

Systematic Review Protocol

This is the fifth version of the protocol, last modified on the 16.03.20 (original: 24.06.19).
This protocol follows the recommendations of the PRISMA-P 2015 statement.(1)(2)

ADMINISTRATIVE INFORMATION

Title

Effects of Clinical Diagnostic Decision Support Systems based on Machine Learning on Physicians' Performance – Protocol for a Systematic Review

Registration

This protocol for a systematic review is registered with the International Prospective Register of Systematic Reviews (PROSPERO). The registration was made on the 24th of June 2019 and not updated since. The registration number is 140075.

Authors

First reviewer: **Baptiste Vasey**, Nuffield Department of Surgical Sciences,
University of Oxford, Oxford, UK
baptiste.vasey@nds.ox.ac.uk

Second reviewers: **Nicole Bilbro**, Maimonides Medical Center, Brooklyn, NY, USA
nicole.bilbro@maimonidesmed.org

Neale Marlow, Nuffield Department of Surgical Sciences,
University of Oxford, Oxford, UK
neale.marlow@trinity.ox.ac.uk

Stephan Ursprung, Department of Radiology,
University of Cambridge, Cambridge UK
su263@cam.ac.uk

Benjamin Beddoe, Faculty of Medicine, Imperial College,
London, UK
benjamin.beddoe15@imperial.ac.uk

Elliott Taylor, Nuffield Department of Surgical Sciences,
University of Oxford, Oxford, UK
elliott.taylor@trinity.ox.ac.uk

Guarantor: **Prof Peter McCulloch**, Nuffield Department of Surgical Sciences,
University of Oxford, Oxford, UK
peter.mcculloch@nds.ox.ac.uk

Corresponding author: **Baptiste Vasey**
Nuffield Department of Surgical Sciences
University of Oxford
Level 5, Room 5402
John Radcliffe Hospital
Headington
Oxford, OX3 9DU

Contributions

BV designed the search strategy, wrote the present protocol and will be first reviewer during the abstracts screening and full texts review phases. NB supported the development of the search strategy, reviewed the protocol and will be second reviewer during the abstracts screening and full texts review phases. NM reviewed the protocol and will be second reviewer during the abstracts screening and full texts review phases. SU reviewed the protocol and will be second reviewer and resolve conflicts during the abstracts screening and full texts review phases. BB and XX will be PM reviewed the protocol, will resolve conflicts during the abstracts screening and full texts review phases and is the guarantor. All authors will contribute to the data extraction and analysis, and to the writing of the final manuscript.

Amendments

All amendments to the present protocol shall be documented under this section and in the PROSPERO record. All amendments shall be complemented by a description, a rational and a date for the change.

13.07.19 Following a request from the PROSPERO administrator, the synthesis plan in the “Data synthesis” section was described in more details.

Old: “Due to the expected heterogeneity of the systematic review’s target studies, the authors do not plan a meta-analysis at the time of writing this protocol. A descriptive synthesis and an analysis of the reported outcomes in line with the systematic review’s objectives will be performed. Subgroups analysis will be performed according to algorithm design, degree of support, medical specialty and any other coherent groups that would emerge from the included studies.”

New: “A narrative synthesis of the reported outcomes in line with the systematic review’s objectives will be performed, including differences in performance between the intervention and control groups as well as between the intervention group and the computer system alone. Underlying factors possibly explaining changes in effect size or direction will be investigated. The authors expect a noticeable variability in the metrics used to assess performance. A summary table of the these metrics will be presented. Qualitative data will be presented descriptively as recommended in the PRISMA elaboration and explanation document.”
“If a subgroup is sufficiently homogenous in term of study population and performance metrics, a quantitative synthesis of the performance metrics will be considered. The minimal

number of studies required for this synthesis will depend on the number of participants in each study.”

