Simple presentation of test accuracy may lead to inflated disease probabilities (letter)
Bachmann, L.M.; Steurer, J.; ter Riet, G.

Published in:
BMJ : British medical journal

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Indiscriminate investigations have adverse effects

Julian H Barth and Richard G Jones

*BMJ* 2003;326:393-
doi:10.1136/bmj.326.7385.393

Updated information and services can be found at:
http://bmj.com/cgi/content/full/326/7385/393

These include:

**References**

This article cites 5 articles, 2 of which can be accessed free at:
http://bmj.com/cgi/content/full/326/7385/393#BIBL

**Rapid responses**

3 rapid responses have been posted to this article, which you can access for free at:
http://bmj.com/cgi/content/full/326/7385/393#responses

You can respond to this article at:
http://bmj.com/cgi/eletter-submit/326/7385/393

**Email alerting service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article

**Notes**

To order reprints of this article go to:
http://www.bmjjournals.com/cgi/reprintform

To subscribe to *BMJ* go to:
http://bmj.bmjjournals.com/subscriptions/subscribe.shtml
Indiscriminate investigations have adverse effects

Editor—The application of evidence based medicine is leading to better treatments by thorough evaluation of treatments based on analyses of risks and benefits. These balance the beneficial clinical gains against the adverse pharmacological and medical effects, using information derived from randomised controlled trials and cost effectiveness studies. In contrast, no such critical approach has been taken for diagnostic tests nor have the consequences and adverse effects of inappropriate investigations been explored. The debate around diagnostic tests has centred largely on minimising the unit costs of the delivery of tests in the light of the enormous increase in the demand for investigations without an obvious and proportionate improvement in health status.

The case report by Krishnan et al highlights an adverse effect of an inappropriate investigation in a woman with hypothyroid induced ascites. The published literature is clear that ascites, and any serous effusion of any aetiology, is associated with raised CA125 concentration. Yet despite this evidence, the interpretation of a false positive result triggered a number of adverse effects and consequences—namely, a clinical positive result triggered a number of adverse consequences. Where the definition of success is in pharmacological studies, we do not change their probability estimates after previously, almost half of the doctors did not change their probability estimates after they were provided with the patient's age.

We also found that the non-technical format resulted in 25 of the 79 general practitioners in group 3 (32% (95% confidence interval 22% to 43%)) multiplying their pre-test probability by exactly 2. This is theoretically incorrect since, for example, a likelihood ratio of 2 changes a pre-test probability of 40% to 57% only, not to 80%, which requires a likelihood ratio of 6. Unfortunately, in our study, this mistake helped those respondents who did not change their pre-test probability after being given the patient's age to get close to the correct value.

**Simple presentation of test accuracy may lead to inflated disease probabilities**

Editor—We found that conveying information on the accuracy of tests in non-technical language improved doctors' ability to estimate disease probabilities accurately. We investigated whether doctors might mis-use such non-technical presentation when considering the probability of endometrial cancer in a patient with positive results on transvaginal ultrasonography. We presented 263 general practitioners in Switzerland with a pre-test probability of 10%, information that the patient was aged 65, and a positive transvaginal ultrasound result. Ninety two general practitioners (group 1) received no information on the test's accuracy; 92 (group 2) were told that the sensitivity of the test was 80% and specificity 60%; and 79 (group 3) were told that a positive result is obtained twice as frequently in women with endometrial cancer as in those without the disease, reflecting a likelihood ratio of 2. The last two statements are numerically equivalent since the likelihood ratio equals sensitivity/(1–specificity). The table shows that the degree of over-estimation of diagnostic accuracy varied with the presentation format. As we found previously, almost half of the doctors did not change their probability estimates after they were provided with the patient's age.

We also found that the non-technical format resulted in 25 of the 79 general practitioners in group 3 (32% (95% confidence interval 22% to 43%)) multiplying their pre-test probability by exactly 2. This is theoretically incorrect since, for example, a likelihood ratio of 2 changes a pre-test probability of 40% to 57% only, not to 80%, which requires a likelihood ratio of 6. Unfortunately, in our study, this mistake helped those respondents who did not change their pre-test probability after being given the patient's age to get close to the correct value.

