保健医療科学　2017　Vol.66 No.1 p.2－6

特集：データに基づく保健医療の計画と展開

＜総説＞

保健医療におけるデータの科学的活用

緒方裕光

国立保健医療科学院研究情報支援研究センター

Scientific utilization of data in healthcare

Hiromitsu Ogata

Center for Public Health Informatics, National Institute of Public Health

抄録

近年の情報通信技術（ICT）の発展にともない，あらゆる分野で大量のデータが容易に収集できるようになってきている．保健医療分野においても，疫学・統計的データに比べて，いくつかの利点を持つ大規模データが科学的根拠として利用可能になりつつある．また，これらの大規模データの中にはレセプト情報など医療費に関するデータも含まれており，これらが十分に活用できれば費用対効果の評価も可能になる．すなわち，ICTの進化により取扱い可能なデータ量が増大することにともない，いわゆる「エビデンスに基づく保健医療」のあり方も，現時点では部分的ではあるが少なくともデータ活用の観点からは大きな変革の時期にきていると言える．

保健医療分野におけるデータの科学的活用に関しては，従来の疫学や統計学を用いた方法の重要性は今後も変わらない．しかし，データの範囲と容量が格段に大きくなることにより，疫学または統計学的な限界を解決しうる有力なデータ活用が可能になる．一方で，これらのデータを科学的に活かすためには，データの質の確保や，分析方法の妥当性評価などが今後の重要な課題となるであろう．

キーワード：保健医療，情報利用，統計学，大規模データベース，費用対効果

Abstract

With the development of information and communication technology (ICT) over the past few years, large amounts of data can easily be collected in all fields. In the healthcare field, large-scale data or big data, which have various advantages over epidemiological/statistical data, are becoming available as scientific evidence. Additionlly, data on medical costs such as receipt information, for example, are included among these large-scale data allowing cost-effectiveness analysis. Therefore, with the increase in the amount of data that can be handled due to the evolution of ICT, the approach to data utilization in the field of health and medical care gradually changes. This means that “evidence-based health care” can be changed from the viewpoint of data utilization.

Regarding the utilization of scientific data in the field of health and medical care, the importance of epidemiological and statistical methods will remain unchanged. However, by significantly increasing the range and amount of data, it is possible to utilize data with very significant advantages from a statistical point of view. Meanwhile, for the utilization of these data, securing quality as scientific data and

連絡先：緒方裕光
〒351-0197　埼玉県和光市南2-3-6
2-3-6, Minami, Wako, Saitama, 351-0197 Japan.
Tel: 048-458-6203
E-mail: ogata.h.aa@niph.go.jp
［平成29年２月７日受理］

evaluating the validity of analysis methods will become important issues for the future.

keywords: healthcare, information utilization, statistics, large scale data base, cost-effectiveness
(accepted for publication, 7th February 2017)

I. はじめに

保健医療におけるデータ活用に関して，科学的方法論にしたがってデータを分析することを前提とすれば，主に以下の三つのアプローチがある．すなわち，１）収集したデータが事前に仮定したモデルや理論に適合するか否かを分析すること，２）不確実性をともなう複雑な現象を理解するためにデータを分析すること，３）何らかの意思決定に直接結び付くようにデータを分析すること，である．

これらのアプローチによって得られた結果は，いずれも保健医療における科学的根拠となりうる．ただし，保健医療分野では，物理や化学などの実験とは異なり，個人または人間集団に関する非常に複雑な現象を取り扱っており，その現象には様々な不確定要素が関係している．したがって，保健医療分野では，上記の三つのアプローチの中でとくに２）と３）の比重が比較的大きいと思われる．２）については，疫学や統計学が有力な方法論であることは言うまでもない．３）についてはいくつかの方法が考えられるが，最近の保健医療分野における代表的なアプローチとして費用対効果を評価するための方法論が挙げられる．