09.10.19 The intervention criteria have been clarified to address uncertainties arisen during abstracts screening.

Old: “Interactive use of a decision support system based on clinical data and machine learning algorithms to improve diagnosis or diagnostic investigations planning. In the context of this review, machine learning algorithms are defined as algorithms that have the ability to independently learn from clinical data knowledge unknown to their programmers and to generate outputs that have not been explicitly programmed.”

New: “Interactive use of a decision support system based on clinical data and machine learning algorithms to improve diagnosis or diagnostic investigations planning. In the context of this review, machine learning algorithms are defined as algorithms that have the ability to independently learn, from clinical data, knowledge unknown to their programmers and to generate outputs that have not been explicitly programmed. Machine learning models considered as general medical statistics, such as linear regression and logistic regression are not included. Diagnosis is defined as “the identification of the nature of an illness” (Oxford Dictionary).”

09.10.19: Two new second reviewers are added.

Benjamin Beddoe, Elliott Taylor

26.11.19: The list of data items to be extracted is modified to reflect the feedback generated during the piloting of the extraction table.

Old: “The following data will be extracted if present:

- study population: number, specialty, seniority
- patient population: in-/outpatient, type of medical conditions, centre size
- dataset: type of sample, sample size and number of events (for training and validation sets), source
- experiment: number of cases per physician, chronology, blinding process, familiarity with the system
- main purpose of the decision support system
- system characteristics: degree of support (tailored information display, highlighted information display, choice of several recommendations, unique recommendation; this scale will be adapted to better reflect the variety of decision support systems encountered), type of recommendations, timing of the recommendation, mathematical model used, attempts to increase the interpretability of the model
- metrics of human performance: type and value of all the metrics used to assess human performance with and without the support system, including, but not limited to, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), positive and negative predictive values, precision, accuracy, recall, position, time to decision, inter- and intra-operator variability, usability, clinical outcomes (mortality, morbidity, adverse events) and institutional outputs (average cost of treatment, length of stay)

- metrics of computer performance: type and value of all the metrics used to assess the performance of the decision support system alone, including, but not limited to, sensitivity, specificity, AUC, positive and negative predictive values, precision, accuracy, recall, position and time to decision.
- study funding: provenance, amount
- existence of a published study protocol”

New: “The following data will be extracted if present:

- study population: number, specialty, seniority
- patient population: type of medical conditions, number of different hospital sites
- dataset: type of sample, sample size and number of events (for training and validation sets), independence of training and test sets
- experiment: task to be performed, experimental design, number of cases per physician, timing of support, gold standard comparison, familiarity with the system.
- main purpose of the decision support system
- system characteristics: mathematical model used, International Medical Device Regulators Forum (IMDRF) risk classification, type of support, , attempts to increase the interpretability of the model
- metrics of human performance: type and value of all the metrics used to assess human performance with and without the support system, including, but not limited to, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), positive and negative predictive values, precision, accuracy, recall, position, time to decision, inter- and intra-operator variability, usability, clinical outcomes (mortality, morbidity, adverse events) and institutional outputs (average cost of treatment, length of stay)
- metrics of computer performance: type and value of all the metrics used to assess the performance of the decision support system alone, including, but not limited to, sensitivity, specificity, AUC, positive and negative predictive values, precision, accuracy, recall, position and time to decision.
- study funding: provenance
- existence of a published study protocol”

26.11.19: One exclusion criterion has been added to strengthen the theoretical approach.

- describing decision support systems based on natural language processing only

16.03.20: The time period considered for inclusion was reduced to 01.01.2010 – 31.05.19. This change was decided for the following reasons. I) The nomenclature used to describe the publications of interest has evolved over time and using the described search strategy over an unrestricted period of time would only yield a partial coverage. II) Several publications describe only the commercial names of the systems tested. With increasing elapsed time since publication, it becomes more and more difficult to contact the authors or the manufacturers to obtain details critical to assess inclusion criteria. III) It is common practice in the field to limit the search to the last few years.

This change was made based on observations obtained during the full text screening phase and before any data extraction started.

Old: “Years: 1806 (PsycINFO) / 1946 (Medline) / 1974 (Embase) to 31.05.2019.”

