根拠に基づく診療ガイドラインの現状

中山健夫 NAKAYAMA Takeo
京都大学大学院医学研究科社会健康医学系専攻健康情報学分野

キーワード 診療ガイドライン，根拠に基づく医療，推奨度

[要約] 根拠に基づく医療（EBM）の手法を用いた診療ガイドラインに対する関心が，医療者のみならず一般の人々の間でも高まっている．診療ガイドラインの定義は「特定の臨床状況のもとで，臨床家や患者が，適切な判断や決断を下せるように支援する目的で体系的に作成された文書」である．本稿では近年，医学領域を中心に進められてきた診療ガイドラインの動向と課題を概観し，看護領域のおける適切なガイドラインの方向性，課題を示す．

はじめに

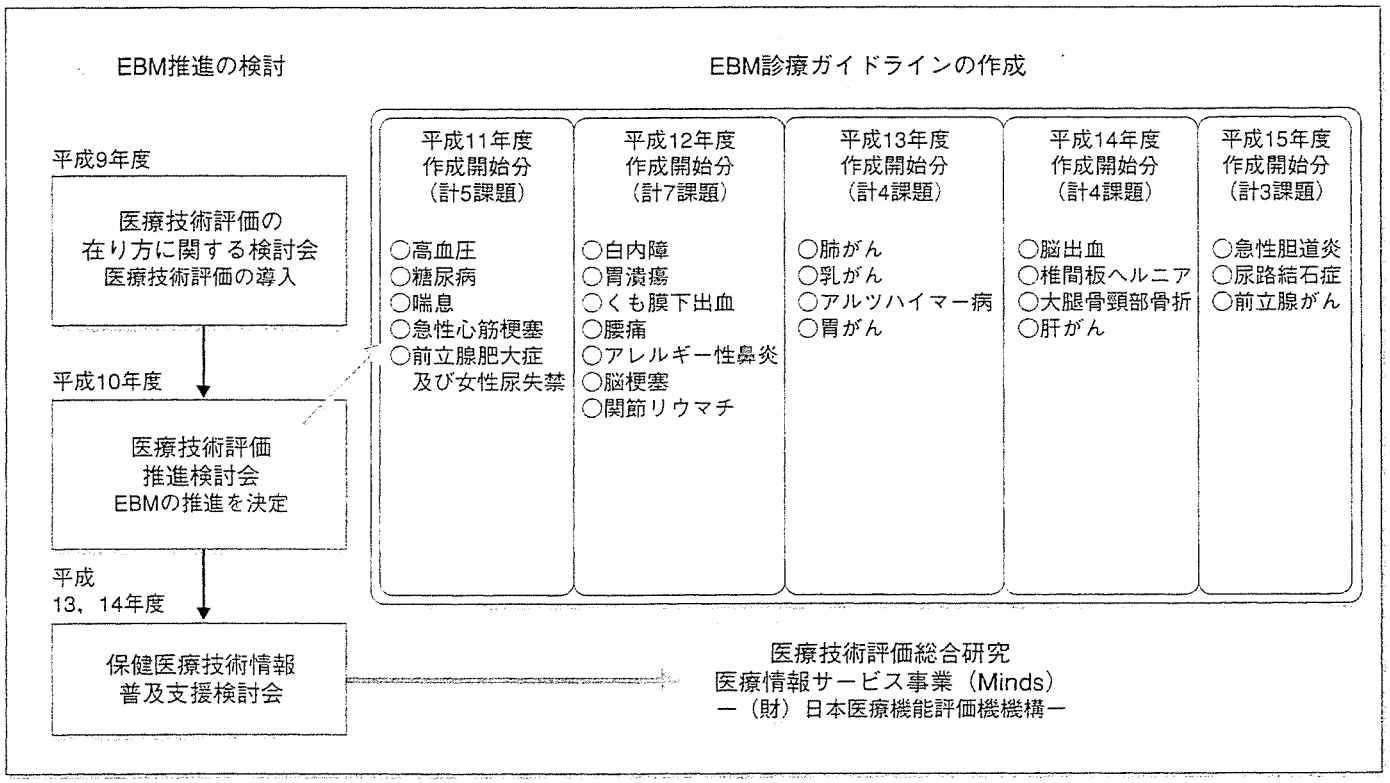
診療ガイドラインは米国のInstitute of Medicineにより「特定の臨床状況のもとで，臨床家や患者が，適切な判断や決断を下せるように支援する目的で体系的に作成された文書」と定義されています[1]．また英国・スコットランドで診療ガイドラインの作成・普及の責任を担っているSIGN（The Scottish Intercollegiate Guidelines Network）は，「私たちの目的は，現在のエビデンスに基づく治療の有効性に関する推奨を含む国レベルの診療ガイドラインの作成，普及を通じて，診療とアウトカムのばらつきを減じ，スコットランドにおける患者ケアの質を向上させることである」と，診療ガイドラインに求められる特性とその役割を簡潔に表現しています．

日本では厚生省（当時）が1999年度の厚生科学研究から，EBMの手法を用いた診療ガイドライン作成を開始しました（●1）[2]．それまでの診療ガイドライン作成においては，根拠とする文献の選択・入手法，評価法，推奨の決定法などが明示されず，主導的立場にある臨床医のコンセンサスによって作られることが一般的でした．厚生（労働）科学研究によるプロジェクトでは臨床疫学者や生物統計学者のような研究方法論の専門家や，適切な文献を検索するために医学図書館員の参加が強く求められました[2]．現在は公的研究費の枠外でも多くの学会が独自に診療ガイドラインの作成を進めています．

本連載では，近年，医学領域を中心に進められてきた診療ガイドラインの動向と課題を概観し，看護領域における適切なガイドラインの方向性，課題を考えていきたいと思います．

診療ガイドラインの構造と作成手順

EBMの手法に基づくガイドラインでは，できる限り客観的なエビデンスに基づいて推奨を行うことが求められており，一定の明示的な作成プロセスを踏むことが重視されています．診療ガイドラインの作成では，まず推奨を示すべき臨床的課題（clinical questions）を挙げ，それに対してシステマティック・レビューの方法に準じて，関連文献の系統的検索と吟味を行います[3]．その結果に基づき推奨度を決



❶ EBM に対する厚生労働省の取り組み

（厚生省健康政策局研究開発振興課医療技術情報推進室監：平成10年度医療技術推進検討会報告書．わかりやすいEBM講座．厚生科学研究所；2000．p.8-28．[2] より）

定するわけですが，推奨度は診療ガイドラインの核心と言えます．あるテーマの文献が収集され，その内容が抽出された「診療エビデンス（clinical evidence）集」は，方向性を示す推奨を含まない点で診療ガイドラインと異なるものです．推奨度は得られたエビデンスのレベルに影響されますが，それだけではなく，さまざまな要素を勘案して総合的に判断されます（❷）[4]．

推奨度の表示形式の例を❸に示します．日本では1990年代前半に米国AHCPR（Agency for Health Care Policy and Research，現 Agency for Healthcare Research and Quality：AHRQ）が提案した方式が多くのガイドラインで用いられましたが，判定が困難な場合も多いことが明らかになりました．特に「推奨度C：行うよう勧めるだけの根拠が明確でない」の解釈についてさまざまな混乱が生じました．

日本脳卒中学会をはじめとする5学会合同の脳卒中治療ガイドラインではC判定を「C1：行うことを考慮してもよいが，十分な科学的根拠はない」と「C2：科学的根拠がないので，勧められない」に区別した新しい提案を行いました．棚橋は，このガイドラインにおける急性期治療の推奨71のうち，Aが0件，Bが6件，C1が53件，C2が8件，Dが4件であり，C1が75％を占めていることを報告しています[5]．これは，必ずしも「日本の臨床医学は，高いレベルのエビデンスに基づいて行われていない」という否定的な意味ではなく，「実際に臨床で行われていることの多くは高いレベルのエビデンスで裏付けられていないものである」という臨床の現実を表しているとも言えるでしょう．カナダの予防医学に関するタスクフォースは，エビデンスが明確でない場合の意思決定・行動の基準を示しています（❹）．

看護を主題とするガイドラインでは，医学的課題

1. エビデンスのレベル
2. エビデンスの数と結論のバラツキ
（同じ結論のエビデンスが多ければ多いほど，そして結論のバラツキが小さければ小さいほど推奨は強いものとなる．必要に応じてメタアナリシスを行う）
3. 臨床的有効性の大きさ
4. 臨床上の適用性
5. 害やコストに関するエビデンス

❷ 推奨度の決め方（以上の要素を勘案して総合的に判断）

米国　Agency for Health Care Policy and Research（1994年）
A：行うよう強く勧められる
B：行うよう勧められる
C：行うよう勧めるだけの根拠が明確でない
D：行わないよう勧められる

脳卒中合同ガイドライン委員会（2004年）
A：行うよう強く勧められる
B：行うよう勧められる
C1：行うことを考慮してもよいが，十分な科学的根拠はない
C2：科学的根拠がないので，勧められない
D：行わないよう勧められる

❸ 推奨度の形式

- 意思決定における患者自身の参加を促すこと
- 害を最小化すること
- 強い必要性が明らかな場合に関してのみ、大きな変化を主張すること
- 不要な「ラベリング」を避けること
- 益の不確かな高価な手技を避けること
- 疾病負担が大きい状況に焦点を当てること
- ハイリスクグループの特別のニーズに配慮すること

❹ エビデンスが明確でない場合の意思決定の基準（カナダ・タスクフォース）

よりも，エビデンスレベルが高いとされる研究は少ないことが推測されます．看護領域であっても，介入の有効性を，ランダム化比較試験（randomized controlled trial：RCT）で評価することが望ましいのですが，RCTだけにエビデンスとしての拠り所を求めることは現実的ではないですし，望ましいことでもありません．医学領域でEBMや診療ガイドライン導入の初期に見られたような，「ランダム化比較試験至上主義」に陥らないよう，看護領域でのガイドラインの議論では十分な留意が望まれます．

