人に依存したシステムは精度が低いと思われる．ところが，臨床試験は，まさに人を対象とし，人が治療をし，人がデータを発生させている．いかに優れたデータ管理システムや解析プログラムを導入・構築しても，完全な臨床試験システムが構築できるはずがない．臨床試験は人を対象とする以上，想定外の事象は必ず起きるものであり，著者らの経験でも，そうした想定外のできごとへの対応を誤ると試験そのものを失敗に導いてしまうこともある．

多施設共同臨床試験のデータセンターの最も大きな役割は，機械的にデータを収集して解析結果を吐き出すことではない．むしろ，想定外の事象への対応を含めて'人'の手に寄らざるをえない部分の臨床試験の支援であるといえる．特に，日常診療の'片手間'で臨床試験に従事しつつ，よりよい治療を1日も早く届けることを念じて臨床試験に携わる試験責任者の研究事務局と施設の臨床医の間の情報共有やコミュニケーションの媒体となることこそが，'人'の手になるデータセンターが果たすべき重要な役割であると著者らは考える．臨床医間のコミュニケーションを支援し促進することは，有害事象情報の共有や，より適切な支持療法のノウハウの共有を通じて，臨床試験に参加してくださった患者さんの安全性を確保することにもつながると考える．臨床試験においても，大事なことは人と人とのコミュニケーションであることを強調して稿を終える．

文　献

1) 福田治彦：研究者主導のがん多施設共同臨床試験におけるデータマネージメント．日本小児臨床薬理学会雑誌　16(1)：76–82，2003.
2) 佐藤暁洋，福田治彦：JCOGについて．分子細胞治療　6(4)：55–59，2007.
3) 福田治彦：エビデンスをつくる人々．EBMジャーナル　2(3)：100–112，2001.
4) 福田治彦：「がん臨床試験の実践～JCOGを例に～」第2回「コンセプトの作成と審査 第三者に試験の意義を理解してもらう」．The MEDICAL Oncologists　3(3)：25–32，2007.

腫瘍内科，3(1)：126-131，2009

連載 臨床試験

臨床試験プロトコールの書き方1*

中村健一** 福田治彦**

Key Words：protocol, clinical trial, medical writing, quality control, cancer

はじめに

今号から数回にわたり，臨床試験プロトコールの書き方について概説する．筆者らは，がんの集学的治療を開発するための研究者主導臨床試験を行っている「日本臨床腫瘍研究グループ(Japan Clinical Oncology Group：JCOG)」(http://www.jcog.jp/)のデータセンター，運営事務局の一員として，JCOG試験の支援や管理に携わっている．JCOGデータセンターとJCOG運営事務局は，計約40名のスタッフにより，計約90の試験を支援・管理しているが，そのうち計画中の約20試験については，JCOG運営事務局の研究支援部門の5名のスタッフ(うち3名が医師)が中心となり，プロトコール作成やプロトコール改訂の支援を行っている．

本シリーズでは，研究支援部門の日常業務における経験に基づいて，プロトコール作成にかかわる最低限の知識とコツを示していきたい．

プロトコールとは？

「プロトコール」の定義であるが，臨床試験の古典的な教科書であるPocockのClinical Trials[1]には，「formal document specifying how the trial is to be conducted：どのように臨床試験が行われるのかが書かれた公式文書」と定義されている．一般的な日本語訳としては，「研究実施計画書」があてられるが，治験においては「GCP(医薬品の臨床試験の実施の基準に関する省令)」で「治験実施計画書」，臨床研究に関する倫理指針では「臨床研究計画書」とされている．いずれにしても，その研究の手順を詳述した「計画書」をプロトコールと呼ぶ．慣習的には，「当科では現在○本のプロトコールを走らせています」，あるいは「当科のプロトコールではシスプラチンを隔週で投与しています」など，「研究/試験」，あるいは「レジメン/治療法」の意味で用いられることもある．しかし，これらは誤りとまでは言えないが，"protocol"の原義が「外交儀典書」や「条約原案」といった「(公式)文書」であることから，混乱を避ける意味では，臨床試験やレジメンをプロトコールと呼ぶのではなく，文書である「計画書」をプロトコールと呼ぶことが推奨される．

臨床研究に関する倫理指針

被験者保護，および研究者が円滑に臨床研究を行えることを目的に定められた「臨床研究に関する倫理指針」(以下，臨床指針)が2008年7月に改正され，2009年4月より施行されることとなった．臨床指針では，あらゆる臨床研究において「研究責任者は，臨床研究計画において，臨床研究の実施計画および作業内容を明示しなければならない」と規定されており，さらに，それらの臨床研究計画が臨床指針に適合しているか否かを，あらかじめ倫理審査委員会に審査を行わせ

* How to write a clinical trial protocol.

** Kenichi NAKAMURA, M.D. & Haruhiko FUKUDA, M.D.：国立がんセンターがん対策情報センター多施設臨床試験・診療支援部〔〒104-0045 東京都中央区築地5-1-1〕；Clinical Trials and Practice Support Division, Center for Cancer Control and Information Services, National Cancer Center, Tokyo 104-0045, JAPAN

なければならないと記述されている．つまりは，プロトコールを書くことなく臨床研究を行ってはならず，プロトコールは倫理審査委員会で必ず審査されなければならないと明記されている．データの単純な集計にとどまる観察研究の場合の扱いはQ&Aに盛り込まれる予定であるため，詳細は厚生労働省のホームページ(http://www.mhlw.go.jp/general/seido/kousei/i-kenkyu/index.html#4)を参照いただきたいが，少なくとも，なんらかの介入を伴う臨床研究(すなわち臨床試験)の場合には，必ずプロトコールを作成し，倫理審査委員会の審査を受けることがこれまで以上に強く求められることになる．実際，改正後の臨床指針には，「3(11)厚生労働大臣等の調査への協力」として，臨床研究機関において臨床研究が臨床指針に従って行われているかどうかの適合性調査(いわゆる「監査」)が行われることが新たに盛り込まれており，実際に施設への訪問調査も行われる見込みである．倫理指針が刑事罰もあり得る連邦法である米国と異なり，わが国の臨床指針は法ではなく“ガイドライン”であって，罰則規定は設けられてはいない．しかし，臨床指針への重大な違反が発覚した場合には，社会的な批判は免れず，民事裁判における判断の根拠となることは容易に予想されることから，実質的にはかなりの強制力をもつと考えられる．

これまでプロトコールと言えば，治験で製薬企業が作成するもの，あるいは，大規模臨床試験でデータセンターが作成するものであって，「プロトコールを書くこと」は自らとは無縁のものと考えてきた臨床医も多いであろうが，今後は，これまでプロトコールなしでも事実上許容されていたと思われる単施設の研究においても，臨床医が自らの手でプロトコールを作成する必要があるであろう．そのため，プロトコールに何を，どのように書くべきか，という知識は臨床研究に携わる医師に欠くべからざるものになったと言える．

プロトコールのもつべき機能

1．多施設，多職種間のコミュニケーションツール

がん領域の治療開発において，第Ⅰ相試験や早期第II相試験といった「早期開発」の臨床試験は，通常がんセンターや大学病院などの専門病院が中心となり，単施設～少数施設の研究として行われる．早期開発の段階では，評価する治療法の開発経緯や過去の安全性情報など事前にわかっている情報が少ないため，プロトコールに盛り込むべき情報は比較的少ない．そして，未知の有害事象が生じる可能性が高いため，治療変更や有害事象に対する対処などは，同じ施設内で情報共有を密にして，臨機応変に対応すべき場面も多い．このような段階の試験では，治療変更規準や支持療法に関するプロトコールにおける取り決めはシンプルかつ柔軟にしておき，むしろ，早期開発の専門家としての担当医の臨床的判断や，担当医間の直接のディスカッションを優先することの方が，登録患者の安全性を確保するためには重要である．

これに対して，後期第II相試験や第III相試験といった「後期開発」の臨床試験は，多数の病院かつ一般病院も参加する多施設共同試験の形をとる．早期開発の段階とは異なり，担当医同士がリアルタイムに口頭で密に情報共有を行うことは不可能であるため，文書による情報共有が基本となる．後期開発では，これまでの治療開発経緯や早期開発段階での安全性情報などの情報量は格段に多くなるし，早期開発で得られたノウハウを治療変更規準や支持療法，検査スケジュールとして具体化することができるため，後期開発のプロトコールに含めるべき内容は必然的に多くなる．さらに，後期開発のプロトコールでは，より多くの施設のより多くの研究者がプロトコールの内容を一義的に理解する必要があることから，内容の正確性や整合性，客観性がより強く求められることとなる．そのため，(プロトコールは分厚ければ良いというものではないが)JCOGの第III相試験のプロトコールは，通常70～100ページ程度のボリュームとなっている．

このように，とくに後期開発ではプロトコールは「多施設」の研究者に読まれることを念頭において作成する必要があるが，プロトコールはさらに「多職種」間のコミュニケーションツールでもある．プロトコールの読み手としては，参加施設の研究者(医師)，臨床研究コーディネー

