Uncertainty in heterogeneity estimates in meta-analyses

John Ioannidis, **Nikolaos Patsopoulos**, and **Evangelos Evangelou** argue that, although meta-analyses often measure heterogeneity between studies, these estimates can have large uncertainty, which must be taken into account when interpreting evidence

John P A Ioannidis professor
Nikolaos A Patsopoulos research associate
Evangelos Evangelou research associate, Clinical Trials and Evidence-Based Medicine Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece
Correspondence to: J P A Ioannidis jioannid@cc.uoi.gr
Accepted: 29 August 2007

An important aim of systematic reviews and meta-analyses is to assess the extent to which different studies give similar or dissimilar results.[1] Clinical, methodological, and biological heterogeneity are often topic specific, but statistical heterogeneity can be examined with the same methods in all meta-analyses. Therefore, the perception of statistical heterogeneity or homogeneity often influences meta-analysts and clinicians in important decisions. These decisions include whether the data are similar enough to combine different studies; whether a treatment is applicable to all or should be "individualised" because of variable benefits or harms in different types of patients; and whether a risk factor affects all people exposed or only select populations. How uncertain is the extent of statistical heterogeneity in meta-analyses? Moreover, is this uncertainty properly factored in when interpreting the results?

Evaluating heterogeneity between studies

Many statistical tests are available for evaluating heterogeneity between studies.[2 3] Until recently, the most popular was Cochran's Q, a statistic based on the χ^2 test.[4] Cochran's Q usually has only low power to detect heterogeneity, however. It also depends on the number of studies and cannot be compared across different meta-analyses.[2 3] Higgins and colleagues, in two highly cited papers,[5 6] proposed the routine use of the I^2 statistic. I^2 is calculated as $(Q–df)/Q\times100\%$, where df is degrees of freedom (number of studies minus 1). Values of I^2 range from 0% to 100%, and it tells us what proportion of the total variation across studies is beyond chance. This statistic can be used to compare the amount of inconsistency across different meta-analyses even with different numbers of studies.[7] I^2 is routinely implemented in all Cochrane reviews (standard option in RevMan) and is increasingly used in meta-analyses published in medical journals.

Higgins and colleagues suggested that we could "tentatively assign adjectives of low, moderate, and high to I^2 values of 25%, 50%, and 75%."[6] Like any metric, however, I^2 has some uncertainty, and Higgins and Thompson provided methods to calculate this uncertainty.[5] Recently, other investigators compared the performance of I^2 and Q in Monte-Carlo simulations across diverse simulated meta-analytic conditions. They found that I^2 also has low statistical power with small numbers of studies and its confidence intervals can be large.[8]

Interpreting heterogeneity in selected meta-analyses

Inferences about the extent of heterogeneity must be especially cautious when the 95% confidence intervals around I^2 are wide, ranging from low to high heterogeneity. Such uncertainty is usually ignored in systematic reviews, however. This can result in misconceptions. For example, a systematic review of corticosteroids for Kawasaki disease found a point estimate $I^2=59\%$.[9] The authors decided to exclude the two studies that were most different, saying that their removal eliminated all of the across study heterogeneity ($Q=5.59$, $P=0.588$, $I^2=0.00$). In fact, the 95% confidence interval for this $I^2=0\%$ estimate still extends from 0% to 56%. With two small randomised trials and six non-randomised comparisons remaining, the meta-analysis concluded that corticosteroids consistently halve the risk of coronary aneurysms. However, the two largest randomised trials on this topic were published after the meta-analysis. Heterogeneity resurfaced: the largest trial found no effect on coronary dimensions,[10] while the other trial showed an 80% reduction in the risk of coronary artery abnormalities.[11]