診療ガイドラインの拘束力

診療ガイドラインは日常診療の対象となる多くの症例に対して参考になりますが，個々の患者の状況，特性に応じて柔軟に利用することが不可欠です．欧米ではガイドラインを“starting point for discussion”としており，関係者のコミュニケーションを進め，意思決定を支援する役割を担うものです．しかし今後，医療者の想定を越える領域でも，医療訴訟の判断基準としてガイドラインが社会的に認識されていく可能性は否定できず，それを確固とした判断の指針と捉える医療消費者も増えていくことが予想されることも確かです．診療ガイドラインの意義と適切な社会的位置付けは医療者だけではなく，法律家や一般の人々とも協力して取り組むべき重要課題と言えます．

ガイドラインの拘束力に関して欧米では，「指令（directive）は推奨（recommendation）よりも強く，推奨（recommendation）は指針（guideline）よりも強い，北米では指針と推奨は同等」と位置付けられています[6]．ガイドラインの役割はあくまで意思決定の支援であり，現場における個別の臨床行為・意思決定を拘束するものではありません．指令よりも強い言葉は「規制（regulation）」であり，この場合はそれに従わないと罰則があるなど，直接的な不利益が生じます．Recommendationは従来「勧告」と訳されていましたが，「勧告」は医療法にも見られる法律用語であり，暗黙の拘束力が意図されています．一般の人々や法律家を交えたガイドラインに関する社会的議論に向けて，誤解を避けるために「推奨」を用いることが適切です．

診療ガイドラインがカバーするケースについて，

Eddyは60～95%の患者にとどまると述べ，95%以上の患者に適応されるものは「スタンダード」，反対に50%ほどの患者にしか適応されないものは「オプション」と区別しています[7]．同じ「診療ガイドライン」でも，慢性疾患で予後への影響要因が多様なものと，進行がんに代表されるように特定の治療が予後に強く影響するものとでは，科学的な意味での拘束力は異なるでしょう．診療ガイドラインに関する議論は総論と共に，特定の臨床状況における各論的な検討を並行して深めていく必要があります．

本来のガイドラインは指示的なものではないと述べましたが，医師向けのガイドラインと異なり，看護師向けのガイドラインではある程度の指示性が望まれるのではないか，という意見もあります．看護ガイドラインはどのような拘束力を持つものか，持つべきものか，または持たないものなのか，今後，関係者の十分な議論が望まれます．

診療ガイドラインの評価

診療ガイドラインの評価に関する国際的な手法としてAGREE共同計画（"Appraisal of Guidelines Research and Evaluation"）が開発した6領域23項目と総合評価から成るツールがあります．内容は下記の通りです[8,9]．

範囲と目的（項目1-3）：ガイドラインの全体的な目的，具体的な臨床問題と対象とする患者集団に関する項目．

利害関係者の関与（項目4-7）：ガイドラインがそのユーザーとして意図された者の見方をどれほど代表するものであるかに関する項目．

開発の厳密さ（項目8-14）：根拠を探し集約するのに用いられた過程と，推奨を導き出す方法，その更新に関する項目．

明快さと提示（項目15-18）：ガイドラインの言葉と形式に関する項目．

適用性（項目19-21）：考えられるガイドラインの適用，組織的・行動的・経済的影響に関する項目．

編集の独立性（項目22-23）：ガイドライン作成グループが利害の衝突を生じる影響力から独立していることに関する項目．

AGREEの方法は臨床医の行動や患者アウトカムの変化ではなく，診療ガイドライン作成の「枠組み」の評価に焦点を当てています．一方，米国から発表されたThe Conference on Guideline Standardization（COGS）は診療ガイドラインの作成者がそれを記述する際の項目・構成の提案であり[10]，発表されたガイドラインの評価ポイントを示すAGREEの方法と表裏一体です[11]．このような国際的な標準様式に沿った形で診療ガイドラインを作成し，報告することが，今後，質の高い診療ガイドラインを世に出すための必須要件となっていくでしょう．

患者の視点の反映

患者・家族と医療者の対話の結節点として診療ガイドラインが機能する可能性を探ることは大きな意義を持ちます[12]．前述のAGREEによる評価項目の一つとして，「患者の視点や選好は考慮されたかどうか」が問われており，「診療ガイドライン開発にあたって，患者の経験と期待に関する情報を知っておかねばならない．そのための方法として，開発グループに患者の代表を含める，患者のインタビューから情報を得る，開発グループが患者の経験に関する文献をレビューする，などがある．この手順が行われたという記述がなければならない」とされています．

患者の代表が診療ガイドライン作成に関与することの意義は当初から主張されていましたが[4]，実質的な取り組みはこれからです．これまで喘息（小児・成人）[13,14]や乳がんに関して，患者向けのガイドライン，またはガイドラインに基づく解説の作成に患者や患者支援者グループが参加した例があります．こ

れらでは，「患者の代表」というよりも，多くの患者の声を集約できる立場の人々が大きな役割を担っていました．医療者向けとされる診療ガイドライン自体の作成に，患者・支援者が直接に関わったケースはまだ見られません．今の状況で何人かの患者がガイドライン作成の会議に参加したとしても，実質的な議論への十分な貢献は期待できないでしょう．それは患者の責任ではなく，「必要な知識を事前に習得する機会が提供されていないこと」と「個人への依存度が高く，システム化されていないこと」が問題と思われます．英国・イングランドでは，診療ガイドラインの作成・普及に責任を持つNational Institute for Health and Clinical Excellence（NICE）の内部に“Patient Involvement Unit（現 Patient and Public Involvement Programme：PPIP）”という専門部局があり，診療ガイドライン作成に参加する患者をコーディネートし，事前にトレーニングセッションを提供しています[15]．これらの海外の先進的な事例を学び，日本における今後の方向性を探る手がかりを得ることは重要です．

稲葉は医療者と患者が共にガイドラインの限界と役割を理解し，医療者は責任と倫理を踏まえて患者の陥りやすい問題を把握し，診療ガイドラインを用いてインフォームド・コンセントを行い，対話の中で治療方針を共に決めていく調和的な医療モデルを提案しています[16]．患者の視点も取り入れて作られた診療ガイドラインは，それを活用すればインフォームド・コンセントを巡るトラブルを回避できるだけでなく，インフォームド・コンセント自体を充実させ，その結果として医療の質・安全性，そして患者満足度を高めるための中核的な手段となり得ます．EBMの手法による診療ガイドラインが，臨床現場にとどまらず，社会的にも適切な認知と支援を受けるためには，EBMを基盤として，そこに臨床倫理や法律的な視点からの検討を加えていくことが望まれます．

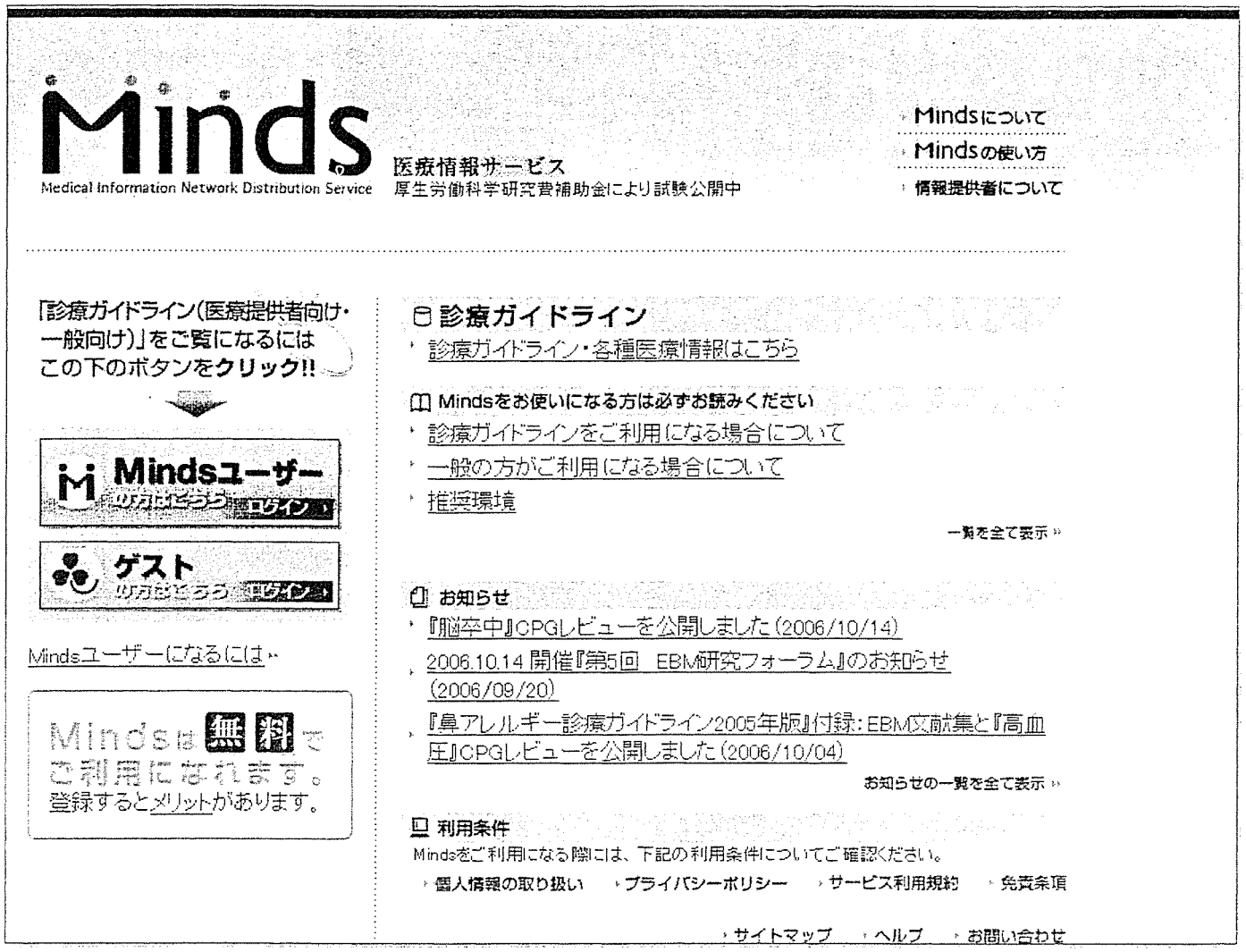
看護ガイドラインでは患者の視点が反映される可能性は医学的ガイドラインよりも大きいと考えられます．今後の看護関係者の取り組みを期待するものです．