**Distributions of attributed likelihood ratios in three groups given different summaries of information on diagnostic accuracy**

<table>
<thead>
<tr>
<th>Group</th>
<th>Median attributed likelihood ratio (25th centile, 75th centile)</th>
<th>Comparison between groups</th>
<th>P value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (n=92)</td>
<td>9 (3.69)</td>
<td>1 v 2</td>
<td>0.0193</td>
</tr>
<tr>
<td>2 (n=92)</td>
<td>6 (2.22)</td>
<td>1 v 3</td>
<td>0.0003</td>
</tr>
<tr>
<td>3 (n=79)</td>
<td>3 (2.9)</td>
<td>2 v 3</td>
<td>0.2934</td>
</tr>
<tr>
<td></td>
<td>1 v (2 v 3)</td>
<td></td>
<td>0.0006</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
<td>0.0013</td>
</tr>
<tr>
<td>stricter analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (n=92)</td>
<td>9 (3.69)</td>
<td>1 v 2</td>
<td>0.0193</td>
</tr>
<tr>
<td>2 (n=92)</td>
<td>6 (2.22)</td>
<td>1 v 3</td>
<td>0.1636</td>
</tr>
<tr>
<td>3 (n=54)</td>
<td>3 (1.17)</td>
<td>2 v 3</td>
<td>0.5682</td>
</tr>
<tr>
<td></td>
<td>1 v (2 v 3)</td>
<td></td>
<td>0.0216</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
<td>0.0599</td>
</tr>
</tbody>
</table>

*Kruskal-Wallis test.

Formula to convert pre-test probability (P1) into post-test probability (P2):

Pre-test odds=likelihood ratio/post-test odds, where pre-test odds=P1/(1–P1) and P2=post-test odds/(1+post-test odds).

Group 1 received no information on the test's accuracy; group 2 were told that the sensitivity of the test was 80% and specificity 60%; group 3 were told that a positive result is obtained twice as frequently in women with endometrial cancer as those without the disease. Actual likelihood ratio associated with the test result was 2.25.
Effect of computerised evidence based guidelines

Computer support is complex intervention

Entror—Eccles et al’s rigorous approach to the evaluation of a computerised decision support system for the management of angina and asthma accounted for many of the flaws in previous trials of computer support. They were no doubt disappointed that no effect was seen, probably due to low usage of the system. Although not discussed in the paper, a possible explanation for this is that, given the comparatively high use of computers required for inclusion in the trial, the practices already used simpler computerised templates to promote collection of process data. Practitioners may therefore have perceived little further to be gained by using the more detailed decision support system, particularly if it did not allow easy switching between the guideline and the clinical system.

The study by Eccles et al shows the complexity of interventions in primary care that incorporate computerised decision support systems. This complexity needs to be fully accounted for in designing and evaluating such interventions. Even with an apparently well developed piece of software, the trial assumed that offering brief training to a minority of practitioners in each practice would be sufficient for it to be incorporated into the increasingly complex care provided in routine general practice consultations.

Trials of computer support in primary care need to acknowledge this complexity by embedding use of the software in a carefully specified model of care. For the high quality management of chronic disease, this model will probably require specialisation within a general practice, as proposed in the new general practitioner contract. Providing focused training to key people in a practice and supporting specialisation through computer decision support may be a more appropriate approach to chronic disease management in primary care. Future use of computer support must consider not only the technical features of the software but also the model of service it is supporting and hence the training requirements of potential users. Theoretically derived measures that predict use of the software by practitioners in these trials could provide further important data on the potential role of decision support in clinical practice. Only then can one truly give computer decision support a fair trial.

Jon D Emery Cancer Research UK clinician scientist Department of Public Health and Primary Care, University of Cambridge, Institute of Public Health, Cambridge CB2 2SR, UK. jde10@medschl.cam.ac.uk

Competing interests: None declared.

Challenge should not be abandoned

Entror—As a coauthor of the trial of COGENT, a clinical decision support system, I would like to correct any misunderstanding this paper may have caused. I hold a large centre for health informatics in the United Kingdom and lead the development of the Prodigy clinical decision support system. The COGENT trial of two computerised guidelines found no differences in a range of measures of the process and outcomes of care, primarily because the system was not used. But these findings should not be extrapolated to other decision support systems.

