THE HOPE AND THE HAZARDS OF USING COMPLIANCE DATA IN RANDOMIZED CONTROLLED TRIALS

STUART J. POCOCK* AND MONA ABDALLA

Medical Statistics Unit, London School of Hygiene & Tropical Medicine, London WC1E 7HT, U.K.

SUMMARY

This paper aims to elucidate both the advantages and limitations of using compliance data in the reporting of treatment differences in clinical trials, illustrating the issues with some recent examples. While analysis by intention-to-treat should remain the principal reporting approach for most major clinical trials, arguments are put forward as to why supplementary analyses taking account of compliance can be of value. However, continued recognition of the potential biases inherent in all such selective analysis is of key importance. Some of the possible analytical approaches are presented along with suggestions on interpretation. Particular emphasis is on one case study, a large European trial in obesity incorporating repeated measures of weight loss, drug plasma level and pill count data. In working on compliance data in clinical trials, the statistician’s main responsibilities may be to undertake a cohesive analysis strategy not influenced by data dredging, to achieve clarity of exposition without undue complexity or oversimplification, and to provide appropriately cautious interpretations which take account of selection biases and data limitations. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

Analysis by intention-to-treat rightly remains the main statistical approach for presenting the comparative results of different treatment policies within a randomized controlled trial.1 Some authors doubt the usefulness of any other analyses comparing treatments because they are prone to selection bias.2–4 However, there is growing interest in exploring more complex statistical approaches which incorporate measures of individual patient compliance with the intended treatment regimens into supplementary comparative analyses.5–7

It is well recognized that crude ‘on treatment’ analyses confined to ‘compliers only’ are prone to produce biased treatment comparisons because non-compliance is not a random occurrence and may well be associated with a poorer (or conceivably a better) outcome. Also, compliance is often not simply a dichotomous measure: there are gradations between full-compliance and total non-compliance which one can attempt to measure, for example, with pill counts or drug plasma levels, and also patients’ compliance status can fluctuate over time.

* Correspondence to: Stuart J. Pocock, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT, U.K. e-mail: S. Pocock@lshtm.ac.uk

CCC 0277–6715/98/030303–15$17.50
© 1998 John Wiley & Sons, Ltd.
Thus, the underlying hope of much current statistical research into non-compliance is that such biases and crudities can be overcome by more sophisticated modelling of the compliance process.\textsuperscript{5,7} The preliminary intention in any given trial should be to understand and document the observed patterns of compliance that occur in each treatment group. One then wishes to quantify the inter-relation between compliance and patient response. Both of these steps lead up to the more important goal of enhancing insight into the ‘true’ treatment effects (and treatment differences) that occur at different levels of compliance, usually with a particular interest in quantifying the magnitude of true treatment difference in patients who are fully compliant.

However, it is important to recognize that any treatment comparisons which allow for non-compliance are potentially hazardous. Only analysis by intention-to-treat can be relied on to provide an unbiased comparison of the treatment policies as implemented. All other analyses deviate from the principle of randomized comparison, need to make assumptions that cannot be fully validated and hence carry a risk of introducing bias.\textsuperscript{8} This is not automatically a heinous crime; after all, observational epidemiology is not based on randomization but few of us would readily dismiss non-trial epidemiology as a useless pastime. Rather we need to recognize the potential for confounding, take what steps we can do to remove it, and always tinge our conclusions with cautious overtones about the limitations of non-randomized inferences.

There is also a danger that the data analyst accepts compliance data as presented, without further exploring the trial’s practical circumstances and the data’s limitations. For instance, appropriate questions might be:

1. Are the individual patient’s compliance data recorded in sufficient detail?
2. How was such compliance data obtained, and was the process identical for both treatment groups?
3. Were reasonable efforts made to document the reasons for patient non-compliance, paying particular regards to adverse events (side-effects) and disease progression?
4. Were reasonable efforts made to continue to collect patient outcome data in non-compliers, or is non-compliance largely linked to missing data?
5. Are the measures of patient compliance reliable? For instance, routinely collected pill count data are prone to over-represent compliance, and it may be unrewarding to invest too much statistical effort into such suspect data.\textsuperscript{9,10}
6. Is the trial of sufficient quality to be worth analysing compliance data in any depth? For instance, if substantial non-compliance has occurred more because of poor trial management and patient education rather than because of sensible and informed patient and investigator behaviour then one might question whether any useful inferences can be drawn.