これからに向けて

診療ガイドラインは医療の質を向上させるための特効薬ではありません．使い方を誤れば瞬時に諸刃の剣に転じるでしょう．その意義，役割，そして限界を医療者と一般の人々がどのように共有していくべきか，開かれたディスカッションを積み重ねていく必要があります．

ガイドラインへのアクセスを容易にするため，米国では国立ガイドライン・クリアリングハウスが，英国では健康情報電子ライブラリーがインターネットを通じて診療ガイドラインを広く一般の人々へも公開しています．日本でも現在，財団法人日本医療機能評価機構の「医療情報サービス事業」，通称“Minds”（Medical Information Network Distribution Service）の整備が進められています（❺）[17]．Mindsでは，診療ガイドラインの中でも看護師が大きな役割を担った「周産期ドメスティック・バイオレンスの支援ガイドライン」[18]がすでに公開されており，続いて「褥瘡局所治療ガイドライン」[19]の掲載準備が進められています．今後，看護領域ガイドラインの整備が進むと共に，それらがMindsで公開され，社会に開かれたガイドラインとして普及，発展していくことが願われます．

これからの連載では，本教室の看護系大学院生と協力して，ガイドラインにおけるエビデンスとコンセンサス，ガイドライン策定への看護の視点の導入，日本発の看護系ガイドラインの可能性と課題，世界の看護系ガイドライン，クリニカルパスとの関係，医療消費者の参加などの問題を扱いたいと考えています．



❺ Minds のホームページ（財団法人日本医療機能評価機構　医療情報サービス事業）
(http://minds.jcqhc.or.jp/to/index.aspx)

日本における看護ガイドラインが良い形で展開されていくことを願いつつ，連載の1回目とさせて頂きました．

文献
1) Institute of Medicine : Guidelines for clinical practice: from development to use. Washington DC, National Academy Press, 1992.
2) 厚生省健康政策局研究開発振興課医療技術情報推進室監：平成10年度医療技術評価推進検討会報告書．わかりやすいEBM講座．厚生科学研究所；2000. p.8-28.
3) 中山健夫：EBMを用いた診療ガイドライン：作成・活用ガイド．金原出版；2004. p.78-80.
4) 福井次矢，丹後俊郎：診療ガイドラインの作成の手順．EBMジャーナル2003；4：28-37.
5) 棚橋紀夫：EBMに基づく治療ガイドラインの作成と応用の問題点．第44回日本神経学会総会プログラム・抄録集 03；61.
6) Naka [illegible] T, Budgell B, Tsutani K : Confusion about the concept of clinical practice guid[illegible] in Japan: On the way to a social consensus. Int J Qual Health Care 2003；15：359-360.

7) Eddy DM : Clinical decision making: from theory to practice. Designing a practice policy. Standards, guidelines, and options. JAMA 1990 ; 263: 3077, 3081, 3084.
8) Appraisal of Guidelines, Research, and Evaluation in Europe (AGREE) Collaborative Group : Guideline development in Europe: An international comparison. Int J Technol Assess Health Care 2000 ; 16: 1039-1049.
9) 長谷川友紀：診療ガイドラインを取り巻く状況：AGREE Collaborationの動向. EBMジャーナル 2003 ; 4 : 294-297.
10) Shiffman RN, Shekelle P, Overhage JM, et al. : Standardized reporting of clinical practice guidelines: a proposal from the Conference on Guideline Standardization. Ann Intern Med 2003 ; 139 : 493-498.
11) 中山健夫：EBMを用いたガイドラインの作成と応用：EBMの考え方と問題点を踏まえて. Progress in Medicine 2003 ; 23 : 3143-3151.
12) Nomura H, Nakayama T : The Japanese healthcare system: The issue is to solve the "tragedy of the commons" without making another. BMJ 2005 ; 24 : 648-649.
13) 宮本昭正監：EBMに基づいた患者と医療スタッフのパートナーシップのための喘息診療ガイドライン2004(小児・成人編). 協和企画 ; 2004.
14) 「ぜんそく診療：患者参加で指針作り，分かりやすく図解」 朝日新聞 平成16年6月13日
15) 鈴木博道：NICEガイドライン開発への患者・介護者の参画. あいみっく 2004 ; 25 : 10-14.
16) 稲葉一人：インフォームド・コンセントを充実させるためのガイドライン.厚生労働科学 EBMを指向した「診療ガイドライン」と医学データベースに利用される「構造化抄録」作成の方法論の開発とそれらの受容性に関する研究（主任研究者・中山健夫）. 2002年度報告書, p.50-58.
17) 特集Minds入門：インターネットによる診療ガイドライン活用方法. 医事新報 2004 ; 4184 : 1-15.
18) 聖路加看護大学女性を中心にしたケア研究班編：EBMの手法による周産期ドメスティック・バイオレンスの支援ガイドライン2004年版. 金原出版 ; 2004.
19) 日本褥瘡学会編：科学的根拠に基づく褥瘡局所治療ガイドライン. 照林社 ; 2005.

INFORMATION

学際領域における評価のデザイン RCTとシステマティック・レビューの現状 第2回シンポジウム

日時：2007年1月12日(金)13：00～15：15
会場：東京大学大学院薬学系研究科・総合研究棟2F講堂（文京区本郷7-3-1）
プログラム：「介護予防のシステマティック・レビュー」(二木立)／「看護のRCTとSR」(片岡弥恵子)
座長：津谷喜一郎
コメンテータ：平岡公一，真田弘美
参加費：無料
参加申込み：事務局にメールで連絡
事務局：〒113-0033 東京都文京区本郷7-3-1 東京大学大学院薬学系研究科医薬政策学講座（高崎）
TEL：03-5841-4828
E-mail：takasaki@mol.f.u-tokyo.ac.jp
URL：http://www.f.u-tokyo.ac.jp/~utdpm/event/rct.html

Comparison of Effects in Randomized Controlled Trials With Observational Studies in Digestive Surgery

Satoru Shikata, MD,† Takeo Nakayama, MD, PhD,‡ Yoshinori Noguchi, MD, MPH,§ Yoshinori Taji, MD,† and Hisakazu Yamagishi, MD, PhD**

Objectives: To compare the results of randomized controlled trials versus observational studies in meta-analyses of digestive surgical topics.

Summary Background Data: While randomized controlled trials have been recognized as providing the highest standard of evidence, claims have been made that observational studies may overestimate treatment benefits. This debate has recently been renewed, particularly with regard to pharmacotherapies.

Methods: The PubMed (1966 to April 2004), EMBASE (1986 to April 2004) and Cochrane databases (Issue 2, 2004) were searched to identify meta-analyses of randomized controlled trials in digestive surgery. Fifty-two outcomes of 18 topics were identified from 276 original articles (96 randomized trials, 180 observational studies) and included in meta-analyses. All available binary data and study characteristics were extracted and combined separately for randomized and observational studies. In each selected digestive surgical topic, summary odds ratios or relative risks from randomized controlled trials were compared with observational studies using an equivalent calculation method.

Results: Significant between-study heterogeneity was seen more often among observational studies (5 of 12 topics) than among randomized trials (1 of 9 topics). In 4 of the 16 primary outcomes compared (10 of 52 total outcomes), summary estimates of treatment effects showed significant discrepancies between the two designs.

Conclusions: One fourth of observational studies gave different results than randomized trials, and between-study heterogeneity was more common in observational studies in the field of digestive surgery.

(*Ann Surg* 2006;244: 668–676)

From the *Department of Digestive Surgery, Kyoto Prefectural University of Medicine, Kyoto, Japan; †Department of Clinical Epidemiology, Kyoto University Graduate School of Medicine, Kyoto, Japan; ‡Department of Health Informatics, Kyoto University School of Public Health, Kyoto, Japan; and §Department of Medicine, Fujita Health University School of Medicine, Aichi, Japan.

Supported by a Health and Labour Sciences Research Grant (Health Technology Assessment) from the Ministry of Health, Labour and Welfare, Japan.

Reprints: Takeo Nakayama, MD, PhD, Department of Health Informatics, Kyoto University School of Public Health, Konoe-cho, Yoshida, Sakyo-ku, Kyoto 606-8501, Japan. E-mail: nakayama@pbh.med.kyoto-u.ac.jp.

ISSN: 0003-4932/06/24405-0668
DOI: 10.1097/01.sla.0000225356.04304.bc

The first randomized controlled trial in medicine was an investigation of streptomycin in 1948.[1] Since then, randomized controlled trials have been widely recognized as offering the gold standard for evaluating treatment efficacy and effectiveness and are classified as providing the highest grade of evidence in the hierarchy of research designs.[2]

Evaluations in the 1970s and 1980s suggested that observational studies may spuriously overestimate treatment benefits, yielding misleading conclusions.[3–6] In recent years, this debate has resurfaced. Some reports have suggested that for selected medical topics, both randomized and observational studies, may yield very similar results.[7,8] Conversely, opposing results have been reported from a large number of diverse medical topics.[9] Although these previous studies have contained some surgical topics, most have assessed topics involving pharmacotherapies. However, pharmacologic and surgical therapies differ in clinical nature, and results for pharmacologic investigations may therefore not apply to surgical fields.

This issue warrants investigation with a focus on the surgical area, and no previous studies appear to have undertaken an exhaustive assessment of a single clinical field. The present study investigated digestive surgery, allowing a systematic search and evaluation.

This systematic and exhaustive search of a large number of diverse articles on digestive surgery seeks to answer the following question: Do observational studies in digestive surgery tend to produce the same results as randomized controlled trials?

METHODS

Search for Meta-Analyses of Randomized Controlled Trials and Selection of Topics

Meta-analyses of randomized controlled trials in digestive surgery that had been published up to April 2004 were selected as topics in this study. Retrieved articles were judged suitable for use as topics only if all the following criteria were met: 1) meta-analysis of randomized controlled trials; 2) investigating digestive surgery; 3) assessing the treatment effects of at least one operative intervention versus any other intervention (operative or nonoperative); and 4) subjects were human. Searches were not limited to English language articles (any language). Studies were excluded if the main purpose was not evaluation of treatment effect, such as diagnosis. A literature search was

performed using the PubMed (1966 to April 2004), EMBASE (1986 to April 2004) and Cochrane Library (Issue 2, 2004) databases. A computer-assisted search was conducted using the following combination of Medical Subject Heading Terms and text words: "surgical procedures, operative," "digestive system surgical procedures," "randomized," "random," "meta-analysis," and "review." A manual search was also performed using references from the retrieved review articles.

