Studies of Quality Monitor Time Series: The V.A. Hospital System

Mike West and Omar Aguilar
Institute of Statistics & Decision Sciences
Duke University, Durham NC 27708-0251
http://www.stat.duke.edu/

June 9, 1997
Revised: April 27, 1998

This report describes statistical research and development work on hospital quality monitor data sets from the nationwide VA hospital system. The project covers statistical analysis, exploration and modelling of data from several quality monitors, with the primary goals of: (a) understanding patterns of variability over time in hospital-level and monitor area specific quality monitor measures, and (b) understanding patterns of dependencies between sets of monitors. We present discussion of basic perspectives on data structure and preliminary data exploration for three monitors, followed by developments of several classes of formal models. We identify classes of hierarchical random effects time series models to be of relevance in modelling single or multiple monitor time series. We summarise basic model features and results of analyses of the three monitor data sets, in both single and multiple monitor frameworks, and present a variety of summary inferences in graphical displays. Our discussion includes summary conclusions related to the two key goals, discussions of questions of comparisons across hospitals, and some recommendations about further potential substantive and statistical investigations.

This work was performed in consultation and collaboration with the VA Management Science Group, Bedford MA. The models presented build on foundational contributions in prior and continuing work by Jim Burgess, Cindy Christiansen, Carl Morris and Ted Stefos.

Contents

1 Introduction

This report describes and summarises statistical modelling developments and analyses of subsets of hospital-level data from the medical center performance monitoring system of the US Department of Veterans Affairs (VA). As discussed in Burgess et al (1996, sections 3 and 4), the performance monitoring system collects, reports and analyses data from across the system of over 170 hospitals, and has as one key focus the general issue of assessing and comparing clinical and health care process performance between facilities. The specific quality monitors of interest here are based on a uniform process of data collection and consolidation across national databases, and encompass a range of inpatient, outpatient and long term care activities at each of the VA medical centers. For each hospital providing care in the area covered by a specific monitor, recorded data include the total numbers of individuals who were exposed to a specific and well-defined outcome in that area, and for how many such individuals that defined outcome occurred. These data are recorded annually. The record also carries an expected, or predicted, number of such outcomes out of the total, based on an assumedly exogenous prediction of the outcome proportion referred to as the *DRG predictor*. This quantity is designed to provide some degree of correction for hospital/monitor specific case-mix and characteristics of the patient population profile.

The work presented here was commissioned by, and performed in consultation with, the Management Science Group of the VA at Bedford MA. The primary statistical objectives are those of adequately defining and accurately estimating underlying measures of hospital-level performance in the monitor-specific areas. Given definitions and estimates, inferences about relative performance over time and across hospitals are in principle addressable. Our involvement in this study follows some long-term work by CN Morris and CL Christiansen (hereafter M&C) who have provided a firm basis for addressing these key questions of definition and appropriate modelling frameworks for estimation of performance measures (Burgess et al 1996). Some brief discussion of this past work is given below, before developing in new statistical directions driven by the motivating and guiding goals of our project. These goals are to explore, model and summarise the

- patterns of variability over time, in periods of single years, in hospital-monitor and area-specific performance measures across a selection of quality monitors, and the
- patterns of dependencies between sets of monitors, in addition to and in combination with assessment of time-variations.

These goals are motivated by policy interests in accurately estimating measures of hospital-level performance in key areas of health care provision, in assessing changes over time in such measures to monitor impact of internal policy changes (or the lack thereof), and ultimately in connection with the development of management and economic incentives designed to encourage and promote care provision at sustained and acceptable levels.

Our data analytic work focuses exclusively on three quality monitors, coded M20, M21 and M22. The outcomes/responses in these monitor areas represent annual numbers of individuals under a binary classification in an area of basic medical or psychiatric health care. Monitor M20 is concerned with *General Psychiatric Discharge* from the hospital; out of the total number of annual discharges, the response recorded is the number of individuals who failed to return for an out-patient visit (appropriately well-defined) within 30 days of discharge. Monitor M21 measures similar outcomes for *Substance Absuse Psychiatric Discharges,* and M22 measures those for *Basic Medical Discharges.* Low return rates are indicative of low "quality" in these specific care areas. The VA data provides information on years 1988 to 1995 inclusive, summaries of which are displayed in the group of figures displayed Section 8.1. Figure 1 displays the raw data in three rows of three frames; in each row, three graphs display aspects of data on the three monitors separately, but combined over all eight years (1988-1995). The graphs plot the observed proportions of successes in each monitor against the total numbers of patients in each case, and then against the DRG-based predicted proportions. Figures 2 and 3 provide related graphs for each year of data separately. There are between 152 and 159 observations in each monitor:year pair, corresponding to the (almost complete) subset of hospitals in the system reporting on each monitor in each year.