一方，この数年間の情報通信技術（ICT）の発展にともない，あらゆる分野で膨大な量のデータが容易に収集できるようになってきている．保健医療分野においても，従来の疫学・統計的データに比べて，様々な利点を持つ大規模データが科学的根拠として利用可能になりつつある．また，これらの大規模データの中にはレセプト情報など医療にかかる費用に関するデータも含まれており，これらが十分に活用できれば費用対効果の評価も可能になる．したがって，ICTの進化により取扱い可能なデータ量が増大することにともなって，上記で挙げた２）と３）のアプローチに少しずつ変化をもたらしているといえる．すなわち，いわゆる「エビデンスに基づく保健医療」のあり方も，現時点では部分的ではあるが，少なくともデータ活用の観点からは大きな変革の時期にあると思われる．

本稿では，保健医療分野における科学的データの活用に関して，従来の疫学や統計学の概念との関係を念頭に置きながら，データ収集，データ解析，データに基づく意思決定，という三つのプロセスについて，今後の課題と展望を概観した．

II. データ収集

１．標本と母集団

保健医療分野で収集されるデータは，原則としてある母集団から抽出された標本に関するデータである．データから導かれる結論は，そのデータ以外の情報がなければ，母集団のみに関する推論である．厳密にいえば，調査対象集団の各要素にある抽出確率と抽出方法を与えたものが母集団であり（すべての要素に等確率を与え独立に標本を抽出する場合が最も単純である），その確率に基づいて抽出された要素の集合が標本である [1]．したがって，標本データを分析する際には，常に標本と母集団の関係を考慮しなければならない．一般に標本サイズが大きくなるほど統計的検出力は上がるが，これらの関係は，統計的検定における有意水準および見つけようとする真の差の大きさに依存する．一例として，独立した２群の母平均の差の検定（t 検定，有意水準５％）において，真の差が0.01σ（２群の母分散は共通でσ^2とする）の場合における，一方の群の標本サイズ（２群の標本サイズを同数とする）と統計的検出力の関係を図１に示した．通常の疫学調査において，２群の母平均の差が0.01σという値は現実的には検出困難といえる微小な差である．この場合，図１でわかるとおり，標本サイズが数十万以上であれば検出力はほぼ100％になる．疫学調査で数十万以上の人を対象とした研究は，経費や時間の面からも非常に稀である（たとえば，日本の原爆被爆者の疫学調査は約12万人のコホート研究である [2]）．

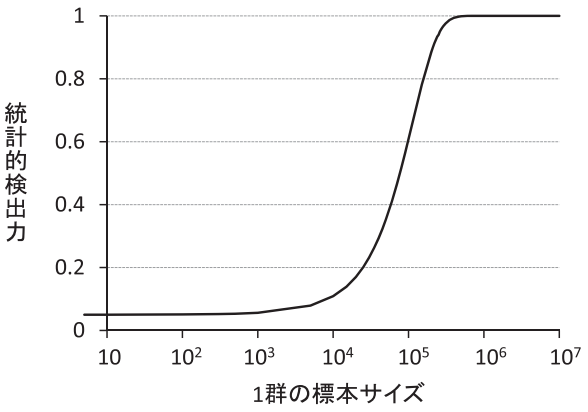

図１　標本サイズと統計的検出力との関係の例
２群の母平均の差の検定（有意水準５％）において真の差が0.01σの場合

一方，近年ではICTの進歩にともなう医療情報の電子化により，病院，地域，国での医療データの共有，関連する医療サービスに関するデータ蓄積，診断情報の二次利用などについて技術開発が進み，大規模データの利用の重要性が高まってきている．すなわち，医療ビッグデータと呼ばれる数万人から数十万人の規模の集団に関する研究が可能になりつつあり，もしこれらのデータの質が十分に確保されているという前提に立てば，検出力が非常に高い（場合によってはほとんど100％に近い）統計的解析が可能になる．