Search for Observational Studies for Meta-Analysis

If meta-analyses of both randomized and observational studies had been performed on the same topic in each selected review article, the results could be used for comparison. However, if meta-analysis of observational studies had not been performed, we attempted to perform that by ourselves. Thus, when a meta-analysis of observational studies could not be identified in the selected review article, we needed to search for such meta-analyses while gathering observational studies under the following process.

For meta-analyses of observational studies, we first searched observational studies for all selected topics. In each topic, the same inclusion criteria used for meta-analysis of randomized controlled trials were used, with the exception of study design. Observational study designs were used if they could be categorized as prospective nonrandomized studies, retrospective cohort studies, case-control studies, case series with control groups, or other unspecified designs (provided a control group was used). A literature search was performed using the PubMed (1966 to April 2004), EMBASE (1986 to April 2004) and Cochrane Library (Issue 2, 2004) databases. PubMed contains no search term for observational studies, so a text-word strategy was used to search for "observational," "nonrandomized," "case series," "case control study," "cohort," "retrospective," and "prospective." In addition, a manual search was performed using references from the retrieved review articles. We also attempted to contact as many experts from the review articles as possible.

Data Extraction and Selection of Outcomes

All available binary data were extracted from the outcomes of the gathered observational studies. Data extraction was performed after translation of the article into English if the article had not been written in English or Japanese. Up to this point, 2 authors (S.S., T.N.) undertook the literature searches and data extraction independently, and disagreements were resolved by consensus.

For final inclusion of a topic in the present evaluation, binary data for the same outcome had to be available from at least one randomized trial and at least one observational study. When primary outcomes had been defined in the review article, these were used for the main comparison. Whenever the primary outcome was unclear, the outcome that was considered a priori as the most clinically important was selected, using consensus among the data extractors. In digestive surgery, mortality was generally given priority in clinical importance over other outcomes.

Statistical Analysis

For all selected topics, data from observational studies were combined. Generally, the fixed-effects model weighted by Peto's odds ratio method or the Mantel-Haenszel method was used for data pooling, followed by a test of heterogeneity.[10,11] Heterogeneity between studies was assessed using Q statistics.[12] Given the low power of this test, a significance level of 0.10 was used, rather than 0.05.[13] If the hypothesis of heterogeneity was accepted, the random-effects model using the DerSimonian-Laird method was used.[14] However, this study sought to compare summary estimates of randomized controlled trials with observational studies under equivalent conditions to the maximum extent possible. Thus, when performing meta-analysis of observational studies, we used the same method that had been used in the meta-analysis of randomized controlled trials. In this study, the quantity I^2 was used for assessing heterogeneity between trials in meta-analyses, calculated as: $I^2 = [(Q - df)/Q] \times 100$, where Q is the χ^2 statistic and df is the degrees of freedom. A value greater than 50% may be considered indicative of substantial heterogeneity.[15]

Although pooled odds ratio or pooled relative risk could be used as the indicator of summary estimates of outcomes, the present study used the same indicator that had been used in the meta-analysis of randomized controlled trials. In this context, odds ratios and relative risks will inevitably be similar in magnitude, as the rates of outcome events are low. Relative risks were therefore considered as odds ratios in comparisons of summary estimates. Confidence intervals were always calculated at 95%. When one arm of an outcome contained no events, this was considered a "zero cell" in the 2 × 2 table. Zero cells create problems in computing ratio measures of treatment effect. This problem was dealt with using a common method of adding 0.5 to each cell of the 2 × 2 table for the trial.[16]

To evaluate concordance between the results of randomized and observational studies, the following analyses were performed: 1) assessment of the number of cases in which the summary estimates of the observational studies suggested an effect at least double that of the randomized trials; and 2) evaluation of whether differences in the summary odds estimates of randomized controlled trials and observational studies for the same topic were larger than what would be expected by chance alone. To accomplish this, Z scores were calculated as follows:

$$Z = [\ln(OR_{RCT}) - \ln(OR_{OBS})] / \{\mathrm{var}[\ln(OR_{RCT})] + \mathrm{var}[\ln(OR_{OBS})]\}^{1/2},$$

where $\ln(OR_{RCT})$ is the natural logarithm of the odds ratio or relative risk of randomized controlled trials, $\ln(OR_{OBS})$ is the natural logarithm of the odds ratio or relative risk of observational studies, and var is variance. A Z score above 1.96 or less than −1.96 suggests a nonrandom difference between randomized controlled trials and observational studies (0.05 level of statistical significance).[17]

All statistical analyses were performed using STATA statistical software version 8.1 (STATA Corporation, College Station, TX).

RESULTS

Characteristics of Topics, Observational Studies

A literature search was first performed to select meta-analyses of randomized controlled trials for the topics, identifying 1184 potentially relevant articles. The process finally identified and selected 15 meta-analyses of randomized controlled trials for digestive surgical topics in this research (Fig. 1).[7,18–31] Three of the 15 reviews contained two topics.[21,30,31] Thus, 18 topics were identified for comparison of summary estimates between randomized controlled trials and observational studies (Table 1).

Meta-analyses of observational studies could not be identified for 10 of the 18 topics (topics 2, 3, 8–13, 17, and 18), so additional meta-analyses were required. Meta-analyses of observational studies had been identified for the remaining 8 topics (topics 1, 4–7, and 14–16), and the results were used for comparisons in this study.

For meta-analyses of observational studies for the 10 topics without existing meta-analyses, a literature search was performed and 111 observational studies were selected from 10,960 articles using the process outlined in Figure 2. Of the 111 selected articles, 17 had not been written in English or Japanese, instead appearing in 7 different languages, and the 2 trial assessors therefore abstracted data from the articles after translation into English by independent translators. A total of 52 common outcomes for both randomized controlled and observational studies were available for comparison in this study.

Using the described processes, 52 outcomes of 18 topics were investigated in 276 original articles (96 randomized trials, 180 observational studies) with a total of 101,170 study patients (Table 1). The 180 observational studies comprised 36 prospective and 144 retrospective studies. Randomized and observational studies on the same topic generally administered treatment in the same way and outcome measures were similarly defined.

Between-Study Heterogeneity

Data on between-study heterogeneity using the I^2 statistic were available for all 10 meta-analyses of observational studies that we performed specifically for the present study (topics 2, 3, 8–13, 17, and 18). Conversely, data had not been described in 8 of the remaining meta-analyses that had been reported (topics 1, 4–7, and 14–16). In primary outcomes of 16 topics, significant heterogeneity was noted between randomized controlled trials in 1 of 9 topics (11.1%). Significant between-study heterogeneity was identified between observational studies in 5 of 12 topics (41.7%). There was no significant difference between the rates of heterogeneity ($P = 0.18$ by Fisher exact test).

Comparison of Primary Outcomes

In almost all topics, the primary outcome defined in the review or decided by author consensus was mortality. However, in topics dealing with safety of procedures, such as appendectomy and operation for fissure-in-ano, one of the complications, such as risk of wound infection or persistence

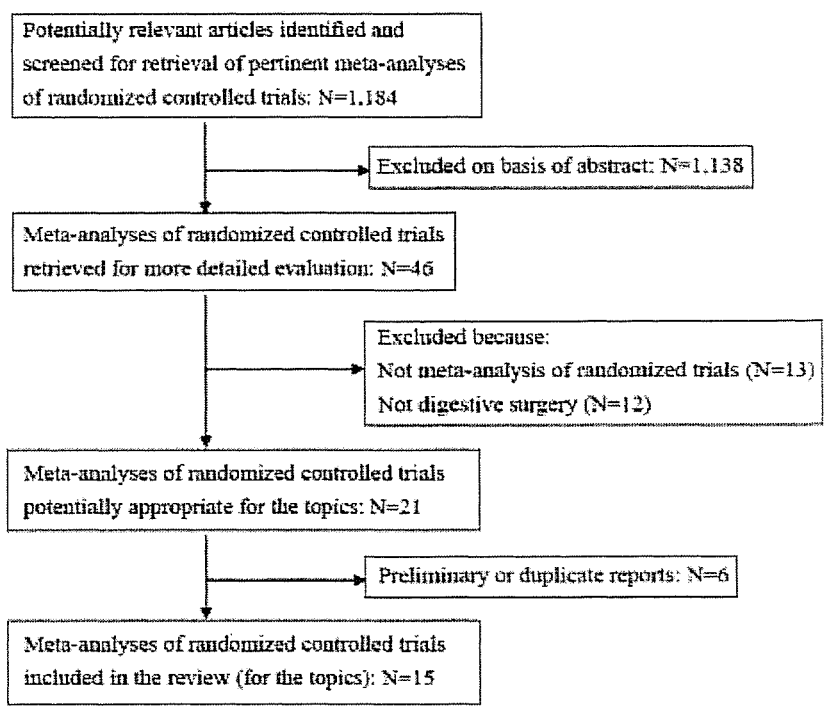


FIGURE 1. Summary profile of search for meta-analyses of randomized controlled trials.