New: “01.01.2010 to 31.05.19”

16.03.20: The assessment of bias strategy was updated to better reflect the specificity of the included publications.

Old: “The risk of bias in individual studies will be assess using the QUADAS-2 tool (3) modified after Riches.(4) QUADAS-2 was developed to assess the risk of bias in studies investigating diagnostic tests and is recommended by by the National Institute for Health and Care Excellence (NICE) and the Agency for Healthcare Regulation and Quality (AHRQ). The tool assesses four different components of the study design independently (patient selection, index test, reference standard, and flow and timing) and does not allow an overall score to be calculated. Riches et al. extended the QUADAS-2 tool by including the source of funding in the bias assessment. A summary of the assessment will be included in the systematic review.”

New: “The risk of bias in individual studies will be assess using the QUADAS-2 tool (3) modified after Riches.(4) QUADAS-2 was developed to assess the risk of bias in studies investigating diagnostic tests and is recommended by the National Institute for Health and Care Excellence (NICE) and the Agency for Healthcare Regulation and Quality (AHRQ). The tool assesses four different components of the study design independently (patient selection, index test, reference standard, and flow and timing) and does not allow an overall score to be calculated. Riches et al. extended the QUADAS-2 tool by including the source of funding in the bias assessment. The subsections and signalling questions from the ROBIN-I assessment tool (10) applicable to the included studies will also be used to complement the risk of bias assessment. This reflects the complex nature of the included studies, evaluating both the performance of a diagnostic test and of an intervention on physicians. A summary of the assessment will be included in the systematic review.”

Support

Outreach Librarian	Tatjana Petrinic (Bodleian Libraries, University of Oxford) supported the development of the search strategy and advised on the systematic review methodology
--------------------	--

Funding	No specific funding was provided for this systematic review.
---------	--

INTRODUCTION

Rationale

The last decade has seen an exponential growth in the number of computational tools using large sets of patient data routinely collected in healthcare settings to perform clinical decision tasks. Previously the prerogative of human physicians, these tasks range from tumour classification to outcome prediction, via radiological diagnostics and triage.

The vast majority of these computational tools are tested for efficiency on specifically designated test datasets or against humans as reference standards, but rarely for the benefit they can have when used as adjunct to clinicians' decision-making. It is unlikely that human physicians will disappear from the medical decision making process in the near future (5) and, as long as the responsibility and liability for patient care remains with them, the human perception of a problem and decision regarding the solution will be crucial factors influencing patient outcomes. Hence, it is important to understand the effects on human performance of this new generation of decision support systems using machine learning algorithms and based on patient data.

Computerized decision support systems are not new in medicine and have already been the subjects of numerous systematic reviews.(6)(7)(8)(4) However, the recent advance in computer sciences has opened the door to a new class of clinical algorithms, which, unlike their predecessors, are not building their recommendations on handcrafted knowledge bases but on their own interpretation of thousands if not millions of data points derived from agnostic clinical data. Whereas this novelty offers the opportunity of increased accuracy and relevance, it also introduces new obstacles related to the interpretability and reliability of the software's outputs. By design these algorithms have the potential to outperform their human operators so that the human contribution can become the limiting factor. The notion of trust and the need to understand how recommendations were produced play a crucial role in bridging the software outputs to actual effects on patient outcomes. Moreover, the usability of a system and its seamless integration into the clinical workflow are important considerations toward a broad deployment of this technology and translating its benefits into improved patient care.

Understanding the impact of specific software design's components, like the mathematical approach or the degree of support provided, on the human perception of the system abilities and access to appropriate metrics to evaluate the non-technical aspects of the human-computer interaction would be useful to orient the development and test of future decision support systems based on machine learning.

Objectives

The primary objective of this systematic review is to evaluate the effects of clinical decision support systems based on machine learning on physicians' performance, focusing on diagnosis or diagnostic investigations planning.

