Misuse of correlation and regression in three medical journals

Alan M W Porter MD PhD

J R Soc Med 1999;**92**:123–128

SUMMARY

Errors relating to the use of the correlation coefficient and bivariate linear regression are often to be found in medical publications. This paper reports a literature search to define the problems. All the papers and letters published in the *British Medical Journal*, *The Lancet* and the *New England Journal of Medicine* during 1997 were screened for examples. Fifteen categories of errors were identified of which eight were important or common. These included: failure to define clearly the relevant sample number; the display of potentially misleading scatterplots; attachment of unwarranted importance to significance levels; and the omission of confidence intervals for correlation coefficients and around regression lines.

INTRODUCTION AND DEFINITIONS

Errors involving the relationship between two quantitative variables are often to be found in medical publications. This paper reports a literature search from three medical journals to define problems relating to correlation and bivariate regression and discusses the correct use and reporting of the two statistics. Full treatment, descriptive and mathematical, of correlation and regression is to be found in statistical textbooks and will not be attempted here. Attention will be confined to relevant definitions.

Correlation

The Pearson correlation coefficient is a summary statistic which measures the direction and magnitude of the association or 'co-relation' between two variables measured on an interval scale, each approximately normally distributed and together having an approximate linear relationship. The Spearman correlation coefficient is the comparable non-parametric ranking statistic for data where at least one of the variables is measured on an ordinal scale or does not form an approximate normal distribution on an interval scale. A correlation coefficient gives an indication of the closeness of the relationship. The null hypothesis, that the sample correlation coefficient has come from a population with a correlation coefficient of zero, may be tested by Student's t statistic. The formula has n in the numerator and r in the denominator. This means that as n increases so does t for the same value of r. Thus a very weak and probably meaningless correlation may be significant if the numbers are large enough. For example, the near-zero correlation coefficient of 0.06 is conventionally significant (P=0.05) if the sample is 1000, and the weak correlation r=0.27 is similarly significant for a sample of 50. Tests of statistical significance involve, therefore, both the magnitude of the observed association and the sample size. It follows that in the context of the biological sciences the probability levels of correlation coefficients are often misleading and should usually be ignored except when numbers are small. Conventional significance indicates that an association is probably real but the question always must be, is the relationship strong enough to be important and meaningful?

Various concerns have been recorded about the dichotomous nature of conventional significance in clinical research[1,2]. Since 1988 medical journals have required authors to place less reliance on probability levels and to use confidence intervals when appropriate[3], but implementation of this policy has not been extended to correlation coefficients. The confidence interval (CI) for a sample correlation coefficient presents an interval with upper and lower limits within which the true population correlation coefficient will probably lie and indicates the extent of the uncertainty attaching to the estimate. The estimation involves use of Fisher's z-transformation (for the mathematical treatment see Altman and Gardner[4]). Few statistical software programs produce the CI output for a correlation coefficient by default.

Calculation by hand is simple and involves the use of only three tables[5] and addition and subtraction. When numbers are small the sampling variation for a correlation coefficient is surprisingly large and the CI correspondingly wide. For example, if n=10 and r=0.7 then the 95% CI spans most of the range of zero to one (0.13 to 0.92). If the sample is doubled to 20 then the 95% CI is still wide (0.37 to 0.87), if further doubled to n=40 it is 0.50 to 0.83.

Redcrest, Heath Rise, Camberley, Surrey GU15 2ER, UK

A recorded correlation coefficient should usually be linked with three other parameters—the sample size, the 95% CI and the probability level. The CI may reasonably be omitted if the correlation is near zero or if the correlation is strong and the sample size large. The *P* value and the confidence interval are not alternatives as has been implied[6].

Regression

Linear regression and correlation are alternative ways of examining the relationship between variables. There are close mathematical relations between the two[7], but their purposes are distinct[4]. Both techniques refer to a linear relationship between two variables. Regression, however, can be extended to non-linear relationships and to more than two variables. The use of these options will not be examined in this paper.

