EBM と生物統計学

Evidence-Based Medicine and Biostatistics

折笠 秀樹

Hideki Origasa

富山医科薬科大学

Toyama Medical and Pharmaceutical University

1. EBM と は

EBM とは Evidence-Based Medicine の略で、日本語では根拠に基づく医療と訳されている（久繁, 1998; 名郷, 1999）。根拠というのは何かというと、それはいわゆる臨床研究のことである。もっと言うと、臨床研究で得られたデータのことである。従って、データに基づいて診療を行うという、きわめて自然なことなのである。データとなると、そこには統計学の入る余地がある。実際にもそういった理由から、臨床研究において統計学や臨床疫学という学問が重視されるようになった。

EBM という言葉が登場したのは 1992 年のことである（Evidence-Based Medicine Working Group, 1992）。カナダにいる Sackett et al. (2000) によって広められた。EBM というのは概念とか枠組みといった抽象論の話ではなく、診療の実践そのものである。従って、EBM の定義は以下のようになされる。目の前の患者さんに対して、その疑問を明らかにしたうえで、それに関連する現在入手できる最新情報を踏まえ、さらにその患者さんの置かれる状況を勘案して、その患者さんにベストな処方を与えるための行動指針である。言われてみると当たりまえの話だが、「最新情報を踏まえ」という点がポイントである。IT 時代にあって情報へのアクセスが容易になったゆえ、EBM もそれによって台頭してきたと思われる。この EBM であるが、表 1 に示したような 4 つのステップからなる。

表 1. EBM 実践における 4 ステップ

ステップ	そこで必要とされること
疑問の定式化	分析能力
情報の収集	IT 利用術
情報の吟味	生物統計学・臨床疫学 (Chance と Bias)
患者への応用	コミュニケーション技術

EBM では根拠となる臨床研究データを扱うわけであるが、それを信憑性の面から分類している。それは主として研究デザインの面（折笠, 1995）から行っている。分類法には様々な提案があるが、表 2 のように信憑性の高い順でまとめた。水準 1 には結果がほぼ一致した RCT（ランダム化比較試験）によるメタアナリシス、大規模な（つまり信頼区間が狭い）RCT、それから臨床研究をするまでもなく明白な事実が含まれている。水準 2 としては、RCT ではなく観察研究によるメタアナリシス、大規模ではない RCT、そしてランダム割付けを伴わない比較試験（CCT: Controlled Clinical Trials）が含まれる。最後の CCT というのは、メタアナリシス又は系統的レビューを国際組織で実施している、コクラングループで考案されたデザイン名である。水準 3 には Cohort 研究が含まれ、水準 4 には Case-control 研究、そして実験研究として前後比較研究が含まれる。水準 5 は Cross-sectional 研究、実験研究として症例集積が含まれる。最後の水準 6 は研究ではなく、単なる経験だけに基づいているものである。

表 2. EBM における根拠水準の順位付け

水準	それに該当する研究デザイン（一部、結果内容も加味）
1a	一致した結果を示す RCT によるメタアナリシス
1b	狭い信頼区間を示す RCT
1c	明白な結果
2a	一致した結果を示す観察研究によるメタアナリシス, RCT
2b	CCT
3	Cohort 研究
4	Case-control 研究, 前後比較研究
5	Cross-sectional 研究, 症例集積
6	単なる意見や経験のたぐい

このように見てくると、生物統計学の中でも研究デザインという観点が EBM で非常に重視されていることがわかる（折笠, 2000; 折笠, 2001）。また、結果として狭い信頼区間というように、統計学の守備範囲である Chance（偶然）という考え方も必須となっている。EBM ではこうした Chance という考え方に加えて、Bias という概念も大切である。後者はどちらかという疫学で重視される概念である。総括すると、臨床疫学を駆使して Bias の入っていないデザインかどうかを選別し、得られた結果が偶然ではなく確かなものかどうかは統計学の Chance という観点から判断する。

結果に Bias の入りにくいデザインとして RCT があるが、EBM ではそれが重要なものとして取り上げられている。ただ RCT であれば大丈夫かという、それは誤りである。そういう理由から、RCT 研究の善し悪しを点検するためのリストが統計家を中心に考案されてきた（Moher et al, 2001）。それを表 3 に示したが、CONSORT 声明と呼ばれている。