Secondary objectives are:

- To compare the performance of the human-computer interactions to the performance of the computer systems alone.
- To identify the evaluation metrics commonly used to evaluate human-computer interaction and performance in the context of medical diagnostic based on machine learning.
- To identify potential gaps in the assessment methodology of human-computer interactions in the context of medical diagnostic based on machine learning.
- To assess if particular strategies for decision support systems' design (mathematical approach, degree of support, timing of support, etc) are consistently associated with better physicians' performance.

METHODS

Eligibility criteria (all should be met)

Study types:	This systematic review will focus on primary research only. This can include, but is not limited to, randomized control trials, case-control trials, cohort studies, before and after studies as well as qualitative research. Case reports and case series will be excluded.
Years:	01.01.2010 to 31.05.2019
Language:	English literature only
Population:	Human medical doctors from all specialties and all levels of seniority, in both in- and outpatient settings, facing a clinical diagnostic decision having a direct impact on patient care. Medical students are not included in the study population.
Intervention:	Interactive use of a decision support system based on clinical data and machine learning algorithms to improve diagnosis or diagnostic investigations planning. In the context of this review, machine learning algorithms are defined as algorithms that have the ability to independently learn, from clinical data, knowledge unknown to their programmers and to generate outputs that have not been explicitly programmed. Machine learning models considered as general medical statistics, such as linear regression and logistic regression are not included. Diagnosis is defined as "the identification of the nature of an illness" (Oxford Dictionary).
Control:	Human medical doctors without the aforementioned decision support system. This includes studies where the same individuals had to perform a task with and without the decision support system.
Outcomes:	Any metrics assessing performance, usability, trust or other components of human-computer interaction.

- Exclusion criteria:
- Will be excluded studies:
- only comparing the outputs of an automated system against human performance without decision support as gold standard
 - describing monitoring or alert systems (including follow up monitoring)
 - describing decision support systems based on handcrafted knowledge or rules bases only (human expert knowledge)
 - describing decision support systems based on natural language processing only
 - describing decision support systems based on validated clinical scores only
 - describing systems uniquely designed to improve the quality of a signal
 - whose target patients are not human

Information sources

The search strategy mentioned in item 10 will be run in Embase (without conference abstracts), Medline and PsycINFO.

Grey literature search will include: The World Health Organization International Clinical Trials Registry Platform, conference abstracts (from 2017 onward), the Cochrane Central Register of Controlled Trials.

Web of Science will be used for forward and backward literature search from included studies.

Search strategy

The following search strategy was developed with the support of an experienced librarian (TP). The initial search has been run on 20.05.19 in MEDLINE and EMBASE and on 12.06.19 in PsycINFO using the Ovid interface. The search will be repeated towards the end of the review process to make sure late indexation are also considered.

As several clinical algorithms are referred to under their trade names in the literature, and might therefore escape our search strategy, trade names will be used in addition to generic search terms to enhance the retrieval where appropriate. These studies will be included as “other resources” in the PRISMA diagram.

- 1 *Decision Making, Computer-Assisted/
- 2 exp Diagnosis, Computer-Assisted/
- 3 *Therapy, Computer-Assisted/
- 4 Drug Therapy, Computer-Assisted/
- 5 exp Decision Support Systems, Clinical/

