Drug Information Journal, Vol. 36, pp. 115–125, 2002
Printed in the USA.

0092-8615/2002

STATISTICAL TECHNIQUES TO DETECT FRAUD AND OTHER DATA IRREGULARITIES IN CLINICAL QUESTIONNAIRE DATA

ROSEMARY N. TAYLOR, BSc, MSc
Consultant, Statistical Services Unit, University of Sheffield, Sheffield, United Kingdom
DAMIAN J. MCENTEGART, BSc, MSc, FIS
Head of Biostatistics & Data Management, Knoll Limited, Nottingham, United Kingdom
ELEANOR C. STILLMAN, BSc, PhD
Lecturer, Department of Probability & Statistics, University of Sheffield, Sheffield, United Kingdom

The detection of fraud and other systematic data irregularities in clinical trials is an important issue. While awareness of the problem is growing and willingness to combat it is clear, there still appears to be a lack of detection procedures suitable for routine implementation by trial coordinators. The shortage is particularly acute for discrete data, since the majority of methods which are available have been developed for continuous responses. In this paper, we examine the suitability of existing methods for discrete outcomes and propose a new technique for questionnaire data in both an informal graphical mode and as a randomization test. This method exploits the underlying correlation structure of a questionnaire and the difficulty in fabricating such details. A data set concerning a trial of a novel drug for treatment of schizophrenia, in which the Brief Psychiatric Rating Scale was used to assess patient mental health, is used for illustration.

Key Words: Fraud; Clinical trials; Clinical questionnaires; Correlation structure; Brief Psychiatric Rating Scale

INTRODUCTION

THE DETECTION OF FRAUD in clinical trials is clearly an important and sensitive issue. Such fraud could result in doubtful conclusions from trials, disadvantaging both the pharmaceutical company and potential patients. The motivation behind this kind of fraud may be laziness, financial gain, or professional recognition. The fraud itself can take many forms including plagiarism, piracy, and altered or ignored inclusion/exclusion criteria, as well as either completely or partially fabricated data. Only fraud involving fabricated data is considered here. More specifically, methods of detecting fabricated data in multicenter studies that include questionnaire responses, together with continuous data, are examined.

Estimates of the prevalence of general fraud vary considerably (1–6) and two surveys (4,6) (though admittedly not providing a systematic study of the area) report that a majority of respondents know of some instance of misconduct. The incidence of fraud

Reprint address: Damian J. McEntegart, Manager Statistics and Special Projects, Clinphone, Lady Bay House, Meadow Grove, Nottingham NG2 3HF, United Kingdom. E-mail: dmcente@clinphone.com.

in multicenter studies uncovered by recent audits has been consistently low with figures of 0.29% for the United States (7), 0.4% for the United Kingdom (8), and 0.43% for Europe and South Africa (2), though some cases may remain undetected. There is also evidence that, in many cases, even deliberately perpetrated fraud may not make an appreciable difference to the conclusions of a study (1).

Nonetheless, the potentially serious consequences, together with the fact that such practice is anathemic to professional scientists, indicate that trial coordinators should implement procedures to check for fraud. Naturally, there is a role for audit in this process, but also necessary is a battery of routinely applied statistical tests (9). These can be seen alongside existing data validation procedures (to which they are in many ways similar), but they have as their specific goal the detection of possibly fraudulent data. A key fact underlying most methods for detecting fabricated data is that it is difficult to artificially generate data which share the mix of structural and random elements typical of real data (1,10,11,12).

In common with other authors (see, for example, 1,9,10,11) many of the methods we will outline for the detection of fraud are techniques that will identify 'general departures from randomness.' Such departures will include protocol deviations, general data irregularities, and the usual, expected data differences between centers in multicenter trials. Fraud is perhaps the least likely explanation for data irregularities but it is often the one with the most serious consequences as it undermines the trial in question and reduces confidence in the industry as a whole. For this reason, we focus on fraud as the potential explanation for data irregularities but accept that follow-up will often find a perfectly satisfactory explanation. Of course, even if no explanation is found, establishment of a deliberate intention to defraud is another matter again, and outside the scope of this paper.

Many of the methods used for detection of fraud (see, for example, 1,11) are familiar exploratory data analysis techniques. However, these techniques give little insight into the type of data that result from questionnaires. Questionnaire data have the particular feature that they are typically categorical or discrete data from questions that are likely to be correlated. The aim of this paper is to examine several existing techniques for their suitability of application to questionnaire data and to develop additional dedicated methods.