We present a chronological discussion of the data exploration and modelling developments, beginning in Section 2 with discussion of the data and our initial perspectives. Section 3 introduces basic random effects hierarchical models and explores summary inferences on the three monitors separately for each year of data; these models are close in spirit and form to those explored by M&C previously. Section 4 then introduces time series concepts to address the notion of systematic patterns of variation over the years in hospital-specific effects on a given monitor. We discuss a class of time series random effects models that necessarily include additional components of unpredictable variability in outcome probabilities as well as time series components. This is developed in the single monitor context, and summary inferences from our analyses of the three monitors separately are discussed and compared to the initial studies in Section 3. Section 5 ties three single monitor models together in a general multiple monitor model: this is a binomial/logit model in which hospital-specific random effects are related through time via a multivariate time series model representing the relationships across monitors. The univariate submodels are precisely those already developed in Section 4; these multivariate models explore and assess aspects of the dependencies among the three monitor series. Section 6 describes some summary inferences for all hospitals on one monitor in one year to illustrate additional possible uses of the models and further aspects of model fit. Section 7 provides some overall summary comments and ideas and recommendations for further work. Section 8 is a catalogue of all graphical summaries of various analyses discussed in the text, and Section 9 provides both a technical appendix and comments on methodology and computation.

2 Data Structure and Initial Perspectives

Our study began with initial examination of the data sets on the monitors separately, guided by the previous work of M&C. These authors have developed a variety of Bayesian hierarchical models for the observed outcomes, including regressions on the DRG predictor and hospital-specific parameters drawn from a hospital population prior (Burgess et al 1996, Christiansen and Morris 1997). These are customised versions of standard random effects generalised linear models with Bayesian interpretations and motivations. We explore some such models for the three Monitors M20, M21 and M22 of interest here, currently restricting attention to 152 of the hospitals that have complete records of outcomes on each of the monitors in each of the eight years.

First, some basic notation and definitions.

Consider data on a single monitor collected in one year across hospitals $i = 1, \ldots, I$. The canonical model for the observed number of responses z_i out of the total n_i in hospital i is that of binomial counts,

$$(z_i | n_i, p_i) \sim Bin(z_i | n_i, p_i) \tag{1}$$

conditionally independently across hospitals. The "success probabilities" p_i are hospital-specific and to be estimated, and the totals n_i are assumed uninformative about p_i. The proportions z_i/n_i are graphed in Figure 1; in each frame we have $8 \times 152 = 1216$ observations for the 8 years of data on each of $I = 152$ hospitals. Super-imposed on the first row of graphs are approximate 99% intervals under the binomial distribution (as n_i varies along the x-axis) assuming $p_i = p$ is fixed at the overall average proportion. Many observations lie outside these bands indicating the very high degree of over-dispersion relative to a single binomial model. This variation is to be explained by models that describe how the individual p_i vary across hospitals (and, later, across years), using a combination of regression on the DRG predictor and random effects. For hospital i, the DRG-based predicted proportion of "successes" d_i is supposed to predict p_i on the basis of system-wide studies of patient case-mix profiles and historical data. Adopting the standard logistic regression framework involves modelling $\mu_i = \log(p_i/(1-p_i))$ as a linear regression on some function of d_i. A natural choice is the mean-corrected logit predictor $x_i = l_i - \bar{l}$ where $l_i = \log(d_i/(1-d_i))$ for $i = 1, \ldots, I$; this is assumed here. Write

$$\mu_i = \alpha_i + \beta_1 x_i \tag{2}$$

for $i = 1, \ldots, I$. This gives a basic linear regression of slope β_1 on the DRG-based predictor and allows for hospital-specific departures through the parameters α_i. As it stands, this is a general model in which the α_i terms stand proxy for *all* sources of extra-binomial variation not adequately captured in the DRG predictor on the scale assumed. Across hospitals, differences and variations in success rates due to differences in patient profiles not adequately mirrored in the DRG predictor are represented in the α_i, and these are necessarily confounded with differences due to policies and practices in the area of care. Without additional covariate information, there is simply no way of "unconfounding" these issues. Assuming the α_i to be exchangeable (random effects) parameters drawn from a hospital-population prior delivers a class of Bayesian hierarchical models. This is the basic modelling framework; all models discussed below are developed as either exact or approximate special cases or generalisations. Some of these include models of M&C, as follows:

- *Normal approximations.* In cases where the n_i are reasonably large and the p_i are not close to zero or one, the binomial likelihood function $p(z_i|n_i, \mu_i)$ can be adequately approximated as a function of μ_i by a function proportional to $\exp(-(y_i - \mu_i)^2/(2s_i))$ where $y_i = \log(z_i/(n - z_i))$ and $s_i = \hat{p}_i(1 - \hat{p}_i)/n_i$ with $\hat{p}_i = z_i/n_i$. This leads to the normal hierarchical model in Burgess et al (1996). Our study has also used this normal approximation in initial data exploration and, more formally, in connection with technical methods of posterior simulation in binomial models. Our models and all analyses are, however, based on the exact conditional binomial sampling distributions.

