

TUTORIAL IN BIOSTATISTICS

Analysis of longitudinal laboratory data in the presence of common selection mechanisms: A view toward greater emphasis on pre-marketing pharmaceutical safety

Jonathan S. Schildcrout^{1,*}, Cathy A. Jenkins¹, Jack H. Ostroff², Daniel L. Gillen³,
Frank E. Harrell¹ and Donald C. Trost²

¹*Department of Biostatistics, Vanderbilt University, Nashville, TN, U.S.A.*

²*Pfizer Global Research and Development, Groton, CT, U.S.A.*

³*Department of Statistics, University of California, Irvine, CA, U.S.A.*

SUMMARY

Pharmaceutical safety has received substantial attention in the recent past; however, longitudinal clinical laboratory data routinely collected during clinical trials to derive safety profiles are often used ineffectively. For example, these data are frequently summarized by comparing proportions (between treatment arms) of participants who cross pre-specified threshold values at some time during follow-up. This research is intended, in part, to encourage more effective utilization of these data by avoiding unnecessary dichotomization of continuous data, acknowledging and making use of the longitudinal follow-up, and combining data from multiple clinical trials. However, appropriate analyses require careful consideration of a number of challenges (e.g. selection, comparability of study populations, etc.). We discuss estimation strategies based on estimating equations and maximum likelihood for analyses in the presence of three response history-dependent selection mechanisms: dropout, follow-up frequency, and treatment discontinuation. In addition, because clinical trials' participants usually represent non-random samples from target populations, we describe two sensitivity analysis approaches. All discussions are motivated by an analysis that aims to characterize the dynamic relationship between concentrations of a liver enzyme (alanine aminotransferase) and three distinct doses (no drug, low dose, and high dose) of an nk-1 antagonist across four Phase II clinical trials. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: pharmaceutical safety; clinical laboratory data; longitudinal data; dynamic treatment effects; pathodynamics

*Correspondence to: Jonathan S. Schildcrout, Department of Biostatistics, Vanderbilt University, Nashville, TN, 37232-2158, U.S.A.

†E-mail: jonathan.schildcrout@vanderbilt.edu

Contract/grant sponsor: Pfizer Global Research and Development, Groton, CT

1. INTRODUCTION

Safety risks believed to be caused by medications have received a substantial amount of attention in the recent past; however, Phase III clinical trials are nearly always designed and powered to establish efficacy. Safety endpoints tend to be considered secondarily and are frequently summarized with suboptimal procedures. For example, adverse events related to hepatotoxicity may be reported as the proportion of subjects who, at some time during follow-up, had liver enzyme activity measures, e.g. alanine aminotransferase (ALT), crossing a specified threshold value (e.g. three times the upper limit of normal). It is well known that dichotomization of continuous variables results in information loss, and often there is no reason to believe that the dichotomization point represents a unique, biologically meaningful threshold. In addition, by ignoring the longitudinal follow-up (e.g. repeated measurements on individuals), dynamic treatment effects are ignored. Considering the large number of adverse events studied simultaneously and concerns regarding type I error inflation, safety signal detection is extremely challenging under the current model of product development. In the recent past, analytical approaches that acknowledge correlation among binary adverse events have been proposed (e.g. see [1, 2]) and have improved our ability to detect these safety signals. However, serious adverse event rates are generally low, and even the improved methods cannot detect many important risk increases.

Post-marketing surveillance is proving to be vital for capturing medication-related adverse events. Using administrative data such as Medicaid, hospital, and insurance claims, many authors have retrospectively constructed follow-up cohort studies to capture excess adverse event risks associated with medications (e.g. [3, 4]). Novel methods for capturing safety signals using spontaneous reporting system data are also promising [5–7]; however, in spite of their utility, post-marketing surveillance data are less reliable than those from clinical trials due to under-reporting, biased reporting, incohesive administrative data systems, or simply because the information collected is observational.

We believe that the larger goal of safety analyses should be clearer understanding of the biological responses to pharmaceutical products. If we can begin to understand how subjects respond to specific products and known (and unknown) classes of products, we are then far more equipped to develop safer medications, or at least better characterize potential risks. The understanding of these mechanisms is not possible using standard approaches where continuous measures are dichotomized, time is ignored, and analyses are conducted without reference to other comparable studies or products. To begin to enhance our ability to characterize safety profiles during development or post-marketing, we must exploit all available information routinely collected in Phases I–III clinical trials, including clinical laboratory data. Summarization of these data is often undertaken to the extent required by a regulatory agency, such as the Food and Drug Administration; however, in many cases, they could be used more effectively.

The Pfizer Historical Research Database is a project that aims at combining clinical data from a large number of Phases I–III clinical trials and using these data for better understanding of pharmaceutical safety. For the purpose of this research, we seek to use clinical laboratory data related to liver function from four Phase II clinical trials of an NK-1 antagonist for which development was discontinued due to several issues, including hepatotoxicity. We measure liver enzyme activity with ALT, which is known to be a relatively specific indicator of acute liver cell damage. All four studies followed subjects for (at least) six weeks, with a schedule of 3–5 follow-up visits.

While the analysis of longitudinal clinical laboratory data can lead to more informed decisions regarding product safety, numerous challenges must be considered in order to make valid inferences in this setting. These include the potential incomparability of patients samples (if multiple studies were included), inter-laboratory variation in measurements (if multiple laboratories were used), and any number of selection mechanisms (e.g. unscheduled visits, outcome-driven or reflex testing, dropout, missing data, treatment discontinuation). In addition, the functional form of the treatment effects over time is unknown; hence, flexible models should be considered. Although heterogeneous study samples may lead to results driven by specific patient subgroups, they also provide opportunities since (valid) inferences based on their analysis can be generalized to broader populations than those based on homogeneous samples. Inter-laboratory variability is not considered in this paper, although some of the challenges associated with it have been discussed [8, 9]. Selection mechanisms can bias results if related to the response process. For example, current values or changes in ALT concentrations may lead to treatment withdrawal, added follow-up visits, or dropout. Much of this research focuses on strategies for the valid analyses of longitudinal clinical laboratory data in the presence of these selection mechanisms.

Another challenge to consider is that safety profiles derived from clinical trials participants may not generalize to the target population and, in fact, may represent the best-case scenario for safety. Study participants can be healthier than target populations (those who will eventually receive the medication), doses are controlled by study protocol, concomitant medications or lifestyle choices (e.g. alcohol use) may be an exclusion criterion, subjects are monitored closely, and the subjects are appropriate for the medication. Once marketed, the lack of regimentation may lead to previously unforeseen adverse reactions. If the goal of clinical trials is ultimately to generalize results to the population of people who may eventually receive the medication, then clinical trial participants often represent a biased sample. Since the extent of this bias is unknown, it is useful to put bounds on uncertainty through structured sensitivity analyses. We describe two sensitivity analyses for longitudinal clinical laboratory data that are intended to be guides for characterizing the 'realm of possibilities' for excess risk and for informed decision-making regarding medication safety.

This paper aims at describing viable approaches for effective utilization of longitudinal clinical laboratory data in characterizing product safety. We discuss selection mechanisms which, if ignored, could lead to inadequate characterizations. We describe two paradigms for treatment effect estimation, inverse probability weighted, semi-parametric estimation, and fully parametric, maximum likelihood estimation, and we propose sensitivity analyses aimed at generalizing study results to target populations. This is not a comprehensive review of all potential approaches. Rather, the goal is to highlight several strategies that could be used to improve (relative to strategies currently used) our ability to make informed decisions regarding product development and safety.

In Section 2, we provide a brief summary of the four studies that are combined for analysis. We discuss the measurement (from which inferences are to be made) and selection models in Section 3, and in Section 4, we describe the potential bias that selection mechanisms can induce, analytical approaches that can lead to valid inferences, and sensitivity analyses aimed at addressing bias due to the non-random selection of clinical trial participants. Finally a discussion and some concluding remarks are provided in Section 5.