Even with those situations in which there exist good quality compliance data in a well-run trial, there is still concern about whether analyses taking account of compliance have sufficient credibility to influence judgements on the relative merits of the different treatments. Many trial protocols require clear prior specification of the primary analyses, whereas analyses incorporating compliance are often done as an exploratory afterthought. No matter how clever the modelling techniques, there remains the potential danger that selective choices of method and presentation could, perhaps subconsciously, enhance treatment differences. Thus, it is important that the reporting of analyses with compliance data aims to avoid (and be seen to avoid) such selection biases.
Taking into account all these important provisos, I think a case can be made for analyses that incorporate compliance data as a valuable complement to the prime formality of analysis by intention-to-treat.

Section 2 illustrates with brief examples some of the common problems inherent in both analysis by intention to treat and in compliance data. The remainder of this paper explores one particular example, an obesity trial, in greater depth in the hope that such a case study will provide some generalizable reality to the debate on how to handle compliance data.

2. BEYOND ANALYSIS BY INTENTION-TO-TREAT

There have been many articles expressing the virtues and the limitations of analysis by intention-to-treat and so we will forgo reiteration of the basic arguments. In general, we feel that such an analysis (when feasible) should be the prime presentation of results for most major randomized trials and the principal conclusion should relate to that analysis. However, that should not deny one the opportunity in supplementary analyses to explore the reasons behind such a prime finding. For instance, the Medical Research Council (MRC) trial of treatment of hypertension in older adults revealed an interesting and unexpected finding from analysis by intention-to-treat; that is, the rates of cardiovascular death or cardiovascular events were significantly higher in the beta-blocker group compared with the diuretic group. Pre-defined strategies for alternative and additional drugs (because of side-effects or inadequate blood pressure control) meant that both the beta-blocker policy and diuretic policy groups experienced substantial individual changes in treatment over time. White and Pocock describe a variety of analytic methods incorporating such treatment changes, and all tended to affirm the original intention-to-treat findings. Such supportive evidence from analyses of compliance (or treatment change) is valuable in reinforcing an analysis by intention-to-treat, but what happens if such supplementary analysis is contradictory?

An interesting case concerns changes in serum cholesterol in the same elderly hypertension trial. One year after randomization, analysis by intention-to-treat revealed small but highly significant mean increases in serum cholesterol in both the diuretic and the beta-blocker groups compared to placebo, mean differences $+0.12 \text{mmol/l}$ ($P = 0.001$) and $+0.08 \text{mmol/l}$ ($P = 0.003$), respectively. This was not unexpected in the diuretic group, but we were puzzled by the apparent lipid effect of beta-blockade. However, 30 per cent of the beta-blocker group were receiving diuretic by one year, either instead of or in addition to beta-blocker. An alternative additive regression model based on treatment actually received at one year (diuretic, beta-blocker, both or neither) for patients in all three treatment groups combined revealed a highly significant increase on diuretic, $+0.11 \text{mmol/l}$ $P < 0.001$, but no evidence of an increase on beta-blocker, $+0.03 \text{mmol/l}$ $P = 0.2$. The analysis by intention-to-treat is not wrong as such, in that it correctly revealed an increase in serum cholesterol in the beta-blocker group, though it is better to think of that group as a policy commencing with beta-blocker alone. However, the logical explanation for that finding is the cholesterol-raising effect of the diuretic within that treatment policy. This example emphasizes the need to clarify what is of prime interest, the policy or the drug; these are often closely related but in complex trials such as the above it is important to disentangle the two.

