Missing clinical trial data: setting the record straight

Urgent action is needed to restore the integrity of the medical evidence base

RESEARCH, p 816
ANALYSIS, pp 809, 811

Elizabeth Loder associate editor, *BMJ*, London WC1H 9JR
Fiona Godlee editor, *BMJ*, London WC1H 9JR
fgodlee@bmj.com

Competing interests: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.
Provenance and peer review: Commissioned; not externally peer reviewed.

Cite this as: *BMJ* 2010;341:c5641
doi: 10.1136/bmj.c5641

Like us, you have probably grown accustomed to the steady stream of revelations about incomplete or suppressed information from clinical trials of drugs and medical devices.[1] If so, this issue of the *BMJ* features a pair of papers that will dismay but not surprise you. Researchers for an official German drug assessment body charged with synthesising evidence on the antidepressant reboxetine encountered serious obstacles when they tried to get unpublished clinical trial information from the drug company that held the data, an experience from which they draw several lessons.[2]

Once they were able to integrate the astounding 74% of patient data that had previously been unpublished, their conclusion was damning: reboxetine is "overall an ineffective and potentially harmful antidepressant".[3] This conclusion starkly contradicts the findings of other recent systematic reviews and meta-analyses published by reputable journals.[4-8] These studies presumably met prevailing standards for the conduct of meta-analyses. Yet we now know that they did not provide a properly balanced view of the harms and benefits of reboxetine. Why? Because they did not combine all of the existing evidence from clinical trials. Furthermore, the difficulties encountered by Wieseler and colleagues in obtaining the reboxetine data show that routine inquiries about missing information, which many authors of meta-analyses make, are probably insufficient.[9] Instead dogged, even heroic, persistence is required, as the Cochrane reviewers trying to untangle the evidence for oseltamivir have found.[10 11]

Research that is conducted but not reported is only part of the problem. Steinbrook and Kassirer point to the rosiglitazone (Avandia) story as an example of problems arising from incomplete access of researchers and others to the raw data within a trial.[12] Problems also arise, they say, with the way in which these data are interpreted or adjudicated. They call for journals and editors to do more, including reserving the right to inspect trial data themselves. This is a contentious topic. Commentator Chris Del Mar applauds this stand,[13] but Nick Freemantle points out that although it is easy to call for unfettered access to data, it is another thing entirely to provide and make use of it.[14]

The reboxetine story and similar episodes must call into question the entire evidence synthesis enterprise. Meta-analyses are generally considered the best form of evidence, but is that a plausible world view any longer when so many of them are likely to be missing relevant information?[15] Existing estimates of treatment benefits are not always altered when previously unpublished clinical trial data become available. At present, however, we do not know the extent to which integration of missing data would support or refute key portions of the existing evidence on which doctors, patients, and policy makers rely.

As Wieseler and colleagues point out, the Food and Drug Administration Amendments Act of 2007 and parallel European efforts will increase the accessibility of clinical trial results and make it more difficult to conceal information.[2] But they do not solve the problem of our current evidence base, which contains incomplete and questionable evidence. So what can be done? At the moment there are no organised efforts to identify missing information and integrate it into the existing evidence base.

The *BMJ* has a particular interest in the impact of unpublished data on the overall verdict regarding the effectiveness of medical treatment. Because we think that it is important to re-evaluate the integrity of the existing base of research evidence, the *BMJ* will devote a special theme issue to this topic in late 2011. A detailed call for papers will follow, but we mention this now because we hope that researchers with such projects under way will feel encouraged. We also hope that other potential authors might begin now to plan suitable projects.

We are especially interested in high quality original research that aims to uncover previously unavailable data and re-evaluate treatments and practice in light of that new evidence. The ideal way to summarise the findings would be a formal meta-analysis, showing how the newly identified information affects the balance of benefit to harm. It is not necessary to conclude that full consideration of all of the evidence in fact changes practice—we will also be interested in papers that conclude that, even with new evidence, nothing should change.

Lost in the sometimes rancorous debate over research transparency, and the reasons for publication and non-publication, is the most important thing: efforts are needed to restore trust in existing evidence. To that end, the *BMJ* is more interested in constructive use of data than finger pointing or blame. We encourage drug companies and device manufacturers, as well as academic researchers, to take advantage of the opportunity afforded by our upcoming theme issue. Full information about previously conducted clinical trials involving drugs, devices, and other treatments is vital to clinical decision making.