- *Poisson approximations.* In cases where n_i is large and p_i is small, appeal to Poisson approximations to the binomial sampling model leads M&C to the class of PRIMM regression models in which z_i is conditionally Poisson with mean $n\lambda_i$ (Christiansen and Morris 1997). In such cases the logistic regression is approximated by a log-linear regression with $\lambda_i = x_i^{\beta_1}\eta_i$ where $\eta_i = \exp(\alpha_i)$. The PRIMM models of M&C adopt gamma distributions for the hospital population priors of the random effects η_i. We note that the low p_i/large n_i condition is clearly valid for some monitors. For Monitors M20, M21 and M22, however, it is evidently not. These monitors have outcome proportions across the full range though concentrated in 0.15-0.85, and widely varying total sample sizes, hence the Poisson approximation is not applicable in these cases.

Our starting models adopt the binomial structure in (1) and are based on normal priors for the population distribution of the random effects, developed in various stages. We begin with the basic normal model

$$\alpha_i \sim N(\alpha_i|\beta_0, w^2), \tag{3}$$

for some mean β_0 and standard deviation w. In terms of the zero-mean random effects

$$\epsilon_i = \alpha_i - \beta_0 \tag{4}$$

we have the equivalent model

$$\mu_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{and} \quad \epsilon_i \sim N(\epsilon_i|0, w^2). \tag{5}$$

The ϵ_i terms represent hospital-specific departures from the system-wide underlying level β_0. We prefer the "hierarchically centred" version (3) with β_0 at the center of the random effects prior for α_i for mainly technical reasons, and posterior inferences for both the absolute random effects α_i and the relative effects ϵ_i are of interest. The level β_0 represents the hospital system-wide average in corrected responses on this logit scale. Management policies that operate across all hospitals, and general improvements (or otherwise) in care provision that impact all hospitals in similar ways are represented by changes in β_0 from year to year. Hospital-specific changes over and above this system-wide measure are then reflected in the ϵ_i. Note that the values $\beta_1 = 1$ and $\beta_0 = 0$ represent a sub-model in which $E(\mu_i) = x_i$; if, in addition, w is small, then the DRG predictions are accurate, $p_i \propto d_i$ in such cases. The key issue of unexplained variability in success probabilities relates to the prior variance w^2, large values of which represent high levels of heterogeneity in outcomes that are not captured via the DRG predictor.

Some summary of key issues mentioned above and some additional comments on the data are pertinent and to be borne in mind throughout all analyses.

- As noted above there is a very marked degree of extra-binomial variation evident in the data on all three monitors. This is true for each of the eight years under study; see Figures 2 and 3 as well as the combined graphs in Figure 1.

- There is an overall suggestion of decreasing levels of observed responses across the eight years and in all three monitors, most marked in Monitor M20 and, to a lesser extent Monitor M21. We see this most clearly in Figure 4 which graphs the simple average responses across years. The average DRG values (not shown) do not show similar decreasing patterns indicating that this is very likely a hospital system-wide feature, perhaps due to VA policy and/or general improvements in care provision over the years. Thus there is an apparent need to consider models in which the overall levels β_0 of outcome responses across all hospitals vary year to year, and in which the regression effects of the DRG variable (however weakly identified they may be) may also vary.

- Across all three monitors and all eight years the values of the DRG predictor variable lie in a very narrow range on the probability scale. This is evident in the third row of graphs in Figure 1. Though there is an apparent positive association between the DRG variable and the observed outcomes, evident in the second row of graphs, it is clear that the lack of variation will lead to a high degree of uncertainty about the regression on DRG in any model. The predictor simply does not vary substantially across the "design" space. For Monitor M21, for example, the DRG variable is very close to 0.4 across all hospitals in all years. We include DRG in all models, but bear in mind that resulting likelihood functions for corresponding regression coefficients will tend to be flat due to this low variability issue.

3 MODEL-1: Exploratory Data Analysis and Modelling

Our analyses began with studies of data from each of the monitors separately, also regarding each year of data on each monitor as a separate data set to be analysed individually. These analyses represent initial studies to investigate the validity of the basic binomial regression model structure.