Readers to whom I have talked have assumed that COGENT guidance and software were based on the current Prodigy system. COGENT used evidence based guidelines from the north of England on the management of asthma and angina and software based on ideas from early Prodigy software. Constraints in the COGENT trial did not allow the software to be tested in practice before the intervention period or the guidance to be reworded for easier comprehension. Major shortcomings were soon apparent, but these problems could not be addressed because the trial method did not accommodate the usual process of software development and guidance formatting. With hindsight, a randomised controlled trial of a new technology (such as a clinical decision support system) should not be undertaken until the technology had been shown to be usable and to be regularly used.

Which way forward? In an increasingly complex world, clinicians overloaded with information need computerised decision support systems if their practice is to be evidence based. The challenge of developing and integrating such systems into clinical workflow should not be abandoned. Not to invest in such systems would be as inappropriate as suggesting that the British army should give up its rifles because of their current technical problems.

Ian N Purves professor of health informatics Sweizer Centre for Health Informatics at Newcastle, University of Newcastle upon Tyne, Newcastle upon Tyne NE2 4AB ian.purves@ncl.ac.uk

Competing interests: IAN is a grant holder, Prodigy contract (Department of Health).

It is good to be honest and say that systems were not used

Entror—The paper by Eccles et al possesses academic integrity, which is widely lacking in computing research. I was the main researcher for the first two phases of the Prodigy project and believe that this project has much to teach the Prodigy team. One of the first detailed reports I wrote on Prodigy in 1998 indicated that Prodigy was actually used very little, about seven times a week, and most of the time (88%) users requested to bypass the system (www.robinth2.free-online.co.uk/ virtualclassroom/chap13/report1.pdf). I am very heartened to see that this type of information is being disseminated rather than suppressed, as was the case with the report I produced.

Robin Beaumont independent health informatics consultant Faculty of Medical Informatics, Royal College of Surgeons of Edinburgh, Edinburgh EH8 9DW robin@iieg.net.co.uk

Competing interests: None declared.
criteria. It did not depend on patient specific information but entry of a more general Read code. It did not contain a reminder to initiate review of patient care or arrange follow up. How far treatment recommendations depended on the patient’s individual clinical review rather than issuing more generic recommendations for treatment is also unclear.

The Prodigy system, the intervention around which this study was based, is an electronic version of a paper guideline that is triggered by entry of a prespecified Read code. By making this the only way in which to enter the computerised guideline the investigators ensured a low level of use during the study. General practitioners are unlikely to continue to enter the same Read code at every consultation as it would mean that each participating patient would have multiple duplicate entries of the same Read code in their electronic record. By excluding any sort of reminder function in their system,2 the investigators have not accounted for a barrier in managing chronic diseases—registration, recall, and regular review of patients. Analysis of factors that operate in managing angina and asthma should have uncovered such barriers before the start of this study.

Other details about the use of the computerised guideline require clarification. What is the definitive number of patients randomised and followed up in each practice for each intervention? What is the number (percentage) of patients in whom the computer guideline went past the first screen? What is the number (percentage) for whom a complete record entry was made? The authors make no comment on the differential use of the electronic guidelines between the two computer suppliers. This study reinforces the fact that passive diffusion of guidelines, in electronic or paper format, is an ineffective way to implement best practice.1 Paying insufficient attention to how a computer interface operates has produced low levels of usage and made the evaluation less useful than it might have been.2 Future studies should take into account the different functions of computer based clinical decision support systems, rather than simply generate suggestions to alter prescribing practice.3

Author’s reply

Ennor—We agree that complex interventions should ideally be developed through an iterative process. Exceptions to this include evaluating a preformed intervention that would not otherwise be rigorously evaluated. This applied at the outset of our study, although our intervention drew heavily on the iterative development of Prodigy software. We conducted an integrated process evaluation to understand the results better. This will appear in the BMJ shortly.

The NHS has invested large amounts of money in information technology, sometimes for little or no benefit. The evaluation of information technology is complex and multifaceted, but a computerised decision support system can be evaluated as a health technology.4 The results of this evaluation may be an important element of software development, until someone comes up with better methods of producing unbiased estimates of effectiveness and efficiency we maintain that all health technologies should be considered evaluable in randomised controlled trials.