It is time to demonstrate a shared commitment to setting the record straight.

1 Cohen D. Rosiglitazone: what went wrong? *BMJ* 2010;341:c4848.
2 Wieseler B, McGauran N, Kaiser T. Finding studies on reboxetine: a tale of hide and seek. *BMJ* 2010;341:c4942.
3 Eyding D, Lelgemann M, Grouven U, Härter M, Kromp M, Kaiser T, et al. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *BMJ* 2010;341:c4737.
4 Ferguson JM, Mendels J, Schwart GE. Effects of reboxetine on Hamilton Depression Rating Scale factors from randomized, placebo-controlled trials in major depression. *Int Clin Psychopharmacol* 2002;17:45-51.
5 Montgomery S, Ferguson JM, Schwartz GE. The antidepressant efficacy of reboxetine in patients with severe depression. *J Clin Psychopharmacol* 2003;23:45-50.
6 Papakostas GI, Nelson JC, Kasper S, Möller HJ. A meta-analysis of clinical trials comparing reboxetine, a norepinephrine reuptake inhibitor, with selective serotonin reuptake inhibitors for the treatment of major depressive disorder. *Eur Neuropsychopharmacol* 2008;18:122-7.
7 Chuluunkhuu G, Nakahara N, Yanagisawa S, Kamae I. The efficacy of reboxetine as an antidepressant, a meta-analysis of both continuous (mean HAM-D score) and dichotomous (response rate) outcomes. *Kobe J Med Sci* 2008;54:E147-58.
8 Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009;373:746-58.
9 Mullan RJ, Flynn DN, Carlberg B, Tleyjeh IM, Kamath CC, LaBella ML, et al. Systematic reviewers commonly contact study authors but do so with limited rigor. *J Clin Epidemiol* 2009;62:138-42.
10 Jefferson T, Jones M, Doshi P, Del Mar C, Dooley L, Foxlee R. Neuraminidase inhibitors for preventing and treating influenza in healthy adults. *Cochrane Database Syst Rev* 2010;2:CD001265.
11 Doshi P. Neuraminidase inhibitors—the story behind the Cochrane review. *BMJ* 2009;339:b5164.
12 Steinbrook R, Kassirer JP. Data availability for industry sponsored trials: what should medical journals require? *BMJ* 2010;341:c5391.
13 Del Mar C. Commentary: but what should journals actually do to keep industry sponsored research unbiased? *BMJ* 2010;341:c5406.
14 Freemantle N. Commentary: journals must facilitate the dissemination and scrutiny of clinical research. *BMJ* 2010;341:c5397.
15 Guyatt GH, Sackett DL, Sinclair JC, Haywood R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX A method for grading health care recommendations. *JAMA* 1995;274:1800-4.

Apgar score and risk of cerebral palsy

Low scores are strongly associated with cerebral palsy and its subtypes

LIBRARY OF CONGRESS/SPL

RESEARCH, p 817

Nigel Paneth university distinguished professor, Departments of Epidemiology and Pediatrics and Human Development, College of Human Medicine, Michigan State University, East Lansing, MI 48824, USA
paneth@msu.edu

Competing interests: None declared.
Provenance and peer review: Commissioned; not externally peer reviewed.

Cite this as: *BMJ* 2010;340:c5175
doi: 10.1136/bmj.c5175

The first large study of Apgar score at birth and the risk of cerebral palsy was the US National Collaborative Perinatal Project. It found that most cases of cerebral palsy occurred in children with normal Apgar scores; that the risk of cerebral palsy was strongly related to an Apgar score lower than 4 in normal weight infants, especially when the low score was prolonged; and that low Apgar scores were less predictive of cerebral palsy in low birthweight infants.[1]

In the linked study, Lie and colleagues confirm—in a Norwegian dataset 10 times larger (540 000 children)—all of these findings but add the new observation that the risk of cerebral palsy by Apgar score depends on the type of cerebral palsy. Hemiplegia was 10 times more common in babies with Apgar score less than 4, diplegia 22 times more common, but quadriplegia 137 times more common.[2]

Although the nature of the insult or insults in cerebral palsy is still unclear, the Norwegian study suggests that for hemiplegia, and perhaps diplegia, the causative exposure might have occurred earlier in gestation than in quadriplegia, where the acute effects of the insult might still be manifesting at the time of birth. Unfortunately, evidence about when a brain damaging event takes place in pregnancy is limited, although neuroradiologists are making progress in this area.[3]

