CLINICAL TRIALS DATA MANAGEMENT AND TRIAL CONDUCT Clinical Trials 2013; 10: 783–806

Application of methods for central statistical monitoring in clinical trials

Amy A Kirkwood[a], Trevor Cox[b] and Allan Hackshaw[a]

Background On-site source data verification is a common and expensive activity, with little evidence that it is worthwhile. Central statistical monitoring (CSM) is a cheaper alternative, where data checks are performed by the coordinating centre, avoiding the need to visit all sites. Several publications have suggested methods for CSM; however, few have described their use in real trials.

Methods R-programs were created to check data at either the subject level (7 tests within 3 programs) or site level (9 tests within 8 programs) using previously described methods or new ones we developed. These aimed to find possible data errors such as outliers, incorrect dates, or anomalous data patterns; digit preference, values too close or too far from the means, unusual correlation structures, extreme variances which may indicate fraud or procedural errors and under-reporting of adverse events. The methods were applied to three trials, one of which had closed and has been published, one in follow-up, and a third to which fabricated data were added. We examined how well the methods work, discussing their strengths and limitations.

Results The R-programs produced simple tables or easy-to-read figures. Few data errors were found in the first two trials, and those added to the third were easily detected. The programs were able to identify patients with outliers based on single or multiple variables. They also detected (1) fabricated patients, generated to have values too close to the multivariate mean, or with too low variances in repeated measurements, and (2) sites which had unusual correlation structures or too few adverse events. Some methods were unreliable if applied to centres with few patients or if data were fabricated in a way which did not fit the assumptions used to create the programs. Outputs from the R-programs are interpreted using examples.

Limitations Detecting data errors is relatively straightforward; however, there are several limitations in the detection of fraud: some programs cannot be applied to small trials or to centres with few patients (<10) and data falsified in a manner which does not fit the program's assumptions may not be detected. In addition, many tests require a visual assessment of the output (showing flagged participants or sites), before data queries are made or on-site visits performed.

Conclusions CSM is a worthwhile alternative to on-site data checking and may be used to limit the number of site visits by targeting only sites which are picked up by the programs. We summarise the methods, show how they are implemented and that they can be easy to interpret. The methods can identify incorrect or unusual data for a trial subject, or centres where the data considered together are too different to other centres and therefore should be reviewed, possibly through an on-site visit. *Clinical Trials* 2013; **10:** 783–806. http://ctj.sagepub.com

[a]Cancer Research UK & UCL Cancer Trials Centre, University College London, London, UK, [b]Cancer Research UK Liverpool Cancer Trials Unit, University of Liverpool Cancer Trials Centre, Liverpool Cancer Research UK Centre, University of Liverpool, Liverpool, UK

Author for correspondence: Amy A Kirkwood, Cancer Research UK & UCL Cancer Trials Centre, University College London, 90 Tottenham Court Road, London, W1T 4TJ, UK.
Email: a.kirkwood@ucl.ac.uk

© The Author(s), 2013
Reprints and permissions: http://www.sagepub.co.uk/journalsPermissions.nav
10.1177/1740774513494504
Downloaded from ctj.sagepub.com at University College London on June 25, 2015

Introduction

Substantial resources are spent conducting clinical trials, due in part to current guidelines and regulations. International Conference on Harmonisation–Good Clinical Practice (ICH-GCP) [1] requires that the data be 'accurate, complete, and verifiable from source documents'. A statement adhered to by many organisations is, 'In general there is a need for on-site monitoring, before, during, and after the trial'. Although site visits can be useful to examine procedures for safety reporting and drug labelling, for verifying pharmacy supplies, and performing other monitoring tasks, considerable effort is typically spent checking trial data with patient records, i.e., source data verification. Although some organisations use source data verification for a random sample of participants (e.g. 20%), others still perform 100% checks. A clinical trial database can never be completely free from errors. Monitoring of trial data is used to minimise errors, but also can be used to check on the progress of a trial and to detect fraud [2].

Fraud is relatively uncommon. In a survey of several thousand US scientists [3], almost 30% admitted to participating in some questionable research activity in their career, but only 0.5% admitted to 'falsifying or "cooking" research data'. Anecdotal evidence of fraud tends to be limited to small projects where the researcher has complete control of the data. Steen [4,5] examined articles between 2000 and 2010 and found that 1 in every 6070 clinical trials was retracted. From 180 assessable retracted articles involving humans, there were 9 clinical trials with >200 participants. Seven [6–12] of these were retracted for fraud; however, the term 'fraud' encompassed a wide range of activities, and only two trials were suspected of falsifying data based on the six retraction statements available [10,11].

On-site monitoring is a core function of many clinical trial organisations, particularly Contract Research Organisations and pharmaceutical companies. However, major data errors are often infrequent, and there is no reliable evidence that on-site visits influence the results and study conclusions. Furthermore, random errors should be balanced between groups in a randomised trial, thus having a negligible effect.

Central Statistical Monitoring (CSM) has been proposed as a cheaper and more efficient alternative to on-site data monitoring of all trial sites (centres) [13]. With CSM, data checks are performed by the coordinating centre in order to minimise the need to visit every site. Although ICH-GCP explicitly allows CSM, the text of that document is not sufficiently permissive: 'however in exceptional circumstances the sponsor may determine that central monitoring … can assure appropriate conduct of the trial'. There is a view that, if full on-site monitoring is not done, the chance of a marketing license being approved is decreased, and so the costs of monitoring are considered justified. However, recent draft guidelines from the Food and Drug Administration encourage the use of CSM [14].

Several authors have described methods for detecting fraud and data errors [15–18], but few publications have described CSM in real clinical trials already completed or in progress and those that do tend to focus on a particular method to check data for anomalies, or have not applied the method(s) to actual trial data.

Al-Marzouki *et al.* [19] investigated fraud in a trial evaluating a dietary intervention for patients with coronary heart disease. They examined the variances and digit preference and suggested that the data had been fabricated. However, they focused on the whole data set, and not by site, which would be one of the main purposes of CSM in clinical trials. Bailey [20] investigated suspected fraud in one laboratory in a multi-centre animal study, using scatter plots to examine correlations between variables. He showed that after a suspicious centre had been investigated, the variance of certain variables increased to a similar level seen in other laboratories.

Herein, we apply CSM methods, provide a suite of R-programs, and show how the output from the programs can be interpreted. We have examined the simplicity and reliability of the methods, including their strengths and limitations.

Methods

We classified data monitoring to be at either (1) trial participant level or (2) site level. A set of R-programs [21] was developed to implement CSM methods; the R-programs are freely available from http://www.ctc.ucl.ac.uk. ("Training" section) R is free and relatively simple to use; the R-programs can be run without spending significant time tailoring the programs for an individual study. We intended to implement a range of data checks without the need for intensive programming by information technology (IT) staff. Tables 1 and 2 list the monitoring checks we examined, their purpose, and the corresponding R-programs. Appendix A – Text 1 describes the methods in more detail.

Our first goal was to detect data errors at the participant level. Buyse *et al.* [16] and Baigent *et al.* [17] suggested calendar checks to find errors such as dates that occur on weekends and holidays (when unexpected) or in an incorrect order, for example, treatment after death. The R-programs to detect outliers, that is, observations that appear too large or too small, were based on univariate (including Grubbs's method) and multivariate approaches (using Euclidean and Mahalanobis distances) [15–17].

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

Table 1. Summary of the methods described for participant-level data monitoring and the situations they can be applied to

	Recording and entry errors	Procedural errors	Fraud	R-program
Participant-level data monitoring				
Dates – order checking[a]	✓	✓	✓	date_order_check
Dates				weekend_hol_check
Weekends[a]	✓		✓	(option: weekends)
National holidays[a]	✓		✓	(option: holiday)
Outliers				outlier_check
Using standard deviation	✓	✓		(option: sd)
Grubbs' method	✓	✓		(option: grubbs)
Multivariate – Euclidean distance	✓	✓		(option: euclid)
Multivariate – Mahalanobis distance	✓	✓		(option: mahal)

[a]Methods developed by the authors.

Table 2. Summary of the methods described for centre-level data monitoring and the situations they can be applied to

	Recording and entry errors	Procedural errors	Fraud	R-program
Centre-level data monitoring				
Digit preference – integers (rounding)	✓	✓		integer_check
Digit preference: (1) Benford's law and (2) comparison with all other sites[a]			✓	digit_preference
Comparison of variable means				mean_check
Chernoff faces		✓	✓	(option: faces)
star plots		✓	✓	(option: star)
Inliers			✓	inlier_check
Adverse event rates[a]		✓		SAE_check
Correlation checks			✓	correlation_check
Variance checks	✓		✓	variance_check
Categorical variables[a]	✓	✓	✓	cat_check

SAE: serious adverse event.
[a]Methods developed by the authors.

Site (centre)-level data monitoring checks aim to (1) identify systematic errors in trial conduct (procedural errors) at a site, which could be due to a genuine misunderstanding of the trial protocol by local staff, and (2) detect fraud resulting from fabricating trial participants and/or data or creating data for missing values of actual participants [16]. We applied these methods to individual trial sites, but they also could be applied to individual investigators or geographical regions. These checks are intended to flag sites discrepant to the other sites by looking for unusual data patterns, possibly triggering an on-site visit to check the procedures and original data. The methods include examining correlation structures [16–18] using a method described by Taylor *et al.* [18], digit preference [15–17] (including Benford's law [22]), and inliers [15], that is, participants with several variables whose values lie close to the mean. We also implemented methods to detect procedural errors to identify sites that rounded too many continuous measurements [18] and for repeated continuous measurements, that is, data for a participant with too little variability over time [15,16,20].

Reporting and monitoring adverse events is an important activity in clinical trials. Over-reporting may indicate that site staff are being overly cautious about classifying adverse events, which creates extra work to process reports. However, under-reporting is potentially serious and could affect the trial conclusions as well as regulatory responsibilities. We examined the number of serious adverse events (SAEs) per site based on the number of participants recruited, and the length of time in the trial. We calculated and SAE rate for each site as the number of participants with at least one SAE divided by the total number recruited, and divided further by trial duration at that site. Another method developed used the time each participant spent in the trial.

The methods described above were applied to three phase III cancer trials, in which overall survival was the main end point:

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

- Study 12 [23]: a double-blind trial of 724 patients with small-cell lung cancer, randomised to receive thalidomide or placebo, in addition to standard chemotherapy. Patients were recruited from 79 centres (2003–2006).
- ABC-02 [24]: an unblinded trial of 324 patients with advanced biliary tract cancer, comparing gemcitabine/cisplatin with gemcitabine. Here, we also manually created data errors and fabricated patients (one person created them, and another who was blind to the errors ran the R-programs to identify them). Patients were recruited from 37 centres (2002–2008).
- TOPICAL [25]: a double-blind trial of 670 patients with non-small cell lung cancer that was ongoing at the time of our assessment of CSM. It compared Tarceva with placebo, among patients considered unfit for chemotherapy. The monitoring findings were checked in real-time with data queries sent to centres. Patients were recruited from 78 centres (2005–2009).

