Asian Journal of Andrology (2013) 15, 159–161
© 2013 AJA, SIMM & SJTU. All rights reserved 1008-682X/13 $32.00
www.nature.com/aja

EDITORIAL

Falling sperm counts and global estrogenic pollution: what have we learned over 20 years?

David J Handelsman[1] and Trevor G Cooper[2]

Asian Journal of Andrology (2013) **15,** 159–161; doi:10.1038/aja.2013.8; published online 18 February 2013

If scandal is the engine of progress in politics, and sensation is that of the media, while anti-modern panic is that of environmentalism, what is the dynamic of medical science? Certainly discovery and invention represent the admirable high road, whereas competition and controversy represent the quotidian low road. This Special Issue is directed to reviewing the lessons learned from the most public, fervent and durable controversy in the short history of Andrology, the claims of world-wide falling sperm counts,[1] due to global pollution by industrial estrogenic chemicals,[2] published two decades ago.

Andrology is a relatively new discipline having become organized only over the last quarter of the twentieth century, although the term and concepts are older. Establishing a new discipline requires demarcating boundaries from previous disciplines, yet such forced separation can lead to stagnation when that means missing out on the refreshing methodological innovation and renewal, the restless flux that is characteristic of mainstream science. Accordingly, Andrology has had areas that lag behind the frontiers of clinical medicine and research. Most notable is semen analysis which has been confined to Andrology laboratories and thereby isolated from mainstream pathology. Although the WHO led test standardisation with its series of WHO Laboratory Manuals for semen analysis from 1980, Andrology labs were slow to adopt the pivotal concepts of interlaboratory quality control (QC) and reference ranges which all chemical pathology labs developed in the mid twentieth century. The crucial need for QC was only recognized decades later in Andrology labs,[3,4] eventually becoming widely adopted at the turn of the twenty-first century,[5,6] but not without surprising resistance for such a self-evident necessity,[7] and compliance remains low.[8,9]

Even slower and still less complete was recognition of rigorous methodology for reference ranges. The complacent perseverance with substandard clinical study methodology, especially the use of convenience samples, was perpetuated by peer acceptance, the undesirable flip side of the comforting sanctuary of isolation. Such insulation from the enhanced rigour in trial design and analysis for observational epidemiology since the 1980s[10–12]—particularly for the retrospective cohort, which Feinstein aptly termed the 'trohoc'[11]—slowed the percolation of the rigorous thinking required to develop valid reference ranges for semen analysis. This required a thorough understanding of the test reference range methodology, which is strongly conditional on defining the reference population. For semen analysis to be comparable with all other pathology tests (and applicable to the general population), the ideal would be a population representative sample of men regardless of fertility status. Assembling such a population of randomly selected men unconcerned with their fertility is, however, virtually unobtainable because of the intrusive requirement for collecting semen samples by masturbation. This major disincentive to participation means that in studies seeking volunteers unconcerned about their fertility, the voluntary participation rate is at best 10%–20%. Such small, biased minority sampling virtually precludes extrapolation of findings to the general male population.[13,14] This represents the central dilemma for semen analysis as an investigative tool for observational population studies,[13–15] that is, wherever the implied inference is that the findings can be extrapolated to other male populations. In the face of these difficulties, the first attempt at valid population reference ranges for semen analysis was only published in 2010[16] as part of the development of the fifth edition of the WHO Manual.[17] Even in that analysis, the dominance of infertility testing in Andrology labs led to the compromise of focusing mainly on populations of recently fertile men, despite their being a biased sample of the general male population, drawn from only the approximately 5% who recently fathered a child.