2. THE STUDIES

We consider four Phase II randomized, placebo-controlled, clinical trials of a single agent (agent Q) from a total of 22 Phase I and II studies. The sample represents approximately 20 per cent of all patients studied on agent Q. We selected these studies due to the similarity in their dosing schedules (0, 50, or 100 mg daily dose for six weeks), scheduled follow-up period (six weeks), and although the studies were conducted on different disease populations (rheumatoid arthritis, psoriasis vulgaris, asthma, and ulcerative colitis), neither the diseases nor the concomitant medications were thought to be related to liver function. Chemotherapy patients, which composed half of all subjects studied on this agent, were not considered due to the effects of chemotherapy on liver function. Other studies were not considered because subjects were exposed to drugs for very short time periods. Throughout this paper, we refer to the 50 and 100 mg daily dose groups as low- and high-dose treatments, respectively. Table I provides a summary of the studies. Liver enzyme concentrations and other blood analytes were measured prior to and usually on the day of medication initiation (day 0). Subjects were scheduled for (at least) four follow-up visits (at approximately days 4, 14, 28, and 42) prior to amendments to the protocol schedules where the first visit was removed. During the course of study 3, a day 35 visit was added to scheduled follow-up which was retained for study 4. Protocols specified a reflex visit schedule wherein the presence of liver enzyme elevations could trigger an increase in the visit frequency. Patient-initiated follow-up was also allowed. In study 1 and for most of study 2, the (non-placebo) treatment was the high dose. During the course of study 2, the protocol was amended and the treatment dose was reduced to low dose. Throughout this paper we refer to 50 mg *bid* (*qd*) per day as high (low) dose treatment.

Table II describes the study populations broken down by the medication administered at study initiation. Approximately half of the subjects were assigned to placebo, while a quarter were assigned to each of the low- and high-dose treatments. Demographic and baseline characteristics were similar with respect to race, baseline ALT concentrations, and weight; however, the low-dose treatment arm was composed of relatively few females who tended to be younger than those in the other treatment arms. Although randomization was performed in all studies, this imbalance occurred because in study 1, where low-dose treatment was not examined, 71 per cent of participants were female and patients tended to be older. In the other studies, approximately 20–25 per cent of participants were female. These imbalances make it impossible to separate a treatment effect from a gender or age effect unless we control for those variables during the analysis. If interest was in treatment by disease interactions, for males and females separately, additional data would be required.

Figure 1 displays a subset of the longitudinal, natural log-transformed ALT response series for each of the assigned treatment groups. The lines connect ALT values for individuals at successive visits, and solid (dotted) lines indicate that the subject was (was not) on assigned treatment the day

Table I. Description of the four studies considered for analysis.

Study	Disease	Groups compared	Follow-up schedule days
1	Rheumatoid arthritis	High <i>versus</i> placebo	4, 14, 28, 42
2	Psoriasis vulgaris	High (low) <i>versus</i> placebo	4, 14, 28, 42
3	Asthma	Low <i>versus</i> placebo	4, 14, 28, (35), 42
4	Ulcerative colitis	Low <i>versus</i> placebo	4, 14, 28, 35, 42

Parentheses indicate a change in the protocol to the value inside parentheses during the course of the study.

Table II. Demographic and baseline characteristics of study participants.

	Placebo	Low dose	High dose	Total
Overall				
<i>N</i>	135	66	76	277
Female (per cent)	48	21	57	44
Ethnicity (per cent)				
White	81	85	84	83
Black	10	6	8	8
Other	9	9	8	9
Age (years)	50 (27, 65)	36 (24, 64)	53 (36, 66)	49 (27, 65)
Weight (kg)	82 (60, 106)	83 (64, 105)	80 (56, 98)	82 (59, 104)
Log-transformed, baseline ALT	2.8 (2.2, 3.5)	2.9 (2.4, 3.5)	2.6 (1.9, 3.4)	2.8 (2.2, 3.5)
Study 1				
<i>N</i>	62	0	59	121
Female (per cent)	74	—	68	71
Ethnicity (per cent)				
White	85	—	83	84
Black	8	—	8	8
Other	6	—	8	7
Age (years)	58 (46, 68)	—	56 (41, 68)	58 (43, 68)
Weight (kg)	75 (56, 102)	—	74 (54, 96)	74 (54, 98)
Log-transformed, baseline ALT	2.6 (2.0, 3.2)	—	2.6 (1.9, 3.1)	2.6 (1.9, 3.2)
Study 2				
<i>N</i>	22	6	17	45
Female (per cent)	23	0	18	18
Ethnicity (per cent)				
White	91	100	88	91
Black	5	0	6	4
Other	5	0	6	4
Age (years)	46 (29, 56)	60 (31, 66)	45 (33, 55)	46 (29, 61)
Weight (kg)	84 (66, 109)	94 (73, 109)	86 (78, 109)	85 (68, 109)
Log-transformed, baseline ALT	3.0 (2.5, 3.5)	2.8 (2.4, 3.7)	3.0 (2.5, 3.6)	3.0 (2.4, 3.6)
Study 3				
<i>N</i>	42	42	0	84
Female (per cent)	29	24	—	26
Ethnicity (per cent)				
White	67	76	—	71
Black	17	10	—	13
Other	17	14	—	15
Age (years)	32 (21, 52)	33 (23, 53)	—	32 (22, 53)
Weight (kg)	86 (70, 111)	86 (69, 107)	—	86 (70, 110)
Log-transformed, baseline ALT	2.9 (2.4, 3.6)	3.0 (2.5, 3.5)	—	2.9 (2.4, 3.6)
Study 4				
<i>N</i>	9	18	0	27
Female (per cent)	22	22	—	22
Ethnicity (per cent)				
White	100	100	—	100
Black	0	0	—	0
Other	0	0	—	0
Age (years)	50 (37, 59)	51 (29, 67)	—	50 (30, 65)
Weight (kg)	80 (54, 97)	76 (63, 95)	—	78 (58, 97)
Log-transformed, baseline ALT	2.9 (2.5, 3.5)	2.9 (2.4, 3.1)	—	2.9 (2.5, 3.3)

Categorical and continuous characteristics are summarized with percentages and 50th (10th, 90th) percentiles, respectively.

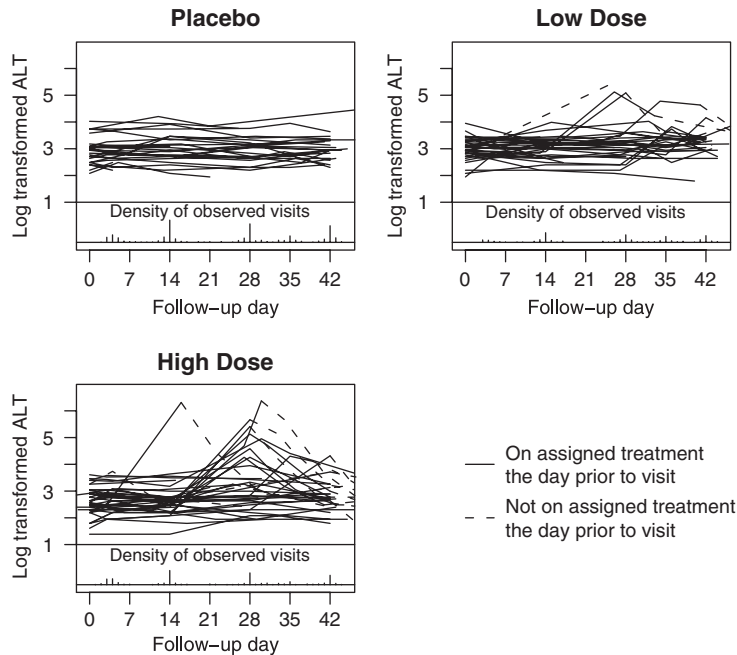


Figure 1. Log-transformed ALT concentrations over time by assigned treatment: All subjects for whom the maximum ALT value was at least twice the baseline value in addition to 25 randomly selected other subjects from each assigned treatment group are shown. Solid (dotted) lines indicate that the subject was (was not) on assigned treatment the day prior to the latter visit. At the bottom of the figure we show the density of the observed visits wherein it can be seen that most visits occurred at or around days 4, 14, 28, 35, and 42.

prior to the latter visit. All subjects for whom the maximum ALT value was at least twice the baseline value in addition to 25 randomly selected subjects from each assigned treatment group are shown. At approximately day 28, a number of subjects in the high-dose treatment arm and fewer subjects in the low-dose arm experienced marked increase in ALT concentrations. Participating physicians applied protocol rules or clinical judgment to discontinue treatment. At the bottom of the figure, we display the observed follow-up density. The majority of visits occurred around days 4, 14, 28, 35, and 42.

