Large scale evidence and replication: insights from rheumatology and beyond

J P A Ioannidis

Many biomedical and research findings fail to be replicated and validated

Increasing concern has been expressed that biomedical research findings many times fail to be replicated and validated. This may occur across the spectrum from basic research to clinical applications. Small studies, poor design, the play of chance, exaggeration in early claims, data dredging, selective publication and reporting of information, strong expert opinions, biases, and financial or other conflicts of interest all probably combine to create uncertainty about whether research findings will stand the test of time and independent replication. Some research findings or beliefs are not even questioned and no effort is ever made to challenge them, but this does not mean that they are true. The recent advent of molecular medicine has only increased the complexity and number of questions that may be asked. It is unclear though, whether the information derived from the various fascinating discovery-driven approaches, massive as it is, is more likely to be possible to replicate. The ratio of sample size to scientific question asked has been decreased to infinitesimal levels, and uncertainty remains as high as ever.

SMALL SAMPLE SIZES: EMPIRICAL DATA

For rheumatology and autoimmune diseases, the danger is as great as for any other medical discipline. Much of the replication difficulties may have to deal with the fact that claims are made or refuted based on limited data in single studies. The median number of patients in randomised controlled trials in systemic sclerosis is only 28,¹ exactly the same as for trials on treatments for systemic lupus erythematosus.² Including also trials on common diseases like rheumatoid arthritis or osteoarthritis, the median sample size for a rheumatology related trial has barely improved, moving from 46 to 88 patients in the past decade.³ Studies in the sample size ranges encountered in

cardiology or oncology mega-trials are unheard of in the rheumatic diseases.

For non-randomised evidence, such as case-control, cohort, or cross sectional studies, investigators are typically content with sample sizes that are usually not much different from those encountered in randomised research. The median sample in genetic association studies encompasses slightly more than 100 cases and a similar number of controls⁴—corresponding to only a couple of dozen subjects carrying the genetic variant of interest in the whole study.

For more molecular level investigations, claims are often made based on an even smaller number of patients. In microarray experiments, thousands of genes are almost always investigated in only a handful of subjects.⁵⁻⁶ The relative rarity of some autoimmune diseases sometimes makes it difficult to assemble even 100 patients, and studies that require transfer of DNA between centres have to face a difficult regulatory environment.

FAILURE TO DISCOVER AND FAILED DISCOVERIES

Given such limited evidence, it is not surprising that research findings are uncertain and often fail their original promise. The typical problem which is claimed when sample sizes are small is that the statistical power to discover something is limited. Thus an effective treatment, a strong risk factor, or an important pathophysiology correlate may be missed, because of limited data. However, small sample sizes lead also to spurious findings. This problem may actually be even more prominent and misleading. Statistically significant findings may still be seen in small studies, when the observed effect gets exaggerated either by chance or through various operating biases, both conscious and unconscious. Hundreds of thousands of small studies are performed, with each of them asking anywhere between one and several thousand research questions. In many cases, investigators are

not even sure about what questions they are asking, or they may change their questions, as they are analysing the data. In fact, such discovery-driven research is increasingly being established as a paradigm of novelty in influential scientific circles.⁶

It is not surprising that given this huge mass of “scientific” testing, several extreme claims may ensue and a good portion of them may eventually be spurious. What is this portion likely to be? This certainly depends on the discipline and type of question and study design, but often it may be high. For genetic associations, for example, it may be that anywhere between 40 and 90% of the discoveries are false positive claims.⁴⁻⁷⁻⁸ The high rate of false discoveries has recently prompted several journals, including *Arthritis and Rheumatism*,⁹ to recommend replication of genetic associations in independent populations in the same publication. For randomised trials, theoretically the “success rate” may be better, but conflicts of interest are strong,¹⁰ so misleading inferences are probably not uncommon at all.