The persistent inability to obtain representative reference ranges of semen analysis from unbiased samples of men represents a major constraint for the valid practical application of semen analysis to toxicology or related analytical population studies of men. In practice, as an investigative tool semen analysis remains largely confined to those undergoing infertility investigations where its application to a captive group with contingent health needs overcomes the resistance to providing semen samples.[18] The only valid exception is for prospective studies using randomisation to balance the unknown as well as known covariables. In that situation, inferences from experiments are focused on the intervention rather than an extrapolation to the general male population. This reflects the remarkable properties of randomisation which, singularly, can guarantee definitive scientific conclusions from experiments compared with the weaker inference from observational studies. In the latter, the proportion of variance explained by known covariables is usually small so that the hunt for unrecognized covariables is essentially endless. Hence, the involvement of experimentation, featuring randomisation to balance unknown as well as known covariables, creates a distinction between strong and weak inference in science. It is no coincidence that the lesser certainties of observational sciences, where truly experimental verification of hypotheses is

[1]ANZAC Research Institute, Concord Hospital & University of Sydney, Sydney, New South Wales 2139, Australia and [2]Tuen Mun, NT, Hong Kong SAR, PR China
Guest Editors for this special issue: Prof. DJ Handelsman (djh@anzac.edu.au) and Dr TG Cooper (ctrevorg@gmail.com)

precluded or very limited, are precisely those that generate the most interminable controversies (evolution, climate, cosmology) even among scientific literati.

These unresolved currents surfaced dramatically with the 1992 Carlsen paper[1] which made startling claims of world-wide falling sperm counts soon followed by the purported explanation: global estrogenic pollution from unidentified industrial chemicals.[2] As Sagan remarked, 'extraordinary claims require extraordinary evidence' and for these claims, the latter never arrived. Yet, such claims fitted perfectly the ideal headline-making attributes—a medical science story, the public's most popular news topic, easily framed into widely understood sensationalistic terms. Almost overnight an alarming claim and cause were projected into the public realm where, in the news media's version of Gresham's Law (that bad money drives out the good), the wilder claims quickly dominated and became fixtures with the staying power that arises from the prevalent journalistic research that comprises recycling of old news clippings. Even in science, the two papers became highly cited with already over 1200 citations each. Yet 20 years later, there is little support among experienced researchers in the field for the validity of either the claimed observation of falling sperm counts or its purported cause. Indeed, the latest review from the proponents of the estrogen pollution hypothesis[2] does not even contain the word 'estrogen'[19] and is without explanation of why their hypothesis was abandoned. Two decades after the publications and ensuing controversy, it is timely to review lessons learned—to ignore one's history is to be condemned to repeat it (Santayana), first as tragedy and then as farce (Marx). Specifically, this review was intended to discern how analytical research in Andrology had gained in sophistication not to repeat such mistakes. To this end, we invited the progenitors of the controversy as well as range of experienced medical scientists familiar with the controversies to reflect on the lessons learned and offer their summations.

In considering the Carlsen paper, technically, a meta-analysis, there were three areas of major flaws—the study population, the laboratory methodology and the data analysis.

Meta-analysis refers to the quantitative pooling of separate, combinable studies with the aim of measuring the same variable by the same methods in similar populations. Originally, meta-analysis occupied the role of the poor man's substitute for the gold standard of a well controlled prospective clinical trial. Its dubious early reputation as '… being to analysis what metaphysics is to physics', a religion,[20] statistical alchemy[21] or 'trying to make one good apple out of a barrel of bad apples' was largely alleviated by thorough methodological standardisation requiring quality checklists.[22–24] As an observational technique, meta-analysis requires scrupulous focus on the quality of input data. This necessitates there being sufficient homogeneity for valid combination, as heterogeneity has unpredictable consequences including biased, invalid results. Formal testing to exclude heterogeneity is an *a priori* validity criterion required for pooling data. This is hard to satisfy without data having uniform common recruitment protocols, as well as in-built internal controls (such as odds ratios from placebo-controlled, randomized trials). By these criteria, the Carlsen data comprising non-comparable studies of screened, volunteer sperm donors, men seeking vasectomy or controls in experimental studies, were highly heterogeneous and invalid for pooling into a meta-analysis.