ター(CRC)，看護師などの"local user"のほかにも，データセンターや運営事務局の中央支援機構の統計家やデータマネージャー，モニターといった"central user"，さらに，プロトコール審査委員会や効果・安全性評価委員会，施設の倫理審査委員会の委員といった"reviewer"としてのユーザーが想定される．プロトコールとは，このように多職種かつさまざまな役割の人間が臨床試験を遂行していく上で用いる「取り扱い説明書(instruction manual)」であり，これらのすべての人に理解可能で，さらに，それらの人たちのニーズに応えたものでなければならない．プロトコールは該当する疾患の専門家が作成することが多いために，とかく難解で専門用語が羅列されているという印象があるかもしれないが，専門用語を用いるのなら，理解を助ける注釈がなければならないし，内容が論理的で読みやすくなければ，ユーザーフレンドリーとは言えず，コミュニケーションツールとして機能しない．

とくに，この"reviewer"としてのユーザーには，患者団体の代表やあるいは患者自身がなり得ることを念頭に置く必要がある．もちろん患者が読み手の中心ではないため，過度に非医療従事者向けの表現にする必要はないが，少なくとも患者が読んで不快に感じるような表現は避けるべきである．JCOG試験においては，求めがあればプロトコールを試験に参加する患者に渡すことも許容しているため，患者が不快に感じる可能性を最小化すべく，たとえば「症例」はできるだけ用いず，文脈に応じて「患者」や「の場合には」など，他の表現に置き換える工夫を行っている．

2．品質管理ツール

PiantadosiのClinical Trials[2]には，プロトコールとは「臨床試験を行うにあたってもっとも重要な品質管理ツールである」と書かれている．多施設共同試験では，異なる施設であっても，同じ適格規準を満たした患者に対して，同じ治療，同じ治療変更規準，同じ評価が行われることが原則である．施設によってばらばらの適格規準によりさまざまな背景の患者が登録され，施設独自のスケジュールや投与量で薬剤の投与が行われ，担当医ごとの治療変更規準により勝手に減量あるいは休止が行われた場合，その臨床試験の結果，たとえ「有効であり安全である」と結論されたとしても，「どういう治療が」有効であり安全であるのかが記述できない．つまり，その「有効性と安全性」を日常診療で再現することが不可能であることから，患者に役立つエビデンスを創ったことにはならない．われわれはこうした試験を「Study for study」と呼ぶ．いくら仮説が正しかろうと，臨床試験の「質」が悪ければ，正しい結果や意味のある結論は導かれないのである．臨床試験の「質」とは，結果・結論の「信頼性」であり，その社会的・科学的な「価値」である．

こうした臨床試験における「質」を保つためのさまざまな工夫・活動を，臨床試験の「品質管理(quality control)」と呼ぶ．品質管理とは，「臨床的仮説→プロトコール→治療→評価→解析→結論」と続く一連の臨床試験の流れのすべての段階で生じ得るエラーを最小化する作業のことであるが，臨床的仮説とそれに対応するプロトコールは，この流れの最上流に位置しており，プロトコールに「エラー」があれば，下流に位置する治療や解析の段階で「エラー」が増幅されつつ拡大再生産されることになってしまう．逆に，プロトコールの完成度が高ければ，あとはそれを遵守できれば，正しい結論が導けるとも言える．プロトコールが，「もっとも重要な品質管理ツール」であるとはそういう意味である．

多施設共同臨床試験のプロトコールを書く研究事務局(試験の責任者)にとっては，おのれがベストと信じるやり方で，すべての登録患者に同じ方針で治療を行い，その上で自分が興味をもっているclinical questionに対する答えを知りたいはずである．しかし，現実には，研究事務局がベストと考える治療法や治療変更規準，支持療法を，細部にわたるまで多施設のすべての担当医が同様にベストと考えるとは限らない．施設の担当医は，医師法に基づく裁量権をもっており，一般診療においては，自らの責任において自らがベストと考える(あるいはその雇用責任者である施設の診療方針に基づいた)医療行為を行っている．多施設共同試験における治療とは，ある意味そうした医師の医療行為の裁量権に制限を課す特殊な医療行為であると言える．数十人，時には数百人の他の医師の医療行為を，

一臨床医である研究事務局の責任において制限するのが多施設共同試験である．よって，治療法，治療変更規準，支持療法をプロトコールに記述する研究事務局には重大な責任があり，記述の不備により患者に生じた不利益は，第一義的には研究事務局の責任である．しかし，その責任は，一臨床医である研究事務局に負わせるには大きすぎる責任であることは論を待たない．そのために，多施設共同試験のプロトコールは，(JCOGでは，当該グループ内での十分な議論に基づいて作成され，さらにプロトコール審査委員会のpeer reviewによる審査・承認を得た上で)各施設の倫理審査委員会での第三者的審査に基づく医療機関の長の承認を得ることが不可欠なのである．プロトコールの審査・承認のプロセスは，研究事務局にとっては煩雑な事務的過程かもしれないが，実は研究事務局の医学的責任を分掌することで，研究事務局を保護する仕組みであることを忘れてはならない．

このことを踏まえれば，多施設共同試験のプロトコール作成にあたって，研究事務局が，事前に十分な議論を行って参加施設のすべての研究者の十分な合意を得ることや，プロトコールの記述が医学的に妥当であるだけでなく，論理的かつわかりやすくなければならないことは自明であろう．それは，責任を分掌してもらう"local user"，"central user"，"reviewer"すべての人に対する研究事務局の義務であり，また礼儀とも言えるであろう．JCOGでは研究支援部門でのプロトコール作成支援やプロトコール審査委員会審査で，誤字・脱字や日本語として不適切な表現の修正も行っているが，それはこうした考えに基づいている．日常診療だけでも多忙な施設の担当医に，プロトコールに規定したとおりの治療あるいは評価を行ってもらうのは容易ではないため，プロトコールが読みやすく，適切で，かつ十分に説得力のもったものであることは最低限の必要条件と言えるだろう．

一方，試験開始後の品質管理活動としては，試験の進捗中に各施設においてプロトコール規定どおりの治療，評価が行われているかどうかを定期的にチェックし，不適格の可能性やプロトコール逸脱，重篤な有害事象などの問題を施設側にフィードバックをかける「モニタリング」がなされる．つまり，プロトコールであるべき姿を示し，モニタリングによってそれらが遵守されていることを確認することが品質管理の大きな流れである．このモニタリングでは，個々の施設単位の問題だけでなく，多くの施設に共通する問題，すなわちプロトコールの記述の不備が見つかることも多い．いかに綿密に作成しても「完璧なプロトコール」はあり得ないため，プロトコールは試験開始後もモニタリングを通じて，常に改善・進歩していくものでもある．

参加する施設の研究者がプロトコールの内容をよく理解していないことや，プロトコール遵守の姿勢に欠ける研究者が中には存在することも事実であるが(明らかにプロトコールを読んでいないと思われる場合もある)，これらを改善する努力は絶えず行わなければならない．そのためには，臨床試験を開始する前に全参加施設の研究者がプロトコールの内容を吟味した上で，それに合意することが必要であり，さらにそれに加えて，試験開始後は正しい結果を出すために，「自分たちの作った」プロトコールを遵守するという意識を高く保つことが，プロトコールを通した品質管理のカギであると考えている．

3．試験の意義を示す声明文

さて，プロトコールとは臨床試験の「取り扱い説明書」だと述べたが，もう一つ，プロトコールはその臨床試験を計画した「意義」を示す声明文的な役割も担っている．米国National Institutes of Health (NIH)のEmanuelら[3)4)]は，臨床研究の8つの倫理要件を提唱しており(表1)，臨床試験のプロトコールでは，この中でも最初の4点，つまりSocial or scientific value, Scientific validity, Fair subject selection, Favorable risk-benefit ratioに十分に考慮されていることを示す必要があるとしている．一般的なプロトコールの章構成は次々回述べる予定であるが，これらのうちもっとも重要な「意義(Social or scientific value)」については，プロトコールの冒頭部分(JCOGでは2章「背景と試験計画の根拠」)に記載され，その臨床試験の「手順」については以降に続けられることが一般的である．これはreviewerにとっては，冒頭部分を読んで，この臨床試験

表1　臨床研究の8つの倫理要件(Emanuelらによる)

倫理要件	詳細要件
①Social or scientific value 社会的/科学的価値	診断・予防・治療の向上に貢献，疾患・健康に有用な知識を得る，すでにある知識や無駄な重複ではない
②Scientific validity 科学的妥当性	一般的に認められた科学的方法論 適切な統計手法・正しいデータ
③Fair subject selection 適正な被験者選択	社会的弱者の保護・過大なリスクのある被験者の除外 利益を受ける集団とリスクを受ける集団が解離しない
④Favorable risk-benefit ratio 適切なリスク/ベネフィットバランス	リスクの最小化・ベネフィットの最大化 被験者のリスクに見合う被験者/社会のベネフィット
⑤Independent review 第三者審査	研究と利害関係をもたない独立した第三者による デザイン・対象・リスク/ベネフィットの評価
⑥Informed consent インフォームドコンセント	研究目的・方法・リスク・ベネフィット・代替治療の充分な説明(information)，理解(comprehension)，自発同意(voluntariness)
⑦Respect for potential and enrolled subjects (候補者を含む)被験者の尊重	同意撤回の自由，プライバシー保護 開始後の新知見や研究結果の説明 継続的な被験者保護の監視
⑧Collaborative partnership 共同パートナーシップ	研究成果はコミュニティで共有 成果のみの搾取は防止 関係者は協調して研究を行う

