Title
Pediatric Pain, Predictive Inference and Sensitivity Analysis

Author
Robert E. Weiss

Publication Date
2011-10-24

Pediatric Pain, Predictive Inference and Sensitivity Analysis

ROBERT WEISS*
Department of Biostatistics
UCLA School of Public Health
Los Angeles CA 90024-1772 U.S.A.
rob@oahu.ph.ucla.edu

August 4, 1994

Abstract

The understanding, prevention and treatment of pain is of great importance to medical science. Children were asked to immerse their hands in cold water until they were unable to tolerate the pain of the cold. The length of time that they kept their hands immersed is a measure of pain tolerance. Two factors were studied; one factor is a child's Style of Coping (CS) with the pain (ATTENDERS pay attention to the pain, DISTRACTERS think of other things) and was assessed at a baseline trial. The other factor is Treatment (T), one of three counseling interventions (a NULL intervention, counseling to ATTEND, or counseling to DISTRACT) and was randomly applied prior to the response. The covariate is a baseline measure of pain tolerance prior to the intervention. Distracters taught to distract tolerated the pain much better than any other group. No strategy improved attenders pain tolerance.

This paper analyzes this data from a predictive Bayesian viewpoint. The assumption of constant variance is not met on the original scale and some of the data is censored, furthermore, the censoring model is unknown. Simultaneous transformation of the response and baseline is used to search for a scale where the variance is constant. Predictive inference is used to provide interpretable inferences. A sensitivity analysis is used to determine whether the treatment of the censored data matter. Without the sensitivity analysis, we are left wondering whether the conclusions rest on a true underlying treatment effect or on cases of questionable quality.

*This research was partially supported by NIH grants #MH 37188 and #GM50011. Thanks to Lonnie Zeltzer and Deb Fanurik for discussion, collaboration and data; to Roderick Little and Dick Berk and Will Thomas for comments. Part of this work was done while on sabbatical at the School of Statistics, University of Minnesota

In spite of a devil's-advocate effort to produce an alternative model which leads to contradictory conclusions, no such model was found.

Key Words: Bayesian Inference, Box-Cox Transformation, Censoring, Conditional Predictive Ordinate, Diagnostics, Influential Observations, Outliers, Pain Tolerance.

1 Introduction

Pain is a symptom of many diseases and also a negative outcome of medical procedures and treatments. The understanding, prevention and eventual treatment of pain is of great importance to medical practice: children with poor pain tolerance may be expected to have further poor outcomes as a result of pain. Avoidance of further necessary medical care is another possible resultant. The long term goal of the research of which this analysis is a part is to develop methods for treating pain and to develop methods for identifying and treating children with poor pain tolerance.

Different people react differently to pain and have different reactions to otherwise similar procedures; children may react differently than adults. What characteristics of people lead to these different feelings is an important part of understanding pain. To develop appropriate treatments, we first need to understand what characteristics of people, diseases, and medical treatments lead to feeling pain, and to pain tolerance, intolerance and avoidance. The experiment analyzed here studies how children dealt with pain in a laboratory setting, and whether a matched or mismatched coping strategy intervention was helpful in coping with the pain.

Children were recruited from an elementary school located on the campus of UCLA. They take part in four cold pressor trials split up into two sessions two weeks apart; there are two trials per session. The cold pressor procedure has children immersing their hands in quite cold water until they are unable to tolerate the pain of the cold. The length of time in seconds that they kept their hands immersed is a measure of pain tolerance. Unlike many procedures for studying pain, the cold pressor is not harmful, is perceived to be fun by the children, and because it is a laboratory procedure, is amenable to designed

experiments. In general, it can be unethical to study pain except as part of an observational study of the pain caused by medical procedures. Since medical procedures vary greatly from trial to trial and may or may not be repeated, it can be quite difficult to draw firm conclusions from such a study. In contrast, the experiment analyzed here is a 2×3 analysis of covariance and is very nearly balanced.

Children vary substantially in their ability to tolerate pain, and a baseline measurement of pain tolerance is used to adjust for individual differences. The covariate is the second cold pressor trial during the first session, the first trial is a 'warm-up'. The response is the second trial during the second session, the intervention occurs between the two trials of the second session. There are two factors in this study; the first, coping style (CS), is an observed characteristic of the children. Coping Style is a summary of the child's approach to coping with the pain, and was assessed by asking the children what they thought about during the baseline cold pressor trials. Raters who did not know of the outcomes of the trials rated each child's response as representative of an attender (A) or distracter (D). Attenders pay attention to the pain, their arm, the procedure and the feelings of their arm in the cold water, while distracters tend to think of other things: the wall, homework, the beach. Of the 61 children with complete data, one half (30) were classified as attenders, and 31 as distracters.

One of three counseling interventions was randomly applied prior to the response trial. The three interventions are a null (N) counseling intervention, counseling to attend (A), or counseling to distract (D). The null intervention functions roughly as a control group. The attend treatment is a matched treatment for attenders and mis-matched for distracters; the reverse holds for the distract treatment. The initial hypothesis was that matched treatments would do better than mismatched treatments. A possible implication, if this hypothesis were to hold up under further study would be to assess children's natural coping strategies, and during painful medical procedures (drawing blood for example), coach them to use their natural coping strategy. The actual outcome was that distracters taught to distract tolerated the pain much better than any other group The other treatments and groups were roughly equivalent. Fanurik, Zeltzer, Roberts and Blount (1993) has further description of the trial.

This data set has several features which make the analysis non-routine in the area of pediatric pain. Children with long tolerances can reasonably be expected to be more variable trial to trial than those with short tolerances; this is born out by graphical analysis. Figure 1 gives boxplots of log(response) − log(baseline). The six groups are identified below the plot by a two character code AA, AD, AN, DA, DD, DN; the first letter stands for the coping style A or D, and the second letter A, D or N stands for the treatment the child was randomized to. Distracters taught to distract are the only group whose lower quartile is greater than zero. The distracters given the null treatment have the upper quartile of the difference of the log tolerances below zero, suggesting that this treatment actually inhibited their ability to stand the pain. A similar plot (not shown) of the data on the original scale response − baseline shows little structure because of extreme outliers in the tails of the distribution of pain tolerance. Figure 1 does not address questions such as how important or significant the differences between DD and the other groups are nor does it address how the treatments might be expected to affect individual responses or average group responses: a full statistical analysis is required.

A second problem is that after 240 seconds of immersion, the child is instructed to remove the arm from the water, and 240 was recorded as the response. Three observations have a censored baseline (id=21, 56, 60) and two have a censored response (id=17, 56). Three potential models of the censoring suggest three different accommodations to this problem. The first is to model the censored times as unknown values known to be longer than 240 seconds. This is the traditional censoring model. The second model says that those children are different from the remainder, and removes those cases from the data set. The third model argues that the children's arms have gone numb from the cold, that is, that they have fallen into a competing risk. Unfortunately, this competing risk does not identify itself except as an observation which lasts much too long. In this last model, had the child correctly responded to the cold probably he or she would have had long tolerances, but the tolerance would not have been as long as 240 seconds. The three models suggest respectively, impute a longer time for the 240 seconds, delete these cases, or impute shorter times for those observations. Accommodation of all three hypotheses at once is

impossible with classical censoring procedures.

A third problem is that case 52 was removed from the original analysis because he always lasted 240 seconds. This case is included in the analyses shown here. An important question is whether this case is influential or outlying. The justification for removing this case implies that the original analysis put some weight on the second censoring model, but other children with an occasional 240 second response were not deleted so the treatment of the censored observations was not consistent.

A fourth problem is that pediatric pain is concerned with individuals as well as groups. A finding of a small but significant treatment effect is a statement about a population, but the goal is to treat individuals. Individual variability may easily outweigh minor advantages of a particular treatment. Thus a classical approach does not give a complete inference.