2．データの構造

通常の調査データは何らかの要素に起因するバラツキを持っており，それらの要因は説明できるものと説明できないものに分けられる．たとえば，データのバラツキが何らかのバイアス（結果や推測の真実からのずれ，またはその過程 [3, 4]）に基づくものである場合，そのバイアスの原因が特定できればそれは説明できる要因になり，特定できなければ偶然誤差によるバラツキとみなすか，またはバイアスの影響が含まれた状態でデータを分析するしかない．通常，調査の最終目的はバイアスの原因を探ることではなく，データによって説明される現象のメカニズムを知ることにある．したがって，調査データを分析する際に，純粋にデータの変動に影響を与えている原因を探るためには，バイアスの影響をできる限り取り除くことが重要である．

データ収集のデザインまたはデータ収集の段階から考慮しなければならないバイアスには，選択バイアス，情報バイアス，交絡がある．選択バイアスと情報バイアスに関しては，データ解析の段階では調整できない．交絡については，交絡因子に関するデータが取得されていれば解析の時点で統計学的に調整することもできる．これらのバイアスうち，選択バイアスは標本抽出の際に生じるものであり，たとえば，研究参加者（標本）の母集団が，当初想定していた母集団とは一致しないような場合がこれに該当する．大規模データと呼ばれるデータの標本サイズは非常に大きく，全数調査データに近いため（標本と母集団のサイズがほぼ同じ大きさ），選択バイアスの影響は非常に小さくなる．さらに，全数調査に近ければ標本データに基づく統計的推論（母集団のパラメータに関する推定や検定）はほとんど意味をなさなくなり，記述統計やパラメータの特定（推定ではない）が主な解析となる．

一方で，一般にビッグデータ（厳密な定義は確立しておらず様々な表現がある [5, 6]）と呼ばれるデータ集合の大部分はもともと研究目的で集められたものではないために，データの正確性やデータ間の関係性などについては深く吟味されていないことが多い．単にデータ容量が膨大というだけでは未整備のデータの集まりであり，科学的根拠として用いるためにはデータの質の確保という課題が解決される必要がある．

III. データ分析

1．データの多様性を活用した分析

質が確保された大容量のデータの集まりは大規模データベースとも呼ばれている．大規模データベースの特徴の一つとしてデータの多様性が挙げられる．すなわち，この大規模データベースが実現すれば，多くの変数に関してそれぞれ大量のデータがそろうことになる．このことは，データ分析の段階で以下に述べるように様々な利点を持つことになる．

まず，データの規模が大きいため，様々な変数に関してグループ別解析や多変量解析などが可能になる．バイアスの一つである交絡はデータ収集の段階でもデータ解析の段階でも調整が可能であるが，データ解析の段階における調整方法としては層化や多変量解析などが代表的である．標本サイズが小さい場合には，層化した時に十分な標本サイズが得られないこと，多変量解析の際に同時に多くの変数を取り扱えないこと，などの欠点があるが，標本サイズが大きければ上記の方法で交絡の影響を非常に小さくすることができる．

また，保健医療データが持つ重要な変数の一つに「時間」があり，複数のデータベースを統合することにより様々な変数に関してデータを時系列的に検討することが可能になる．別々の目的で構築された複数のデータベースでそれぞれ時系列的に蓄積されたデータを再構成し，分析できるようにするシステムは一般にデータウェアハウスと呼ばれている [7]．このようなシステムは，とくに地域の保健医療行政おいてその地域の状況に応じて独自に地域診断や将来予測などの分析を進める際に非常に有用であると思われる．

