comparison is of two groups treated concurrently, trials using inferior methods of allocation are not acceptable to the *BMJ*.[4] Assessment bias is prevented either by having an objective outcome measure or by blinding the assessor to the treatment identity (hence when both the patient and the clinician are so blinded the trial is "double blind").

Because in non-randomised, non-blind comparisons allocation and assessment bias have not been addressed methodologically an attempt is usually made to address them logically. That is, a rational argument is advanced that there is no reason for any such biases to exist or be large enough to explain the differences seen. Low P values are cited in support of the causal conclusion.

Is the *BMJ* operating a double standard in requiring much more rigorous methodology for treatment comparisons which are explicitly labelled as "research" but not applying such rigorous criteria for studies not so labelled but which attempt to draw causal conclusions? I think it is.

1 Smith R. Audit and research. *BMJ* 1992;305:905-6.
2 Penney GC, Glasier A, Templeton A. Multicentre criterion based audit of the management of induced abortion in Scotland. *BMJ* 1994;309:15-9
3 Pocock SJ. Randomised clinical trials. *BMJ* 1977;i:1661.
4 Altman DG. Randomisation. *BMJ* 1991;302:1481-2.

Decline in sperm counts: an artefact of changed reference range of "normal"?

Peter Bromwich, Jack Cohen, Ian Stewart, Andrew Walker

Abstract

***Objective*—To investigate a reported fall in sperm counts during 1940-90 in relation to the reduced lower reference value of "normal" during the same period by assuming the null hypothesis that no change had occurred in the probability distribution of the sperm concentration.**

***Design*—Analysis by using various mathematical models of the probability distribution of sperm concentration together with experimental data which supported a model employing a logarithmic distribution.**

***Subjects*—235 men presenting for stimulated in vitro fertilisation at Midland Fertility Services, Aldridge, in 1992 together with samples of 20 ejaculates from each of five men attending the same centre during 1992-3.**

***Results*—The effect of the change in lower reference value for the "normal" sperm concentration (from 60×10^9 to 20×10^9/l) depended on the probability distribution of the concentration in the population. If that distribution was normal or uniform, then very little of the reported decline was a consequence of the change in lower reference value. If it was heavily skewed, then most or all of the reported decline may have been a consequence of that change. The limited experimental data available indicate that the distribution was heavily skewed.**

***Conclusions*—Depending on the actual distribution of sperm concentration in the population, the reported decline in concentration may have been accounted for entirely or in part by the change in lower reference value. The original evidence does not support the hypothesis that the sperm count declined significantly between 1940 and 1990.**

See editorial by Farrow

Midland Fertility Services, Court Parade, Aldridge WS9 8LT
Peter Bromwich, *medical director*
Andrew Walker, *embryologist*

Centre for Ecosystems Analysis and Management, University of Warwick Coventry CV4 7AL
Jack Cohen, *visiting senior research fellow*

Nonlinear Systems Laboratory, Mathematics Institute, University of Warwick, Coventry CV4 7AL
Ian Stewart, *professor*

Correspondence to:
Dr Cohen.

BMJ 1994;309:19-22

Introduction

Carlsen *et al* performed an extensive analysis of historical data on human sperm concentrations.[1] By using linear regression analysis on 61 different sets of data obtained between 1938 and 1990 they reported a significant ($P<0.0001$) decrease in mean sperm concentration—from 113×10^9/l in 1940 to 66×10^9/l in 1990, a decline of more than 40%. Their conclusion received widespread recognition, including coverage by the media.[2] They also reported a marginal ($P<0.03$) decrease in mean seminal volume, from 3·4 ml in 1940 to 2·8 ml in 1990. As further support of the hypothesis of a decline in sperm quality, they noted that "the lower reference value for a 'normal' human sperm count has changed from 60×10^6/ml in the 1940s to the present value of 20×10^6/ml." Note that their use of the term "sperm count" refers to concentration, not total numbers of sperms in an ejaculate, and in keeping with other authors we occasionally use it in this sense below.