Important methodological issues exist about the timing and duration of such evaluations, and we agree with Purves that they should be performed on stable systems. Given the cyclical nature of software development and the self belief and enthusiasm of developers, such points must be specified and enacted to avoid self perpetuating iterative cycles of development with the constant promise of jam tomorrow.

Our description of the system that we evaluated is accurate, and none of the authors dissented from it up to the point of publication.

Data were collected from November 1997 to September 2000, with the intervention running during the last 15 months of this time period. The trial was paused for six months while the software team worked on improvements. The rates of presentation of patients we reported equated to opportunities for the system to be used between twice a day and every other day in some practices. By the start of the intervention period, Prodigy software had become available and was delivered to trial practices alongside the study software. Our feedback from practices indicated that at least some asked for the Prodigy software to be turned off. This echoes Beaumont’s letter and implies that increasing the number of guidelines offered may not be the remedy that Purves suggests.

Two correspondents identified the importance of the issue of training. Contrary to Purves’s letter, two people from each practice were invited to a one day training session and the software was installed within 10 weeks by the computer supplier of two thirds of the trial practices. For the second supplier this interval was almost double, owing to unforeseeable commercial considerations in the company. We acknowledged the importance of training while suggesting that what happened was representative of the real world of primary care. We still believe this to be true but supplement Emery’s and Purves’s call for better training in service settings. Fahey et al say that the low levels of use of the system were partly due to requiring the entry of a single Read code and lack of responsiveness to patient specific information. Initially the system could be triggered automatically by a range of specified Read codes in the patient record. It could also be triggered by a clinician entering Read codes selected by the practice and was therefore not a passive method of dissemination. But this was changed in response to requests from the study practices. The automatic triggering was removed and a customisable Read code entry method was used for the final eight months of the intervention. Thus the system did rely on patient specific information.

Emery said that we may have had a ceiling effect due to practices currently using computerised templates. This seems unlikely because only 26% of practices already had

2 Randolph A, Haynes RB, Wyatt JC, Cook DJ, Guyatt GH. Users’ guides to the medical literature. XVIII. How to use an article evaluating the clinical impact of a computer-based clinical decision support system. JAMA 1999;282:97-74.
Assortative mating may explain spouses' risk of same disease

I ntrinsically, Hippisley-Cox et al observed significant similarities for disease between spouses in a large sample of 8836 couples recruited through general practice. They think that shared environmental factors may cause these similarities but reject assortative mating as an explanation.

In a sample from the Netherlands twin register we could not replicate their spouse similarities for asthma, depression, diabetes, and cardiovascular disease, possibly because of our smaller sample size of 2122 spouse pairs. When we examined health behaviour in a larger sample we found good associations between spouses for smoking, alcohol problems, and exercise behaviour, even after controlling for age and body mass index of both spouses.

The duration of the relationship influenced these associations between spouses (figure). Except for alcohol problems, spouse similarities in health behaviour decreased as the duration of the relationship increased. This implies that assortative factors are based on similarity at the time dating began and highlights the importance of determining similarities in disease status at the time of dating, as suggested by Hippisley-Cox et al.

Novartis was not in breach of code for “inventing” disease

E ntror–Ferriman’s news item is incorrect on at least two counts.

Firstly, it is not true to state that the authority had imposed no penalty on the company for issuing misleading literature. Novartis, like all companies ruled in breach of the code in future.

Secondly, the impression from the heading to the article, as above, is misleading. Novartis was not ruled in breach of the code for inventing a disease. Novartis was ruled in breach of the code for giving a misleading impression of the effect of Starlix on cardiovascular mortality and risk as detailed in the main body of the article.
and myocardial ischaemia; if, as is said by Ikeda et al, the hypertension and impaired consciousness are assumed to be related to a brain lesion, the delays in obtaining imaging investigations could lead to delays in instituting potentially life-saving treatment.

The possibility of illicit drug ingestion should be considered in any young, hypertensive patient presenting to an emergency department with reduced consciousness, so that appropriate management can be started without delay.

Kim Whelan  registrar in toxicology
kim.whelan@gst.thames.nhs.uk

Alison Jones  consultant physician
Paul Morgan  registrar in toxicology
National Poisons Information Service, Guy's and St Thomas' NHS Trust, London SE1 5ER


Unit of analysis errors should be clarified in meta-analyses

Entro—Weingarten et al present a comprehensive study in what is a complex area of research. We were, however, unclear whether any of the included primary studies had unit of analysis errors and how the authors dealt with such studies in their meta-analysis.

