエートされて、プロモータによって腫瘍になって、プログレッサによって浸潤・転移するというような、そういう発がんプロセスがちゃんと現象として観察されるまでには幾つかの加速度的なプロセスが起こった上で発生することが知られています。その場合は多段階で加速度的に起こるというメカニズムを捉えたようなモデル化が望ましいです。つまり、がんの状態に近いところまで反応が起これば起こるほど、ドーズの影響はより大きくなるはずである、というのを捉えようとしたものが、ここにある線形化多段階モデルですが、エクスポネンシャルのさっきのQuantal-Linearに近いのですが、1乗の項だけではなくて、ドーズが高くなれば高くなるほど、2乗項とか3乗項とかK乗項まであるのですけれども、その何乗とやっている影響が大きくなるようにつくられているモデルということです。

(PP)

これをもとに特にドーズが小さいときにも、マルチステージモデルは数理的にいろいろ有用であることが知られています。ドーズが小さいときには、エクスポネンシャルの中の2乗以上の項は物すごく小さいので無視できるということで、テイラー展開をして近似をすると、ここの全体というのが $\beta_{ix}$ に近似されます。それが理由で発がんスロープファクターというのがあって、今までのリスク解析とかでも使われてきました。モデルの妥当性がスロープファクターに関しては過去に議論をされています。動物実験データというか、他の知見と比べてスロープファクターだけを使うとミニマムのドーズが100~1,000倍くらい違うというようなこととかも出てきてしまって騒ぎを起こした理由の一つなのですけれども、それ以来、その多段階モデルはやみくもに使われている状況は少なくなりました。

幾つかのがんで閾値が見られないことが理由だったり、適合性だけで下向きの外挿が厳

しいということなのですが、つまりよくフィットするからというだけで使われ過ぎているということだったり、あるいは変異原性の試験とかを見ると低用量まで直線性を認めるとか閾値なしの仮定とか安全側の推定をするべきということがあって、スロープファクターは必ずしもリスクアセスメントで規制下限値を決めるのにはピットフォールが幾つもある、ということが議論されてきました。(PP)

それで、これまで話してきたことを話した順番順に並べたものがこちらのスライドです。ワイブル分布からマルチステージまでを見ていただきました。分類という列に書いているものの中で、「機構」というものは背景にあるメカニズムが意外に明らかであって、少なくとも数学的には説明可能なものがあるものです。「分布」と書いているものはロジスティックとプロビットが典型ですけれども、もともと数理的に決まっていたものを外から当てはめているというようなイメージのものです。それぞれ由来を書いています。

ワイブルが故障とか、ガンマは指数分布の繰り返しというような、このそれぞれの分布に関して、どんなイメージで皆さんに知っていただきたいかというのがそこで、リスク解析をする上での特性が右端の列で書いているものです。プロセスに記憶があるというようなものは簡単なのですけれども、例えばプロビットとロジスティックの違いでテールの厚さの違いがあるとか、あるいは多段階発がんモデルとかの活用経験があるのでマルチステージがよいとか、いろいろな理由があって、こういうのが使い分けられているというところを見ていただければと思います。一番大変なところがここまでなので、一旦ここで止めたいと思います。

○川村座長 ありがとうございます。

数学好きにはたまらなくおもしろい課題だろうと思いますけれども、たくさんの紹介をしていただきました。根本的なモデルというものの考え方の話から始まって、各モデルの成立の経緯から特性まで、いろいろとお話をいただいております。委員の先生方はいかがでしょう。このモデルの成り立ちや特性について、何か御質問はありませんでしょうか。

岡田先生。

○岡田専門委員 ここに書いておられるものはたくさんあるのですけれども、例えば体内動態が千差万別でいろいろなことがあったとしたら、本来はそれぞれにまた新しいモデルをつくらなければいけないものなのですか。

○西浦専門参考人 モデルの使用の原則が背景にある現象を全て捉えられないといけないとすると、一つひとつの物質が起こす現象に関してモデルをつくり直すというのが理想です。けれども、特にその化学物質のように種類が多い場合というのがそうなのですが、9割くらいになるかもしれないと思いますけれども、たくさんの化学物質と対象とするエンドポイントの間で生体内の反応がどういうメカニズムに従って起こっているのかが分からない場合がありますので、そのときはいろいろなモデルを試してみて、どれの予測がよいかという話に置きかえながら検討されている、というのがリスクアセスメントの現状であると理解しています。

○川村座長 ありがとうございます。それはかなり本質的なお話だと思います。およそ科学が発達したからと言って全てのことがわかっているわけではなくて、結局、真理というのは何かというと、論理的に考察して到達する部分もないではないけれども、実際に仮説を立てて当てはまりを見て、それで矛盾が生じなければ真実というふうにサイエンスとしてはなると思います。光が粒子か波かという、そんなのは誰も見たことはないけれども、光の速さが一定であるという、そういう物理の根源的なところまで、そういう仮定をすると全ての現象がうまく説明できるということなので、そこから始まっているということを考えると、今日のお話を聞くまでは、それはメカニズムが先でしょうと思ったのだけれども、それは言い切れない話なので、やはり経験則に基づいて当てはまりを見て、これのほうが説明しやすいということで、今までの話から、分布というのもやはり極めて重要な要素だということになるのかなと思いました。

他に御質問はいかがでしょうか。佐藤先生。

○佐藤委員長 よく理解できていないのだと思うのだけれども、ガンマ分布と指数ワンヒットというのはどこがどう違うか、もう一回教えていただけますか。

○西浦専門参考人 指数分布というのは1つの細胞に1つの核があって、化学物質が1個反応するとがん細胞になりますよというような、そんなイメージのものをモデル化すると、ドーズが横軸にあるとこんな形になりますというのが指数分布です。1つのイベントの1回の刺激による発生確率ですね。一方で、ガンマ分布というのはそういった刺激が何回も起こらないといけなくて、同じようなことが続いていくうちに、最終的にがんの細胞になってしまうのですけれども、そういうプロセスを必要回数分を繰り返すと、当該ドーズに対してガンマ分布に従って反応することになる、ということが知られています。その違いです。だから、指数分布は言いかえるとワンヒットと呼ばれます。1回損傷を起こして、どこかでがんが出てくればよいというものなのですけれども、ガンマに関してはここの分だけ、まず全部がん化してしまわないとイベントが観察されないようなものを相手にしているので、マルチヒットモデルと呼ばれます。

○佐藤委員長 それさがさき畳み込みとおっしゃっていた話になるわけですね。

○西浦専門参考人 はい。

○川村座長 その右に重なるほど、確率が上がっていくというと、それがマルチステージになっていくと。

○西浦専門参考人 そうです。

○川村座長 2乗、3乗とかという繰り返しがより起こりやすくなって、記憶というわけではないでしょうけれども、ある条件が積み重なっていくとプロバビリティが上がっていくモデルがそのマルチステージということですか。

○西浦専門参考人 マルチステージに関しては、こちらと対比するのではなくて、ドーズの量の影響の強さを直接モデル化したものになります。今、ドーズが低い量なのか、あるいは中くらいの量なのか、高い量なのかというときに、ドーズの違いによって、ステージ