In many clinical trials analysis by intention-to-treat is not feasible, since for patients who withdraw from follow-up visits, hopefully few in number, the necessary visit-related outcome measures cannot be obtained. This raises two sometimes neglected issues: (i) the likely close relation between non-compliance with treatment and non-compliance with follow-up (that is,
withdrawal); and (ii) the distinction between visit-related outcomes (for example, quantitative measures, symptoms) and events during continuous follow-up (for example, death, development of serious non-fatal disease). The former cannot usually achieve full intention-to-treat (it is a case of how near can one achieve) whereas for the latter it should be possible in a well-run trial.\textsuperscript{15}

An example of the former is the Multi-centre Anti-Atheroma Study (MAAS)\textsuperscript{16} which compared simvastatin with placebo over 4 years follow-up in 381 patients with coronary artery disease. The two primary outcome measures were the mean changes over 4 years in the mean and minimum lumen diameter of pre-selected segments of coronary arteries, as measured by quantitative coronary arteriography. It was recognized in advance that not all patients would attend the 4 year visit, and hence there ensued an interesting debate by the trial steering committee on how exactly to define these primary endpoints. The basic choice is whether to only use 4 year data (that is, a pure endpoint, but on fewer patients) or for those patients not reaching the 4 year visit to substitute data from an earlier angiogram if available (that is, a more mixed endpoint but on more patients). In fact, the latter choice was made by (a) using pre-intervention angiograms for patients undergoing PTCA or CABG before 4 years and (b) using 2 year angiograms for other patients without 4 year data. It turned out that 345 patients (91 per cent) contributed a primary endpoint: 248 at 4 years, 35 pre-PTCA or CABG and 62 substituted from 2 years. Such a ‘last observation carried forward’ approach is based on the judgement that adding in earlier data in some sense gets one nearer to the underlying true treatment difference than simply deleting such patients from analysis. It is hard to check the wisdom of the judgement, but if those missing a visit are less compliant and/or experience a poorer response and there is not too marked an overall trend in response over time, then it is likely to get one nearer to the unachievable analysis by intention-to-treat, that is, a fair comparison of the randomized policies. Other approaches for handling missing data with informed censoring are the assignment of scores or formal penalties to missing data.\textsuperscript{15} The former also introduces arbitrariness into the analysis and the latter produces tests that are more conservative than the usual approaches to missing data that assume censoring is random.

The likelihood that non-compliers will also withdraw from follow-up can pose a particular problem if they are unevenly split across treatment groups. For instance, in a Dutch trial\textsuperscript{17} comparing inhaled corticosteroid with placebo in 116 children with asthma (both groups received inhaled beta-2 agonists) more patients withdrew on placebo than on corticosteroid (26 and 3, respectively) largely due to exacerbations of asthma. This makes it difficult to interpret quantitatively the results for the pre-defined primary outcome measures forced expiratory rate in one second (FEV\textsubscript{1}) per cent predicted and the dose of methacholine necessary to reduce FEV\textsubscript{1} by 20 per cent of its baseline value (PD\textsubscript{20}) as measured every 2 and 4 months, respectively, over 2 years follow-up, since clearly those for whom measurements exist become increasingly non-representative of those randomized. However, a qualitative inference was made easy by the fact that in an ‘on treatment’ analysis the corticosteroid group had significantly higher mean FEV\textsubscript{1} and PD\textsubscript{20} than the placebo group at all time points. In other examples with such informative censoring the on treatment analysis and the withdrawal pattern may not be so mutually supportive, in which case we see no real quantitative solution without making quite strong assumptions which might well not be justified.

