ITALIAN JOURNAL OF PUBLIC HEALTH

Meta-analysis in public health: potentials and problems

John P.A. Ioannidis

Department of Hygiene and Epidemiology, University of Ioannina, School of Medicine, Ioannina, Greece and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts University School of Medicine, Boston, USA

Correspondence to: John P.A. Ioannidis, MD, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece. E-mail:jioannid@cc.uoi.gr

Abstract

Meta-analysis has evolved as a primary tool for evidence-based medicine. Initially, meta-analysis was seen as a technique that could improve statistical power in a research world of small, underpowered studies. We increasingly recognize that meta-analysis is a critical tool that can help us measure and understand not only summary effects, but also heterogeneity (diversity) and bias. Here I discuss some key themes and challenges for "meta-epidemiology". These include the contrast between randomized and observational evidence; the evolutionary nature of biomedical evidence; the contrast between small and larger studies; the difficulties in appraising study "quality" and its potential impact on the study effects; and the scandal of missing even minimal, key information on the harms of interventions that are otherwise postulated to be effective. I discuss a general outlook about the validity of the evidence in medicine and public health. I suggest that we should learn to live with uncertainty, since the evidence that is available is often limited, biased, or both. This means that we should be prepared to dismiss big chunks of biomedical dogma, including perhaps whole specialties and sub-specialties of current medicine, as false, erroneous, irrelevant or even potentially dangerous for the health of individuals and populations. An effort should be made to shift the accumulation and synthesis of evidence towards answering critical public health-related questions.
This paper is based on a lecture presented at the European Public Health Association 2005 annual conference in Graz, Austria.

Key words: meta-analysis, evidence based medicine, public health, bias, uncertainty

Meta-analysis: why? why not?

The advent of evidence-based medicine (EBM) has challenged our understanding about the relative merits of various sources of information.

Expert opinion has had major prestige and influence in clinical and public health decision-making in the past. In the EBM framework, expert opinion was found to be often extremely misleading [1]. Therefore hierarchies of evidence developed in the 1990s not only displaced expert opinion from its prominent place, but also ranked it at the worst possible tier of evidence.

Conversely, meta-analysis, the systematic and rigorous quantitative integration of information on the same research question, reached the top tiers of these hierarchies [2].

While this prominence is now widely accepted, meta-analysis has also given us insights about the limitations of evidence, including its own limitations. We increasingly appreciate that study design alone does not suffice for ensuring the credibility of the evidence [3]. Different questions may require different designs to answer them. Moreover, a well-done observational study may sometimes be preferable or superior to a poorly done, conflicted randomized trial or a meta-analysis of several such trials.

A bad meta-analysis of poorly done studies may do more harm currently than many ignorant experts. Meta-analysis has been critical to its own self. This has led to its gradual transformation.

Initially, meta-analysis was seen as a promising method that can help compile data in a mathematically appropriate way [4]. Combing data could improve statistical power, when there were several small studies on a specific question, but all of them were largely underpowered, when seen in isolation.

While this is a temping advantage of synthesizing information, we should be careful not to overemphasize this "synthetic" function of meta-analysis.

Many times the data to be synthesized are problematic in one or several ways. Simply putting problematic data together will not overcome their problems. Superficial combination of poor data entails the risk of sanctifying poor information.

This difficulty has led some researchers to think that meta-analysis should be applied only when the

constituent data are "perfect": large, well-done studies with no biases and with very similar results. However, I find this application of meta-analysis rather boring, if not superfluous. With such perfection, even a single study should be able to give the answer to the question-at-hand. Thus, it is unlikely we will know much more through meta-analysis compared to what each single study can tell us. Needless to stress, this scenario is also uncommon. Perfect biomedical research is a rare specimen by definition, given the inherent difficulties of conducting such research.