POSSIBILITIES FOR LARGE SCALE EVIDENCE

This situation suggests that it is essential to generate large scale evidence on clinical and research questions of interest, regardless of how strong the data seem on preliminary small studies. It is also very important not to accept evidence, no matter how substantial, at face value, but to scrutinise it for potential errors and biases.

“Even apparently good evidence should be closely examined for possible errors and biases”

There are several different approaches to obtaining large scale evidence. Single centres can occasionally generate large scale information, provided that massive data can be collected meticulously over many years. For example, we recently managed to collect data on almost 800 patients with primary Sjögren’s syndrome followed up at a single centre over a period of 20 years and the evidence was strong enough to suggest that primary Sjögren’s syndrome is probably two different diseases, one with a high risk of lymphoproliferative disease and death and, another, far more common, with no different risk from the general population.¹¹ However, in the vast majority of settings, single centre experiences would be fragmented snapshots. Multicentre collaborations should be encouraged. For example, such multicentre collaborations have recently been able to establish databases on 1000 well characterised

patients with systemic lupus erythematosus¹² and 1000 patients with antiphospholipid syndrome.¹³

Once data are obtained from different centres and different settings, it is important to try to dissect and estimate the variability that exists between the various pieces of information. In a recent collaborative project on systemic sclerosis, we accumulated information on 3311 patients and 19 990 patient-years of follow up.¹⁴ The most interesting finding perhaps was the large diversity that existed between the different cohorts contributing data. In some cohorts we could identify patients with a low risk profile who had a risk of death similar to that of the general population, while in other cohorts with stronger referral patterns, patients with a similar "low risk profile" actually had a very high risk of death. Even for well established diseases with clear criteria, different centres may have very different types of patients. Demonstrating and measuring these differences can provide important lessons, even for diseases that have had well established and widely accepted definition criteria for many years now. For several autoimmune diseases, large scale evidence may be used to improve the definitions of the diseases/syndromes themselves, and in most cases large scale multicentre evidence would be indispensable to improve the standardisation of disease outcome.¹⁵

"BIG" SCIENCE IS NOT NECESSARILY "GOOD" SCIENCE

We should be cautious and realise that large sample sizes are not necessarily synonymous with high quality research and offer no guarantee that the research is scientifically or clinically relevant. Some randomised trials funded by industry manage to achieve very large sample sizes.

"Large sample sizes do not guarantee that the results are reliable"

This primarily serves the purpose of being certain that statistically significant results will be achieved eventually, even if the absolute treatment effects for the selected outcomes are actually trivial from a clinical and scientific viewpoint. Sample size seen in isolation is not a reliable marker of high scientific standards.

EVIDENCE IS AN EVOLVING PUZZLE

Prospective large scale evidence is preferable to collation of already available retrospective data. However, for most research questions in rheumatology and beyond, it is unlikely that we will ever have the single, perfect, definitive, prospective study. Even if we do, it is likely

that other prospective and retrospective studies may also be conducted on the same or similar questions. Thus it is becoming increasingly important to view evidence in its totality and to examine its continuous evolution over time as new data accumulate. This suggests that meta-analytic methodologies are likely to have increasing impact in the future.¹⁶

Meta-analysis, the rigorous and systematic quantitative synthesis of information on the same question, was initially promoted as a tool for pooling data and improving power.¹⁷ Although this is useful, simple pooling does not suffice. As described above, errors and biases abound and naive meta-analysis may do more harm than good by simply summing up erroneous information and leading to erroneous confidence in spuriously "precise" results.

We need more emphasis on methods and diagnostics that would examine rigorously the accumulating evidence for the presence of heterogeneity, bias, or both.¹⁸ Meta-analysis provides a unique opportunity for a critical appraisal of the quality of the data, the potential gaps, sources of diversity, and sources of errors. Feedback from meta-analyses on these identified sources of diversity and problems may also be used to shape and correct the future research agenda.