Experience suggests that simple procedures that can be performed widely have a greater impact on health than more complex and demanding procedures that are less widely applied. The Apgar scoring system works because it comprises just a few components that can easily be memorised, and requires no equipment and modest training. Each component of the five part scale scores 0, 1, or 2, providing a score range from 0 to 10. The Apgar score is used in delivery rooms around the world, and is even recorded on birth certificates in many countries,[4] making it the only general clinical assessment recorded on entire populations, and potentially available for linkage to medical records.

The anaesthesiologist Virginia Apgar created her system for scoring a baby's condition at birth in 1953 to ensure that doctors and nurses in the delivery room would look at the baby.[5] She recognised that the intense focus on the mother could at times lead to neglect of the infant in the crucial first minutes after birth and that if systematic assessment of babies at birth was to become routine, an assessment tool was needed that was simple enough to compete with the hectic environment of the delivery room.

The Apgar score is, as Lie and colleagues highlight, a measure of the elusive quality, "vitality." But where does that vitality comes from? The five components of the score—colour, heart rate, respiration, reflex irritability, and muscle tone—seem to be weighted towards the cardiorespiratory system. But a closer look shows that the Apgar score is also a neurological examination. Tone and reflex irritability are measures of the intactness of the nervous system, and respiration, and therefore colour, depend on the central drive to breathe.

A low Apgar score, especially when it persists beyond the first minute of life, is therefore an indicator of central nervous system depression, so it is not surprising that it can predict later neurological dysfunction. Although most infants with low Apgar scores recover quickly, in a small fraction these neurological abnormalities persist and even worsen to result in neonatal encephalopathy—a syndrome of altered levels of consciousness; abnormal tone; diminished movement; and, in severe cases, seizures, hypoventilation, and abnormal primitive reflexes.[6] All, or nearly all, children with low Apgar scores who develop cerebral palsy have probably also experienced persistently abnormal neurological findings in the first days of life.[7]

The leap from observing signs of neurological depression at birth to assuming that asphyxia occurred at or around birth must be resisted. Objective measures of impaired gas exchange around the time of birth correlate modestly with a depressed Apgar score at five minutes,[8] and they also correlate poorly with later neurological and cognitive outcomes, as Apgar and colleagues were among the first to show.[9]

What is the lesson for clinical practice? Although complex and expensive technologies are becoming increasingly available, they are no replacement for skills in clinical observation. A low Apgar score (<4) at five minutes in a baby of normal weight is an important clue that the baby has an increased risk of death and disability, even though most infants with such scores recover quickly and do well. Such babies must be watched closely for the persistence or development of encephalopathic signs, especially in the light of robust evidence that babies with encephalopathy may benefit from head or body cooling.[10]

1 Nelson KB, Ellenberg JH. Apgar scores as predictors of chronic neurologic disability. *Pediatrics* 1981;68:36-44.
2 Lie KK, Grøholt E-K, Eskild A. Association of cerebral palsy with Apgar score in low and normal birthweight infants: population based cohort study. *BMJ* 2010;341:c4990.
3 Bax M, Tydeman C, Flodmark O. Clinical and MRI correlates of cerebral palsy: the European Cerebral Palsy Study. *JAMA* 2006;296:1602-8.
4 Hamilton BE, Martin JA, Ventura SJ. Births: preliminary data for 2007. National vital statistics reports. National Center for Health Statistics. 2009. www.cdc.gov/nchs/data/nvsr/nvsr57/nvsr57_12.pdf.
5 Apgar V. A proposal for a new method of evaluation of the newborn infant. *Curr Res Anesth Analg* 1953;32:260-7.
6 Fenichel JM. Hypoxic-ischaemic encephalopathy in the newborn. *Arch Neurol* 1983; 40: 261–6.
7 Ellenberg JH, Nelson KB. Cluster of perinatal events identifying infants at high risk for death or disability. *J Pediatr* 1988;113:546-52.
8 Victory R, Penava D, da Silva O, Natale R, Richardson B. Umbilical cord pH and base excess values in relation to adverse outcome events for infants delivering at term. *Am J Obstet Gynecol* 2004;191:2021-8.
9 Apgar V, Girdany BR, McIntosh R, Taylor HC Jr. Neonatal anoxia. I. A study of the relation of oxygenation at birth to intellectual development. *Pediatrics* 1955;15:653-62.
10 Edwards AD, Brocklehurst P, Gunn AJ, Halliday H, Juszczak E, Levene M, et al. Neurological outcomes at 18 months of age after moderate hypothermia for perinatal hypoxic ischaemic encephalopathy: synthesis and meta-analysis of trial data. *BMJ* 2010;340:c363.