DATA

The data set used to illustrate the paper relates to a multicenter parallel group study of a novel drug for the treatment of schizophrenia conducted by Knoll Limited. The study included data from 13 sites, however, since some sites have as few as two patients, only the data from the larger sites are used here. These sites, A-F, had sizes $n_A = 13$, $n_B = 24$, $n_C = 14$, $n_D = 16$, $n_E = 10$, and $n_F = 16$. Patients were treated with either Treatment 1, the established treatment, or new Treatment 2. Each patient was assessed at a baseline visit and again after one, two, four, six, and eight weeks. A visit consisted of a physical examination, taking details such as weight, blood pressure, and pulse; laboratory tests of blood and urine samples; and completion of a number of questionnaires by the investigator. These questionnaires related to the severity of symptoms and each was scored on a discrete scale. The one on which we shall focus in this paper is the Brief Psychiatric Rating Scale (BPRS) (13), which gives an overall assessment of a patient's mental health. It consists of 18 questions, each scored on a scale from 1 to 7. Before we proceed it is important to note that there is no suspicion of fraud in this data set; it is used merely for illustration.

EXISTING METHODS

Exploratory Multivariate Methods

Multivariate graphical techniques are useful for comparing aspects of the data visually. Although these methods are intended for

continuous data, they can be used perfectly adequately for discrete questionnaire data. We look at some examples below.

Chernoff Faces (14)

These use a picture of a face with a number of different characteristics, each varying in accordance with an underlying variable. Routines to plot the faces are available in S-PLUS (15).

Clearly, it would not be feasible to produce such plots for each patient in large studies, but they can give useful indications of intersite differences if used to plot mean responses for each site. Figure 1 shows the six sites plotted in this way (giving only the first 15 questions in the BPRS, pooled over all visits). The responses at Site F are seen to be perhaps slightly more severe than the others.

These plots show the data in an appealing way, but some variables have more emphasis than others. For example, the size of the face may have more impact on the impression given to the viewer than the distance between the eyes. This impression may vary between individuals as part of the general subjectivity of the technique. Symbolic faces are also limited to illustrating around 15 parameters.

Star Plots (15)

These plots overcome two of the difficulties with faces mentioned above, accommodating as many variables as necessary and with equal emphasis, however, their interpretation is still subjective. They illustrate the data by representing each variable by a radial line of length proportional to the size of response. The ends of these lines are joined, producing a 'star.' Figure 2 shows the site means (this time for all 18 BPRS questions) and again, Site F perhaps has slightly more severe symptoms.

Inliers (11)

When scanning entered data for possible errors or implausible data, it is usual to look for outliers. When looking for invented data, however, the reverse is often more appropriate. The perpetrator of the fraud is ex-

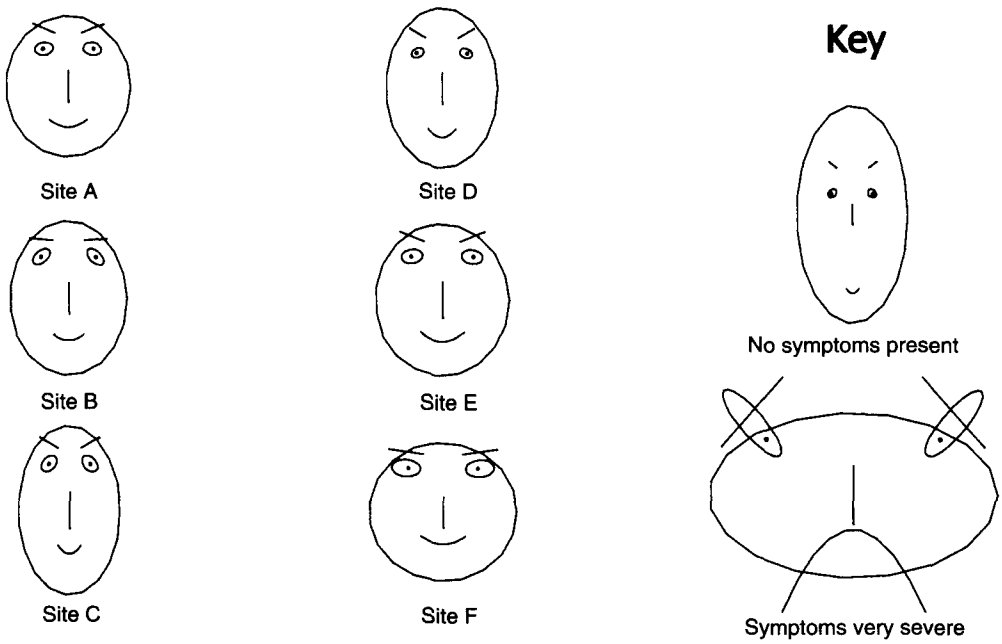

FIGURE 1. Chernoff faces plot of the first 15 questions of the BPRS data split by site.