3. THE MODELS

In this section, we describe the measurement model from which inferences are to be made, and the postulated models for dropout, sampling rates, and treatment discontinuation.

3.1. The measurement model

Let $i \in \{1, \dots, N\}$ denote subject, $Y_i(t_{ij})$ be the response for subject i at the time t_{ij} of his/her j th visit ($j \in \{1, \dots, n_i\}$), $\mathbf{X}_i(t_{ij})$ be the $p+1$ vector of covariate values, and $\varepsilon_i(t_{ij})$ be a mean zero

error term. The generic model from which we base inference is

$$Y_i(t_{ij}) = \mathbf{X}_i^T(t_{ij})\boldsymbol{\beta}(t_{ij}) + \varepsilon_i(t_{ij}) \quad (1)$$

The parameter vector, $\boldsymbol{\beta}(t_{ij})$, is time dependent and it characterizes the strength of the relationship between the response and the covariate series. If the goal is simply estimation of $\boldsymbol{\beta}(t_{ij})$, then in the absence of selection bias, additional assumptions regarding the error process (e.g. correlation among repeated responses) other than finite variance are not required, and an empirical estimation approach such as generalized estimating equations (GEE) [10] could be used. If an additional goal is to increase flexibility in the functional form of $\boldsymbol{\beta}(t_{ij})$, then non-parametric or semiparametric models could be considered. Several authors have studied these models and have relied on counting process theory to derive asymptotic distributions of cumulative covariate effects $B(t_{ij}) = \sum_i \sum_j \beta(t_{ij})$ (e.g. see References [11–14]). A likelihood-based approach generally assumes a multivariate Gaussian distribution on the error process. The corresponding estimates are at least as efficient as estimating equations' approaches and are more robust to selection bias; however, proper specification of second- and higher-order moments is required for valid inference.

Since the majority of observations in the studies we consider occur at discrete time points and the interpretation of the non-parametric cumulative coefficients is challenging, we parameterize treatment effects using main effects and interactions with natural spline functions of time (with knots located at days 14, 28, 35, and 42). Maximum likelihood-based and empirical estimation will both be considered.

The specific mean model from which we wish to base inference is given by

$$\begin{aligned} E\{Y_i(t_{ij}) | \mathbf{x}_i(t_{ij})\} = & \beta_0 + \text{ns}(t_{ij})\boldsymbol{\beta}_t + \{\beta_{\text{low}} + \text{ns}(t_{ij})\boldsymbol{\beta}_{\text{low},t}\} \cdot I\{\text{Tr}t_i(t_{ij}) = \text{low}\} \\ & + \{\beta_{\text{high}} + \text{ns}(t_{ij})\boldsymbol{\beta}_{\text{high},t}\} \cdot I\{\text{Tr}t_i(t_{ij}) = \text{high}\} + \dots \end{aligned} \quad (2)$$

where $\text{ns}(t_{ij})$ denotes the natural spline design matrix, $I(\cdot)$ is 1 if \cdot is true and 0 if not, and $\text{Tr}t_i(t_{ij})$ denotes the treatment received by subject i on the day preceding day t_{ij} with possible values: 'high', 'low', and 'placebo or none'. The values $\{\beta_{\text{low}} + \beta_{\text{low},t}\text{ns}(t_{ij})\}$ and $\{\beta_{\text{high}} + \beta_{\text{high},t}\text{ns}(t_{ij})\}$ reflect dynamic treatment effects relative to placebo or no treatment and are the inferential targets. For analysis, we also include baseline covariates, sex, race, and a smooth function of age to control for potential confounding.

3.2. Selection bias due to dropout and irregular sampling rates

We now consider the potential for selection bias due to dropout or due to an alteration in the rate at which subjects are followed. Let (1) C_i be an indicator for inclusion in the clinical trial; (2) $R_i(t_{ij})$ be an indicator that the subject is in the study and at risk (not dropped out) at time t_{ij} ; and (3) $S_i(t_{ij})$ be an indicator that the subject was observed (sampled) at time t_{ij} .

To describe the relationship between the response \mathbf{Y}_i and covariates \mathbf{X}_i , one could use the joint distribution $f(\mathbf{y}_i, \mathbf{x}_i) \equiv f(\mathbf{y}_i | \mathbf{x}_i)f(\mathbf{x}_i)$. However, it is standard to leave the marginal distribution of \mathbf{X}_i unspecified and to capture the relationship between the response and covariates using $f(\mathbf{y}_i | \mathbf{x}_i)$. If subjects participating in the studies represent a random sample from the target population, then $f(\mathbf{y}_i | \mathbf{x}_i) = f(\mathbf{y}_i | \mathbf{x}_i, C_i = 1)$, and all inferences generalize to the population. That is, study results are externally valid. For ease of exposition, we drop

the condition, $C_i = 1$, which for now should be considered implicit. Internal validity is possible if the selection mechanisms, such as dropout and sampling rates, are uninformative. Assume that the observed response history, $\mathcal{H}_i^y(t_{ij}) \equiv \{y_i(t_{i1}), \dots, y_i(t_{ij-1})\}$, can be used to describe within subject correlation. According to the terminology of Little and Rubin [15] and following Lipsitz *et al.* [16], if $f(\mathbf{y}_i | \mathbf{x}_i) = \prod_{j=1}^{n_i} f\{y_i(t_{ij}) | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij})\}$ and $\Pr\{S_i(t_{ij}) = 1, R_i(t_{ij}) = 1 | y_i(t_{ij}), \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij})\} = \Pr\{S_i(t_{ij}) = 1, R_i(t_{ij}) = 1 | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij})\}$, then the dropout and sampling time mechanisms are ‘at random’ and are uninformative. This can be shown by noting

$$\begin{aligned} & f\{y_i(t_{ij}) | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij}), S_i(t_{ij}) = 1, R_i(t_{ij}) = 1\} \\ &= \frac{\Pr\{S_i(t_{ij}) = 1, R_i(t_{ij}) = 1 | y_i(t_{ij}), \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij})\} \cdot f\{y_i(t_{ij}) | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij})\}}{\Pr\{S_i(t_{ij}) = 1, R_i(t_{ij}) = 1 | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij})\}} \\ &= f\{y_i(t_{ij}) | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij})\} \end{aligned}$$

As discussed in Reference [16], likelihood-based estimation will be valid in this setting without explicit acknowledgment of the selection mechanisms as long as the covariance among responses has been properly acknowledged. An unadjusted analysis is not valid with empirical approaches, such as GEEs since $f\{y_i(t_{ij}) | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij}), S_i(t_{ij}) = 1, R_i(t_{ij}) = 1\} = f\{y_i(t_{ij}) | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij})\}$ does not imply $f\{y_i(t_{ij}) | \mathbf{x}_i(t_{ij}), S_i(t_{ij}) = 1, R_i(t_{ij}) = 1\} = f\{y_i(t_{ij}) | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_i(t_{ij})\}$; hence, in general, $E\{Y_i(t_{ij}) | \mathbf{x}_i(t_{ij}), S_i(t_{ij}) = 1, R_i(t_{ij}) = 1\} \neq E\{Y_i(t_{ij}) | \mathbf{x}_i(t_{ij})\}$. Many authors have discussed estimating equation-based approaches to acknowledge dropout (e.g. [17]) and irregular follow-up (e.g. [13, 14, 18–20]) in longitudinal data analyses. The approaches to acknowledge longitudinal dropout use Horvitz–Thompson inverse probability weights [21], while many of the approaches to irregular follow-up use inverse intensity modeling approaches during analysis. In all cases, estimation is based on solving

$$\sum_{i=1}^N \mathbf{X}_i^T \mathbf{W}_i \mathbf{P}_i^{-1} \{\mathbf{Y}_i - E(\mathbf{Y}_i | \mathbf{X}_i; \boldsymbol{\beta})\} = \mathbf{0} \quad (3)$$

for $\boldsymbol{\beta}$, where \mathbf{W}_i is the inverse of a working covariance matrix, and \mathbf{P}_i is a diagonal probability or intensity of being observed matrix. Estimates based on inverse probability of weighting can be inefficient if sampling probabilities vary drastically (e.g. if some observations are weighted very heavily).