In order to avoid bias, Carlsen *et al* restricted their study to men with proved fertility (39 studies) or "normal" men of unknown fertility (22 studies). However, the mean sperm concentration is very sensitive to the form of the probability distribution for sperm concentration, either in individuals or across the population. It is also sensitive to cut offs introduced by selection of subjects. Similar comments apply to other measures of fertility, such as the total number of sperms in an ejaculate, but in keeping with Carlsen *et al* we restrict attention to the concentration.

In particular, the reduction of the lower reference value for a normal sperm concentration is likely to lower the observed mean, because men with mean sperm concentrations between 20×10^9/l and $60\times^9$/l would tend to be excluded in studies performed in the 1940s but be included in later studies.

To what extent might the apparent decline reported by Carlsen *et al* be a consequence of the change in the lower reference value? If lower sperm counts are more probable than high ones, then discarding subjects with low sperm counts has a disproportionate effect on the mean. Thus the effect of such a change depends on how the probability distribution of the sperm concentration interacts with the effects of averaging over accumulated results for individual patients or over accumulated results for a population.

Models

The usual statistical models employed to analyse these questions are probability distributions, usually defined by a probability density function.[3] The distribution is normal if its probability density function is the usual bell shaped curve[4] and uniform if the probability density function is constant within some range and zero outside it. It is a power law distribution if the probability density function is proportional to X^{-s} for some positive constant s. When $s=1$ such a distribution is said to be logarithmic, because the cumulative probability distribution is given by a logarithm. Because the total probability must be 1, power law distributions must be reduced to zero at some upper cut off when $s<2$ or otherwise modified. Power law distributions, and in particular logarithmic ones, are heavily skewed towards lower values.

Surprisingly little is known about the probability distributions of sperm concentrations, either for individuals or for populations. The effect on the mean of a change in the lower reference value depends sensitively on the underlying distribution. To illustrate this dependence we have considered four hypothetical cases: uniform, normal, logarithmic, and power law distributions.

For this analysis we assume a simple model as null hypothesis. This is that the probability distribution of sperm concentration across the male population remained unchanged between 1940 and 1990; that men with a concentration below 60×10^9/l were rejected in the 1940s; but that only men with a concentration below 20×10^9/l were rejected in the 1980s-90s. In practice the effect of exclusion would not be as clear cut, because individual variability implies that men who are included in a trial may subsequently produce ejaculates whose concentrations lie in the excluded range. This effect could be incorporated into a refined model. It would have a moderate quantitative effect on our calculations, but we would not expect it to change the main conclusions.

We analyse the effect of the change in lower reference value as follows. Given a probability distribution, we first cut it off at a lower limit of 60×10^9/l and adjust parameters to reproduce the 1940 mean of 113×10^9 obtained by Carlsen *et al.* We call this the "assigned" mean. Leaving the parameters unchanged, we then include data values between 20×10^9/l and 60×10^9/l and recompute the mean. We refer to this as the "predicted" mean appropriate to the lower reference value adopted in 1990. The table gives the results. Mathematical details are available separately from JC.

If the distribution is logarithmic, then the predicted 1990 mean is 76×10^9/l, so that the change of the lower reference value accounts for almost all of the apparent decline in mean sperm concentration. With uniform or normal distributions the change in lower reference value cannot account for the apparent decline. For some power law distributions the change in the lower reference value leads to a bigger decline than that reported by Carlsen *et al.*

In general, any distribution that is heavily skewed towards smaller values will lead to a disproportionate decrease in the mean concentration in response to the change in the lower reference value without any actual change in the distribution of sperm concentration across the population. This is the main point of our argument, and it is not affected by fine details of models.

Predicted change in observed mean sperm concentration between 1940 and 1990, assuming no change in underlying distribution and fitting 1940 value

Distribution	1940 Mean (assigned)	1990 Mean (predicted)
Uniform	113×10^9/l	93×10^9/l
Normal*	113×10^9/l	93×10^9/l or higher
Power law ($0\geq s>1$)	113×10^9/l	76×10^9/l or higher
Logarithmic ($s=1$)	113×10^9/l	76×10^9/l
Power law ($1\geq s>2$)†	113×10^9/l	38×10^9/l—76×10^9/l
Power law ($s\geq2$)	113×10^9/l	38×10^9/l

*Value for normal distribution is stated as lower limit because it depends on SD.
†Values for power laws depend on s when $1\geq s>2$.