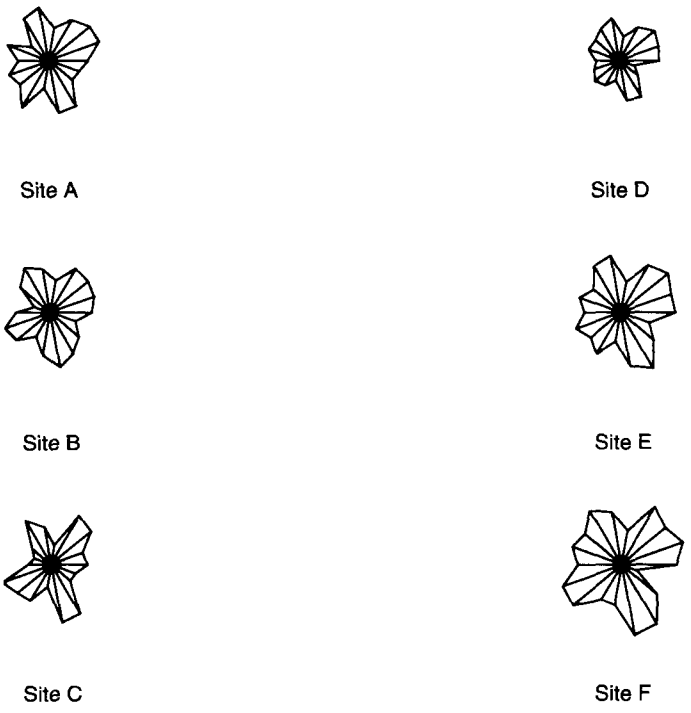

FIGURE 2. Star plot of the BPRS data split by site.

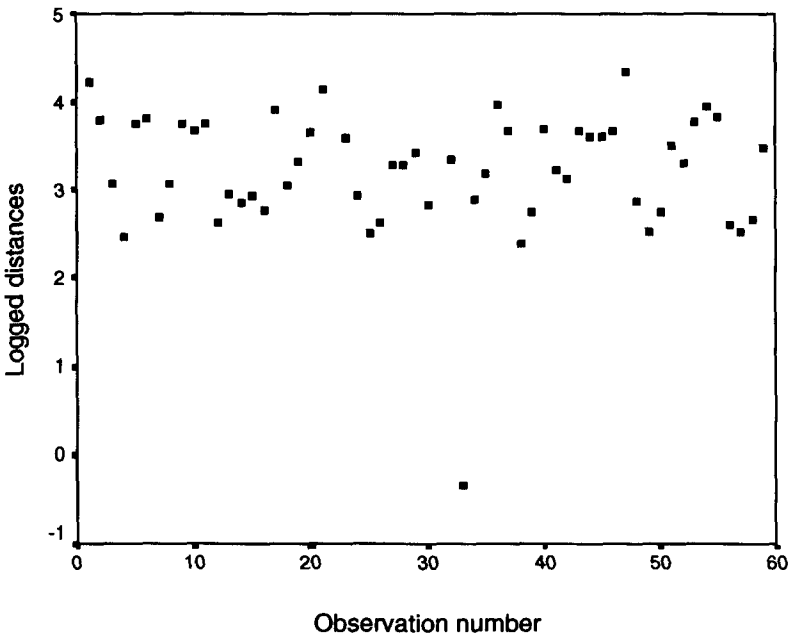

FIGURE 3. The logged distances for Site A together with the invented patient (observation number 33).

pected to be cautious in order not to draw attention to the imputed values and thus suspect data are those lying close to the mean of the distribution. Of course, the error in this thinking is that while it would not be surprising to find a patient with average weight, say, it would be most unusual for all their measurements to be average (11). Thus, a check for inliers, that is, observations too close to their multivariate mean, may be useful in detection of fraud.

The patient-based distance measure proposed by Evans (11) is to use the sum of squared z-scores of each observation from its mean, that is,

$$d_i = \sum_j \left(\frac{y_{ij} - \bar{y}_j}{s_j} \right)^2,$$

where y_{ij} = response of patient i to question j, j = 1, . . . , k
$\bar{y}_j$ = the mean response to question j, j = 1, . . . , k
s_j = the standard deviation of the responses to question j, j = 1, . . . , k over the group being considered.

Unusual values are highlighted more easily by plotting distance on a logarithmic scale (10). If required, an (approximate) formal test can be based on comparing the distance with the lower tail of a chi squared distribution (11).

We illustrate that this method has satisfactory sensitivity for discrete data by the following demonstration. We took Site A BPRS data and added a fictitious patient whose response to each question was the nearest integer to the real patients' mean response for that question. The distance measure for each patient was calculated and is plotted in Figure 3. The fictitious patient, who has observation number 33, is clearly visible.