Because Study 12 and ABC-02 had closed already, data could be examined only retrospectively; thus, queries about anomalous or suspicious data were not sent to sites that participated in those trials. We describe how output from the R-programs which apply CSM methods were interpreted. We also provide a summary of their main strengths and limitations in Table 3. We refer to 'real' data as those observed from the trials, as opposed to fabricated data that were manually created to evaluate the programs.

Results

Participant-level data monitoring

Dates: ordering, weekends, and national holidays

We examined whether dates of randomisation, blood test results, chemotherapy, and follow-up appointments occurred on weekends or holidays and whether there were obvious date ordering errors. The R-program produces tables listing the discrepancies. Very few errors were detected (Table 4) particularly in the Study 12 database, which already had been cleaned and analysed. Weekend and holiday dates from the TOPICAL trial were queried, some were found to be correct (often inpatient treatment), and some were data errors. For Study 12, after inspecting paper case report forms (CRFs) several dates would have been changed. However, there were only 13 living patients for whom any of these dates would be used in survival analyses and corrections would likely have a negligible effect on the results.

Outliers

Several univariate continuous variables were examined; Figures 1(a) and (b) provide examples of the output (outliers are shown as solid points). The number of data points considered to be outliers using the $\pm k$ Standard Deviation (SD) method was 478 (0.93% of all data values) and 148 (1.6%) in Study 12 and TOPICAL respectively. Using Grubbs's test, these numbers increased to 2056 (4%) in Study 12 and 425 (3.3%) in TOPICAL. However, none of these variables were used in the primary analyses.

Twenty fabricated values were added to the variable 'haemoglobin' in the ABC-02 trial by an independent statistician who made them 'extreme'. The R-program detected 13 out of 20 using a ± 3 SD cut-off; the other 7 were found after the 13 had been replaced with their genuine values. All 20 were detected immediately when the cut-off was ± 2 SDs (Figure 1(b)) or when Grubbs's test was used at ± 3 SDs because the larger false values were removed after each iteration of the program and no longer masked the less extreme outliers. No genuine data values were picked up as outliers with either method. Both methods require the data to be Normally distributed, so our R-programs produce Normal probability plots and histograms. When the data clearly are not distributed Normally, the program could be re-run to detect outliers that lie more than $k \times$ inter-quartile range, that is, above and below the upper and lower quartiles.

The R-programs also can be used to check several continuous variables for participants simultaneously (multivariate outliers). We used data from several CRFs for each of the three trials (example in Figure 2). D values, where D is the sum of either the Normalised Euclidean distances or the Mahalanobis distance from the mean (Appendix A), which exceed ± 2 SDs from the mean D are automatically shown in red. A list of participants with large D values is also produced. For example, only 13 of 658 patients (2%) using the Euclidean distance and 8 (1.2%) using the Mahalanobis distance were flagged as multivariate outliers in the TOPICAL trial using 19 pretreatment variables simultaneously. A participant identified in both the univariate and multivariate outlier programs could be flagged for particular attention.

Site-level data monitoring

Rounding and digit preference

The R-program to identify rounding [18] was applied to all of the continuous variables on two CRFs in Study 12. In the example shown in Figure 3, two sites showed a monotonic increase, indicating no evidence of systematic rounding, but site 68 reported only integer values after the first observation.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

Table 3. Strengths and limitations of the methods

	Strengths	Limitations
Participant-level data monitoring		
Dates		
Order checking Weekends National holidays	• A simple spreadsheet is produced containing all the information needed to query the results • Can check a large number of variables quickly	• May pick up unimportant errors which take time to query • Within the weekend and national holiday programs, care needs to be taken to only include dates that you would not expect to fall on weekends and national holidays • Will currently only detect UK national holidays; the code could be adapted for use in other countries
Outliers		
Standard Deviation (SD) method Grubbs' test Inter-quartile range (IQR) method Multivariate – using Euclidean distance Multivariate – using Mahalanobis distance	• An easy-to-read spreadsheet is produced containing all the information needed to query the outliers • Can be arranged by site so that sites consistently reporting outlying values can be flagged • The SD cut-off can be chosen by the user and changed for different data sets, perhaps lowered for variables to be used in the primary analyses • Several options (univariate: SD, Grubbs, IQR; multivariate: Euclidean and Mahalanobis distances) are available depending on whether the data are Normally distributed	• May not find all errors if smaller outliers are masked by bigger ones (SD method), especially in small data sets • May not be applicable with non-Normally distributed data (SD, Grubbs) • The SD cut-off needs to be carefully chosen to avoid flagging too many real data points • The multivariate methods may not pick up errors detected using the univariate methods
Centre-level data monitoring		
Digit preference: integers (rounding)	• Significant rounding can be easily detected by eye	• Somewhat subjective; however, alterations to the code may be possible so that only sites with a certain percentage of rounded values or several consecutive rounded values would be shown in the output. • Many plots to examine if a trial has many sites
Digit preference		
Benford's law		• Benford's law is only applicable for variables where the leading digit can be 1–9, so many clinical trial variables would not be consistent with this (e.g., blood pressure). Also, the variable should not be Normally distributed
Comparison with all other sites (Either first or last digit)	• Can also be used to detect rounding, for example, do sites have lots of continuous measurements ending with 5 or 0? • No underlying assumptions about the variables (unlike Benford's law)	• May detect small (unimportant) differences in large sites. Care needs to be taken to use several methods or examine several CRFs (comparing which numbers appear preferred) to ensure there really are problems with a site • Assumes that investigators falsifying data may show digit preference
Comparison of variable means		
Chernoff faces Star plots	• Graphical display of several variables simultaneously	• Difficult to interpret • Small differences in some features stand out more than large differences in others (face plots) • Difficult to read when many variables are used (star plots) • Both are difficult to read with many sites (may work better in trials with a small number of sites, 5–10 perhaps)

(continued)

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

Table 3. (Continued)

	Strengths	Limitations
Inliers	• Can detect participants with several values lying too close to the mean • Produces Easy-to-read spreadsheets and plots which give the values of each variable for inlying participants along with the mean of the variable	• The program may not automatically flag participants if several are created lying close to the means; however, they should show up on the plots. • The cut-off needs to be carefully chosen so that not too many participants are flagged • Will not flag falsified participants if their data values do not lie close to the means • Should not be used in sites with small numbers of participants • Assumes falsified data would lie close to the means (i.e., made to look as similar to the real data as possible)
Adverse event rates	• Can detect sites with small numbers of SAEs in comparison to the number of participants and the time in the trial • Produces easy-to-interpret plots and spreadsheets	• Need to carefully examine the plots and tables (from the output) to consider whether flagged sites need investigating
Correlation checks	• Adjustable significance limits • May detect falsified data which appear acceptable in univariate analysis	• Can only be used for sites with sufficient numbers of participants (at least 10) so that the correlation matrix is sufficiently reliable • Grey-scale plots are subjective to interpret • Variables with zero or very small SDs are not appropriate • Problems if there are small (very close to zero) variances within some variables, within a site; these variables or sites will need to be removed • The more sites examined, the more likely that a statistically significant p-value would be found by chance, hence we suggest using a low p-value cut-off of 0.01
Variance checks	• Second option (to look for no change between observations) also available	• Large outliers may cause problems • Some subjectivity when interpreting the output • Needs a reasonably large number of observations per participant
Categorical variables	• Output is easy to interpret	• Can only be used in sites with reasonably large numbers of participants in each level (expected values of at least 5 in each cell to run a chi-square test)

SAE: serious adverse event; CRF: case report form.

Table 4. Numbers of potential date errors found in the TOPICAL and Study 12 trials

Trial/test	Number of data sheets	Number of variables	Number of data values	Number of errors found (%)
Study 12				
Date order	11	29	25,903	61 (0.23)
Weekends	5	21	45,325	308 (0.67)
National holidays	5	21	45,325	50 (0.11)
TOPICAL				
Date order	21	47	30,964	191 (0.62)
Weekends	7	15	13,336	288 (2.2)
National holidays	7	15	13,336	49 (0.37)

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

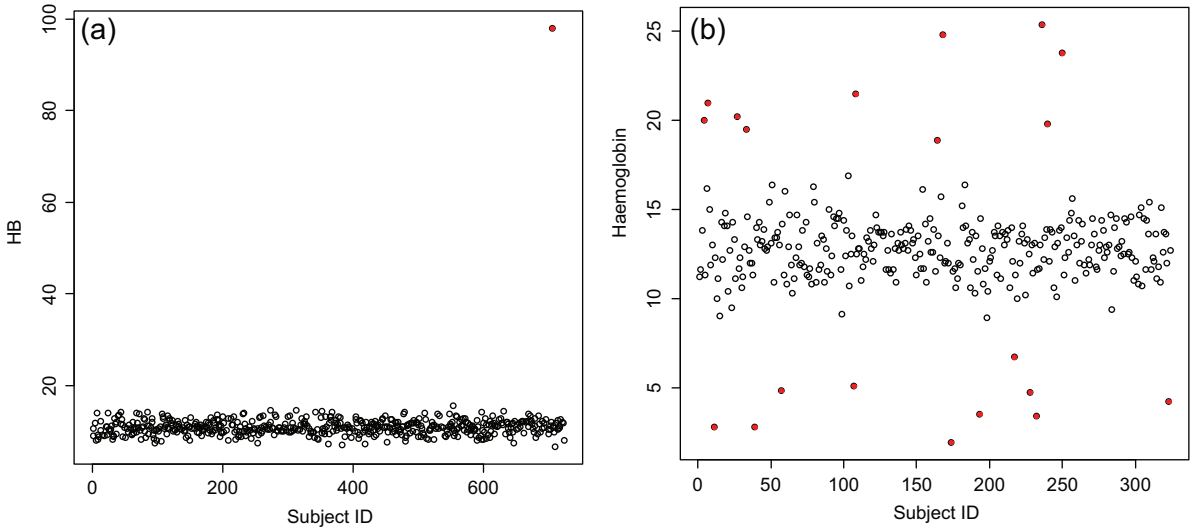

Figure 1. (a) A scatter plot for the Study 12 trial and one continuous variable 'haemoglobin' (HB) at baseline. There is only one outlier, automatically shown as a solid point (coloured red in the R output). (b) A scatter plot for the ABC-02 trial and one continuous variable, haemoglobin, at baseline. The points automatically shown as a solid point (coloured red in the R output) are outliers (lying $> \pm 2$ SD from the mean). These were all 20 fabricated values that were added to the data set.

Benford's law [22] was applied to numeric variables from several CRFs within Study 12 to identify digit preferences. When comparing the observed distribution of leading digits with that expected from Benford's law, we flagged sites that had p-values ≤ 0.01 (chi-square test). Table A1 (Appendix A) is the output for Study 12, where 16 out of 65 sites were flagged. However, Benford's law should be used only for variables where the leading digit can range between 1 and 9, may take values across a range of several orders of magnitude, and are not Normally distributed; assumptions that are unlikely to hold for many clinical trial variables. In Study 12, for example, the data overall did not fit Benford's distribution ($p < 0.001$).