I present a Bayesian analysis of this data that accounts for the problems just presented. Transformation of the baseline and response values is used to accommodate the heteroskedasticity. The purpose of transforming the response is to adjust for the nonconstant variance. Box and Cox (1964) presented an analysis of response transformations, and Box and Tidwell (1962) analyzed a model where the covariate was transformed to provide a better predictor of the response. Since our covariate and response are repeated measures, symmetry suggests that the transformations of the covariate and response be identical leading to a combined Box-Cox (1964) Box-Tidwell (1962) model.

In this paper I treat the unknown transformation parameter as unknown, the majority of traditional approaches estimate the unknown transformation parameter and then make inference conditional on that estimate. This has the unfortunate effect of ignoring some of the uncertainty in the model. Exactly how to do inference in transformation models has been a subject of much discussion in statistics (Bickel and Doksum 1981; Box and Cox 1982, Carrol and Ruppert 1981; Hinkley and Runger 1984). Frequentist solutions have depended strongly on asymptotic theory. Traditional inference both Bayesian and frequentist relies on having interpretable parameters for their inferences. Because of the unknown transformation parameter, the model is very non-linear in the parameters and appropriate interpretation is difficult. In contrast, a predictive

Bayesian approach gives small-sample, simple and interpretable inferences.

Prediction is used to unify the analyses throughout the paper. A prediction in classical analysis is usually a point estimate of the future value of an observation. A measure of uncertainty is usually, though not always attached; often the uncertainty estimates only the uncertainty because parameters are unknown and does not include the uncertainty associated with the fact that a new person can be expected to differ from the average or median response of all individuals with similar covariates. In this paper, predictions explicitly account for the random error associated with measuring a new person. Predictive distributions and summaries are used for inference. Thus treatment effects which appear substantial for individuals can be expected to persist over a population; the reverse is not necessarily true. Predictive diagnostics are used to identify influential and outlying observations and to assess the sensitivity of the inferences to the impact of the known (censoring) and suspected (outliers, influential observations) possible model misspecifications.

The paper is structured as follows: section 2 contains the inferential data analysis; section 3 contains the sensitivity analysis. The paper closes with discussion. The appendix contains the theoretical development of the statistical tools used: the model is discussed in appendix A; predictions are discussed in appendix B; outlier diagnostics are covered in appendix C; influence analysis is in appendix D; and general sensitivity analysis is in appendix E.

2 Data Analysis

An earlier analysis (Fanurik, Zeltzer, Roberts and Blount 1992) used least squares to estimate treatment and coping style effects, essentially this is a normal model. As discussed in the introduction, the current analysis improves the original model by adding a transformation parameter. Complete details of the model and computations are given in Appendix A. The discussion here covers the major points. Both the baseline b_i and the response r_i are transformed using the power transformation popularized by Box and Cox (1964). Formally, the data are modeled as

$$r_i^{(\lambda)} = x_i\beta_1 + b_i^{(\lambda)}\beta_2 + \varepsilon_i,$$

$$\varepsilon_i \sim N(0, \tau^{-1} I).$$

where x_i is a parameterization of the indicators for the 6 groups; b_i is the baseline tolerance; r_i is the response tolerance; β_1 is a vector of the treatment parameters; and β_2 is the coefficient of the baseline. The transformation is taken as the Box-Cox (1964) power family transformation

$$r^{(\lambda)} = \begin{cases} (r^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log r & \lambda = 0 \end{cases} .$$

Rather than follow the traditional approach of substituting a point estimate for λ, λ is treated as unknown to avoid possible underestimation of variability due to estimating the transformation.

The data along with case diagnostics are given in tables 1 (CS=1, attend) and 2 (CS=2, distract). Within treatment and coping style, observations are ordered by L_1, an influence diagnostic. The case diagnostics will be discussed later.

Computations are based on a simple random sample of size 4000 from an approximation to the posterior. The idea is that the inference from a Bayesian analysis is a probability density called the posterior. We can take a simple random sample (srs) from the posterior; summaries of the srs estimate summaries of the posterior. For example, a histogram of the λ component of the srs provides a quantitative idea of the range of values of the transformation parameter preferred by the data. Details are given in appendix A. Other summaries of interest of the posterior are developed later and in appendices A and B.

Sampling from $p(\beta, \tau|\lambda, Y)$ is straightforward. Given λ, the density of τ and β is the posterior resulting from normal theory linear regression, and this posterior is quite familiar, see for example, Box and Tiao (1973). Sampling from it is also straightforward. Sampling from $p(\lambda|Y)$ is more difficult, however, an approximating normal with mean .142427 and variance .0142605 provided a very good fit to $p(\lambda|Y)$. The approximation is quite good, the L_1 norm between the approximation and the exact distribution is only .018, which was considered too small to worry about. The L_1 norm can be thought of as the mimimum amount of probability that would have to be moved around to change the approximating normal density into the true density. The density $p(\lambda|Y)$ at $\lambda = 0$

is high, however following Rubin (1984) the transformation to normality, which is unknown and the scale of interest which is the original untransformed scale are kept separate. Samples from predictive distributions can be obtained by first sampling from the posterior of the parameters, then sampling a new observation conditional on those parameters, and then repeating the entire process. Appendix A gives the mathematical details.

Figure 2 gives predictive distributions for new observations with a baseline tolerance of 24 seconds, the median of the baseline measurements. Figure 2 is a kernel density estimate (Silverman 1986) based on samples of size 4000 from the predictive distribution. These distributions represent the possible observations we expect to see given the model and data for a new child with a baseline of 24 seconds from each of the 6 groups. Each distribution has a long right tail. It is difficult to distinguish amongst the predictions for the three attender groups AA, AD, and AN; in the distracters group, DD have substantially longer tolerances than DA which is somewhat longer again than DN. Plots for baselines of 6 and 120 seconds show essentially identical structure except for changes in location.

Table 3 summarizes Figure 2 with predictive means and standard deviations for all six groups and baselines of 6, 24 and 120 seconds. From table 3 and an exploratory regression of the standard deviations on the means, one can find that the standard deviations are consistently 2.7 more than 54% of the means for all table entries. In other words, the standard deviations quite closely track the means. This does not happen in the untransformed model where the standard deviations increase only as a function of the leverage of the observations. Unlike more common estimative inferences, the predictive standard deviation also includes individual variability, and thus shows how poorly we can predict a single individual's outcome. From table 3 and figure 2, we see that for a baseline of 6 seconds, the improvement over baseline for DD is only $21(= 27 - 6)$ seconds. In contrast, for the analysis (not given here) on the original scale without the transformation parameter, the estimated improvement due to the distract treatment for distracters is nearly one minute for a baseline of 6 seconds. There is also a regression to the mean effect, where the predictive means for a very low baseline of 6 seconds all show a predicted increase in response, while at 120 seconds, only the DD group has a predicted mean longer than the baseline.

ID	CS	TMT	B	R	L_1	100CPO
61	A	A	35.31	11.71	0.284	0.714
33	A	A	11.92	44.72	0.263	0.325
53	A	A	32.84	25.21	0.076	2.263
35	A	A	23.29	20.67	0.074	2.747
7	A	A	19.03	30.37	0.073	1.959
13	A	A	13.47	15.98	0.073	3.605
54	A	A	30.66	38.47	0.064	1.672
3	A	A	23.41	31.38	0.059	2.041
31	A	A	30.84	37.03	0.059	1.775
10	A	A	26.30	28.64	0.050	2.307
34	A	D	12.99	34.76	0.182	0.807
28	A	D	11.42	27.44	0.138	1.423
60	A	D	16.50	11.12	0.128	3.252
57	A	D	23.18	14.16	0.126	2.694
41	A	D	16.44	12.63	0.106	3.506
47	A	D	13.41	21.19	0.067	2.784
21	A	D	42.22	41.44	0.060	1.599
26	A	D	18.13	19.33	0.059	3.197
24	A	D	18.77	20.34	0.057	3.078
59	A	D	27.61	27.00	0.050	2.422
19	A	N	240.00	116.68	0.12	0.592
2	A	N	10.00	8.27	0.118	4.800
23	A	N	6.24	7.13	0.114	6.300
46	A	N	38.85	48.42	0.113	1.050
45	A	N	33.54	22.65	0.075	2.538
16	A	N	20.03	26.82	0.071	2.197
25	A	N	9.63	15.28	0.071	3.778
6	A	N	11.05	13.86	0.070	4.241
55	A	N	11.19	15.51	0.067	3.834
37	A	N	16.87	18.88	0.059	3.286

Table 1: Data and case diagnostics, part 1: attenders. Id number; CS, coping style, A=attend, D=distract; TMT, treatment, A=attend, D=distract, and N=null; B =baseline tolerance, seconds; R = response tolerance, seconds; L_1; $100 * \text{CPO}_i$.