の「意義」が理解できない場合や意義が乏しいと判断される場合には，その後の「手順」の記述を審査する価値がないということを意味する．また，local userにとっては，まず臨床試験の「意義」を十分に理解してから，実際の治療にあたるべきであることを意味している．このように，重要な「意義」の部分にSocial or scientific valueが十分かつ論理的に記述されている必要があるため，必然的に「背景」の記述は，ある程度の分量とならざるをえない(JCOGのプロトコールでは，通常10～15ページ程度)．

①～④の倫理要件を満たす内容が十分に記載されており，また説得力をもっているかどうかが，第三者により審査されて承認が得られた(Emanuel 8要件のIndependent review)のち，個々の患者に「充分な説明」を行って，「理解」してもらった上で，「自発同意」によるインフォームド・コンセントを得て(Emanuel 8要件のInformed consent)，はじめて臨床試験による治療の開始が正当化される．

プロトコールは，研究者のみが理解できればよいという内部文書ではなく，第三者の審査の洗礼を受ける対外的文書，すなわち「他人が読むための文書」である．とかく臨床試験のプロトコールというと，内部の取り決めのみが書かれた文書と捉えられがちであるが，他者に臨床試験の意義を理解させるためのツールでもあることは強く意識すべきである．

だれが書くか？

プロトコールの核となる部分の初稿は，もちろん実地臨床に携わり，その臨床試験におけるclinical questionをもっとも切実に感じている研究事務局が記載すべきである．具体的には，試験の意義，患者選択規準，治療，治療変更規準，安全性情報，評価スケジュールなどは臨床試験を立案した研究事務局自身が書くべき事項としてあげられる．しかし，実地臨床で多忙な臨床医が，時には100ページにもなるプロトコールを独力で作成することは不可能であり，また，多くの場合はじめて本格的なプロトコールを書く研究事務局が，プロトコール作成のノウハウをもたないことは当然である．JCOGでも古くは研究事務局自身がプロトコールを作成し，それがそのままプロトコール審査委員会へ提出されていたが，あまりにプロトコールの質のばらつき

が大きく，かつ標準化されていなかったため，データセンター長がプロトコール審査委員会提出前に，レビューもしくは直接部分執筆を行うようになり，その後プロトコールコーディネーター(医師)がデータセンター/運営事務局内での意見を取りまとめるとともに，自らも部分執筆を行う分担執筆体制へと変化してきた．そのため，現在ではプロトコール作成に関するノウハウは，JCOGの中でもプロトコールコーディネーターに蓄積されており，プロトコールコーディネーターは研究者と質疑応答を繰り返しながら不整合の修正，不足部分の追記を行って，研究者と二人三脚でプロトコールを完成へ近づけるとともに，これらのノウハウを最新のプロトコールに反映することで，プロトコールの質の向上ならびに標準化を図っている．

また，たとえば品質管理や品質保証，倫理や規制用件にかかわる定型的な部分や，データ管理や解析方法など，むしろcentral userであるデータセンター/運営事務局が記載すべき箇所も多い．現在，JCOGでは医師以外(non-MD)の研究支援部門スタッフが定型的な箇所を埋めて，書式を整えるとともに誤字・脱字や日本語表記のチェックを行っている．また，解析方法についてはデータセンターの統計家が記載し，当該試験担当のデータマネージャーがとくにデータ管理の部分に関してレビューを行っている．プロトコールコーディネーターはこれらのcentral userの意見をまとめ，研究事務局との窓口となってプロトコールを完成に導く役割を担っている．

現在，JCOGでは3名の医師がプロトコールコーディネーターを務めているが，米国の代表的な多施設共同試験グループであるSouthwest Oncology Group(SWOG：JCOGの兄貴分とも言える)のプロトコールコーディネーターは，すべてnon-MDである．全国的に医師不足が叫ばれている中，JCOGデータセンター/運営事務局の医師も限られているため，JCOGでもnon-MDのプロトコールコーディネーターの養成が必要と考えている．高度な医学的知識が求められることから，non-MDがプロトコールコーディネーターとなるためには，一定期間の修練が必要であろうが，プロトコール作成のスピードアップや，安定したプロトコールの質の担保といったメリットも大きく，JCOGでもnon-MDのプロトコールコーディネーター養成に着手したところである．

おわりに

臨床試験が多くの組織で行われるようになり，それと呼応するように臨床指針の改正がなされるなど，プロトコールにかかわる規制用件も厳しくかつ細かくなってきた．必然的にプロトコールに書かれるべき事項は増え，倫理審査委員会の審査も厳しくなってきており，「量」，「質」ともにこれまで以上の水準を求められている．多施設共同臨床試験のプロトコールは，もはや研究者ひとりの力で完成させられるものではなく，多施設・多職種の人々(これらの人々はユーザーでもある)が協力して完成させるものへと完全に移行したと言えるだろう．

今号では，プロトコールの「定義」，「機能」および「書き手」について概説を行った．次号では，引き続き「良いプロトコールの条件」や「プロトコールの標準化」などについて具体的な例をあげつつ解説を行う予定である．

文　献

1) Pocock SJ. Clinical Trials : A Practical Approach. Chichester : John Wiley ; 1983.
2) Piantadosi S. Clinical Trials : A Methodologic Perspective. 2nd ed. New York : John Wiley ; 2005.
3) Emanuel EJ, Wendler D, Grady C. What Makes Clinical Research Ethical? JAMA 2000 ; 283 : 2701-11.
4) Emanuel EJ, Wendler D, Killen J. What Makes Clinical Research in Developing Countries Ethical? The Benchmarks of Ethical Research. J Infect Dis 2004 ; 189 : 930-7.

*　　*　　*

臨床試験グループの現状と展望(JCOG)

Current status and future perspectives of Cooperative Group (JCOG: Japan Clinical Oncology Group)

中村健一[1]　福田治彦[2]　柴田大朗[2]

Key words：臨床試験，肺癌，Cooperative Group，intergroup study

1. JCOGについて

a. 歴　史

がんの後期治療開発を担う研究者主導の多施設共同研究グループは"Cooperative Group"と呼ばれる．米国では1950年代からCooperative Groupによる臨床試験という方法が，がんの治療開発に有用であるとの認識の下，国策として数多くのCooperative Groupが組織され，現在までに多くのエビデンスが発信されてきた．

これに対して日本では厚生省(当時)のがん研究助成金指定研究「がんの集学的治療の研究」班(末舛班：1978-86年)によって，本格的に多施設共同の前向き介入試験が開始されたが，その開始時期は米国から遅れること約30年，1985年頃である．この研究班は，その後「固形がんの集学的治療の研究」班(下山班：1987-98年)に引き継がれ，Cooperative Groupとしての体制整備が進められた．この下山班を中核としたCooperative Groupは，1990年に自らを「日本臨床腫瘍研究グループ(JCOG)」と命名し，1991年にはJCOG統計センター(現データセンター)を設置，中央登録によるランダム化比較試験を開始した[1]．その後，データマネージメントや統計解析を中心としてデータセンターの機能，人員を徐々に拡充し，現在約90の臨床試験を管理・運営する組織となった(登録中の試験は約25)．現在，JCOG全体での年間登録患者数は約1,500名である．

b. 組　織

SWOG(The Southwest Oncology Group)やEORTC(European Organization for Research and Treatment of Cancer)など，欧米のCooperative Groupは基本的にはすべて同じ基本構造を有している．JCOGも同様に，①研究実施主体である「臨床研究者集団」，②データ管理・解析などを行う「支援機構」としてのデータセンターと運営事務局，③「監視機構」としての役割をもつ各種委員会の3つの基本構造を有しており，これをJCOG代表者，JCOG運営委員会が統括する，という体制をとっている．

①の「臨床研究者集団」としては，1978年にはリンパ腫グループと食道がんグループの2グループしか存在しなかったが，現在は14の臓器別(専門領域別)サブグループがあり，グループ単位で試験を計画・実施している．全グループ合わせると，200弱の医療機関から約400の診療科，3,000名を超える研究者がJCOGに参加している．

②の「支援機構」としては，JCOGデータセンター(データ管理部門，統計部門，システム部門，総務部門)とJCOG運営事務局(研究支援部

[1]Kenichi Nakamura: JCOG Operations Office, Clinical Trials and Practice Support Division, Center for Cancer Control and Information Services, National Cancer Center　国立がんセンター　がん対策情報センター　臨床試験・診療支援部　JCOG運営事務局　[2]Haruhiko Fukuda, Taro Shibata: JCOG Data Center　同　JCOGデータセンター

0047-1852/08/¥40/頁/JCLS

門，企画調整部門，品質保証部門)から構成されており，現在合わせて約40名のスタッフにより運営されている．

③の「監視機構」は，プロトコールの審査・レビューを行う「プロトコール審査委員会」や，研究実施中の安全性の監視や中間解析の審査を行う「効果・安全性評価委員会」をはじめとした恒常的委員会(standing committee)に加え，「放射線治療委員会」や「病理委員会」に代表される専門委員会(discipline committee)，個々の問題に対処するための一時的小委員会(ad hoc committee)(例：遺伝子解析ポリシー作成のための「遺伝子研究小委員会」)からなっている．JCOGには14グループの様々な領域の専門家が存在するため，プロトコールの審査や有害事象の審査は，当該グループ以外の研究者による"peer review"形式を取っていることが特色である．そのため，例えば肺がん内科グループのプロトコールは，大腸がん外科グループやリンパ腫グループといった他グループの研究者にも理解可能かつ，説得力をもったものでなければ，コンセプト審査やフルプロトコール審査で不承認となりかねない．JCOGは試験の計画から開始までの時間がかかりすぎる，と批判されることがあるが，このpeer review systemは，臨床試験の基本である科学性・倫理性を担保するために有効な仕組みであると考えている．