TABLE 1. Topics of Meta-Analyses Considering Both Randomized Controlled Trials and Observational Studies

Identification No.	Topic	Randomized Controlled Trial		Observational Study (no. of studies)		No. of Comparable Outcomes
		Meta-Analysis	No. of Studies (no. of patients)	Prospective/ Retrospective	Total (no. of patients)	
1	Closed postoperative peritoneal lavage vs. no lavage for generalized peritonitis	Leiboff et al[18] (1987)	4 (173)	2/6	8 (1034)	1
2	Splenorenal shunt vs. endoscopic sclerotherapy in prevention of variceal rebleeding	Spina et al[19] (1992)	4 (310)	0/2	2 (344)	1
3	Routine drainage vs. no drainage after elective colorectal surgery	Urbach et al[20] (1999)	4 (414)	0/5	5 (1767)	4
4	Anal stretch vs. sphincterotomy for fissure-in-ano	Nelson et al[21] (1999)	6 (328)	0/4	4 (537)	2
5	Open vs. closed lateral sphincterotomy for fissure-in-ano	Nelson et al[21] (1999)	2 (140)	0/4	4 (1365)	2
6	Laparoscopic vs. open appendectomy for acute appendicitis	Benson et al[7] (2000)	16 (1703)	3/4	7 (1502)	1
7	Transthoracic vs. transhiatal resection for carcinoma of the esophagus	Hulscher et al[22] (2001)	3 (138)	3/18	21 (2466)	6
8	Hand-sewn vs. stapled esophagogastric anastomosis after esophagectomy	Urschel et al[23] (2001)	5 (467)	2/8	10 (3196)	3
9	Posterior vs. anterior route of reconstruction after esophagectomy	Urschel et al[24] (2001)	6 (342)	0/3	3 (329)	4
10	Pyloroplasty vs. no drainage in gastric reconstruction after esophagectomy	Urschel et al[25] (2002)	3 (347)	0/2	2 (111)	1
11	Primary repair vs. fecal diversion for penetrating colon injuries	Singer et al[26] (2002)	5 (467)	4/29	33 (5745)	4
12	Stapled vs. hand-sewn methods for colorectal anastomosis surgery	Lustosa et al[27] (2002)	9 (1233)	2/13	15 (3894)	6
13	Stapled vs. conventional hemorrhoidectomy	Sutherland et al[28] (2002)	7 (591)	2/5	7 (910)	3
14	Extended vs. limited lymph node dissection for adenocarcinoma of the stomach	McCulloch et al[29] (2003)	3 (1729)	5/8	13 (4058)	2
15	Open (Hasson type) vs. closed (needle/trocar) access in laparoscopic surgery	Merlin et al[30] (2003)	4 (302)	4/6	10 (20,664)	3
16	Direct trocar vs. closed (needle/ trocar) access in laparoscopic surgery	Merlin et al[30] (2003)	3 (665)	0/2	2 (1575)	3
17	Early vs. delayed open cholecystectomy for acute cholecystitis	Papi et al[31] (2004)	9 (916)	2/16	18 (37,475)	3
18	Early vs. delayed laparoscopic cholecystectomy for acute cholecystitis	Papi et al[31] (2004)	3 (228)	7/9	16 (3705)	3
Total			96 (10,493)		180 (90,677)	52

of fissure, was considered as a more appropriate primary outcome.

In 16 of 18 topics, primary outcomes could be compared between observational studies and randomized controlled trials. These summary estimates and associated 95% confidence intervals are shown in Figure 3. One of 16 primary outcomes displayed a magnitude of effect in the combined observational studies that was outside the 95% confidence interval for the

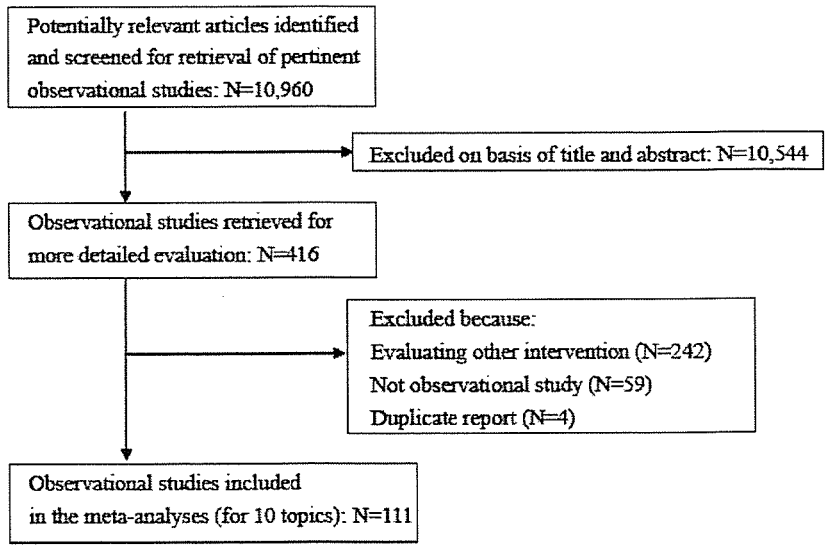


FIGURE 2. Summary profile of search for observational studies.

combined randomized controlled trials (topic 14). In 4 of 16 primary outcomes, summary estimates from observational studies were at least double those from randomized controlled trials (topics 7, 8, 15, and 17). The converse occurred in 3 topics (topics 11, 14, and 16) (exact $P = 0.45$ by Wilcoxon test). Evaluation by Z score revealed significant discrepancies between randomized trials and observational studies for 4 of 16 primary outcomes (topic 7, $Z = -4.28$; topic 8, $Z = -2.36$; topic 11, $Z = 2.19$; topic 14, $Z = 4.34$).

Comparison of All Outcomes

All summary estimates for 52 outcomes of 18 topics are shown in Table 2. Three types of calculation model were used: random effects calculation using the DerSimonian-Laird method; and fixed effects calculation using Peto's odds ratio method or the Mantel-Haenszel method. In 21 of 52 outcomes, relative risk was evaluated rather than odds ratio in meta-analyses of observational studies, as the original meta-analyses of randomized controlled trials had used relative risks for evaluations.

In 9 of 52 outcomes, summary estimates from observational studies were at least double those from randomized controlled trials. The converse occurred in 10 outcomes (exact $P = 0.943$ by Wilcoxon test). Evaluation by Z score revealed significant discrepancies between randomized trials and observational studies in 10 of 52 outcomes.

Overall, these data suggest that about one fourth of observational studies gave different results than randomized trials.

DISCUSSION

Using data from 276 articles in 18 topics, summary estimates were compared between randomized controlled trials and observational studies in digestive surgery. Significant between-study heterogeneity occurred more often between observational studies than between randomized controlled trials. One fourth of the summary estimates of treatment effects in randomized controlled trials and observational studies differed significantly from each other. From this study, observational studies in digestive surgery tend to have similar results to those by randomized controlled trials. At least, they do not tend to overestimate or underestimate more than randomized controlled trials.

Our findings support the conclusions of earlier evaluations in the 1970s and 1980s.[3–6] In 2001, Ioannidis et al investigated 45 diverse pharmacologic and surgical topics in 408 articles and concluded that observational studies tend to indicate larger treatment effects (28 of 45 topics vs. 11 of 45 topics) and between-study heterogeneity is more frequent among observational studies than among randomized controlled trials (41% vs. 23%).[9] On the other hand, previous studies by Benson and Hartz[7] and Concato et al[8] reached the opposite conclusion. Benson and Hartz[7] investigated 19 diverse pharmacologic and surgical treatments in 136 articles and found little evidence of larger or differing estimates of treatment effects in observational studies compared with randomized controlled trials. Concato et al[8] evaluated 5 clinical topics and 99 articles, concluding that well-designed observational studies do not systematically overestimate the magnitude of treatment effects when compared with randomized controlled trials on the same topic.

All these previous studies have made substantial contributions toward identifying the problems caused by differing study designs. However, conclusions have inevitably been in the form of general statements, as the studies addressed diverse topics in various clinical fields. The present study was limited to a single clinical field, digestive surgery and thus offers two advantages over previous studies: a more exhaustive search is possible in studies of diverse clinical fields; and higher applicability to clinical practice than a general statement.

In 25% of digestive surgical topics, summary estimates of treatment effects in observational studies yielded different results than randomized trials, but both designs reached similar results in the remaining topics. This may be attributable to various factors. First, quality of surgical randomized controlled trials is low according to some review articles and

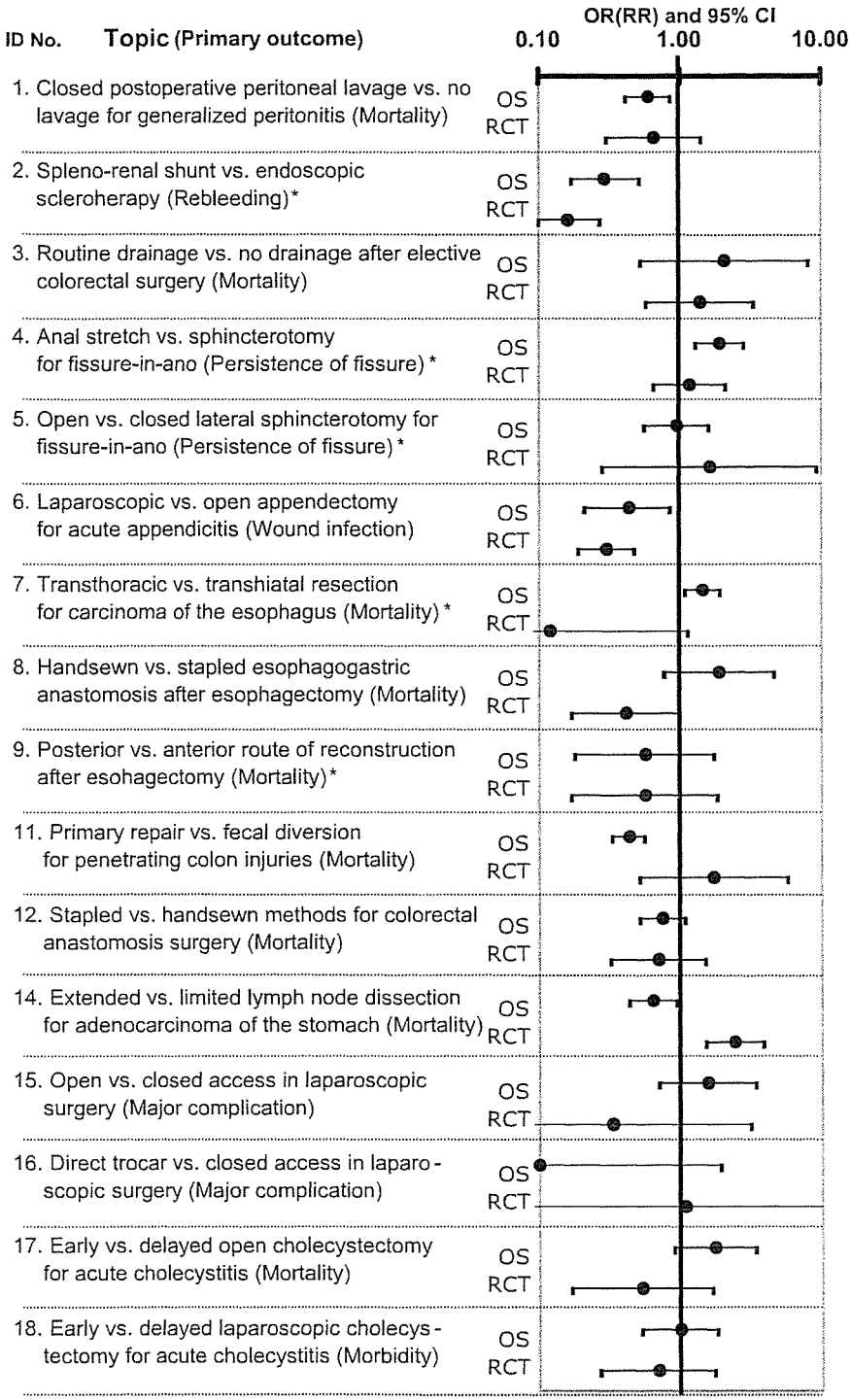


FIGURE 3. Comparison of primary outcomes between observational studies and randomized controlled trials. This figure is based on data from 13 review articles[7,18–24,26,27,29–31] and 10 meta-analyses of observational studies by the authors. OR, odds ratio; RR, relative risk; CI, confidence interval. *Outcome reporting relative risk rather than odds ratio.