We should pay more attention to the whole picture and the problems that it reveals. For example, 15 years ago, it was clear that we already had large scale evidence, although fragmented, on non-steroidal anti-inflammatory agents in rheumatoid arthritis and that several problems were commonly encountered in the design and reporting of these trials.¹⁹ Fifteen years later we continue to perform a lot of trials on this area, but we continue also making many of the same mistakes, while there is also the potential for new types of error as the standard requirements increase.^{3 20}

CONCLUSION

In all, this editorial should be read as a plea for evidence. Solid evidence requires large scale and good quality information. It also requires a thorough appreciation of the whole picture and this may be constantly changing over time as more data accumulate. The systematic and critical monitoring of the big picture can be a very important research endeavour on its own and we have a lot to learn from it.

Ann Rheum Dis 2005;**64**:345–346.
doi: 10.1136/ard.2004.027979

Correspondence to: Dr J P A Ioannidis, Ioannina School of Medicine, Ioannina, Greece; jioannid@cc.uoi.gr

Accepted for publication 14 September 2004
Published Online First 30 September 2004

REFERENCES

- 1 Kyriakidi M, Ioannidis JP. Design and quality considerations for randomized controlled trials in systemic sclerosis. *Arthritis Rheum* 2002;**47**:73–81.
- 2 Karassa FB, Tatsioni A, Ioannidis JP. Design, quality, and bias in randomized controlled trials of systemic lupus erythematosus. *J Rheumatol* 2003;**30**:979–84.
- 3 Hill CL, LaValley MP, Felson DT. Secular changes in the quality of published randomized clinical trials in rheumatology. *Arthritis Rheum* 2002;**46**:779–84.
- 4 Ioannidis JP. Genetic associations: false or true? *Trends Mol Med* 2003;**9**:135–8.
- 5 Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* 2003;**362**:1439–44.
- 6 Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* 2004;**4**:309–14.
- 7 Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001;**29**:306–9.
- 8 Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003;**33**:177–82.
- 9 Huizinga TWJ, Pisetsky DS, Kimberly RP. Associations, populations, and the truth: recommendations for genetic association studies in Arthritis and Rheumatism. *Arthritis Rheum* 2004;**50**:2066–71.
- 10 Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ* 2003;**326**:1167–70.
- 11 Ioannidis JP, Vassiliou VA, Moutsopoulos HM. Long-term risk of mortality and lymphoproliferative disease and predictive classification of primary Sjogren's syndrome. *Arthritis Rheum* 2002;**46**:741–7.
- 12 Cervera R, Khamashta MA, Font J, Sebastiani GD, Gil A, Lavilla P, et al. Morbidity and mortality in systemic lupus erythematosus during a 10-year period: a comparison of early and late manifestations in a cohort of 1,000 patients. *Medicine (Baltimore)* 2003;**82**:299–308.
- 13 Cervera R, Piette JC, Font J, Khamashta MA, Shoenfeld Y, Camps MT, et al. Antiphospholipid syndrome: clinical and immunologic manifestations and patterns of disease expression in a cohort of 1,000 patients. *Arthritis Rheum* 2002;**46**:1019–27.
- 14 Ioannidis JPA, Vlachoyiannopoulos PG, Haidich A-B, Medsger TA Jr, Lucas M, Michet CJ, et al. Mortality in systemic sclerosis: an international meta-analysis of individual patient data. *Am J Med* (in press).
- 15 Wolfe F, Lasserre M, van der Heijde D, Stucki G, Suarez-Almazor M, Pincus T, et al. Preliminary core set of domains and reporting requirements for longitudinal observational studies in rheumatology. *J Rheumatol* 1999;**26**:484–9.
- 16 Olkin I. Meta-analysis: reconciling the results of independent studies. *Stat Med* 1995;**14**:457–72.
- 17 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.
- 18 Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998;**351**:123–7.
- 19 Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Control Clin Trials* 1989;**10**:31–56.
- 20 Gotzsche PC. Reporting of outcomes in arthritis trials measured on ordinal and interval scales is inadequate in relation to meta-analysis. *Ann Rheum Dis* 2001;**60**:349–52.