We proposed an alternative method, which compares the observed distribution of leading digits within each site with that from all other sites. This does not have the same limitations as Benford's law, so all continuous variables could be included (Appendix A – Table A1). For example, the data from site 11 (Table 5) would be flagged using Benford's law, but when compared with all other sites, its distribution does not appear to be discrepant; indicated by the very different p-values, $p < 0.001$ versus $p = 0.77$. The number of sites that were flagged with this method was only 3 of 66 (data from the Study 12 chemotherapy CRF) and 1 of 34 (data from the TOPICAL pretreatment CRF). One site in Study 12 was flagged based on data from the randomisation and chemotherapy CRFs but, on closer inspection, did not appear suspicious because the differences were small and the digit patterns were not similar between the two CRFs. When a large number of

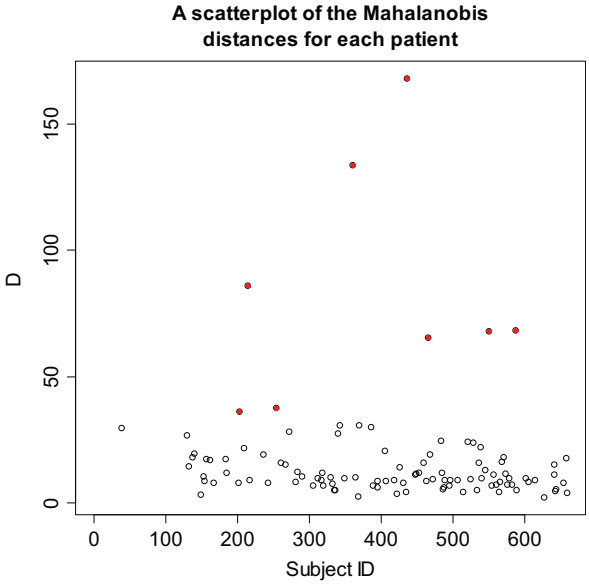

Figure 2. A scatter plot for the TOPICAL trial, based on 19 continuous variables (on the case report form completed at the start of treatment). There are 8 potential multivariate outliers (where *D* exceeds ± 2 SDs), shown in red.

data values were examined, even a small difference between the observed and expected proportions sometimes produced a small p-value, so we concluded that the size of the differences should be examined as well as the p-values. Our R-program can also use the last digit of continuous variables instead of the first digit when examining rounding. No sites were flagged using this test on data from

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

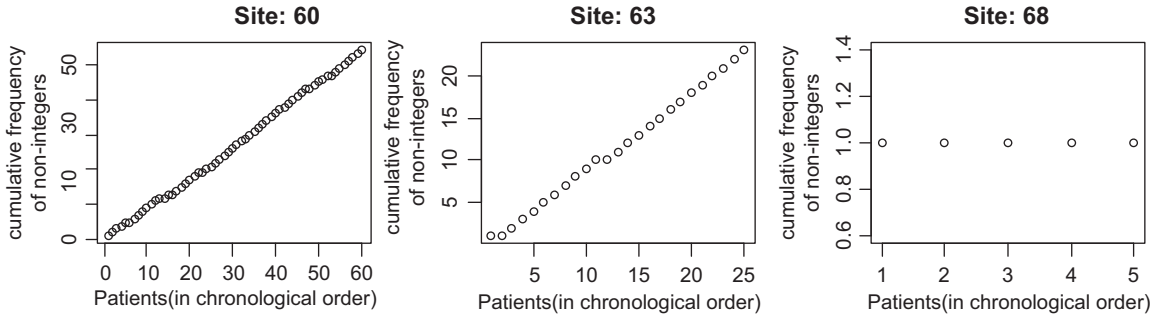

Figure 3. Plot to check for rounding in the variable 'white blood cell count' in three different sites in Study 12.
Site 68 shows evidence of rounding, which in an ongoing trial should be investigated further.

Table 5. Output from the R-program for two sites, comparing the distribution of the first significant digit with the expected distribution based on (1) Benford's law and (2) the distribution from all other sites (Study 12 trial)

First digit	Expected proportion (Benford's law)	Site number 11		
		Number of data values (using 26 variables)	Observed proportion	Observed proportion from all other sites
1	0.301	343	0.298	0.324
2	0.176	180	0.157	0.155
3	0.125	164	0.143	13.6
4	0.097	155	0.135	0.121
5	0.079	86	0.075	0.072
6	0.067	65	0.057	0.055
7	0.058	54	0.047	0.048
8	0.051	47	0.041	0.043
9	0.046	56	0.049	0.046
p-value – Benford's law			<0.001	
p-value – compared with all other sites			0.77	

either the Study 12 chemotherapy CRF or the TOPICAL pretreatment CRF.

Comparing means of variables among sites

We implemented three methods for simultaneously comparing the means of continuous measurements among centres: Chernoff face plots [26,27], star plots [27], and parallel coordinate plots (see Appendix A). However, the results from application of these methods were difficult to interpret, and two methods were particularly influenced by outliers. In addition, numerically small differences in the means of variables between one site and another could appear large when displayed as Chernoff faces.

Inliers (data values too close to the means, possibly indicating fraud)

Figure 4 shows a plot used to identify inliers for the TOPICAL trial (based on 19 pretreatment blood values). Similar to the multivariate outlier program, D, the sum of the Euclidean distances from the mean, is calculated for each participant, but here we focus on participants with unusually *small* D values, which are more apparent on a log scale. Only one patient (from site 20, circled in Figure 4) had a $\log(D_i)$ value that exceeded 2.5 SDs below the overall mean $\log(D)$). The number of inliers for the other two trials was also low. Closer inspection of data from patients which were flagged as inliers showed they came from CRFs with small numbers of continuous variables with just one or two of the values lying close or equal to the mean.

Because few inliers were found in Study 12 and TOPICAL, we fabricated patients in ABC-02 which had data close to the mean values (Appendix A – Text 1, 'Inliers'). In a site with few patients (n = 9), a cut-off of 2 SDs identified all 6 fabricated patients, and 2.5 SDs flagged four of them. In the site with a medium number of patients (n = 18), no fabricated patients were found using a cut-off of 2.5 SDs, but three were found using 2 SDs. Five fabricated

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

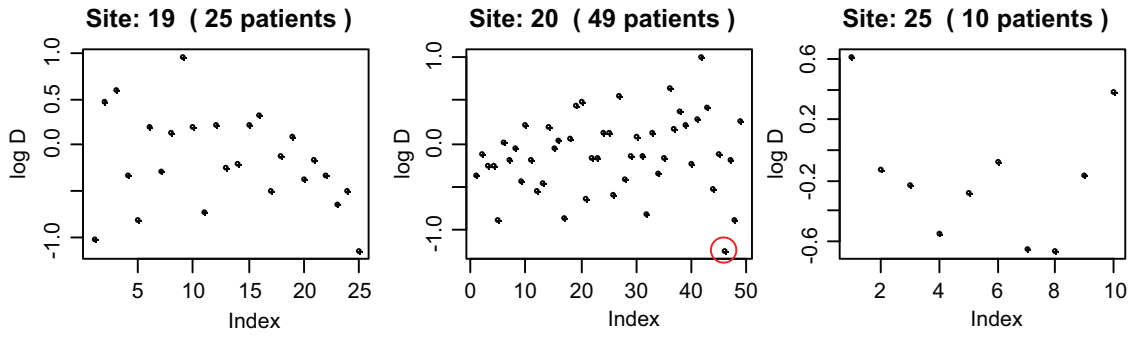

Figure 4. Examining inliers in TOPICAL for 3 sites and 19 variables (i.e., participants who have values too close to the means).
Inliers are identified as having large negative values for log *D* (y-axis). Each point represents a trial participant, and the point circled (in site 20; in the R-program output this shows up as a red circle) is a patient whose *d* value lies more than 2.5 standard deviations (SDs) below the mean. The program automatically circles inliers.

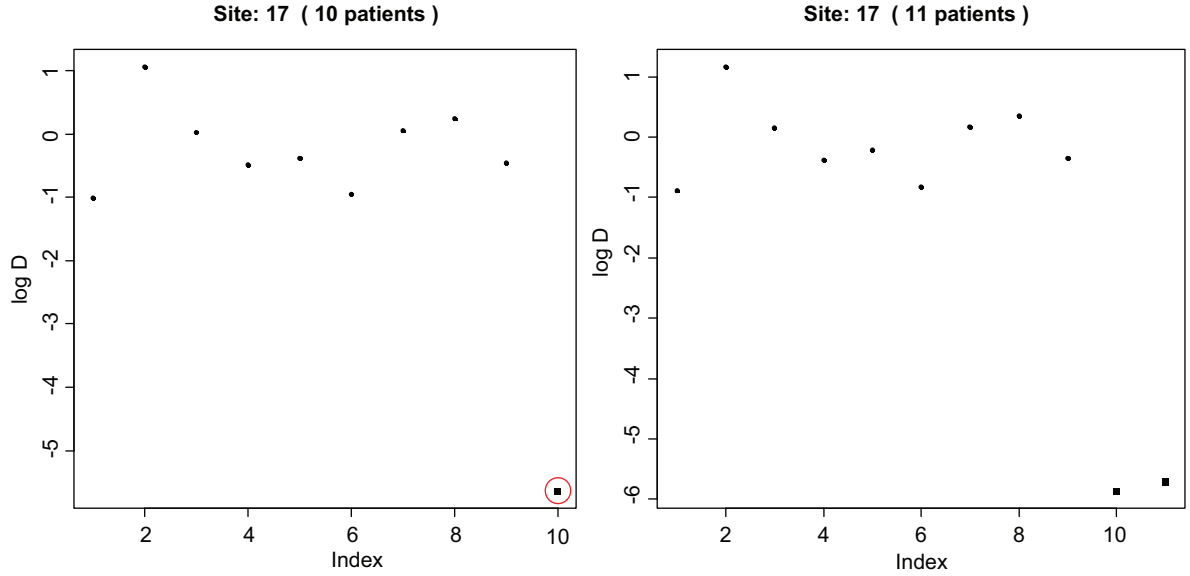

Figure 5. Inliers in one site in ABC-02, in which patients (shown as black squares) were manually created to be similar to the means of several variables. In the left-hand figure, there is one fabricated patient automatically flagged (circled) by the program (−2.5 SDs away from the mean *D*; in the R-program, output appears as a red circle). But in the right-hand figure, there are two fabricated patients. Together they have decreased the variance of *d*, so they are no longer flagged as inliers (they are not circled in red by the program), but they are apparent on inspection.
SD: standard deviation.

patients in the large site (n = 49) were detected using 2.5 SDs, and all using 2 SDs. However, when several data values are fabricated within a site, they may skew the overall mean and SD of log(*D*). In this situation fabricated data may not be flagged as inliers, but they may appear to be discrepant on visual examination (Figure 5). Grubbs's test may be helpful for identifying both true inliers and outliers.