3.3. Selection bias due to treatment discontinuation

In safety analyses of longitudinal clinical laboratory data, our focus is rightfully on the treatment received rather than on intended treatment (typical of efficacy studies). However, such analyses pose challenges since the treatment received at time t_{ij} may be predicted by response history, $\mathcal{H}_i^y(t_{ij})$, even after conditioning on covariate history, $\mathcal{H}_i^x(t_{ij})$. That is, $f\{x_i(t_{ij}) | \mathcal{H}_i^y(t_{ij}), \mathcal{H}_i^x(t_{ij})\} \neq f\{x_i(t_{ij}) | \mathcal{H}_i^x(t_{ij})\}$. In such cases, the treatment received is said to be statistically endogenous with respect to the response process, and response history both mediates and confounds the relationship between the response \mathbf{Y}_i and treatment received \mathbf{X}_i (see References [22, 23]). Similar to the approach taken to acknowledge the dropout and sampling rate mechanisms, observations

for an estimating equations approach can be weighted by the inverse probability of being on the treatment received at each time, t_{ij} . By reweighting observations in this manner, we effectively construct a pseudo-population in which $\mathcal{H}_i^y(t_{ij})$ mediates but does not confound the relationship of interest. That is, if we are only concerned with the bias resulting from treatment discontinuation, we can solve equation (3) with \mathbf{P}_i , the diagonal probability of treatment matrix. A very important caveat with inverse probability weighting in this setting is that the full covariate conditional mean $E\{Y_i(t_{ij}) | X_i(t_{i1}), \dots, X_i(t_{ini})\}$ is not equal to the cross-sectional mean $E\{Y_i(t_{ij}) | x_i(t_{ij})\}$; hence, to ensure that unbiased parameter estimates result, a diagonal working covariance matrix (e.g. independence estimating equations) must be used [23–25].

3.4. Alternative approaches

In the previous subsections we discussed selection mechanisms related to dropout, irregular sampling rates, and treatment discontinuation. Our primary consideration has been that the selection mechanisms are related to the response series through response history and covariate values. We assume that selection is conditionally independent of current responses. To obtain valid inferences with empirical approaches, models for the selection mechanism must be correct. We now propose two additional approaches that partially alleviate the need to model selection mechanisms explicitly when using empirical estimation strategies (ES). In the first approach, a function of response history ($\mathcal{H}_i^y(t_{ij})$) is captured in the measurement model. We refer to this measurement model as a conditional model, whereas equation (2) is a marginal model. A conditional model lowers the potential risk for selection bias by reducing or eliminating the dependence of the selection mechanisms and the current response on ‘leftover’ response history. This model is useful and is often of scientific interest; however, we recognize that it may not capture the causal relationship between the response series and treatment if a portion of the treatment effect is mediated by $\mathcal{H}_i^y(t_{ij})$. Another approach we consider is intended to address challenges directly related to treatment discontinuation. In this approach, all response values succeeding treatment discontinuation are dropped from the analyses. We expect physicians to discontinue treatment based on observed response values (e.g. response history), and this is supported by Figure 1. Thus, with this approach, we induce a dropout mechanism that is response history dependent (uninformative), and we eliminate concerns related to statistical endogeneity of treatment. In cases where a large number of subjects have been removed from treatment, this approach will likely be inefficient, and if there is reason to believe that treatment received would no longer be conditionally independent of current responses, this approach would not be valid. However, for the current analyses, we believe this to be a reasonable approach.

4. ANALYSIS

In this section, we discuss the analysis of the data described in Section 2 using methods described in Section 3. We first focus our attention on study 1. We describe assessment of the dependence of the selection mechanisms (dropout, sampling rates, and treatment received) on response history in Section 4.1, and we contrast several ES in Section 4.2. We then combine all four studies and describe the assessment of study comparability in Section 4.3 and analyze the combined studies’ data considering response history-induced selection bias in Section 4.4. Finally,

in Section 4.5, we discuss sensitivity analyses that are vital for external validity of study results, e.g. when $f(\mathbf{y}_i | \mathbf{x}_i) \neq f(\mathbf{y}_i | \mathbf{x}_i, C_i = 1)$.

4.1. Assessing the impact of response history on the dropout, sampling, and treatment received mechanisms

In this and the next subsection, we focus on study 1. Since the dropout, follow-up, and treatment discontinuation distributions are likely to be influenced by response history, but (we assume) are conditionally independent of current response values, selection mechanisms are uninformative. The selection model we consider includes all three of these components and can be decomposed as follows:

$$\begin{aligned} & \Pr\{R_i(t_{ij}) = 1, S_i(t_{ij}) = 1, X_{i,\text{rec}}(t_{ij}) = x_{i,\text{rec}}(t_{ij}) | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_{i,\text{sel}}(t_{ij})\} \\ &= \Pr\{R_i(t_{ij}) = 1 | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_{i,\text{sel}}(t_{ij})\} \end{aligned} \quad (4)$$

$$\times \Pr\{S_i(t_{ij}) = 1 | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_{i,\text{sel}}(t_{ij}), R_i(t_{ij}) = 1\} \quad (5)$$

$$\times \Pr\{X_{i,\text{rec}}(t_{ij}) = x_{i,\text{rec}}(t_{ij}) | \mathcal{H}_i^y(t_{ij}), \mathbf{x}_{i,\text{sel}}(t_{ij}), R_i(t_{ij}) = 1, S_i(t_{ij}) = 1\} \quad (6)$$

where $X_{i,\text{rec}}(t_{ij})$ denotes treatment received, and $\mathbf{X}_{i,\text{sel}}(t_{ij})$ corresponds to selection model covariates that may or may not contain measurement model covariates $\mathbf{X}_i(t_{ij})$. $\mathbf{x}_{i,\text{sel}}(t_{ij})$ denotes the observed values of those covariates. The complete selection model is composed of an at risk model (4), a sampling model conditioned on being at risk (5), and a treatment received model conditioned on being at risk and sampled (6). The design matrix, $\mathbf{X}_{i,\text{sel}}(t_{ij})$, need not be the same in each of the selection models, although they will be for this analysis. While other functional forms of $\mathcal{H}_i^y(t_{ij})$ could be considered, results among several of them were qualitatively similar. We report results for the first-order lagged response, e.g. $\mathcal{H}_i^y(t_{ij}) = y_i(t_{ij-1})$, in this report.

The linear predictor included in each of the above selection models is given by

$$\eta_i(t_{ij}) = \gamma_{y_i(t_{ij-1})} y_i(t_{ij-1}) + \mathbf{x}_{i,\text{sel}}^T(t_{ij}) \boldsymbol{\gamma} \quad (7)$$

where $\mathbf{x}_{i,\text{sel}}(t_{ij})$ is composed of covariates: assigned treatment, a smooth (natural spline) function of time ($ns(t_{ij})$), log-transformed baseline ALT, sex, race, and age. It is worth noting that randomized treatment and time did not appear to modify the effect of $y_i(t_{ij-1})$ in these selection models. Logistic regression analyses were used to obtain the estimated probabilities of being at risk $\{R_i(t_{ij})\}$, of being sampled $\{S_i(t_{ij})\}$ conditioned on being at risk, and of being on the treatment received $\{X_{i,\text{rec}}(t_{ij})\}$ conditioned on being at risk and sampled. Figure 2(a) describes the effect of $y_i(t_{ij-1})$ on each of these models, and we display the estimated values of $\gamma_{y_i(t_{ij-1})}$ from equation (7) associated with a 0.4 unit increase in $y_i(t_{ij-1})$ (e.g. a 50 per cent increase in the lagged value of the untransformed ALT value). Response history appears to be positively associated with being at risk (not dropped out), with being sampled (if still at risk), and with being on placebo treatment (if still at risk and sampled). Among these, the effect size of $y_i(t_{ij-1})$ was largest in the selection model involving treatment received. To be sure that we make valid inferences using empirical (e.g. GEE) procedures, these mechanisms should be acknowledged explicitly.