さらに，通常の疫学研究(観察研究)では，研究によって条件や観察指標が様々に異なるために，システマティック・レビュー [8, 9] を行って研究論文を抽出しても質の異なる研究が多く存在する．したがって，メタ・アナリシス [9, 10] を実施できるほどの十分な数の同質の疫学研究結果を得ることができない場合が多い．メタ・アナリシスの特徴の一つは，複数のエビデンスをその信頼性の大きさに応じて重みづけしたうえで統合することにあり，この方法は科学的根拠に基づく保健医療を実施する上で非常に重要な考え方を含んでいる．すでに臨床試験に関してはデータベース化がなされているが，疫学研究（観察研究）も含めてあらゆる調査研究データについて，メタ・アナリシスが可能な状態でデータベース化されるようになれば疫学データの有効利用につながるであろう．

2．その他の課題

現在ではインターネットが普及しており，様々な情報がインターネット上にあふれている．これらの情報の質は多種多様であり，科学的に正確かつ高度な情報もあれば噂や風評などの情報も混在している．したがって，こ

れらのインターネット上の膨大な容量のデータをどのように有効活用するかは今後の興味深い課題である．一部の分野ではすでに多くの研究が行われており，その解析方法についても様々な方法が発案されている．しかしながら，インターネット上の情報は標本と母集団との関係が明確ではないため，従来の疫学的方法，統計学的方法が使えないという欠点があり，少なくとも定量的な解析方法が確立されているとはいえない．

その他，保健医療分野における大規模データベースやデータウェアハウスに関する課題として，非公開情報を分析する際の個人情報保護やデータの目的外利用のためのルール整備の必要性など，情報処理技術や分析方法とは次元の異なる課題も残されている．

IV. 意思決定

保健医療におけるデータ活用の最終段階は，データ分析の結果を何らかの意思決定に反映させることである．現実の意思決定は様々な要素のバランスの上に行われており，必ずしも科学的根拠のみが重視されているわけではない．たとえば，意思決定の根拠となる情報には一般に科学的情報と経験的情報があり，科学的根拠の情報量が少なければ，後者の比重が大きくなり，主観的要素が意思決定に大きな影響を与えることになる．以下では，科学的情報の活用に限定して意思決定のための分析方法について述べる．

保健医療分野の意思決定に結びつく代表的かつ現実的な科学的アプローチの一つとして，費用便益分析がある．たとえば，ある健康リスクを減少させるためにそのリスク要因への曝露を制限する必要があるとすれば，その制限を確立させるためには何らかの費用を必要とする．一般にリスクをゼロに近付けようとすればそれに要する費用は高くなる．リスクを減少させることは言い換えれば便益（ベネフィット）を増大させることであり，費用と便益は相反する要素となる．意思決定のための費用便益分析の具体的な例として，図２に放射線防護における集団線量とリスク（発がんリスク）および防護にかかる費用との関係を示す [11]．なお，図２では，文献 [11] に記載された図に信頼区間の幅を模式図的に追加してある．このとき，「リスクにともなう損害を金額に換算した費用」と「防護に要する費用」との合計が最も小さいときの集団線量がリスク受容レベルの最適解ということになる．しかし，この分析を意思決定につなげるためには以下の点が課題となる．すなわち，まずリスクによる損害を合理的な方法で費用に換算できること，次に推定値の信頼区間または不確実性を考慮すること，さらに費用に換算できない要因も考慮すること，などである．データ活用の観点から言えばデータの正確性が重要であり，バイアスが強く影響していれば不確実性は大きくなり最適解の存在する範囲も大きくなる（精度の低い解しか得られない）．なお，実際の放射線防護では，費用に換算されない社会的要因なども考慮されている [11]．