Correlation checks

We examined whether a site appeared different from others within the same trial using a set of continuous variables; the output is a grey-scale grid of squares, where each square represents the correlation between two variables. (Colour could be used in place of the grey scale.) Pairs of variables with a correlation coefficient of 1 are indicated by black squares, those with a coefficient of −1 as white squares, and everything else as a shade of grey. Discrepant sites tend to have more light or dark squares than other sites. A formal statistical test using simulations was also applied (see Appendix A – Text 1, 'Correlation checks'). None of the sites in Study 12 or ABC-02 were flagged, that is, had $p < 0.01$. In the data from the TOPICAL pretreatment CRF, only two sites had $p < 0.01$; but on closer inspection, neither

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

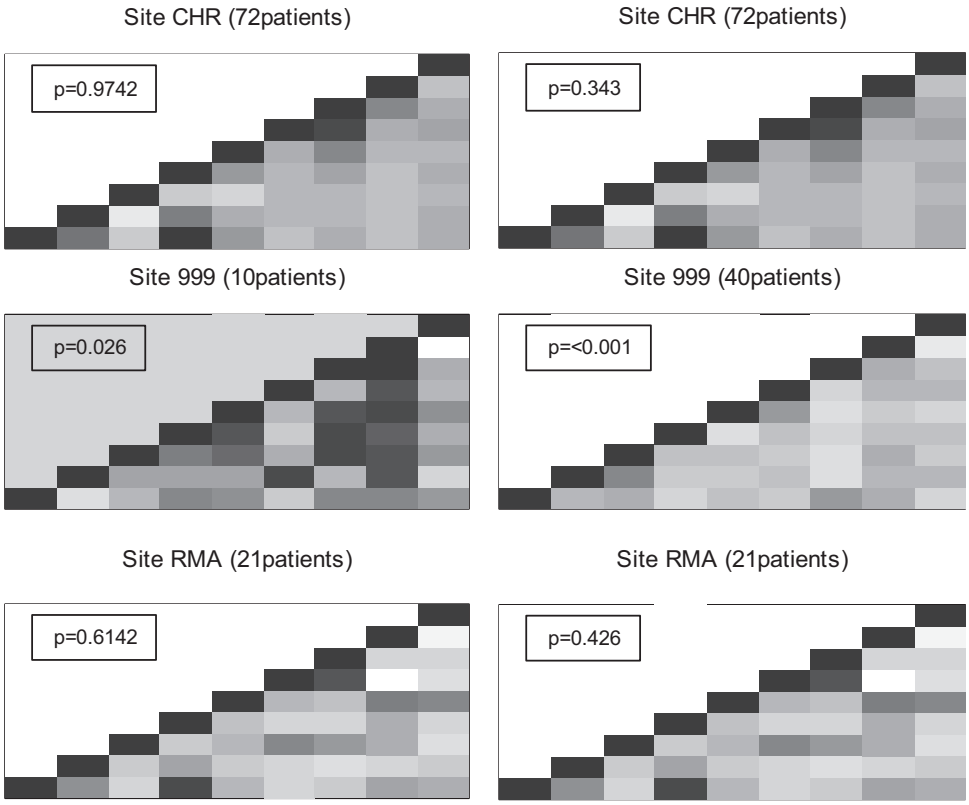

Figure 6. Correlation checks for the ABC-02 trial for 9 variables. Each square represents the correlation between a pair of variables (highly positive = dark colour, highly negative = light colour). Both panels show two real sites (CHR and RMA) and a fabricated site (site 999) generated using randomly chosen values (left panel) or by choosing values close to the mean of each variable (right panel). The p-value for each site is based on simulations which count the number of times a site with a correlation matrix as extreme as that observed could be generated using randomly chosen patients from all other sites.
In both panels, the correlation structure in site 999 appears discrepant. In the left-hand panel, there is a solid dark block (8th variable along), though the p-value is just above the statistical significance cut-off of 0.01 (i.e., 0.026), because of few patients (n = 10). However, in the right-hand panel, the shading is generally lighter overall, showing little correlation between variables, in contrast to all other sites, where there are clear correlations (e.g., 4th variable along and 1 up; 7th variable along and 6 up). With n = 40 patients, the comparison is statistically significant.

site particularly stood out on the grey-scale plots. We concluded that the small p-value may have been influenced by one or two correlations that were weaker or stronger than in the other sites (Appendix A – Figure A1). Both the output display and p-values must be examined when interpreting the results of these checks.

We tested the program by adding fabricated sites using two different methods (Appendix A – Text 1). When sites were created by randomly picking values for each variable from the values seen in other sites, one could create plots that appeared strange, with whole columns that looked strikingly different, but a p-value just above 0.01, particularly in fabricated sites with small numbers of patients (e.g., Figure 6, left panel). When sites were created to have values around the means of each variable, the correlation plots tend to have an overall light grey colour, indicating little correlation. Large fabricated sites (>25 patients) tend to produce a p-value < 0.01. When we ran the R-program 10 times for sites with 30, 35, and 40 fabricated patients, a fake site with $p < 0.01$ was flagged 19 of 30 times. Close inspection of the output revealed a clear absence of strong correlations that appeared in all other sites (Figure 6, right panel).

Variance checks for repeated measures data

Variance checks were made for each of four values from laboratory assays of blood samples taken up to 6 times in Study 12, and 7 such values from up to 18 times in ABC-02. The R-program produced a table showing the percentage of patients in each site which fell into the bottom 2.5% of variances (based on the number of patients checked), and the displays as in Figure 7 (for platelets in ABC-02). In

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

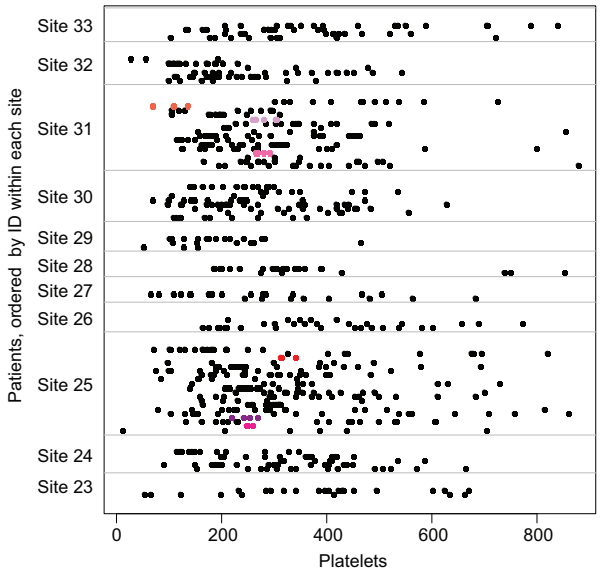

Figure 7. Variance checks for ABC-02. There are three fabricated patients with small variances in site 25 and three in site 31 (shown in shades of red and pink).
Dots of the same colour belong to the same patient. The R-program automatically assigns a non-black colour to participants with outliers.

neither trial was a site with patients in the bottom 2.5% of concern. Few sites had more than one outlying patient, often with only a few repeated measurements; and those sites that did had relatively large numbers of patients in total, such that the variances did not appear to be unusual.

In the ABC-02 trial, 11 fabricated patients were added to two sites by an independent statistician; both sites were flagged by the R-program. Seven of the 11 patients had variances in the bottom 2.5% of the distribution for at least 1 of the 7 variables tested, including one fabricated patient who had variances which fell in the bottom 2.5% for 4 different variables. The two sites with fabricated patients appeared in the bottom 2.5% of variances for more variables than any actual trial site. Figure 7 shows the output for sites 23–33; three fabricated patients with low variances are highlighted in sites 25 and 31.

The variance check method involves visual inspection of potentially many displays, with several plots for each variable checked. The number of participants per site, the size of the observed variances, and where the flagged participants fall in relation to others, must be considered when identifying anomalies. For example, several participants with low variance who cluster within a site may indicate fabricated data [20]. Correct transcription of data from CRFs and other measurements for the flagged participants should be checked.

Table 6. Example of output from the checks of a categorical variable, using the baseline stage in Study 12.

(a) Details of the test for each site.

Stage	Site frequency	All other sites' frequency
Site: 41		
Limited	1 (7.7)	367 (51.6)
Extensive	12 (92.3)	344 (48.4)
Chi-square p	0.004	
Site: 45		
Limited	12 (46.2)	356 (51.0)
Extensive	14 (53.9)	342 (49.0)
Chi-square p	0.77	

(b) p-value for each site tested (can be examined quickly, example shows the 5 sites).

Site	p-value
36	0.332
37	0.186
41	0.004
44	0.328
45	0.775

Baseline stages were compared using a chi-square test. Each site (with enough patients) was compared to all of the remaining sites combined. Results are output in two spreadsheets.

Categorical variables

Categorical variables were checked using chi-square tests which compared each site with all of the remaining sites combined. Two categorical variables were checked in Study 12. For Eastern Cooperative Oncology Group (ECOG) scores and stages, only 5 of 79 and 22 of 79 sites, respectively, had sufficient numbers to be checked. No site had a $p<0.01$ for ECOG score but 1 did for stage (Table 6). Closer inspection of the observed frequencies revealed no cause for concern. This R-program is particularly useful for large trials with many participants in each site in which the main outcome measure is categorical, for example, response to treatment. Numbers of deaths or disease progressions also could be investigated using this method.

Adverse events

The R-program flags sites that have very few or too many participants with adverse events compared with other sites. We focused on SAEs, but the program could be adapted to examine any type of adverse event. A summary table was produced to show sites from ABC-02 (in which we fabricated data) that are in the lowest and highest 10th centile of SAE rates. The rate was calculated as the number of participants with an SAE in the site divided by the number of participants and the time the site had been recruiting; an

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

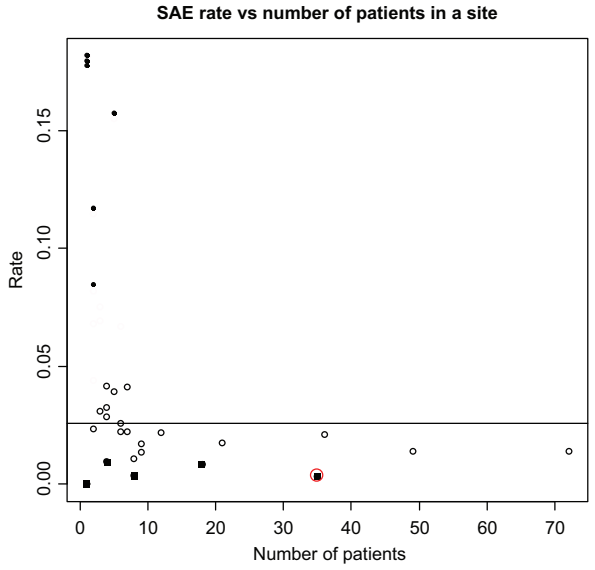

Figure 8. Examining SAE rates (ABC-02).
SAE: serious adverse event.
Each point represents a site. The y-axis is the SAE rate per site, allowing for the number of patients and the time the site has been recruiting. The lowest 10% of SAE rates are shown as black squares, and the highest 10% as solid black circles. The circled observation is one of the fabricated sites added during the simulations, and identified as having a low SAE rate compared to the average for all sites (horizontal line). Sites furthest towards the bottom right could have on-site monitoring checks.

example is shown in Tables A2 and A3 (Appendix A). Any centile value can be specified in the R-program. Sites that required further investigation had a reasonable number of participants followed for several months, but had few or no SAEs. We ran simulations by adding a fabricated site to the ABC-02 data 625 times, covering all combinations of numbers of patients (5, 10, 25, 35, or 45), lengths of time in the trial (5, 10, 15, 30, or 45 months), and numbers of patients with SAEs (from 1 to all patients).