The common scenario is that we have studies of various strengths, designs, quality, and problems. Besides getting summary estimates, meta-analysis is more useful in this situation for listing and possibly dissecting sources of bias, quantifying heterogeneity, and proposing some potential explanations for dissecting genuine heterogeneity from bias [5]. For public health-related interventions, in particular, heterogeneity and bias are major forces. Ignoring them would be inappropriate.

In this presentation I will discuss some of the key challenges that we face when we try to address heterogeneity and bias in the framework of meta-analysis:

1. The contrast of randomized vs. observational evidence: what kind ofstudies can we trust to be certain?
2. The contrast of large vs. small studies: how extensive does the evidencehave to be to be certain?
3. The contrast of early vs. late studies: when can we be certain?
4. The contrast of good vs. bad quality studies: does quality matter?
5. The contrast of safety vs. efficacy: doing some good or doing no harm?

Randomized vs. observational evidence

Theoretically randomized trials are susceptible to fewer biases than non-randomized observational studies in evaluating the efficacy of therapeutic and preventive interventions. However, this does not mean that they are always worse. Two groups of investigators probed 5 and 18 medical questions respectively where both randomized trials and observational studies had been conducted [6,7]. Surprisingly, these evaluations concluded that the results of the two designs generally agree with each other and randomised trials actually show larger between-study heterogeneity [6-8].

Examining 45 different questions where both randomized and observational studies existed on the same question, we observed the contrary [9]: heterogeneity was more prominent in observational evidence. Moreover, in about 1 of 5 questions the two types of evidence disagreed beyond chance and in a much larger percentage the difference in the effect size was enough to bear careful consideration, even if this difference were not formally statistically significant.

We should caution that the questions where both randomized and observational data exist are rather limited and their existence is determined by very strong selection forces. Overall, the observed agreement may be even exaggerated. Moreover, both randomized and observational data may sometimes get the wrong answer, so even if they do agree among themselves, this does not prove that they are correct. In all, the comparison of randomized and observational evidence already provides us hints that the evidence is sometimes unreliable.

Does evidence change over time?

Given this uncertainty even in seemingly robust study designs, when can we be sure that an intervention works? Or that an intervention does not work? When can we decide that we have enough evidence and no more studies are indicated? Can we be fooled even when large amounts of evidence have been accumulated?

Moreover, can we suspect that we are being fooled?

We have accumulated direct data and indirect hints that the available published evidence at any time may be only a selected part of the total possible evidence. Besides clear-cut publication bias [10], time lag bias may cause the delayed publication and dissemination of the least favourable "negative" results [11]. We have empirical evidence that while trials with formally statistically significant and those with non-significant results take the same time to complete - with certain exceptions - the latter trials are delayed publication after their completion [11].

In the world of non-randomized evidence, the selective publication forces may be even more prominent. Significance-chasing is probably ubiquitous and may distort the evolving literature over time. Significance-chasing may not necessarily result in publication bias. A "negative" study may not necessarily disappear in the world of unpublished literature. It may be transformed into some kind of "positive" results, by selective use of outcomes, analyses and distorted presentation. What is worse, typically we have no clue what the original protocol has been so as to safeguard ourselves against such manipulation.

Finally, what I call here "manipulation" might be what some other researchers would call "imaginative analysis". Evidence of selection biases in the observational literature has been obtained along several lines, including prognostic studies and genetic studies [12,13].

This is probably just the tip of the iceberg.

Effect sizes may fluctuate over time as more evidence appears in the published literature on the same question. Fluctuations may be of two kinds.

In the more benign scenario, it is only an issue of the uncertainty surrounding the summary effect [14,15]. This uncertainty goes beyond what is conveyed typically by the 95% confidence intervals of a summary estimate by a traditional random effects meta-analysis model.

We know that even when we have accumulated data on 2,000 randomized subjects, the next trial may change the summary relative risk by about 25% [15]. Given that most effect sizes are relatively modest, uncertainty for the mere presence of a treatment effect is likely to exist even when several thousand subjects have been randomized. In the majority of clinically important questions, considerable uncertainty is the rule.

