Human Reproduction, Vol.30, No.8 pp. 1747–1748, 2015
Advanced Access publication on June 24, 2015 doi:10.1093/humrep/dev146

human reproduction EDITORIAL

The war on error

Johannes L.H. Evers
Editor-in-Chief
E-mail: jlh.evers@gmail.com

Richard M. Sharpe, Deputy Editor
Edgardo Somigliana, Deputy Editor
Andrew C. Williams, Managing Editor

> *Who knows, asked Robert Browning, but the world may end tonight? True, but on available evidence most of us make ready to commute on 8:30 the next day.*
>
> (*Austin Bradford Hill, 1965*)

Risk is the world's biggest industry and permeates almost every aspect of social and economic life (Adams, 1995). For the search term 'risk' Google will return 893 million hits, as compared with only 491 million for 'disease', but 1670 million for 'sex'. Risk and risk determining factors (risk factors) are the domain of clinical epidemiology. The term 'clinical epidemiology' was coined by Alvan Feinstein in 1968 (Feinstein, 1968). He described it as: 'The clinicostatistical study of diseased populations; the intellectual activities of clinical epidemiology include the following: the occurrence rates and geographic distribution of disease; the patterns of natural and post-therapeutic events that constitute varying clinical courses in the diverse spectrum of disease; and the clinical appraisal of therapy'. Clinical epidemiology aims to describe what the burden of disease is in a community and to provide the methods to help understand causes, risk factors and contributing or predisposing conditions that influence disease (Morrow, 2015). Risk factors in essence are however simple fractions made up of a numerator and a denominator. Fractions are all they are, and risk is what they represent, a statistic, or a predictor at best. But even 'the claim that past behavior is the best *predictor* of future behavior does not mean that past behavior *causes* future behavior' (Wikstrom, 2007). This quotation should be on every clinician's office wall. Clinical epidemiology has developed progressively more sophisticated tools to estimate and express risk, frequently in so-called 'adjusted' odds ratios. But the more sophisticated these tools became the less critical researchers, readers, journalists and policy makers appear to have become. Details—and especially limitations—of new research findings yielded to over-simplistic (and biologically implausible) headlines in the popular press: 'IVF causes skin cancer!'

And now we have entered the world of bioinformatics, with its mammoth datasets, subjected to unfathomably complex high-level statistical software packages. Information technology has proliferated dramatically and the ability to store, retrieve and analyze large amounts of data has increased concomitantly. There are two major forms of clinical databases: administrative databases (often maintained by health insurance companies or health care policy makers) that focus on incidence rates and costs; and clinical registries (by doctors or hospitals) that focus on quality issues and treatment outcome (Cook and Collins, 2015). Both may enable research as a secondary activity. A third novel form of big database, as a primary research activity, is found in genome research. Genome Wide Association Studies (GWAS) examine the association between genetic variants, usually single-nucleotide polymorphisms (SNPs), and a trait, usually a disease, by comparing the DNA of cases (with the disease) and controls (without disease). Immense numbers of genetic variants become available—typically hundreds of thousands to one million or more of tested SNPs—with fold changes as the critical yardstick. Poorly defined cases, insufficient sample size and absence of adjustment for multiple testing are but a few of the problems encountered (Pearson and Manolio, 2008). And also here, the finding that a certain genetic variant co-segregates with a certain cancer does not mean that that particular SNP *causes* the cancer. All database research has in common that the data are not always being used for the purpose they were designated for. Another problem is that the large number of statistical tests performed presents an unprecedented potential for false-positive results (Pearson and Manolio, 2008). The more comparisons are made, the greater is the chance that something will turn out statistically significant, even when no relationship exists. Checking for twenty different, normally distributed risk factors will—by definition—produce one significant finding if a *P*-value of 0.05 is set as the limit of significance. Checking for one million SNPs demands even more statistical restraint. *P*-values of 0.00000005 or less are the rule in this field of research.

In the current issue of the journal we present such a big data study (Reigstad *et al.*, 2015). The authors studied, with impeccable epidemiological techniques, a population-based cohort consisting of all women registered in the Medical Birth Registry of Norway. Cancers occurring in this group of women were identified by linkage to the Cancer Registry of Norway. Although neither of the databases was developed for the purposes of this research, the authors established that of the total study population of more than 800 000 women, 16 525 gave birth following ART and 22 282 women were diagnosed with cancer. Of the latter, 338 were ART women and 21 944 non-ART women. The authors performed sophisticated statistical analyses, corrections and adjustments, including Cox proportional hazard analysis, testing of the assumption of proportional hazards using Schoenfeld residuals, and adjustment for multiple testing by Benjamini-Hochberg correction. The results showed an elevated cancer risk in one out of seven sites for ART women. The hazard ratio (HR) for cancer of the Central Nervous System was 1.50 (95% CI 1.03– 2.18), and among those specifically subjected to IVF (but not ICSI) the HR was 1.83 (95% CI 1.22–2.73). Analysis of risk of overall cancer gave an HR of 1.16, which—not remarkably—reached significance due to the large numbers involved (95% CI 1.04–1.29). However, all findings became statistically non-significant after correction for multiple analyses (Reigstad *et al.*, 2015).