Figure 8 is an example of one of the output displays of the rate of SAEs versus the number of participants at each site. The circled black square is the fabricated site (specified as being open for 45 months, having recruited 35 patients but having recorded only 6 SAEs).

From the simulations, we also attempted to specify the maximum number of SAEs that a site could have but still appear in the bottom 10% of rates (Appendix A – Table A4); for example, in a site with 25 patients which had been open for 30 months, there could be $\leq$9 patients with an SAE and it would still fall within the bottom 10% of all sites. Such sites, which would fall in the bottom right-hand side of the display in Figure 8 (small SAE rates and relatively large numbers of patients), may warrant further investigation.

The method described above uses an estimated overall time (i.e., time from first randomisation until last randomisation plus 'x' months, where 'x' is specified as the number of months in which SAEs are expected, i.e. slightly longer than the treatment time. We also used another method of calculating the SAE rate, based on the time spent in the trial by each participant, from date of randomisation through date last seen (Appendix A – Text 1, 'Adverse events'). Using this approach, the two largest sites were flagged by the R-program because of the long times during which they accrued participants (Appendix A – Figure A2). However, close inspection of the output revealed that their rates were actually larger than the overall rate, although below the median per site, and the total number of SAEs recorded was relatively high, with one site having recorded 65 SAEs for 46 patients and the other 54 SAEs for 32 patients.

Strengths and limitations

The strengths and limitations of each R-program are summarised in Table 3. In most cases, the main limitation to interpretation of output from the programs is the size of the trial. The methods to detect data errors can be applied to all trials; however, many of the methods that aim to detect fraud would be difficult to apply reliably in small trials, or even larger trials when small numbers (<10) of participants are recruited within each site.

Furthermore, the programs that examine possible fraud were created using certain assumptions about the way fabricated data would be generated; when these assumptions are incorrect, that is, a researcher is 'too good' at faking data, fraud may not be detected.

Another difficultly is that several methods rely on a visual assessment of output displays, and for these, we added formal statistical tests to aid interpretation. However, the displays should be readily interpreted once the user has gained experience in applying the R-programs, and we stress that the output from no single program should be used as definite evidence of any irregularity at a site.

Finally, we have applied the methods, implemented in the R-programs, only to trials in which fraud was unlikely to have occurred; therefore, we had to fabricate data ourselves to evaluate some methods. We encourage readers to help refine and improve these R-programs by applying the methods to their own databases.

Discussion

We have summarised several methods of CSM by illustrating their application to data from three trials

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

and the main strengths and limitations of each method (Table 3). Although the methods are not perfect, they are easy to implement and interpret. Importantly, they identify certain data errors, such as incorrect dates and outliers, far more quickly and easily than one would during usual data editing and correction processes. They also may help to detect fraudulent sites or participants when fraudulent data are generated in a manner consistent with the assumptions that are the basis of the methods.

The R-programs can be executed automatically and the output should be examined by suitably trained staff. On-site data monitoring visits could be targeted for sites that appear discrepant to the others. During visits to other sites, resources could focus on other on-site activities, such as staff training, helping with accrual, documentation of informed consent, and pharmacy and adverse event checks.

Participant-level data monitoring often is performed during data entry, using automatic validation checks within the database. However, the methods we describe here are not easily programmed into many database systems. Methods such as the correlation check, the inliers, and digit preference program could be applied to as many variables as desired to look for unusual patterns, but we recommend that checks to identify errors and rounding and testing the distributions of categorical data be limited to variables associated directly with safety, treatment compliance, and the primary efficacy end points.

Central site-level monitoring considers both visual assessments of data displays and formal statistical tests. Sites will, by chance, be flagged by one of the methods. Therefore, to avoid too many centres being flagged as suspicious, no single data check should be used automatically trigger an on-site visit. CSM should guide the depth of investigation of anomalies. Consistency of findings from several checks should be used to determine whether a particular centre is identified as suspicious [18]. When we applied the four tests designed to detect potential fraud (digit preference, inliers, correlations, and variance checks) to Study 12, only two sites were flagged by more than one test. Both sites were flagged by the digit preference program and both contained a patient inlier. Further investigation and checks ruled out fraud. More detailed checks also could be performed when a centre is flagged as suspicious before undertaking on-site investigation. For example, if a site reported a particularly low number or rate of SAEs, a first step could be to compare the baseline characteristics of participants enrolled at this site with participants at other sites.

Our R-programs were based on methods suggested by others [15–18]. Although these articles discussed types of fraud and possible methods of detection, only one applied some of the methods to real data [18]. O'Kelly [28] fabricated participants and a blinded statistician applied methods for inliers, outliers, and unusual correlation structures [18] comparing the results between sites. We agree with O'Kelly who suggested that the actual values of the correlation matrix should be examined, rather than just looking at how different the overall structure is. We did not cover Statistical Process Control (SPC) for clinical trials [29,30]. Most SPC methodology is based on the use of control charts for monitoring a process over time. Similar methods could be used to monitor clinical trials, for example, tracking the average number of SAEs per participant in each trial arm over time.

At present, paper CRFs are used to report data in many trials, including all trials run within our Clinical Trials Unit (CTU) and 57% of Canadian trials [31]. In our vision for the future, a Personal Data Assistant (PDA) would be used for collecting participant data. The PDA would check the data being entered and warn when a possible error was being made or for missing data. At the end of data entry, the forms would be automatically transmitted to the database in the coordinating centre. Participant and centre-level checks, and trial level summaries, would be generated automatically. Trial staff could choose from a menu that allows interactive looks at the data of the types we summarise in our article. For example, for participant-level monitoring, a panel of tables and graphs could appear with demographic details, event dates, missing data, and time plot of when SAEs were recorded. At the site level, there would be a panel showing recruitment graphs, a control chart for SAEs, a control chart for determining the level of on-site monitoring and various fraud test results, and tables and graphs of differences between centres.

In conclusion, CSM can be a cost-effective and worthwhile alternative to on-site source data verification. It can identify anomalous participant data that may be incorrect or fabricated, and sites where the data considered together are quite different from all other sites and therefore should be investigated. The methods are relatively simple to implement and interpret.

Acknowledgment

Selected for poster presentation at the Annual Meeting of the Society for Clinical Trials (May 2011) and the MRC Clinical Trials Methodology Conference (October 2011). The trial registration details were as follows: Study 14: ISRCTN77341241; TOPICAL: ISRCTN77383050; and ABC-02: ISRCTN82956140.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest

None declared.

References

1. Guideline for good clinical practice E6(R1). In *International Conference on Harmonisation*, 1996. Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6_R1/Step4/E6_R1__Guideline.pdf (accessed 2011).
2. **Marinez YN, McMahan A, Barnwell GM, Wigodsky HS.** Ensuring data quality in medical research through an integrated data management system. *Stat Med* 1984; **3**: 101–11.
3. **Martinson BC, Anderson MS, de Vries R.** Scientists behaving badly. *Nature* 2005; **435**(7043): 737–38.
4. **Steen RG.** Retractions in medical literature: How many patients are put at risk by flawed research? *J Med Ethics* 2011; **37**: 688–92.
5. **Steen RG.** Retractions in medical literature: How can patients be protected from risk? *J Med Ethics* 2011; **38**: 228–32.
6. **Paventi S, Parafati MA, Luzio ED, Pellegrino CA.** Usefulness of two-dimensional echocardiography and myocardial perfusion imaging for immediate evaluation of chest pain in the emergency department. *Resuscitation* 2001; 49: 47–51. Retraction in *Resuscitation* 2002; **54**(1): 107.
7. **Ben-Gal Y, Moshkovitz Y, Nesher N,** *et al*. Drug-eluting stents versus coronary artery bypass grafting in patients with diabetes mellitus. Retraction in *Ann Thorac Surg* 2007; 84(2): 712.
8. **Jacobs MJ, van Eps RG, de Jong DS, Schurink GW, Mochtar B.** Regarding 'Prevention of renal failure in patients undergoing thoracoabdominal aortic aneurysm repair'. *J Vasc Surg* 2006; **43**(2): 428–29; discussion 429.
9. **Cheng B-Q, Jia CQ, Liu CT,** *et al*. Chemoembolization combined with radiofrequency ablation for patients with hepatocellular carcinoma larger than 3 cm: A randomized controlled trial. *JAMA* 2008; 299(14): 1669–77. Retraction in DeAngelis CD, Fontanarosa PB. *JAMA* 2009; **301**(18): 1931.
10. **Nakao N, Yoshimura A, Morita H,** *et al*. Combination treatment of angiotensin-II receptor blocker and angiotensin-converting-enzyme inhibitor in non-diabetic renal disease (COOPERATE): A randomised controlled trial. Retraction in *Lancet* 2009; **374**(9697): 1226.
11. **Shafer SL. Notice of retraction.** *Anesth Analg* 2009; **108**(4): 1350.
12. **Abu-Omar AA.** Prevention of postpartum hemorrhage safety and efficacy. *Saudi Med J* 2001; 22: 1118–21. Retraction in *Saudi Med J* 2008; 29(10): 1523.
13. **Bakobaki JM, Rauchenberger M, Joffe N,** *et al*. The potential for central monitoring techniques to replace on-site monitoring: Findings from an international multi-centre clinical trial. *Clin Trials* 2012; 9: 257–64.
14. Guidance for industry oversight of clinical investigations – A risk-based approach to monitoring, draft guidance. Available at: http://www.fda.gov/downloads/Drugs/./Guidances/UCM269919.pdf (accessed December 2012).
15. **Evans SJW.** Statistical aspects of the detection of fraud. In Lock S, Wells F (eds). *Fraud and Misconduct in Biomedical Research* (2nd edn). BMJ Publishing Group, London, 1996, pp. 226–39.
16. **Buyse M, George SL, Evans S,** *et al*. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Stat Med* 1999; **18**: 3435–51.
17. **Baigent C, Harrell FE, Buyse M, Emberson JR, Altman DG.** Ensuring trial validity by data quality assurance and diversification of monitoring methods. *Clin Trials* 2008; **5**: 49–55.
18. **Taylor RN, McEntegart DJ, Stillman EC.** Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Inf J* 2002; **36**: 115–25.
19. **Al-Marzouki S, Evans S, Marshall T, Roberts I.** Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *BMJ* 2005; **331**: 267–70.
20. **Bailey KR.** Detecting fabrication of data in a multicenter collaborative animal study. *Control Clin Trials* 1991; **12**: 741–52.
21. **R Development Core Team.** *R: A Language and Environment for Statistical Computing*. 2005.
22. **Benford F.** The law of anomalous numbers. *Proc Am Philos Soc* 1938; **78**(4): 551–72.
23. **Lee SM, Woll PJ, Rudd R,** *et al*. Anti-angiogenic therapy using thalidomide combined with chemotherapy in small cell lung cancer: A randomized, double-blind, placebo-controlled trial. *J Natl Cancer Inst* 2009; **101**(15): 1049–57.
24. **Valle J, Wasan H, Palmer DH,** *et al*. Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. *N Engl J Med* 2010; **362**(14): 1273–81.
25. **Lee SM, Khan I, Upadhyay S,** *et al*. First-line erlotinib in patients with advanced non-small-cell lung cancer unsuitable for chemotherapy (TOPICAL): A double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol* 2012; **13**(11): 1161–70.
26. **Chernoff H.** The use of faces to represent points in k-dimensional space graphically. *J Am Stat Assoc* 1973; **68**(342): 361–68.
27. **Wolf P, Bielefeld U.** aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, and some slider functions. R package Version 1.2.6, 2012. Available at: http://CRAN.R-project.org/package=aplpack
28. **O'Kelly M.** Using statistical techniques to detect fraud: A test case. *Pharm Stat* 2004; **3**: 237–46.
29. **Svolba G, Bauer P.** Statistical quality control in clinical trials. *Control Clin Trials* 1999; **20**: 519–30.
30. **McNees P, Dow KH, Loerzel VW.** Application of the CuSum technique to evaluate changes in recruitment strategies. *Nurs Res* 2005; **54**: 399–405.
31. **El Emam K, Jonker E, Sampson M, Krleža-Jerić K, Neisa A.** The use of electronic data capture tools in clinical trials: Web-survey of 259 Canadian trials. *J Med Internet Res* 2009; **11**(1): e8.
32. **Barnett V, Lewis T.** *Outliers in Statistical Data* (3rd edn). John Wiley & Sons, Chichester, 1994.
33. **Hill TP.** A statistical derivation of the significant-digit law. *Stat Sci* 1995; **10**: 354–63.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