3. USE OF COMPLIANCE DATA IN AN OBESITY TRIAL

A European multi-centre double blind randomized trial compared dexfenfluramine (dF) with placebo over 1 year’s follow-up in 822 obese patients. Following the initial publication\textsuperscript{18} we were
asked to re-analyse the trial data paying particular regard to the measures of patient compliance. Of prime interest were the plasma concentrations of fenfluramine (F) and its metabolite norfenfluramine (nF) taken at 6 and 12 months, and also pill count data were available at 1, 2, 4, 6, 8, 10 and 12 months. We hope this brief report is of general interest in illustrating how relatively straightforward analysis techniques can be used to explore compliance data and its relation to patient response (weight loss), and to air discussion on the value and limitations of such an exercise in making inferences about treatment effects. Obesity is a particularly interesting field since substantial numbers of patients are prone not to fully comply either with drug treatment or dietary advice, and missing data, both on outcome and compliance measures, are an added complication.

Our starting point is the original publication’s main display of results (Figure 1). To their ‘completers only’ analysis (that is, the 256 dF and 227 placebo patients with all weight measures taken) we have added the mean and standard error of weight loss as a percentage of initial weight for all available data at each visit. The latter is more complete, involves differing numbers of patients at each time point and slightly reduces the mean weight loss in both groups at all time points. However, the pattern of observed treatment differences is similar in both analyses, with the dF group having around an extra 3 per cent weight loss by 6 months ($P < 0.001$) which is maintained at 12 months.

Our aims were to investigate the impact of non-compliance on the observed treatment effect, to try and estimate the ‘true’ treatment effect in full compliers (that is, patients taking the intended 15 mg oral dF twice daily) and to determine whether the observed lack of additional weight loss
between 6 and 12 months is due to non-compliance or true non-effect, while paying due regard to
the limitations and assumptions inherent in such data explorations. We concentrate primarily on
0 to 6 month weight loss, but also briefly summarize findings for the 0 to 12 month period. It is
useful to clarify exactly what one is trying to estimate here. While the average effect in the study
population if all patients were fully compliant may be conceptually interesting it has little
practical meaning and is difficult to estimate in view of selection biases. Thus, it is probably more
meaningful and realistic to estimate the true effect in the subset who would be fully compliant in
practice. This is then highly relevant to the patient considering starting active treatment who can
be informed ‘if you comply fully, the predicted effect is as follows’.

3.1. Plasma levels and weight loss

The short-term plasma marker of dexfenfluramine (dF) is the fenfluramine (F) concentration, and
this metabolizes into norfenfluramine (nF) which has a longer half life. Both F and nF plasma
levels were measured in 277 and 220 dF patients at 6 and 12 months, respectively, which
represents 86 per cent and 75 per cent of patients attending their 6 and 12 month visits. Some
centres did not collect plasma samples consistently, but patient refusal may also have played
a part in the missingness. While nF level is less prone to short-term variation (that is, time since
last dF dose) it is more prone to laboratory variation since it is on average around half the F level.
Hence, we use F + nF level in ng/l as liable to be the most stable simple plasma measure of drug
metabolism.

Figure 2 shows the relation of F + nF level at 6 months with weight change at 6 months in 277
dF patients. The choice of data display has an impact on the initial perception of the strength of
association. The scatter diagram, like most such diagrams with hundreds of points, suggests
a weak association though the fitted regression line has a highly significant negative slope
\( P < 0.001 \). Note that 32 patients had zero F + nF level indicating their total recent non-
compliance. Also in Figure 2, we show the mean and SEM of per cent weight change for patients
ranked and then grouped according to their F + nF level. The ten groups are of intended equal
size (around 28 per group) but this varies slightly due to tied F + nF levels. There is a clear inverse
linear trend except for the highest group (F + nF level > 64 ng/l). Those dF patients with missing
plasma measures \( n = 46 \) have a mean weight loss near to the mean for all dF patients, while the
mean weight loss for placebo patients \( n = 310 \) is similar to that for dF patients with zero plasma
level. Note that a further 81 dF patients and 108 placebo patients did not have weight measured
at 6 months.

The results of a multiple regression applied to these data are shown in Table I. To induce
adequate linearity, 11 patients with F + nF > 80 ng/l were given a value equal to this cut-off
point. The results confirm a highly significant inverse association between F + nF level and per
cent weight change. The intercept of 7.6 per cent weight loss, is not surprisingly, close to the mean
for zero compliant patients, and the non-significant regression coefficient for placebo indicates
their similarity to the intercept. The coefficient for the missing plasma group suggests a mean
weight loss compatible with an unobserved F + nF level of 28 ng/l, though with a wide
confidence interval.