We explored analyses of the model comprised of equations (1) and (5) under standard diffuse reference priors for the additional parameters (β_0, β_1, w). Reanalyses under ranges of diffuse but proper priors for these three parameters were performed to assess sensitivity. Across several choices of seemingly uninformative though proper priors, the posterior results confirm those from the reference prior based analyses. Hence we are satisfied with reporting the reference analysis. Model adequacy has been explored and verified through post-hoc residual analyses, examining plots and features of posterior samples of well-defined residuals. This is not illustrated here, but is in the more elaborate final models below. Additional ranges of analyses were performed using the normal approximation to the binomial likelihood functions. In general, very similar posterior results for (β_0, β_1, w) were delivered, indicating that the normal approximations are indeed very good for inferring these parameters and with these data sets. However, a small number of hospital/monitor pairs have very low sample sizes n_i in some years, and there are some in which the observed outcome proportions are close to zero or one. In such cases, the posteriors for the corresponding α_i parameters cannot be expected to be necessarily reliable approximations, and any errors in approximation must be expected to impact on inferences about other parameters. Hence we prefer to remain in the theoretically sound binomial framework rather than to adopt the (technically and computationally more manageable) normal approximations directly. However, we note that for much of the data the normal model represents a sound approximation.

We refer to this basic single-year, single-monitor model as MODEL-1. Figures displayed in a group in Section 8.2 display a selection of posterior summaries from these individual analyses. We comment on these graphs and make summary initial conclusions for each of the three monitors over the eight years. Most of the graphs display box-plots of posterior distributions for selected model parameters. The boxes are centred at posterior medians, have boxes drawn out to posterior upper and lower quartiles, and lines from the quartiles to notches at points no further than 1.5 times the interquartile range from the quartiles. Points are plotted outside these final notches to represent the very extreme tails of the posteriors–most of these can be simply ignored, as almost all the posterior probability lies between the notches (well over 99% under a normal posterior).

Hospital population levels

Our main comments relate to inferences on the β_0 parameters. At a mean DRG variate $x_i = 0$, parameter β_0 represents the average response probability, on the logit scale, across the hospital system. Differences across years for a single monitor require explanations, possibly in terms of changes in system-wide policies and practices, or in terms of otherwise uncontrolled "drifts" in levels of quality. We refer to the upper graph in each of Figures 5, 8 and 11.

There are meaningful differences in the β_0 parameters across the eight years in each of the three monitors. The main feature is a general decreasing trend in β_0 over the years for all three monitors, more markedly so for Monitors M20 and M21. This corresponds to generally decreased probabilities of return for out-patient visits within 30 days of discharge, and the apparent similarities between Monitors M20 and M21 are consistent with the two being related areas of care, each related to psychiatric discharges. Monitor M22, General Medical Discharges, exhibits a quite abrupt increase in β_0 in 1995, after decreasing and levelling off in 1993-4; this requires further consideration and interpretation from VA personnel.

DRG regression effect

Now consider the β_1 regression parameters, referring to the second graph in each of Figures 5, 8 and 11. Again there are apparent differences over the years within each monitor, although the values are relatively stable compared with β_0 and have a lesser impact on overall conclusions.

Dispersion in hospital population of random effects

Now consider the standard deviation w that determines the dispersion in the random effects distribution on the population of hospitals. We refer to the lower graph in each of Figures 5, 8 and 11. Within each monitor, w appears essentially constant over the years and the range of values indicated supports appreciable variation consistent with the high degree of extra binomial variation apparent in the data. This variability is highest for Monitor M20, at s.d.

levels of $0.6 - 0.7$ on this logit scale, then Monitor M21, at levels around $0.4 - 0.5$, and with levels around $0.3 - 0.4$ for Monitor M22. As a benchmark for interpretation, consider a case in which the DRG predicted proportion is 50% (so zero on the logit scale) and assume $\beta_0 = 0$ so that baseline population probabilities are around 0.5. A standard deviation of $w = 0.5$ leads to an 95% interval for the outcome probability of roughly $0.27 - 0.73$. Thus, even ignoring the additional binomial variation about the outcome probability, the random effects distribution covers much of the observed range of the data.

Random effects for three example hospitals

The remaining graphs display inferences on hospital-specific random effects for three arbitrarily selected hospitals, those with station numbers 515, 552 and 592. In each year we graph the posterior boxplots for the corresponding hospital-specific levels α_i and, separately, those for the zero-mean relative levels ϵ_i. These graphs are given in Figures 6 and 7 for Monitor M20, Figures 9 and 10 for Monitor M20, and Figures 12 and 13 for Monitor M22. These provide examples of the kinds of patterns of variations exhibited by the random effects within individual hospitals. One notable feature is that of evident dependence over the years in the ϵ_i within specific hospitals, especially in terms of sustained signs. Though the series of length eight are very short, this is supportive of systematic dependence structure over time that is naturally expected: a hospital that has tended to be below the population norm in terms of its proportions of outcomes in recent years will be expected to maintain its below average position this year, so that the ϵ parameters of this hospital will tend to be of the same sign. We are therefore clearly interested in some form of time series structure to describe and incorporate such dependencies, and in order to explore and assess their effects on estimation of all model parameters.