34. **Geyer CL, Williamson PP.** Detecting fraud in data sets using Benford's law. *Commun Stat Simul Comput* 2004; **33**(1): 229–46.
35. **Cox TF.** *An Introduction to Multivariate Data Analysis*. Hodder Arnold, London, 2005.
36. **Warnes GR.** (Includes R source code and/or documentation contributed by: Bolker B, Bonebakker L, Gentleman R, *et al*.) gplots: Various R-programming tools for plotting data. R package Version 2.11.0, 2012. Available at: http://CRAN.R-project.org/package=gplots

Appendix A

To run the R-programs automatically, the variables must appear in a specific order within the data set. To avoid the problem of different trials using different variable names, all of the programs are based on the order in which the variables are given; that is, the participant identification (ID) number must be in the first column and the site name/number in the last (in most cases).

Text 1

Details of the methods used

Participant-level data monitoring

1. Date checks

For each participant, all dates should occur after the first participant was randomised (or registered for single arm studies), and before final events such as death and the end of the trial (defined as the date the database was locked or the current date). Another check that can be made is whether certain dates fall on weekends or national holidays, when for some trials registration, randomisation, and some clinic appointments are unlikely to fall on these days. Care must be taken in choosing which dates to check, because dates of death, emergency treatment, or some clinic visits may occur at any time. The *date_order_check* program performs checks to detect dates which do not fall in the correct order and the *weekend_hol_check* program looks for dates which fall on a weekend or a national holiday.

2. Outliers

Outliers are observations which appear to be inconsistent with the rest of the data, usually appearing as too large or too small. Occasionally, we expect to observe some extreme values (especially from highly skewed distributions), and so these would not really be outliers. The methods we summarise here apply to any continuous measurement, and they compare the observed value for a single participant to those from all other participants. Outliers at the participant level are more likely to result from errors rather than fraud, because those who fabricate data try not to make them stand out too much from the rest to avoid detection ('Inliers'). Barnett and Lewis [32] comprehensively cover the theory of outliers and associated tests. Both univariate and multivariate checks can be examined using the *outlier_check* R-program.

Univariate outliers (each variable considered separately):

One method involves finding the mean and standard deviation (SD) for a single continuous variable over all sites and comparing each participant's value against these. For example, when checking body weight, we specify a value '*k*' in the *outlier_check* program, and all participants whose value is more than $\pm k$ SDs from the mean are flagged (Standard Deviation method). A value of $k = 2$ should yield approximately 5% (expected from a Normal distribution) and so could represent too many data checks. We therefore used $k = 3$ (though $k = 2.5$ is acceptable). The second method is based on Grubbs's test, which is similar to the Standard Deviation method, but it examines outliers iteratively; the most extreme value is identified and then removed before looking for the second most extreme value, and so on. Removing outliers each time will reduce the SD and prevent masking (i.e., extreme observations hidden or 'masked' by those even more extreme).

Multivariate outliers (several variables considered simultaneously):

Buyse *et al*. [16] suggest using the squared Euclidean distance from the mean for the detection of both multivariate inliers and outliers. This method aims to identify participants who have outliers for several variables (these may or may not represent data errors because a participant with a genuine extreme measurement for one variable might have similar extreme values for other variables). The value x_{ij} represents a measurement for participant i and variable j. The difference between x_{ij} and the mean of that variable over all participants (in all sites) (*i.e.* $\bar{x}_j$) is then divided by the SD for all participants (S_j) and squared

$$D_i = \sum_j \left(\frac{x_{ij} - \bar{x}_j}{S_j} \right)^2$$

Each participant then has a value D_i (based on several variables), and the mean of these over all participants is D. We can specify '*k*' (e.g., $k = 3$) in order to list participants who have a high D_i (i.e., more than k SDs from the overall mean D). Large D_i

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

indicates that the participant's data values considered together are more extreme. Aspects such as eligibility criteria could thus be checked.

We also developed a program to calculate the Mahalanobis distance. This is similar to the Euclidean distance, but also takes into account the correlation structure of the data: the distance, D_i, of the vector observation, x_i, to the mean vector, x, is calculated as

$$D_i = \sqrt{(x_i - \bar{x})' S^{-1} (x_i - \bar{x})}$$

where S is the observed covariance matrix. Observations with large D_i values may be potential outliers. The Mahalanobis distances should follow a chi-square distribution (for Normally distributed data), so the extreme values can be selected as those which exceed a critical value (based on a probability level specified when running the program).

Centre (site) level data monitoring

Checks performed at the centre level have two purposes. First, to identify systematic errors in trial conduct at a site (procedural errors), for example, due to the protocol not being followed correctly, incorrect measurement or recording of some variables, or adverse events are over or under-reported. These could be due to a genuine misunderstanding of the trial protocol by local staff. The second purpose is to detect fraud, that is, creating trial participants and associated data, or inventing data for real participants who have missing values [15]. The methods described below aim to flag a site that is discrepant to the rest, in order to consider an on-site visit for checking the procedures and original data.

1. Rounding and digit preference

We used two methods to identify whether too many data values are being rounded to a whole integer (procedural error), or whether data are being created (fraud).

The first, proposed by Taylor *et al*. [18], implemented with the *integer_check* R-program, examines the pattern of integers recorded over time (details in Table A5). A plot for each site is produced, in which there is a monotonic increase (line of identity) if all the numbers are non-integers; too much rounding shows up as horizontal lines.

The second method uses Benford's law [22] to identify falsified data. In many naturally occurring situations, the distribution of the first significant digit of any numeric variable is as follows: the probability that the first digit is i (where i = 1–9) is approximately $\log_{10}(i+1) - \log_{10}(i)$ [33]. Therefore, there should be about 30% leading 1s and only 6% leading 7s. The rule works best when examining many different variables at once. Benford's law is useful because people tend to be poor random number generators, so if they create false data, the law should break down. The *digit_preference* R-program examines all the specified numeric variables on a case report form (CRF) and applies a chi-square test to see whether the observed distribution of leading digits is very different from that expected [34]. However, Benford's law should only be used when the first digit can range between 1 and 9, so measurements restricted to say 100–300 cannot be used. In addition, the law tends to not fit some distributions (e.g., Gaussian). Because of these limitations, we also compared the distribution of leading digits from each site with that from all other sites together, using a chi-square test:

$$\chi^2_{\text{test}} = \sum_{j=1}^{9} \left(\frac{(S_{i,j} - A_{i,j})^2}{A_{i,j}} \right) N_i$$

$$S_{i,j} = \frac{\#\textit{leading digits } j \textit{ in site } i}{\#\textit{ observations in site } i}$$

$$A_{i,j} = \frac{\#\textit{ leading digits all sites } i \neq j}{\#\textit{observations in all sites } i \neq j}$$

$$N_i = \#\textit{ observations in site } i$$

2. Comparing means of variables between centres

Comparing the means of several variables between centres is best done graphically, for which there are various techniques [16,18,35]. One method is Chernoff faces, where each site is represented by a cartoon face, and the means of the variables determine the size and shape of various features of the face (eyes, nose, and so on). We developed programs to examine many variables simultaneously. Chernoff faces (*mean_check* R-program, using the aplpack [27] package) are potentially useful because people should be able to recognise differences between faces easily. If a particular site has mean values that are very different from the others, it might indicate that some participants have been fabricated, or those recruited are so different to other centres that they require investigation. The plots could also be used to examine baseline measurements, to see whether eligibility criteria are being followed correctly. Star plots (*mean_check*, using the aplpack [27] package) have a principle similar to Chernoff faces, in which each branch and each point of a star (a site) represent a different variable, and the length of these two features is determined by the means.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

Parallel coordinate plots simply plot the means of each variable at each site, joining the points together with a different colour line for each site.

Text 2

Comparing means of variables between centres

Chernoff face plots were obtained for 15 blood measurements in Study 12 (Figure A3). Sites 11, 59, and 78 appear discrepant to the others. The R-program produces summary statistics to examine the data further in unusual sites (Table A6). Much of the differences were due to the presence of extreme outliers for a few participants. While Chernoff faces are visually appealing, they have several limitations. First, the assessment is subjective. Second, the facial characteristics are heavily influenced by outliers, which if not identified and corrected beforehand, would make a site appear more discrepant than it really is. Removing extreme outliers from a site will change the features of the face both at the site and, to a lesser extent, all other sites (Figure A4). Site 11 does not appear as visually different to the other sites in Figure A4, as it does in Figure A3. Third, large differences could be missed depending on which facial feature is selected to represent each variable; for example, visually examining differences in the height of the eyes or nose width is more difficult than, for example, width of the mouth or hair. Therefore, the most important variables could be linked to features whose shape is more readily distinguishable. Fourth, although the human eye can often identify unusual patterns, choosing discrepant sites for trials with many (e.g., >50) centres is not easy. Fifth, Chernoff faces work with exactly 15 variables (because of 15 different features to manipulate). If there are more than 15 variables, only the first 15 are included, and if there are fewer, the first variables will be included multiple times until there are 15 in total.