It is of interest to estimate the mean weight loss of patients who are fully compliant with the dF
regimen. One difficulty with plasma levels is that they do not only measure compliance but will
also reflect drug bioavailability as affected by the individual’s metabolism. Previous shorter-term
and more tightly controlled studies of the same dF dosage, 15 mg twice daily, are of use. One

Figure 2. Per cent weight change at 6 months by plasma F + nF level at 6 months for 277 patients randomized to dF (a) as a scattergram with regression line fitted, and (b) mean (and SEM) of per cent weight change by mean plasma level for subjects in 10 approximately equal groups ranked according to plasma level. Placebo and missing plasma groups also plotted.

small study \((n = 11)\) with tight monitoring of drug taking gave a mean \(F + nF = 56.1\) ng/l (SD 16.6) whereas several other small studies with subjectively declared compliance had mean 46.4 ng/l (SD 15.2). Hence, for simplicity we will use 50 ng/l as the plausible mean \(F + nF\) level in fully compliant patients. This compares with an observed 35.3 ng/l and 31.6 ng/l in the current trial’s dF group at 6 and 12 months, respectively.

From the regression model in Table I the estimated mean weight loss for \(F + nF = 50\) ng/l is 11.03 per cent (standard error 0.46 percent) assuming random partial compliance. Compared to the mean placebo group weight loss of 7.08 per cent (SEM 0.36 per cent), we can estimate an
observed weight difference between ‘fully compliant’ dF patients and all placebo patients of 
−3.96 per cent with 95 per cent confidence interval −2.8 per cent to −5.1 per cent. This is 
somewhat greater than the −2.8 per cent observed weight difference at 6 months between all dF 
and all placebo patients as shown in Figure 1.

The key issue here is how we interpret such an estimate of apparent effect in full compliers. 
Some might argue it has no purpose since what matters is the policy of prescribing dF to all 
patients and the consequent average effect across all patients, that is, stick to the intention-to-
treat philosophy. Actually one cannot achieve that global an estimate since some patients in both 
groups withdraw and were not weighed at intended visits. Also, the policy as implemented in the 
trial may have led to more non-compliance and withdrawal than could be achieved in a partic-
ularly highly motivated centre’s patients. At least conceptually it is worth considering more ideal 
circumstances than those achieved. Another consideration is the perspective of the individual 
patient who might wish to argue ‘Never mind all those other non-compliant patients who lower 
the average benefit. Suppose I comply properly, how much can I expect to benefit’. This point of 
view carries more weight (excuse the pun!) when the reasons for non-compliance are primarily 
a matter of individual patient choice rather than side-effects of treatment, as is to some extent the 
case here.

Provided one accepts the regression model assumptions and the estimated plasma levels 
achieved in full compliers, then the weight loss estimate in full compliers is unbiased in the sense 
that it can plausibly be related to a representative sample of full compliers. However, such a sample 
(and its corresponding population beyond the trial) is a post hoc selection since one cannot determine in advance who will be fully compliant. More importantly, the consequent 
comparison with placebo patients should ideally relate to placebo compliers, with the additional 
assumption that the compliance selection process is the same in both randomized groups, 
principles that underlie the work of Efron and Feldman. Of course with plasma level markers 
there is no equivalent compliance data for the placebo group, in which case we need to consider 
the validity of interpreting the ‘dF compliers versus all placebo’ contrast. One useful fact is the 
similarity between the intercept and the placebo mean in Figure 2. Since placebo non-compliers 
and dF non-compliers should have identical weight loss (both receive nothing in a double blind 
trial) this adds some indirect support to the idea that the placebo group’s weight loss/compliance 
relationship would be very weak (or possibly absent), adding plausibility to a simple interpreta-
tion that the observed dF effect in compliers is due to the treatment itself.