ID	CS	TMT	B	R	L_1	100CPO
27	D	A	10.51	22.80	0.149	1.645
49	D	A	52.01	20.16	0.145	1.733
43	D	A	12.42	8.06	0.140	4.300
56	D	A	240.00	104.50	0.095	0.685
4	D	A	85.91	60.30	0.073	1.129
50	D	A	14.47	14.53	0.070	4.034
44	D	A	12.58	15.63	0.069	3.772
17	D	A	17.53	21.73	0.068	2.724
36	D	A	49.00	43.00	0.067	1.496
11	D	A	23.93	20.00	0.059	3.101
15	D	D	41.72	240.00	0.668	0.001
58	D	D	36.43	180.19	0.402	0.021
9	D	D	11.75	13.29	0.267	1.091
1	D	D	24.22	20.30	0.241	0.721
22	D	D	44.94	35.97	0.185	0.727
38	D	D	25.13	31.04	0.134	1.358
52	D	D	240.00	240.00	0.130	0.307
48	D	D	42.58	48.94	0.093	1.143
8	D	D	41.20	78.00	0.068	0.935
29	D	D	29.51	63.12	0.063	1.131
42	D	D	29.35	55.27	0.051	1.314
30	D	N	88.89	6.67	0.684	0.016
20	D	N	44.16	65.42	0.265	0.191
14	D	N	45.41	44.31	0.140	0.927
12	D	N	41.20	40.78	0.133	1.041
32	D	N	10.12	7.62	0.098	6.543
51	D	N	24.51	12.19	0.091	4.081
5	D	N	18.95	20.35	0.088	2.549
18	D	N	20.29	11.89	0.083	4.512
40	D	N	16.75	14.66	0.069	3.987
39	D	N	38.89	20.90	0.061	2.961

Table 2: Data and case diagnostics, part 2: distracters. Id number; CS, coping style, A=attend, D=distract; TMT, treatment, A=attend, D=distract, and N=null; B =baseline tolerance, seconds; R = response tolerance, seconds; L_1; $100 * \mathrm{CPO}_i$.

Table 4 gives the probability $P(Z_{new,j} > Z_{new,j'}|Y, B)$ that giving a child treatment j leads to a longer tolerance than giving that same child treatment j'. This is the probability that treatment j is better than treatment j' for that particular child. This probability depends on individual characteristics of the child entered into the model. For distracters, it is clear that the distract treatment leads to longer pain tolerance, compared to the other treatments, while for attenders no treatment is clearly longer. Mathematical details of the predictive inferences are given in Appendix B.

A classical analysis at this point would have compared treatments and groups using statistical significance. This leads to statements such as the p-value is the chance of seeing a statistic as large or larger, given the null hypothesis, in repeated samples. The statement says nothing about the data set at hand, rather it is a statement about the statistical procedure used (hypothesis testing). In contrast, the summaries of the treatment effects in table 4 are explicitly about the treatments in this data set, are relatively easy to understand, and answer questions of direct scientific interest.

The computation in table 4 is appropriate for covariates which can be manipulated by the experimenter. It assumes that the individual's error will not be affected by the assigned treatment. Separate probabilities are given for each coping style since coping style is an observed characteristic of the individuals. Each calculation assumes that the baseline measure, coping style and random predictive person error is the same under either treatment. The probabilities do not depend upon the baseline. These probabilities do happen to approximately equal the one sided p-value for the hypothesis test that the coefficient for the difference in treatment effects is equal to zero and may be interpreted using our intuition about the p-value scale. Thus if it is useful to continue using the usual language, we can say that for distracters, the distract treatment effect is significantly greater than either the attend or null treatements and that no other differences were significant. The most interesting conclusion is that $P(DD > DA) \approx 1$ and $P(DD > DN) = 1.000$. The next section will check to see if this result is stable under model perturbations, if it is, we can be comfortable with our conclusion that the distract treatment is better than the other two treatments.

CS/TMT	6	24	120
DD	27.1	56.9	146.2
	(17.4)	(34.2)	(81.8)
DA	11.4	26.0	74.0
	(8.3)	(16.7)	(44.3)
DN	8.6	19.8	58.9
	(6.8)	(13.3)	(36.0)
AD	13.0	29.7	82.9
	(9.8)	(19.5)	(50.3)
AA	14.3	31.2	87.1
	(10.5)	(19.8)	(48.3)
AN	12.3	26.7	78.7
	(9.7)	(17.3)	(46.4)

Table 3: Predictive means and (standard deviations) for the six groups, for baselines of 6, 24, and 120 seconds. Based on samples of size 2000.

Table 5 gives the probability that a person from any CS-TMT group combination has greater tolerance than someone else from another CS-TMT group. This computation is appropriate for observational characteristics not under the control of the experimenter. As in table 4 results do not depend on the particular baseline tolerance. The differences between tables 4 and 5 show how the predictive approach taken here distinguishes between randomized treatments versus observational studies; the inference is much stronger for the treatments than for observational characteristics. Two different (investigator allocated) treatments can be observed in the same individual, thus the random error is the same in both measures (one of which is necessarily hypothetical) but observational characteristics must necessarily have two independent random errors because it requires two people to make the comparison.

Table 4 is not quantitatively very different from the inference that might be made by someone using the analysis of Bickel and Doksum (1981). Interestingly, their method of inference is perhaps the least popular of the various approaches to classical inference. However, the inferences displayed in table 5 are quite different from the inference advocated by anyone previously in a transformation model.

	Coping Style			
	Distracters		Attenders	
Treatment	Attend	None	Attend	None
Distract	0.99925	1.000	.35875	.59075
Attend		0.8355		.7285

Table 4: Posterior probability that placing a person on the row treatment will lead to greater tolerance than putting that same person on the column treatment. Based on samples of 4000.

	dd	da	dn	ad	aa	an
dd	—	0.840	0.900	0.798	0.768	0.817
da		—	0.615	0.438	0.397	0.466
dn			—	0.327	0.291	0.353
ad				—	0.458	0.528
aa					—	0.570

Table 5: Posterior probability that placing a person from the coping style and group from the row has greater tolerance than a different person from the coping style and treatment group given in the column headings. The ordering of the groups is the same as in table 3, dd $\gg$ aa $>$ ad $>$ an $>$ da $\gg$ dn. Based on a sample of 4001.

3 Sensitivity Analysis

This section explores the robustness of the conclusions of the previous section to changes in model specification. Two kinds of model changes are explored, case deletion using case diagnostics, and a sensitivity analysis to determine the effect of altering the likelihood contribution of the censored cases with baseline or response tolerances of 240 seconds.

The purpose of sensitivity analysis is to search through a large space of a priori plausible alternative models at low cost. An alternative model can be ignored if it leads to the same conclusions as the original model or if it is not supported by the data (Weiss 1994). If no alternative models are found which meet both these criteria then we stand by the basic model and its conclusions. If we find some models which lead to different conclusions and are supported by the data, then we will identify a few extreme models to use to fit to the data. The method used here for identifying if the conclusions are changed is not sensitive to particular conclusions of the analysis, it is a global measure of differences

between the conclusions from the original model and the conclusions from the alternative model. Thus the next step fits these alternative models and compares the specific conclusions of importance with those from the original model. If the main conclusions are similar to those from the basic model, then we can be much more comfortable in the conclusions. If conclusions are qualitatively different, we conclude that the data does not support strong conclusions. An alternative to the sensitivity analysis is to fit a mixture model encompassing all possible perturbations to the basic model. This requires substantially more work, and would be sensitive to the prior probabilities specified in the mixture model. Mathematical details of the diagnostics used are in appendices C, D, and E.