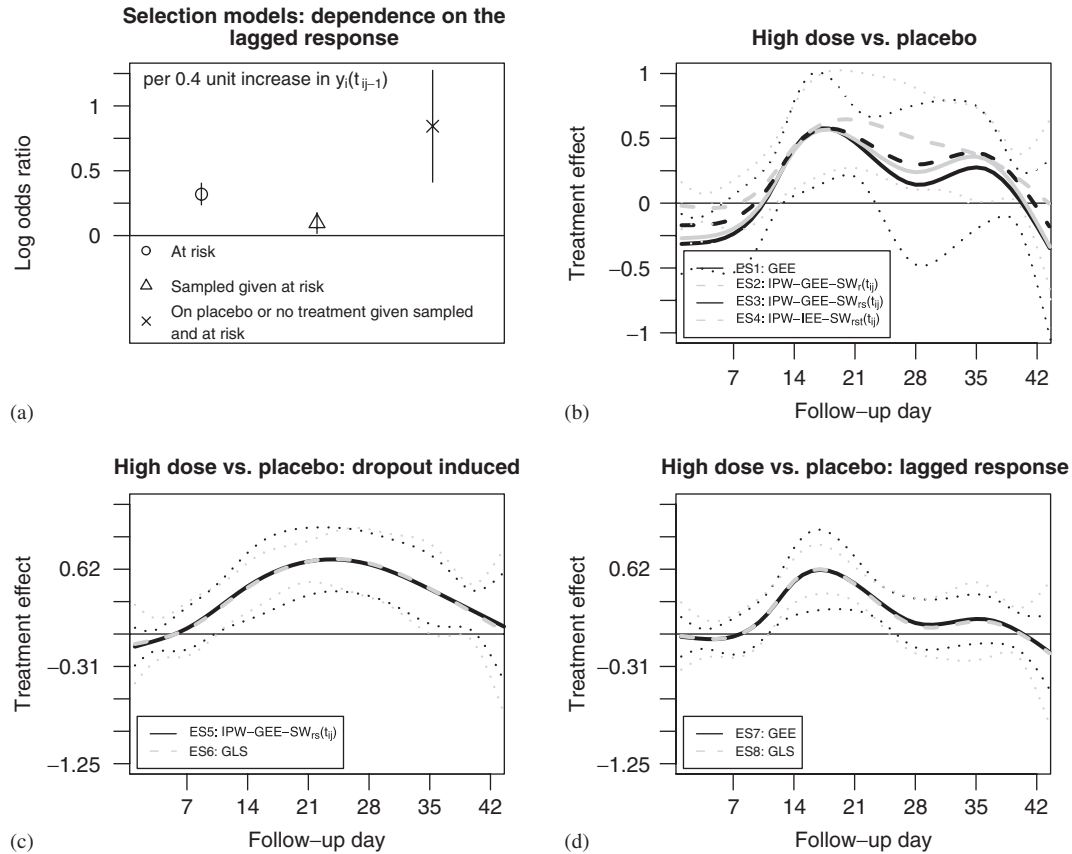


Figure 2. Analysis of study 1: (a) Estimates of $\gamma_{y_i(t_{ij-1})}$, based on the logistic regression model with linear predictor given in equation (7). The estimate is the log odds ratio associated with a 0.4 unit increase in $y_i(t_{ij-1})$, which corresponds to a 50 per cent increase in the untransformed ALT scale. In (b)–(d) the treatment effect of high-dose treatment (*versus* placebo or no treatment) among the eight estimation strategies (ES). In (b), we compare estimating equation approaches on the impact of ignoring the three selection mechanisms, but for ease of exposition we show confidence intervals only for ES1 and ES4 (dotted black and grey lines, respectively). In (c), we display the analyses in which dropout was induced once treatment was discontinued, and in (d), a response history covariate was included in the measurement model. In (b)–(d), treatment effects are on the log-transformed ALT scale.

4.2. Analysis of study 1

We now consider eight ES, derived from models described in Section 3, for the analysis of study 1. While we compare the estimates, we leave interpretation of measurement model results to Section 4.4 where all four studies are combined. The ES are outlined in Table III. ES1–ES6 correspond to the marginal model described in equation (2), whereas ES7–ES8 pertain to the conditional model where $\mathcal{H}_i^y(t_{ij})$ was included in the measurement model. ES1–ES5 and ES7 use GEE approaches, whereas ES6 and ES8 use maximum likelihood-based generalized least squares (GLS) with first-order autoregressive and exchangeable correlation structures, respectively.

Table III. Summary of the estimation strategies considered for analyses.

Strategy	Mean model	Data used for analysis	Estimation procedure
ES1	$\mathbf{X}_i^T(t_{ij})\boldsymbol{\beta}(t_{ij})$	All available	GEE
ES2	$\mathbf{X}_i^T(t_{ij})\boldsymbol{\beta}(t_{ij})$	All available	IPW-GEE-SW _r (t _{ij})
ES3	$\mathbf{X}_i^T(t_{ij})\boldsymbol{\beta}(t_{ij})$	All available	IPW-GEE-SW _{rs} (t _{ij})
ES4	$\mathbf{X}_i^T(t_{ij})\boldsymbol{\beta}(t_{ij})$	All available	IPW-IEE-SW _{rst} (t _{ij})
ES5	$\mathbf{X}_i^T(t_{ij})\boldsymbol{\beta}(t_{ij})$	Preceding treatment discontinuation only	IPW-GEE-SW _{rs} (t _{ij})
ES6	$\mathbf{X}_i^T(t_{ij})\boldsymbol{\beta}(t_{ij})$	preceding treatment discontinuation only	GLS
ES7	$\mathbf{X}_i^T(t_{ij})\boldsymbol{\beta}^*(t_{ij}) + \beta_{\text{lag}}Y_i(t_{ij-1})$	All available	GEE
ES8	$\mathbf{X}_i^T(t_{ij})\boldsymbol{\beta}^*(t_{ij}) + \beta_{\text{lag}}Y_i(t_{ij-1})$	All available	GLS

GEE is generalized estimating equation and IEE assumes a diagonal working covariance matrix. IPW-GEE is inverse probability weighted GEE, and GLS is generalized least squares. The correlation structures for all GEE estimates other than ES4 were autoregressive with decay rate a linear function of time separations. The values SW_r(t_{ij}), SW_{rs}(t_{ij}), and SW_{rst}(t_{ij}) correspond to the stabilized weights for the *j*th observation of subject *i*, occurring at time t_{ij} based on the at-risk, the at-risk and sampled, and the at-risk, sampled and on observed treatment selection models, respectively. In ES1–ES6, a marginal mean model is considered, X_i(t_{ij})β(t_{ij}), and in ES7–ES8 a conditional model, X_i(t_{ij})β*(t_{ij}) + β_{lag}Y_i(t_{ij-1}), is used. In ES5–ES6, all observations succeeding treatment discontinuation are dropped from analyses.