may be so low that the essential contents of randomized trials do not differ from those of observational studies.[32,33] Second, for most topics, sample sizes may be too small to detect clinically important differences between the results of two types of study. Actually, 12 of 18 topics used fewer than 500 randomized patients. Combined with the use of a rare endpoint, mortality, we could expect to see very large confidence intervals in the randomized evidence. The wide confidence intervals mean that demonstrating any significant discrepancy between the two designs will be very difficult.

This study examined not only primary outcomes, but also the secondary outcomes. Generally, results about

TABLE 2. Summary Estimates for All Outcomes

Identification No.	Topic	Outcome	Randomized Controlled Trial		Observational Study		Calculation Model
			No. of Studies	Summary Estimate OR (95% CI)	No. of Studies	Summary Estimate OR (95% CI)	
1	Closed postoperative peritoneal lavage vs. no lavage for generalized peritonitis	Mortality	4	0.65 (0.30–1.40)	8	0.59 (0.41–0.85)	M-H
2	Splenorenal shunt vs. endoscopic sclerotherapy in the prevention of variceal rebleeding	Rebleeding*	4	0.16 (0.10–0.27)	2	0.29 (0.17–0.51)	M-H
3	Routine drainage vs. no drainage after elective colorectal surgery	Mortality.	4	1.38 (0.57–3.31)	2	2.05 (0.52–8.11)	M-H
		Anastomotic leak	4	1.47 (0.71–3.06)	5	1.99 (1.12–3.53)	M-H
		Pulmonary complication	4	0.81 (0.41–1.59)	2	0.94 (0.24–3.73)	M-H
		Wound infection	4	1.70 (0.87–3.30)	2	0.94 (0.24–3.73)	M-H
4	Anal stretch vs. sphincterotomy for fissure-in-ano	Persistence of fissure*	6	1.16 (0.65–2.08)	4	1.89 (1.28–2.81)	M-H
		Flatus incontinence*	4	6.63 (2.06–21.3)	4	1.34 (0.79–2.27)	M-H
5	Open vs. closed lateral sphincterotomy for fissure-in-ano	Persistence of fissure*	2	1.61 (0.28–9.28)	4	0.94 (0.55–1.58)	M-H
		Flatus incontinence*	2	0.79 (0.29–2.13)	4	1.16 (0.94–1.51)	M-H
6	Laparoscopic vs. open appendectomy for acute appendicitis	Wound infection	16	0.30 (0.19–0.47)	7	0.43 (0.21–0.84)	M-H
7	Transthoracic vs. transhiatal resection for carcinoma of the esophagus	Mortality*	3	0.12 (0.04–1.12)	20	1.43 (1.08–1.89)	M-H
		Cardiac complication*	2	0.77 (0.30–1.99)	5	1.19 (0.70–2.01)	M-H
		Pulmonary complication*	2	0.85 (0.53–1.38)	10	1.20 (0.99–1.46)	M-H
		Anastomotic leak*	3	1.20 (0.34–4.25)	14	0.49 (0.38–0.64)	M-H
		Vocal cord paralysis*	2	0.98 (0.14–6.59)	9	0.51 (0.33–0.78)	M-H
		3-yr survival*	1	1.83 (0.70–4.78)	8	1.44 (1.12–1.86)	M-H
8	Hand-sewn vs. stapled esophagogastric anastomosis after esophagectomy	Mortality	4	0.41 (0.17–0.98)	3	1.87 (0.76–4.57)	D-L
		Anastomotic leak*	5	0.79 (0.44–1.42)	10	1.77 (1.22–2.56)	D-L
		Anastomotic stricture*	4	0.60 (0.27–1.33)	7	0.79 (0.41–1.50)	D-L
9	Posterior vs. anterior route of reconstruction after esophagectomy	Mortality*	3	0.56 (0.17–1.82)	3	0.56 (0.18–1.72)	D-L
		Anastomotic leak*	4	1.01 (0.35–2.94)	3	0.28 (0.10–0.79)	D-L
		Pulmonary complication*	3	0.67 (0.34–1.33)	3	0.81 (0.50–1.34)	D-L
		Cardiac complication*	3	0.43 (0.17–1.12)	2	0.87 (0.44–1.74)	D-L
10	Pyloroplasty vs. no drainage in gastric reconstruction after esophagectomy	Pulmonary complication*	2	0.69 (0.42–1.14)	2	4.07 (0.91–18.3)	D-L
11	Primary repair vs. fecal diversion for penetrating colon injuries	Mortality	5	1.70 (0.51–5.70)	25	0.43 (0.33–0.55)	Peto
		Morbidity	5	0.28 (0.18–0.42)	20	0.73 (0.60–0.90)	Peto
		Intraabdominal infection	5	0.59 (0.38–0.94)	20	0.60 (0.49–0.74)	Peto
		Wound infection	5	0.55 (0.34–0.89)	18	0.78 (0.62–0.98)	Peto
12	Stapled vs. hand-sewn methods for colorectal anastomosis surgery	Mortality	7	0.69 (0.32–1.49)	12	0.74 (0.51–1.07)	Peto
		Anastomotic leak	9	0.99 (0.71–1.40)	11	1.16 (0.82–1.64)	Peto
		Anastomotic stricture	7	3.59 (2.02–6.35)	5	3.78 (1.40–10.2)	Peto
		Hemorrhage	4	1.78 (0.84–3.81)	2	0.59 (0.08–4.19)	Peto
		Reoperation	3	1.94 (0.95–3.98)	3	0.18 (0.12–0.26)	Peto
		Wound infection	6	1.43 (0.67–3.04)	5	1.28 (0.83–1.97)	Peto

(Continued)

TABLE 2. (*Continued*)

Identification No.	Topic	Outcome	Randomized Controlled Trial		Observational Study		Calculation Model
			No. of Studies	Summary Estimate OR (95% CI)	No. of Studies	Summary Estimate OR (95% CI)	
13	Stapled vs. conventional hemorrhoidectomy	Thrombosis of external piles*	2	0.56 (0.19–1.61)	2	0.71 (0.14–3.58)	M-H
		Urinary retention*	3	0.59 (0.28–1.24)	3	0.41 (0.23–0.72)	M-H
		Anal stenosis (2–6 wk)*	2	1.07 (0.36–3.17)	3	0.55 (0.16–1.83)	M-H
14	Extended vs. limited lymph node dissection for adenocarcinoma of the stomach	Mortality	2	2.39 (1.50–3.82)	2	0.63 (0.43–0.93)	M-H
		5-yr survival	2	0.92 (0.72–1.17)	2	1.17 (0.97–1.42)	M-H
15	Open (Hasson type) vs. closed (needle/trocar) access in laparoscopic surgery	Major complication	1	0.33 (0.04–3.13)	6	1.54 (0.70–3.40)	M-H
		Minor complication	2	0.82 (0.44–1.54)	5	0.52 (0.26–1.05)	M-H
		Conversion to laparotomy	2	0.32 (0.05–1.96)	4	0.43 (0.16–1.21)	M-H
16	Direct trocar vs. closed (needle/trocar) access in laparoscopic surgery	Major complication	1	1.07 (0.07–16.9)	1	0.09 (0.00–1.90)	M-H
		Minor complication	3	0.19 (0.09–0.40)	2	0.07 (0.04–0.14)	M-H
		Conversion to laparotomy	1	1.17 (0.16–8.58)	1	0.09 (0.00–1.90)	M-H
17	Early vs. delayed open cholecystectomy for acute cholecystitis	Mortality	9	0.53 (0.17–1.66)	13	1.73 (0.89–3.37)	D-L
		Morbidity	9	0.95 (0.66–1.38)	12	0.95 (0.59–1.54)	D-L
		Common bile duct injuries	9	0.66 (0.20–2.17)	3	1.35 (0.24–7.61)	D-L
18	Early vs. delayed laparoscopic cholecystectomy for acute cholecystitis	Morbidity	3	0.69 (0.27–1.73)	11	0.98 (0.53–1.80)	D-L
		Common bile duct injuries	3	0.70 (0.07–6.19)	11	1.27 (0.56–2.87)	D-L
		Conversion to laparotomy	3	0.62 (0.32–1.19)	16	0.39 (0.14–1.07)	D-L

OR, odds ratio; CI, confidence interval; M-H, Mantel-Haenszel method; D-L, DerSimonian-Laird method; Peto, Peto's odds ratio method.
*Outcome reporting relative risk rather than odds ratio.

concordance of different studies may vary depending on whether primary or secondary outcomes are examined. Discrepancies may be less apparent for secondary outcomes than for primary outcomes because secondary events are likely to be too uncommon to show any significant difference between arms except in extremely large trials (mega-trials).[17]

One possible explanation for the greater frequency of between-study heterogeneity in observational studies than in randomized trials is that each observational study usually includes a wide spectrum of subjects from the population at risk. In contrast, randomized trials use specific inclusion criteria and may not be representative of populations seen in clinical practice.