First case diagnostics were checked. The conditional predictive ordinate (CPO) outlier statistic, and L_1 norm case influence statistic, were calculated; these values were given in tables 1 and 2. The CPO statistic was first proposed by Geisser (1980) and has been analyzed by Geisser (1987, 1989, 1990), Pettit (1985, 1990), Pettit and Smith (1985) and Weiss (1994). The CPO_i is a Bayes factor in favor of the original model against the model which deletes case i. The Bayes factor BF for a model M_0 against model M_1 updates the prior odds $p(M_0)/p(M_1)$ of the two models to the posterior odds $p(M_0|Y)/p(M_1|Y) = \text{BF} * p(M_0)/p(M_1)$. The smaller CPO is, that is, the smaller the Bayes Factor BF is, the more outlying is the observation and the more the data support M_1, the case deleted model. A factor of 10 difference in two values of CPO indicates roughly that the data indicate that the observation with the smaller CPO has 10 times greater odds of being an outlier than the other observation. The CPO statistic can be converted to the $(0, 1)$ probability scale. This probability is large only for cases 15, 30, and 58, which had probabilities of .76, .07 and .05 of being outliers. All other cases all have probabilities of being outliers that are less than .006. More information about CPO is included in appendix C

The L_1 influence diagnostic is one of Csiszár's (1967) divergences between densities. It measures how different two posteriors are. Earlier we used the L_1 divergence to assess the difference between the exact posterior distribution $p(\lambda|Y)$ and an approximating normal distribution. The Kullback divergence Kullback (1958 p. 6) is perhaps the most commonly used divergence in statistics

but the L_1 divergence seems easier to interpret (Weiss 1994). Appendix D includes a short description of L_1 and gives a predictive interpretation for it. There is a diagnostic for each case in the data set; the values are given in tables 1 and 2. The L_1 diagnostic ranges from 0 to 1; zero indicates no influence: the model with all the data in it and the reduced data model lead to identical conclusions in all respects. An L_1 of less than .1 is considered to be relatively uninfluential, around .3 is moderately influential, above .5 is high, and an L_1 of 1 indicates that the posterior from the case deleted model does not share any support with the basic model. The L_1 influence measure is a global measure, and is susceptible to influence on aspects of the posterior that may not be of direct interest.

Case 52 was neither influential nor outlying by itself. Consequently, deleting on its own was not considered further.

Three cases $30, 58, 15$ are identified as outlying by CPO, and these three are also very influential by L_1, with values over .4 for each case. As part of the final stage of the analysis, these three observations were deleted as a set to form one perturbation.

Next I assess influence of the values for the censored observations. Several children kept their arms immersed for the maximum permitted time of 240 seconds on either the baseline or the response. The baseline is 240 seconds for cases 19, 52, and 56; the response is 240 for cases 15 and 52; additionally, $y_{58} = 180.19$. The next largest observation is $y_{19} = 116.68$. The standard censoring model treats these observations as conditionally normal given the parameters, but with unobserved values known to be larger than 240. However, alternative models are possible. The reason for the long tolerances is quite probably that the arms become numb; the children are no longer responding to pain or cold as a stimulus. The numbness is a competing risk, but not one whose onset is observed. More appropriate might be to take 240 as an upper bound on the responses. As a lower bound, we might pick either the largest, or second largest uncensored observation, $y_{58} = 180.19$ or $y_{19} = 116.68$. Since y_{19} is only 12 seconds larger than the next observation, while 180.19 was a full minute larger than 116.68, it was also considered as a possible numbness induced measurement. To assess the influence of alternative values, values of

120, 150, 180, 240, and 300 seconds were substituted for the censored values. For values of 120 and 150, case 58 was both perturbed and unperturbed.

Initially, this was intended to be the first stage of a more complicated analysis. The thought was to find a base level such as 120 seconds, and then to consider what happened if the actual imputed values were allowed to vary from case to case. However, the perturbations from 240 to 300 and from 240 to 180 were surprisingly uninfluential. Initially, these were expected to be very influential changes. It was judged that the further perturbations were unlikely to cause major changes to the posterior, given that changing the value 240 by a full minute was so uninfluential. It was decided that the additional sensitivity analysis were not needed.

Two multiple case deletions were considered, the four cases with baseline or response of 240 and the three influential outliers from the single case diagnostics.

Information about these model perturbations are given in table 6, with L_1 influence statistic and CPO outlier statistics. For different values of the perturbed maximal tolerances, CPO can be sensibly be compared to each other. For the case deletion values, the CPO values have been adjusted to make them comparable to each other and to the other values in the table. Appendix C explains about the case deletion adjustments of CPO. The smaller CPO is, the more the data prefer that perturbation to the original unperturbed model.

The perturbation of values from 240 to 300 is actually less supported by the data than the null model since CPO > 1, but the other perturbations from 240 to shorter values are more supported than the null. The value of CPO decreases smoothly as the perturbed tolerance measure decreases towards 120, and as all 5 observations with baseline or response greater than 120 are changed to 120. All of the changes are influential, and the influence is increasing with decreasing CPO.

Two further analyses are suggested by the sensitivity analysis: one deleted the three outlying cases, and one where the large x and y values are perturbed to 120 seconds. These two perturbations are quite different. In the families of perturbations considered, these are the most influential and the most outlying. If the major conclusions do not change under these perturbed models, then we can be more comfortable that the major conclusions are not sensitive to the

Type of perturbation	No. of cases	cases	new value	L_1	CPO
Increase 240's	4	15, 19, 52, 56	300	.3	8.12
No perturbation	0			0	1
decrease 240's	4	15, 19, 52, 56	180	.365	.077
decrease 240's	4	15, 19, 52, 56	150	.511	.022
decrease 240's	4	15, 19, 52, 56	120	.597	.0090
decrease 240's and 180	5	15, 19, 52, 56, 58	150	.648	.0044
decrease 240's and 180	5	15, 19, 52, 56, 58	120	.805	.00028
delete cases	3	15 30 58		.944	3.94e-8
delete cases	4	15, 19, 52, 56		.664	2.32e-4

Table 6: Summary of perturbation results. Lists type of perturbation; number of cases involved; case numbers; new baseline or response value for changed 240 and 180 values; L_1 influence statistic; CPO outlier statistic. For the case deletion perturbations, the CPO statistic has been adjusted as discussed in Appendix C.

assumptions either about the cases with $y = 240$ or $x_{2i} = 240$ or with the outliers left in the model. Interest lies especially in checking table 4 for these perturbed models. The results are given in table 7. There are no major changes in the outcomes of distracters counseled to distract versus counseled to attend or none. The largest change seems to be for comparing da to dc which is not of great interest because both treatments are second rate treatments.

4 Discussion.

The predictive inference tools provide interpretable inferences which distinguish between observational characteristics of the subjects and manipulatable treatments.

The sensitivity and diagnostic tools greatly strengthen the certainty of the conclusions that can be drawn from this data set. In spite of the devil's-advocate efforts to produce an alternative model which leads to contradictory conclusions, no such model was found. Without the diagnostic and sensitivity analysis, we are left wondering whether the conclusions rest on a true underlying treatment effect or on the cases of questionable quality. Thus the conclusion that the distract treatment is substantially better than the attend or the none for dis-

		Model		
				240 →
CS	TMT > TMT	Orig	Del3	120
DIS	DIS > ATT	1.0	1.0	.99
	DIS > NON	1.0	1.0	1.0
	ATT > NON	.84	.56	.91
ATT	DIS > ATT	.36	.36	.37
	DIS > NON	.59	.65	.50
	ATT > NON	.73	.77	.63

Table 7: Predictive probabilities that one treatment is better than another under the original and two perturbed models, the delete 3 outliers model, and the model that adjusts all large observations down to 120 seconds.

tracters appears solid based on this analysis. On the other hand, sensitivity analysis for the results in table 3 suggests that these values have additional uncertainty that is not quantified by the standard errors. The proper model is uncertain, and different models will lead to different numerical conclusions. Only the sensitivity of the qualitative conclusions in table 7 were shown; the sensitivity analyses for 3 weren't shown here.