表 3. RCT 論文のチェックリスト: CONSORT 声明

表題と要旨

「ランダム割付け」や「ランダム化（無作為化）」という用語を含む。

緒論

どうしてこの臨床試験が必要かの根拠を述べる。

方法

対象の適格条件を述べる。

介入の内容を述べる。

研究仮説を述べる。

主要・副次エンドポイントを定義する。

症例数設定の根拠を述べる。

ランダム化コード票作成法, 具体的割り付け手順の隠蔽化を述べる。

盲検化について述べる。

統計解析法について述べる。

結果

患者プロフィールの流れ図を示す。

登録と追跡の期間について述べる。

群ごとにベースラインデータを示す。

解析対象数を明記する。

エンドポイントについて効果サイズと信頼区間を示す。

多重性を意識し, 層別解析や調整解析の結果を示す。

重要な副作用について述べる。

コメント

バイアス, 症例数不足, 多重性について述べる。

一般化可能性について述べる。

総合的に結果を解釈する。

EBM では RCT と同様に, メタアナリシスという研究も重視される。こちらもデザインがメタアナリシスだからといって, 一概に信用できるものではない。RCT の場合と同様にして, メタアナリシス論文用のチェックリストも作成された (Moher et al, 1999)。それは QUOROM 声明と呼ばれており, 表 4 に日本語訳を示した。これらのチェックリストを利用して, 要は Bias と Chance の点からそれぞれの研究の信憑性を評価することである。

表 4. メタアナリシス論文のチェックリスト: QUOROM 声明

表題

「メタアナリシス」又は「系統的レビュー」という用語を含む。

抄録

構造化抄録を用いる。

目的 臨床疑問を明示する。

データ 用いた研究のリストなど、情報源を示す。

方法 選択基準、妥当性評価法、データ要約法、研究特徴、統合のための統計手法を示す。

結果 選択した RCT, 除外した RCT を表示し、知見（点推定と信頼区間）を示す。また層別解析の結果も示す。

結論 主たる結果を述べる。

緒論

臨床疑問とその介入について調査する理由、総合評価する根拠を述べる。

方法

検索 情報源と検索条件を示す。

選択 集団、介入内容、評価項目、研究デザインに関して選択基準と除外基準を示す。

評価 収集された研究に関して、妥当性評価の基準と手法を示す。

要約 データ要約の手順について示す（2人独立に、など）。

研究 各研究のデザイン特徴を示す項目を示す。

統合 効果指標、統合手法、欠損の取り扱い、異質性の評価法、出版バイアスの点検法について示す。

結果

流れ図 メタアナリシスの実施に関する流れ図を示す。

研究 研究ごとにデザイン特徴を示す。

統合 データ統合の結果を示し、要約指標を示す。

考察

結果の妥当性について、バイアスの可能性について述べる。

2. 生物統計学との関連性

EBM において生物統計学 (Biostatistics) がどのように関連するか。それは臨床研究の結果について、正しく評価するさいに必要となる。得られた結果は P 値で示され、それは $P < 0.05$ であれば通常有意として発表される。この P 値をみると、それはたまたま良い結果になったわけではなく、Chance を超えて良い結果だとわかる。しかし、この P 値だけではどのくらい良かったかということまでは不明である。そこで、P 値と並んで信頼区間が使われるわけである (舟喜と折笠, 2001)。信頼区間を見れば、悪くてもこのくらい、良ければこのくらい、といった目安がわか

る。このように、EBM において研究結果が Chance (偶然) で説明できるのか、そうでないかを見分けるには生物統計学の素養が必要となる。

もう一方の素養は、いわゆる臨床疫学 (Clinical Epidemiology) と呼ばれるものである。折笠 (1995) によると、生物統計学の計画部門と位置付けることもできよう。これは、Chance に対して Bias (偏り) という用語で象徴される。Bias がないかを見分けるためには研究デザインの知識に加えて、その臨床領域の基礎的知識、そして現場でどのようにして実施されたかを知らないとは判断が難しい。