- 6 *Algorithms/
- 7 (CDSS* or CCDSS* or "decision support" or "decision making" or "diagnos* support" or "computer aided" or CAD* or "computer assisted" or "digital assistance" or algorithm*).ab,kw,ti.
- 8 1 or 2 or 3 or 4 or 5 or 6 or 7
- 9 exp Artificial Intelligence/
- 10 exp Latent Class Analysis/
- 11 exp Pattern Recognition, Automated/
- 12 ("artificial intelligence" or AI or "machine learning" or "deep learning" or "neural network" or "support vector machine" or "Bayesian network" or "nearest neighbour" or "decision tree" or "random forest" or "patient similarity" or "pattern recognition" or "natural language processing" or (supervised adj2 learning) or (unsupervised adj2 learning) or "reinforcement learning").ab,kw,ti.
- 13 9 or 10 or 11 or 12
- 14 (doctor* or residen* or physician* or clinician* or surgeon* or registrar* or "house officer*" or fellow* or medics or consultant* or attending or practitioner* or oncologist* or pathologist* or radiologist* or ophthalmologist* or neurologist* or cardiologist* or urologist* or gynecologist* or gastroenterologist* or pneumologist* or dermatologist* or endocrinologist* or psychiatrist* or pediatrician* or internist* or anesthesiologist* or orthopedist*).ab,kw,ti.
- 15 (safety or trust or usability or confidence or reliability or performance or outperform* or metrics or measure* or evaluat* or assess* or effective* or precision or recall or accuracy or "patient* outcome*" or "clinical outcome*" or "surgical outcome*" or "term outcome*" or mortality or morbidity or complication*).ab,kw,ti.
- 16 8 and 13 and 14 and 15
- 17 limit 16 to (editorial or letter or "review" or "systematic review")
- 18 16 not 17

Study records

- Data management: Deduplication will be carried out both automatically and manually using the EndNote software. The abstracts screening, study selection and data extraction will be performed with the Covidence systematic review online tool.(9)
- Selection process: Abstracts screening will be performed by at least two independent reviewers. The first reviewer will screen through all of the abstracts. Conflicts will be resolved by a third reviewer. Abstracts meeting the inclusion criteria or possibly meeting the inclusion criteria will be selected for full text review and pdf files will be uploaded to the systematic review library.
- Full text review and inclusion will be performed by at least two reviewers. The first reviewer will review all the selected publications. Conflicts will be resolved by discussion and a third reviewer will adjudicate any unresolved conflict.
- Data collection process: Data extraction and collection will be performed by at least two independent reviewers. Data will be collected using a standardised extraction sheet designed by the first reviewer and containing all the items mentioned in Item 12. Reviewers will attend a practical introduction to ensure consistency of the data collection. Conflicts will be resolved by discussion and a third reviewer will adjudicate any unresolved conflict.
- Given the expected high heterogeneity of measured outcomes, authors will not necessarily be contacted to obtain missing data.

Data items

The following data will be extracted if present:

- study population: number, specialty, seniority
- patient population: type of medical conditions, number of different hospital sites
- dataset: type of sample, sample size and number of events (for training and validation sets), independence of training and test sets
- experiment: task to be performed, experimental design, number of cases per physician, timing of support, gold standard comparison, familiarity with the system.
- main purpose of the decision support system
- system characteristics: mathematical model used, International Medical Device Regulators Forum (IMDRF) risk classification, type of support, , attempts to increase the interpretability of the model
- metrics of human performance: type and value of all the metrics used to assess human performance with and without the support system, including, but not limited to, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), positive and negative predictive values, precision, accuracy, recall, position, time to decision, inter- and intra-operator variability, usability, clinical outcomes (mortality,

morbidity, adverse events) and institutional outputs (average cost of treatment, length of stay)

- metrics of computer performance: type and value of all the metrics used to assess the performance of the decision support system alone, including, but not limited to, sensitivity, specificity, AUC, positive and negative predictive values, precision, accuracy, recall, position and time to decision.
- study funding: provenance
- existence of a published study protocol

Outcomes and prioritization

The main outcome is the physicians' performance with and without the described decision support systems. We expect the metrics used to quantify performance to vary depending on the main purpose of the decision support system described. These metrics include, but are not limited to, sensitivity, specificity, AUC, positive and negative predictive values, precision, accuracy, recall, position, time to decision, inter- and intra-operator variability, usability, clinical outcomes (mortality, morbidity, adverse events) and institutional outputs (average cost of treatment, length of stay). Qualitative performance assessment will also be considered.

The additional outcomes are:

- the performance of the computer system alone. We expect the metrics used to quantify performance to vary depending on the main purpose of the decision support system. The metrics include, but are not limited to, sensitivity, specificity, AUC, positive and negative predictive values, precision, accuracy, recall, position and time to decision.
- the qualitative and quantitative evaluation of the decision support system by the human operators. We expect that only few studies address this point and the metrics used for the evaluation to be heterogeneous.