2. 肺がん内科グループ

a. 歴　史

肺がん内科グループの歴史は1984年に始まる．最初に行われたスタディはJCOG8502で，当時の参加施設は5施設のみであった[2]．JCOG8502は小細胞肺癌を対象としたCAV療法，PE療法，CAV/PE併用療法の3群300名のランダム化比較試験であった．JCOG8502ではCAV/PE併用療法の有用性が確認され，その結果はJournal of the National Cancer Instituteに公表されている[3]．その後，肺がん内科グループの参加施設は38施設にまで増え，40以上の臨床試験を通じて，肺がん薬物療法に関する日本発のエビデンスを発信し続けている．

b. Pivotal study(JCOG9511)

新しい標準治療が確立するには，理想的には2つ以上の独立したランダム化比較試験の結果が必要であるとされるが，そうした，標準治療を変える根拠となる臨床試験のことを"pivotal study"と呼ぶ．JCOG肺がん内科グループが行ったpivotal studyとして代表的なのがJCOG9511である．

JCOG9511は，進展型小細胞肺癌患者に対し，それまでの標準治療であったエトポシド+シスプラチン併用療法(EP)に比して，新治療である塩酸イリノテカン+シスプラチン併用療法(IP)が全生存期間において上回るかどうか(優越性)を検証する第III相試験であった．当初は230名が予定登録数であったが，3年あまりで154名が登録された時点で第2回中間解析が行われ，多重性を考慮したうえで統計的有意にIP群の優越性が示されて有効中止となった．生存期間中央値はEP群の9.4カ月に対し，IP群で12.8カ月であり(片側層別ログランク $p=0.002$)，IP療法は進展型小細胞肺癌患者に対する新たな標準治療であると結論された．この結果は2002年のThe New England Journal of Medicineに掲載され[4]，小細胞肺癌の治療の発展に大きく貢献した．

c. 現在進行中，計画中の臨床試験

JCOG肺がん内科グループは，小細胞肺癌，非小細胞肺癌という分類のほかに，病期，年齢，初発/再発などによって複数の治療開発の対象を設定し，それぞれの対象に対して治療開発を行っている．現在進行中，計画中の試験の概要を**表1**に示す．

小細胞肺癌については，初発例では限局型と進展型に対象を分け，再発例はsensitive relapseとrefractory relapseに分けて臨床試験を行っている．早期非小細胞肺癌については肺がん外科グループでの治療開発が主となるが，進行非小細胞肺癌に対しては肺がん内科グループが主体となっている．対象の母集団が少ない試験を除いて，ほとんどが第III相試験の設定となっており，いずれの試験もpositive resultとなれば，世界の標準治療を変えるpivotal studyとな

表1 JCOG肺がん内科グループで進行中，計画中の臨床試験

	JCOG number	phase	対 象	標準治療	試験治療	primary endpoint	予定登録数
小細胞肺癌	JCOG0202	III	限局型(初発)	EP+RT→EP	EP+RT→IP	全生存期間	250
	JCOG0509	III	進展型(初発)	IP	AP	全生存期間	282
	JCOG0605	III	sensitive relapse	NGT	PEI	全生存期間	180
	PC705	II	refractory relapse	—	amrubicin	奏効割合	80 (予定)
非小細胞肺癌	JCOG0402	I/II	若年者局所進行	—	VNR+CDDP →gefitinib	安全性 (治療完遂割合)	37
	JCOG0301	III	高齢者IIIA，IIIB期	RT単独	CBDCA+RT	全生存期間	200
	PC704*	III	高齢者IV期	DOC単剤	DOC+CDDP	全生存期間	385 (予定)

*WJOGとの共同試験予定.
PC: コンセプト段階の試験.
EP: etoposide+cisplatin, RT: radiation therapy, IP: irinotecan+cisplatin, AP: amrubicin+cisplatin, NGT: nogitecan, PEI: cisplatin+etoposide+irinotecan, VNR: vinorelbine, CDDP: cisplatin, CBDCA: carboplatin, DOC: docetaxel.

表2 JCOG肺がん外科グループで計画中の臨床試験

	JCOG number	phase	対 象	標準治療	試験治療	primary endpoint	予定登録数
非小細胞肺癌	PC603*	II	臨床病期IA(<2cm)の末梢型画像的非浸潤癌	—	肺楔状切除	無再発生存期間	300 (予定)
	JCOG0802*	III	上記以外の臨床病期IA(<2cm)末梢型肺癌	肺葉切除	肺区域切除	全生存期間	1,100 (予定)
	JCOG0707	III	病理病期IA(>2cm), IB期	UFT(2年間)	S-1(1年間)	全生存期間	960 (予定)

*WJOGとの共同試験予定.
PC: コンセプト段階の試験.

りうる.

3. 肺がん外科グループ

a. 歴 史

肺がん外科グループは肺がん内科グループとほぼ同時期の1985年に設立された．最初に行われたスタディはJCOG8601で，病理病期IIIA期の非小細胞肺癌患者209名に対する，経過観察群とシスプラチン+ビンデシンによる術後補助化学療法群を比べるランダム化比較試験であった．残念ながら，この試験では術後補助化学療法の優越性が示されない結果に終わったが[5]，その後，肺がん外科グループでは術前や術後の補助化学療法を中心に約10試験が継続的に行われている．最近ではsuperior sulcus tumorに対して，術前化学放射線療法(マイトマイシン+ビンデシン+シスプラチン併用)+手術が有望な新治療であることを示した第II相試験の結果がJournal of Clinical Oncologyに掲載された[6]．

b. 現在計画中の臨床試験

肺がん外科グループの治療開発戦略には大きく，早期の肺がんに対して侵襲性を軽減する新しい治療としての縮小手術の開発と，治癒率や延命効果の向上を狙う補助化学療法の開発の2つがある．現在グループで計画中の3つの臨床試験を表2に示す．

1) 縮小手術に関する臨床試験

縮小手術に関しては，「胸部薄切CT所見によ

り，リンパ節郭清が不要な非浸潤癌は予測可能である」という仮説の下に，JCOG0201「胸部薄切CT所見に基づく肺野型早期肺癌の診断とその妥当性に関する研究」が行われた．腫瘍径が3cm以下で，胸部薄切CT上の充実濃度(C)の大きさの腫瘍最大径(T)に占める割合(C/T比)が0.5以下のものを「画像的非浸潤癌」とし，これが「病理学的非浸潤癌」と十分に一致するのであれば画像的非浸潤癌を縮小切除の対象としてよいという仮説を検証すべく811名を登録したが，primary endpointである特異度(病理学的浸潤癌のうち術前胸部薄切CTにより正しく浸潤癌と診断された患者の割合)が96.4%と，当初設定した判断規準よりも下回る結果となった[7]．その後の探索的な検討により，「腫瘍径が2cm以下かつC/T比が0.25」という新たな規準が，十分に高い特異度が得られる，つまり病理学的浸潤癌を「誤って」非浸潤癌と画像で評価してしまうことで過小切除に終わってしまう結果となる危険性が十分に低い規準であることが示唆された．

現在計画中の第II相試験(PC603)は，この新たな画像診断規準によって画像的非浸潤癌を判定し，この対象に対して試験治療である楔状切除の有用性を問う臨床試験である．また，画像的非浸潤癌「以外」の肺野型小型(<2cm)臨床病期IA非小細胞肺癌に対しては依然としてリンパ節郭清が必要と考えられるため，標準治療である肺葉切除に比して，試験治療であるリンパ節郭清を伴う区域切除が生存期間において劣っていないこと(非劣性)を検証する第III相試験を計画中である(JCOG0802)．

なお，これら手術に関する2つの臨床試験はJCOG肺がん外科グループのみでは十分な患者数が確保できないことが予想されるため，後述するように，WJOG(West Japan Oncology Group)との共同試験として計画中である．

2) 補助化学療法に関する臨床試験

縮小切除の対象になるような早期非小細胞肺癌を除いた病理病期IA(>2cm)，IB，II，IIIA期は術後補助化学療法の改良が治療開発の主要なテーマとなっている．詳細は他稿に譲るが，2006年のASCOで発表された，5つのシスプラチン併用術後補助化学療法の臨床試験のpooled analysisであるLACE(Lung Adjuvant Cisplatin Evaluation)の結果から，II期，III期ではシスプラチン併用術後補助化学療法の有用性が示された．しかし，IB期に関しては，その有用性は明らかではなく，肺がん外科グループでは前述の縮小手術の臨床試験の対象とならない病理病期IA(>2cm)，IB期の患者を対象に術後補助化学療法の第III相試験を計画中である(JCOG0707)．病理病期IA期，IB期の肺腺癌に対しては，日本肺癌術後補助化学療法研究会(Japan Lung Cancer Research Group: JLCRG)が，手術単独群を比較対照として，UFTの2年服用群の優越性を第III相試験によって示しているため[8]，JCOG0707での標準治療はUFT(2年間)とし，試験治療は最近他領域の術後補助化学療法でも高い有効性が示されているS-1(1年間)としている．予定登録数はこれまでの肺がん内科，肺がん外科グループの試験を通じて最多の1,000名近い登録数を予定しており，2008年前半には試験開始予定である．

4. WJOG(West Japan Oncology Group: 西日本がん研究機構)との共同試験

a. 計画中の試験

最近のJCOG肺がん内科，肺がん外科グループでの大きなトピックは，WJOGとの国内グループ間共同試験(intergroup study)の計画である．

計画中の試験は以下のとおり(数字はすべて計画段階のもの)．いずれの試験も2008年中には順次開始予定である．

[肺がん内科グループ]

PC704: 高齢者進行非小細胞肺がんに対するドセタキセルとドセタキセル+シスプラチン併用を比較する第III相ランダム化比較試験(予定登録数385名，登録期間4年，追跡期間1年)．

[肺がん外科グループ]

JCOG0802: 肺野型小型非小細胞肺癌に対する肺葉切除と縮小切除(区域切除)の第III相試験(予定登録数1,100名，予定登録期間3年，追

跡期間5年).