Regression is a method of estimating a numerical relationship. The two variables involved in a regression are the dependent (or outcome) variable (y), which is scaled conventionally on the ordinate (upright axis) of a scatterplot, whereas the independent (or predictor) variable (x) is scaled on the abscissae. Regression is both descriptive and predictive, whereas correlation is only descriptive. The correct use of each depends on whether 'relationship' or 'relationship and prediction' is the primary aim of the investigator. The nature of the descriptive element in each is, however, different. Whilst correlation only indicates the closeness or otherwise of the relationship, the regression coefficient will allow the amount of positive or negative change in the dependent variable to be related to a unit increase or decrease in the independent variable. Regression also provides a means of prediction of the value of variable y which corresponds to a given value of variable x and calculation of the associated confidence interval. The information yielded about the relationship by regression is in addition to that yielded by the correlation coefficient but does not substitute for it. It is good practice always to check and record that the usual assumptions associated with linear regression (linearity; normal distribution of residuals with a mean of zero and a constant variance) have not been greatly violated[7].

Regression is closely related to correlation. If there is no linear relationship between x and y (r=0), then the slope of the regression line is zero. The hypothesis may be tested by taking the ratio of the slope to its standard error as a t-statistic. The square of the correlation coefficient (r^2), or the 'coefficient of determination', represents the proportion of the variability which is explained by the regression model. It is a measure of the goodness of fit of a particular model. Thus if the correlation coefficient is 0.6 then 36% of the variability is explained by the model. If there is a linear relationship between two variables a straight line through the points can be used to summarize the data. The model for a simple linear regression relating y and x is characterized as:

$$y=a+bx$$

where a is the intercept (value for y when x=0) and b is the slope or regression coefficient (amount by which y increases for unit increase in x). Methods of estimating a and b are discussed below.

There is one caveat about the use of regressions which is often ignored. It is wrong to extrapolate a regression line derived from one cloud of data points (an array) to a distant and outlying data point or to a second array beyond and distinct from the domain of the first array[8]. If numbers are small the parameters of a regression equation may be highly influenced by the inclusion of an outlier[9], which then qualifies as an 'influential point'. A regression model should be used only for that part of the data that excludes an influential outlier[10] or alternatively a lognormal or loglog scale should be used to draw the outlier into the array. The same caution should be exercised if a correlation coefficient is derived for a cloud of points which includes an outlier (Figure 1c). The statistical literature is not, and cannot be, specific about the definition of an outlier[8], which is more easily recognized than defined. The inclusion or exclusion of an outlier by researchers, after careful examination and consideration of the data, should always be explained and justified.

Use of regression is rare in medical papers though the term is often erroneously used in respect of correlation.

Line-fitting

There are three main ways in which a regression line may be fitted through a cloud of points—use of the least square regression (LSR), the major axis (MA) or the reduced major axis (RMA)[7]. The choice of regression model is not a trivial consideration as the different techniques usually yield different regression equations unless the correlation coefficient is high, when the slopes of the three lines converge.

The question of the best line to use in different circumstances is far from resolved and agreed among biometricians. Many, however, are now using the LSR in all circumstances[9], though this may reflect in part a failure to appreciate its erroneous assumptions or lack of knowledge of alternatives[11]. The anthropological and biometrical literature contains many anxious discussions and analyses of the issues (see Smith[9] for a bibliography). The medical literature, however, is silent and medical researchers are seemingly unaware that a problem exists. It must be suspected that all the statistical programs used by medical researchers compute the LSR, but in the medical papers