Eight systematic reviews published in the *BMJ* between 1 July 2005 and 1 January 2006 performed meta-analyses of randomised trials and seven of them performed some statistical analysis of heterogeneity between studies (table on bmj.com).[12-18] Each review stated that they had tried to interpret heterogeneity, and seven meta-analyses provided enough information for us to calculate the 95% confidence interval of I^2. The lower 95% confidence interval was always as low as 0% (rounded to integer percentage), with one exception. The upper 95% confidence interval always exceeded the 50% threshold, and in four cases it also exceeded the 75% threshold. A conclusive statement was feasible in only one case, where I^2 was 69%, the 95% confidence interval was 40% to 80%, the Q statistic had $P<0.001$, and the authors justifiably concluded that "there was significant heterogeneity among these trials."[13] This meta-analysis had 15 studies, so the power of both Q and I^2 was good. In all other meta-analyses (two to 12 studies each), strong statements in interpreting heterogeneity would be difficult to make. Only one review presented 95% confidence intervals for an I^2 estimate.[12] The authors concluded that "we could not observe significant heterogeneity." Indeed the Q statistic had $P=0.19$. However, with only five studies, the power

to detect heterogeneity was negligible. The I^2 statistic was 35% and the 95% confidence interval ranged from 0% (no heterogeneity) to 76% (high heterogeneity).

Uncertainty in I^2: large scale survey of meta-analyses

This limitation is not confined to the selected examples presented here—it is probably the rule rather than the exception. We used two large datasets of meta-analyses to evaluate empirically the extent of uncertainty in I^2 estimates. Firstly, we looked at meta-analyses of the *Cochrane Database of Systematic Reviews* (Issue 4, 2005) that had four or more synthesised studies and binary outcomes. Because each Cochrane review may include several meta-analyses, we looked only at the one with the highest number of studies; in the case of ties, we used the one with the largest sample size. We did not look at meta-analyses of two or three studies. Such studies form a sizeable proportion of the Cochrane Library,[19] but their 95% confidence intervals of I^2 almost always span a wide range of heterogeneity, unless the studies are large and they give very different results. In total, we calculated the I^2 statistic and its 95% confidence intervals for 1011 meta-analyses. The second dataset was a previously described database of 50 meta-analyses of gene-disease associations that had found a nominally statistically significant effect ($P<0.05$) for the proposed genetic risk factors.[20]

Figure 1 shows the upper and lower 95% confidence intervals of I^2 for the two sets of meta-analyses. The pattern is similar. Of the meta-analyses where I^2 is ≤25% (low heterogeneity), 83% of the Cochrane meta-analyses and 73% of the genetic risk factor meta-analyses have upper 95% confidence intervals that cross into the range of large heterogeneity (I^2 ≥50%). Of the meta-analyses where I^2 is ≥50% (large heterogeneity), 67% of the Cochrane meta-analyses and 52% of the genetic risk factor meta-analyses have lower 95% confidence intervals that cross into the range of low heterogeneity (I^2 ≤25%).

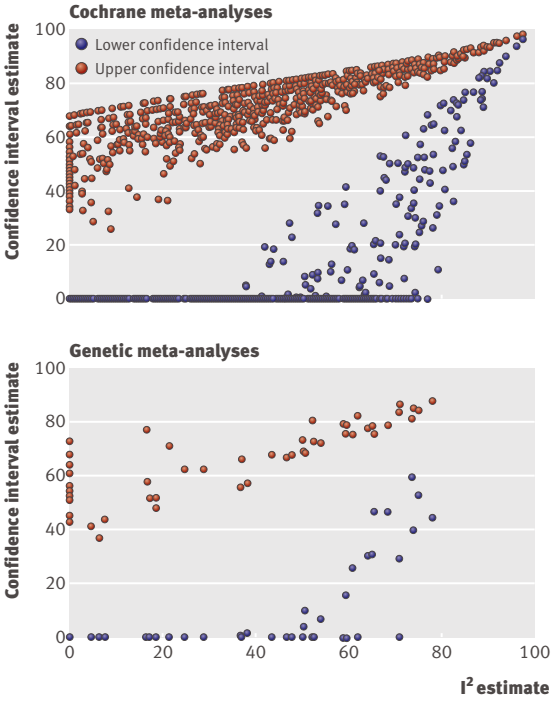

Fig 1 | Confidence intervals for estimated I^2 in 1011 Cochrane meta-analyses and 50 meta-analyses of genetic risk factors. The median number of studies was 7 (interquartile range 5-11) and 20 (13-26), respectively, and the median total sample size was 1112 (512-2691) and 4660 (2823-8761), respectively. The median I^2 was 21% (0-50%) and 38% (5-60%), respectively