Zeger and Liang offer a method for estimating the slope of the weight change at 6 month in 
relation to compliance in the placebo group, assuming that the intercept of both the dF and 
placebo groups are equal, that the weight change relates linearly to compliance and that the

---

**Table I. Multiple regression relating 6 month per cent weight loss to plasma F + nF level**

<table>
<thead>
<tr>
<th>Regression coefficient</th>
<th>Standard error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−7.61</td>
<td>0.69</td>
</tr>
<tr>
<td>F + nF level* (ng/ml)</td>
<td>−0.069</td>
<td>0.017</td>
</tr>
<tr>
<td>Missing plasma</td>
<td>−1.91</td>
<td>1.15</td>
</tr>
<tr>
<td>Placebo</td>
<td>0.53</td>
<td>0.77</td>
</tr>
</tbody>
</table>

* F + nF exceeding 80 mg/ml set to that cut-off
compliance distribution is the same in both dF and placebo groups. This method actually gives a positive estimate for the placebo slope, 0.015 per cent per ng/ml, suggesting that our comparison of the ‘dF compliers versus all placebo’ is conservative in relation to a comparison of ‘dF compliers versus placebo compliers.

However, some concern about possible undetectable selection biases must inevitably remain, which leads to the view that all such compliance analyses should be interpreted as descriptive supplements of an exploratory nature which cannot provide formal inferences of the same certitude as analyses by intention-to-treat.

### 3.2. Are pill counts any use?

There is considerable scepticism regarding the value of pill counts as an indicator of patient compliance. The number of pills that the investigator records as having been returned by the patient ideally should be subtracted from the number prescribed to give the number taken between planned visits, but there are many aspects of patient (and investigator) behaviour that cast doubt on the validity of this subtraction. Despite such individual ‘measurement error’ one can argue that in group terms the number of pills recorded as returned contains some (limited) insight into the patterns of non-compliance and hence is worth analysing. In this trial we first consider the relation between pill counts and weight change at 6 months, and then we combine this evidence with the plasma level data.

Patient visits were either 30 days or 60 days apart and corresponding totals of pills dispensed were either 96 or 192, respectively. For each visit the estimated proportion of capsules taken (PCT) since the previous visit = (number dispensed – number returned)/number dispensed. For a patient who is fully compliant, that is, 2 tablets every day, then PCT should equal 0.625, provided the time between visits was exactly 30 or 60 days and reporting was accurate. In principle, one could have adjusted PCT to reflect the exact number of days between visits but unfortunately dates of visits were not reliably kept on all patients. For each patient PCT was calculated at each visit and Figure 3 shows the distributions at 1 month and 6 months. There is a peak in the distributions at the full-compliance level of 0.625 but with substantial scatter. The other peak at PCT = 1 may be explained by the investigator and/or the patient declaring that no pills were returned. While this might conceivably be due to substantial over-compliance, it may well be a recording error perhaps reflecting investigator lack of interest in pill counts. At 6 months a small proportion returned more than were dispensed last time, perhaps indicating delayed confession of longer-term non-compliance. Are such data worth analysing? Well, let’s see.

With such an unusual distribution it seemed appropriate to group patients into six compliance categories as follows:

<table>
<thead>
<tr>
<th>PCT</th>
<th>Poor</th>
<th>Moderate</th>
<th>Good</th>
<th>Over</th>
<th>All</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.4</td>
<td>0.4–0.54</td>
<td>0.55–0.69</td>
<td>0.7–0.99</td>
<td>1</td>
<td>Unknown</td>
<td></td>
</tr>
</tbody>
</table>

To avoid selection bias in choice of grouping, this was done without reference to any data on treatment group, weight or plasma levels. Figure 4 shows for dF and placebo group the mean per cent weight change at 6 months by mean PCT for these 6 compliance categories. Numbers of patients and standard errors are also given, indicating that while ‘good compliers’ are the largest group they are only 30 per cent of the total. Poor compliers have the same mean weight loss in both groups, but then there is an increasing trend of treatment difference with increased compliance. The ‘over’ and ‘all’ groups are curious in that they have significantly less weight loss.
Figure 3. Distributions of the estimated proportion of capsules taken (PCT) at 1 month and 6 month visits, for both groups combined

than the poor compliers in the placebo group, but a marked difference between treatment groups. Clearly there are strong selection effects by both patients and investigators, which link to compliance with both randomized treatment and diet.