ES1 ignores all selection mechanisms, and GEE with a transition-type working covariance matrix is used. That is, working correlation is structured so as to be a linear function of time separations among visits. This working correlation structure is used for all GEE estimates except ES4 which applies a diagonal working correlation matrix (e.g. independence estimating equations or IEE) due to the reasons discussed at the end of Section 3.3. In ES2–ES4, the selection mechanisms are acknowledged using inverse probability weighted GEE (IPW-GEE) [17] with stabilized weights [22, 26]. In ES2, ES3, and ES4, the diagonal elements of P⁻¹ in equation (3) are given by

$$SW_r(t_{ij}) = \frac{\Pr\{R_i(t_{ij}) = 1 \mid \mathbf{x}_i(t_{ij})\}}{\Pr\{R_i(t_{ij}) = 1 \mid \mathcal{H}_i^y(t_{ij}), \mathbf{x}_{i,\text{sel}}(t_{ij})\}}$$

$$SW_{rs}(t_{ij}) = SW_r(t_{ij}) \cdot \frac{\Pr\{S_i(t_{ij}) = 1 \mid \mathbf{x}_i(t_{ij}), R_i(t_{ij}) = 1\}}{\Pr\{S_i(t_{ij}) = 1 \mid \mathcal{H}_i^y(t_{ij}), \mathbf{x}_{i,\text{sel}}(t_{ij}), R_i(t_{ij}) = 1\}}$$

$$SW_{rst}(t_{ij}) = SW_{rs}(t_{ij}) \cdot \frac{\Pr\{X_{i,\text{rec}}(t_{ij}) = x_{i,\text{rec}}(t_{ij}) \mid \mathbf{x}_i(t_{ij}), R_i(t_{ij}) = 1, S_i(t_{ij}) = 1\}}{\Pr\{X_{i,\text{rec}}(t_{ij}) = x_{i,\text{rec}}(t_{ij}) \mid \mathcal{H}_i^y(t_{ij}), \mathbf{x}_{i,\text{sel}}(t_{ij}), R_i(t_{ij}) = 1, S_i(t_{ij}) = 1\}}$$

respectively, where x_i(t_{ij}) represents covariates in the measurement model. The numerator ‘stabilizes’ the inverse probability weight and improves the efficiency of parameter estimates to the extent that covariates in the measurement model predict the selection mechanisms. ES5 and ES6 are estimated in the setting where all follow-up visits succeeding treatment discontinuation are dropped. As described in Section 3.4, the induced dropout mechanism is believed to be uninformative, and there is no bias due to treatment discontinuation. ES5 acknowledges selection

(dropout and sampling) explicitly using stabilized inverse probability weights, and ES6 acknowledges selection implicitly since it involves maximum likelihood. In ES7 we used GEE without further acknowledgement of selection, and in ES8 we used GLS with an exchangeable correlation structure. For all analyses the R programming language [27] was used, in particular, the nlme [28], geepack [29], and Hmisc [30] packages.

Figure 2(b)–(d) displays the treatment effects based on various ES. Figure 2(b) depicts the impact that ignorance of the selection models can have on estimates of the treatment effects. For example, by ignoring the selection mechanisms using empirical approaches, parameter estimates tended to be less positive and were negative at early follow-up times. Figure 2(c) pertains to models in which dropout following treatment discontinuation was induced. Results for ES5 and ES6 were similar. Figure 2(d) depicts the conditional mean model estimates (e.g. lagged responses included in the measurement model). By controlling for lagged responses, ES7, which ignores selection and is potentially vulnerable to biases, produces very similar estimates to ES8, which should be robust to these mechanisms. That is, explicit modeling selection using empirical approaches may not be necessary if the response history is captured in the measurement model. In both cases we assumed that controlling for response history alleviates the need to consider the bias induced by treatment discontinuation.

4.3. Testing the appropriateness of combining information from multiple trials

We now turn attention to the analysis of all four studies combined. One of the primary concerns with combining multiple studies is that the patient populations and study protocols may not be comparable. The studies examined here were chosen in part because schedules for follow-up were similar. However, patients from the four studies had different diseases, and while it would seem that they may be comparable with respect to liver function, further examination is appropriate. To explore the extent to which the populations from the four studies/disease classes were comparable, we stratified on the treatment received immediately following study initiation (e.g. placebo, low dose, and high dose) and examined the impact of disease (i.e. study) in longitudinal data models. Thus, we sought to examine the extent to which disease class was associated with ALT concentrations within treatment arms. Strong effects would indicate that the study samples may not be comparable. GLS was used to regress the log-transformed ALT values on subject-specific covariates: baseline log-transformed ALT, sex, age, time (natural splines with knots at days 14, 28, and 42), disease class, and the disease by time interaction. Since likelihood-based approaches were used, we ignored the sampling and dropout mechanisms. This approach is intended to provide only a general guideline regarding sample comparability. Within each strata, we compared Akaike's Information Criterion (AIC) in models that ignored disease class, that incorporated disease class indicators, and that included disease class by time interactions. According to the AIC, in all disease strata, models that included neither disease class nor disease class by time interactions were optimal. For the placebo arm models, the AICs for models without disease class indicators, with disease class indicators, and with disease by time interactions were 30.6, 38.9, and 71.2, respectively. For low-dose treatment they were 268.0, 276.0, and 282.2, respectively, and for high-dose treatment they were 561.5, 563.9, and 567.0. These results suggest that there is little to no evidence in support of within-treatment disease class effects or disease class by time interaction effects. In the combined study analyses we make no further adjustment for disease.

4.4. Combined study analysis

We now combine the data from the four studies to assess the effect of agent Q on liver function as measured by ALT. We followed the approaches outlined in Section 4.2, but limited the discussion to the most viable ES (ES4–ES8). Figure 3 displays treatment effects for marginal mean models (Figure 3(a) and (b)) and conditional mean models in which we adjust for lagged responses (Figure 3(c) and (d)). Figure 3(a) and (c) corresponds to the low-dose treatment effects *versus* placebo or off treatment, and Figure 3(b) and (d) relates to high-dose treatment effects. Note that the treatment effects displayed in Figure 3(a) and (b) correspond to common model estimates (as do Figure 3(c) and (d)). That is, ES6 estimates in Figure 3(a) and (b) are the estimated low- and high-dose treatment effects from a single model.

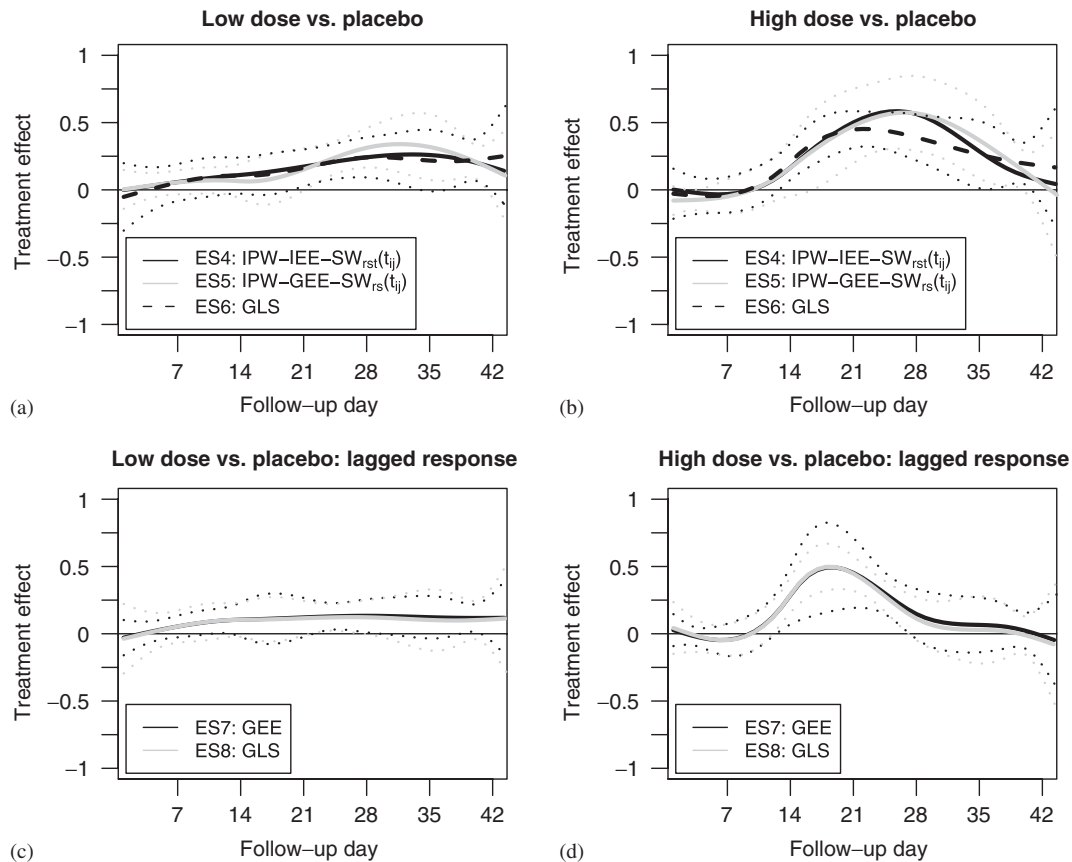


Figure 3. Analyses of all four studies: (a) and (b) Low-dose and high-dose treatment effect estimates for marginal models, respectively. (c) and (d) The estimates for conditional models. For a given estimation strategy, the low-dose and high-dose treatment effect estimates were obtained from a single model (e.g. treatment effects using ES6 in (a) and (b) are based on a single model fit). For ease of exposition, in (a) and (b), we display pointwise confidence intervals only for ES5 and ES6. All treatment effects are on the log-transformed ALT scale.