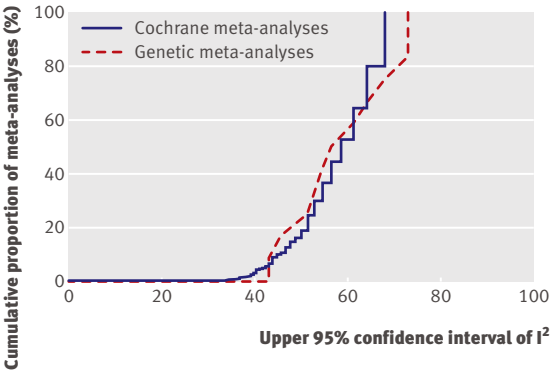

Fig 2 | Proportion of meta-analyses with estimated I^2=0% whose upper 95% confidence interval of I^2 is lower than a given value

Meta-analyses where I^2 is estimated at 0% are affected by an especially important misconception. Many reviews interpret this as absence of heterogeneity, but the upper 95% confidence interval may be substantial (as in the Kawasaki example discussed above[9]). Figure 2 shows the uncertainty for the upper 95% confidence interval of I^2 for the two sets of meta-analyses, limited to those with I^2=0% (n=373 for Cochrane reviews, n=12 genetic studies). The upper 95% confidence interval exceeds 33% in all these meta-analyses. For 81% of the meta-analyses with I^2=0%, the 95% confidence intervals are 50% or higher. Because of the way that research is currently reported, considerable heterogeneity between studies cannot be excluded with confidence in most meta-analyses. Some heterogeneity between studies is probably present in most meta-analyses. Claims for homogeneity may sometimes be stronger than the evidence allows. Trusting a non-significant P value for the Q statistic and an I^2 estimate of 0% may sometimes lead to spurious certainty about the comparability and similarity of study results.

Technical aspects

The confidence interval of I^2 can be calculated by several methods.[5] Two methods, a test based approach and a non-central χ^2 based approach, have been implemented in Stata (heterogi module). The performance of these two methods is comparable, although the test based approach often gives lower values for lower and upper confidence intervals, so that the non-central χ^2 based approach may be preferable.

Concluding comments

All statistical tests for heterogeneity are weak, including I^2. The clinical implications of this are considerable and

must be examined on a case by case basis. Putting too much trust in homogeneity of effects may give a false sense of reassurance that one size fits all. Lack of evidence of heterogeneity is not evidence of homogeneity. Conversely, putting too much trust in the presence of heterogeneity of effects may lead to spurious subgroup and exploratory analyses. Given that I^2 is not precise, 95% confidence intervals should always be given.

Contributors and sources: JI has a longstanding interest in meta-analyses and heterogeneity and had the original idea for this article. NP and EE collected the data. NP performed statistical analyses with help from JI and EE. JI wrote the manuscript and NP and EE commented on it. JI is guarantor.

Competing interests: None declared.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Lau J, Ioannidis JPA, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998;351:123-7.
2 Sutton A, Abrams K, Jones D, Sheldon T, Song F. *Methods for meta-analysis in medical research*. Chichester: Wiley, 2000.
3 Petitti DB. Approaches to heterogeneity in meta-analysis. *Stat Med* 2001;20:3625-33.
4 Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954;10:101-29.
5 Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539-58.
6 Higgins JPT, Thompson SG, Deeks J, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.
7 Mittlbock M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat Med* 2006;25:4321-33.
8 Huedo-Medina TB, Sánchez-Meca F, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychol Methods* 2006;11:193-206.
9 Wooditch AC, Aronoff SC. Effect of initial corticosteroid therapy on coronary artery aneurysm formation in Kawasaki disease: a meta-analysis of 862 children. *Pediatrics* 2005;116:989-95.