Risk of bias in individual studies

The risk of bias in individual studies will be assessed using the QUADAS-2 tool (3) modified after Riches.(4) QUADAS-2 was developed to assess the risk of bias in studies investigating diagnostic tests and is recommended by the National Institute for Health and Care Excellence (NICE) and the Agency for Healthcare Regulation and Quality (AHRQ).

The tool assesses four different components of the study design independently (patient selection, index test, reference standard, and flow and timing) and does not allow an overall score to be calculated. Riches et al. extended the QUADAS-2 tool by including the source of funding in the bias assessment.

The subsections and signalling questions from the ROBINS-I assessment tool (10) applicable to the included studies will also be used to complement the risk of bias assessment. This reflects the complex nature of the included studies, evaluating both the performance of a diagnostic test and of an intervention on physicians.

A summary of the assessment will be included in the systematic review.

Data synthesis

Given the expected heterogeneity of the systematic review's target studies, the authors do not plan a meta-analysis at the time of registering this protocol.

A narrative synthesis of the reported outcomes in line with the systematic review's objectives will be performed, including differences in performance between the intervention and control groups as well as between the intervention group and the computer system alone. Underlying factors possibly explaining changes in effect size or direction will be investigated. The authors expect a noticeable variability in the metrics used to assess performance. A summary table of these metrics will be presented. Qualitative data will be presented descriptively as recommended in the PRISMA elaboration and explanation document.

Subgroups analysis will be performed according to the mathematical model used, the degree of support and the physicians' level of seniority. Any other coherent groups emerging from the included studies could also be subject to a subgroup analysis. If a subgroup is sufficiently homogenous in terms of study population and performance metrics, a quantitative synthesis of the performance metrics will be considered. The minimal number of studies required for this synthesis will depend on the number of participants in each study.

Meta-bias

The Clinical Trial Register at the International Clinical Trials Registry Platform of the World Health Organisation will be searched to look for unpublished trials (publication bias) or partial reporting of outcomes (outcome reporting bias). Due to the expected heterogeneity of the systematic review's target studies, the authors do not plan to perform funnel plots.

The overall provenance of funding will also be considered in the assessment of meta-bias.

Confidence in cumulative evidence

If quantitative summary statistics are performed, the confidence in cumulative evidence will be assessed according to the Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology.(11)

REFERENCES

1. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* [Internet]. 2015;4(1):1. Available from: <https://doi.org/10.1186/2046-4053-4-1>
2. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ Br Med J* [Internet]. 2015 Jan 2;349:g7647. Available from: <http://www.bmj.com/content/349/bmj.g7647.abstract>
3. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011 Oct;155(8):529–36.
4. Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The Effectiveness of Electronic Differential Diagnoses (DDX) Generators: A Systematic Review and Meta-Analysis. *PLoS One* [Internet]. 2016;11(3):1–26. Available from: <https://doi.org/10.1371/journal.pone.0148991>
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* [Internet]. 2019;25(1):44–56. Available from: <https://doi.org/10.1038/s41591-018-0300-7>
6. Jaspers MWM, Smeulders M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *J Am Med Inform Assoc*. 2011 May;18(3):327–34.
7. AX G, NJ A, McDonald H, al et. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *JAMA* [Internet]. 2005 Mar 9;293(10):1223–38. Available from: <http://dx.doi.org/10.1001/jama.293.10.1223>
8. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. *Ann Intern Med*. 2012 Jul;157(1):29–43.
9. Veritas Health Innovation. Covidence systematic review software. Melbourne, Australia;
10. Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* [Internet]. 2016 Oct 12;355:i4919. Available from: <http://www.bmj.com/content/355/bmj.i4919.abstract>
11. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* [Internet]. 2008 Apr 24;336(7650):924 LP-926. Available from: <http://www.bmj.com/content/336/7650/924.abstract>