One shortcoming of the sensitivity analysis presented here is that the different plausible perturbations define separate models, and the possibility of joint perturbation has not been considered. This in principle invites a combinatorial explosion of models as models with 2, 3 and more outliers are considered, plus combinations with other perturbations. An infinite regress of analyses is always possible and the analysis shown here was deemed satisfactory for this data set since the sensitivity analyses strongly support the qualitative conclusions from the basic analysis.

An additional value to the sensitivity is for design of future trials. Future studies may have the children remove their arms after three or even two minutes. The sensitivity analysis shows that while details of the conclusions will change, the overall qualitative conclusions will not.

References

[1] Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations

revisited. *JASA*, 76 296-311.

[2] Box, G. E. P. (1980). Reply to comments on "Sampling and Bayes' inference in scientific modelling and robustness" *JRSS-A*, 143, 425- 430.

[3] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations (with discussion). *JRSS-B*, 26, 211-252.

[4] Box, G.E.P. & Cox, D.R. (1982). An analysis of transformations revisited, rebutted. *JASA*, 77, 209-210.

[5] Box, G. E. P. and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley Publishing Co., 1973

[6] Box, G.E.P. & Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, 4, 531-550.

[7] Carlin, B. P. & Polson, N. G. (1991). An expected utility approach to influence diagnostics. *JASA*, 86, 1013-1021.

[8] Carlin, B. P. & Polson, N. G. (1992). Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics 4*, pp. 577-586, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, (Eds). Oxford: Clarendon Press.

[9] Carrol, R. J. and Ruppert, D. (1981). On prediction and the power transformation family. *Biometrika*, 68, 609-615.

[10] Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* **2**, 299-318.

[11] Fanurik, D., Zeltzer, L. K., Roberts, M. C. and Blount, R. L. (1993). The relationship between children's coping styles and psychological interventions for cold pressor pain. *Pain*, in press.

[12] Geisser, S. (1980). Comments on "Sampling and Bayes' inference in scientific modelling and robustness" *JRSS-A*, 143, 416- 417.

[13] Geisser, S. (1987). Influential observations, diagnostics and discordancy tests. *J. Applied Statistics*, 14, 133-142.

[14] Geisser, S. (1989). Predictive discordancy tests for exponential observations. *Canadian Journal of Statistics*, 17, 19-26.

[15] Geisser, S. (1990). Predictive approaches to discordancy testing. In *Bayesian and likelihood methods in statistics and econometrics: Essays in honor of George A. Barnard*, S. Geisser;J. S. Hodges, S. J. Press, A. Zellner, (Eds), pp. 321-335. Amsterdam: Elsevier-North Holland.

[16] Geisser, S. (1991). Diagnostics, divergences and perturbation analysis. In *Directions in Robust Statistics and Diagnostics I*, W. Stahel and S. Weisberg, (Eds), pp. 89-100. Springer-Verlag, New York.

[17] Geisser, S. (1992). Bayesian perturbation diagnostics and robustness. In *Bayesian analysis in statistics and econometrics, Lecture notes in Statistics, Vol. 75*, P. K. Goel and N. S. Iyengar, (Eds)., pp. 298-301.

[18] Gelfand. A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398-409.

[19] Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57,1317-1339.

[20] Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data (with discussion). *JASA*, 79, 302-320.

[21] Kass, R. E., Tierney, L. & Kadane, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika*, **76**, 663-674.

[22] Kullback, S. (1959). *Information Theory and Statistics*, Gloucester, MA: Peter Smith.

[23] Pettit, L. I. & Smith, A. F. M. (1985). Outliers and Influential Observations in Linear Models. In *Bayesian Statistics 2*, Eds. J. M. Bernardo, M. DeGroot, D. Lindley, and A. F. M. Smith, pp. 473-94 Amsterdam: North Holland.

[24] Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution. *JRSS-B*, 52, 175-184.

[25] Rubin, D.B. (1984). Comments on "The analysis of transformed data". *JASA*, 79, 309-312.

[26] Rubin, D. B. (1983). A case study of the robustness of Bayesian methods of inference: Estimating the total in a finite population using transformations to normality. *Scientific Inference, Data Analysis, and Robustness*, Eds. G.E.P. Box, Tom Leonard, Chien-Fu Wu. Academic Press, 213-244.

[27] Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The Statistician*, 41, 169-178.

[28] Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

[29] Stigler, S. M. (1980). Comments on "Sampling and Bayes' inference in scientific modelling and robustness". *JRSS-A*, 143, 424-425.

[30] Weiss, R.E. (1994). Bayesian sensitivity analysis. Submitted for publication.

A The Model

The general model considered here is

$$\begin{aligned} Y^{(\lambda)} &= X_1\beta_1 + X_2^{(\lambda)}\beta_2 + \varepsilon, \\ &\quad \varepsilon \sim N_n(0, \tau^{-1}I) \end{aligned} \tag{1}$$

where $Y^{(\lambda)}$ is a vector of length n with i^{th} element $y_i^{(\lambda)}$, a transformation of y_i parameterized by λ. Similarly, $X_2^{(\lambda)}$ is a baseline measure with individual elements $x_{i2}^{(\lambda)}$; $X_1, n \times p$, a matrix of fixed covariates; β_1 is a vector of regression coefficients; β_2 is a scalar regression coefficient; and ε is a vector of iid normal errors. The analysis of data with power transformations of both response and predictor variables is not new, especially in the Econometrics literature (see Sakia 1992 for a review). This section treats the analysis of (1) generally. For the pain data, X_1 is a parameterization of the design matrix of a 2 by 3 analysis of variance, while X_2 is the baseline measure.

Let $\beta^T = (\beta_1^T, \beta_2), and X_\lambda = (X_1 | X_2^{(\lambda)})$ has a subscript to remind us of the dependence of X on λ. The likelihood resulting from (1) is

$$L(\beta, \tau, \lambda) \propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}\left(Y^{(\lambda)} - X_\lambda\beta\right)^t\left(Y^{(\lambda)} - X_\lambda\beta\right)\right\} J(Y, \lambda)$$

where

$$J(Y, \lambda) = \prod_{i=1}^{n} J(y_i, \lambda) = \prod \partial y_i^{(\lambda)}/\partial y_i$$

is the Jacobian of the transformation. For the Box-Cox transformation, $J(y_i, \lambda) = y_i^{\lambda-1}$. A heuristic for including the jacobian is as an adjustment so that the sampling density of y integrates to 1 back on the untransformed scale. The maximizers of $L(\beta, \tau, \lambda)$ for fixed λ are the usual maximum likelihood estimates

$$\hat{\beta}(\lambda) = (X_\lambda^T X_\lambda)^{-1} X_\lambda^T Y^{(\lambda)}$$

and

$$\hat{\tau}(\lambda) = n(\mathrm{RSS}(\lambda))^{-1}$$

where $\mathrm{RSS}(\lambda) = (Y^{(\lambda)} - X_\lambda\hat{\beta}(\lambda))^T(Y^{(\lambda)} - X_\lambda\hat{\beta}(\lambda))$. The profile log likelihood of λ is

$$\ell(\lambda) = \ell(\hat{\beta}(\lambda), \hat{\tau}(\lambda), \lambda) = \frac{n}{2}\log\left(\frac{n}{\mathrm{RSS}(\lambda)}\right) - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^{n} \log y_i \ .$$

The profile log likelihood of λ is the log likelihood $\ell(\beta, \tau, \lambda)$ with $\hat{\beta}(\lambda)$ and $\hat{\tau}(\lambda)$ substituted for β and τ. This can be maximized and plotted as in Box and Cox (1964), and the asymptotic χ^2 test based on $\chi^2 = 2(\ell(\hat{\lambda}) - \ell(\lambda_0))$ can be used to test $H_0 : \lambda = \lambda_0$ and to produce confidence intervals.