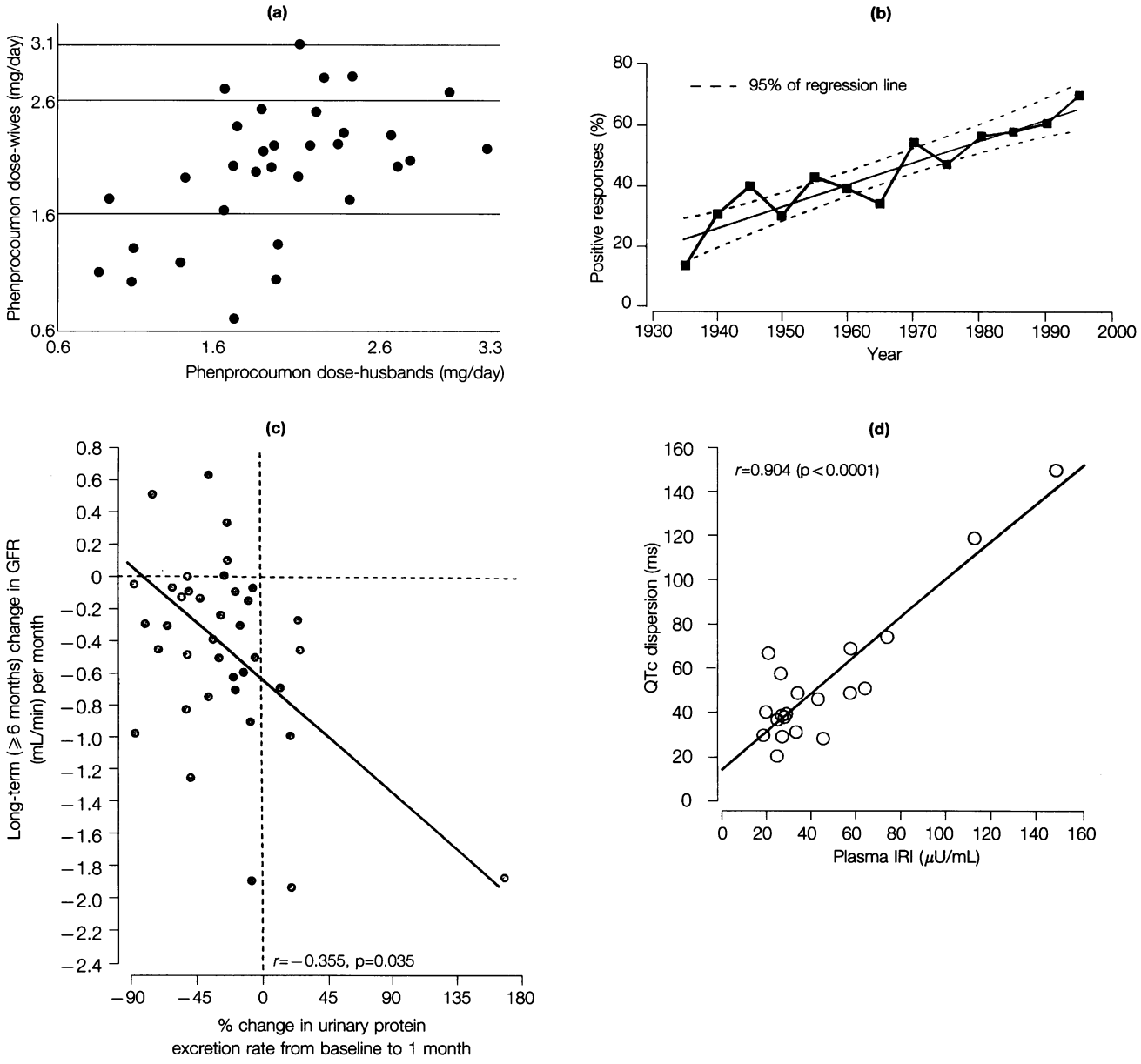

Figure 1 **Four illustrative plots taken from the papers reviewed, with original legends omitted.** (*a*) As only correlation is involved a regression line has correctly not been drawn through the points[32]. The plot enables the reader to note that the relationship is probably linear and that there are no outliers. Details of the correlation coefficient (*n*, *r*, CI, *P*) need to be added either beside the plot or in the legend. (*b*) A regression equation is given in the legend and the regression line is legitimate, but its nature (LSR, MA, RMA) has not been defined[22]. Unremarked heteroscedasticity (see text) is evident. (*c*) The plot shows an irrelevant regression line and a distant outlier[20]. Though a significant correlation coefficient of $r=-0.36$ is recorded, the authors' claim of an inverse correlation between the two variables is doubtful, for if the outlier is excluded then $n=35$, $r\simeq-0.22$, $P=0.2$. (*d*) The authors have given some information about the correlation coefficient beside the plot but have failed to record the number of data points or to calculate and record the CI[21]. The regression line is irrelevant and the two outliers are unexplained. It is possible that the positive association between the two variables is only weak. The plots are published by courtesy of the *BMJ* and *The Lancet*

searched (see below) the matter has not once been defined. The differences and indications for the three methods of line-fitting will not be discussed here.