The general conclusions made above hold up in further analyses with more structured models for exploration of dependencies in quality monitors across monitor areas and across years, as we shall see. There are obviously differences in the summary inferences as we change the models, but the differences in important characteristics remain very small.

4 MODEL-2: Models for Single Monitor Time Series

The next stage of investigation concerns the structure of single monitor series over time. The initial modelling efforts just described indicate apparent variability over time in the hospital/monitor population parameters β_0 and β_1, as well as the random effects α_i that explain much of the observed extra-binomial variability. Nevertheless there are indications of relative stability of the ϵ_i across years in some hospitals. This is consistent with expectation. Unless policies and protocols in the monitor care areas are radically changed from one year to the next, there should be stability in these quantities as representing true quality levels; any changes beyond this reflect random variations due to the characteristics of the patient sample presenting at the hospital. We move to simple time series models to incorporate the view that the ϵ_i are expected to remain stable within each hospital over such a short number of years, but that unexplained sources of variability at the hospital level are expected to induce random changes year to year.

We stress that our developments do not currently constrain the hospital/monitor population parameters β_0 and β_1. We introduce one pair of these parameters for each year separately, and explore analyses that do not impose any further structure, simply allowing for the estimation of these parameters. For example, the apparent decrease over time in the overall levels of response on Monitor M20 will then be largely accounted for by estimating separate β_0 parameters for each of the eight years, but it is not anticipated in the model. This model does not therefore provide us with a structure to explore or model dependence over time in the β parameters, and so is not useful for prediction to future years without modification or intervention. This is not a criticism; our goals are to evaluate patterns over time in the eight years of data, and to explore inferences about changes in hospital-specific effects over the years. We have no current interest in, nor remit for, the development of predictive models for the β parameters; such would require the evaluation of expert opinion about the reasons behind any inferred time evolution and the use of this in phrasing appropriate predictive models. Such developments are technically feasible and could be incorporated into the study at a further stage, as necessary.

Maintaining the focus on a single monitor, we now write α_{it} for the logistic random effects parameter of hospital i in year t, with $t = 1, \ldots, 8$ representing our years 1988 to 1995 inclusive. The corresponding relative effects are now $\epsilon_{it} = \alpha_{it} - \beta_{0t}$ where β_{0t} and β_{1t} are the population parameters of the logistic regression in each year $t = 1, \ldots, 8$. Our studies use autoregressions, specifically AR(1) models for the series α_{it} over time t, considered conditionally independently over hospitals. This simple time series model is the first logical linear departure from the random sampling models of the initial study. With such a short time span (8 years) more complex models are largely untenable. In any case, the AR(1) model is a natural, interpretable model that describes this year's α and "close to" last year's but with a degree of "noise" added. It also turns out to be adequate as a model of dependence through time for the three monitors under study, and, as mentioned below, has a desirable consequence in that the annual marginal distributions of the hospital-specific effects are the same across the short span of eight years.

Modelling any form of time series dependence over the years in the α_{it} quantities introduces partial stochastic constraints so that the α_{it} are no longer as free to vary as in the single year models of Section 3 where they are essentially unconstrained, viewed simply as independent across years. Hence it is necessary to consider that some of the evident variation in the logit parameters μ_{it} will be unexplained once we extend to this time series structure. We need to account for this, and do so by including in the linear model for the logit parameters additional terms representing residual, unexplained variation.

Specifically, our extension of the single monitor model in equations (1) to (5) is as follows. Independently across hospitals $i = 1, \ldots, I$, and over all years $t = 1, \ldots, 8$, the data are assumed to arise from the set of $8I$ binomial models

$$(z_{it}|n_{it}, p_{it}) \sim Bin(z_{it}|n_{it}, p_{it}) \qquad (6)$$

with logistic regression on the DRG predictors and random effects,

$$\mu_{it} \equiv \log(p_{it}/(1 - p_{it})) = \alpha_{it} + \beta_{1t}x_{it} + \nu_{it}. \qquad (7)$$

This regression is similar to equation (2) but now the subscripting makes explicit the year and hospital-specific parameters. In addition, we have introduced residuals

$$\nu_{it} \sim N(\nu_{it}|0, v^2) \qquad (8)$$

to model residual variation not explained by the regression and hospital-specific random effects α_{it}. The time series components assume that, again independently across hospitals i, the α_{it} follow first-order autoregressions centred around the current population level β_{0t} in year t. For $t > 1$,

$$\alpha_{it} = \beta_{0t} + \phi(\alpha_{i,t-1} - \beta_{0,t-1}) + \omega_{it}, \qquad (9)$$

where $\omega_{it} \sim N(\omega_{it}|0, u^2)$ independently over hospitals i and years t. At $t = 1$,