PC603: 胸部薄切CT所見に基づく肺野型早期肺癌に対する縮小切除の第II相試験(予定登録数300名，登録期間6年，追跡期間5年).

b. なぜ共同試験か？

海外では一般的なintergroup studyが，これまで国内では行われてこなかった．日本における共同試験といえば，むしろ国内の施設が海外の企業主導の治験や製造販売後臨床試験(市販後臨床試験)に参加したり，中央支援機構が確立された海外グループの試験に参加する，といった「国際」共同試験が主流であり，本格的な国内のCooperative Group間の共同試験は行われてこなかった．国内のグループ間で共同試験が行われてこなかった背景としては，日本ではCooperative Groupという概念自体が確立されておらず，各Cooperative Groupの中央機構の体制整備が不十分であったこと，グループ間の情報交換が不十分であったことなどがあげられる．

c. 国内共同試験のメリット，デメリット

共同試験を行うメリットとしては次の3点があげられる．

(1) 集積期間の短縮：肺がん領域に限らず，患者に貢献するエビデンスを産み出すのは早いに越したことはないが，特に肺がん治療の分野においては，海外から数千例規模のランダム化比較試験の結果が次々に公表される現状であり，それらを目の当たりにすると，登録スピードを上げて早期に結論を得るべく共同試験を計画するというのは自然な成り行きである．共同試験を行い果たして想定どおりの集積期間の短縮が得られるかどうかは未知であるが，肺がん分野での代表的なCooperative GroupであるJCOGとWJOGの共同試験で十分な成果が上げられるかどうかは，肺がん臨床試験における日本の「国力」を測るうえでも興味深いところである．

(2) 重複を防ぐ：同じコンセプトの試験を国内の2つのグループで同時に行うことは，2倍の登録患者，2倍の研究者・データセンターの労力が必要となり，しかも結論が出るのに時間がかかる，という大きなデメリットがある．UMIN-CTRなどの臨床試験登録システム(http://www.umin.ac.jp/ctr/index-j.htm)の整備により，類似した試験の重複は減っていくと期待されるが，臨床試験登録システムに情報が掲示されるのは試験開始時点であることが多い．実際には試験開始までには，試験計画段階からプロトコール作成に至るまで膨大な労力と時間が費やされるため，同じコンセプトの試験であれば試験立案段階から情報を共有し，共同試験を計画することが，リソースの無駄使いを減らすという意味では理想的である．その意味では今回の共同試験を一つの端緒としてグループ間の情報交換を密にすることで，無駄な試験の重複を減らすことにもつながることが期待される．

(3) グループの運営システムの質の向上：もう一つ重要なメリットに，相互のグループの運営システムを知ることによって，お互いの長所を取り入れ，結果的にグループ自体の質の向上が相互に望めるという点があげられる．共同試験ではお互いの運営システムが異なるため，多くの事項について擦り合わせが必要となる．例えば，重篤な有害事象発生時の迅速な情報共有の方法，定期的な安全性情報の交換・検討の方法，登録・割付方法，効果・安全性評価委員会の審査システム(改訂審査，中間解析審査，有害事象報告の審査など)，解析方法，結果公表時のauthorshipなど，運営面でクリアすべき問題は多い．これらの擦り合わせは試験計画段階で非常に苦労する点であるが，これらの検討を通して，双方のグループのシステムの長所・短所を相対的に把握することができ，結果的に自グループの運営システムの質の向上につながると考えられる．

共同試験を行うデメリットは，(3)で述べたことの裏返しであるが，上記の運営システムの手順を擦り合わせる煩雑さと，それに伴う試験の質の低下への危惧である．それらのデメリットを最小化するために，上記の手順について現在綿密な相互検討を行っているところである．

おわりに

本稿では，Cooperative Groupの一つである

JCOGの現状と展望をJCOG肺がん内科グループ，肺がん外科グループの活動を中心に概説した．JCOGでは，本稿で紹介した肺がん内科・外科グループによるWJOGとの共同試験のほかにも，胃がん外科グループが韓国との国際共同試験を間もなく開始する予定であり，他グループとのintergroup studyのための体制整備は当面のJCOGの重要課題である．これらのintergroup studyを行うグループの共通点は，「臨床医の片手間ではない恒常的なデータセンターを有すること」であり，この要件を満たすかぎりにおいて，今後は他の専門領域における他グループとの共同試験もありうる．JCOGが現在の形を取り始めた1980年代末には，まだまだ臨床医の手弁当による臨床試験が「美徳」であったことを思えば，20年を経て，上記の要件を満たすintergroup studyが実際に現実となったことには隔世の感がある．Cooperative Groupの整備が米国に30年遅れた日本であったが，ようやく，科学的・倫理的な臨床試験によるエビデンスに基づいて，臨床医ががん患者の治療に自信をもってあたれる時代が来たといえるのかもしれない．

謝辞　本稿の内容は主として，厚生労働省がん研究助成金指定研究班17指-2(主任研究者：西條長宏)ならびに17指-5(主任研究者：福田治彦)の研究に基づくものである．この場をお借りして，執筆に際して情報提供をいただいた，17指-2主任研究者(JCOG代表者)西條長宏先生，JCOG肺がん内科グループ代表者田村友秀先生，JCOG肺がん外科グループ代表者加藤治文先生に深謝いたします．

■文　献

1) 佐藤暁洋，福田治彦：JCOGについて．分子細胞治療　6(4)：55-59，2007.
2) JCOGホームページ(http://www.jcog.jp).
3) Fukuoka M, et al: Randomized trial of cyclophosphamide, doxorubicin, and vincristine versus cisplatin and etoposide versus alternation of these regimens in small-cell lung cancer. J Natl Cancer Inst　83(12)：855-861，1991.
4) Noda K, et al: Irinotecan plus cisplatin compared with etoposide plus cisplatin for extensive small-cell lung cancer. N Engl J Med　346(2)：85-91，2002.
5) Ohta M, et al: Adjuvant chemotherapy for completely resected stage III non-small-cell lung cancer. J Thorac Cardiovasc Surg　106(4)：703-708，1993.
6) Kunitoh H, et al: Phase II trial of preoperative chemoradiotherapy followed by surgical resection in patients with superior sulcus non-small-cell lung cancers: report of Japan Clinical Oncology Group trial 9806. J Clin Oncol　26(4)：644-649，2008.
7) Suzuki K, et al: Evaluation of radiologic diagnosis in peripheral clinical IA lung cancers—A prospective study for radiological diagnosis of peripheral early lung cancer (JCOG 0201). 42nd Annual Meeting of the American Society of Clinical Oncology, Jun 2006.
8) Kato H, et al: A randomized trial of adjuvant chemotherapy with uracil-tegafur for adenocarcinoma of the lung. N Engl J Med　350：1713-1721，2004.

腫瘍内科，1(2)：144-148，2007

特集 臨床試験

中間解析*

吉村健一**

Key Words：clinical trials, interim analysis, group sequential analysis, alpha-spending function, multiple comparison

はじめに

現在実施中の臨床試験の途中結果をみたところ，新治療に割り付けられた対象者の成績が標準治療に割り付けられた対象者のそれに比べて明らかに優越しているようにみえる状況に直面した場合，研究者としてどういうアクションをとるべきであろうか．あるいはこれとは反対に，試験計画時に抱いた期待に反して新治療群の成績が標準治療群に比べて明らかに劣っているように見える状況であった場合，研究者としてどういうアクションをとるべきであろうか．どちらの状況においても実際に観察された群間差の大きさに応じて適切なアクションをとるべきであることは当然であるが，試験を早期中止すべきかどうかについて臨床的観点，統計的観点，倫理的観点を総合して悩ましい決断を迫られる状況が現実には意外に多く存在する．近年，試験の実施中にその時点までに累積した結果を統計的に適切な方法を用いて評価すること(これを中間解析という)がしだいに一般的となっている．中間解析の目的は，その結果に基づいて試験継続か早期中止かを判断することである．臨床試験における統計的原則を述べた国際的ガイドライン〔International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) E9〕[1]においても以下のように定義されている．

中間解析　Interim Analysis

試験の正式な完了以前に，有効性または安全性に関して試験治療群間を比較することを意図して行われるあらゆる解析．

このガイドラインに従えば，臨床試験の途中で群間比較を行うことを中間解析といい，中間解析で行うことは生存期間などに代表される有効性に関する群間比較のみに限られず，毒性などの安全性に関する群間比較も含まれる．本稿では試験の正式な完了以前に行うこの種の群間比較のことを中間解析，試験の正式な完了以後に中間解析による早期中止がなかった場合に当該試験の主たる結論を下すために行う群間比較のことを最終解析と呼ぶこととする．

ここでは，この中間解析についてその概要を解説することを目的とする．

倫理的観点からみた中間解析

臨床試験ではヒトを対象者としており，統計的側面からのみではなく倫理的側面からも適切な研究デザインを採用することが要請される．

これから実施しようとする研究において中間

* Interim analysis.