の進展が線形になる保証はなくて、低いドーズではそんなに反応が起こっていなかったけれども、高いドーズになると反応度が線形よりも極めて高くなりますよというようなメカニズムをモデル化したいときに、そのマルチステージが向いているということになります。

○川村座長 では、その履歴は関係ないということですか。

○西浦専門参考人 そうです。マルチステージに履歴は関係なくて、用量と反応の関係で加速度が起こるか否かというところがマルチステージとワンヒットの違いです。

○川村座長 分かりました。他にいかがでしょうか。あと一つ、こういう評価で大事なのは微量領域のところですよ。外挿という言い方もあるかもしれませんが、そういう threshold のありなしとか、プロビットなどだと割と最初は地面をはいつくばって立ち上がりみたいに見えるところが出やすいということなので、イベントの発生メカニズムと立ち上がりとの関連というのは、先生はどう考えていらっしゃいますか。

○西浦専門参考人 ロードーズでの立ち上がりとイベントの発生との関係というのは、ロードーズを詳しく見たときと、用量反応曲線の特性の左裾を厳密に検討して、皆さんは決められています。ロードーズのときの話をするとき弱いのが指数分布です。ワンヒットだと必ず縦に立ち上がらないといけないので、まあまあ不向きであるということが言われていて、ロードーズに関しては左裾がどんな形かが重要になります。観察データが途中のドーズまでしかない中で、現実に近いロードーズの挙動を捉えられているか否か、というのが、この8つのモデルの比較のときの勝負になることで、それと Restriction とかのパラメータの話がまた関連するので、その辺はこの後に触れさせていただければと思っています。

○川村座長 モデルを基準値の決定に採用するときは全体の実験とか疫学研究の全ての領域のデータを使って書くのですけれども、注目をしているのはやはり微量領域なのです。だから、全体の当てはまりと部分の当てはまりとといいますか、その矛盾とかが起きることはあるのでしょうか。

○西浦専門参考人 いい質問ですね。全体の観察されている部分、こういうリスクアセスメントのよく起こっていることというのは、高用量領域の動物実験などのデータを通じて数理モデルを適用することによって、低用量の推測をできるだけするというようなものです。当てはまりに関して、今、評価ができるというのは、高用量のほうの領域しかデータがないのだったら、高用量のデータに基づいて決まっています。低用量がそれで正しいかどうかというのは高用量の現存のデータとこの関数の想定に依存しています。なので、その想定が悪かったら悪くなるし、正しければ正しいということになるのですけれども、低用量の当てはまりを評価することは、データがない限りはできないというのが原則です。

○川村座長 分かりました。結局そうですね。済みません、確認させていただきました。よろしいでしょうか。

○赤堀専門委員 すごく素人的な質問ですが、先ほどの $\gamma$ のお話のところ少し理解できなかったのですが、あの $\gamma$ の値というのは例えばリスクアセスメントをするときに $10^5$ と

か、あの話をしているのか、そうではなくて全く別の値を設定しているものなのかというところをまず聞きたいです。あとBMDモデルの中で今は8種類のモデルが当てはめられているのですが、他にもいっぱいモデルがあるけれども、この8個になっている理由は何かあるのであれば、教えていただければと思います。

○西浦専門参考人 ヲの御質問が少し理解できなかったのですが。

○赤堀専門委員 あのヲというのはどこから来る値になるのですか。

○西浦専門参考人 ヲのスライドに何かあった話でしょうか。

○赤堀専門委員 例えばスライド9のところで、式の中にBMDモデルにするときには下の8つのモデルに関してはヲを入れてやっているのがBMD特異的な処置であるということだったので、このヲが意味しているものが何なのかということです。

○西浦専門参考人 パラメータヲが意味していることは、ドーズレスポンスカーブを描くと、とてもよく分かります。ドーズを増やしていったときに、本来の分布はドーズがゼロであると値がゼロになるもので、S字型の分布なのですが、観察データは特に疫学データなどがそうなのですが、全くエクスポージャがないときでも発がんしていることがあって、ドーズがゼロでも反応はゼロより大きなところから始まって、そこを基点にS字曲線が描かれる、ということがありますが、その接点の部分を立ち上げているのがヲです。だから、ドーズゼロでヲから始まってくださいというような処置をこの6つの分布でやっているということです。

○赤堀専門委員 必ずしもデータドリブンで来るわけではなくて、仮定の値を入れてやるということになるのですか。疫学みたいな形であれば、発がんとかであれば、バックグラウンドで自然発生的に出てくる発がんの発生率を例えば、そこに当てはめるというイメージになるのかなと思うのですが、そういう理解でよろしいですか。

○西浦専門参考人 そうです。

○赤堀専門委員 ありがとうございます。もう一つは、何でBMDモデルの中はこの8つなのですかという簡単な質問です。

○西浦専門参考人 これはアメリカのEPAとかヨーロッパにもあったり、あるいは日本でもBMD法の推定パッケージをつくられている方がいらしたりすると思いますけれども、そこでデフォルトでこれだけ比較したらというおススメで任意で抽出されているモデルです。それは統計の分布で言うと2パラメータでスタンダードディストリビューションと言われるもので、パラメータ数が少ないにも関わらず観察現象に当てはまりがよくてコンベンショナルに使われているもの、というのが大体この8つに当てはまります。

より詳しい微細のデータがあつて、パラメータ数が増えたものというのはもっとたくさん分布がありますけれども、多くの場合において動物実験のデータはサンプル数が限られていて、例えばガンマ分布でも詳しいジェネライズドガンマというのがあるのですが、そういうのにすると5パラメータとかになってしまいます。でも、サンプル数や観察データポイントの不足などによって5パラメータが持つ豊富な情報量を動物実験のデータ

では与えることができない、ということがよく起こり得るのですけれども、そんな限られたデータの中でもできるだけ観察パターンを再現できるようにしたのが、こういう2パラメータディストリビューションです。パラメータ数が少なくても便利だからというので、とりあえずこの程度の分布の組合せの検討をやっておこうかというので選ばれているものだと私は理解をしています。

○赤堀専門委員 ありがとうございます。

○佐藤委員長 それに関連していいですか。最初のほうで先生はリニアでもいいとかおっしゃっていたのですが、ホッケースティックなどというのものもあるのだけれども、ああいうのが排除されてしまった理由は何かあるのですか。

○西浦専門参考人 経験的に一般化しやすい分布がこういう分布だからというふうにヒストリカルには理解しています。何というか、リニアというのは例えば近似でこういう分布ではなくて、「ある用量」までは直線的に増加するけれども、「ある用量」以上ではプラトーに達しx軸に平行な直線になるというような分布もできるわけですが、そういったモデルですと、やはり、その「ある用量」をどうやって実験医学的に決めるかという、すごく重大な責任と恣意性が入らざるを得ないのです。それと比べると現象をより詳細に捉えているように見えるほうが良さそう、というのが今までの経験則から得られていることで、それで連続的な分布が使われるようになったのだと思います。