$$\alpha_{i1} \sim N(\alpha_{i1}|\beta_{01}, w^2) \qquad (10)$$

Here ϕ is the autoregressive parameter and will generally be close to one, lying in part of stationary region $0 < \phi < 1$, and we have the relationship $w^2 = u^2/(1 - \phi^2)$. Equivalently, we could work in terms of the annual deviations of the hospital-specific effects from the annual levels, $\epsilon_{it} = \alpha_{it} - \beta_{0t}$, related over time by

$$\epsilon_{it} = \phi\epsilon_{i,t-1} + \omega_{it}. \qquad (11)$$

Notice that we assume constant values of ϕ and u in the time series components. This is supported on the basis of exploratory and confirmatory analyses, though could be relaxed to allow for differing variances across hospitals and/or years as may be desirable for other applications. This model class has some important characteristics, as follows.

- The hospital-specific relative effects ϵ_{it} are now related over time within each hospital according to the AR model. With an appropriately large value of ϕ, this implies high positive correlations between the ϵ_{it} in a given hospital over the years. This is consistent with the view that, for example, a hospital that is been generally "good" in a specific monitor/care in one year will have a high probability of remaining "good" the next year, and vice versa. This is true irrespective of the system-wide changes in quality levels; they are captured in changes in the β_{0t} sequence, and the ϵ_{it} represent departures above or below the system levels β_{0t}.

- The constancy of ϕ and u over years is sufficient (though not necessary) to induce (desirable) stationarity into the model for the relative effects ϵ_{it}. Thus, for example, considering any one chosen year t we have the implied marginal distribution

$$\epsilon_{it} \sim N(\epsilon_{it}|0, w^2) \tag{12}$$

  with $w^2 = u^2/(1-\phi^2)$, independently over i. Equivalently, $\alpha_{it} \sim N(\alpha_{it}|\beta_{0t}, w^2)$ for each year t. So, in each year, the relative random effects are a random sample from a zero-mean normal distribution. This is consistent with a view of no global changes in the hospital population makeup, variability in expected levels being essentially constant over the short period of years once the DRG predictor and any system-wide changes are accounted for through β_{1t} and β_{0t}, respectively. Changes in relative performance of hospitals can therefore be assessed across years.

- Residual variation in the logit probabilities not explained by either the DRG-based regression or the correlated random effects is contributed by the ν_{it} in equation (8). Estimation of the residual variance parameter v^2 together with all other parameters will provide indications of the extent of this "unexplained" variability. With respect to the exploratory, single year model with no time series structure in Section 3, we are basically "splitting" the original α parameters of equation (2) into two components: the new α_{it}, still hospital-specific random effects but now structurally related over time, and the purely residual and unpredictable components ν_{it}. The model degenerates to the earlier non-time series model if $\phi = 0$, in which case the α_{it} and ν_{it} terms are not distinguished. Our analyses to follow indicate relatively high time series dependencies with appropriate values of ϕ that are positive and significantly large. As a result, the "systematic" (though random) variation explained by the α_{it} is isolated and identified separately from the unexplained variation in the ν_{it}.

To summarise, this MODEL-2 framework with logit probabilities

$$\mu_{it} \equiv \log(p_{it}/(1-p_{it})) = \beta_{0t} + \beta_{1t}x_{it} + \epsilon_{it} + \nu_{it} \tag{13}$$

is a class of single monitor, hierarchical random effects time series models and has the following components of variation underlying the observed quality monitor outcomes:

1. Sampling variability from the data-level binomial distributions (1);
2. Unconstrained year-to-year variation in the population level of logit-probabilities, β_{0t};
3. Unconstrained year-to-year variation in the regression coefficient on the DRG-based predictor, β_{1t};
4. Structurally related hospital-specific random effects parameters ϵ_{it} that represent hospital/year departures from the DRG-corrected population level of probability of response in the monitor care area;
5. Residual, unexplained components ν_{it} in the same index.

One final feature to note, not primary though of some interest, concerns the time series structure of the combined hospital-specific random effects $\epsilon_{it} + \nu_{it}$ above. This combined term represents a time series whose structure is autoregressive, moving-average of order one, ARMA(1,1). The addition of the residual/noise terms ν_{it} to the AR(1) process ϵ_{it} acts to modify the correlation structure. Of course if v is small compared to w the modification is small.

Summaries of analysis of each of the three monitors separately appear in the figures in Section 8.3, in formats similar to those from the preliminary analysis under `MODEL-1`. The following comments summarise the key and relevant features of these graphs.

Hospital population parameters

Summaries of the graphs in Figures 14, 17 and 20 are as follows for the hospital population parameters β_{0t} and β_{1t} in each year t, and for the constant hyperparameters w, v and ϕ, one set for each monitor analysis.