Another major flaw was the participation bias noted previously arising from the requirement for semen analysis from unselected men from the general population. Specifically, the use of data from infertility clinics is inherently invalid as a sample of the general male population. In this singular respect, the Carlsen meta-analysis got it right—they collected studies only of non-infertile men. The participation by self-selected and screened volunteers with an inherently low rate of recruitment of non-infertile men creates intractable bias in either or both negative and positive directions, dependent on men's perception of their fertility. This was demonstrated in a study which could identify from a single city over the same period of time either stable levels of sperm output at different median levels or a falling sperm output, presumably depending on different recruitment strategies for the unpopular task of providing semen samples.[13]

The still unresolved difficulties to define population reference ranges for semen analysis make it obvious that retrospective collections of semen samples such as those collected by Carlsen could not be representative of their time or place of origin. This applies not only to the collection of single site studies assembled by Carlsen, but equally to the many multicentre studies which followed it. Although some better organized, prospective post-Carlsen studies usually had laboratory QC to ensure comparability of the semen analysis end points, none had any QC on recruitment. Hence, it remains clear that the composition of each study centre's population could not be validly compared with any other centres, given low recruitment rates and likely differences in recruitment and representativeness to their own background populations. Simply put, if study centre/clinic A1 located in city B1 and country C1 is to be compared with results from study centre/clinic A2 located in city B2 and country C2, how can one verify that they validly represent A, B or C in order to compare them? Is it realistic to accept that the inherently low, biased recruitment rates with over 80% of volunteers declining to participate can constitute an ignorable bias in such comparisons? Such multicentre comparisons, uncontrolled for the differences in recruitment success and representativeness, simply extend the single-site temporal fallacy of comparability to one involving both multisite geographical and temporal fallacies of comparability.

Before the development of semen analysis, investigating the male contribution to fertility was a disreputable concept to be tested, if at all, by the semiquantitative post-coital test which proved that sexual intercourse deposited spermatozoa in the cervical mucus. As a laboratory test, semen analysis has a relatively short history with the first paper published in 1929[25] and the first large sample studies in infertile couples in 1951[26] and among non-infertile men in 1974;[27] the publication of the second large study allowing comparison with the first, immediately created the first claim of falling sperm counts.[27] Only in 2013 have first guidelines been proposed for systematic evaluation of clinical studies that feature semen analysis.[28]

Methodological defects in the use of laboratory methods were another problematic area. For some of the period, semen analysis methods were not even standardized, let alone covered by interlab QC to ensure reproducibility between studies or locations. While large sample sizes could overcome to some degree the problems of random errors, such as those from poorly standardized lab methods, no larger sample size can overcome systematic errors from differences in methods or poor standardisation, particularly over the long time span of the study's temporal framework.

Finally, the data analysis methods of Carlsen were unsatisfactory. Although well known since the involvement of a statistician (Ruth Gold), in MacLeod's classical series of papers[29] that semen data are always skewed and require power[30] or log[31] transformation to create a Gaussian distribution before parametric statistical analysis. Yet, Carlsen used the arithmetic mean which in itself distorted the findings[32] and when a valid central measure (median) was used instead on the Carlsen data, the findings were not statistically significant.[33]

In considering the global estrogen pollution hypothesis, the proposal had at the outset a major problem with biological plausibility in that it remains hard to understand how traces of any chemicals could have adverse estrogenic effects during pregnancy, given the massively estrogenic background of pregnancy from placental steroid secretion. While not excluding other teratogenic or carcinogenic mechanisms for chemicals, this fact renders implausible prior claims of estrogenic (i.e., estrogen receptor-mediated) birth defects. Even the unquestionable transplacental carcinogenicity of diethylstilbestrol (DES) for girls[34] must involve additional non-estrogenic genetic or environmental mechanisms, as the classical vaginal adenocarcinoma only occurs in a small minority (<1 : 1000) of girls exposed to DES *in utero*.[35] The empirical nail in the coffin of this hypothesis where it concerned male reproductive function was provided by the 1995 classic DES follow-up study reporting a meticulous 40-year follow-up of boys exposed *in utero* to DES (or placebo).[36] Wilcox *et al.* reported no reduction (and possibly an increase) in male fertility, despite the mother's taking during pregnancy DES doses that were on average equivalent to twice her body weight in the equivalent of DDT as a prototype estrogenic chemical pollutant. The claims of male reproductive pathologies in some[37] but not all[38] follow-up studies may relate to the higher prevalence of preterm birth,[35] a major risk factor for cryptorchidism. The definitive Wilcox study refutation reduced all other studies of trace estrogenic chemicals on prenatal human male reproductive development to mere idle commentary.