** Kenichi YOSHIMURA, Ph.D.: 国立がんセンター がん対策情報センター 臨床試験・診療支援部/医学統計室長〔〒104-0045 東京都中央区築地5-1-1〕; Biostatistics and Epidemiology Section/Clinical Trials and Practice Support Division, Center for Cancer Control & Information Services, National Cancer Center, Tokyo 104-0045, JAPAN

解析を予め計画しておかないことは，研究者が試験を実施するより前にもっていた有効性や安全性に関する予想が誤りであった場合に対する倫理的な担保が欠落していることに対応する．先行する研究や生物学的な根拠にいくら精緻に基づいて当該試験を計画しようとも，実際には事前の期待を超えるほど大きく新治療群が標準治療群に比べて優れることや，これとは反対に事前の期待に反して新治療群が標準治療群よりも著しく劣ることなどは往々にしてありうる．前者の事前の期待を超えて新治療群が優る場合，最終解析時点よりも早期の時点であっても十分に検証的な結果を得ることが可能である．中間解析を実施することによって，実施しない場合に比べてより早期の段階で十分に検証的な結論を得ることができるのであれば倫理的な観点より好ましいと言える．また，試験途中で新治療群が著しく劣っていることが明らかになった場合に試験を早期中止できるように中間解析を適切に計画しておくことも倫理的な観点より好ましいと言える．世界医師会によるヘルシンキ宣言(ヒトを対象とする医学研究の倫理的原則)[2]においても以下のようにされている．

> *医師は，内在する危険が十分に評価され，しかもその危険を適切に管理できることが確信できない場合には，ヒトを対象とする医学研究に従事することを控えるべきである．医師は，利益よりも潜在する危険が高いと判断される場合，または有効かつ利益のある結果の決定的証拠が得られた場合には，すべての実験を中止しなければならない．*

臨床試験において中間解析を適切に計画する必要性が倫理的な観点からも支持されていると言える．

一方で，実際には早期中止にたる結果が得られていないのにもかかわらず，不適切な中間解析による結果に基づいて早期中止することも倫理的な問題となりうる．ここで不適切な中間解析というのは，統計的観点から不適切なもの，実施運営上不適切なものなど，不適切な手順で行われたもの全てを指す．たとえば，中間解析時の結果において見かけ上は臨床的に大きな差がある場合でも，後述する統計的に適切な調整を伴うと十分には検証されているとは言えない状況は大いにありうる．一般には，試験早期に行った中間解析であるほど臨床的印象と統計的調整を伴った結果の間での食い違いが生じやすい．統計的調整を伴うと早期中止に至る結果が得られないということは現在得られているものが偶然の範囲内であるということを一般に意味する．つまり，そのまま試験を継続した場合には中間解析時の印象とまったく異なるような結果が将来の解析時に得られる可能性が決して小さくないということである．このような状況では臨床的印象だけにとらわれることなく，統計的観点，倫理的観点をも総合した上で当該試験に対する適切な判断を行うべきである．

統計的観点からみた中間解析

中間解析を計画した試験デザインを採用する場合，最終解析に加えて中間解析を実施するという意味で計2回以上の群間比較を同一試験内で行うことになる．これより，統計的観点からは検定の多重性(multiplicity)と呼ばれる問題が生じる．

検定の多重性とは検定を2回以上行うことによって生じる．これが生じた場合には後述する適切な方法を用いない限り，試験全体でのαエラーの確率を事前に定めたもの以下に保つことができない．αエラーとは本当は差がないのに差があると誤って判断しまう誤りのことである．p値が5％以下となった場合に統計的有意と判断する(これを有意水準5％という)検定を1回のみ行う場合にαエラーを犯す確率は5％に等しくなる．これは仮説検定の定義から導かれるものであり，実際に得られたp値にはよらない．一方で，2つの検定をそれぞれの有意水準5％で行うとαエラーを犯す確率は5％より大きくなることが知られている．簡単に，例として男女どちらにおいてもそれぞれ新治療と標準治療を比較する状況，つまり男性のみに限って1回比較し，さらに女性のみに限って1回比較するという計2回の検定を行う状況考える．当然ながら1人の対象者が重複して男女どちらの属性ももつということはないため，統計的にはこの2回の検定は互いに独立であるという．独立なk

個の検定をそれぞれの有意水準を5％として行った場合，本当はどちらにおいても差がないのにも関わらず，少なくとも1つ以上を誤って統計的有意であると判断してしまう確率(これを試験全体でαエラーを犯す確率という)は$1-(1-\alpha)^k$により求められる．ここでαは個々の検定で用いる有意水準を表す．前述の$k=2$の状況においては試験全体でαエラーを犯す確率が$1-(1-0.05)^2=9.8\%$となる．つまり，本当は男女ともに新治療と標準治療の間に差がないのにもかかわらず，少なくとも男女のどちらか一方以上で統計的有意差があると誤って判断してしまう確率が，それぞれの検定で設定した有意水準5％の約2倍にあたる約10％であることを意味する．この試験全体でαエラーを犯す確率は，式から求められるとおり，検定の回数kが増えれば増えるほど上昇する．

中間解析における検定の多重性とは，前述した通り，中間解析と最終解析あわせて2回以上の群間比較を同一試験内で行うことから生じる．例として，全生存期間をプライマリーエンドポイントとして新治療と標準治療を比較するランダム化試験を考える．この試験では最終解析時点までに両群あわせて200イベントを観察するように計画されているとする．ここでは全生存期間に興味があるため，サンプルサイズではなくイベント数を単位とする．ログランク検定やCox回帰に代表される生存時間解析ではサンプルサイズにかかわらずイベント数が大きいほど一般により高い検出力が得られる．両群あわせて200イベントが観察された場合，標準治療に比べた新治療のハザード比0.67に対応する群間差に対して片側有意水準を5％とするログランク検定の検出力は約90％となる．ハザード比は群間差を表す指標の1つであり，生存時間解析において一般によく用いられるものである．さて，この例において100イベントが観察された時点で中間解析を1回行うこととする．この中間解析時点は最終解析時点に得られるイベントの100/200＝1/2が得られた時点である．これより，この中間解析は情報時間(information time)0.5の時点で行う解析であると統計学的に表現されることが一般的である．同様にして，最終解析は情報時間1の時点で行う解析であると表現できる．これらの設定の下，中間解析および最終解析どちらにおいても有意水準5％の検定により群間比較を行った場合，本当は差がないのにもかかわらず，少なくとも一方の解析時点で統計的有意と誤って判断してしまう確率(同様に試験全体でαエラーを犯す確率という)を数値計算により求めると約8％となる．独立な検定に対して求めたものに比べ，試験全体でのαエラーを犯す確率の上昇分がやや小さくなっているのは，2つの検定が独立ではなく相関をもっているからである．中間解析の際に解析対象となったイベントは当然ながら最終解析の際にも重複して解析対象となるという点において両者は独立でない．参考として，1試験内で行う中間解析回数ごとの試験全体でのαエラーを犯す確率をコンピュータ・シミュレーションにより求めたものを表1に示す．ここでは中間解析はすべて等間隔で行い，中間解析も最終解析もすべて有意水準5％の検定を用いて群間比較を行うという設定の下，それぞれ10,000,000個の疑似臨床試験データをコンピュータ上で発生させることによって数値的に求めた．

表1　中間解析を等間隔に行った場合の実施回数と試験全体でαエラーを犯す確率

中間解析実施回数	αエラーを犯す確率
0	5.0
1	8.0
2	10.1
3	11.7
4	13.0
5	14.1
6	15.0
7	15.8
8	16.5
9	17.2
10	17.8

検定の多重性の問題を調整するために用いる統計手法を一般に多重性調整法と呼ぶ．この多重性調整法では一般に試験全体でαエラーを犯す確率が名目水準以下となるように，個々の検定で用いる有意水準としてこの名目水準より小さめの値を用いる．よく知られる多重性調整法としてボンフェローニ法がある．検定が独立でない状況では，検定の間の相関が強くなるほど