Perhaps the most useful estimate is the dF — placebo difference in mean per cent weight loss for good compliers, which is 3.3 per cent with 95 per cent CI from 1.57 to 5.05 per cent.

So far we have only used the 6 months PCT which relates to the previous 2 months but there also exist PCT estimates at 1, 2 and 4 months, that is, four compliance indicators per patient each
in 6 possible categories. Under the assumption that all four PCT indicators are equally important in their relation to weight loss, then each patient’s four contributions to the six categories are split evenly. For instance, if a patient was classified as good, poor, good and missing at 1, 2, 4 and 6 months, respectively, then they are counted as $\frac{1}{2}$ good, $\frac{1}{2}$ poor and $\frac{1}{2}$ missing. (One possible refinement would be to weight visits according to the time period covered, but we felt it more appropriate to use equal weights on the basis that each visit’s PCT may be an equally valuable marker of compliance.) The consequent results for the regression model incorporating these ‘compliance factor weightings’, treatment and their interaction, have been converted into the estimated mean per cent weight changes (and SEM) by compliance factor and treatment shown in Figure 5. Each mean in Figure 5 can be interpreted as the estimated weight change in a patient who is in that compliance category on all four occasions (assuming the relation of compliance to weight change at 6 months is the same for all pill count occasions).

The observed pattern is somewhat similar to Figure 4, though the more extensive use of pill count data has made the ‘over’ and ‘all’ compliance categories in the placebo group more similar to poor compliance. Also there is now an increasing trend in weight loss from poor to good compliance in the placebo group. The corresponding trend in the dF group is also much steeper so that the estimated mean per cent weight loss at 6 months in a good complier at all four time points is 13.94 per cent (SEM 0.92 per cent). The estimated dF versus placebo difference in per cent weight loss in such good compliers is also increased to 4.16 per cent (SEM 1.27 per cent). Note that standard errors are increased in Figure 5 compared to Figure 4, possibly due to the fact that patients contribute to more than one compliance category thus reducing discrimination (and
also reducing misclassification which explains why the estimated treatment effects and trends are greater).

Despite the well-known unreliability of pill count data this example does suggest that in a large clinical trial some insights into treatment effects and their relation to compliance can still be deciphered amongst the misclassifications that occur. However, one should recall that pill counts tend to be overoptimistic so that estimates of effect in so called ‘good compliers’ relate to a selective category some of whose compliance is probably not as good as that label implies. The dF versus placebo inferences for ‘good compliers’ are based on post hoc selection into that category, and its interpretation is more useful if this selection process is the same in both treatments, an assumption that is difficult to verify.

Once again, exploratory description rather than formal inference is the appropriate attitude to adopt.

3.3. An Overall Synthesis

The relation of F + nF plasma levels at 6 months to pill counts at 6 months is shown in Figure 6. The trend in mean F + nF across the poor, moderate, good and over categories is as expected, though individual variation is substantial. The ‘all’ and ‘missing’ categories have low mean plasma levels compatible with rather poor average compliance. This adds further support to the idea that pill count data make some sense, though the mean plasma level in the ‘good’ and ‘over’ compliers is somewhat lower than the 50 ng/l average that true perfect compliance should realize. One oddity is the fact that pill count overcompliers have higher mean plasma levels but poorer mean weight loss. One can speculate that some of them are particularly devious, taking excessive doses shortly before visits while being poorly compliant the rest of the time!