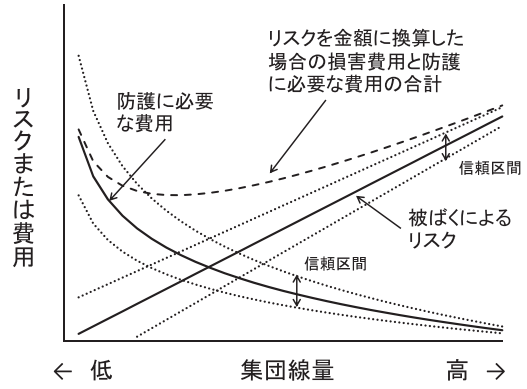

図２　放射線被ばくに関わるリスクと防護に必要な費用との関係
（文献 [11] に記載された図を編集した．）

さらに意思決定に結び付くもう一つの現実的な方法として，投入した費用に対してどの程度の効果が得られるかを分析する費用効果分析がある．前述の費用便益分析はいわゆる最適解を求めるための方法であるのに対して，費用効果分析は，複数の選択肢の中から費用対効果の低い選択肢を排除するため，あるいは複数の選択肢を比較するために用いられる．実際に，保健医療分野では，薬剤，医療機器，医療技術などについては，費用対効果の評価方法論に関してすでに多くの議論や研究がなされている．とくにレセプト，臨床試験，診療経過など，費用に関するデータソースが利用可能になるにともない，効果指標の確立も含めて標準的な方法論が整いつつある [12-14]．当然ながら保健医療におけるあらゆる行為（診療，保健事業など）には費用が発生しており，この費用対効果の評価は保健医療分野における意思決定の際の重要な科学的根拠となる．一方，地域や国で行われる保健事業全般に関しては，保健事業の効果の測定が容易ではなく，費用対効果の評価のための方法論はまだ十分には確立されていない．しかしながら，限られた予算の中で保健事業を展開する以上は，データソースの確保に関する課題も含めてより合理的な観点から費用対効果を評価する方法を確立することが望まれる．

V. おわりに

保健医療において様々なデータが科学的根拠として十分に活用されるためには，データの容量に関わらずデータ収集および分析の方法が科学的でなければならない．その前提として，収集されるデータの質がある程度確保されていることが重要である．一般的には分析対象のデータの範囲と量が大きくなるほど，個々のデータの質について合理的な評価を行うことが難しくなる．したがって，本稿で中心的課題として述べた大規模データを十分に活用するためには，単に大容量のデータを機械的に取り扱うだけでなく，データの妥当性や信頼性の評価

を行うことが大きな課題となる．その他に，情報処理に関する技術的な側面だけでなく個人情報保護やデータの二次利用のルール化などについても今後検討していく必要がある．

参考文献

[1] 林知己夫．データの科学．東京：朝倉書店；2001.
[2] 西信雄，児玉和紀．広島・長崎放射線影響研究所コホート研究．医学のあゆみ．2008;224(2):157-161.
[3] Armitage P, Colton T. Encyclopedia of Biostatistics, 2 nd edition. Chichester: Wiley; 2005.
[4] Last JM. A Dictionary of Epidemiology, 2nd edition. New York: Oxford University Press; 1988.
[5] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. June 2011. http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (accessed 2017-02-04)
[6] 総務省．平成24年版情報通信白書． http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/html/nc121410.html (accessed 2017-02-04)
[7] Inmon WH. Building The Data Warehouse, 4th edition. Chichester: Wiley; 2006.
[8] Gray JAM. Evidence-Based Healthcare, 2nd edition. London: Churchill Livingstone; 2001.
[9] Chalmers I, Altman DG. Systematic Reviews. London: BMJ Publishing Group; 1995.
[10] Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for Meta-Analysis in Medical Research. Chichester: Wiley; 2000.
[11] ICRP. The optimisation of radiological protection: broadening the process. ICRP Publication 101b. Annals of the ICRP. 2006;36(3):65,71-104.
[12] 福田敬．医療経済評価手法の概要．保健医療科学．2013;62(6):584-589.
[13] 白岩健．「医療経済評価研究における分析手法に関するガイドライン」の解説．保健医療科学．2013;62(6):590-598.
[14] 福田敬，白岩健，池田俊也，五十嵐中，赤沢学，石田博，他．医療経済評価研究における分析手法に関するガイドライン．保健医療科学．2013;62(6):625-640.