Scatterplots

Preliminary plotting of bivariate data allows the nature of the relationship between two variables to be examined and, if correlation or regression statistics are contemplated, confirms that the relationship is probably linear. Three main types of scatterplot are to be found in published papers: (1) a plot with data points but without a regression line (Figure 1a); (2) a plot with data points and with a regression line (Figure 1b,c,d); and (3) a plot with data points with a regression line and with the boundaries of the confidence interval drawn (Figure 1b).

For a correlation, printing a plot merely offers a visual representation, sometimes misleading, of the correlation coefficient. The judged association of the two variables in a plot by the reader is vulnerable to perceptual or 'display' factors relating to the scaling and the size and orientation of the cloud of points[12]. There is a tendency for the judged association to increase as the absolute size of the cloud of points decreases. It will be even more misleading if a regression line, irrelevant when only correlation is involved, is drawn through the data points (Figure 1c,d). In this role such a line deceives the eye into an illusion of definitiveness and precision when the underlying relationship may well be weak and obscure.

The Pearson correlation coefficient and linear bivariate regression are relatively uncommon statistics in medical publications, but their correct use is important. This paper scrutinizes published papers from three medical journals for examples of their use and misuse.

METHODS

All the papers and letters from three weekly general medical journals, the *British Medical Journal* (*BMJ*), *The Lancet* and the *New England Journal of Medicine* (*NEJM*), published in 1997 were searched for examples of correlation and bivariate linear regression either in the text or in one or more of any plots. Qualifying papers and letters comprise the data set[13–36] (*BMJ* n=13, *Lancet* n=5, *NEJM* n=6).

RESULTS

Fifteen errors were identified and are listed below. Eight of these errors are important because either they occur frequently or else they are major. The frequency of occurrence of all errors is recorded for each of the three journals with examples and sometimes comments.

Important errors

Error 1

Failure to state clearly in the text the number of cases used in a correlation coefficient (*BMJ*=3)[15–17].

Error 2

Citing a correlation coefficient without giving the 95% confidence interval when appropriate (*BMJ*=6, *Lancet*=4, *NEJM*=5)[13,15,16,18–21,23,25–29,31,32]. This error occurs often. There was only one paper where the CIs were given[33] and these were not irrelevant since the correlations were all near zero.

Error 3

The inappropriate use in a scatterplot of a regression line through an array when only correlation is involved (*BMJ*=2, *Lancet*=4, *NEJM*=5)[13,15,18–21,25–29] (Figure 1c,d). The insertion of a line through a cloud of points is again a frequent error, but there were authors who correctly did not draw one (*BMJ*=6)[16,17,23,24,31,32] (Figure 1a).

Error 4

Failing to explain and justify the inclusion of one or more outliers in a plot and computations (*BMJ*=1, *Lancet*=4, *NEJM*=3)[18–21,24,27–29] (Figure 1c,d).

Error 5

Use of a correlation coefficient when numbers are very small or when plotting of the data indicates a probable non-linear trend (*BMJ*=1)[36].

Error 6

Use of the Pearson correlation coefficient when the Spearman is more appropriate (*BMJ*=3)[13–15]. In the three examples each contains an ordinal measurement.

Error 7

The appearance of apparent heteroscedasticity in a plot (a progressive increase or decrease of the spread of the residuals around the regression line) with no comment in the text (*BMJ*=2, *Lancet*=1, *NEJM*=1)[15,22,23,25] (Figure 1b). This violates the equality of variance assumption of a regression analysis.