Variation in caesarean delivery rates

Specific risk groups should be monitored at a local level

RESEARCH, p 818

Marian Knight senior clinical research fellow, National Perinatal Epidemiology Unit, University of Oxford, Oxford OX3 7LF, UK
marian.knight@npeu.ox.ac.uk
Elizabeth A Sullivan associate professor, Perinatal and Reproductive Epidemiology Unit, School of Women's and Children's Health, University of New South Wales, Sydney, NSW, Australia

Competing interests: None declared.
Provenance and peer review: Commissioned; not externally peer reviewed.
Cite this as: *BMJ* 2010;341:c5255
doi: 10.1136/bmj.c5255

Rising rates of delivery by caesarean section are a cause of concern worldwide. Wide variation has been noted between countries—for example, caesarean delivery rates are 15% in the Netherlands but 38% in Italy.[1] More than twofold differences in primary caesarean delivery rates have also been reported across regions in Canada,[2] and between hospital delivery units in the United States and Australia.[3 4] Although there is no consensus concerning the optimal caesarean delivery rate, it is clear that poor access to emergency obstetric care, and hence poor access to caesarean delivery, can harm both mother and infant.[5] Conversely, high rates of operative delivery may result in poorer maternal and infant outcomes for the current or subsequent births.[6 7] Variations in caesarean delivery rates have been attributed to differences in the characteristics of women giving birth. In the linked study, Bragg and colleagues assess whether the variation in unadjusted caesarean section rates between NHS trusts in England can be explained by maternal characteristics and clinical risk factors.[8]

Previous studies have shown that women are more likely to be delivered by caesarean section if they are in their first pregnancy; older; have previously delivered by caesarean section; have a breech presentation; deliver preterm; or have other complications of pregnancy or medical problems, including diabetes, hypertensive disorders of pregnancy, or obesity.[9 10] However, comparisons of rates of caesarean delivery often fail to take these factors into account, with rates not adjusted for population differences in these characteristics. Bragg and colleagues show significant variation in the rates of caesarean delivery among NHS trusts in England, after adjustment for several of these factors.[8] Although they were unable to investigate some potential explanations—including maternal obesity, indication for caesarean section, gestational age at delivery, and models of care—persisting differences in these

BSIP, BOUCHARLAT/SPL

factors are unlikely to account for the greater than twofold difference in caesarean delivery rates that they calculated—adjusted rates varied between 15% and 32% among the units investigated.

Other suggested reasons for variation in caesarean delivery rates include contrasting medico-legal environments, private compared with public healthcare systems, differences in delivery volumes, and differences in the training of junior obstetricians. These factors are unlikely to explain variation between units within the relatively uniform climate of NHS hospitals in England. Although maternal choice has also been cited as a potential reason for increases in rates of caesarean delivery, there is little evidence to suggest that

this accounts for much variation in caesarean rates between hospitals. Variation is most probably related to differences in thresholds for intervention at institutional and practitioner levels and variations in the preferred models of care.

This research indicates, at a minimum, the need for more informed surveillance of caesarean sections at a hospital, regional, and national level. Several approaches could achieve this. Perhaps the most straightforward approach is the use of the "standard primipara," whereby units collect specific data on a defined group of low risk women only. The "standard primipara" is a 20-34 year old woman, who is giving birth for the first time, free of obstetric and specific medical complications, and has a singleton term pregnancy with a non-small for gestational age infant in a cephalic presentation. Comparison of intervention rates in this group of women effectively controls for differences in population or case mix between units, and it has been used to show the impact of guidelines on intrapartum care.[11] An extension of this approach is to divide women into 10 population subgroups according to specific combinations of distinct characteristics: parity, multiple pregnancy, fetal presentation, type of labour onset, gestation, and previous caesarean delivery. This approach allows comparison of caesarean delivery rates within comparable population subgroups, but it also allows units to establish the contribution to the total caesarean delivery rate made by women in each cohort,[12] and hence to target approaches to reduce the total rate. Both of these approaches require specific data collection. Enhancement of routine data to improve the monitoring of obstetric care remains an option, but it is unlikely that compliance with evidence based practice can ever be monitored this way.