統計学的有意というのは $P < 0.05$ になった結果を言うが、たとえ統計学的有意になったとしても、臨床的には意味がないこともある。それを知るためには信頼区間が必要である。例えば、糖尿病治療のための新薬と標準薬があり、新薬では HbA1c を 0.5% 下げ、標準薬では 0.3% 下げたとしよう。両者の差は 0.2% (95% 信頼区間: 0.1~0.3%) であったとする。この結果は統計学的には有意である。帰無仮説の 0% を含んでいないからである。しかし、高く見積もっても新薬は 0.3% より下げるにすぎない。こうなると、臨床的に有意な結果かどうかは疑わしくなる。また、次のことも考える余地があろう。患者背景が HbA1c 9% など高い集団だとすると、0.3% ではあまり意味がないかもしれない。一方、HbA1c 7% 近辺では 0.3% でも臨床的に意味があるかもしれない。

最後に、EBM で用いる特別の統計量 (効果指標) について考えたい。その代表例が NNT (Number Needed to Treat) という指標である。Laupacis et al. (1988) がはじめて提唱した統計量であり、2 群のイベント発生率の絶対差の逆数で定義される。例えば、新薬のイベント発生率が 5%、標準薬のそれが 3% としよう。絶対差は 2% ($= 0.02$) である。その逆数をとると、 $1 \div 0.02 = 50$ 人となる。この 50 人というのが NNT になる。すなわち 50 人に 1 人の割合で、新薬によりイベントを回避できることを意味する。NNT が小さいほうがその効果は強いことになる。ここで、絶対差の信頼区間は二項分布あるいは正規近似などで容易に求められる。しかし、その逆数の信頼区間となると Delta method など、さらに近似を用いる必要が生じる。さらに面倒なことに、この NNT の信頼区間を図示すると、それは図 1 のように断絶された形となる。0 人をはさんでプラス側では救うほう (有効性) を示し、マイナス側はイベントを増やすほう (危険性) を示す。さらに、その大きさが連続的ではなく、0 人に近づくにつれて有効性・危険性ともに大きくなる。このため変則的な信頼区間になるのである。

信頼区間が途中で断絶されないようにするには、図2のように軸を変えないといけない。図2では危険から有効へと連続的な軸となっているが、NNTに関してみると変則的となってしまうている。信頼区間(-500人から+100人)は、通常なら-500から-300, -100, 0を経て+100まで移行すると考えるだろう。しかし、ここでは-500から-1000, $-\infty$ (+0人と同じ)を経て+100人へ移行するのである。解釈としては、悪く見積もると500人に1人の割合でイベントを増やす可能性、よく見積もると100人に1人の割合でイベントを減らすとなる。NNTというEBM特有の統計量に関する信頼区間の構成法については、Altman (1998) 及び Bender (2001) を参照されたい。また、イベント発生までの時間データでのNNTについては、Altman and Anderson (1999) が参照になろう。これらの例から、EBMに対する統計学からのチャレンジはさらに必要であることが伺われる。

疫学の分野では Person-year (人年) という単位がよく使われる。これについては統計モデル解析として、Poisson 回帰などの手法が開発されてきた。上のNNTなどの解析でも、そのBehaviourが変わったものなので、特別な解析法の開発が今後進むことを期待したい。

3. 結果の表示法をめぐる

EBMでは特に治療法の臨床成績を示すとき、様々の指標が提案されてきた。時には、その違いに惑わされることもある。1988年に様々の効果指標がReviewされているが(Laupacis et al., 1988), そのとき初めてNNTという指標が提案された。NNTというのは主として有効性のほうを示しているものだが、そうではなくNegativeな結果が報告されることもある。いわゆる有害作用を報告した論文である。このときには何人救えるかではなく、何人に危害を及ぼすかということになる。それにあわせて、NNTH (Number Needed To Harm) という指標が提案された。そうであれば、NNTのほうも本来はBenefitだからということで、NTTB (Number Needed To Benefit) と呼ぶこともある。この2つの用語はAltman (1998) の中で初めて見られた。有害作用のほうのNNTHについては、Bjerre and LeLorier (2000) や折笠 (2000a) に見られる事例を参照されたい。これらは臨床家に有用とされる、すべて絶対的な指標である。ただ治療法同士を比較するには、相対的な指標が便利なこともある。その中でNRR (Number Remaining at Risk) という指標も最近提案された (Massel and Cruickshank, 2002)。ここではEBMでよく登場する

これらの指標、つまり RRR, ARR, NTT, そして NRR について, GUSTO スタディでの例で比較してみる。

表 5. GUSTO スタディ結果を用いた各 EBM 指標の比較

Arm	Event rate			
tPA	6.3 %			
SK	7.3 %	RR 0.86	→	RRR 14 %
			→	ARR 1 %
			→	NTT 100 人
		$NRR = RR / (1 - RR) = 6.3$		