○佐藤委員長 ただ、逆に言えば、そういうブレイクポイントみたいなものはあった方が楽は楽という感じもするのです。

○西浦専門参考人 そうなのですけれども、BMD10とかの話よりも、より難しくなると思います。

○川村座長 底が深いことが分かりました。引き続いて、次の選択のところに行ってくださいでしょうか。

○西浦専門参考人 次にモデル選択に関して、お話をさせていただきます。特に使用モデルが決まらないようなことがあった時にどうすればいいですかというクエスチョンを事前にいただいたので、その辺も考えながら御説明させていただきます。

(PP)

前回のワーキンググループで川村先生、広瀬先生のお二人から説明をいただいていたと思いますけれども、BMDのモデル選択に関して、今どのように実施されているのかということのおさらいで、まず確認をさせていただきます。今、BMD法でモデル選択をする上では2つの選択するプロセスが行われています。1つはEPAとかで出てきているBMDSというソフトウェアを使用するときには棄却基準というのがあって、適合度検定をします。観察データとモデルの間でフィットがいかどうか適合度検定をして、p値が0.1未満のモデルをまず却下することにしています。あるいはBMDとBMDLの比が10以上でNOAELから外れるということの蓋然性が高い場合は却下しているということです。

もう一個のプロセスですけれども、モデルが残ったときにそこからどうしているかとい

うことですが、「Survivedの」と書いているのは幾つかのモデルが適合度検定などで却下されて棄却されて外してしまっただ後に残ったモデルのセットでBMDLが最も低いものをとるか、あるいはモデル依存性がそんなに目立っていなければ、最もAICが低いものか、あるいはAveragingとってモデルの平均化などを検討するというふうな、その2段階のプロセスをしながらモデル選択が行われて、ロードーズの議論が行われているというのが現状です。

(PP)

次に、モデルの選択に関してです。モデル選択についてもそうなのですが、「考え方や方法論を知りたい」と聞いていただいたのは本当にうれしかったのですが、ゴールドンブックがありまして、Burnham&Andersonがモデル選択の専門本として知られています。『Model Selection and Multimodel Inference』という本があるのですが、情報量基準と言うのですが、この範囲では実は皆さんはAICとつき合っていて御存じかもしれませんが、赤池先生は統計数理研究所の所長を務められた方ですが、日本は情報量基準の最先端を行っていた国です。赤池先生自身が論文を出されて、少したってから、すごい論文が日本から出ていたことがわかっていたのですが、今はこうやって国際的にたくさんの情報量基準ができて、どのようにモデルを選択すべきかという基本的な根本論も専門書になって、統計モデルや数理モデルの研究をするメンバーでは共有されています。

(PP)

モデル選択をする話をプリンシプルで話させていただくのですが、モデル選択の科学で受け入れてもらわないといけないことが一つだけあります。モデル選択をするというのはどうやって選択をしているかということ、予測を重視するということです。モデルが当てはまってい、そのモデルが未来またはサンプルサイズをふやしても何回繰り返しても当てはまりのよいデータになるというような、そんなモデルを選びたいというのがモデル選択のプリンシプルです。大体普段やることというのは手元にある限定的な観測データ、すごくサンプルサイズが少ないようなデータも含めて予測のためのモデルを構築するので、モデル選択の目的は「予測がよい」モデルを見つけることです。

そこは背景にある現象というのを捉えているか否か、真のモデルかどうかというのは実は統計学ではモデル選択においては余り重視していません。モデル選択というもののサイエンスのプリンシプルは予測がうまくいくモデルで、かつ、パラメータ数が少なくて済むような簡単なモデルがよい。パラメータ数がふえ過ぎると観察データポイントと同じ数のパラメータを使えば完璧にデータを説明してしまえるので、パラメータもたくさん使うべきではないのですが、簡単なモデルで予測がうまくいくものがあるので、それを評価するための評価指標がAIC、DICと言われるモデル選択基準です。

(PP)

そのモデル選択を簡便に済ませてしまうための基準なのですが、AICというものがどういうものかを一緒に振り返らせていただきたいのですが、8つのBMD法で使うモデルがあって、観察データにフィットしたときに最も好まれるモデルというのはここで計算する

AICが最小値になるようなモデルです、と言われていています。AICというのは、この「 $2k$ 」というのはパラメータ数が $k$ で、パラメータ数 $\times 2$ のことです。「 $2\ln(L)$ 」というのは対数尤度を負にした部分なのですけれども、これは観察データをモデルが説明できる確率の対数値です。かける2ですね。だから、何か目の前に動物実験の結果があって、モデルをフィットしたときに、このモデルで観察データがこのデータが出てくるという確率の対数をとったものです。確率の対数なので負の値になってしまいますから、そこにマイナスをつけると正の値になるので、これは何かと言うとパラメータ数をペナルティーにしながら、フィットのよさを評価するというような、そんな指標になるということです。

パラメータ数がふえると「 $2k$ 」の部分が大きくなり過ぎるので、AICがでかくなります。一方でパラメータ数が少ないとき、 $2k$ の部分は小さいかも知れませんが、ライクリーフッドと言って、その観察部分の確率のところが低くなるので負の対数尤度が大きくなってしまいます。 $2k$ というのは、オーバーフィッティングを避けて簡単なモデルでAIC最小になるものをよしとしましょうというコンベンションで、これが出てきました。だから、最も当てはまりのよいモデルはパラメータ数の多いモデルなのですけれども、AICを使うと当てはまりだけではなくて、パラメータ数もほどよくミニマムになっているというような、そんな比較をした上で選択ができるということです。

(PP)

BMDとかのソフトウェアがEPAなどでありますけれども、モデル選択では、まずモデル選択の話はこの先からちょっと深く入るのですが、先に2段階の評価をしているということを理解していただきたいと思っています。

1段階目は先ほどお話をしました適合度検定をします。これは何をしているかという、モデルで有意にデータから外れているものを棄却したいので、有意に乖離しているモデルを捨てたいということです。いろいろな選択とかの議論をする前に、まず捨てたいということです。そのときに $p$ が10%未満のモデルはフィットさえしていないから捨てたいという意味でやっています。注意しないといけないのは、その捨てた後はこの $p$ 値は以後参照しないということです。 $p$ 値は捨てるための話であって、Neyman-Pearsonの原理と言うのですけれども、仮説検定をしているのです。仮説検定をして捨てるか否かの話をして、その後選択の話に持っていくのですけれども、その選択をする部分の選抜にさえ漏れるような却下されるべきモデルを決める、というのが第1段階です。その $p$ 値はその後の過程で参照する必要はないです。

2段階目で残ったモデル間でAICを比較していただくということになります。そのときはAICの比較だけではなくて、BMDLを保守的に最小の信頼区間をとるとか、あるいはモデルの意味論で選択するとかいうような幾つか方法があるとは思いますが。

(PP)

よくある誤解なのですけれども、モデルを選択するときに検定もする場合があります。それはモデル選択というのがAICに沿っているという点から言うと、必ずしも理路整然とし