It is now 10 years since the national sentinel caesarean section audit in the UK, which examined practice in detail,[10] and yet wide variation still exists. Unwarranted variation in clinical practice has been cited as an indication of a poor quality service.

Bragg and colleagues' study provides the impetus for ongoing work to investigate and tackle the reasons for regional and subregional variations in caesarean section practice. As with the original audit, women and their families, clinicians, planners, policy makers, and hospitals would benefit from a more detailed examination of variations in caesarean delivery practice and the generation of the high quality evidence needed to inform practice guidelines. High quality population based observational studies can provide robust evidence where randomised controlled trials are not possible or unethical, and such studies should be encouraged. There is no place for poor guidelines based on poor evidence.

1 EURO-PERISTAT Project. European Perinatal Health Report, 2008. www.europeristat.com/bm.doc/european-perinatal-health-report.pdf.
2 Hanley GE, Janssen PA, Greyson D. Regional variation in the cesarean delivery and assisted vaginal delivery rates. *Obstet Gynecol* 2010;115:1201-8.
3 Clark SL, Belfort MA, Hankins GD, Meyers JA, Houser FM. Variation in the rates of operative delivery in the United States. *Am J Obstet Gynecol* 2007;196:526.e1-5.
4 Victorian Government, Department of Health. Victorian Maternity Services Performance Indicators. Complete set for 2008-9. 2010. www.health.vic.gov.au/maternitycare/matpeform-ind-0809.pdf.
5 Paxton A, Maine D, Freedman L, Fry D, Lobis S. The evidence for emergency obstetric care. *Int J Gynaecol Obstet* 2005;88:181-93.
6 Knight M, Kurinczuk JJ, Spark P, Brocklehurst P. Cesarean delivery and peripartum hysterectomy. *Obstet Gynecol* 2008;111:97-105.
7 Betran AP, Merialdi M, Lauer JA, Bing-Shun W, Thomas J, Van Look P, et al. Rates of caesarean section: analysis of global, regional and national estimates. *Paediatr Perinat Epidemiol* 2007;21:98-113.
8 Bragg F, Cromwell DA, Edozien LC, Gurol-Urganci I, Mahmood TA, Templeton A, et al. Variation in rates of caesarean section among English NHS trusts after accounting for maternal and clinical risk: cross sectional study. *BMJ* 2010;341:c5065.
9 Laws P, Sullivan EA. Australia's mothers and babies 2007. Perinatal statistics series no 23. AIHW National Perinatal Statistics Unit, 2009. www.aihw.gov.au/publications/per/per-48-10972/per-48-10972.pdf.
10 Royal College of Obstetrics and Gynaecology. Clinical Effectiveness Support Unit. The national sentinel caesarean section audit report. RCOG Press, 2001.
11 Alfirevic Z, Edwards G, Platt MJ. The impact of delivery suite guidelines on intrapartum care in "standard primigravida." *Eur J Obstet Gynecol Reprod Biol* 2004;115:28-31.
12 Brennan DJ, Robson MS, Murphy M, O'Herlihy C. Comparative analysis of international cesarean delivery rates using 10-group classification identifies significant variation in spontaneous labor. *Am J Obstet Gynecol* 2009;201:308.e1-8.

Are measures of patient satisfaction hopelessly flawed?

No, but they need further refinement

RESEARCH, p 820

Jeannie L Haggerty McGill research chair in family and community medicine, St Mary's Hospital Center, 3830 Lacombe Avenue, Montréal, QC, Canada H3T 1M5
jeannie.haggerty@mcgill.ca

Competing interests: None declared.
Provenance and peer review: Commissioned; not externally peer reviewed.