- Across years, inferences for the β_{0t} and β_{1t} quantities are essentially the same as under `MODEL-1`, as is to be expected as we are not constraining these quantities; the displays of yearly boxplots for these quantities in each of the three single monitor analysis appear in the upper two rows of Figures 14, 17 and 20.
- For each of the monitor analyses, boxplots of posteriors for the monitor-specific w appear in the lower row of Figures 14, 17 and 20. The supported ranges are lower that those of the `MODEL-1` analyses in each case. This is to be expected if, as we shall confirm is the case below, the variances v^2 of the "unpredictable" components of variation in the random effects model are non-negligible.

- For the monitor-specific autoregressive parameters ϕ, posterior boxplots appear in the lower row of Figures 14, 17 and 20. These indicate highly significant dependence structures in each case, with inferred values of ϕ in the ranges $0.7 - 0.8$ for M20, $0.6 - 0.75$ for M21 and $0.8 - 0.9$ for M22. The posteriors have some overlap though do suggest that the ϕ parameters may be different across monitors. The dependence in the random effects time series is high in each case, but perhaps higher for M22 than M20 or M21, the latter two being more comparable. There are thus apparent differences here between M22 and the other two monitors, perhaps associated with the fact that latter two are more closely related health care areas, both involving psychiatric discharges. These ϕ values lead to smoothing in the estimation of the individual random effects series for each hospital within each analysis, noted below.

- For the monitor-specific standard deviation v of the unpredictable/residual components of variation, posterior boxplots appear in the lower row of Figures 14, 17 and 20. These indicate non-negligible values, in the ranges of $0.3-0.37$ for M20, $0.27-0.33$ for M21 and $0.12-0.16$ for M22. By comparison with the posteriors for the w parameters this indicates that the unpredictable/residual components of the model contribute significantly to the variability in outcome probabilities. In terms of the variance ratio $v^2/(v^2+w^2)$, the ϵ_{it} residuals contribute, very roughly, about $20 - 25\%$ variation for M20, about 30-35% for M21, but only about 15% for M22. Again we see quantitative differences between M22 and the other two monitors, perhaps related to the underlying nature of the differences in health care area.

Random effects for three example hospitals

Summary inferences for the random effects α_{it} and ϵ_{it} for the three selected hospitals are again given in terms of posterior boxplots for each year in each of the single monitor analyses. These appear in Figures 15 and 16 for M20, Figures 18 and 19 for M21, and Figures 21 and 22 for M22.

The most immediate features relate to the apparent smoothing over the years in the patterns of the relative random effects ϵ_{it} within each monitor. Relative to the corresponding figures from the exploratory analyses under `MODEL-1`, consecutive values of the ϵ_{it} within each hospital are linked by the autoregressive model and, as fairly high positive values of the autoregressive parameters are inferred, this leads to a reasonable degree of shrinkage as each ϵ_{it} is now estimated partly by the data in neighbouring years as well as by that in year t. The resulting smoothing is also evident in the patterns of the absolute levels α_{it}.

Residual structure analysis

Additional graphs in Figures 23 to 26 inclusive are illustrative of graphical assessments of model adequacy via more-or-less standard Bayesian “residual analysis”. Our primary concern is with identifying and investigating any apparent residual structure in the data following model fitting. Referring to the approximate logistic normal version of the binomial model in Section 2, note that the model implies approximate (standard) normality and conditional independence of the standardised data residuals $e_{it} \equiv (y_{it} - \mu_{it})/s_{it}$ where, extending the subscripts to the current model, y_{it} is the logit transform of the observed outcome proportion z_{it}/n_{it} and s_{it} the corresponding approximate standard deviation, for each i, t. In our simulation-based analysis of the exact binomial model here, we repeatedly simulate the full joint posterior distribution for all model parameters and random effects. This means that we also obtain a sequence of samples of the set of μ_{it}, which represent draws from their posterior distribution. These lead to trivially computed values of the e_{it} which similarly represent samples from the posterior distribution of the standardised data residuals. Using any one of these samples of residuals we can compute numerical and graphical measures of concordance with, or departures from, the theoretical normal distribution, and so assess model fit. The graphs in Figures 23 and 23 display some such graphs using just two sets of sampled residuals on monitor M21 in year $t = 1995$, the final year of the data. These are representative of all monitor/year residuals and repeated sampling from the posteriors produces very similar graphs. For these samples, the plots include a simple graph against hospital number, a graph against the corresponding sample sizes n_{it}, a normal quantile plot and a simple histogram. None of these graphs indicate any kind of meaningful departures from normality, and as this is maintained across residual samples, it provides support for model adequacy.