On the basis of these analyses, we find that there is a clear dose–response relationship, the functional form of the marginal and conditional model treatment effects are substantially different, and while there is a high-dose treatment effect during the course of follow-up, it does not appear to persist to the end of the studies. The adaptive response is interesting (or perhaps puzzling) and may warrant further investigation. Such an effect would be hidden in (standard) analyses that compare the proportion of subjects crossing threshold values at some time during follow-up in each treatment arm.

The ES considered here all yielded qualitatively similar conclusions regarding the treatment effect of agent Q. The functional form of the high-dose treatment effect with ES6 was a bit different from ES4 and ES5, but for conditional models ES7 and ES8 estimates were very close. For the most part, and assuming GLS correlation structures were specified appropriately, likelihood-based approaches tended to be more efficient than empirical approaches (a well-known result), although this was not the case at earlier time points.

4.5. Sensitivity analyses

To this point we have considered analyses that would permit valid inferences if trials represent a random sample from the target population, e.g. if $f(\mathbf{y}_i | \mathbf{x}_i) = f(\mathbf{y}_i | \mathbf{x}_i, C_i = 1)$. In fact, for a variety of reasons, this is not likely to be the case. In cases of non-random selection, postulating a number of reasonable selection mechanisms and examining the sensitivity of study results to these mechanisms are appropriate. All conclusions should then be made in light of the sensitivity analysis results.

For the purpose of safety, our interest focuses on a selection mechanism that would lead to an over- or under-representation of subjects who are at risk for experiencing adverse outcomes. Using the longitudinal response data, there are a number of ways to categorize subjects into risk categories using subject-specific response profiles [31, 32]. However, for our purposes, categories of risk can be captured with simple summaries of subject-specific response profiles. For example, we expect that subjects exhibiting substantial response variation or whose overall response values tend to be high may be particularly susceptible to severe reactions. On the basis of the average and standard deviation of individual response profiles, we grouped subjects into four categories of risk with a k-means clustering algorithm. The results from the cluster analysis are displayed in Figure 4(a), and Figure 4(b) displays representative response profiles of individuals from each of the clusters. For our purposes cluster 3, which pertains to subjects with the highest means and standard deviations, is of primary interest.

We propose sensitivity analyses based on the results of the clustering routine. The analyses postulate that, relative to the target population, subjects in cluster 3 (high risk) may be over- or under-represented in our sample. We are perhaps most interested in the case where they are under-represented. In the first analysis, we consider a tilted, non-parametric bootstrap approach (see, for example, [33]) wherein resampling weights for subjects in cluster 3 are altered from the usual value of one (the resampling weight for all other subjects). We consider resampling weights, $\pi_s \in \{0.5, 2.0, 4.0\}$, for all subjects in cluster 3. $\pi_s > 1$ implies resampling of subjects in cluster 3 with probability greater than subjects in clusters 1, 2, and 4. Summaries across 200 bootstrap samples were used to estimate the associated treatment effects. The second approach is similar in spirit to the first; however, rather than resampling, a single model is estimated, and the weights given to subjects in cluster 3 are altered, $\pi_w \in \{0.5, 2.0, 4.0\}$. Thus, when $\pi_w = 4$, cluster 3 subjects' observations are given four times the weight that they would be given in a standard analysis.

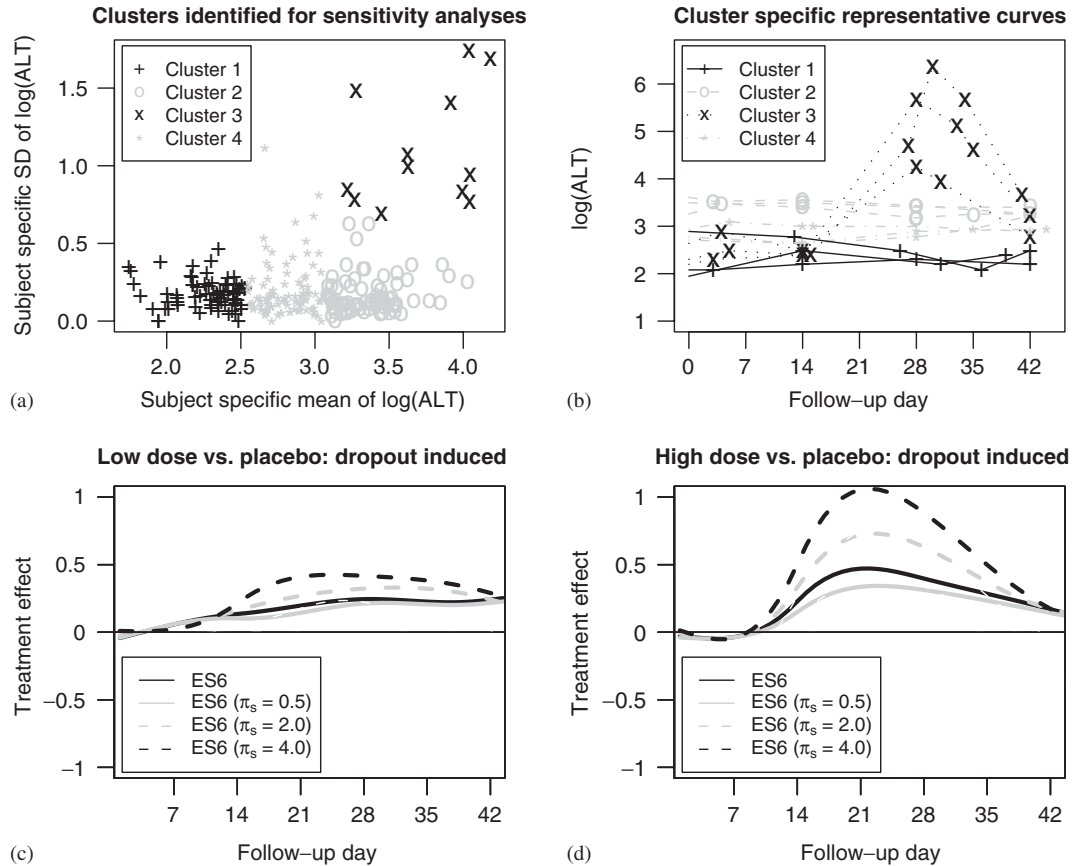


Figure 4. Sensitivity analyses: (a) The clusters identified by the k -means clustering algorithm, and (b) shows three corresponding representative curves. (c) and (d) Results from the sensitivity analyses in which the relative influence of cluster 3 was altered. Tilted non-parametric bootstrapping results are shown with sampling weights equal to 0.5, 2.0, and 4.0 for cluster 3. All other subjects had sampling weights equal to 1. For each of the analyses, averages among 200 bootstrap samples were used to summarize treatment effects. In (c) and (d), treatment effects are on the log-transformed ALT scale.

The parameter estimates obtained from these approaches were effectively indistinguishable; hence while we report the results for the first approach only, they apply to both.