The article gave rise to an animated debate at several subsequent weekly Editorial Team meetings, the expert reviewers and the Associate Editor raised major concerns, but the authors either rebutted or adjusted their analysis and interpretation. External expert advice was sought, and finally, after two rounds of major revisions, it was decided to publish the paper. Some of us argued that the outcome of the study was reassuring, namely no increased cancer risk was found. Others were afraid that the authors' assiduous interpretation of their findings, that the study indicated 'a possible elevation of risk of CNS cancer as well as a slightly increased risk of overall cancer for women who give birth following ART, although the risk estimates were not statistically significant after adjusting for multiple analyses', might still give rise to headlines such as 'IVF increases cancer by 16%'. Therefore, to put the study and its findings into perspective we then invited two authorities in the field to comment on the Norwegian study. Their commentaries are published in this same issue of the journal (Grimes, 2015; Van Wely, 2015). Madelon van Wely points at possible confounding and early detection bias, but feels encouraged by the study as no increases in hazard rates persisted after correction for multiple testing (Van Wely, 2015). David Grimes cautions against the inappropriate use of administrative databases for epidemiologic research. He rejects the claim of an increased cancer risk in the Reigstad study for the small magnitude of the increase, the lack of a clear, specific, and measurable exposure, for inadequate control for potential confounding factors, for the weakness of the association and for the lack of biological plausibility. Hazard ratios below 2, according to Grimes, are more likely due to bias than to causation. Upsetting infertility patients about brain tumors is unwarranted on the basis of this study, and is unethical (Grimes, 2015).

We at *Human Reproduction* feel it's time for serious reflection. Risk ratios above 2 are rare in our field. Carefully performed prospective observational studies of a focused hypothesis with a biological plausibility, drawing attention to hitherto unknown associations may still have their value. On the other hand, linking non-dedicated administrative databases is fraught with error. Or, to quote Warlow, on stroke studies: 'Observational epidemiological studies, whether cohort or case–control, have revealed an amazing number of associations (risk factors) with stroke, most of which cannot possibly be on the causal pathway. In any event, new and eagerly reported associations are commonly not confirmed in later but usually less prominent studies, or they turn out to be due to confounding' (Warlow, 1998).

What does this all mean in daily practice for *Human Reproduction*, its authors and readers? Perhaps most important is for us all to change the way we look at epidemiological studies, in essence to look from the opposite perspective to that used currently. Thus, our basic assumption should be that we will *not* find any statistically significant association (after applying the appropriate corrections). If we do find a significant association, our first thought should be could this have arisen by chance—does it mean that our corrections are deficient? Did we overlook any bias or confounder? Only after rigorously trying to 'correct the observation away', and failing to do so, should we then consider that it might have meaning. This would be the most self-critical approach, but one which runs counter to most of our current concepts and instincts. Nonetheless, as research moves ever more deeply into the mining of large datasets we need to change our mindsets and expectations accordingly if we are to win the war on error and avoid traveling down many false paths.

References

Adams J. *Risk*. London: University College London Press, 1995.

Cook JA, Collins GS. The rise of big clinical databases. *Br J Surg* 2015; **102**:e93–e101.

Feinstein A. Clinical epidemiology I: The population experiments of nature and of man in human illness. *Ann Intern Med* 1968;**69**:807–820.

Grimes DA. Epidemiologic research with administrative databases: red herrings, false alarms, and pseudo-epidemics. *Hum Reprod* 2015;**30**: 1749–1752.

Hill AB. The environment and disease: association or causation? *Proc Royal Acad Med* 1965;**58**:295–300.

Morrow R. Epidemiology: health and disease in populations. www.eolss.net (accessed 26 May 2015).

Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA* 2008;**299**:1335–1344.

Reigstad MM, Larsen IK, Myklebust TA, Robsahm TE, Oldereid NB, Omland AK, Vangen S, Brinton LA, Storeng R. Cancer risk among parous women following assisted reproductive technology. *Hum Reprod* 2015;**30**:1952–1963.

Van Wely M. Assisted reproduction and cancer risk: how useful are national databases? *Hum Reprod* 2015;**30**:1753–1754.

Warlow CP. Epidemiology of stroke. *Lancet* 1998;**352**:1–4.

Wikstrom PO. In search of causes and explanations of crime. In: King R, Wincup E (eds). *Doing Research on Crime and Justice*, 2nd edn. Oxford: Oxford University Press, 2007.