All topics examined in this study were comparisons in the form of A versus B. Generally, A represented a new procedure while B represented an accepted method, but deciding which was newer was difficult in some topics. Most trials in medicine estimate the benefits of pharmacologic effects, whereas 50 of 52 outcomes in this study estimate risks of operations, such as mortality and morbidity. Discrepancies in summary estimates were estimated accordingly between randomized trials and observational studies. For example, the greatest statistical discrepancy between the two types of study design was topic 14 (mortality), comparing extended and limited lymph node dissections for adenocarcinoma of the stomach ($Z = 4.34$). In this topic, although the summary estimate from observational studies was one fourth that from randomized trials (0.63 vs. 2.39), this represented an underestimation of risks, not of benefits.

The authors revealed that one fourth of observational studies gave different results to randomized trials and between-study heterogeneity was more common in observational studies in the field of digestive surgery. Furthermore, even if clinical applicability is improved by combining a large number of observational studies, estimations of treatment effect sometimes differ from those obtained from randomized controlled trials. The present study confirmed such tendencies in the well-defined area of digestive surgery. However, observational studies offer several advantages over randomized controlled trials, including lower cost, greater timeliness, and a broader range of patients.[34] These benefits remain worthy of attention in real clinical settings, particularly where random allocation is not easily accepted by either clinicians or patients. In the field of digestive surgery, large

observational studies may actually be more reliable than small underpowered randomized controlled trials. To clarify how to interpret the findings of observational studies and randomized controlled trials, further analyses in other fields are eagerly awaited.

REFERENCES

1. Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. *BMJ.* 1948;2:769–782.
2. Preventive Services Task Force. *Guide to Clinical Preventive Services: Report of the U.S. Preventive Services Task Force*, 2nd ed. Baltimore: Williams & Wilkins, 1996.
3. Chalmers TC, Matta RJ, Smith H Jr, et al. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med.* 1977;297:1091–1096.
4. Sacks HS, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med.* 1982;72:233–240.
5. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I. *Med Stat Med.* 1989;8:441–454.
6. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II. *Surgical Stat Med.* 1989;8:455–466.
7. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med.* 2000;342:1878–1886.
8. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med.* 2000;342:1887–1892.
9. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA.* 2001;286:821–830.
10. Yusuf S, Peto R, Lewis J, et al. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis.* 1985;27:335–371.
11. Mantel N, Haenszel WH. Statistical aspects of the analysis of data from retrospective studies of diseases. *J Natl Cancer Inst.* 1959;22:719–748.
12. Fleiss JL. *Statistical Methods for Rates and Proportions.* New York: Wiley, 1981.
13. Fleiss JL. Analysis of data from multiclinic trials. *Control Clin Trials.* 1986;7:267–275.
14. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials.* 1986;7:177–188.
15. Julian PTH, Simon GT, Jonathan JD, et al. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327:557–560.
16. Matthias E, George DS, Douglas GA. *Systematic Reviews in Health Care: Meta-Analysis in Context.* 2nd ed. London: BMJ Books, 2001.
17. Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons of meta-analyses and large trials. *JAMA.* 1998;279:1089–1093.
18. Leiboff AR, Soroff HS. The treatment of generalized peritonitis by closed postoperative peritoneal lavage: a critical review of the literature. *Arch Surg.* 1987;122:1005–1010.
19. Spina GP, Henderson JM, Rikkers LF, et al. Distal spleno-renal shunt versus endoscopic sclerotherapy in the prevention of variceal rebleeding: a meta-analysis of 4 randomized clinical trials. *J Hepatol.* 1992;16:338–345.
20. Urbach DR, Kennedy ED, Cohen MM. Colon and rectal anastomoses do not require routine drainage: a systematic review and meta-analysis. *Ann Surg.* 1999;229:174–180.
21. Nelson RL. Meta-analysis of operative techniques for fissure-in-ano. *Dis Colon Rectum.* 1999;42:1424–1428.
22. Hulscher JB, Tijssen JG, Obertop H, et al. Transthoracic versus transhiatal resection for carcinoma of the esophagus: a meta-analysis. *Ann Thorac Surg.* 2001;72:306–313.
23. Urschel JD, Blewett CJ, Bennett WF, et al. Handsewn or stapled esophagogastric anastomoses after esophagectomy for cancer: meta-analysis of randomized controlled trials. *Dis Esophagus.* 2001;14:212–217.
24. Urschel JD, Urschel DM, Miller JD, et al. A meta-analysis of randomized controlled trials of route of reconstruction after esophagectomy for cancer. *Am J Surg.* 2001;182:470–475.
25. Urschel JD, Blewett CJ, Young JE, et al. Pyloric drainage (pyloroplasty) or no drainage in gastric reconstruction after esophagectomy: a meta-analysis of randomized controlled trials. *Dig Surg.* 2002;19:160–164.
26. Singer MA, Nelson RL. Primary repair of penetrating colon injuries: a systematic review. *Dis Colon Rectum.* 2002;45:1579–1587.
27. Lustosa SA, Matos D, Atallah AN, et al. Stapled versus handsewn methods for colorectal anastomosis surgery: a systematic review of randomized controlled trials. *Sao Paulo Med J.* 2002;120:132–136.
28. Sutherland LM, Burchard AK, Matsuda K, et al. A systematic review of stapled hemorrhoidectomy. *Arch Surg.* 2002;137:1395–1406.
29. McCulloch P, Nita ME, Kazi H, et al. Extended versus limited lymph nodes dissection technique for adenocarcinoma of the stomach. *Cochrane Database Syst Rev.* 2003;(4):CD001964.
30. Merlin TL, Hiller JE, Maddern GJ, et al. Systematic review of the safety and effectiveness of methods used to establish pneumoperitoneum in laparoscopic surgery. *Br J Surg.* 2003;90:668–679.
31. Papi C, Catarci M, D'Ambrosio L, et al. Timing of cholecystectomy for acute calculous cholecystitis: a meta-analysis. *Am J Gastroenterol.* 2004;99:147–155.
32. Solomon MJ, McLeod RS. Surgery and the randomized controlled trial: past, present and future. *Med J Aust.* 1998;169:380–383.
33. McCulloch P, Taylor I, Sasako M, et al. Randomized trials in surgery: problems and possible solutions. *BMJ.* 2002;324:1448–1451.
34. Feinstein AR. Epidemiologic analyses of causation: the unlearned scientific lessons of randomized trials. *J Clin Epidemiol.* 1989;42:481–489.

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2006; 15: 813–822
Published online 18 September 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/pds.1317

ORIGINAL REPORT

Effectiveness and safety of ritodrine hydrochloride for the treatment of preterm labour: a systematic review†

Yukari Yaju MPH[1] and Takeo Nakayama MD, PhD[2]*

[1]*Department of Epidemiology and Preventive Health Sciences, Graduate School of Medicine, University of Tokyo, Tokyo, Japan*
[2]*Department of Health Informatics, School of Public Health, Kyoto University, Kyoto, Japan*

SUMMARY

Purpose To analyse the available data on the effectiveness and safety of ritodrine hydrochloride in delaying delivery and in decreasing the incidence of preterm birth.

Methods Systematic review of randomised controlled trials (RCTs) that compared the effectiveness and safety of ritodrine hydrochloride with a placebo or with no treatment. Main outcome measures were relative risks (RRs) for perinatal mortality, neonatal respiratory distress syndrome (RDS), delivery within 48 hours or 7 days, preterm birth before 37 weeks gestation and low birth weight. We searched computerised databases (MEDLINE, CENTRAL, Ichushi Web) from their inception to October 2004, and searched the references of eligible trials.

Results Seventeen RCTs were included and meta-analysis was conducted. Pooled RRs relative to placebo for delivery within 48 hours or 7 days for parenteral ritodrine hydrochloride were 0.74 (95%CI (confidential interval): 0.56, 0.97), 0.85 (95%CI: 0.74, 0.97). There was no significant decrease in perinatal mortality, the proportion of RDS, preterm birth and low birth weight infants. Maternal side-effects significantly increased in patients receiving ritodrine with respect to those receiving a placebo. Pooled RRs relative to placebo for oral ritodrine hydrochloride showed no significant decrease in primary and secondary endpoints.

Conclusions The effectiveness of parenteral ritodrine hydrochloride for tocolysis in preterm labour is limited to short-range prolongation of gestation. The effectiveness of maintenance tocolytic therapy with oral ritodrine hydrochloride was not proved. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS—systematic review; meta-analysis; ritodrine hydrochloride; preterm labour; tocolysis

Received 5 August 2005; Revised 5 April 2006; Accepted 22 July 2006

INTRODUCTION

Prevention of preterm birth is currently a primary concern in perinatal medicine. The rate of preterm delivery ranges from 5 to 10% in the developed countries, and it has been increasing at an alarming rate in recent years. Approximately 30% of the threatened premature delivery is believed to be caused by multiple pregnancy, while 20–25% is caused by uterine disorders such as premature rupture of the membranes (PROM: rupture of membranes which occurs before labour begins and has been associated with complications such as intrauterine infection) and cervical incompetence, 15–20% by the general physical condition of the mother or factors of the foetus and 30% by unknown factors.[1] Although there

*Correspondence to: Dr T. Nakayama, Department of Health Informatics, School of Public Health, Kyoto University, Konoe-cho, Yoshida, Sakyo-ku, City of Kyoto, Kyoto 606-8501, Japan. E-mail: nakayama@pbh.med.kyoto-u.ac.jp
†No conflict of interest was declared.

is a general consensus that threatened premature delivery should be treated causally when causes of it are known, most cases of uterine contractions that occur before 37 weeks gestation are treated with tocolytic drugs. According to the American College of Obstetricians and Gynecologists' guidelines, while tocolytic drugs may be useful for short-term prolongation of pregnancy, they do not improve perinatal outcomes (such as perinatal mortality, neonatal respiratory distress syndrome (RDS), etc.). The primary objective of tocolytic therapy is to prolong pregnancy long enough to provide time for the administration of antenatal steroids to improve foetal lung maturity and maternal transport to a neonatal intensive care unit (NICU).