た考え方ではないです。モデルAICを検討するとき横にPバリューが出ているというのは慣習的に出している間はいいのですけれども、モデル選択の基準にしている間はちょっとおかしくて、検定はNeyman-Pearson的には仮説検定の考え方で、モデル選択の情報量基準とは考え方が異なるということです。あるいは検定すべきモデルを、モデル選択を通じて選ぶというもおかしかったりするので、プロセスの順番を逆にしたりしても、また変なことが起こってしまうということです。

別の注意点ということで一つ書いていますけれども、選択されるモデルのパラメータの95%信頼区間に注意するのが重要です。ということで、先ほどパラメータ $\gamma$ の質問をいただきましたけれども、低用量のデータがないときに、こうやって切片を決めるのは結構難儀します。特にS字曲線の立ち上がり部分しかほとんどデータがないというのが多くの事例ですから、 $\gamma$ の信頼区間がすごく広がることがあります。 $\gamma$ の推定がうまくいかないというときに、実はそのモデルからBMDLを計算してみるとBMDLの信頼区間下限も物すごく広がってたりするということも起こり得るのですけれども、そういうことが起こり得るので、パラメータの信頼区間に注意するのもモデル選択のときには必要になります。モデル選択云々の前にパラメータの信頼区間が広過ぎやしないかということをチェックしないと、そのモデルが今、得たい情報をモデル化するために妥当なパラメータ推定値を与えた上でフィットされているかどうかの評価できませんので、信頼区間も見てくださいということです。

(PP)

AICで選択したモデルというのは、ちょっとしつこいですが、たくさんあるモデルの中で最良の予測が得られるモデルです。これは最良の予測とはどういうことかと言うと、もう一回ランダムに同じようなデータをサンプリングして、もう一回同じような頭数の動物で実験をしても、たくさんあるモデルの中でも最も適合しやすいと考えられるようなモデルということです。これは真のモデルを選んでいるわけではないです。ここが、私たちがずっと頭を悩ますところなのですけれども、予測で当てはまりがずっといいでしょうと選ぶというのです。

ただ、一つ注意しないといけない点が、AICは最強かと言うと、そういうわけではなくて、モデル選択はサンプルサイズに影響されていることも注意しないといけません。実はサンプル数が少ないと単純なモデルのほうがAICは最小になりやすいということがありますし、サンプルサイズが多いと複雑なモデルが選ばれやすいというような、AICも必ずしも万能ではないということも知っておいてください。AICの兄弟分は小サンプルのときにはAIC<sub>c</sub>とかいうのがあったり、あるいはBICと言ってベイズ推定をしたときに使えるようなものもあったり、同じような情報量基準に従っています。ただ、今のところはBMDのソフトウェアだとAICを提供するものがほとんどではないかなと思います。

(PP)

モデル選択をする際にAICがほんの少しでも小さければいいのかというような議論をす

ることがあるのですけれども、原則、最小AICがいいということです。ただ、AICにもサンプリングのエラーが付きまとうのでばらつきがあるので、コンベンションで使われているのはここに書いているとおりなのですが、AICがモデル間で差が0~2くらいの間だったら、より母集団に近いモデルが正しく選択されないリスクがあるということですが、モデル選択でAICが最小だったけれども、1とか2の差に集中し過ぎていて、実はAICもエラーがあって別のモデルのほうが本当はよかったということも起こり得ます。4~7だとちょっと危険とか10以上だったら大丈夫とか、これはコンベンションにすぎないです。BurnhamとAndersonが今までの数値実験でやっているものです。そういうこともあるので、どんなときでも近いAICがもしあるのだったら、その2番手、3番手の候補も必ず列挙して比較することが大事です。先ほども言いましたけれども、BMDLみたいにローエストリミットに興味がある場合はパラメータ値の信頼区間もちゃんと比較しておく必要が生じます。

(PP)

ただ、モデルを予測だけに基づいて選択するというのも必ずしも正しい方法というわけではないので、メカニスティックな解釈による選択もできます。つまり、本来の対象としている現象を捉えたようなモデルが理想ですから、AICによるモデル選択だったら興味のあるエンドポイントのある現象の科学的根拠を特定できないので、モデルのメカニズムとして当てはまっているモデルを優先して選択するというのもありなのではございますけれども、それにこだわり過ぎるのも結構ミスが起こりやすいピットフォールがあります。

例えば、がんと言うと勉強でどうしてもマルチステージとマルチヒットを皆さんは覚えてしまうので、がんになると、このモデルだけでいいのではないかというくらいの議論さえ起こってしまいがちなこともあるのですけれども、その考え方は必ずしも正しいものではないです。マルチステージとかマルチヒットというものががんをエンドポイントにしたときに当てはまりやすいというのは経験的に知られているような物質であればいいですが、そうではなくて、得られたパラメータも背景にあるメカニズムを捉えているかどうかを定かではないというときに、必ずそれをやみくもに使ってしまうのは余り適切ではなくて、モデルメカニズムの仮説どおりに推定がはまっていなくて意味が失われかねません。

データが発生するメカニズムは不明の間はModel uncertaintyと呼ばれるものが残っている状態です。生体内でどうなっているかがわからないので、複数のモデルで分析して、そのModel uncertaintyに対応すべきということです。

BMDLの検討でいつも問題になるのですけれども、Quantal-Linearがいつも問題だから除外するというような場合もアメリカとかで議論をされることがあるのですが、これも誤りです。除外理由の恣意性はできるだけ入れないようにしたいということです。

(PP)

メカニスティックモデルが優先的に選択されるときがあるということではございますけれども、そういう場合はどういうときかをリストしたのがこのスライドです。私の考えではございますけれども、物質レベルで生命現象とともに実証経験がある場合です。体内で多段階発現の線形モデル

があって、それがどうしてこういう線形なのかということも含めて、ある程度実証をされている場合は必ずそれは使っていいということになると思います。そういう事例を私はほとんど見たことがないのですけれども、感染症に関しては例外があって、ベータポワソンがなぜドーズレスポンスのモデル化にいいのかとか経験則があったりします。

2番がほとんどの事例だと思いますが、経験的に特定の分布が特定の現象に適合しやすいというようなエビデンスがある場合があります。例えば自分は大好きなのですが、急性感染症の潜伏期間は対数正規分布に従うということは経験的によく知られていますし、腸管感染症のドーズレスポンスカーブはベータポアソン分布が最適であるということもオランダの研究グループとかを通じて、よく知られているものです。生体内の現象が既に説明が実施されていて、どうして正規分布かということも含めて、一定の数理的メカニズムというのがもう議論をされていて、そのエビデンスが数値的にも繰り返されているという場合は、何度も繰り返されているは場合こういう選択を優先的にするというのは正しいプラクティスだと思います。

だから、背景のメカニズムがないといけないことに注意しないといけません。ふざけているわけではないのですけれども、都市伝説なのですが、宇多田ヒカルさんがデビューしたころのCDの売り上げは時刻とともに対数正規分布に従いやすいということがインターネットのマニアの人たちの間で話題になって、自分もフィットしたらそのとおりだったのですけれども、それは偶然です。単に対数正規分布に従うような何かトレンドが自然に起こったのですけれども、そのメカニズムはわからないまま一旦育児休業とかに入られたのですが、そこは背景の議論が起こっていないので単なる経験のみとかいうのだと、必ずこれを使うというほどの保証をしないということです。