SUMMARY POINTS

The extent of between study heterogeneity should be measured when interpreting results of meta-analyses

Meta-analyses rarely document uncertainty in estimates of heterogeneity

Our evaluation of a large number of meta-analyses shows a wide range of uncertainty about the extent of heterogeneity in most

Confidence intervals of I^2 should be calculated and considered when interpreting meta-analyses

10 Newburger JW, Sleeper LA, McCrindle BW, Minich LL, Gersony W, Vetter VL, et al; Pediatric Heart Network Investigators. Randomized trial of pulsed corticosteroid therapy for primary treatment of Kawasaki disease. *N Engl J Med* 2007;356:663-75.
11 Inoue Y, Okada Y, Shinohara M, Kobayashi T, Kobayashi T, Tomomasa T, et al. **A multicenter prospective randomized trial of corticosteroids** in primary therapy for Kawasaki disease: clinical course and coronary artery outcome. *J Pediatr* 2006;149:336-41.
12 Maier PC, Funk J, Schwarzer G, Antes G, Falck-Ytter YT. **Treatment of** ocular hypertension and open angle glaucoma: meta-analysis of randomised controlled trials. *BMJ* 2005;331:134.
13 Dennis CL. **Psychosocial and psychological interventions for** prevention of postnatal depression: systematic review. *BMJ* 2005;331:15.
14 Devereaux PJ, Beattie WS, Choi PT, Badner NH, Guyatt GH, Villar JC, et al. How strong is the evidence for the use of perioperative beta blockers in non-cardiac surgery? Systematic review and meta-analysis of randomised controlled trials. *BMJ* 2005;331:313-21.
15 Taylor SJ, Candy B, Bryar RM, Ramsay J, Vrijhoef HJ, Esmond G, et al. Effectiveness of innovations in nurse led chronic disease management for patients with chronic obstructive pulmonary disease: systematic review of evidence. *BMJ* 2005;331:485.
16 Webster AC, Woodroffe RC, Taylor RS, Chapman JR, Craig JC. Acrolimus versus ciclosporin as primary immunosuppression for kidney transplant recipients: meta-analysis and meta-regression of randomised trial data. *BMJ* 2005;331:810.
17 McDonald MA, Simpson SH, Ezekowitz JA, Gyenes G, Tsuyuki RT. Angiotensin receptor blockers and risk of myocardial infarction: systematic review. *BMJ* 2005;331:873.
18 Glass J, Lanctot KL, Herrmann N, Sproule BA, Busto UE. **Sedative** hypnotics in older people with insomnia: meta-analysis of risks and benefits. *BMJ* 2005;331:1169.
19 Ioannidis JP, Trikalinos TA, Zintzaras E. Extreme between-study homogeneity in meta-analyses could offer useful insights. *J Clin Epidemiol* 2006;59:1023-32.
20 Ioannidis JP, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol* 2006;164:609-14.

Improving the quality of care with performance indicators

Effective improvements in health care require methods to evaluate professional practice. **Azeem Majeed**, **Helen Lester**, and **Andrew Bindman** examine the assessment of quality

The quality of services provided by primary care doctors varies widely, and there is often a large gap between optimal primary care services and actual practice.[1] This quality gap can have serious health consequences, including deaths from medical errors, increased rates of complications in chronic disease, hospital admissions for adverse drug reactions and interactions, and outbreaks of potentially preventable infectious diseases such as measles. It also has large financial costs for the healthcare system, national governments, and society, as well as affecting patients' quality of life.

The reasons for the quality gap are not always within the doctors' control. Sometimes the cause can lie with the public—for example, parents who refuse to allow their child to receive the measles, mumps, and rubella vaccine because of concerns about side effects. Even when the doctor and patient agree to follow a healthcare plan that meets the highest standard for quality, structural barriers related to the design or financing of healthcare systems can prevent the timely receipt of that service—for example, screening mammography for an appropriately aged woman. Nevertheless, the focus of this article and others in the series is on measuring the performance of doctors. Causes of the quality gap that lie with the doctor include being unaware of best practice and the latest guidance on managing a condition or being wary about using certain interventions, such as warfarin to reduce the risk of cerebrovascular disease, because of the fear of adverse events.