Bayesian inference in this model permits simple graphical and numerical posterior summaries, including summaries that can replace frequentist multiple comparison procedures. The posterior

$$p(\lambda, \beta, \tau|Y) = p(\lambda|Y)p(\tau|\lambda, Y)p(\beta|\tau, \lambda, Y)$$

can be explicitly calculated up to the normalizing constant for $p(\lambda|Y)$ which, since λ is often a scalar, can be dealt with quite easily. Here we take a uniform improper prior $p(\beta, \tau, \lambda) \propto \tau^{-1}$. In the pain example, this is useful for imitating the original classical analysis. It also was sensible because it was felt that there was not much useful prior information to put into the prior. The regression coefficients β are then conditionally normal:

$$[\beta|\tau, \lambda] \sim N\left(\hat{\beta}(\lambda), (\tau X_\lambda^T X_\lambda)^{-1}\right) , \tag{2}$$

while τ given λ has the gamma distribution

$$[\tau|\lambda, Y] \sim \Gamma\left(\frac{n-p}{2}, \frac{\text{RSS}(\lambda)}{2}\right) \tag{3}$$

with density

$$p(\tau|\lambda, Y) \propto \tau^{\frac{n-p}{2}-1} \exp\left\{-\frac{\text{RSS}(\lambda)\tau}{2}\right\}$$

and the density of λ is proportional to

$$p(\lambda|Y) \propto |X_\lambda^T X_\lambda|^{-1/2}(\text{RSS}(\lambda))^{-(\frac{n-p}{2})}(\prod y_i)^{\lambda-1} . \tag{4}$$

Sampling-based Bayesian inference (Gelfand and Smith 1990) is used for the computations in this model. First sample from $p(\lambda|Y)$, then from $p(\tau|\lambda, Y)$ and finally from $p(\beta|\lambda, \tau, Y)$. Sampling from $p(\lambda|Y)$ can be accomplished either by importance sampling (Geweke 1989) or discrete approximation (see Rubin 1983 for an example in the Box-Cox transformation model). In the pain data a normal approximation to $p(\lambda|Y)$ suffices. Repeating the sampling M times produces M independent samples $\theta^{(m)} = (\beta^{(m)}, \tau^{(m)}, \lambda^{(m)}), m = 1, \ldots M$ from the posterior.

B Predictive Inference.

The predictive distribution of a new observation Z at a particular covariate value x_f is the integral of the conditional density of Z given θ with respect to the posterior of θ given the observed data. In symbols, $f(Z|x_f, \theta)$ is

$$f(z|x_f, \theta) = \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\left\{-\frac{\tau}{2}\,(z^{(\lambda)} - x_{f,\lambda}^T\beta)^2\right\} z^{\lambda-1}$$

where $x_{f,\lambda}$ is the covariate vector after transforming the baseline by λ, and $z^{\lambda-1}$ is the jacobian $J(z,\lambda)$ from transforming from the $z^{(\lambda)}$ scale where normality holds back to the scale z where we take observations and our intuition works. Algebraically, the predictive distribution is

$$f(z|Y, x_f) = \int\int\int f(z|x_f, \theta)p(\beta, \tau, \lambda|Y)d\beta d\tau d\lambda\ .$$

Predictive distributions are easily simulated by sampling w from a $N(x_{f,\lambda^{(m)}}^T\beta^{(m)}, (\tau^{(m)})^{-1})$, then backtransforming w into $z_f^{(m)}$. When the transformation is the standard power family transformation

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y_i & \lambda = 0 \end{cases}$$

then

$$z_f^{(m)} = \begin{cases} (\lambda w + 1)^{1/\lambda} & \lambda \neq 0 \\ e^w & \lambda = 0 \end{cases}$$

If $(\lambda w + 1) \leq 0$ then resample w again. As statisticians have usually done with the Box-Cox model, it is assumed that the response is non-negative and the censoring due to this constraint has negligible impact on the inference.

Inference for comparing different treatments or treatment combinations can be done in both a predictive graphical fashion and with appropriate summary statistics. This has several advantages over usual point estimation and testing procedures. It avoids the uninterpretability of marginal estimated coefficients and standard errors (Box and Cox 1982, Carroll and Ruppert 1981; Hinkley and Runger 1984). For example, if the analysis is done in the square root scale, then a treatment effect coefficient is a difference of square root times, which is not too easy to interpret. If the transformation is on an unknown scale, then the coefficient doesn't even have that limited interpretation. The predictive approach

also permits assessment of practical differences on the original scale of measurement, and permits uncertainty due to the transformation to be propagated. The predictive distribution on the original measurement scale can be a great aid in communicating with subject matter researchers who have difficulty interpreting log, square root and other transformed scales. Another advantage of the predictive framework is that it makes all parameterizations of the ANOVA model equivalent, and we need not worry about proper interpretation of coefficients. Relevant summary measures can easily be estimated, for example means and standard deviations of predictions and for differences in predictions. Taking a predictive approach permits assessment of practical as well as statistical impact of treatments.

The predictive distribution of $z|x_f$ for observations in different treatment groups may be computed for equal values of the baseline covariates to assess treatment differences. We also may be interested in the difference in response of either the same or two different people in two different groups denoted by x_{f1} and x_{f2}. Let z_{f1} and z_{f2} be conditionally independent predictions at x_{f1} and x_{f2}. A plot of the densities of z_{f1} and z_{f2} is the current best guess as to the responses of two people with covariates x_{f1} and x_{f2}. One can also consider the posterior of $z_{f1} - z_{f2}$. These densities are relevant when the two groups are distinguished by unchangeable characteristics of the individuals. A summary measure of the difference between z_{f1} and z_{f2} of particular interest is

$$\begin{aligned} P(z_{f1} > z_{f2}) &= P((x_{f1}^t\beta + \epsilon_{f1})^{(\lambda)} > (x_{f2}^t\beta + \epsilon_{f2})^{(\lambda)} \\ &= P((x_{f1} - x_{f2})^t\beta > \epsilon_{f1} - \epsilon_{f2}). \end{aligned}$$

The dependence of the coefficients $x_{fk}, k = 1, 2$ on λ is suppressed to make the formulae readable. Suppose $(x_{f1} - x_{f2})^t\beta = \beta_j$, as happens in the pain data when we are comparing two predictive distributions that differ only by a single indicator variable such as a treatment effect; the baseline measurements for both sets of covariates are assumed equal. Then

$$P(z_{f1} > z_{f2}) = E\left[\Phi\left((.5\tau)^{1/2}\beta_j\right)\right], \tag{5}$$

where $\Phi(a)$ is the standard normal cumulative distribution function. The prob-

ability (5) can be approximated by

$$P(z_{f1} > z_{f2}) \approx \Phi\left(\frac{\hat{\beta}_j}{(\mathrm{Var}(\hat{\beta}_j) + 2\hat{\tau}^{-1})^{.5}}\right).$$

where $\mathrm{Var}(\hat{\beta}_j)$ is the posterior variance of β.

When the difference between x_{f1} and x_{f2} is a treatment which can be modified by the experimenter, we may be interested in a different comparison. Let z^*_{f1} and z^*_{f2} be corresponding predictions at x_{f1} and x_{f2} whose unobserved normal errors ϵ^*_{f1} and ϵ^*_{f2} are equal with probability 1. Then $P(z^*_{f1} > z_{f2}|Y)$ is the probability that treatment 1 produces a larger response than treatment 2 on one person and, again assuming $(x_{f1} - x_{f2})^t\beta = \beta_j$ then

$$P(z^*_{f1} > z^*_{f2}|Y) = P(\beta_j > 0|Y) \approx \Phi\left(\frac{\hat{\beta}}{(\mathrm{var}(\hat{\beta}_j))^{.5}}\right) \tag{6}$$

holds for monotone transformations and error structures that are location-scale models after transformation. Equation (6) provides a predictive interpretation to the coefficient β_j. The approximation is equal to the classical one sided p-value for testing $H_0 : \beta_j = 0$. Neither probability (5) and (6) depends on the baseline measurements. From (6) compared to the approximation to (5) we see that we have a stronger inference for treatment effects than for the coefficients of observed variables.