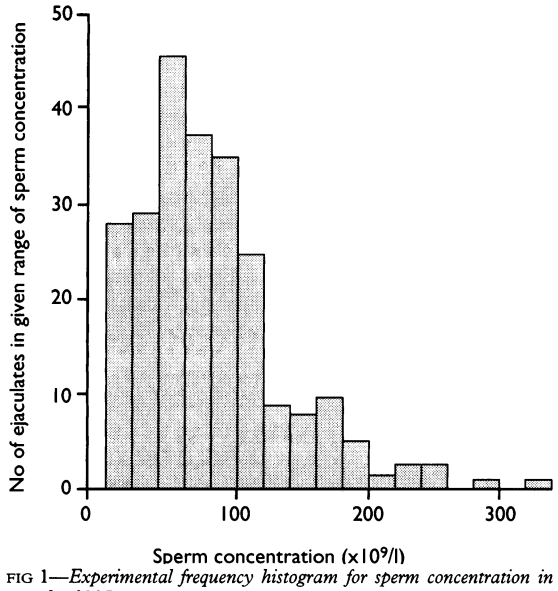

FIG 1—*Experimental frequency histogram for sperm concentration in sample of 235 men*

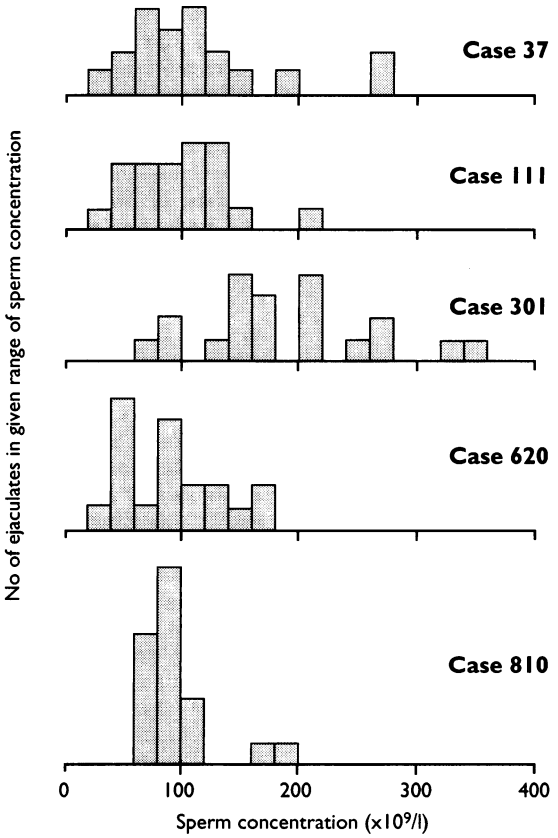

FIG 2—*Experimental frequency histograms for sperm concentration in series of 20 samples from five men*

Experimental data

How are sperm concentrations actually distributed? Because of the widespread use of the mean sperm count (or concentration) as a clinical variable many papers have reported data on mean values. However, information on the distribution about the mean is scarce. Figure 2 in Carlsen *et al* (our fig 4 (left)) summarises how percentages of men with sperm concentrations in five ranges changed between 1930 and 1990. We show below that their data are consistent with our hypothesis that there has been no change in the distribution of sperm concentration in the population. Cohen considered possible distributions.[5] The folklore observation that "twice as many sperms is as likely as half as many" suggests that the distribution is logarithmic. In order to offer (limited) experimental support for the skewness of sperm concentration data we shall discuss new data supplied by Midland Fertility Services, Aldridge. These data are of two kinds: (*a*) sperm concentrations for 235 men presenting for stimulated in vitro fertilisation at Midland Fertility Services, Aldridge, in 1992; (*b*) sperm concentrations in 20

ejaculates from each of five selected men attending the same centre during 1992-3.

DISTRIBUTION FOR POPULATIONS

Figure 1 shows the observed distribution of sperm concentrations. Figure 1 is drawn as a histogram, whose shape provides an approximation of the probability density function. To do this, the possible values of the concentration have been divided into intervals, or "bins," and the number of men whose concentration falls in a given bin is plotted as the height of the corresponding vertical bar. The histogram clearly possesses a key feature of logarithmic and power law distributions. It is highly skewed towards low values.