Unit of analysis errors occur in cluster randomised trials when individual patients’ data are analysed as if there was no clustering in the provider, practice, or units randomised to the intervention groups (patients’ data are analysed as independent observations). Standard statistical methods that do not account for cluster effects in cluster randomised trial data result in the overestimation of the significance of an intervention (artificially extreme P values and overly narrow confidence intervals). Correspondingly, the inclusion of studies with unit of analysis errors in a meta-analysis will give greater weight to the results of such studies.

The table of included studies reported by Weingarten et al indicated that the unit of analysis differed from the unit of randomisation in 22 cluster randomised trials, but it was not clear from the report how often unit of analysis errors occurred in these studies or how the authors dealt with studies with such errors in the meta-analysis. Methods exist for re-analysing studies with such errors.

We recently completed a systematic review of guideline dissemination and implementation strategies; 51 out of 110 cluster randomised trials had unit of analysis errors, and reanalysis was possible in only one study. Poor reporting of cluster randomised trials has led to a proposed extension to the CONSORT statement, which is currently under discussion. Systematic reviews of studies with unit of analysis errors should clearly state how they handled such studies in a review.

Ruth E Thomas  research fellow
r.e.thomas@abdn.ac.uk

Craig Ramsay  senior statistician
Health Services Research Unit, University of Aberdeen, Aberdeen AB25 2ZD

Laura McAuley  review group coordinator
Cochrane Effective Practice and Organisation of Care Group, Cochrane Effective Practice and Organisation of Care Group, Institute of Population Health, University of Ottawa, 1 Stewart Street, Ottawa, ON, Canada K1N 6N3

Jeremy M Grimshaw  director, clinical epidemiology programme
Institute of Population Health, University of Ottawa, Ottawa Health Research Institute, 1053 Carling Avenue, Ottawa, ON, Canada K1Y 4E9


GPs can separate oncological wheat from chaff

Entro—Summerton’s editorial on identifying symptoms potentially indicating an underlying cancer in primary care calls for research based theory. Only community based studies will help general practitioners to decide on the importance of a symptom or physical sign reported by their patients as hospital series are unrepresentative.

Although selection bias is part of the problem, it may be comparatively minor as most patients with cancer are treated in secondary care, even though their disease is diagnosed in primary care. Perhaps more important is bias from the development of symptoms in the time from the first symptom appearing to presentation in primary care and finally hospital.

General practitioners are already quite successful in sifting out the wheat from the chaff (although the comparative rarity of cancer in primary care indicates that a better farming analogy would be finding a needle in the haystack). For example, an estimated positive predictive value for colorectal cancer of rectal bleeding in the community is 0.1%.

1 Summerton N. Symptoms of possible oncological significance: separating the wheat from the chaff. BMJ 2002;325:254-5. (50 November.)


Look before you flush

Entro—Moayyedi and Ford described recent advances in gastroenterology. The national programme for early detection of colon cancer uses the following statement to raise public awareness for early detection of rectal bleeding: “Look at it before you flush it.” It is really helpful to make people aware of looking for early signs of colon cancer, but I have noticed some important things that might hinder this national programme.

The lavatory disinfectants now sold in supermarkets are mostly blue in colour and change water blue, which makes looking for any blood quite difficult. I suggest that we stop selling colouring agents and replace them by colourless ones or even use reagents that turn a certain colour in the presence of minor blood amounts. Can we? MOURAD IBRAHIM HABIB  clinical research fellow
St James’s University Hospitals, Leeds LS9 7 TF mouradhabib@hotmail.com

Correction

Open letter to Tony Blair: Call to prevent escalating violence

An editorial error occurred in this open letter to Tony Blair (p 220, 25 January). By adding “the” to the authorship line we implied that the letter had been signed by all staff, students, and alumni of the London School of Hygiene and Tropical Medicine. The authorship line should have read: “On behalf of 500 staff, students, and alumni of the London School of Hygiene and Tropical Medicine. The authorship line should have read: “On behalf of the staff, students, and alumni…” as published.”