Treatment and group difference assessments can be evaluated predictively by taking predictive means within groups as well as on the individual level given here. For example one could compute the probability that the average response on treatment 1 is greater than the average response on treatment 2. If substantive differences are found for individuals, then necessarily there will be differences for groups; thus an analysis that shows effects for individuals will show effects for groups. The converse need not hold. Group level evaluations may be pertinent for global decisions by policy makers, however individual differences in utility can easily outweigh global policy recommendation, something less likely to happen when a treatment has a beneficial effect that is noticeable on the individual level. Formula (6) shows that for location-scale models with monotone transformations inference about treatments is the same for an individual and a population.

C Outlier Diagnostics.

Diagnostics can and should be implemented in any new model. An advantage of the tools explained here is that they are readily applied in other models. This section discusses an outlier diagnostic while the next section discusses an influence diagnostic.

It is of specific interest to identify outliers in pediatric pain research: these children may have low or high tolerance and may be at risk for (not necessarily respectively) low or high usage or avoidance of medical procedures. However, the outliers of interest are from marginal models where the distribution of the baseline measure is also modeled; not outliers conditional on the baseline measure as in model (1); if both the baseline and response are small, the case will not be identified as an outlier yet these children are specifically of interest for further intervention. In the current analysis, outlier statistics are used to identify observations which may be discordant and may affect the analysis.

A Bayesian outlier statistic is CPO, the conditional predictive ordinate of Geisser (1980, 1987, 1989, 1990), Pettit and Smith (1985), Pettit (1990) and Weiss (1994). Define

$$\text{CPO}_i = f(y_i|Y_{(i)}) = \int f(y_i|\theta, x_i)p(\theta|Y_{(i)})d\theta,$$

which is also the normalizing constant in the updating version of Bayes theorem

$$p(\theta|Y) = \frac{p(\theta|Y_{(i)})f(y_i|\theta)}{\text{CPO}_i} \tag{7}$$

and also a computational byproduct of the influence diagnostics of the next section. Carlin and Polson (1991, 1992) estimate CPO_i directly from (7), requiring n samples from each of the n densities $p(\theta|Y_{(i)})$. By solving (7) for $p(\theta|Y_{(i)})$, CPO can be computed as

$$\text{CPO}_i^{-1} \approx M^{-1}\sum_{m=1}^{M}[f(y_i|\theta, x_i)]^{-1}.$$

which only requires a single posterior sample from $p(\theta|Y)$.

The diagnostic CPO is dependent upon the scale of analysis (Box 1980, Geisser 1980, Stigler 1980); this is solved by choosing a particular measurement

scale for analysis: for the pain data, the original scale in seconds is chosen. Changing the analysis scale from seconds to hours also causes a change in the absolute level of CPO_i, this can be normed internally by realizing that most of the data is good data, and that most of the CPO statistics are representing good data, not bad. The lower quartile of CPO for this data set is .01, which is a convenient number to use: all CPO statistics for case deletion have been multiplied by 100 in tables 1 and 2. In data sets with non independent and identically distributed data, the maximum possible value $\sup_z f(z|x_i, Y_{(i)})$ for any one observation i is dependent on the covariates. In normal theory multiple linear regression with known σ, this maximum value is $((1-h_i)/(2\pi h_i \sigma^2)^{-1})^{.5}$. For this data set and after transforming the baseline variables by the mean value of λ, the maximum value varied only by a factor of 2, which was relatively unimportant given the wide range of the CPO_i themselves.

For multiple case deletion, with only a few specially select sets of deleted cases, it is less easy to provide an internal norming of the values. A small simulation was undertaken, where 100 sets of 4 observations were deleted, and CPO computed. The five number summary (min, lower quartile, median, upper quartile and max) of CPO from this sample was ($2.43e$-11, $2.21e$-08, $1.01e$-07, $3.55e$-07, $2.97e$-06). The set of four observations with censored x and y values have a CPO that is 10.5 times smaller than the smallest CPO in the 100 sets of 4, suggesting that this set is outlying. To assess the outlyingness of the three case deletion, first note that the 25^{th} percentile of the single observation CPOs is .01, while the same percentile of the sets of 4 is $2e$-8, or approximately the fourth power of the single observation CPOs, suggesting that for sets of 3 observations, approximately, a typical value of CPO could be expected around $1e$-6. In contrast, $\text{CPO}_{15,30,58} \approx 4e$-14, a factor of $4e$-8 smaller than $1e$-6 and 2000 times smaller than the smallest CPO from the sets of four. Thus the data appear to strongly support the deletion of the three outliers over the deletion of the 4 children with long tolerance times. The values of CPO in table 6 have already been adjusted for the two multiple-case-deletion perturbations .

D Influence Analysis.

Influential cases can be identified using the L_1 distance influence statistic of Weiss (1994). See also Geisser (1991,1992).

$$L_{1i} = \frac{1}{2}\int \left| P(\theta|Y) - P(\theta|Y_{(i)}) \right| d\theta$$

where the parameter vector $\theta = (\beta, \tau, \lambda)$ and $p(\theta|Y_{(i)})$ is the posterior density given the data omitting the i^{th} case. The statistic L_{1i} is bounded between 0 and 1; 0 indicating no influence and 1 indicating that the two posteriors $p(\theta|Y)$ and $p(\theta|Y_{(i)})$ do not share support. If interest lies in a particular parameter or predictor L_{1i} can be replaced by

$$L_{1i,\gamma} = \frac{1}{2}\int \left| p(\gamma(\theta)|Y) - p(\gamma(\theta)|Y_{(i)}) \right| d\gamma \ .$$

However $0 \leq L_{1i,\gamma} \leq L_{1i} \leq 1$, so that L_{1i} small guarantees $L_{1i,\gamma}$ small. Also if a vector of predictions $\underline{z}_m$ are of interest then taking $\theta^* = (\beta, \tau, \lambda, \underline{z}_m)$ and

$$L^*_{1i} = \frac{1}{2}\int \left| p(\theta^*|Y) - p(\theta^*|Y_{(i)}) \right| d\theta^*$$

then

$$L^*_{1i} = L_{1i}$$

and so $L_{1i} \geq L_{1i,\underline{z}_m} \geq 0$. In fact, an even stronger statement is possible which gives L_1 a predictive interpretation. Let $\underline{z}_m$ be an set of m predictions for each m and assume that the ratio $f(y_i|\underline{z}_m, Y_{(i)})[f(y_i|\theta)]^{-1}$ converges pointwise almost everywhere to 1. Then

$$L_{1i,\underline{z}_m} \nearrow L_{1i}.$$

In other words, if the limiting information in the predictive densities is equivalent to knowing the parameter θ, then the limiting influence on the predictions is equal to the influence on the posterior.

If interest lies in a posterior or predictive expectation $E[\gamma(\theta^*)|Y]$, then influence may be assessed by the quantity

$$\begin{aligned} d_i &= E\left[\gamma(\theta^*)|Y_{(i)}\right] - E\left[\gamma(\theta^*)|Y\right] \\ &= \int \gamma(\theta^*)\left[P(\theta^*|Y_{(i)}) - p(\theta^*|Y)\right] d\theta^* \\ &= E\left[\gamma(\theta^*)(\frac{\mathrm{CPO}_i}{f(y_i|\theta, x_i)} - 1)\right] \ . \end{aligned}$$

so

$$d_i = \text{CPO}_i * \text{Cov}\left(\gamma(\theta^*), [f(y_i|\theta, x_i)]^{-1}\right) \ . \tag{8}$$

If γ is an indicator function, then $d_i = L_{1i,\gamma}$. Table 7 permits us to compute d_i for γ's of particular interest.

E General Perturbations

Several observations have baseline x_i or response y_i values censored at 240 seconds. The appropriate values are difficult to model, instead a sensitivity analysis will be undertaken to see if deleting these observations as a set or changing their values leads to a change in one of the major conclusions of the analysis. Define a perturbation function (Weiss 1994; Kass, Tierney and Kadane 1989) for example by

$$h(\theta) = \frac{f(y_i + \omega|x_i, \theta)}{f(y_i|x_i, \theta)} \ ,$$

to assess the influence of changing y_i to $y_i + \omega$. Then the perturbed posterior is

$$p_h(\theta|Y) = p(\theta|Y)h(\theta)[E[h(\theta)]^{-1}.$$

The influence of this perturbation can be assessed by L_{1i} or d_i, and the previous formulas and interpretations apply directly without modification except that the perturbation has changed from single case deletion to y_i or x_i perturbation. The perturbations can also be more complicated than $h(\theta)$ above; most perturbations in table 6 are combinations of x and y perturbations for four or five cases.