Digit Preference Checks

Typically, if data are invented 'by hand' the perpetrator may leave a 'signature' in terms of personal preference for particular digits or patterns (16) revealed in the trailing digits of any measurements. Of course, once again there may be legitimate reasons for unequal distributions of digits, for example, blood pressure measured to 5 mmHg or pulse data measured for 30 seconds and doubled, but this should be actively considered. Indeed, it may be appropriate to direct investigators explicitly to the level of accuracy required. Simple patterns may be revealed in a stem-and-leaf diagram, by a chi-squared test for equifrequency, or by more complicated 'poker hands' or runs (17). Buyse et al. (1) cite the use of Benford's law that states that the probability of the first significant digit being equal to D (D = 1, . . . , 9) is approximately given by a particular logarithmic distribution. For questionnaire data, where responses are typically a single integer from a small range, these tests are of limited value. It may, however, be possible, as in our own case, to conduct the tests on associated continuous variables (eg, the vital signs and laboratory data values taken at each visit) in order to detect wholly invented patients/visits.

One useful elaboration of the basic technique is to look for change in digit preference over time. This may reveal a change of investigator, simplification of the measurement method, or invention of patients' readings after a certain time point (the latter two situations perhaps indicating fraud due to laziness).

For example, we investigated whether the frequency of obtaining integer weights changed over time. Recoding integer weights as one and others as zero, arranging these in chronological order, and plotting the cumulative sum should result in a straight line if the frequency is unchanging.

In Figure 4 the plot for Site F is shown. There is a clear break of slope in the center. Further investigation showed that this corresponded to the period September 3 to 16 and is explained by an alternative investigator substituting for the main investigator during a holiday.

Date Checking

One final method that we mention briefly is date checking (1,9,10,18). Clearly, this is not

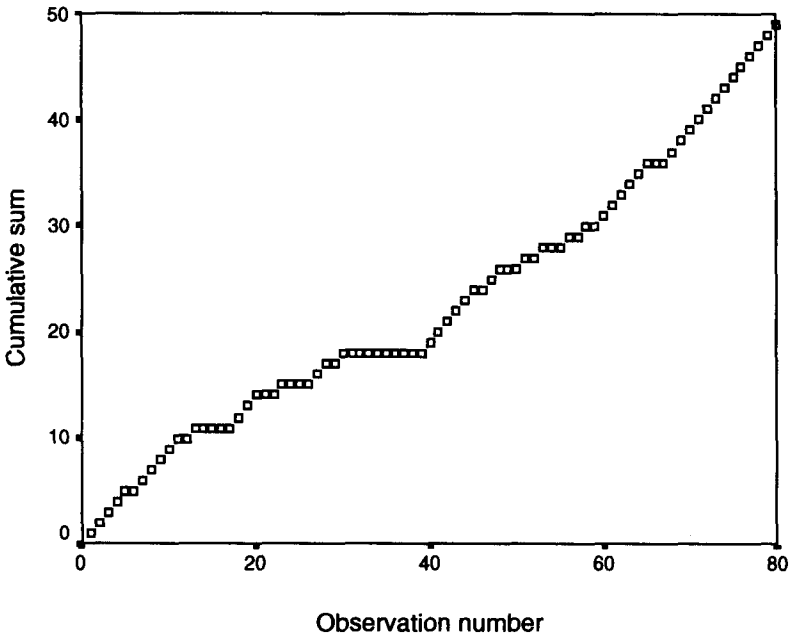

FIGURE 4. Cumulative sum plot of occurrence of integer weight values over time. A change in frequency occurs around observation number 30.

related to the nature of the variables and should be carried out for both continuous and discrete measurements. It is usual to check that routine measurements are not taken at weekends or bank holidays and to verify that the ordering and spacing of tests meets the study guidelines. In checking our data, it was found that laboratory test results frequently predated the actual recorded visit data. This can be explained by samples being taken and analyzed prior to a visit so that results are available to the clinician at that time.

A CORRELATION-BASED METHOD FOR QUESTIONNAIRE DATA

Many papers giving methods for detection of fraud (eg, 1,11) advocate checking that correlations between recorded variables are neither implausibly weak nor strong. However, no guidelines are presented on how this should be carried out in practice on what is likely to be a very large number of variables. The categorical or ordinal nature of the data also deters investigators from calculating ordinary (Pearson) correlation coefficients for questionnaire-based studies. We show below that examination of these coefficients can simply and usefully be conducted for questionnaire data. Clearly, both graphical and randomization versions of the test given below would also be appropriate for other discrete or continuous data such as laboratory data, vital signs, or multiple measures of efficacy.