Azeem Majeed professor of primary care, Department of Primary Care and Social Medicine, Imperial College Faculty of Medicine, London W6 8RP

Helen Lester professor of primary care, National Primary Care Research and Development Centre, University of Manchester, Manchester M13 9PL

Andrew B Bindman professor of medicine, epidemiology, and biostatistics, Division of General Internal Medicine, University of California San Francisco, San Francisco General Hospital, San Francisco, CA 94110, USA

Correspondence to: A Majeed
a.majeed@imperial.ac.uk

What is quality and how do we measure it?

The Institute of Medicine defines quality as: "The degree to which health services for individuals and populations increase the likelihood of desired health

outcomes and are consistent with current professional knowledge."[2] To measure how well health services meet this goal, a range of performance indicators (sometimes described as quality indicators or quality measures) have been developed.[3]

Indicators are measurable elements of practice for which there is evidence or consensus that they reflect quality and hence help change the quality of care provided. Indicators are often based on routinely collected data, data from electronic medical records, and sometimes data from surveys.[4]

This is the first article in a series looking at use of performance indicators in the UK and elsewhere.
This series is edited by Azeem Majeed, professor of primary care, Imperial College London (a.majeed@imperial.ac.uk) and Helen Lester, professor of primary care, University of Manchester (helen.lester@manchester.ac.uk).

Current initiatives

In England in the 1990s, the use of performance indicators initially developed ad hoc, with different regions developing their own indicators. The introduction of performance indicators was accompanied by various other quality improvement initiatives including a series of national service frameworks, which set out objectives for the health service, and the establishment of the National Institute for Clinical Excellence (now the National Institute for Health and Clinical Excellence), which provides guidance on promoting good health and preventing and treating ill health.

During the past decade, the development and implementation of performance indicators has been largely driven by an increased interest in the quality of care and the arrival of computerised administrative and clinical databases that, for the first time, could provide routine information on quality. Performance indicators have become increasingly sophisticated—for example, moving in the UK from relatively simple indicators based on administrative or claims data to more sophisticated measures based on clinical information from electronic medical records. In the United States, the development of quality measurement was initially driven by the rapidly increasing costs of health care and purchasers' need to know they were getting value for money. Other important factors were the desire to make performance data available publicly and developments in health informatics, which have reduced the cost of producing performance indicators while steadily increasing their sophistication.[5]

In April 2004, the UK government took the bold step of introducing standardised performance indicators across the country and linking performance to general practitioners' pay.[6] The quality and outcomes framework in the resulting new contract for general practitioners includes a range of performance measures for clinical, organisational, and other areas (such as cervical screening and contraceptive services) and also patient experience. Early results suggest that most general practices achieved high scores across the different parts of the framework. However, as indicators within the framework change and thresholds for achievement alter, we may begin to see greater variation between practices in measured quality of care.

Public disclosure of performance data

Public disclosure of performance data is becoming increasingly common in both the UK and the US. In the UK, practice data from the quality and outcomes framework is published online, giving the public access to standardised information on general practices for the first time (www.qof.ic.nhs.uk/). At present, this does not seem to have a large role in how patients choose their general practice, although public disclosure of performance data has been shown to encourage provider organisations to improve quality.[7] However, use may change as the range of information on general practices increases and patients become more skilled at using the internet to view performance data.

What can we learn from other countries?

Performance monitoring systems and performance indicators have largely been developed in a manner that makes them unique to each country. This means it can be difficult to compare quality of care and to transfer performance indicators directly between different health systems, clinical practices, and cultures.[8] However, although some indicators will inevitably be country specific, others, such as glycated haemoglobin concentrations in diabetic people or percentage of patients with coronary heart disease prescribed statins, need not be, and could be designed in a way that makes international comparisons possible.

Quality indicators can also be used to benchmark performance across countries to gain insights into what is achievable and how to improve quality.[9] However, this requires investment in information systems to support measurement of quality. For example, it would have been difficult to implement a system of performance indicators throughout the UK without widespread computerisation of medical records in primary care. Furthermore, although the practice of primary care has many similarities in different countries, differences in the way in which clinical data are collected and coded complicates comparisons.[10] For example, the UK uses Read codes, the US uses international classification of disease (ICD-9) codes, and many other countries use international classification of primary care (ICPC) codes. An internationally accepted set of data standards for coding diagnoses and other clinical data is needed as a first step towards routine comparisons of quality of health care across countries.