DISTRIBUTION FOR INDIVIDUALS

Figure 2 shows data consisting of 20 concentration measurements performed on five different men (identified by case numbers at Midland Fertility Services, Aldridge). These men were selected as potential donors and thus may have had higher mean sperm counts and concentrations than would be typical of the population as a whole. The distribution of sperm concentrations for individuals was highly skewed towards lower values. Because the number of observations was small the skewness was shown by the thinning out of bars at higher values as well as in the peaks at lower values. When the data were pooled (fig 3) the skewness became more apparent. For this analysis the relevant feature of figure 2 is that it shows considerable variability within individuals. The precise nature of that variability cannot be deduced with confidence from the limited data presented here, but it is likely to be skewed in the same general manner as the pooled data.

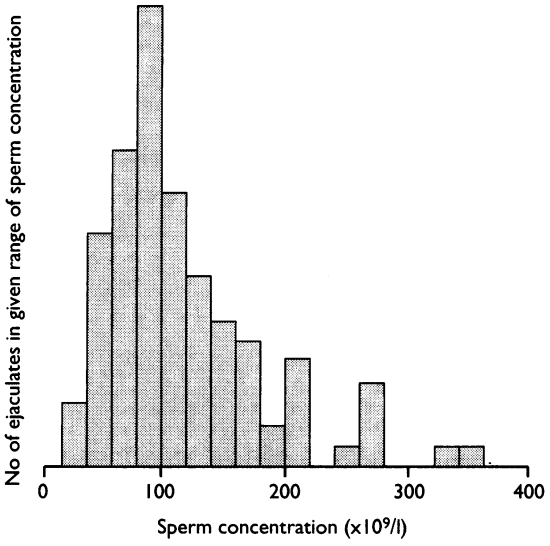

FIG 3—*Pooled data from figure 2*

Figure 2 in Carlsen *et al* (our fig 4 (left)) shows how the percentages of men in their study whose sperm concentrations fell into one of five ranges varied between 1940 and 1990. The ranges were <20, 20-40, 41-60, 61-100, and >100 (all $\times 10^9$/l). There was a dramatic decline in the percentage for the highest range from 50% in 1940 to 16% in 1990. However, this apparent decline is consistent with the hypothesis that the distribution of sperm concentration had not changed but that the lower cut off had. If we assume a logarithmic distribution with an upper cut off of 190×10^9/l, as in the table, then figure 4 (right) shows how the percentages are predicted to change, assuming our logarithmic model, when the lower cut off is reduced from 60×10^9/l to 20×10^9/l. With a lower cut off of 60×10^9/l, the percentage in the highest range is 55%, but this drops to 28% when the lower cut off is changed to 20×10^9/l. It is clear why: the inclusion in 1990 of large numbers of men with low sperm concentrations, who would have been excluded in 1940, drastically reduced the percentage in the higher ranges. There is an especially dramatic effect on the highest range, which is very broad.

Our simple model assigns zero percentages to all ranges below the 1940 and 1990 cut offs. In contrast, figure 4 (left) shows non-zero percentages below these cut offs. The discrepancy can be explained in terms of individual variability, as discussed above. A man may on one or more occasions produce a sperm concentration high enough to be included in a study, even though his mean concentration is below the lower reference value that is in force. Some subsequent observations may then be lower than this value. The result would be to spread out the lower cut off assumed in our model.

These data suggest that logarithmic or power law distributions provide more appropriate models for sperm production than normal or uniform distributions. This break with tradition can to some extent be justified biologically. For example, we have found simple models of sperm recruitment to ejaculates that generate various power law distributions. Details of these are also available separately from JC.