(PP)

モデル選択をするときには、Lowest BMDLと言ってBMDLの最小値かAICか、どちらを使うかということですが、ケース・バイ・ケースで議論をいただくことが必要だと思いますが、ここに利点と欠点について書いています。Lowest BMDLを使うときには保守的に最小のBMDLを与えるモデルを選ぶので、保守的なモデルの決定が可能です。そういう判断をすると信頼区間ができるだけ狭いような実験デザインをしないといけないので、よい実験をするように皆さんへのインセンティブが働くことになります。ただ、こういう方法をとるとサンプルサイズに依存して、すごく低いドーズがBMDLとして出てくることもありますから、ちょっと注意しないといけないということです。必ずしもLowest BMDLのプラクティスがユニバーサルに基準とするとプラクティカルかどうかというと、そこには疑問があるかもしれません。さっきパラメータの話をしましたけれども、妥当性が不明なフィットとすると、そのドローバックがあるということです。

AICを使うと最も当てはめのよいモデルが選択可能なのですけれども、先ほど川村先生から質問をいただいたとおりです。Lowest doseのあたりが知りたいのですけれども、その辺の判断に関しては恣意性がどうしても入ってしまうということです。高用量のデータしか

ないので、低用量の部分は高用量のデータがこうだったとしたときに低用量はこの辺でしようというようなスペキュレーションをしているというのが現状ですので、どんな場合もどうしても恣意的になってしまう。欠点はパラメータの信頼区間を気にしていないということと、サンプルサイズが小さければAICも誤りがあるということです。

(PP)

ちょっと駆け足で行きます。

(PP)

モデルの平均化についてです。そういうことがあるので、選ばれたモデルをどこか一つだけを選ぶというのがとても難しいので、アンサンブルという考え方があります。アンサンブルに関連して、予測のサイエンスでは天気予報が大先輩です。感染症とかをやっているでも太刀打ちできないくらい天気予報が進んでいて、ベイズ推定が古くから活用されてきており、アンサンブル予測というのがあります。これは何かと言うと、いま実際に近づいてきているようですが、台風があつて初期のパラメータを変えると2回微分するようなモデルで台風のコースを予測しているのですが、その予報が全く違うトラジェクトリーをたどってしまうということが知られています。アンサンブル予報というのが、皆さんがニュースで見ているものです。たくさん予報がある中で平均をとって、もっともらしいところを特定して、それを皆さんが偉そうな顔で何時に来るでしょうとかいう話をやっているというのが平均化です。

(PP)

平均化話がBMDで出てきているのは、AICとかBICのモデル選択が不確実で、その不確実性が推定量の性質に及ぼす影響を考慮しないと、誤った推論が行われがちなので、平均化をしたらどうですかということです。今、BMD法ではWheelerさんというアメリカが中心になって、こういう議論をしています。いまだに平均化を何でしているかというのは、あえてここで言及していますが、真のモデルかどうかを議論するのは一旦諦めるというのがモデル平均化のフィロソフィーでもあります。真のモデルがどうかとかは一回考えるのをあきらめて、「予測がいい」当てはまりの良いモデルをとにかく追求するために、正しい天気予報を得るために複数のモデルを平均化して、パラメータの区間推定値、つまりBMDLを含めて追求した結果がモデル平均化の話です。

(PP)

複数のモデルでBMDLというのが出て、不確実性を考慮するために初期にやられていたのはどういうことかと言うと、こういうふうにBMDLではあるモデル系で出てきたときに、そのモデル系の重みをつけて平均をとります。実は重みはAICで与えられるのですけれども、それぞれのモデルのAICの重みをつけて足してしまったものをある程度何らかの形で割り算をすると、BMDLモデルアベレージングと言うのですけれども、平均化したモデルでのBMDLが出るというコンセプトです。ただし、BMDLアベレージングだけを繰り返してみたところ、これは再現性が悪いということがわかってきて、BMDLだけの指標をそれぞれのモデルから

抜き出した上で平均化をすると信頼性に乏しいことがわかりました。つまり、再現性が悪いのですが、再現性が担保されないので今どんなことが行われているかということ、ドーズレスポンスの曲線自体を平均化するということです。

(PP)

BMRが得られているので、BMRをAICの重みでつけて足し合わせて、その上で得られた最後のモデルアベレージした曲線が1本できますので、それをもとにさらにブートストラップ法とか、あるいはベイズ推定をするとBMDLの95%の下限が出ます。BMDLの不確実性を正確に捉えたBMDLの決定が可能ということです。BMDLの決定のみだったら、最良のモデル一つを選ぶよりも優れているのではないかという考えの下で、シミュレーションベースでそういう議論が行われています。ただ、BMDL決定のときにアメリカでたくさん議論が行われていますが、Quantal-Linearモデルが平均化の信頼性が低く、再現性が乏しいということが知られています。信頼性が低いというのは再現性が乏しいということなのです。

(PP)

それに関して懐疑的な方もいるので、妥当なのですかという質問に一定の見解を示してくださいということでした。それをまとめたのがここです。私からの返答なのですが、どんなモデルも誤りなので、何かのモデルをフィットしたときに「妥当ですか」というのは実はアンフェアなクエスチョンなのがプリンシプルです。BMD法によるモデル化の哲学というので8つ比較しているのから見ていただけたらと思います。真のモデルで追求すると、真のモデルを選びたいというのは否定しないけれども、結局はいろいろ分からないので、よい予測モデルを追求することを第一にして、このリスク評価が行われています。真のモデルという、つまり妥当なモデルというのはそういうことを表すのかもしれませんが、平均化前も後も決して妥当なモデルはどこにもないので、当てはめのよいモデルをより当てはめをよくしようというのがモデル平均化のフィロソフィーのようです。それは一定の検証を経た方法に基づいていて、アクセス可能なデータで数値的に最も妥当だと考えられるようなBMDLの決定を実現しているというのが、追求するところになっているというところがモデル平均化に基づくBMDLに関して、自分の言えることです。

一旦ここで止めさせていただきます。

○川村座長 ありがとうございます。

非常に重要なモデルの選択、核心に入ってきたところですが、御質問はありますでしょうか。モデルの当てはまりを見ているわけですが、特定の一つの実験で当てはまりを決めてしまっているのかということがすごく気になって、いろいろな実験を通して、このモデルが当てはまりがいいと決めるのだったら分かるのですが、特定の実験というのは偶然のばらつきがいっぱい入っているので、それで当てはまりがいい悪いと言ってしまっているのだろうかという疑問があります。

○西浦専門参考人 おっしゃるとおりだと思います。一つの限られたサンプルサイズの実験データ、ワンセットから得られるモデルというのは、そのワンセットが微分のときの最

良のモデルしか示すことができなくて、最も予測が良いというのは、それがリピートされるほうがよりよいというのは当然で、それは例えば疫学領域で言うメタアナリシスのようなことといます。こういうモデルでも平均化するフィロソフィーで使われることが、微生物のリスクアセスメントなどではやられるのですけれども、複数の実験でアップデートしていくというのが一番理想だとは思いますが。