The CPO statistic can also be computed for these perturbations (Weiss 1994). Consider the perturbed posterior as resulting from a perturbed model, then CPO is the Bayes factor against the perturbed model in favor of the original model. If $h(\theta)$ is a ratio of products of sampling densities, then $\text{CPO}_h = [E[h(\theta)]$.

For all perturbations $h(\theta)$ to a model we calculate both influence and outlier statistics. If CPO is small, the perturbation is supported by the data, and if L_1 is small, then the perturbation has no effect on the inference. As with the influence and outlier statistics for individual cases, we only need to worry about perturbations which are both influential and outlying.

Figure Captions

Figure 1. Box plots of log response minus log baseline for the six groups.

Figure 2. Predictive distributions of new observations for each of the 6 groups, given a baseline measurement of 24 seconds.

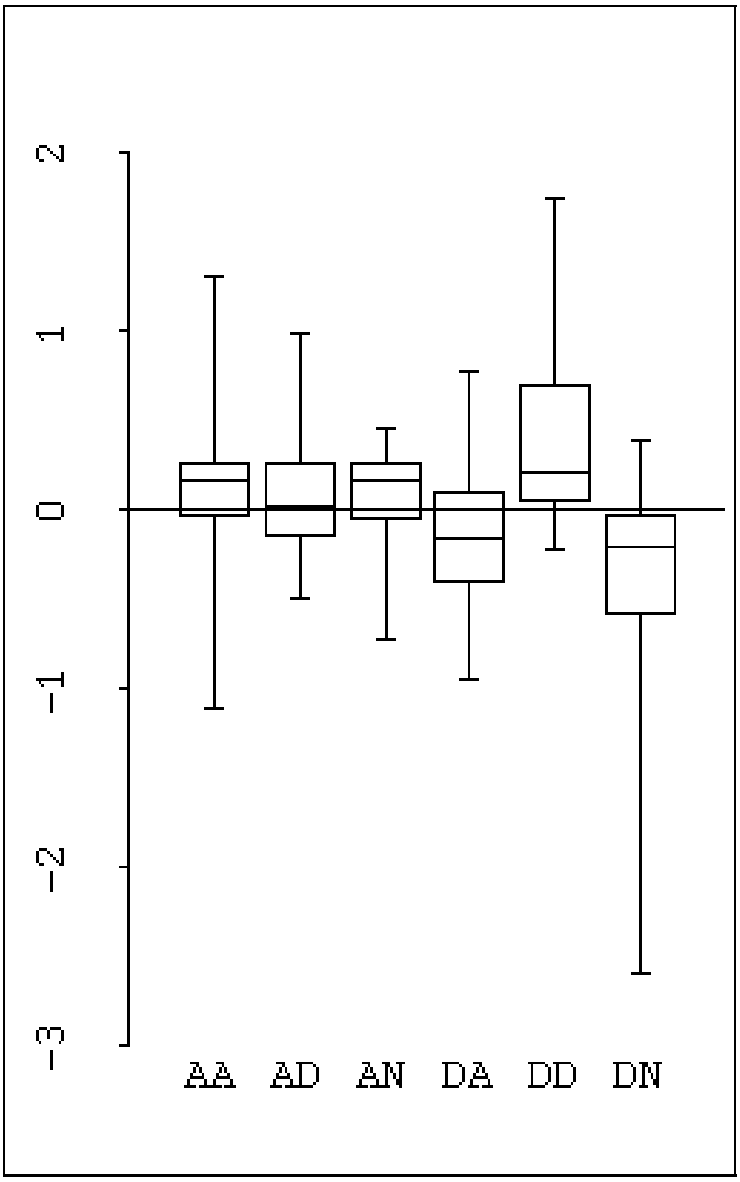

Figure 1.

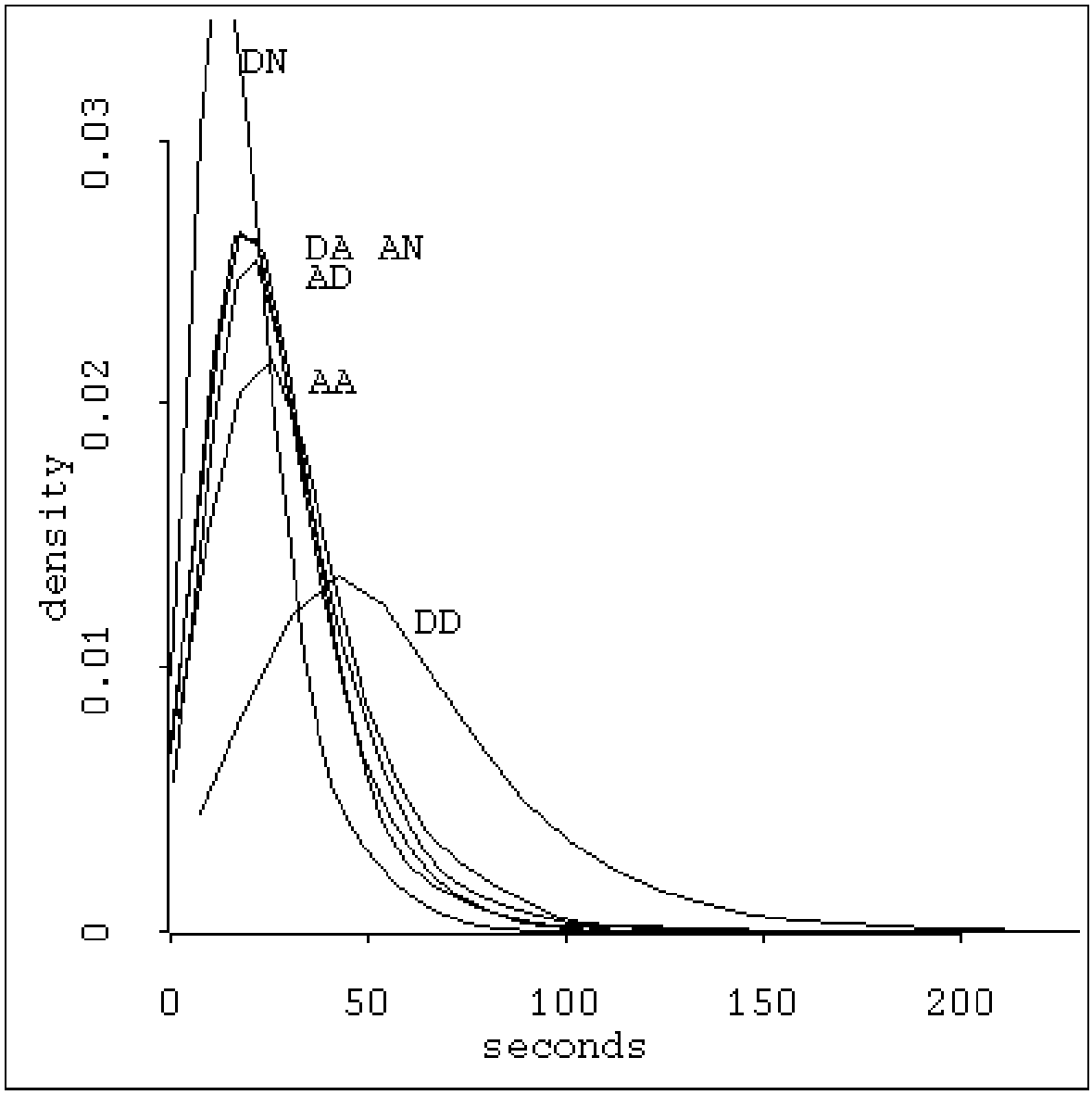

Figure 2.