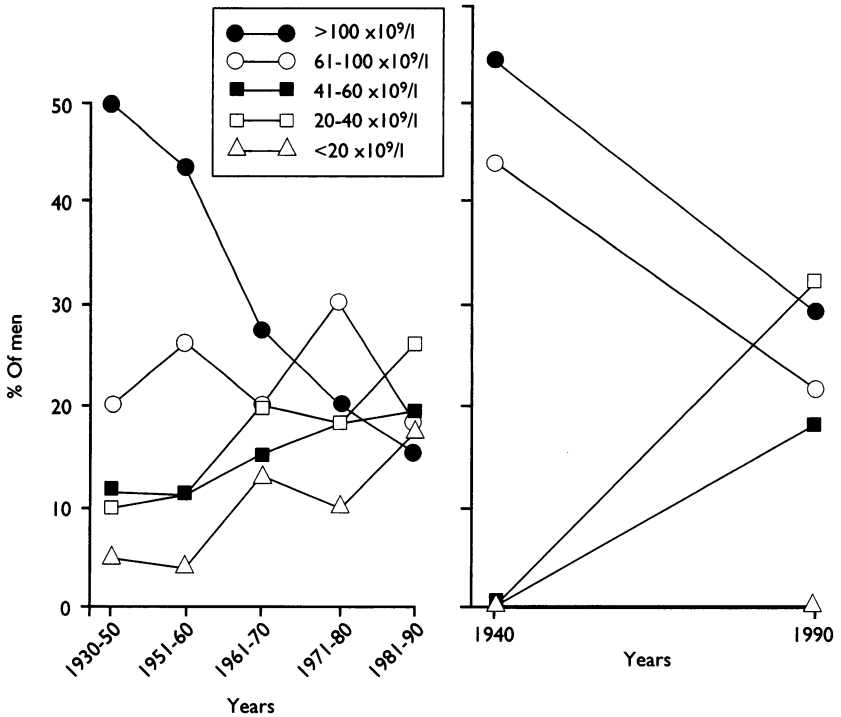

FIG 4—*Changes in percentages of men with sperm concentrations in ranges <20, 20-40, 41-60, 61-100, and >100 (all x10⁹/l) between 1940 and 1990. Left: historical data from Carlsen et al.[1] Right: Theoretical changes predicted by logarithmic distribution with upper cut off 190x10⁹/l.*

Discussion and conclusions

The data from Midland Fertility Services, Aldridge, support the hypothesis that the probability distribution of sperm concentration, both in individuals and in populations, resembles a logarithmic distribution. In particular, they are heavily skewed towards lower values and do not resemble either a normal or a uniform distribution. The observed mean value for sperm concentration is therefore very sensitive to the choice of cut offs at lower values—that is, the lower reference value for a normal sperm count (measured as a concentration). The analysis summarised in the table and figure 4 indicates that nearly all of the observed decline in mean sperm count may be a consequence of the reduction of the lower reference value and that the evidence presented by Carlsen *et al* for a decline in sperm quality is unconvincing.

Clinical implications

- Sperm concentrations in successive samples from one man, and aggregate data from many patients, are highly skewed and closer to a logarithmic distribution than a normal distribution
- The evidence for a long term decline in sperm concentrations, based on historical data, is unconvincing
- Lower reference values of normal (of 60×10^9/l or 20×10^9/l) should not be applied uncritically
- The pattern of individual variability means that averages may be poor measures of fertility
- Geometric means may be more appropriate clinical variables than arithmetic means but are unreliable and require validation

Similar reasoning applies to any sufficiently skewed distribution, so we would not expect improved data to change the general line of our argument. However, a decline that was considerably smaller than that reported by Carlsen *et al* could be consistent with our analysis and might be detectable with confidence, given better data. More extensive data are needed to establish with greater precision the probability distributions of sperm concentration in populations and in individuals.

It is standard to use arithmetic means of sperm counts and concentrations as clinical variables. However, if the hypothesis of near logarithmic distributions is confirmed, then the geometric mean would be a more appropriate statistic.

The level of significance ($P<0{\cdot}0001$) reported in the linear regression analysis of Carlsen *et al* represents only the confidence that the observed mean has changed. It does not indicate the cause of that change. It can be accounted for by a change in the lower reference value for normal sperm count, provided that the distribution for sperm production is sufficiently skewed towards lower values. In particular, a change in sperm concentration from 113×10^9/l to 76×10^9/l can be entirely accounted for in this way by using a logarithmic distribution, which is supported by the available data. The remaining discrepancy between 76×10^9/l and 66×10^9/l is unlikely to be significant.