The above factual background is provided in order to enable readers to arrive at their own conclusions, considering that we are all entitled to our private opinions but not to private facts. Bearing in mind an adaptation of Tukey's maxim that it is '…better to ask the right question, for which the answer may be vague and approximate but might always be improved upon, than to ask the wrong question for which the exact answer will always be wrong',[39] you are invited to review the comments in this issue—and to arrive at your own opinion.

1 Carlsen E, Giwercman A, Keiding N, Skakkebaek NE. Evidence for decreasing quality of semen during past 50 years. *Br Med J* 1992; **305**: 609–13.

2 Sharpe RM, Skakkebaek NE. Are oestrogens involved in falling sperm counts and disorders of the male reproductive tract? *Lancet* 1993; **341**: 1392–5.

3 Neuwinger J, Behre HM, Nieschlag E. External quality control in the andrology laboratory: an experimental multicenter trial. *Fertil Steril* 1990; **54**: 308–14.

4 Cooper TG, Neuwinger J, Bahrs S, Nieschlag E. Internal quality control of semen analysis. *Fertil Steril* 1992; **58**: 172–8.

5 Keel BA, Quinn P, Schmidt CF Jr, Serafy NT Jr, Serafy NT Sr *et al*. Results of the American Association of Bioanalysts national proficiency testing programme in andrology. *Hum Reprod* 2000; **15**: 680–6.

6 Cooper TG, Bjorndahl L, Vreeburg J, Nieschlag E. Semen analysis and external quality control schemes for semen analysis need global standardization. *Int J Androl* 2002; **25**: 306–11.

7 Jequier AM. Is quality assurance in semen analysis still really necessary? A clinician's viewpoint. *Hum Reprod* 2005; **20**: 2039–42.

8 Penn HA, Windsperger A, Smith Z, Parekattil SJ, Kuang WW *et al*. National semen analysis reference range reporting: adherence to the 1999 World Health Organization guidelines 10 years later. *Fertil Steril* 2011; **95**: 2320–3.

9 Mallidis C, Cooper TG, Hellenkemper B, Lablans M, Uckert F *et al*. Ten years' experience with an external quality control program for semen analysis. *Fertil Steril* 2012; **98**: 611–6.e4.

10 Sackett DL. Bias in analytical research. *J Chron Dis* 1979; **32**: 51–63.

11 Feinstein A. Clinical Epidemiology: The Architecture of Clinical Research. Phildelphia, PA: WB Saunders Company; 1985. p812.

12 Rothman K, Greenland S. Modern Epidemiology. 2nd ed. Phildelphia, PA: Lippincott Willams & Wilkins; 1998. p738.

13 Handelsman DJ. Sperm output of healthy men in Australia: magnitude of bias due to self-selected volunteers. *Hum Reprod* 1997; **12**: 2701–5.

14 Muller A, de la Rochebrochard E, Labbe-Decleves C, Jouannet P, Bujan L *et al*. Selection bias in semen studies due to self-selection of volunteers. *Hum Reprod* 2004; **19**: 2838–44.

15 Cohn BA, Overstreet JW, Fogel RJ, Brazil CK, Baird DD *et al*. Epidemiologic studies of human semen quality: considerations for study design. *Am J Epidemiol* 2002; **155**: 664–71.