A more problematic type of fluctuation is when effect sizes change in the same direction as more evidence accumulates. In the most common scenario, the addition of more data tends to shrink the treatment effect.

While genuine heterogeneity may be operating in some of these cases, the most common explanation is that here we are dealing with biased early estimates of effect that get dissipated as better data gradually accumulate.

Many early effects may disappear with more careful scrutiny and additional data. In the epidemiological literature, this may be particularly prominent, and in discovery-oriented research with massive hypothesis testing it may be the rule [16,17].

Small vs. larger studies

Theoretically, large and small studies should give the same results, even if we let aside for a moment the inherent difficulty of defining what qualifies for a large study.

Evaluating treatment effect as a function of study size or precision may be considered across a variety of meta-analyses. Empirical evidence suggests that for randomized trials, studies with over 1000 subjects tend to have similar results as smaller studies, but discrepancies that are beyond what can be accounted by chance may still occur in 10-30% of the cases [18-21]. In these situations, large studies tend to give more conservative results, but this is not always the case. Differences in the exact estimate of the effect sizes are even more common; in many cases these discrepancies do not reach formal statistical significance probably due to low power alone.

Discrepancies between small and larger studies may offer a hint for either genuine heterogeneity or bias. It is misleading to consider that such discrepancies are necessarily due to publication bias. The discriminating ability of precision-based tests for publication bias is unknown [22], but it may be low. For most important biomedical questions, only a couple of randomized trials are performed, if at all, so probing for publication bias retrospectively is a search in the dark.

Empirical evidence suggests that discrepancies between small and larger studies also tend to be more frequent for secondary than for primary endpoints [21]. Other reasons to consider include genuine design differences between large and smaller studies, and field-specific issues that vary across different research questions.

"Quality" of studies

Early empirical evaluations suggested that effect sizes may depend on aggregate quality scores [23]. This claim has stimulated thinking about these issues, but it may be problematic. There are so many quality scales and scores [24] that inferences are widely different depending on how quality is defined and summarized [25]. Moreover, "quality" of research is often non-transparent when only a published report is available and may remain partly non-transparent even when a detailed protocol is available. Quality may often be very difficult and/or subjective to measure.

Other empirical evaluations suggested that specific quality items such as lack of blinding and lack of allocation concealment in RCTs may inflate treatment effects [26]. Now it seems more likely that such quality deficits may be associated either with inflated or with deflated treatment effects [27]. It is difficult, if not impossible, to know in advance how a specific quality defect will modify the results of a study. It is also practically impossible to "correct" the results of a poor study based on its perceived quality deficits.

The whole research about the quality of clinical, including public health, research, has sensitized us to the multiple problems that may arise in designing, conducting, analyzing, and reporting a study. Hopefully, this sensitization will lead to better studies in the future [28].

With the risk of oversimplifying, this empirical research has also led to an understanding that there are broadly two kinds of "bad" quality. First,

quality may be bad on purpose. In this situation, investigators, sponsors, or other people and forces involved in a study are aware of these poor choices, but nevertheless they implement them. Here the typical motive is to inflate treatment effects and bad quality usually means also a shift in the results in one specific direction. Second, quality may be bad because of ignorance.

Ignorance about clinical research methods is unfortunately highly prevalent. In this case, the poor choice is due to lack of knowledge and familiarity with research methods. These quality deficits may also be a surrogate of other deficits in the design, conduct and analysis of a study. Here the effect sizes may be affected in either direction, and actually it may be more common to get deflated rather inflated effect sizes when poor quality is due to ignorance. Admittedly separating conflicted knowledge from ignorance is not always easy.

Harms

Decision making in health care requires a balancing of efficacy information against safety information. Unfortunately there is increasing evidence that the recording of adequate safety information in medical research, in particular randomized trials, is neglected [29]. This would be a particular concern for interventions in public health, where it is very important that harms to the general population are minimized. Empirical data suggest that the space devoted to harms in the Results section of randomized trial reports is less than the space that is given to write the author names and their affiliations in these same articles [29,30]. There are very few interventions for which large-scale evidence is available on well-defined potential harms [31]. Explicit recommendations have been made on how to help improve this situation, at least at the reporting phase [32].