Error 8

Attaching undue importance to a significant outcome in the context of correlation (*BMJ*=2, *Lancet*=2, *NEJM*=2)[18,20,24,25,27,36].

Examples

'We found a significant linear relation ... (P=0.002)'[18]. Comment: n=53, r=−0.44 (95% CI −0.20 to −0.64). If an outlier is discounted the correlation coefficient and lower limit of the CI would be even smaller.

'... percentage reduction in proteinuria was inversely correlated with decline in GFR (P=0.035)'[20]. Comment: n=36 r=−0.36 (95% CI −0.03 to −0.61). If the extreme outlier is discounted (Figure 1c) then $r \simeq -0.22$, P=0.2.

'All pairs of measures were highly correlated (all P values <0.0001)'[24]. Comment: inspection of the correlation matrix shows that for two of the ten relationships r=0.59 and −0.66, which are only 'moderate' and not 'high' correlations. The sample numbers are not clear and thus the CIs cannot be calculated.

'There was a significant positive correlation between the plasma level of HIV RNA and the resting energy

expenditure ... (r=0.404, P=0.011)'[25]. Comment: for n=36 the 95% CI for r is 0.09 to 0.65.

'... the degree of reduction in PAI-1 levels correlated significantly with the degree of increase in D-dimer levels (r=−0.541 P=0.002)'[27]. Comment: for this relationship if n=24 then the 95% CI is −0.18 to −0.78.

'Plasma ascorbate concentration decreased with increasing age (r=−0.17)' and again '... also correlated with the dietary intake of carotenes (r=0.159)'[36]. Comment: such weak correlations are largely meaningless even with large numbers of cases (n=1605). Some sort of association between any two biological variables taken from a large sample may usually be anticipated.

The above authors made more robust claims than their data supported. The reason for this was the attachment of undue importance to the probability level and the failure to calculate confidence intervals and note their wide range.

Other errors

Error 9

Failure to enter, adjacent to a plot or in the legend of the plot, the number of data points (*BMJ*=4, *Lancet*=4, *NEJM*=5)[15–19,21–23,25–29]. This omission obliges the reader to make a tedious and often uncertain search in the text for the information or else try to count the number of data points in the plot.

Error 10

Confusion between correlation and regression (*Lancet*=1)[18]: 'To test whether changes in mean HbA_{1c} percentages. ... were related to changes in mean serum IGF-1 concentrations, we did linear regression analysis on these data .... We found a significant linear relation'[18]. A plot is displayed with an irrelevant regression line through the array. A correlation coefficient is given but no regression equation. The computation relates to correlation but not regression. This error is often implied by the use of a regression line in a plot when regression is not involved (error 3).

Error 11

Justifiable use of a scatterplot to demonstrate a relationship but no details of the correlation coefficient given (*BMJ*=1)[17].

Error 12

Failure to state the type of regression line (LSR, MA, RMA) drawn through an array. The line was always undefined whether used inappropriately (error 3) or appropriately (*BMJ*=1, *Lancet*=1)[22,34] (Figure 1b). The LSR line was probably used in all examples.

Error 13

Failure to draw the hyperbolic confidence interval lines on either side of the regression line when a regression line has been appropriately drawn through an array (*BMJ*=1)[34].

Error 14

Parallel confidence interval lines, or one parallel line and one hyperbolic one, around a regression line (*NEJM*=1)[35]. Comment: CI lines are hyperbolic around a regression line, reflecting greater uncertainty at the extremes of the distribution (Figure 1b).

Error 15

Comparison of the slopes of two undefined regression lines by inspection rather than by use of the analysis of covariance statistic (*NEJM*=1)[26].

Errors not found in this data set included combining data from discontinuous arrays and the use of the correlation coefficient to test the extent of concordance between original and replicated measurements or two methods of measuring the same variable. Bland and Altman[37] have emphasized the difference between 'agreement' and 'correlation'. The two are not necessarily the same. An intraclass coefficient of reliability[38] is preferable to the correlation coefficient in this context.