In developing these methods we take advantage of the strongly defined structure of questionnaire data. For example, the BPRS consists of 18 questions, each of which rates the level of a symptom on a scale from 1 (absent) to 7 (extremely severe). The total score can be used as a single measure of mental health, but here we focus on the 18 individual scores recorded for each patient. It is reasonable to assume (unless a questionnaire has been specifically designed to measure orthogonal features) that the responses are not independent and in many cases it is plausible that the correlations between the answers to individual questions will remain fairly con-

stant even if the actual scores vary. Our test is based on assessing the magnitude of variation between correlation structures.

Depending on the actual design of the study, comparisons of various types may be of interest, but here we consider the evidence for a difference in correlation structure across sites only. We, thus, calculate our correlations between pairs of questions at each site, combining data over visits and treatments. This is based on the idea that fraud is most likely to occur at a site level and that, while treatment effects are clearly likely to influence the questionnaire responses, and perhaps even their correlation structure, there is little reason to suspect differential effects between sites. If one had concerns about between-center differences in demographics or other prognostic variables such as baseline BPRS score, inpatient/outpatient ratio, duration of schizophrenia, and prior use of concomitant medicine, one could always, subject to sufficient patient numbers, perform the analyses on subgroups, for example, between sites within treatments. Similarly, the exercise could be performed for specific visits.

The correlation coefficient between questions t and s for each site is calculated as

$$r_{t,s} = \frac{\left\{\sum_{i=1}^{N} (y_{it} - \bar{y}_t)(y_{is} - \bar{y}_s)\right\}}{\left[\left\{\sum_{i=1}^{N} (y_{it} - \bar{y}_t)^2\right\}\left\{\sum_{i=1}^{N} (y_{is} - \bar{y}_s)^2\right\}\right]^{1/2}}$$

where y_{ij} = response to jth question for patient i, j = t,s

$\bar{y}_j$ = mean response for question j, j = t,s

N = number of observations, that is, the total number of sets of 18 responses for that site.

Graphical Approach

As there are 18 questions on the BPRS questionnaire there are 153 (= 18 × 17/2) pairs of questions and so 153 different correlation coefficients relating one question with another. It is a useful preliminary step to display the figures in matrix form using grey levels to indicate the strength of correlations (eg, using 15). For the purposes of this display the absolute values of the correlations were used, as there were very few negative correlations. In the general case it would be possible to allow for negative correlations, perhaps by the use of different colors for positive and negative values. In the test below the actual correlations were used. Comparisons between sites from the displays in Figure 5 are easily possible.

Here no clear between-site differences are apparent, but to illustrate what might happen if data were invented, the data for Site A were permuted as follows. All responses for each question were permuted and randomly assigned as new responses for that question, but each question was treated separately. Thus, responses to each question were each individually plausible, but they are no longer linked to 'corresponding' responses to the other questions. Comparison of the correlation plots for the original and permuted data (Figure 6) shows the appreciably weaker correlation structure of the permuted set.

A Test Procedure

To formalize the graphical comparison performed above, a randomization test approach (19) was adopted. We regard our site correlation coefficients as points in 153-dimensional space. For a site to be declared unusual it must lie in the extremes of this space, typically far from some central point. We propose using as a central point the correlation coefficient calculated from all six sites and measuring distance from it by (squared) Euclidean distance, that is,

$$d^* = \sum_{\substack{t,s \\ t \neq s}} \left(r_{t,s} - r^*_{t,s}\right)^2$$

where $r_{t,s}$ = the Pearson correlation coefficient between questions t and s for that site and $r^*_{t,s}$ = the correlation coefficient between questions t and s calculated from data for all six sites. To identify whether a site lies an