Instead of confirming the apparent decline in sperm count, as Carlsen *et al* assert, the change in lower reference value may well be responsible for it.

1 Carlsen E, Giwercman A, Keiding N, Skakkebæk NE. Evidence for decreasing quality of semen during past 50 years. *BMJ* 1992;**305**:609-12.
2 *Horizon* "Assault on the male" BBC2, 31 Oct 1993.
3 Feller W. *An introduction to probability theory and its applications*. New York: Wiley, 1957.
4 Moore DS, McCabe GP. *Introduction to the practice of statistics*. New York: Freeman, 1993.
5 Cohen J. The comparative physiology of gamete populations. In: Lowenstein O, ed. *Advances in comparative physiology and biochemistry*. Vol 4. New York: Academic Press, 1971:267-380.

(Accepted 11 April 1994)

Commentary

Importance of empirical evidence

Niels Keiding, Aleksander Giwercman, Elisabeth Carlsen, Niels E Skakkebæk

Bromwich *et al* point out that the distribution of sperm count is skewed to the right and that if a differential selection of skewed distributions is applied over the years this will bias the observed time trends. Both of these assertions are correct; indeed, in all 16 of the 61 publications cited in our original overview for which both median and mean were given the median was smaller than the mean, confirming the skewness.[1]

Bromwich *et al* present some elaborate, although rather elementary, statistical points about skewed distributions and differential selection from these, but they fail to give any empirical reference that might support their suggestion of differential selection. One possibility is that they believe that the lower reference values for sperm counts of 60×10^6/ml in the 1940s and 20×10^6/ml at present had been used as truncation values for the reported distributions over the years. If Bromwich *et al* had actually studied the reports they would have found that there were plenty of values under these limits in even the oldest articles. The article by MacLeod and Gold in 1951, based on 1000 men, is particularly important in this respect.[2] This early paper is largely responsible for the high historical values and is thus responsible for a considerable part of the observed decline. However, the authors of this paper were surprised about the low values contained in it. This paper was presumably the first to explicitly mention that it is "obvious to many that this figure [60×10^6/ml] is too high."

There are many problems with historical overviews (meta-analyses), but the article by Bromwich *et al* amounts to discussing time trends detached from the relevant empirical evidence. Thus, the most cautious conclusion that can be drawn from the existing data is still that semen quality has declined significantly between 1940 and 1990. After several years of published evidence being ignored, the increasing incidence of abnormalities of male genital organs (including a highly significant increase in incidence of testicular cancer[3]) has finally attracted the attention of the scientific world. We hope that the paper of Bromwich *et al*, which is apparently based on wrong assumptions, will not bring confusion or divert attention from the urgent need for more research into the threat to male reproductive functions.[4]

Statistical Research Unit, University of Copenhagen, Panum Institute, DK-2200 Copenhagen, Denmark
Niels Keiding, *professor*

University Department of Growth and Reproduction, Rigshospitalet 5064, DK-2100 Copenhagen
Aleksander Giwercman, *senior registrar*
Elisabeth Carlsen, *research fellow*
Niels E Skakkebæk, *professor*

1 Carlsen E, Giwercman A, Keiding N, Skakkebæk NE. Evidence for decreasing quality of semen during past 50 years. *BMJ* 1992;**305**:609-12.
2 MacLeod J, Gold RZ. The male factor in fertility and infertility. II. Spermatozoön counts in 1000 men of known fertility and in 1000 cases of infertile marriage. *J Urol* 1951;**66**:436-49.
3 Møller H. Clues to the aetiology of testicular germ cell tumours from descriptive terminology. *Eur Urol* 1993;**23**:8-15.
4 Skakkebæk NE. Giwercman A, de Kretser D. Pathogenesis and management of male infertility. *Lancet* 1994;**343**:1473-9.

Correction

Management of female prisoners with abnormal cervical cytology

An authors' error occurred in this paper by G P Downey *et al* (28 May, pp 1412-3). P Curtis, senior registrar in obstetrics and gynaecology at the Royal Free Hospital, was omitted from the list of authors. The authors of this paper are therefore G P Downey, G Gabriel, A R S Deery, J Crow, P Curtis, and P G Walker.