For the purposes of exposition we will only consider ES6, and the results of these analyses are displayed in Figure 4(c) and (d), respectively. It is clear that varying the influence of those ‘high-risk’ subjects, by either changing weights or sampling probabilities, altered the estimates of treatment effects. With greater weight given to cluster 3 subjects, much larger treatment effects were obtained. However, the qualitative conclusions from the analyses remained unchanged. There appears to be a clear dose–response relationship to treatment, but according to our analyses, the apparent adaptive response persists and treatment effects disappear by the end of the study.

5. DISCUSSION

The goal of this research was to examine a number of strategies that could be used for the analysis of the longitudinal clinical laboratory data routinely collected during clinical trials. Frequently, these data are not utilized effectively in that continuous data are often dichotomized, longitudinal follow-up is ignored, and information from multiple studies are rarely combined for the purpose of characterizing product safety. Much of the discussion focused on selection bias. Toward internal study validity, we considered three selection mechanisms (dropout, visit frequency, and treatment discontinuation) which if related to the response process can lead to biased treatment effect estimates. We assumed that selection was related to the response process only through its history (e.g. past response values), which we believe to be reasonable in many clinical trial settings. Methods for handling these mechanisms include inverse probability weighted approaches (for all mechanisms) and direct likelihood-based approaches (for dropout and visit frequency mechanisms). Although not considered here, there are a large number of other ES (e.g. propensity score based estimators, multiple imputation-based estimators, doubly robust estimators, etc.) that could certainly be applied. To avoid at least some of the bias associated with selection, we also suggested alternative approaches wherein we capture response history in the measurement model or we induce dropout when treatment has been discontinued due to response elevations. By capturing response history in the measurement model, we alter the estimation target; however, if the new target is of interest, the residual dependence of selection and the current response on response history are likely to be diminished substantially, reducing the potential for severe biases. The induction of dropout following response history-dependent treatment discontinuation alleviates potential bias due to statistical endogeneity, which permits direct application of likelihood-based methods.

Since the ultimate goal of safety analyses is to generalize results to the patients who will eventually receive treatment, non-random selection is common. This is caused by the relative healthiness of clinical trial populations, exclusion criteria for dangerous concomitant medications and lifestyle choices (e.g. alcohol use), and the close monitoring of patients, among other reasons. Since the extent of the induced bias is untestable, we proposed sensitivity analyses in which the selection mechanism is based on a surmised measure of risk (e.g. instability in longitudinal profiles) for a serious adverse event such as severe liver injury. With this approach we can test the extent to which the informative selection mechanism could impact study conclusions.

In the analysis of agent Q, we observed a dynamic dose–response relationship between ALT concentrations and treatment. The treatment effect was largest at approximately week four, but had abated by the end of week six. Product development was terminated for a number of reasons; however, if the medication was (1) highly effective and/or (2) could be used on a short-term basis, and if the decision to terminate its development been made based exclusively on standard analyses of ALT concentrations, then we would suggest that further investigation is warranted. The dynamic treatment effect we observed is not captured with standard approaches, and we believe that the analyses we propose are vastly superior toward characterizing the safety profile of this and other products.

While the models we have considered capture the impact of treatment on expected log-transformed ALT concentrations, they do not tell the entire story. Most subjects on low- and high-dose treatment did not experience elevations at all during study follow-up, and subject-specific information that can be used to identify such subjects is unavailable. Models that are able to incorporate (latent) mixtures of responders and non-responders [34] or analyses that could be used

to identify responders (perhaps by incorporating biomarker or genotypic information) could be very useful in this setting.

The data used for the analyses we discussed are routinely collected in clinical trials. In addition to being utilized more effectively toward the characterization of pharmaceutical safety in individual products being brought to market, they can also be used toward a number of other aims, e.g. retrospective confirmation of post-marketing signals, identification of subjects for future studies regarding individualized sensitivity to therapy, etc. The key is that the data have already been collected, and we believe that far more information can be extracted from them than is by standard practices.

ACKNOWLEDGEMENTS

This research was funded by a contract from Pfizer Global Research and Development, Groton, CT, as part of a collaborative project on statistical and graphical tools for the analysis of pharmaceutical safety.

REFERENCES

1. Berry SM, Berry, DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics* 2004; **60**:418–426.
2. Agresti A, Klingenberg B. Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Applied Statistics, Journal of the Royal Statistical Society, Series C* 2005; **54**:691–706.
3. Ray WA, Stein CM, Daugherty JR, Hall K, Arbogast, PG, Griffin MR. COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. *Lancet* 2002; **360**:1071–1073.
4. Cooper WO, Hernandez-Diaz S, Arbogast PG, Dudley JA, Dyer S, Gideon PS, Hall K, Ray WA. Major congenital malformations after first-trimester exposure to ACE inhibitors. *New England Journal of Medicine* 2006; **354**:2443–2451.
5. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician* 1999; **53**:177–190.
6. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* 2001; **10**:483–486.
7. Rothman KJ, Lanes S, Sacks, ST. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiol Drug Saf* 2004; **13**:519–523.
8. Harris EK, Boyd JC. *Statistical Bases of Reference Values in Laboratory Medicine*. Marcel Dekker Inc.: New York, 1995.
9. Trost DC. Multivariate probability-based detection of drug-induced hepatic signals. *Toxicology Review* 2006; **25**:37–54.
10. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
11. Scheike TH, Zhang MJ. Cumulative regression function tests for regression models for longitudinal data. *The Annals of Statistics* 1998; **26**:1328–1355.
12. Martinussen T, Scheike TH. A semiparametric additive regression model for longitudinal data. *Biometrika* 1999; **86**:691–702.
13. Lin DY, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 2001; **96**:103–126.
14. Sun Y, Wu H. Semiparametric time-varying coefficients regression model for longitudinal data. *Scandinavian Journal of Statistics* 2005; **32**:21–47.
15. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 2002.
16. Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Gelber R, Lipshultz S. Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* 2002; **58**:621–630.
17. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**:846–866.
18. Martinussen T, Scheike TH. A nonparametric dynamic additive regression model for longitudinal data. *The Annals of Statistics* 2000; **28**:1000–1025.

19. Martinussen T, Scheike TH. Sampling adjusted analysis of dynamic additive regression models for longitudinal data. *Scandinavian Journal of Statistics* 2001; **28**:303–323.
20. Lin H, Scharfstein DO, Rosenheck RA. Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Statistical Methodology, Journal of the Royal Statistical Society, Series B* 2004; **66**:791–813.
21. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**:663–685.
22. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* 2001; **96**:440–448.
23. Diggle P, Heagerty PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press: Oxford, 2002.
24. Schildcrout JS, Heagerty PJ. Regression analysis of longitudinal binary response data with time-dependent environmental covariates: bias and efficiency. *Biostatistics (Oxford)* 2005; **6**:633–652.
25. Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics, Part B—Simulation and Computation* [Split from: @J(CommStat)] 1994; **23**:939–951.
26. Robins JM. Marginal structural models. In *ASA Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association, Alexandria, VA, U.S.A., 1997; 1–10.
27. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: Austria, 2006.
28. Pinheiro J, Bates D, DebRoy S, Sarkar D. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-77, 2006.
29. Yan J, Fine J. Estimating equations for association structures. *Statistics in Medicine* 2004; **23**:859–874.
30. Harrell FEH. *Hmisc: Harrell Miscellaneous*. R package version 3.1-2, 2006.
31. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 2002; **97**:611–631.
32. Otey ME, Parthasarathy S, Trost DC. Dissimilarity measures for detecting hepatotoxicity in clinical trial data. *SIAM International Conference on Data Mining*, Bethesda, MD, U.S.A., 2006.
33. Hesterberg TC. Bootstrap tilting confidence intervals and hypothesis tests. *Computing Science and Statistics*, Interface Foundation of North America: Fairfax Station, VA, U.S.A., 1999; **31**:389–393.
34. Xu W, Hedeker D. A random-effects mixture model for classifying treatment response in longitudinal clinical trials. *Journal of Biopharmaceutical Statistics* 2001; **11**:253–273.