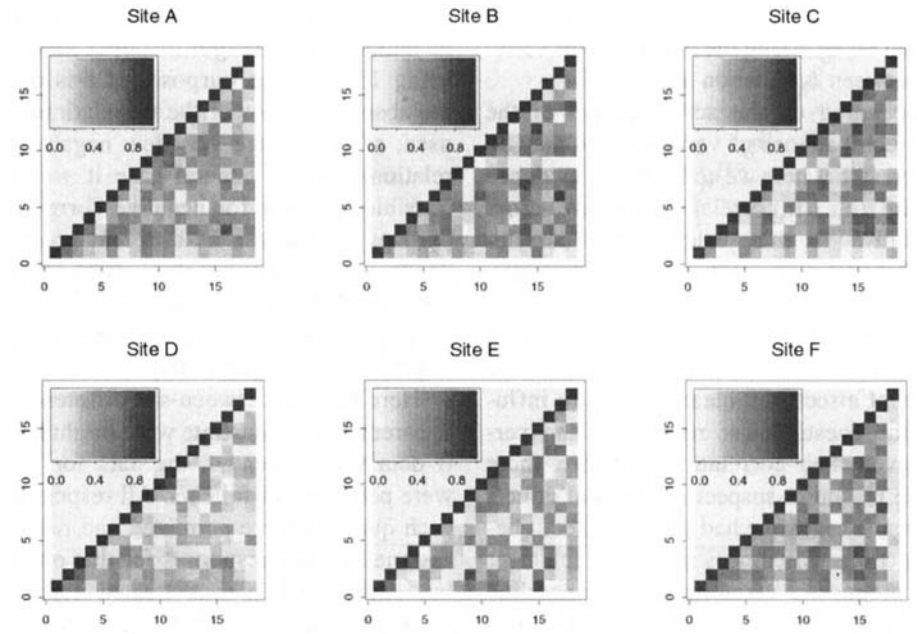

FIGURE 5. The correlation coefficients for all 153 pairs of questions in the BPRS data displayed using grey scale intensity to code for strength.

unusually large distance from the central point, while recognizing the different site sizes, we proceed as follows. Consider a check on the validity of Site A, which consists of data on n_A patients. Suppose we create a 'pseudo Site A' composed of n_A individuals selected at random without replacement from all those in the study, and then calculate its between-question correlation vector and squared distance from the central point as defined above. This d* is then a random value from the distribution of the squared

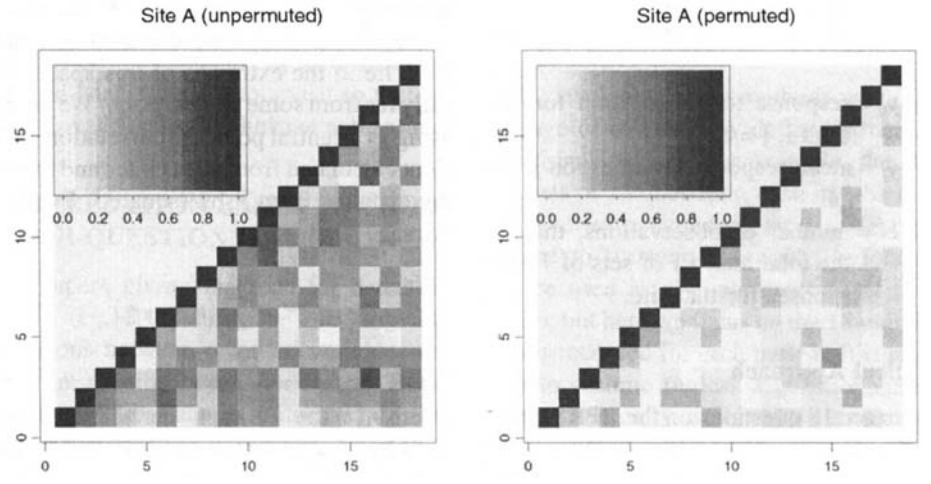

FIGURE 6. Grey scale coded display of the genuine and permuted Site A data. The appreciably weaker correlation structure in the permuted set is clearly visible.

distance statistic under the null hypothesis of only random variation between the site correlation vectors. By repeating this procedure many times, we build up a picture of the entire null distribution. We can then look at the position in this distribution of the real Site A.

A similar procedure can be carried out for each of the other sites. Note that in each case the random selection is at the level of the individual and data from all visits for that individual are taken together for calculation of the correlations. This is in line with our earlier decision not to differentiate between visits.

The procedure was implemented in S-PLUS (15) and the following plot (Figure 7) shows the null distributions based on 5000 (as calculated from the recommendations in 19) simulations for each site. The squared Euclidean distance, d*, and the proportion of observations greater than the real site, q, are given in each case. We can examine the plots to establish whether any site exhibits unusual behavior. Using a 5% test with a Bonferroni correction for the multiple sites, none of the real sites shows strong evidence of an atypical correlation structure (though site D is borderline; see below).

A similar procedure was adopted for the permuted Site A data, Ap, used in Figure 6 (though here the selection was over, and central point r* calculated from, the data sets for individuals in Ap, B, . . . , F, ie, the real Site A was excluded) and the resulting distribution is also given in Figure 7. The 'fraudulent' data are clearly extreme.