Cite this as: *BMJ* 2010;341:c4783
doi: 10.1136/bmj.c4783

Measures of patient satisfaction and the patient experience—as instituted in the UK Quality and Outcomes Framework for primary care—supply feedback that helps health professionals provide patient centred care; they also give insight about the interpersonal dimension of quality of care as a complement to the technical quality of care. In the linked study, Salisbury and colleagues explore whether responses to questions in patient surveys that claim to assess the performance of general practices or doctors reflect differences between the practices, the doctors, or the patients themselves.[1] The analysis separates the variance in patient satisfaction and patient experience into that attributed to differences between practices and those between doctors. The study found that when patients were asked a single question about how satisfied overall they were with their practice, only 4.6% of the variance in their satisfaction ratings was a result of differences between practices; the remaining variance resulted from differences between patients plus random error. In contrast, when asked to report on their experience with usual time they had to wait for an appointment, more than 20% of the variance in responses was a result of differences between practices. The authors conclude that for the purpose of discriminating performance between practices, it is better to ask patients to report on their experience rather than ask for satisfaction ratings.

Still, the observation that measures of patient satisfaction and patient experience vary widely even among patients with the same doctor or practice raises questions about their use for evaluating practice performance. How can we make sense of such variance between patients? Is the use of satisfaction ratings a hopelessly flawed approach to evaluating

practice performance? And what are the implications for both measurement and performance evaluation?

Firstly, the variance in satisfaction scores is not surprising given the multidimensional nature of health care and patient satisfaction. Although satisfaction is seen as a judgment about whether expectations were met, it is influenced by varying standards, different expectations, the patient's disposition, time since care, and previous experience.[2] None the less, qualitative research shows that patients will give positive satisfaction ratings even in the face of a negative experience unless they believe that the poor care is under the direct control of the person they are evaluating.[3 4] For example, they may be unhappy about hurried communication with their doctor but still give an adequate rating because they attribute this to time constraints not a lack of intrinsic skills. Consequently, positive satisfaction ratings include both true positives and false positives. This compromises sensitivity in a diagnostic test and by the same token reduces the precision of satisfaction ratings. In contrast, negative satisfaction ratings tend to be truly negative (or highly specific in the analogy of diagnostic accuracy) and reflect important incidents, such as a lack of respect or medical errors.[4 5] The implication is that the representation of satisfaction and satisfaction ratings needs to be changed. It is better to report the proportion of patients who are less than totally satisfied rather than the average satisfaction. High satisfaction ratings indicate that care is adequate not that it is of superior quality; low ratings indicate problems and should not be masked by reporting average scores.

Secondly, a defining characteristic of primary care is its high degree of variety and variance, even within the practice of one doctor.[6 7] On a technical note, it is important to remember that analytical modelling that separates the variance into practice, doctor, and patient levels cannot separate variance between patients from random error. Part of this random error comes from the variation within practices and within doctors, which is to be expected, given the complexity of primary care. It is not surprising that such complexity can be only partially captured by a short questionnaire about experience and satisfaction. Despite this, patient assessments of health care work surprising well. Salisbury and colleagues show that assessments of access explain more variance between practices than they do between doctors, which makes sense for an attribute related to organisational arrangements. Conversely, assessments of communication explain more variance between doctors than between practices. Other studies have also found that patient assessments appropriately detect more variance between practices for organisational attributes and between doctors for personal care attributes.[8 9] The implication is that the differences between practices and between doctors seen in the current analytical models underestimate the true differences that occur at the practice and doctor levels, and although Salisbury and colleagues are right in advocating prudence in interpreting small differences between practices, we can be confident that statistically significant differences are real and clinically relevant.

Thirdly, these results have implications for improving the science of measurement. Although it is difficult to measure patients' perceptions of health care, it is most appropriate that patients should assess the interpersonal dimension of quality of care because they are the ones to whom we are ultimately accountable. It is therefore crucial that patient surveys are refined to maximise precision and minimise bias. The research community needs to develop and refine robust and comparable measures, bearing in mind that deficiencies in the measurement of satisfaction are more common in newly devised instruments.[4]

Measures of patient satisfaction need to be refined, but they are not hopelessly flawed. When they detect problems, these are real and important. They should be presented in a way that highlights the informative negative assessments, and they need to be combined with reports (such as experience) of components that can be benchmarked to recognised best practices.