Star plots (Figure A5) and parallel coordinate plots (Figure A6) are shown for Study 12, based on the same variables as in Figure A3. An advantage of both of these is that variables have equal weight on the plot, compared to Chernoff faces where the matching of variables to features is subjective. However, the star plot is difficult to interpret because there are many branches, and there is no easy way to tell which variable is associated with each branch. Although the parallel coordinate plot appears similarly confusing at first, it can easily be investigated interactively, but only if the appropriate software is available (e.g., JMP, SAS). However, this cannot be done in R automatically. Star and parallel coordinate plots, like Chernoff faces, are also affected by a few participants with outlying data values.

3. Inliers

Examining inliers could identify data that are too similar to the rest (instead of too different, that is, outliers). However, if data are fabricated, people tend to choose values close to the mean of the other observations, so that they would not be readily noticed [15–18]. It would be unusual for a genuine participant to have most or all of their variable values around the mean. Inliers can be detected using the same calculations as in multivariate outliers for participant-level data checks ('Outliers'), that is, for each participant i and several variables, we have the sum D_i. Differences show up more clearly on a logarithmic scale, where the smaller the differences from the overall mean, the larger and more negative the value of $\log(D_i)$. Having several participants with very small D_i in the same site may require further investigation. In our program *inlier_check*, we specified -3 SDs lower than the mean of $\log(D)$, over all participants.

Because few inliers were found with the observed data, we fabricated participants and data (in ABC-02) close to the mean values, using values for each variable that fell within $k\sigma$ of the mean of the observed data, where σ is the observed SD of the variable and k was a small value (one fabricated participant for each of $k = 0.5$, 0.4, 0.3, 0.25, 0.2, and 0.1). This test was performed for one small site (9 participants), one medium site (18 participants), and one relatively large site (49 participants).

4. Correlation checks

The method here (*correlation_check* R-program) examines correlations between variables within a site and then compares the correlation structures between sites [18,28]. The premise is that even if a researcher has created false data and used sensible values for a single variable, it is difficult to fabricate several variables that together are consistent with real data.

We can select a set of continuous variables from one CRF, and the R-program calculates the pairwise correlation matrix between these variables for each site ($r_{x,y}$). These can be displayed using a grey-scale plot (using the gplots [36] package). This plot creates a grid of squares (one square represents a correlation, and the colour depends on the size of the correlation coefficient. Sites with high positive correlations are black, high negative correlation are white, and everything else a shade of grey, with lighter greys for the correlations nearer -1 and darker greys for the correlations closer to $+1$.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

The plots are a quick way to see whether any site appears different. The program can also perform a formal test to compare the correlation structure of each site with the overall correlation structure.

The formal test calculates a value of d^*, where d^* is the sum of the squared differences between the correlations in each site and the correlations in all the data (using the pairwise correlation matrix for all data, $R_{x,y}$)

$$d^* = \sum_{\substack{x,y \\ x \neq y}} \left(r_{x,y} - R_{x,y}\right)^2$$

The program then generates 1000 'pseudo' sites for each real site. For example, if there are 30 participants in site Z, 1000 'pseudo' site Zs are created by randomly selecting 30 participants out of all those in the trial. The value of d^* for each site is calculated, and a count is set up to record how many of these d^* values exceed the d^* for our chosen site. If fewer than 1% (or another pre-specified value) are larger (this percentage is our p-value) we conclude there may be reason to be suspicious and should perhaps have a closer look at the data.

To examine this method further, we created sites (for the ABC-02 trial) and compared them with the genuine ones. This was done by removing one large site from the data set and creating new ones with varying numbers of patients (10, 15, 20, 25, 30, and 35). The new sites were created in two different ways. In the first method, we generated patients by randomly choosing a value for each continuous variable from all of the values of that variable in the removed site. The second method involved randomly generating values for each variable from a Normal distribution with the mean and SD calculated from the removed site. Any values below or above the lower and upper limits observed in the removed site were generated again. This is perhaps a more realistic method of data generation as it would favour values close to the mean. Both methods involve creating data for each variable separately, without considering its correlation with any other variable. The results were similar using either method: the fabricated sites were the only ones that produced a p-value < 0.01.

5. Variance checks for repeated measures

Evans [15] suggested looking at the variance of repeated measurements, and we used a technique [20] which examines variances when there are multiple records for each participant (*variance_check* R-program). Important continuous variables are selected, and the variance of these for each participant is found. The mean and SD of the variances between participants is calculated, and any individual whose variance is more than $\pm k$ SDs from the mean are flagged (e.g., $k = 3$). If data are fabricated, it might be difficult to do so over several time points, because the false values may not vary enough compared to real data.

Sometimes, the SD of the variances could be large in comparison to the mean, so participants with small variances are unlikely to be detected. To overcome this, if the mean minus SDs is less than zero, the lowest 2.5% of participants in that site are also identified.

The continuous variable is plotted against an index number which orders participants, first by site then by date of randomisation, in a scatter plot. Participants with extreme values are automatically plotted in different colours to make it easier to see whether several appear in the same site. Small variances could indicate fabricated data, where the data are too similar to the other measurements for the participant, and these are coloured in shades of red/pink. Participants with large variances (possible data errors) are coloured in shades of blue.

An option is included in the R-program to look for zero change between repeated measurements (rather than small variance). This removes the problems caused by large outlying values distorting the overall variance; however, it may only be useful to identify procedural errors (investigators writing down previous test results) rather than falsified data.

6. Categorical variables

Most of the data checks above are meant for continuous variables. The R-program *cat_check* can be used for categorical variables. It compares the number of observations at each level in a given site with the frequency distribution seen in all of the other sites. Frequency tables for each site are provided in the output, including a chi-square test on sites that have an expected count of ≥ 5 participants in each level (or other specified minimum), and p-values.

7. Adverse events

In trials of participants who are already ill (patients), interest is often in serious adverse events (SAEs), and these are the ones that must be reported to the coordinating centre. In studies of healthy participants, SAEs might be uncommon, so there could also be interest in mild or moderate events. The following method can be applied to any definition of adverse event, but the purpose is to identify sites that appear to have too few, which may require an on-site visit.

When examining the number of SAEs per site, we need to allow for the number of participants recruited, as well as the length of time in the trial.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

We developed a simple approach, where for each site, we calculated the SAE rate as the number of participants with at least one SAE divided by the total number recruited, which is further divided by trial duration at that site. The trial duration at each site is taken as the time between the first randomisation and the last expected SAE report date/date of data dump (whichever comes first). The date of the last expected SAE report is calculated as the date of the last randomisation plus the number of months we expect to receiving SAEs; for example, if a treatment lasts 6 months, we may add 7 months for the SAE reports to be sent to the coordinating centre. This SAE rate is plotted against the number of recruited participants in the site. The *SAE_check* R-program identifies centres with the highest and lowest 10% of rates (or alternative proportion) and shows all sites in relation to the average SAE rate for the whole trial.

We also had another approach where the duration in the trial for each participant was used instead of overall duration at the site. However, we still limited the number of days each participant could contribute to the total duration, by only covering the period in which we expect to receive SAEs (i.e., slightly longer than the time the participant is on treatment). This is because in trials where participants have long life expectancies, those sites which opened early may have many participants with long follow-up times, falsely deflating the SAE rate in comparison to those which had opened more recently.

The key aspect is to flag centres that are relatively large, but with a low number of SAEs (because low rates could arise by chance from centres with few participants).

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

Table A1. Output from the R-program examining leading digit preference, using 26 variables on a chemotherapy case report form from the Study 12 trial (of 79 recruiting sites, 66 were examined, and 37 of these are shown below)

Site ID number	Number of data values	Number of patients	Chi-square p-value comparing observed % with Benford's law (26 variables)	Chi-square p-value comparing observed % with observed % from all other sites (26 variables)
1	295	8	0.006	0.277
2	359	10	0.167	0.616
3	337	8	0.19	0.345
4	436	11	0.048	0.062
5	548	15	0.025	0.781
6	379	9	<0.001	0.130
7	380	14	0.197	0.759
8	278	7	0.205	0.608
9	359	10	0.263	0.174
10	138	3	0.399	0.916
11	1150	27	<0.001	0.767
12	358	10	0.043	0.509
14	1041	31	<0.001	0.003
15	194	7	0.046	0.058
16	525	13	0.009	0.192
17	295	8	0.209	0.213
19	338	10	0.261	0.453
⋮	⋮	⋮	⋮	⋮
58	605	17	0.005	0.186
59	188	5	0.022	0.001
60	530	12	0.006	0.003
62	117	4	0.132	0.084
63	244	7	0.007	0.011
64	160	4	0.144	0.119
66	186	4	0.386	0.383
69	212	6	0.011	0.109
70	120	3	0.073	0.061
71	252	7	0.005	0.140
72	501	12	0.469	0.989
73	356	9	0.105	0.801
75	282	7	0.204	0.391
76	291	7	0.189	0.761
78	161	7	0.654	0.868
79	224	5	0.101	0.776
81	469	13	0.163	0.850
83	250	6	0.007	0.251
88	192	6	0.146	0.359
93	161	4	0.524	0.783

ID: identification.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

Table A2. A fabricated site ('FAKE' in table) with 25 patients, 9 SAEs and 30 months between the first and last randomisation was added to the data

Site ID code	Rate	Patients with an SAE	Total number of SAEs	Number of patients	Time in trial (months)
Lowest 10%					
ARI	0	0	0	1	5.5
BRI	0	0	0	1	5.5
SOU	0.004	1	1	8	34.08
UCL	0.008	5	5	18	33.29
WES	0.009	1	1	4	26.43
FAKE	0.010	9	9	25	35.46
Highest 10%					
DRI	0.082	2	2	2	12.20
SAL	0.115	2	2	2	8.72
BEL	0.155	5	6	5	6.45
DUN	0.182	1	1	1	5.5
GLA	0.182	1	1	1	5.5
WRE	0.182	1	1	1	5.5

SAE: serious adverse event; ID: identification.
Output from the R-program that lists sites that have SAE rates in the lowest and highest 10th centile in ABC-02.

Table A3. A fabricated site with 35 patients, 6 SAEs and 45 months between the first and last randomisation

Site ID code	Rate	Patients with an SAE	Total number of SAEs	Number of patients	Time in trial (months)
Lowest 10%					
ARI	0	0	0	1	5.5
BRI	0	0	0	1	5.5
FAKE	0.003	6	6	35	50.48
SOU	0.004	1	1	8	34.08
UCL	0.008	5	5	18	33.29
WES	0.009	1	1	4	26.43
Highest 10%					
DRI	0.082	2	2	2	12.20
SAL	0.115	2	2	2	8.72
BEL	0.155	5	6	5	6.45
DUN	0.182	1	1	1	5.5
GLA	0.182	1	1	1	5.5
WRE	0.181818	1	1	1	5.5

SAE: serious adverse event; ID: identification.