CONCLUSIONS

Bivariate linear regressions are not often used in medical research and most of the above problems attach to the use of correlation coefficients. Some mistakes continually recur. The first is a puzzling disregard of confidence intervals for correlation coefficients. With three exceptions the CIs for correlation coefficients should always be calculated and entered. The exceptions are very weak or very strong coefficients and large sample numbers. A plot or its legend, therefore, should contain details of the four parameters relating to a correlation coefficient (n, r, CI, P). The second is the omission, either in the text or in a plot, of the sample size applying to a particular correlation coefficient. The relevant number should always be clearly defined. The third common error is to accept and print a default output from a statistical program which draws a regression line through a cloud of points when only correlation is involved. The fourth is confusion about the difference between correlation and regression and their correct use. The fifth relates to outliers. All outliers should be explained and justified and it is good practice to recompute and record the correlation coefficient and its CI with the outlier(s) omitted. Confronted with a plot and a line extended to an outlier, such as Figure 1c, the reader should place a thumb over the outlier[39], mentally remove the line and then decide what if anything is demonstrated by way of a relationship. The sixth

and most important mistake relates to paying undue attention to a probability level when interpreting a correlation coefficient. This widespread preoccupation with probability levels often leads to unwarranted conclusions.

A scatterplot to display visually the relationship between two variables is reassuring for the reader as it permits a judgment to be made about linearity, the strength of any association and the existence of one or more outliers. With strong relationships a plot is probably unnecessary, provided that the authors confirm linearity in the text. A line should never be drawn through an array unless the authors are using regression either to define the incremental changes of the dependent variable with changes of the independent variable or else to fulfil the predictive role defined earlier. In this case the regression equation should always be shown in the text, beside the plot or in the legend. The line used should be defined (LSR, MA or RMA).

The findings of this study suggest that editors and referees should demand a more disciplined, standardized and structured approach to the use and presentation of correlation and linear bivariate regression statistics in medical research.

Acknowledgments I thank Mr Ivan Collier and Mr Paul Seed for reading the text and a referee for helpful suggestions.

REFERENCES

1 Gardner MJ, Altman DG. Confidence intervals rather than *p* values: estimation rather than hypothesis testing. *BMJ* 1986;**292**:746–50

2 Sweeney KG, MacAuley D, Gray DP. Personal significance: the third dimension. *Lancet* 1998;**351**:134–6

3 Editorial. Estimating with confidence. *BMJ* 1988;**296**:1210–11

4 Altman DG, Gardner MJ. Calculating confidence intervals for regression and correlation. *BMJ* 1988;**296**:1238–42

5 Diem K, Lentner C, eds. *Documenta Geigy: Scientific Tables.* Macclesfield: Geigy, 1975

6 Greenhalgh T. Statistics for the non-statistician. II "Significant" relations and their pitfalls. *BMJ* 1997;**315**:422–5

7 Sokal RR, Rohlf FJ. *Biometry*. San Francisco: W H Freeman, 1981

8 Brooks DG, Carroll SS, Verdini WA. Characterizing the domain of a regression model. *Am Stat* 1988;**42**:187–90

9 Smith RJ. Regression models for prediction equations. *J Hum Evol* 1994;**26**:239–44

10 Weldon KL. *Statistics: a conceptual approach*. New Jersey: Prentice-Hall, 1986

11 LaBarbera M. Analyzing body size as a factor in ecology and evolution. *Annu Rev Ecol System* 1989;**20**:97–117

12 Cleveland WS, Diaconis P, McGill R. Variables on scatterplots look more highly correlated when the scales are increased. *Science* 1982;**216**:1138–41

13 Maggiorini M, Bartsch P, Oelz O. Association between raised body temperature and acute mountain sickness: cross sectional study. *BMJ* 1997;**315**:403–4

14 Weir C, Dyker A, Lees K. Participants required for trial of treatment with glucose and insulin [Reply to a Letter]. *BMJ* 1997;**315**:811

15 Payne N, Saul C. Variations in use of cardiology services in a health authority: comparison of coronary artery revascularisation rates with prevalence of angina and coronary mortality. *BMJ* 1997;**314**:257–61