1 Salisbury C, Wallace M, Montgomery A. Patient experience and satisfaction in primary care: secondary analysis using multilevel modelling. *BMJ* 2010;341:c5004.
2 Crow R, Gage H, Hampson S, Hart J, Kimber A, Storey L, et al. The measurement of satisfaction with healthcare: implications for practice from a systematic review of the literature. *Health Technol Assess* 2002;6:1-244.
3 Schneider H, Palmer N. Getting to the truth? Researching user views of primary health care. *Health Policy Plan* 2002;17:32-41.
4 Collins K, O'Cathain A. The continuum of patient satisfaction—from satisfied to very satisfied. *Soc Sci Med* 2003;57:2465-70.
5 Taylor B, Marcantonio ER, Pagovich O, Carbo A, Bergmann M, Davis RB, et al. Do medical inpatients who report poor service quality experience more adverse events and medical errors? *Med Care* 2008;46:224-8.
6 Love T, Burton C. General practice as a complex system: a novel analysis of consultation data. *Fam Pract* 2005;22:347-52.
7 Katerndahl DA, Wood R, Jaén CR. A method for estimating relative complexity of ambulatory care. *Ann Fam Med* 2010;8:341-7.
8 Haggerty JL, Pineault R, Beaulieu M-D, Brunelle Y, Gauthier J, Goulet F, et al. Practice features associated with patient-reported accessibility, continuity and coordination of primary health care. *Ann Fam Med* 2008;6:116-23.
9 Rodriguez HP, Scoggins JF, von Glahn T, Zaslavsky AM, Safran DG. Attributing sources of variation in patients' experiences of ambulatory care. *Med Care* 2009;47:835-41.

Misleading communication of risk

Editors should enforce transparent reporting in abstracts

Cite this as: *BMJ* 2010;341:c4830
doi: 10.1136/bmj.c4830

In 1996 a review of mammography screening reported in its abstract a 24% reduction of breast cancer mortality[1]; a review in 2002 claimed a 21% reduction.[2] Accordingly, health pamphlets, websites, and invitations broadcast a 20% (or 25%) benefit.[3] Did the public know that this impressive number corresponds to a reduction from about five to four in every 1000 women, that is, 0.1%? The answer is, no. In a representative quota sample in nine European countries, 92% of about 5000 women overestimated the benefit 10-fold, 100-fold, and more, or they did not know.[4] For example, 27% of women in the United Kingdom believed that out of every 1000 women who were screened, 200 fewer would die of breast cancer. But it is not only patients who are misled. When asked what the

Gerd Gigerenzer director
sekgigerenzer@mpib-berlin.mpg.de
Odette Wegwarth research scientist
Markus Feufel postdoctoral fellow, Harding Center for Risk Literacy, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Competing interests: None declared.
Provenance and peer review: Commissioned; not externally peer reviewed.

"25% mortality reduction from breast cancer" means, 31% of 150 gynaecologists answered that for every 1000 women who were screened, 25 or 250 fewer would die.[3]

In 1995, the UK Committee on Safety of Medicines issued a warning that third generation oral contraceptive pills increased the risk of potentially life threatening thrombosis twofold. The news provoked great anxiety, and many women stopped taking the pill, which led to unwanted pregnancies and abortions––some 13 000 additional abortions in the next year in England and Wales—and an extra £46m (€55m; $71m) in costs for the NHS.[5] Yet how big was the twofold risk? The studies revealed that for every 7000 women who took the earlier, second generation pills, one had a thrombosis, and this number increased to two in women who took third generation pills. The problem of misleading reporting has not gone away. In 2009, the *BMJ* published two articles on oral contraceptives and thrombosis; one made the absolute numbers transparent in the abstract,[6] whereas the other reported that "oral contraceptives increased the risk of venous thrombosis fivefold."[7]


DAVID PAGE PHOTOGRAPHY/ALAMY

These two examples illustrate a general point. Absolute risks (reductions and increases), such as from one to two in 7000, are transparent, while relative risks such as "twofold" provide incomplete and misleading risk information.[3 8] Relative risks do not inform about the baseline risk—for example, whether twofold means from one to two or from 50 to 100 in 7000—and without this information, people overestimate benefits or harms.[3 9] In the case of the pill scare, the losers were women, particularly adolescent girls, taxpayers, and the drug industry. Reporting relative risks without baseline risk is practised not only by journalists because big numbers make better headlines or by health organisations because they increase screening participation rates. The source seems to be medical journals, from which figures spread to press releases, health pamphlets, and the media.