Table A4. The maximum number of patients with an SAE that a site can have for it to fall within the bottom 10% of all sites, for specified numbers of patients and lengths of time in the ABC-02 trial. The numbers in the table below would differ for other trials, where the SAE rate is different

Patients	Time in trial[a] (months)				
	5	10	15	30	45
5	0	0	1	1	2
10	1	1	2	3	5
25	2	4	5	9	13
35	3	5	7	13	19
50	5	8	11	19	27

SAE: serious adverse event.
[a]Time between the first and last recruitment + 5.5 months.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

Table A5. Illustration of the method by Taylor *et al.* [5] to identify rounding

Participant index	Body weight (kg)	Date (in order)	X	Cumulative sum of X
1	75.5	11 January 2010	1	1
2	65.3	13 January 2010	1	2
3	68	22 January 2010	0	2
4	89.6	01 February 2010	1	3
5	90	02 February 2010	0	3
6	64	18 February 2010	0	3
7	78.8	17 March 2010	1	4
8	55.7	24 March 2010	1	5
⋮	⋮	⋮	⋮	⋮

The table is based on participants from the same site.
For example, consider baseline body weight. The weights are ordered by the date recorded, and a binary variable (X) is generated to record whether it is an integer (X = 0) or not (X = 1). The cumulative sum of X is then plotted against the participant index. If all the body weights were non-integers, this would form the line y = x. The line will be horizontal where an integer is recorded. If a site appears to have periods with many integer values in a row or more frequent integer values than other sites, then this site could be investigated to check that their staff understands the correct level of accuracy required.

Table A6. Summary statistics for site number 78 in Study 12 that appeared to be unusual, when examining the Chernoff faces diagram (Figure A3). The table shows the individual data values for 7 patients and 4 variables

Patient ID	wbc	alt	alp	LDH
1	17.53	N/A	N/A	N/A
2	13.69	N/A	203	317
3	19.74	24	301	270
4	10.4	21	284	520
5	15.12	N/A	130	N/A
6	15.24	N/A	1728	4848
7	18.47	N/A	N/A	577
Site means	15.74	22.5	529.2	1306.4
Overall means	10.66	41.40	148.10	732.69
Facial feature controlled	Face; width	Eyes; height	Eyes; width	Hair; width

ID: identification.
A single patient with large outliers (highlighted) has skewed the site means for two variables (pre-treatment alkaline phosphatase (alp) and lactate dehydrogenase (LDH)). These patients were also picked up by the outliers R-program. The 'alt' variable (alanine aminotransferase) had a site mean about half the overall mean, but this was based on just two values: there was much missing data. The 'wbc' (white blood cells) variable has a high mean relative to the overall value but no outliers – this site could be investigated to see whether the values were being recorded correctly.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

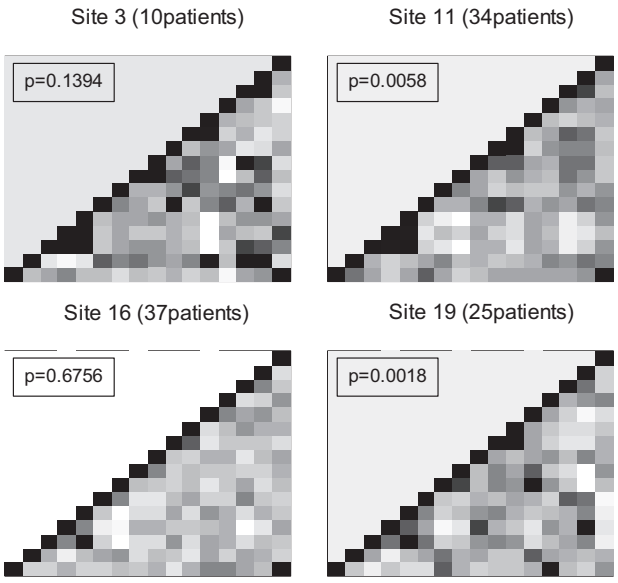

Figure A1. Correlation checks for the TOPICAL trial using 16 variables (each square represents a correlation between a pair of variables). Sites 11 and 19 were the only two that had p-values below 0.01.

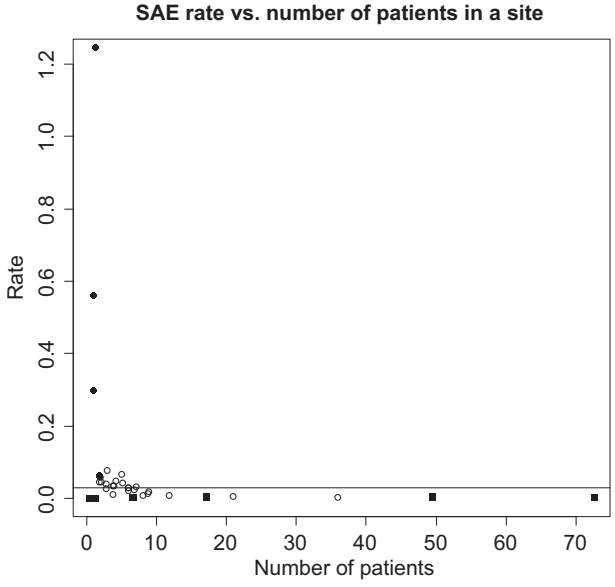

Figure A2. Examining SAE rates (ABC-02).
SAE: serious adverse event.
Again each point represents a site, the y-axis is the SAE rate per site, allowing for the number of patients and the time in the trial. The lowest 10% of SAE rates are shown as black squares and the highest 10% as solid black circles. However, in this version, the time in the trial is taken as the sum of the time each patient has been in the trial (cut-off at 5.5 months per patient) to create the rate rather than the time the site has been open. Note that the two biggest sites fall within the bottom 10% (shown as black squares).

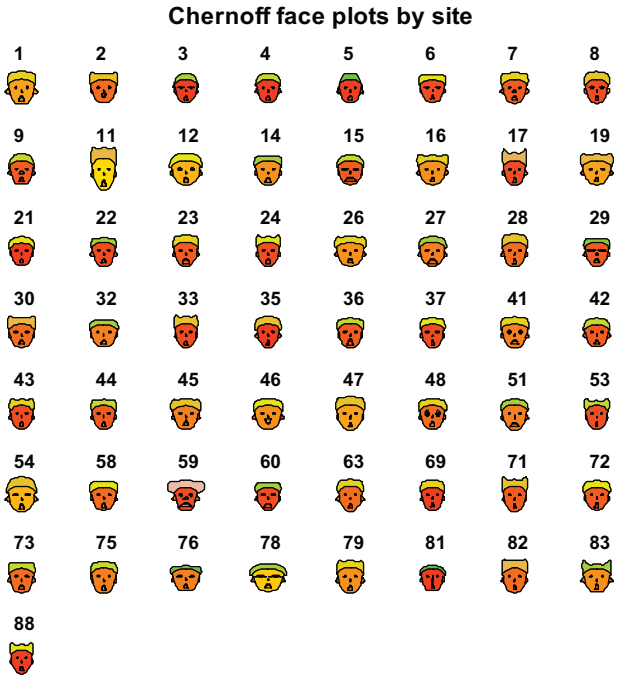

Figure A3. Chernoff faces for a single case report form (start of chemotherapy) for the Study 12 trial (15 variables).
Three sites (11, 59 and 78) appear visually to be discrepant to the others.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015

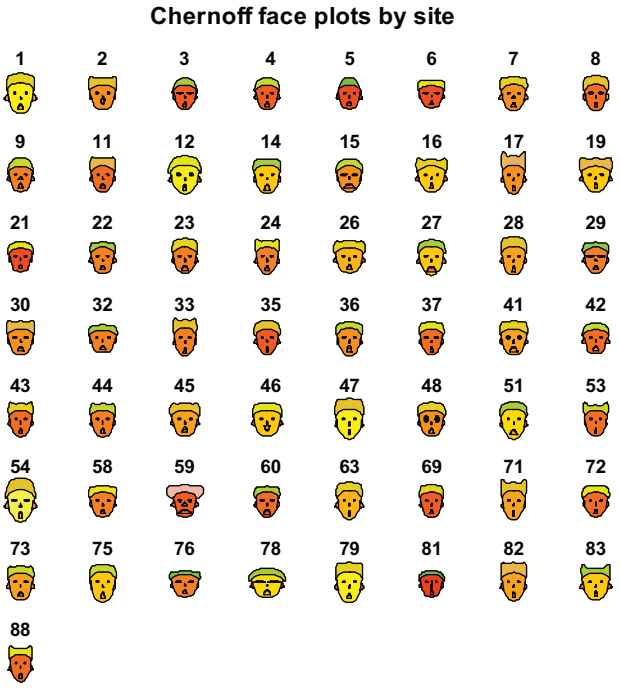

Figure A4. Chernoff faces for a single case report form for Study 12 (start of chemotherapy form; 15 variables) after one participant with a single outlier for one variable in site 11 had been corrected.
Site 11 appeared unusual in Figure A3, but not so much now.

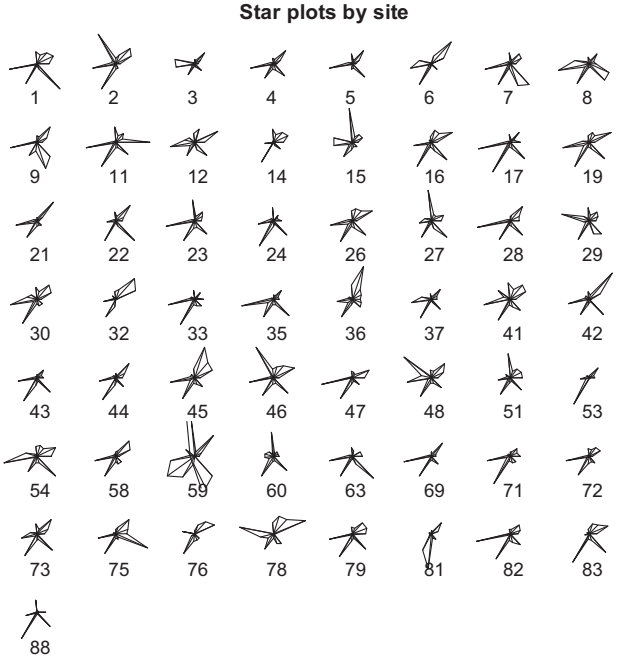

Figure A5. Star plot for a single case report form (start of chemotherapy) for Study 12 trial (15 variables).

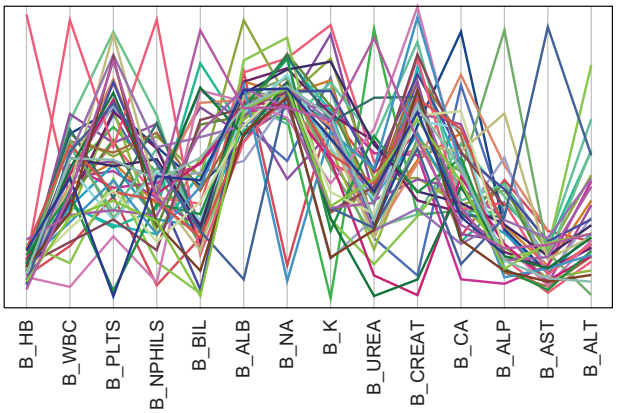

Figure A6. Parallel coordinate plot for a single case report form (start of chemotherapy) for Study 12 trial (14 variables).
Each coloured line represents a different centre.

Downloaded from ctj.sagepub.com at University College London on June 25, 2015