16 Seglen PO. Why the impact factor of journals should not be used for evaluating research. *BMJ* 1997;**314**:498–502

17 Wilkinson RG. Health inequalities: relative or absolute material standards? *BMJ* 1997;**314**:591–5

18 Acerini CL, Patton CM, Savage MO, Kernell A, Westphal O, Dunger DB. Randomised placebo-controlled trial of human recombinant insulin-like growth factor 1 plus intensive insulin therapy in adolescents with insulin-dependent diabetes mellitus. *Lancet* 1997;**350**:1199–204

19 Nagasaka S, Iwamoto Y, Ishikawa S, Kuzuya T, Saito T. Efficacy of troglitazone measured by insulin resistance index. *Lancet* 1997;**350**:184

20 GISEN Group. Randomised placebo-controlled trial of effect of ramipril on decline in glomerular filtration rate and risk of terminal renal failure in proteinuric, non-diabetic nephropathy. *Lancet* 1997;**349**:1857–63

21 Watanabe T, Ashikaga T, Nishizaki M, Yamawake N, Arita M. Association of insulin with QTc dispersion. *Lancet* 1997;**350**:1821–2

22 Hall JC, Platell C. Half-life of truth in surgical literature. *Lancet* 1997; **350**:1752

23 Wilson P, Dunn LJ. Risk analysis can identify those patients needing isolation. *BMJ* 1997;**315**:58

24 Wilson M, Daly M. Life expectancy, economic inequality, homicide, and reproductive timing in Chicago neighbourhoods. *BMJ* 1997; **314**:1271–5

25 Mulligan K, Tai VW, Schambelan M. Energy expenditure in human immunodeficiency virus infection. *N Engl J Med* 1997;**336**:70–1

26 Roberts JD, Fineman JR, Morin FC, *et al.* Inhaled nitric oxide and persistent pulmonary hypertension of the newborn. *N Engl J Med* 1997; **336**:605–10

27 Koh KK, Mincemoyer R, Bui M, *et al.* Effects of hormone-replacement therapy on fibrinolysis in postmenopausal women. *N Engl J Med* 1997; **336**:683–90

28 Doty RL, Cheng L, Mannon LJ, Yousem DM. Letter. Olfactory dysfunction in multiple sclerosis. *N Engl J Med* 1997;**336**:1918–19

29 Hartman BW, Wagenbichler P, Soregi G. Maternal and umbilical-cord serum leptin concentrations in normal, full-term pregnancies. *N Engl J Med* 1997;**337**:863

30 Jacobson JM, Greenspan JS, Spritzler J, *et al.* Thalidomide for the treatment of oral aphthous ulcers in patients with human immunodeficiency virus infection. *N Engl J Med* 1997;**336**:1487–93

31 Colhoun H, Ben-Shlomo Y, Dong W, Bost L, Marmot M. Ecological analysis of collectivity of alcohol consumption in England: importance of average drinker. *BMJ* 1997;**314**:1164–8

32 van Haeften TW, de Vries J, Sixma JJ. Concordance of phenprocoumon dosage in married couples. *BMJ* 1997;**314**:1386

33 Dent OF, Sulway MR, Broe GA, *et al.* Alcohol consumption and cognitive performance in a random sample of Australian soldiers who served in the second world war. *BMJ* 1997;**314**:1655–7

34 Clarke R, Frost C, Collins R, Appleby P, Peto R. Dietary lipids and blood cholesterol: quantitative meta-analysis of metabolic ward studies. *BMJ* 1997;**314**:112–17

35 Goodnough LT, Monk TG, Andriole GL. Erythropoietin therapy. *N Engl J Med* 1997;**336**:933–9

36 Nyyssonen K, Parviainen MT, Salonen R, Tuomilehto J, Salonen JT. Vitamin C deficiency and risk of myocardial infarction: prospective population study of men from eastern Finland. *BMJ* 1997;**314**:634–8

37 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10

38 Fleiss JL. *The Design and Analysis of Clinical Experiments*. New York: John Wiley, 1986

39 Cruess DF. Statistics in journals [Letter]. *Lancet* 1991;**337**:432