Information systems that can produce comparable information on process and outcomes of care are also needed to enable international comparisons. The need for substantive baseline data to compare change in quality against pre-existing trends, and the development of supporting educational strategies for health professionals are also issues that other countries may wish to consider if introducing a national system of performance indicators.

Web resources

Quality and outcomes framework data for general practices (www.qof.ic.nhs.uk/)—Provides information on the performance of general practices in England

Quest for Quality and Improved Performance (http://212.72.48.4/QQUIP/index.aspx?Chapterid=19691)—Independent data and commentary about the quality and performance of health care in England

National Quality Measures Clearing House (www.qualitymeasures.ahrq.gov/)—A US Department of Health and Human Services sponsored public repository for evidence based measures of quality

National Quality Forum (www.qualityforum.org) —A private, not for profit organisation created to develop and implement a national strategy for measuring and reporting healthcare quality in the US

SUMMARY POINTS

The performance of primary care doctors is being monitored more closely in the US and UK

Purchasers of health care are also starting to link performance to pay

Public access to performance data is allowing more informed decisions when choosing doctors and health care

Computerisation will enable the development of increasingly sophisticated performance indicators

Greater standardisation of clinical data and performance indicators is needed for more meaningful international comparisons

Finally, using performance indicators also has some potential adverse consequences. These include doctors declining to accept patients who could be difficult to manage; overtreatment of patients who may not benefit greatly from an intervention; and neglect of areas not covered by performance monitoring. Doctors may have to spend more time on collecting the performance data and less on dealing with patients. However, despite the pitfalls,[11] performance measures in primary care are here to stay and will be used increasingly for quality improvement and performance management in the UK, Europe, United States, and elsewhere. Other articles in this series will discuss how quality measures have been used in the United States; the patient perspective on measuring quality; whether quality of care is determined by more than what is measurable; and future directions in measuring quality in primary care.

Competing interests: AM's department has received funding for work on developing methods of measuring quality of care from the Department of Health and Dr Foster Intelligence. HL provides academic advice to the BMA and employers' negotiating teams on the development of the quality and outcomes framework.

Contributors and sources: AM has research interests in the measurement of healthcare quality using administrative and clinical databases. HL has written about pay for performance and has a long research interest in health quality and inequalities. AB performs research on the impact of polices on low-income persons' access to and quality of care in the United States. This article was based on previous reviews of the measurement of healthcare quality completed by the authors.

Provenance and peer review: Commissioned; externally peer reviewed.

1 Institute of Medicine. *Crossing the quality chasm: a new health system for the 21st century*. Washington, DC: National Academy Press, 2001.
2 Institute of Medicine. *IOM definition of quality*. www.iom.edu/CMS/8089.aspx.
3 Majeed FA, Voss S. Performance indicators for general practice. *BMJ* 1995;311:209-10.
4 Marshall M, Campbell S, Hacker J, Roland M. *Quality indicators for general practice*. London: Royal Society of Medicine, 2002.
5 Goldfield N, Gnani S, Majeed A. Profiling performance in primary care in the USA. *BMJ* 2003;326:744-7.
6 Roland M. Linking physician pay to quality of care—a major experiment in the United Kingdom. *N Engl J Med* 2004;351:1448-54.
7 Marshall M, Shekelle PG, Leatherman S, Brook RH. The public release of performance data. *JAMA* 2000;283:1866-74.
8 Marshall M, Roland M, Brook R, McGlynn E, Shekelle P. *Measuring general practice. A demonstration project to develop and test a set of primary care clinical quality indicators*. London: Nuffield Trust, 2003.
9 Lester HE, Hobbs FDR. Major policy changes for primary care: potential lessons for the US new model of family medicine from the quality and outcomes framework in the UK. *Fam Med* 2007;39:90-6.
10 Bindman AB, Britt H, Crampton P, Forrest CB, Majeed A. Diagnostic scope of and exposure to primary medical care in Australia, New Zealand, and the United States: Results from three national surveys. *BMJ* 2007;334:1261.
11 Heath I, Smeeth L, Hippisley-Cox J. Measuring performance and missing the point? *BMJ* (in press).