The next step is to simultaneously relate weight loss to treatment group, plasma level and pill counts. One approach to this is to build on the regression model for Figure 5 by adding one
additional explanatory variable, plasma F + nF level. This has then been used to estimate the mean weight loss in a dF patient with both F + nF = 50 ng/l and a ‘good’ pill count compliance record throughout, and compare this with the equivalent placebo good pill count complier.

Table II shows these results for 6 month weight change compared to the previous models described in Section 3. Table II also shows the corresponding results for 12 month weight loss, using plasma levels at 12 months and pill count data at all seven visits up to 12 months.

At 6 months, this latest model estimate for the treatment effect in a patient who complies well in both plasma level and pill count measures is 4.66 per cent weight loss (SEM 1.27 per cent) compared to a good placebo complier. This is greater than the estimates using each compliance measure separately, but also carries a much larger standard error than the estimate based on plasma level alone. The 12 month results suggest slightly increased weight loss for patients with a plasma level indicative of good compliance. However, the quality of pill count data may have deteriorated further, as indicated by their weak association with plasma level compared to 6 months so that additional use of pill count data did not enhance the treatment effect at 12 months in ‘good compliers’.

4. CONCLUDING REMARKS
The case study in Section 3 illustrates some of the relatively straightforward statistical approaches to incorporating quite complex compliance data into supplementary analyses of a clinical trial. Of course, any case study has its limits of generalizability to other studies. For instance, interpretation would become more difficult if there had been a significant placebo effect (Table I) and the analytical approach would need extending for trials comparing two active agents both with compliance measures such as drug plasma levels. Thus, we would encourage further illustrative examples to be explored as a practical complement to the substantial methodological
Table II. Estimated treatment effect from a variety of models

<table>
<thead>
<tr>
<th></th>
<th>Estimated per cent mean weight change (and SEM)</th>
<th>Treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DF-Placebo</td>
<td>DF-Placebo</td>
</tr>
<tr>
<td><strong>At 6 months (n = 633)</strong></td>
<td>[n] 633</td>
<td></td>
</tr>
<tr>
<td>All patients*</td>
<td>- 9·88 (0·35)</td>
<td>- 2·80 (0·50)</td>
</tr>
<tr>
<td>Plasma F + nF = 50 ng/ml†</td>
<td>- 11·04 (0·46)</td>
<td>- 3·96 (0·58)</td>
</tr>
<tr>
<td>Good pill count compliers‡</td>
<td>- 13·94 (0·92)</td>
<td>- 4·16 (1·27)</td>
</tr>
<tr>
<td>Good compliers and F + nF = 50 ng/ml§</td>
<td>- 14·44 (0·92)</td>
<td>- 4·66 (1·27)</td>
</tr>
<tr>
<td><strong>At 12 months (n = 564)</strong></td>
<td>[n] 564</td>
<td></td>
</tr>
<tr>
<td>All patients</td>
<td>- 9·36 (0·45)</td>
<td>- 2·92 (0·65)</td>
</tr>
<tr>
<td>Plasma F + nF = 50 ng/ml</td>
<td>- 11·70 (0·68)</td>
<td>- 5·26 (0·82)</td>
</tr>
<tr>
<td>Good pill count compliers</td>
<td>- 13·79 (1·43)</td>
<td>- 3·67 (2·02)</td>
</tr>
<tr>
<td>Good compliers and F + nF = 50 ng/ml§</td>
<td>- 14·71 (1·45)</td>
<td>- 4·59 (2·04)</td>
</tr>
</tbody>
</table>

* As from Figure 1
† As from Figure 2
‡ As from Figure 3
§ As from Figure 4

The statistician’s role of data analyst (issues 1 and 2) may consume much time, but it is important that we also assert our broader scientific role by ensuring that due restraint is exercised in issues 3 and 4.

ACKNOWLEDGEMENTS

We are very grateful to the dexfenfluramine trial investigators for permission to present recent findings on compliance. We especially appreciate the stimulus provided by Bruce Campbell, Bernard Guy-Grand and the late Paul Turner in encouraging us into this field.

REFERENCES