○川村座長 ありがとうございます。いかがでしょうか。前回、私が説明する中でモデルはアプリアリに論理で決めるべきだと堂々と主張してしまったのですけれども、今の話を聞いて大分軌道修正が必要だと反省をしているところです。私は臨床予測モデルをよくつくるのですけれども、結果的には当てはまりというか、プレディクションのパフォーマンスを見て決めているので、考えたらやっていることは同じことで、やはりよい予測モデルを選ぶというのがある意味必然だと改めて考えているところです。ありがとうございました。

○小関専門委員 先生の意見を教えていただきたいのですけれども、よい予測というのは何をもってよいとしますか。結局、低用量域で外挿しなければいけない部分はいずれにしても、どこまで行っても真実がわからないところで、その部分を含めて、よい予測というのはどういうことを考えたらよろしいでしょうか。

○西浦専門参考人 観察データがどのドーズで得られているかにかかわらず、最も適合がよいと判断されるモデルがここで言う最もよいモデルです。ロードーズが最もよいかどうかというのは、ここではデータがないから議論ができないというのが一つのプリンシプルなものと、ロードーズに関する妥当性をどこで保証するかというのは、今までのNOAELとかのデータとBMDを比較しているようなものがありますよね。そういったプラクティスを重ねることによってBMD10とNOAELの比較というのが、ちょっとエビデンスが出てくると、S字曲線でロードーズ領域を予測していることの妥当性が少しずつ補完されてくる。そういうのを通じてやっていくものなので、そこは切り分けないといけないです。

○小関専門委員 ありがとうございます。

○川村座長 そうですね。モデルの当てはまりは全体の関数としてのことなので、ロードーズだけとかBMDLの当てはまりはある意味、歴史的に見るしかないかもしれないですね。よろしいでしょうか。

では、次へ進んでください。

○西浦専門参考人 長くなって済みません。駆け足で進みます。残っている部分ですけれども、あとはBMRの設定についてとRestriction ONというものがモデル選択をした後のBMDLの決定とかで行われるので、その話をさせていただきます。

(PP)

BMRの設定なので、BMDではBMRで推移される反応の増加に伴って推定されるもので、ドーズがこれを満たすというようなもので、BMRというのが頻繁にスライド内にあるこれらの値です、というのは前回でも触れていただいたところです。これはPバリューの

決定と似ていて、恣意的に1%、5%、10%という分布の裾の分位点を選ぶことによって決定が行われています。いずれも恣意的なものですけれども、コンベンションには10%でソフトウェアに設定されていることが多いです。これをコンベンションでデフォルトにするべきかどうかというと、ヨーロッパもアメリカもソフトウェアを使う人たちはちょっとずるくて、10%の設定しているのだけれども、これはデフォルトではありませんと言うのです。責任は取りませんという感じなのです。ただ、10%がコンベンションで受け入れられているというのが今の状況です。疫学データだけNOAELとBMD10という10%を切ったときの間の関係が少しぶれるので、1%とかを使ったほうが良いというようなサジェスションをするような研究もあります。BMDLというのはBMDの信頼区間の下限を指します。

(PP)

BMRを設定する上でBMD10を設定していますけれども、これはリスクが10%超過していますということがベースですというのが知っていただきたいことです。例えばバックグラウンドレートと言ってドーズがゼロのときにがんとかの発生率が15%で、追加リスクが8.5%だと超過リスクは $8.5 \div (100 - 15)$ とか、もともとのバックグラウンド分を差し引きしてあげないといけないので、それで10%になっているということに注意しないといけないということです。

さっきも言いましたけれども、BMD10はデフォルトではないとは言われていますけれども、コンベンションで受け入れられているのですが、それはなぜかと言うと、このテーブルは『EFSA Journal』からとってきたものですが、Quantal dataというので、ここがBMDのどの分位点くらいに相当するのを書いてあるものですが、NOAELが大体BMD10よりも上側にある事例がいっぱいありますよというようなのが経験則で今は得られていることが多くて、それをもとにNOAELより下を保証するために使えますよというので提唱されているものです。ただ、全ての化学物質についてユニバーサルにそんなことがわかっているわけではないです。経験的に10%でBMD10を定義すると大丈夫なことが多い。連続データだと5%を使ってくださいというサジェスションがありますということです。

(PP)

もう図を書いているのでお分かりかとは思いますが、BMD10はExtraのIncidenceをとるということです。Lowest doseで反応がゼロではないので、それをもとにここをゼロとしたときの10%の分位点をいつもとってプラクティスしているということ覚えていてください。

(PP)

BMRの設定に関してEPAが指示していることというのは、前回のワーキンググループとかでも広瀬先生がスライドで出してくださっていたことです。2値データというのがそういうドーズレスポンスのクオンタルレスポンスに相当するのですが、BMRの10%はスタンダードですということを言われているのですが、その理由はバイオアッセイの感度限界とかに相当して、それくらいでしょうということを言われていて、もっと低いものも生物

学的な理由があったら使ってくださいと、1%、5%も使ってくださいということなのですけれども、ただ、10は常にプレゼンすべきですというようなサジェスションが得られているというのが今のところでは。

連続データはデータによって特性が違うのですけれども、エンドポイントに応じて生物学的有意性を根拠に決めてくださいと。そうでなければ2値データに変換してでもいいから、分かりやすい分析をすることですということ。アイデアがなかったら、1SDの変化とかをもとに決めてくださいということでは言われているのですけれども、これは事例によって変えていくしかないというのが今の状況です。

(PP)

それでコンベンショナルに10%が使われているのですけれども、BMRの設定でその問題点なのですが、サンプルサイズの決定がそんなにたくさんフォーミュレートされているわけではないです。先ほどの議論から分かると思いますけれども、サンプルサイズが小さいとパラメータの信頼区間が広がりますので、信頼する下限値が低くなりますので、BMDLも低くなってしまいます。だから、最適なサンプルサイズというのは用量反応関係によって異なるわけで、そのときの低用量の反応によって得られる結論は結構変わってきますよということ。

例えばBMRの設定のためにミニマムに必要とされるサンプルサイズがどれくらいなのかというので、OECDとかのガイドラインを踏まえて物質別でどれくらいなのかというと、まだオープンクエストです。こういうのは研究で埋めていかないといけないのだろうと思います。動物実験で1群20頭とか、あるいはもっと少ないものを使って、どのドーズにどうやって割り振るのかというのは、実はそんなにきれいにBMD法を念頭にフォーミュレートされているわけではないので、まだこのあたりは群雄割拠しています。特にそのBMDというコンセプトが分位点の下の信頼区間なので、解析的に扱えるわけではないから結構難しいということです。ベイズ法とかでシミュレーションがたくさん必要で、一般化することが恐らく難しいのではないかと予期しています。

(PP)