GUSTO スタディでは、血栓溶解剤である tPA と SK を比較している (表 5)。イベント発生率は tPA 群で 6.3%, SK 群で 7.3% であった。絶対値で見ると、1% tPA のほうが優れる結果であった。相対指標であるリスク比 (RR: Risk Ratio) は 0.86 であったので、相対リスク減少率 (RRR: Relative Risk Reduction) は 14% になる。また、絶対リスク減少率 (ARR: Absolute Risk Reduction) は 1% であるから、NTT は 100 人になる。すなわち tPA を使用することで、100 人に 1 人の割合でイベントを回避できるというわけである。臨床論文では RRR という相対指標で結果を示すことが多いが、現場主義の EBM では絶対指標のほうを重視する。それはなぜかという、相対では有効性を感覚的に把握できないためであろう。一方、14% リスクを減少するという相対指標では、リスク発生のベースライン値の大きさを超えて比較しやすい面がある。ここで新しく 2002 年に提案された NRR を考えてみよう。NRR は RR を $(1 - RR)$ で割った値であるので、これは相対指標になる。この例では 6.3 と算出される。これはどういった意味があるのか。tPA 治療によって救われる人と救われない人の比である。tPA による治療があっても、まだ 6.3 倍の人が救われず残っている。当然、この値が小さいほど治療効果が高いことになる。しかし、この指標は相対であるので、やはり現場で有用な指標にはならない。すなわち、目の前の患者へ応用する EBM にはあまり役立たない。しかし、治療法間の比較にはこうした相対指標は有用であろう。例えば、他の抗血栓療法で SK に対する NRR が 5 倍だとすると、この tPA よりも優れていることがわかる。このように、NRR は医療技術評価の尺度として有用かもしれない。このような様々の指標を開発することで、異なった解釈を可能にすることも考えてもらいたい。

ここには述べなかったが、その他にも Number Needed to Screen というような統計量も提案された (Rembold, 1998)。例えば、乳がんを 1 例検診で回避するには何名検診しないといけないかなので、検診の有効性を評価するための指標になる。しかし形式的に見れば、これは先に述べた NNTH と同様である。

4. EBM と診療ガイドライン

EBM というのは診療実践そのものであるが、近年 Evidence に基づいた診療ガイドライン策定も進んでいる。診療ガイドラインの目標とは、診療の最低水準を引き上げて、どこでも同じ医療が Jpn J Biomet Vol. 24, Special Issue, 2003

受けられるようになることである。さらに言えば、医療がより透明になることも期待される。そういったガイドラインは専門家のコンセンサスで作成することも可能であるが、そこに Evidence を引用しながら作成するようになってきた（折笠, 2000b）。ガイドラインを策定するにおいては、臨床疑問をまず明らかにする。こういった内容を書くかということであり、これは主として医師が担当しなければならない。続いて、その疑問に答えるようなデータを収集するわけだが、そこではリサーチ・ライブラリアンが貢献する。もちろん、研究デザイン知識については統計家がサポートし、臨床用語については医師が協力する。得られたデータを吟味するさいには、いわゆる Evidence 水準を明らかにするわけであるが、その際には統計家や医師が関与する。最後に、ガイドライン声明を書き上げる段階では主として医師が関与することになるだろう。このように、Evidence としての臨床データに基づくガイドライン策定において、統計家が協力する余地は大きい。実際、著者も糖尿病及び脳卒中全般のガイドライン策定において、情報収集及びその吟味において協力をしてきた。

5. EBM の問題点

EBM というのは臨床データに基づいて、その時点でベストと考えられる標準治療を万人に提供するのが目標である。しかし臨床データがさらに整ってくると、個人別治療も可能になってくるだろう。すでに高脂血症などでは、それ以外にリスクファクターの個数で治療法を変化させている。これは、Framingham スタディという Cohort 研究データなどを参考にして設定したそうである。科学を駆使してここまで進んだかという感がするわけである。このように見てくると、EBM というのは完璧で譲ることのできないものに思えるわけだが、それが万能ということはない。Evidence ではなく Experience をより重視する医療として、Narrative based medicine (NBM) もよく聞かれるようになってきた（Greenhalgh, 1999）。Evidence により画一化するのではなく、患者のエピソードや診療経験で変えるべきというものである。EBM も最終的に個別医療を目指すので、結果的には目標は同じと思われる。しかし、EBM 発祥の原典であるイギリスから、EBM とは相反する NBM が出てきたのは決して悪いことではないと思う。