An analysis of the articles published in the *Annals of Internal Medicine*, *BMJ*, *JAMA*, *Journal of the National Cancer Institute*, *Lancet*, and the *New England Journal of Medicine*, 2003-4, showed that 68% (150/222) failed to report the underlying absolute risks in the abstract. Among those, about half did report the absolute risks elsewhere in the article, but the other half did not.[10] Similarly, an analysis of 119 systematic reviews in *BMJ*, *JAMA*, and *Lancet* from 2004 to 2006 showed that every second article discussed only relative risks or odds ratios.[11]

Conveying relative risks without baseline risk is the first "sin" against transparent reporting. The second is mismatched framing—reporting benefits, such as relative risk reductions, in big numbers and harms, such as absolute risk increases, in small numbers.[3] If we use the example of a treatment that reduces the probability of getting disease A from 10 to five in 1000, whereas it increases the risk of disease B from five to 10 in 1000, authors who use mismatched framing would report the benefit as a 50% risk reduction and the harm as an increase of five in 1000; that is, 0.5%. Medical journals permit mismatched framing. One in three articles in the *BMJ*, *JAMA*, and *Lancet* from 2004 to 2006 used mismatched framing when both benefits and harms were reported.[11]

Have editors since stopped non-transparent reporting? To check the current situation, we examined the abstracts of all free accessible research articles published in the *BMJ* in 2009 that reported drug interventions. Of the 37 articles identified, 16 failed to report the underlying absolute numbers for the reported relative risk measures in the abstract. Among these, 14 reported the absolute risks elsewhere in the article, but two did not report them anywhere. Moreover, absolute risks or the number needed to treat (NNT) were more often reported for harms (10/16) than for benefits (14/27).

How can those who are responsible for accurate communication of risk do better? And who should be monitoring them to ensure that they do? Steps can be taken to improve the transparency of risk communication.[12] Firstly, editors should enforce transparent reporting in journal abstracts: no mismatched framing, no relative risks without baseline risks, and always give absolute numbers such as absolute risks and NNT.

Secondly, institutions that subscribe to medical journals could give journal publishers two years to implement the first measure and, if publishers do not comply, cancel their subscriptions.

Thirdly, writers of guidelines, such as the CONSORT statement, should stipulate transparent reporting of benefit and harms in abstracts.

1 Larsson LG, Nyström L, Wall S, Rutqvist L, Andersson I, Bjurstam N, et al. The Swedish randomised mammography screening trials: analysis of their effect on the breast cancer related excess mortality. *J Med Screen* 1996;3:129-32.
2 Nyström L, Andersson I, Bjurstam N, Frisell J, Nordenskjöld B, Rutqvist LE. Long-term effects of mammography screening: Updated overview of the Swedish randomised trials. *Lancet* 2002;359:909-19.
3 Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients to make sense of health statistics. *Psychol Sci Public Interest* 2007;8:53-96.
4 Gigerenzer G, Mata J, Frank R. Public knowledge of benefits of breast and prostate cancer screening in Europe. *J Natl Cancer Inst* 2009;101:1216-20.
5 Furedi A. The public health implications of the 1995 "pill scare." *Hum Reprod Update* 1999;5:621-6.
6 Lidegaard Ø, Løkkegaard E, Svendsen AL, Agger C. Hormonal contraception and risk of venous thromboembolism: national follow-up study. *BMJ* 2009;339:b2890.
7 Van Hylckama Vlieg A, Helmerhorst FM, Vandenbroucke JP, Doggen CJM, Rosendaal FR. The venous thrombotic risk of oral contraceptives, effects of oestrogen dose and progestogen type: results of the MEGA case-control study. *BMJ* 2009;339:b2921.
8 Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. *BMJ* 2003;327:741-4.
9 Covey J. A meta-analysis of the effects of presenting treatment benefits in different formats. *Med Decis Making* 2007;27:638-54.
10 Schwartz LM, Woloshin S, Dvorin EL, Welch HG. Ratio measures in leading medical journals: structured review of accessibility of underlying absolute risks. *BMJ* 2006;333:1248-52.
11 Sedrakyan A, Shih C. Improving depiction of benefits and harms: analyses of studies of well-known therapeutics and review of high-impact medical journals. *Med Care* 2007;45:523-8.
12 Gigerenzer G, Gray JAM. Launching the century of the patient. In: Gigerenzer G, Gray JAM, eds. *Better doctors, better patients, better decisions: envisioning healthcare 2020*. MIT Press [forthcoming].