最後はそのパラメータの制約に関して、Restrictionを書いていますけれども、これは広瀬先生が和訳してくださっているスライドと同じものをEFSAのジャーナルからとってきているものです。右側の端っこにConstraintsと言ってパラメータ制約を書いています。それぞれの関数について、こういう範囲でというのがあるのですけれども、こういうソフトをつくとやみくもにそのデータに8つのモデルをとにかく投入するということが多くて、それで出てきたパラメータを余り解釈せずについてしまうことがあるから、こういう制約をわざわざOnとかOffも両方してくださいというのをつくってくれているというのが背景の気を遣ってくれている感じです。

(PP)

なぜかと言うと、体育会系的に違うモデルをコレラとかに投げたのがこれです。『PLOS

Computational Biology』からとってきていますけれども、ドーズが変わって、いろいろなモデルをフィットしてみましたというのが左で、観察データはサンプルサイズ別にこの丸で書いているものです。上のほうのドーズと下のドーズが1点しかないのですけれども、すると予測するモデルで指数分布という紫のもので予測しているようなものと、ワイブルとここで書かれているものは全く違うようなものになってしまう。それを右側の図に直すと、予測されるモデルが全く性格の違うようなものになってしまうということです。この解釈をいつも生じるようにするのではなくて、一定の度合いで修飾した上で推定したいというのが、パラメータ制約がそういうソフトウェアについているところの目的とするものです。

(PP)

だから、非現実的なフィットを避けるということなのですからけれども、そのConstraintsというのですが、それはいろいろなものがあります。何かと等しくするとか、何かより大きくするとか、あるいは整数にするとか、いろいろなことがあるのですけれども、このConstraintsというのは実は数値計算法という領域で、統計よりもさらに数値の計算をコンピュータとかでやる人たちが専門にしている領域で、最適化プログラミングと言うのですけれども、どういうふうに数値をからからとコンピュータの中で回していくのが最適な計算かという議論で使われていて、詳しく研究されているのはそちらのフィールドです。数理的にはフィットがいいけれども、明らかに間違ったカーブで収束すると困るから、こういうことをするということなのですから、それでBMD法について言及されているのは、ワイブルでRestrictionがないときにパワーパラメータがうまくいかないですよということですが、100%未満のところではPlateauに達してしまうようなモデルがあったりすると、これはワイブルと違いますけれども、100%未満のところではPlateauに達するモデルがあると、それはそもそも観察された現象から対応していないとかいうようなことがあるのです。

(PP)

この右側の図がEFSAジャーナルからとってきたものですが、そういうRestrictionが必ずしもうまくいかないよという議論がされていて、それは前回のワーキンググループでも少し発言をしていただいたと思うのですが、もともとのデータを見ていると、すごく指数分布のような分布に見えるのですけれども、これがあるから、この立ち上げ部分に対応するために最初のスロープの部分が無限大にでかくなってしまふようなことが起こるのですが、ワイブルのフィットをそのまま何もせずにとここが無限になるとか、あるいはsteepness parameterというのがあるのですけれども、それが推定する情報量はほとんどないです。

なので、ここに制約をかけて観察データにフィットしましょうというのでRestrictionをやっていたら、ケアフルな研究者がもっとロードーズをじっくり見たら下の図のような感じになっているのではないかと、ログをとって確認をしたのがこの図です。下のドーズで反応が起こらない部分も含めて見てみると、きれいなS字曲線になっていて、Restriction

をかけずにこのデータにフィットさせたら、実はRestrictionがなくてもワイブルがうまくいったということなので、さっきのインフィニットな上がり方、無限な上がり方をするものに、Restrictionをかけるとロードーズの部分は誤った結論が出かねないですよというのをここに引用しているトキシコロジーの雑誌で議論されていますけれども、BMDLを人為的に高くしてしまうようなことにつながるので、ここはすごく注意をして観察しないといけないということが言われています。

(PP)

パラメータ制約はとても専門的で、関数の形がわかっていないといけないのですが、私も、私が今、研究室で立ち上げ2年目で20人くらい教えているのですが、その大学院生の人たちにいつもどういうConstraintsに関する指導をしているかというのがこの最後のスライドです。Constraintsがないと収束できない場合は、収束理由がちゃんときちり説明できないと、それは間違っていますよと教えています。あるいはConstraintsがないと収束できないモデルが誤っているときもあるから、落ち着いてデータをしっかりまず一緒に眺めましょうという話をしています。何が言いたいかというと、わからないまま絶対にこのConstraintsに関しては使ってはいけないということです。どうしてもわからない場合はONをデフォルトにして無批判に受けるよりもOFFもいつも同時に見るというような姿勢をするしかないというのが、今こういう立場からできるサジェスションです。

こういうこともありますし、今日何度か受けたロードーズ領域で正しいモデルは何かという話が、最終的に皆さんが追求する部分ですけれども、こういうときにどういうデータが今はふさわしいとわかっているのかを再確認した上でRestrictionも考えるべきですけれども、BMD法にふさわしいデータというのは観察データにちゃんとしたドーズレスポンスの量反応関係のトレンドが見られて、イベントが見られる2点以上と、できればコントロールというのがあって、イベントは低用量データができれば望ましくて、特に高用量から低用量の予測ばかりではなくて、LOAEL、NOAEL付近でたくさん実験をしているようなデータでサンプル数が多いものはやはり有用です。そういったデータをもとにやってみて、Constraintsのあるなしがまだ話として残るかどうかを見てみると、そんなに頭を悩ます問題ではないのかもしれませんということです。

以上です。済みません、長くなってしまいました、失礼しました。

○川村座長 ありがとうございます。

今のところはBMRの数値をどうするかということと、Restrictionをかけるイエス・オア・ノーという課題でしたけれども、BMRがかなり経験的にというか、ずっと1、5、10でされることが多かったと。10はスタンダードでしょうということで外すことはしないけれども、それがベストかどうかは議論の余地があるということだと思います。御質問はいかがでしょうか。

BMRの設定とか特に量的な指標の場合は1SDとかいうことになっていますけれども、それと本当の臨床的な意義との対比をやった研究というのはあるのでしょうか。

○西浦専門参考人 臨床的な意義というのは具体的に言うと。

○川村座長 臨床というか生体の本当のイベントと、イベントとは限らないですけども、主観的なものでも結構ですけども、例えば神経伝導速度などを評価に使ったことがあるのですが、どれだけ遅延するということがわかったとして、一体それがどれだけの意味を持つのかと。有意に落ちているけれども、それは意味があることかというようなことがかって評価の中で出てきたことがあるのです。

○西浦専門参考人 なるほど。そういう現象とはエンドポイント以外は対比をされていません。BMDLというのは単なる分位点の信頼区間にすぎないです。だから、選ばれたモデル、あるいは平均化したモデルが正しいとしたときに、その平均化したモデルの分位点の信頼区間下限なので、その平均化したモデル、あるいは選ばれたモデルを100回シミュレーションとかしていても、95%以上の確率でその分位点は必ずその点よりも上にあるというようなところがBMDLに相当するという、その統計学的な意味合いだけだと思います。