Notwithstanding the lack of clear statistical significance, the results for Site D were examined further. All sites had been comprehensively monitored according to Good Clinical Practice, including verification of source data in the investigators' files. These monitoring records and other relevant data were

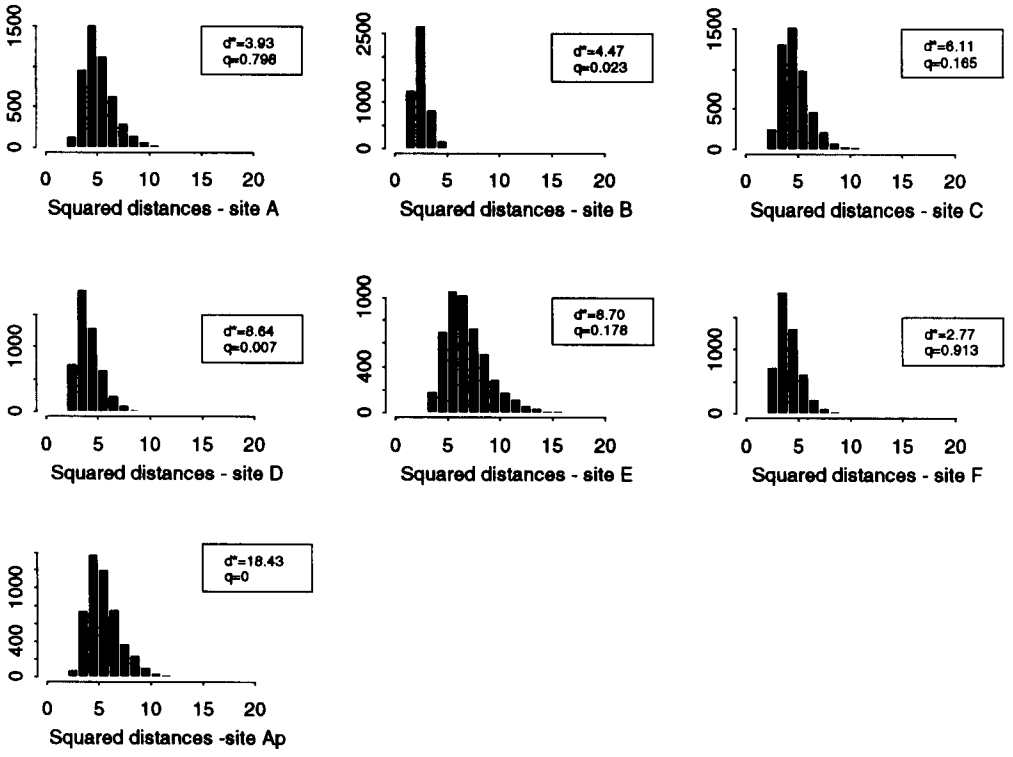

FIGURE 7. Distribution of squared distances from the central point for each site.

examined, but there was nothing to suggest that any fraud had occurred. It was noted, however, that this site had three investigators conducting assessments. For the main part, each patient was assessed by the same investigator, but this was not always the case. It is, therefore, considered most likely that any atypical correlation structure is a reflection of this fact. Training sessions to ensure consistency in ratings between investigators were conducted before the study, but it is likely that some differences would have remained.

DISCUSSION

Many of the techniques described in this paper will identify general 'departures from randomness.' Such departures will include systematic irregularities in data quality and other protocol deviations in data collection, measurement, or recording. The Chernoff faces and star plots have been used to compare sites on mean responses. Such differences between sites are to be expected due to variation in demographics, disease characteristics, standard therapies, site procedures and a variety of other considerations. Despite the expectation of their existence though, intersite differences should still be examined as a source of possible fraud. Such checks have led to the discovery of a well-known fraud case in a multicenter animal experiment (20). Such cases will be the exception, however, and thus it is important to state again that there will be a variety of explanations other than fraud for data highlighted for further examination by the use of these techniques.

This paper has focused on fraud as this is the cause with most concern and ramifications for the public (concern that treatments are effectively evaluated), the sponsor (eg, delay in drug approval process), and the industry in general. Other causes are, of course, also of interest, especially if they lead to remedial action, for example, more training of site staff while the study is ongoing. It should also be noted that extensive exploration of data for departures from randomness may detect many false positives by chance.

Our test for departures from random variation in the correlation structure has been derived using the principles of randomization testing. Further work could be conducted to establish its exact properties, but we note that as an exploratory data evaluation technique these are not critical. The test could also be contrasted with tests for the strict equality of correlation matrices (eg, 21). An alternative graphical technique for displaying correlation matrices is available (22) but the displays presented here are considered easier to produce and visualize.

CONCLUSION

Generally, the main responsibility for detecting fraud lies with the clinical monitors and quality assurance group. Measures for these groups to prevent and detect fraud have been given elsewhere including in previous issues of this journal (23,24). Statistical techniques also have an important role in the detection of fraud. They can be used as screening mechanisms to identify sites for audit or other examination. They can also be used as tools for further investigation of sites that fall under suspicion following an audit. With respect to questionnaire data, it is more difficult for clinical and quality assurance teams to detect fraud because of the large number of data items to assimilate and the coarse grading scale for individual responses. This article has illustrated some existing and novel techniques that can be utilized. Standard programs could be written for use of these methods within each company, either as routine or as special investigations.