このボンフェローニ法は過度に調整する傾向があるというデメリットをもつ．ただしその調整手順の単純さから得られるメリットとして，多くの状況に対してユニバーサルに用いることができる．ボンフェローニ法では個々の検定で用いる有意水準として，名目上定めた試験全体でαエラーを犯す確率をこれから行おうとする検定の数で割り算したものを用いる．つまり，名目上定めた確率を等分割してそれを個々の検定の有意水準として用いる．たとえば，試験全体でαエラーを犯す確率を0.05に定めた下でこれから10個の検定を行おうとする場合，ボンフェローニ法によって調整された個々の検定の有意水準は0.05/10＝0.005となる．これは個々の検定のp値が0.005以下となった場合に統計的有意，そうでない場合に統計的有意でないと判断することに対応する．検定が独立の場合には前述した式から試験全体でαエラーを犯す確率を求めると$1-(1-0.005)^{10}=0.0489$となり，確かにこれが名目上定めた確率0.05を超えないことが確認できる．行おうとする独立な検定が100個あろうとも，同様に個々の検定の有意水準0.05/100＝0.0005とすれば$1-(1-0.0005)^{100}=0.0488$となり，確かに名目上定めた確率0.05を超えない．ただし，検定が独立でない状況ではボンフェローニ法は必ずしも適切な多重性調整法とならない．一般に検定間の相関が強くなるほど，実際の試験全体でαエラーを犯す確率は過度に小さくなる．この確率を過小になるに従ってβエラーを犯す確率は過大となり，これにより個々の検定における検出力が低下し，さらには群間差の推定における精度が低下してしまう．

前述した通り，中間解析と最終解析の間には相関があるため，ボンフェローニ法などの簡易な多重性調整法は適切でない．当該試験において中間解析を複数回実施しようとする場合には，それら中間解析の間にも相関がある．中間解析を行うにあたっては中間解析に特化した多重性調整法である群逐次解析法(group sequential analysis)を用いることが一般的である．群逐次解析法としてさまざまな方法がこれまでに提案されているが，その中でもよく用いられる方法はα消費関数(alpha-spending function)と呼ばれる関数を利用する方法(これをα消費関数法という)である．ここではα消費関数法の数理的説明は省かせていただくものの，多重性調整の概要はボンフェローニ法の際に述べたものに通ずるものがあり，名目上定めた試験全体でαエラーを犯す確率を中間解析や最終解析などを行う度に分割して用いる方法である．時間が経過するにつれて累積した結果に基づいて中間解析の実施回数は増えていくとともに，中間解析ごとに試験全体でαエラーを犯す確率を小出しに使っていき，あたかも中間解析ごとにαを消費しているようにみえる．これより他の多重性調整法と同様に，個々の検定で用いる有意水準は試験全体でαエラーを犯す確率より小さめの値を用いることになる．ボンフェローニ法と大きく異なる点は，試験全体でαエラーを犯す確率の分割の仕方である．群逐次解析法では，検定ごとに等分割することは必ずしも一般的ではなく，また前述の通り検定が独立ではないことから検定の間にある相関を適切に考慮した上で個々の検定でも用いる有意水準を定める．α消費関数の関数型としては一定の範囲内であれば任意のものを用いることが実際上は可能であるが，がん領域で頻繁に用いられるものはO'Brien-Flemingタイプのα消費関数である．このO'Brien-Flemingタイプのα消費関数は試験早期には早期中止する可能性が非常に低い一方で，試験が進むにつれて早期中止する可能性が徐々に高くなる傾向をもつ．試験早期ほどサンプルサイズも小さく，これより必然的に情報量も少なくなる．そのような試験早期の状況で早期中止する可能性を低くする一方で，情報量がより多くなる最終解析に近い時点であるほど早期中止されやすいことは一般的に臨床的にも受け入れやすい性質であるといえる．もちろん中間解析においても統計的に偶然誤差を加味した上で判断を行うわけであるが，結果の一般化可能性の観点から考えるに情報量の多い結果の方が系統的誤差(偏り，バイアス bias)が少ないと一般的に考えることができるためである．

ほかにもO'Brien-Flemingタイプに並列してテキストに紹介されることが多いα消費関数の関数型としてはPocockタイプがある．しかしながら，これが用いられる事例はがん領域に限らずともきわめて稀である．参考までに，Pocockタ

イプのα消費関数を用いた場合，それまでに累積した情報量によらず等しくαを消費する．つまり，たとえ試験早期であろうとも最終解析であろうとも等しくαを消費するため，O'Brien-Flemingタイプの関数に比べて試験早期であっても早期中止する可能性が高い．

通常用いられる群逐次解析法では，中間解析時点や中間解析回数が結果に依存しないことを前提にして多重性調整を行っていることには注意が必要である．たとえば，1回目の中間解析の結果，見かけ上の群間差が臨床的には十分過ぎるほどに存在していたものの，統計的調整を伴うとp値が有意水準よりもわずかに大きかったため統計的有意とは判断できないという結果を得たとしよう．この場合，研究者としてはこれほどの群間差が存在するならば可能な限り早く試験を中止して結果を公表すべきであるため，次回の第二回中間解析時期の前倒し，あるいは事前に計画されていなかった中間解析の追加を考えたくなるかもしれない．しかしながら，このように実際の結果に依存した中間解析時期の決定，あるいは中間解析の追加を行うとαエラーを犯す確率が名目上定めたものを超えてしまう．つまり，検定の多重性が存在する状況と同様に，本当は差がないのにもかかわらず誤って群間差があると判断してしまう確率を一定範囲内に制御することができなくなってしまう．たしかに本当にこれほどの群間差があるのであれば倫理的にも試験を早期中止すべきであると考えられるかもしれないものの，良かれと思って行ったそのような対応によって結果的にαエラーを犯す確率が高まってしまう危険性を有するのである．後述する独立データモニタリング委員会によって当該試験に関わる研究者とは独立にこのような判断を行おうともこの危険性は避けられない．つまり，このような状況においても決して安易な対応をするべきでなく，その必要性を臨床的観点，統計的観点，倫理的観点を総合して吟味した上で適切な判断を行うべきである．

実施運営の観点からみた中間解析

中間解析の結果として統計的有意ではなく，これにより試験継続となった場合にも，その中間解析の結果を当該試験に関わる研究者が見てしまうと，有形無形を問わず，試験の適切な運営に影響を与えてしまう可能性がある．たとえば，中間解析時に統計的有意でないものの，新治療が標準治療に比べて見かけ上優れているような結果が得られたとする．もしもその結果を当該試験に関わる研究者が見てしまった場合には，中間解析前後で試験に登録される対象者の属性などが大きく異なってしまいうる．このような場合，最終解析の結果が一般性に乏しい結果となってしまうかもしれない．あるいは新治療が一般的にも用いることが可能な治療なのであれば，その試験はいまだ検証的結果を得てはいないのにもかかわらず，中間解析を境に登録なされにくくなってしまうことで試験自体が検証的結果を得られないものになってしまうかもしれない．中間解析の実施運営上，その結果を評価する目的で当該試験に関わる研究者とは独立な委員会(これを独立データモニタリング委員会という，効果安全性評価委員会等と表現されることもある)を設置することが一般的である．通常，中間解析結果は独立データモニタリング委員会の構成メンバーによって詳細に評価，検討され，試験継続に関する結論のみが当該試験に関わる研究者に伝わるような形式をとる．構成メンバーには当該領域に精通した臨床家や統計家らを含むことが望ましいとされる．統計家も当該試験に関わっておらず独立であることが好ましいという意見もあるものの，必ずしも世界的にコンセンサスが得ているわけではない．統計家の独立性に関しては現在でもその良し悪しに関する議論が存在しており，世界的にみると必ずしも独立となっていないのが現状である．

文　献

1) 厚生省医薬安全局審査管理課．臨床試験のための統計的原則（平成10年11月30日付医薬審第1047号）．1998.
2) World Medical Association. Declaration of Helsinki. Human Participant Protections Education for Research Teams.(In：日本医師会・訳．ヘルシンキ宣言．ヒトを対象とする医学研究の倫理的原則)．2004.

腫瘍学における統計学

医学領域で用いられる統計学は一般に生物統計学とよばれ，腫瘍学（とくに臨床腫瘍学）の発展にも大きく寄与してきた。

臨床腫瘍学に関する臨床研究の例として，ここではKudohら（2008）により非小細胞肺癌患者を対象者としてゲフィニチブ市販後に実施された観察研究を取り上げる。この研究では，EGFRチロシンキナーゼ阻害薬であるゲフィチニブ（gefitinib）と急性肺障害・間質性肺炎（ILD）の間の関連の評価を主な目的の1つとし，ゲフィチニブ投与の有無によらず結果としてILDを発症した患者（これをケースという）1例に対してILDを発症しなかった患者（これをコントロールという）4例を選択してそれぞれについて詳細なデータが収集された。

EGFR：epidermal growth factor receptor（上皮成長因子受容体）

ILD：interstitial lung disease（急性肺障害・間質性肺炎）

オッズ比

この研究の結果を【表1】に示した。もしもゲフィチニブ投与とILD発症の間に関連がないならば，ケースにおけるゲフィチニブの投与割合はコントロールにおけるそれと同程度になることが期待できる。ケース・コントロール研究では，この関連を評価・定量するために一般にオッズ比とよばれる指標を用いる。以降，オッズ比とは【表2】で定義されるものをさす。このオッズ比はゲフィチニブ投与とILD発症に関連がない場合に値が1となり，ゲフィチニブ投与によってILD発症が増えるのであれば1よりも大きな値となる。前述のデータから【表3】のようにオッズ比を求めると2.35となる。