Although terbutaline sulphate and magnesium sulphate are used as first-line tocolytic agents in most Western countries, commonly used drugs for tocolysis vary in each country and obstetric ward. The first-line tocolytic agent which is safe and effective at the true endpoint has not yet been established.[2]

In Japan, the low birth weight rate increased from 5.7% in 1970 to 8.6% in 2000. Approximately two-thirds of the low birth weight cases are attributed to either preterm labour or PROM. The rate of preterm labour in Japan also increased from 4.1% in 1980 to over 5% after 2000. Although ritodrine hydrochloride (hereinafter referred to ritodrine) is not commonly used for tocolysis in Western countries, it has been used to prolong pregnancy for more than 7 days in most obstetric wards in Japan. Furthermore, oral ritodrine are currently used as maintenance therapy following acute tocolysis in 10% of obstetric wards in Japan.[3] Ritodrine is one of the betamimetics which are widely used especially in resource-poor countries.[4]

Considering the possibility of serious adverse reactions (e.g. pulmonary edema), re-evaluation of the effectiveness and safety of ritodrine is essential for improving drug therapy for preventing preterm birth.

METHODS

Searching and selection

A systematic electronic search of MEDLINE (1966-October, 2004), the Cochrane Central Register of Controlled Trials (CENTRAL in Cochrane Library 2004, issue 3) and Igaku Chuo Zasshi(Japan Centra Revuo Medicina) Web (hereinafter referred to as Ichushi Web; 1983-October, 2004). This was supplemented by searching the reference lists of all retrieved articles. The bibliographic record and abstracts of each of the retrieved articles were checked against the inclusion/exclusion criteria (hereinafter referred to as selection criteria) (Table 1).

We limited our search to English-language abstracts obtained on MEDLINE and CENTRAL search. Text words that were applied to the search included 'preterm labour', 'premature labour', tocolysis and ritodrine (Table 2).

Assessment of methodological quality

The methodological quality of individual trials was assessed independently by two investigators (Y.Y., T.N.) using five selection criteria (Table 3). The five criteria were established based on *the Cochrane*

Table 1. Inclusion/exclusion criteria

Category	Inclusion/exclusion criteria
Patients	Pregnant women
Disease	Threatened preterm delivery (the following studies were excluded) • Trials of intravenous ritodrine hydrochloride involving only women with multiple pregnancy • Trials of oral ritodrine hydrochloride involving only women with multiple pregnancy and trials of oral ritodrine hydrochloride involving only women with premature rupture of the membranes (PROM)
Intervention	Acute tocolysis with intravenous ritodrine hydrochloride or maintenance therapy after acute tocolysis with oral ritodrine hydrochloride
Outcomes	Trials measuring either one of the following outcomes: perinatal mortality, proportion of neonatal RDS, proportion of preterm birth, proportion of birth within 48 hours/7 days of treatment initiation
Year of publication	MEDLINE: 1966–October, 2004 The Cochrane Library 2004 issue 3 Ichushi Web: 1983–October, 2004
Language	English, Japanese
Study design	RCTs (trials that include description of random allocation) using either a placebo control or a non-treatment control

Table 2. Search strategy for MEDLINE

No.	Most recent queries	Results
#1	Search 'Ritodrine'[MeSH] OR ritodrine	936
#2	Search 'Labor, Premature'[MeSH] OR 'preterm labor' OR 'preterm labour' OR 'premature labor' OR 'premature labor' OR tocolysis	11 805
#3	Search #1 AND #2	571
#4	Search #1 AND #2 Field: All Fields, Limits: Randomized Controlled Trial	65
#5	Search #3 AND randomi*	73
#6	Search #3 AND 'double blind*'	29
#7	Search #4 OR #5 OR #6	101

Reviewers' Handbook (Section 6. Assessment of study quality). Because the validity of measuring the quality of individual trials by means of a quality scoring system is questionable, the methodological quality was determined by reviewing the individual trials to examine whether or not and to what extent each criterion was met.

Data extraction

The following information was extracted independently by two investigators: Bibliographic records, study methods, subjects, interventions and outcomes. A predefined set of data were extracted from the selected trials to create structured abstract tables. The primary endpoints were perinatal death and incidence of respiratory distress syndrome (RDS). The secondary endpoints were birth within 48 hours/7 days of treatment initiation, preterm delivery (birth before 37 weeks gestation), the low birth weight (live-born infants weighing <2500 g), recurrent uterine contractions after acute tocolysis (extracted only for trials with oral ritodrine) and the adverse reactions (palpitation, chest pain).

Assessment of conceptual heterogeneity

The selected trials were analysed and compared with respect to study quality and subjects characteristics in order to assess conceptual heterogeneity.

Table 3. Criteria for methodological quality

(1) Method of allocation concealment: description of random allocation methods
(2) Masking (blinding): description of double-blinding
(3) Sample size calculation: description of methods used to determine sample size
(4) Completeness of follow-up: follow-up rate is more than 80%
(5) Explicit definition of threatened preterm delivery

Quantitative data synthesis

To pool dichotomous outcomes, we calculated estimates of relative risk (RR) and odds ratio (OR) using a random effects model described by DerSimonian and Laird, as well as a Mantel-Haenszel fixed effects model with the Review Manager version 4.2.7.[5,6] We report in this analysis RRs and those 95% confidential intervals (CIs) calculated with random effects model. Data were analysed using an intention-to-treat (ITT) approach. In order to conduct a sensitivity analysis, the selected trials were grouped in six different ways (pooling groups) based on the types of inter-study conceptual heterogeneity. In order to assess statistical heterogeneity of study results in each of the six pooling groups, we calculated Q-statistics, p-value for heterogeneity test and I^2 (as a measure of inconsistency across study results). The p-value <0.1 and/or $I^2 > 50\%$ was considered significant. Publication bias was visually assessed using the funnel plot.

RESULTS

Searching and selection

The title and abstracts of 101 articles retrieved via MEDLINE were reviewed against the selection criteria. A total of 12 articles (i.e. 8 on intravenous ritodrine and 4 on oral ritodrine) met the selection criteria. Of the eight articles on intravenous ritodrine, two were the same trial, and thus, one of them was excluded. A total of 11 articles were selected.

Although 140 randomised controlled trials (RCTs) were found in the CENTRAL database, those articles that met the selection criteria turned out to be exactly the same articles as the ones selected from MEDLINE search results. However, four RCTs were found in the Cochrane systematic review by Anotayanonth *et al.*,[4] which was listed in the Cochrane Database of Systematic Reviews (CDSR). Because each of these four RCTs was unpublished, the data were extracted

via the Cochrane review and a review article by Merkatz.[7]

Articles including the terms 'ritodrine' and '*Utemerin*' were searched in the Ichushi Web. After excluding case reports and reviews, a total of 353 articles were retrieved. The titles and abstracts were reviewed against the selection criteria. Only one article on oral ritodrine met the selection criteria. One RCT of intravenous ritodrine which did not meet the selection criteria was included in the analysis for the following reasons: the trial was conducted among the Japanese women and provided certain information on adverse reactions to ritodrine. Thus, a total of two articles from Ichushi Web were selected.

Overall, 17 RCTs were included in the analysis. The process of literature search and selection is illustrated in Figure 1.

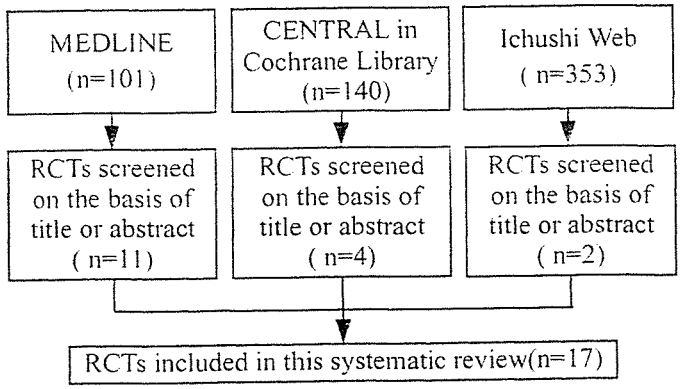


Figure 1. Flow diagram of RCTs retrieved from database search

Assessment of methodological quality

The individual articles were checked to confirm whether or not and to what extent each criterion listed in Table 3 was met. Some articles lacked an adequate description of the random allocation methods, while others did not describe sample size calculation. Although the quality of the retrieved articles was generally low, no significant variation was observed in the quality of the selected studies. Thus, all 17 selected articles were included in the analysis.

Characteristics of eligible trials—conceptual heterogeneity

Seventeen RCTs included in the analysis are listed in Table 4.

Study quality. In terms of sample size, an article by CPLIG [8] (The Canadian Preterm Labor Investigators Group) had the highest quality of all RCTs selected. No significant variation was observed in the study quality of other articles. Two articles by Larsen [13,11] provided no information regarding the outcome data of patients excluded from the analysis. Considering

Table 4. Characteristics of randomised controlled trials on effectiveness of riodrine hydrochloride

Treatment	Author	Year	No. of participants		Participants characteristics			Antenatal steroids
			Treatment	Control	Weeks of gestation	Multiple pregnancy	PROM	
Acute tocolysis (IV)	CPLI group[8]	1992	352	356	20–35	T	Included	Used
	Garite TJ[9]	1987	39	40	25–30	S	PROM only	Not used
	Leveno KJ[10]	1986	54	52	24–33	T	Not included	Not used
	Larsen JF[11]	1986	63	62	20–36	S	Included	N/S
	Tohoku group[12]	1984	22	25	24–36	N/S	Not included	N/S
	Larsen JF[13]	1980	150	49	20–36	S	Not included	Used
	Barden TP[7]	1980	12	13	20–36	S	N/S	N/S
	Hobel CJ[7]	1980	16	15	20–36	S	N/S	N/S
	Mariona A[7]	1980	5	6	20–36	S	N/S	N/S
	Scommegna A[7]	1980	15	17	20–36	S	N/S	N/S
	Spellacy WN[14]	1979	14	15	20–36	S	Included	N/S
	Wesselius CA[15]	1971	43	38	20–36	N/S	Included	N/S
Maintenance therapy (oral)	Holleboom CA[16]	1996	50	45	~34	T	Not included	N/S
	Ricci JM[17]	1991	25	25	24–34	S	Not included	Used
	Sakamoto S[18]	1985	98	95	24–37	T	Not included	N/S
	Creasy RK[19]	1980	35	35	20–36	T	Not included	N/S
	Walters WA[20]	1977	24	23	28–32	N/S	Not included	Not used

N/S, not stated; T, including twins; S, singleton only.