Acknowledgments—This work was carried out while Rosemary N. Taylor was in receipt of an Engineering and Physical Sciences Research Council studentship. We would like to thank Jonathan Mitchell and Dr. Raymond Bratty for making the data available, and Professor Clive Anderson, Dr. Paul Blackwell and two anonymous referees for their valuable comments.

REFERENCES

1. Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, Lesaffre E, Murray G, Edler L, Hutton

J, Colton T, Lachenbruch P, Verma BL. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. Report of the International Society of Clinical Biostatistics Subcommittee on Fraud. *Stat Med.* 1999;18(24):3435–3451.
2. Schmidt J, Gertzen H, Aschenbrenner KM, Ryholt-Jensen S. Detecting fraud using auditing and biometrical methods. *Appl Clin Trials.* 1995;4(5):40–49.
3. O'Donnell P. Facing up to fraud. *Appl Clin Trials.* 1993;2(3):36–40.
4. Lock S. Research misconduct: a resume of recent events. In: Lock S, Wells F, eds. *Fraud and Misconduct in Medical Research.* 2nd edition. London: BMJ Publishing Group; 1996:14–39.
5. Wells F. Investigating fraud—again. *Appl Clin Trials.* 2000;10(2):26–27.
6. Ranstam J, Buyse M, George SL, Evans S, Geller NL, Scherrer B, Lesaffre E, Murray G, Edler L, Hutton JL, Colton T, Lachenbruch P. Fraud in medical research: An international survey of biostatisticians. *Control Clin Trials.* 2000;21:415–427.
7. Weiss RB, Vogelzang NJ, Peterson BA, Panasci LC, Carpenter JC, Gavigan M, Sartell K, Frei E, McIntyre OR. A successful system of scientific data audits for clinical trials. *JAMA.* 1995:270;459–464.
8. Hone J. Combating fraud and misconduct in medical research. *Scrip Magazine.* 1993:March;14–15.
9. Collins M, Evans S, Moynihan J, Piper D, Thomas P, Wells F. *Statistical Techniques for the Investigation of Fraud in Clinical Research.* London, England: Association of the British Pharmaceutical Industry Fraud Statistics Working Party; 1993.
10. Evans SJW. Detection of Fraud. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics.* Chichester: John Wiley & Sons; 1998: 1583–1588.
11. Evans SJW. Statistical aspects of the detection of fraud. In: Lock S, Wells F, eds. *Fraud and Misconduct in Medical Research.* 2nd edition. London: BMJ Publishing Group; 1996:226–239.
12. Mosimann JE, Wiseman CV, Edelman RE. Data fabrication: Can people generate random digits? *Account Research.* 1995;4:31–55.
13. Overall JE, Gorham DR. The Brief Psychiatric Rating Scale. *Psychol Rep.* 1962;10:799–812.
14. Chernoff H. The use of faces to represent points in k-dimensional space graphically. *J Am Stat Assoc.* 1973;68(342):361–368.
15. S-PLUS. Seattle, WA: Statistical Sciences, Inc; 2000.
16. Preece DA. Distribution of final digits in data. *The Statistician.* 1981;30(1):31–60.
17. Newman TG, Odell PL. *The Generation of Random Variates.* No. 29 of Griffin's Statistical Monographs & Courses. Stuart, A, ed. London; Griffin; 1971.
18. Horowitz AM. Fraud and scientific misconduct in the United States. In: Lock S, Wells F, eds. *Fraud and Misconduct in Medical Research.* 2nd edition. London: BMJ Publishing Group; 1996:144–165.
19. Manly BFJ. *Randomization and Monte Carlo Methods in Biology.* London: Chapman & Hall; 1991.
20. Bailey K. Detecting fabrication of data in a multicentre collaborative animal study. *Control Clin Trials.* 1991;12:741–752.
21. Larntz K, Perlman MD. A Simple Test for the Equality of Correlation Matrices. In: Gupta SS, Berger JO, eds. *Statistical Decision Theory and Related Topics IV, Vol* 2. New York, NY: Springer-Verlag; 1988:289–298.
22. Koziol JA, Alexander JE, Bauer LO, Kuperman S, Morzorati S, O'Connor SJ, Rohrbaugh J, Porjesze B, Begleiter H, Polich J. A graphical technique for displaying correlation matrices. *Am Statistician.* 1997:51;301–304.
23. Walsh RC. Systematic measures for the prevention and early detection of investigator fraud. *Drug Inf J.* 1994;28:1161–1165.
24. Mackintosh DR, Zepp VJ. Detection of negligence, fraud and other bad faith efforts during field auditing of clinical trial sites. *Drug Inf J.* 1996;30:645–653.