ここで求めたオッズ比の値をもって，ゲフィチニブ投与によりILD発症が増えたと結論づけてもよいであろうか。答えは"No"である。その理由として，①観察された結果はまったくの偶然であり，真実としてはゲフィチニブ投与とILD発症の間には関連がないかもしれないこと（偶然性），②病状が進んだ患者はそもそもILDを発症しやすく，そのような患者にゲフィチニブが投与される傾向があるかもしれないこと（交絡），③ILDの診断は難しいことからケースも真にはILDではないかもしれないこと（誤分類）があげられる。少なくともこの3つの可能性を十分に排除できない限り，前述の結論は必ずしも適切でないといえる。以下ではこの例を題材にして①～③の順に統計学の枠組で議論を行ってみることにする。

95%信頼区間

①は観察結果に含まれる偶然性を評価することの重要性を表している。真実としては関連がないのにもかかわらず偶然的な誤差（誤差的な挙動）によって2.35と観察された可能性が高いのであれば，前述の結論を出すのは早計であろう。

統計学を用いれば，観察結果に含まれる偶然性を定量化し，これを範囲・区間として明示することも可能である。医学研究では，この区間として慣習的に95%信頼区間を用いることが多い。95%信頼区間とは，仮に何度も研究を繰り返し，それら研究ごとに95%信頼区間を求めたとした場合に正しい値（真値）を100回中95回はその区間内に含むように構成したものである。ただし実用上は，「求めた95%信頼区間内に真値を含む確率が95%である」との解釈を行ったとしても大きな問題とはならない。

適切な統計手法を用い，【表1】のデータに基づいてオッズ比の95%信頼区間を求めると1.56～3.52となる。偶然性を考慮して構成された区間にオッズ比1を含まないため，まったくの偶然のみで2.35が観察された可能性は十分に低いと判断できるであろう。

交絡と層別解析

次に②は交絡（confounding）という現象を表しており，この交絡の原因となる因子を交絡因子という。病状が進むとゲフィチニブが投与される傾向があるとともに，ゲフィチニブ投与の有無によらずILDを発症しやすくなるという構造【図1】がある場合，病状が交絡因子となっており，ゲフィチニブ投与とILD発症の間に観察された関連はこの病状を介した見かけ上のものであり，必ずしも真にこのような関連があるわけでない。

適切な統計的手法を用いれば，この交絡の影響を減らすことができる。例えば，【図1】の構造が正しいのであれば病状が進んでいない患者のみに限って先の【表1】のように結果をまとめた場合，病状が同程度である患者の間ではこの影響が小さくなることから交絡の影響を減らすことができるであろう。一方で病状が進んだ患者のみに限って評価を行っても同様のことがいえる。このようなアプローチを層別解析といい，交絡を考慮・調整する目的でしばしば用いられる。層別解析のほかにも，これを一般化したものに相当するロジスティック回帰など統計モデルを用いて調整を行うことも一般的である。Kudohら（2008）は，モデルを用いて年齢，全身状態，罹病期間，心疾患，合併症，喫煙状況などの交絡因子となりうる主なものを調整して求めたオッズ比として3.23（95%信頼区間1.94～5.40）も報告している。この結果に基づけば，これらを調整したとしても，ゲフィチニブ投与とILD発症の間の関連は否定できないと判断できるであろう。

表1　ゲフィチニブ投与の有無とILD発症の関連

（人）

	ゲフィチニブ投与 有	無	計
ケース	79	43	122
コントロール	252	322	574

表2　オッズ比の算出方法

（人）

	ゲフィチニブ投与 有	無
ケース	a	b
コントロール	c	d

$$\text{オッズ比} = \frac{a}{b} \Big/ \frac{c}{d} = \frac{ad}{bc}$$

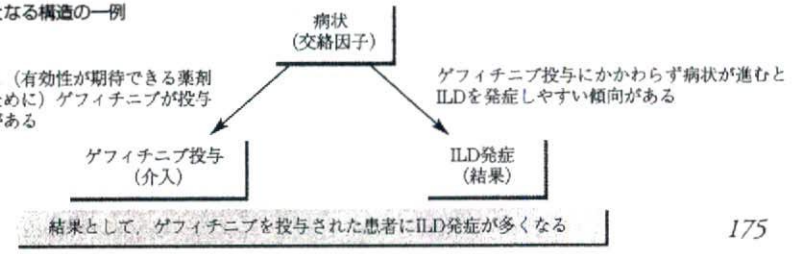
表3　ゲフィチニブ投与の有無とILD発症の関連に関するオッズ比

（人）

	ゲフィチニブ投与 有	無	
ケース	79	43	→ ケースにおける投与オッズ＝79/43
コントロール	252	322	→ コントロールにおける投与オッズ＝252/322

オッズ比 ＝（79/43）／（252/322）＝2.35
オッズ比の95%信頼区間：1.56～3.52

図1　病状が交絡因子となる構造の一例



感度・特異度

最後に③は誤分類とよばれる現象を表す。ここで誤分類とは真にはILDを発症していない患者がILDと診断されること（偽陽性），または真にはILDを発症している患者がILDと診断されないこと（偽陰性の両方）をさす。誤分類が存在する場合，正しいオッズ比とは異なる偏り（バイアス）のある結果が得られる可能性が高まることから注意が必要である。

誤分類の程度を表す指標として感度（sensitivity）と特異度（specificity）がある。ここで感度とは真にILDを発症している患者を正しく診断できた割合，特異度とは真にILDを発症していない患者を正しく診断できた割合にそれぞれ対応する【表4】。感度，特異度ともに値が大きいほどその診断能力が高いこと，すなわち誤分類が少ないことを表す。

適切な統計手法・研究デザインを用いれば誤分類による影響を減らすことも可能である。高度な統計的考え方を必要とするために詳細は省略するが，Kudohら（2008）では誤分類を考慮した解析の結果も提示してその評価を行っている。

ここでは1つの研究例を用いて解説を試みたが，統計学が臨床腫瘍学の発展に大きく寄与してきたことが少しでも伝われば幸いである。エビデンスに基づく医療（EBM）を実践する観点からも，適切な統計手法を用いてデザインされた研究，およびそれらのデータを解析した結果に対して適切な解釈が行える素養を身につけることは重要であり，その重要性はますます高まっている。また実際に一研究者として研究を行う場合にも統計学的考え方を念頭におき，適切に研究を行うことがますます求められている。

EBM：evidence-based medicine（エビデンスに基づく医療）

表4　真のILD発症と診断されたILD発症

（人）

		実際に診断されたILD発症 有	無
真のILD発症	有	x	y
	無	z	w

感度 $= x / (x+y)$
特異度 $= w / (z+w)$

Basic Point　public health

●罹患率

疾病に新たにかかることを罹患といい，特定の集団においてある期間に発生した特定の疾患の新規患者数を，その疾病に新たにかかるリスクにあった人口で除したものを罹患率（incidence）という。特定の集団においてある時点で特定の疾患を有する者の割合を表す有病率（prevalence）とは異なる。予防対策に用いる指標としては有病率よりも罹患率を用いるほうが適切である。

Basic Point　public health

●死亡率

特定の集団において単位時間における死亡者数の割合を死亡率（粗死亡率）という。単位時間としては1年が用いられることが多い。集団間で年齢構成が異なるような場合，年齢構成による交絡のため，粗死亡率では適切な比較が行えない。例えば，昭和60年と平成15年の間で死亡率の比較を行う場合がこれに該当する。この場合には，直接法により率を標準化（standardization）することで得られる年齢調整死亡率や，間接法により率を標準化することで得られる標準化死亡比（standardized mortality ratio: SMR）を用いることが一般的である。

Level up View

●がん対策基本法（法令番号：平成18年法律第98号）

がん対策基本法は第164回通常国会において提出された議員立法であり，2006年6月6日成立し，2007年4月1日に施行された法律である。これを期に，わが国は更なるがん対策を推進することとなった。

がん対策基本法の第一条には厳しい現状認識とともにその目的がうたわれている【表A】。

がん対策基本法には，①がんの予防と早期発見の推進，②がん医療の均てん化の促進，③がん研究の推進が本法の3本柱として明示されており，行政としてこれらを推進していくことが期待される。

表A　がん対策基本法 第一条

この法律は，我が国のがん対策がこれまでの取組により進展し，成果を収めてきたものの，なお，がんが国民の疾病による死亡の最大の原因となっている等がんが国民の生命及び健康にとって重大な問題となっている現状にかんがみ，がん対策の一層の充実を図るため，がん対策に関し，基本理念を定め，国，地方公共団体，医療保険者，国民及び医師等の責務を明らかにし，並びにがん対策の推進に関する計画の策定について定めるとともに，がん対策の基本となる事項を定めることにより，がん対策を総合的かつ計画的に推進することを目的とする。

○文献

1) Kudoh S, et al.：Interstitial lung disease in Japanese patients with lung cancer: a cohort and nested case-control study. Am J Respir Crit Care Med,177:1348-1357, 2008.

Self Check

- □ 95%信頼区間は偶然性を考慮して結果を提示する目的で用いられる。
- □ 交絡が存在する場合，観察結果には偏り（バイアス）が含まれる。
- □ 感度，特異度は診断能力の指標として用いられる。
- □ 感度とは，真には疾患をもつ患者に対して正しく診断できる確率である。
- □ 特異度とは，真には疾患をもたない患者に対して正しく診断できる確率である。

〈吉村健一〉