However, since changes are unlikely to be instantaneous, one should be very cautious in applying evidence where the harms have not been properly quantified, even if the benefit seems quite well-documented.

Overall validity of the evidence

The more we scrutinize the evidence, the more we realize that the evidence upon which medicine and public health are built is often thin and unreliable. It may be that the large majority of biomedical research findings are false [33]. This situation is not likely to improve in the current high-tech, molecular era of discovery-oriented research. If anything, the typical false discovery rate may be gradually escalating. We should get used to living with this uncertainty.

It can be shown [33] that the post-study odds of a true finding are small when effect sizes are small; when studies are small; when fields are "hot" (many teams work on them); when there is strong interest (and possibly conflicts) in the results; when databases are rich; and when analyses are more flexible. This reflects actually the picture of current biomedical research.

The credibility of many research claims may be very low. Many medical sub-specialties are built on thin air. I would not be surprised if entire specialties of medical practice disappear one day soon, after which we will not believe that we had allowed them to flourish all along and to receive fame and public - as well as private - money and attention. The main barrier to the dissolution of big components of biomedical dogma may be our own inability to accept that our profession might often have been way wrong, irrelevant to health, or even harmful. Some prestigious medical sub-specialties may be surviving out of pure bias and professional self-conflicts of interest to the detriment of human kind.

Some final remarks

Meta-analysis can give some answers towards optimizing the biomedical knowledge base and eventually its application to health care and public health. We should not forget that evidence is not simply numbers. In health care, it must be interpreted with a critical mind, and applied with humane compassion. Meta-analysis has taught us many lessons about the need to live with a lot of uncertainty and healthy scepticism. Evidence is often uncertain despite all good intentions. What is worse, not even good intentions can be ascertained all the time in the generation of the evidence.

Even when evidence seems certain, we may still be misled.

This scepticism does not necessarily mean that action has to wait until evidence becomes certain. Unfortunately, for many important questions of public health portend, evidence may remain limited, fragmented and biased to some extent. The research agenda may sometimes continue to spend millions of Euros on petty research questions for petty drugs that simply inflate the assets of companies or the CV of petty investigators rather than answer questions of major public health importance.

While public health practitioners feel this great injustice, we still need to act, but be prepared for major surprises, both positive and negative. Highlighting the uncertainty that we have for

many public health issues may also help to draw more attention to them. This may then help drive persuasive arguments that we need better and more convincing research.

Very little research is conducted for public health-oriented questions, although these are of major importance. The evidence, whatever exists, is often irrelevant. For example, with a third of the global burden of disease carried by Africa, only 0.2% of randomized trials have been conducted in Africa. As a monument to greed and irrationality, the vast majority of these few trials that have been conducted in Africa still pertain to drug interventions for diseases that affect affluent economies [34].

Developing countries would benefit particularly for the dissemination of timely, critical systematic reviews [35], but these systematic reviews also need relevant primary evidence to be generated [36].

I have also not discussed here how meta-analyses and systematic reviews can be most efficiently integrated in public health practice. This is a major challenge on its own [37] that goes beyond the scope of this presentation.

However, it is likely that the same meta-analyses may have small or large impact depending on how they are used. As public health struggles with the often inadequate sufficiency, stability, and credibility of the evidence for important questions [38], we need to find a way to ensure that useful meta-analyses find their way into public health practice and practitioners.

Meta-analysis also creates a tightly disciplined scientific basis for medicine and public health. Perhaps this happens for the first time in the history of medicine and public health. The basic biological sciences have already detached themselves from medicine, as they have become extremely powerful in terms of their research agenda and support in the last 20 years.