さらに、EBM の総本山とも言える Oxford 大学臨床試験部の Peter Sleight 博士は、EBM に関して多くの問題点を指摘している（桑島, 1999）。例えば、EBM はあまりに単純すぎて、正しい根拠が不十分である。死亡率ばかりがエンドポイントに設定されている。RCT 万能なので、リスクの低い患者に偏りすぎている。もう一方の万能であるメタアナリシスには様々のバイアスが入りやすい。EBM の良い点を主として本稿で述べてきたが、そうではない点もあることを知っていてほしい。

6. ま と め

EBM において生物統計学はあまり関係ないという声もよく聞くが、本稿で生物統計学が寄与してきた、及び寄与すべき領域について述べた。Evidence 水準について整備すること、RCT やメタアナリシスであっても良質かどうかを見分ける基準を策定すること、EBM 特有の指標（統計量）について統計学的性質を検討すること、医療技術評価や診療ガイドライン策定にもっと貢献

すること, 以上の可能性を促したつもりである.

文 献

- Altman, D. G. (1998). Confidence intervals for the number needed to treat. *Br Med J* 317: 1309-1312.
- Altman, D. G. and Anderson, P. K. (1999). Calculating the number needed to treat for trials where the outcome is time to an event. *Br Med J* 319: 1492-1495.
- Bender, R. (2001). Calculating confidence intervals for the number needed to treat. *Contr Clin Trials* 22: 102-110.
- Bjerre, L. M. and LeLorier, J. (2000). Expressing the magnitude of adverse effects in case-control studies: "the number of patients needed to be treated for one additional patient to be harmed". *Br Med J* 320: 503-506.
- Cook, R. J. and Sackett, D. L. (1995). The number needed to treat: a clinically useful measure of treatment effect. *Br Med J* 310: 452-454.
- Evidence-Based Medicine Working Group (1992). Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA* 268: 2420-2425.
- Greenhalgh, T. (1999). Narrative based medicine in an evidence based world. *Br Med J* 318: 323-325.
- Laupacis, A., Sackett, D. L., and Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 318: 1728-1733.
- Massel, D. and Cruickshank, M. K. (2002). The number remaining at risk: an adjunct to the number needed to treat. *Can J Cardiol* 18: 254-258.
- Moher D., Cook D. J., Eastwood S., Olkin I., Rennie D., Stroup D., for the QUOROM Group (1999). Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. *Lancet* 354: 1896-1900.
- Moher D., Schulz K. F., Altman D. G., for the CONSORT Group (2001). The CONSORT statement. *JAMA* 285: 1987-1991.
- Rembold, C. M. (1998). Number needed to screen: development of a statistic for disease screening. *Br Med J* 317: 307-312.
- Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., and Haynes, R. B. (2000). Evidence-based medicine, 2nd edition. Edinburgh: Churchill Livingstone.
- 折笠秀樹 (1995). 臨床研究デザイン. 東京: 真興交易医書出版部.
- 折笠秀樹 (2000). EBM の実践に必要な臨床統計学のキーワード. 今月の治療 8(6): 636-643.
- 折笠秀樹 (2000a). 薬物治療に伴う有害事象に関する論文の読み方. 日本薬剤師会雑誌 52(8): 1121-1126.
- 折笠秀樹 (2000b). EBM に則った診療ガイドライン. *Mebio* 17(2): 42-46.
- 折笠秀樹 (2001). EBM 実践にさいして必要な統計学. IN: 「EBM 循環器疾患の治療 2001-2002 (三

- 田村・山科・川名・桑島, 編集)」, pp.1-20, 東京: 中外医学社.
- 桑島巖 (1999). EBM は医療の実践に不可欠か. Medical Tribune 8 月 19 日号: 18.
- 名郷直樹 (1999). EBM 実践ワークブック. 東京: 南江堂.
- 久繁哲徳・監訳 (1998). 根拠にもとづく医療. 東京: じほう.
- 舟喜光一, 折笠秀樹・訳 (2001). 信頼性の統計学—信頼区間および統計ガイドライン—. 東京: サイエンティスト社.