16 Cooper TG, Noonan E, von Eckardstein S, Auger J, Baker HW *et al*. World Health Organization reference values for human semen characteristics. *Hum Reprod Update* 2010; **16**: 231–45.

17 World Health Organization. WHO Laboratory Manual For The Examination and Processing of Human Semen. 5th ed. Geneva: WHO; 2010. p128.

18 Handelsman DJ, Cooper TG. Semen Analysis in 21st Century Medicine special issue in *Asian Journal of Andrology. Asian J Androl* 2010; **12**: 7–123.

19 Sharpe RM, Skakkebaek NE. Testicular dysgenesis syndrome: mechanistic insights and potential new downstream effects. *Fertil Steril* 2008; **89**: e33–8.

20 Meinert CL. Meta-analysis: science or religion. *Control Clin Trials* 1989; **10**: 257S–63S.

21 Feinstein AR. Meta-analysis: statistical alchemy for the 21st century. *J Clin Epidemiol* 1995; **48**: 71–9.

22 Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD *et al*. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000; **283**: 2008–12.

23 Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D *et al*. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM Statement. *Onkologie* 2000; **23**: 597–602.

24 von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC *et al*. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007; **147**: 573–7.

25 Macomber D, Sanders MB. The spermatozoa count: its value in the diagnosis, prognosis and treatment of sterility. *N Engl J Med* 1929; **200**: 981–4.

26 MacLeod J, Gold RZ. The male factor in fertility and infertility. II. Spermatozoan counts in 1000 men of known fertility and in 1000 cases of infertile marriage. *J Urol* 1951; **66**: 436–39.

27 Nelson CM, Bunge RG. Semen analysis: evidence for changing parameters of male fertility potential. *Fertil Steril* 1974; **25**: 503–7.

28 Sanchez-Pozo MC, Mendiola J, Serrano M, Mozas J, Bjorndahl L *et al*. Proposal of guidelines for the appraisal of SEMen QUAlity studies (SEMQUA). *Hum Reprod* 2013; **28**: 10–21.

29 MacLeod J, Gold RZ. The male factor in fertility and infertility. VII. Semen quality in relation to age and sexual activity. *Fertil Steril* 1951; **4**: 194–209.

30 Handelsman DJ. Optimal power transformations for analysis of sperm concentration and other semen variables. *J Androl* 2002; **23**: 629–34.

31 Berman NG, Wang C, Paulsen CA. Methodological issues in the analysis of human sperm concentration. *J Androl* 1996; **17**: 68–73.

32 Bromwich P, Cohen J, Stewart I, Walker A. Decline in sperm counts: an artefact of changed reference range of "normal". *Br Med J* 1994; **309**: 19–22.

33 Handelsman DJ. Estrogens and falling sperm counts. *Reprod Fertil Dev* 2001; **13**: 317–24.

34 Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *N Engl J Med* 1971; **284**: 878–81.

35 Hoover RN, Hyer M, Pfeiffer RM, Adam E, Bond B *et al*. Adverse health outcomes in women exposed *in utero* to diethylstilbestrol. *N Engl J Med* 2011; **365**: 1304–14.

36 Wilcox AJ, Baird DD, Weinberg CR, Hornsby PP, Herbst AL. Fertility in men exposed prenatally to diethylstilbestrol. *N Engl J Med* 1995; **332**: 1411–6.

37 Palmer JR, Herbst AL, Noller KL, Boggs DA, Troisi R *et al*. Urogenital abnormalities in men exposed to diethylstilbestrol *in utero*: a cohort study. *Environ Health* 2009; **8**: 37.

38 Palmer JR, Wise LA, Robboy SJ, Titus-Ernstoff L, Noller KL *et al*. Hypospadias in sons of women exposed to diethylstilbestrol *in utero*. *Epidemiology* 2005; **16**: 583–6.

39 Tukey JW. The future of data analysis. *Ann Math Stat* 1962; **33**: 1–67.