Medicine and public health used to lead these scientific disciplines, but now the reverse scenario unfolds. Evidence-based medicine can allow medicine and public health to maintain a rigorous empirical research philosophy and an independent standpoint as scientific disciplines.

This should hopefully allow developing and maintaining healthy and mutually fruitful critical links with the rapidly developing front of basic biomedical sciences and to deal more efficiently with diseases of major population impact.

References

1) Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA 1992;268:240-8.

2) West S, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Rockville, MD: Agency for Healthcare Research and Quality, 2002:51-63. (AHRQ publication No 02-E016).

3) Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.

4) Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med 1987;316:450-5.

5) Lau J, Ioannidis JPA, Schmid CH. Summing up evidence: one answer is not always enough. Lancet 1998;351:123-7.

6) Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878-86.

7) Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887-92.

8) Ioannidis JPA, Haidich A-B, Lau J. Editorial: Any casualties in the clash of randomised and observational evidence? BMJ 2001;322:879-80.

9) Ioannidis JPA, Haidich A-B, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and non-randomized studies. JAMA 2001;286:821-30.

10) Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991;337:867-72.

11) Ioannidis JPA. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 1998;279:281-6.

12) Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. J Natl Cancer Inst 2005;97:1043-55.

13) Pan Z, Trikalinos TA, Kavvoura FK, Lau J, Ioannidis JP. Local literature bias in genetic epidemiology: an empirical evaluation of the chinese literature. PLoS Med 2005;2:e334.

14) Ioannidis JPA, Lau J. Evolution of treatment effects over time: empirical evidence from recursive cumulative meta-analyses. PNAS 2001;98:831-6.

15) Trikalinos TA, Churchill R, Ferri M, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. J Clin Epidemiol 2004;57:1124-30.

16) Ioannidis JPA, Ntzani EA, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet 2001;29:306-9.

17) Ioannidis JPA. Genetic associations: false or true? Trends Mol Med 2003;9:135-8.

18) Villar J, Carroli G, Belizan JM. Predictive ability of meta-analyses of randomised controlled trials. Lancet 1995;345:772-6.

19) LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997;337:536-42.

20) Cappelleri JC, Ioannidis JPA, Schmid CH, et al. Large trials versus meta-analysis of smaller trials: how do their results compare? JAMA 1996;276:1332-8.

21) Ioannidis JPA, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. JAMA 1998;279:1089-93.

22) Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315:629-34.

23) Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609-13.

24) Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995;16:62-73.

25) Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282:1054-60.

26) Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408-12.

27) Balk EM, Bonis PAL, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973-82.
28) Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663-94.
29) Ioannidis JPA, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. JAMA 2001;285:437-43.
30) Papanikolaou PN, Churchill R, Wahlbeck K, Ioannidis JP. Safety reporting in randomized trials of mental health interventions. Am J Psychiatry 2004;161:1692-7.
31) Papanikolaou PN, Ioannidis JP. Availability of large-scale evidence on specific harms from systematic reviews of randomized trials. Am J Med 2004;117:582-9.
32) Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, Schulz K, Moher D; CONSORT Group. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004;141:781-8.
33) Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
34) Isaakidis P, Swingler GH, Pienaar E, Volmink J, Ioannidis JPA. Burden of disease and randomized evidence in sub-Saharan Africa: survey of research. British Medical Journal 2002;324:702-5.
35) McMichael C, Waters E, Volmink J. Evidence-based public health: what does it offer developing countries? J Public Health (Oxf) 2005;27:215-21.
36) Waters E, Doyle J. Evidence-based public health practice: improving the quality and quantity of the evidence. J Public Health Med 2002;24:227-9.
37) Dobbins M, Cockerill R, Barnsley J, Ciliska D. Factors of the innovation, organization, environment, and individual that predict the influence five systematic reviews had on public health decisions. Int J Technol Assess Health Care 2001;17:467-78.
38) Muellerleile P, Mullen B. Sufficiency and stability of evidence for public health interventions using cumulative meta-analysis. Am J Public Health 2006;96(3):515-22. Epub 2006 Jan 31.