○川村座長 では、余り臨床的というか、あるいは生物学的なことと対比したことは余りされていないということですか。

○西浦専門参考人 私が理解している限りでは、そのローエストリミットの議論と現象との間の関係というのが何か論文になっているというのは、自分自身は知らないです。

○川村座長 他に御意見をお願いします。

○岡田専門委員 BMDLのときに、そのパラメータの信頼区間の下限でとるという話ですけども、その信頼区間の誤差は、例えば動物実験のときにある1匹をポジティブに判断するか、ネガティブに判断するかを非常に迷われていて、結局どちらかにしたと。そういうような影響は入ってきていないわけですか。与えられたデータはきちんとなってしまうと。

○西浦専門参考人 もし迷われるようなことがあったら、その影響が混入します。だから、エンドポイントという縦軸で対象とするイベントで発がんかどうかでどれくらい起こるか否かの定義にかかわっている問題なので、その定義の判断基準は物すごく明示的でないといけません。もし判断を迷うようなことがあると、そこは不確実なデータが混入してしまうことになるので、信頼区間云々の話よりもそのデータをどうハンドリングするかの問題になってしまうのではないかと思います。

○岡田専門委員 それで私がもし心配性だったら、もともとのデータにさかのぼって1件ずつプラスしてみて感度解析みたいな、そんなのもあり得るかなとは思うのですけれども、それとこのお話とは一応別の話という感じですか。

○西浦専門参考人 BMDLの設定というよりも、それは実験のデザイン側の問題だとは思いますが。感度分析でどちらかにプラスしてモデルを適合することはもちろんできますけれども、その場合はどちらをどうやって選択しようかという話が出てくるので、まあまあ難しいかなとは思いますが。

○川村座長 よろしいでしょうか。どうぞ。

○山田専門委員　ここで聞くことがふさわしいかどうかはわからないのですが、例えば動物試験とかであるLOELの相当する用量で複数の所見が出てきた場合に、どの所見を選べばいいかということについて、統計学的に見たときに注意すべき点とかがあれば、教えていただきたいなと思います。

○西浦専門参考人　特定のエンドポイントを判断するために2つの判断基準が出てくるということですか。

○山田専門委員　一般毒性だったら複数の臓器に影響が出てくる、あるいは血液学の数値データ、あるいは病理みたいに非常に動物数が少ないデータのときもあると思いますけれども、そういった時にどれを選べばいいかということについて。

○西浦専門参考人　何らかの病理所見の有無とかで検討するときとかで、複数のエンドポイントがあるときは、もうエンドポイントを複数のドーズレスポンスカーブを描くというのが一番なのだと思いますけれども、それもどちらかと言うと実験計画の話に関連する話なのですが、どの臓器の毒性をより重視するかでエンドポイントを1種類カットしてしまうのが一番分かりやすいのですが、どうしても複数を見たい場合は複数が見られるものを、エンドポイントの軸をこちらが腎毒性でこちらが肝毒性でここがドーズというような、そんなのもつくりようと思ったら、データが豊富であればできるので、それをやるというのは最終的に目的が合致するならやるべきではあるとは思いますが、普通はどこかエンドポイントをきれいに決めてしまって、その定義の一つについて次元にしたほうが簡単だとは思いますが。

○川村座長　ありがとうございます。統計学的というよりは実験計画とかですね。そのエンドポイントの合目的性というか、そういうようなところまでいく話だなと思います。

他にはよろしいでしょうか。では、吉田先生。

○吉田委員　これは先生の御意見というか御感想を伺いたいのですが、今のどういうエンドポイントを選ぶかということは実験モデルで、私も本当にそう思うのですが、私は毒性病理という顕微鏡を覗く仕事をしておりまして、その判断というのはある意味では非常に数値データのようにクリアなものではないですよ。そういう視覚で見た、ある意味では質的なもののようなものにこういったエンドポイントにということに対する先生の御感想といいますか、このように今まで感じていらしたということでもよろしいのですが、御感想などをお聞かせいただければ非常にありがたいです。

○西浦専門参考人　エンドポイントの定義はいつもたくさんこうやって共同研究をして現場の声を聞けば聞くほど、私たちは頭を悩ませるものです。統計学的なモデリングの研究でも疫学の研究でもそうですけれども、何らかの症例のデータを扱うときはやはりケースデフィニションをきれいにメソッドのセクションで書かないと、そういう分野の論文はリジェクトされてしまいますので、どれくらい客観的にディフェンドできるような定義をするかというのは、データをとる前に決めるのが一番理想的なのだと思いますけれども、とった後に最もはまりやすいものを専門家と一緒に議論することも往々にしてあるのですが、そのと

きに最もディフェンドが論理的にしやすいものを探していく作業をしないとイケないのですけれども、どこまで病理所見とかの目を見たものとかが客観化できるかというのが、そういうときに一番キーになるものだと理解しています。

○川村座長 ありがとうございます。極めて難しい質問になるかと思いますが、判断の標準化とか、そういう別の次元の話になるかと存じます。

他に御意見はよろしいですか。恐らく聞きたいことはいっぱいあると思うのですけれども、進行も少しビハインドになっていますので、このあたりで一つの区切りをつけさせていただければと思います。それでは、西浦先生、どうもありがとうございました。(拍手)

それでは、本日のまとめをしまいたいと思います。西浦先生からは、統計学的な見地からベンチマークドーズ法を適用する際の論点とそれぞれに対する統計学的な視点からの考え方について基本的な話から、そして、プラクティカルな話まで御説明いただいたと思います。今のお話の中から我々として疑問に思っている点、着地点が見出せそうなものもありますし、これからまだまだ議論を深めていかなければいけない問題もあるように思います。

今後の評価技術企画ワーキンググループでは、今のお話を基点として決定して、おおよそこれでいけそうだとするところは決定の方向に近づけますし、これではまだ決め切るには至らないということに関しては補強するための情報収集も必要であろうと思います。ワーキンググループはそういう方法論をきちんと詰めて、そして、最終的には我々が採用すべき手法を固定することになると思いますので、今のお話を基点として、完全に1回で分かるわけではないけれども、分かりかける進展は十分にあったと思いますので、これを各自十分に咀嚼していただいて、引用文献の分厚さは全然わからないけれども、2点ほど書籍も御紹介いただいたので、そういうところにも当たって各委員の使命として、それぞれの頭を整理して意見をまとめていただきたいと思います。いかがでしょうか。そういう方針についてはよろしいですか。

今日は本当に貴重な一歩だと思います。進め方についてもある程度、御納得いただいたようですので、今後について、次回以降はこういう視点で議論を詰めてまいりたいと思っております。事務局、準備のほうはよろしいですか。

○窪崎課長補佐 承知いたしました。次回の議事及びお集まりいただく日時につきましては、日程調整の上、決まり次第お知らせします。

○川村座長 ありがとうございます。まだもちろん完全に消化できたわけではないと思いますが、頭の中に素材が十分に入ってきたと思いますので、本日の議事としてはこれにて終了させていただきます。予定時間を大分超過しましたが、中身が濃かったので勘弁していただきたいと思います。午後の仕事に差し障りがある人もいらっしゃるかもしれませんが、中身の濃さで御勘弁いただきまして、今日の議事はこれにて終了いたしたいと思っております。どうもありがとうございました。