More formally, Figures 25 and 26 summarise the posterior distributions of the actual data residuals e_{it} for monitor M21 in year $t = 1995$. Figure 25 displays graphical summaries of this posterior distribution in terms of marginal posterior 95% intervals for each hospital, with posterior means marked. The hospitals are ordered here according to the posterior medians of the underlying outcome probabilities p_{it} on this monitor in this year. There are no outlying hospitals or other worrisome features evident in this display. This comforting conclusion is supported by additional displays of aspects of the posterior distribution of the *ordered* observation residuals. Figure 26 displays a normal quantile plot and a histogram of the posterior means of the *ordered* observation residuals for all hospitals i. The curve superimposed on the histogram is the standard normal density function. The quantile plot includes vertical lines representing approximate posterior 95% intervals showing the uncertainty in the marginal posteriors for the ordered residuals. The conformity with normality here is excellent.

5 MODEL-3: Models for Multiple Monitor Time Series

The MODEL-2 class appears to provide satisfactory description of the dependence structure over time in hospital-specific quality measures. The second key goal of this study is to explore possible dependencies in quality monitor outcomes across monitor areas. Hence we are interested in possible dependencies among the hospital-specific random – both the ϵ and ν terms – across the three monitors. We know that M20 and M21 relate to two areas of psychiatric care, so might anticipate that the quality levels are positively correlated between these two. M22 relates to general medical care and so also might be expected to be positively related as representing some form of "overall" quality at the specific hospital. The next step therefore is to tie the three sets of monitor data together into a model that allows for the investigation of the structure of the joint distribution of sets of relative effects across the three monitors, and of the nature of dependencies in changes over time in these effects. To begin we need to extend the notation to make explicit the consideration of several monitors simultaneously.

This new class of models, referred to as MODEL-3, considers monitors $j = 1, 2, 3$ in each year $t = 1, \ldots, 8$ and for each hospital $i = 1, \ldots, I$. This development is obviously general and can be applied to any numbers of monitors and years. On the three monitors, we have observed outcomes $\mathbf{z}_{it} = (z_{i1t}, z_{i2t}, z_{i3t})'$, representing three conditionally independent binomial responses out of totals $\mathbf{n}_{it} = (n_{i1t}, n_{i2t}, n_{i3t})'$ and with "success" probabilities $\mathbf{p}_{it} = (p_{i1t}, p_{i2t}, p_{i3t})'$, respectively. The joint density of these data is

$$p(\mathbf{z}_{it}|\mathbf{n}_{it}, \mathbf{p}_{it}) = \prod_{j=1}^{3} Bin(z_{ijt}|n_{ijt}, p_{ijt}) \tag{14}$$

conditionally independently over hospitals i and years t.

The general multiple-monitor hierarchical/random effects model structure of MODEL-3 is as follows.

Regression and hierarchical/random effects structure

On the logit scale, $\boldsymbol{\mu}_{it} = (\mu_{i1t}, \mu_{i2t}, \mu_{i3t})'$ with

$$\mu_{ijt} \equiv \log(p_{ijt}/(1 - p_{ijt})) = \beta_{0jt} + \beta_{1jt}x_{ijt} + \epsilon_{ijt} + \nu_{ijt} \tag{15}$$

for each monitor $j = 1, 2, 3$. Here x_{ijt} is the logit transform of the corresponding DRG predicted proportion, with the mean for year t across all hospitals i within monitor j subtracted. The β parameters are collected in vectors as $\boldsymbol{\beta}_{0t} = (\beta_{01t}, \beta_{02t}, \beta_{03t})'$ and $\boldsymbol{\beta}_{1t} = (\beta_{11t}, \beta_{12t}, \beta_{13t})'$. With the 3×3 design matrices $\mathbf{X}_{it} = \text{diag}(x_{i1t}, x_{i2t}, x_{i3t})$, we then have

$$\boldsymbol{\mu}_{it} = \boldsymbol{\beta}_{0t} + \mathbf{X}_{it}\boldsymbol{\beta}_{1t} + \boldsymbol{\epsilon}_{it} + \boldsymbol{\nu}_{it} \tag{16}$$

where $\boldsymbol{\epsilon}_{it} = (\epsilon_{i1t}, \epsilon_{i2t}, \epsilon_{i3t})'$ and $\boldsymbol{\nu}_{it} = (\nu_{i1t}, \nu_{i2t}, \nu_{i3t})'$. In each year, marginally, the hospital-specific random effect vectors $\boldsymbol{\epsilon}_{it}$ are conditionally independent over hospitals i and

$$\boldsymbol{\epsilon}_{it} \sim N(\boldsymbol{\epsilon}_{it}|\mathbf{0}, \mathbf{W}) \tag{17}$$

for some constant 3×3 matrix $\mathbf{W}$. This matrix represents the within-year variability in the systematic components of corrected quality levels across the entire hospital population, and the within-year dependencies between such quality measures across the three monitors. Notice that essentially arbitrary correlations are admitted across monitors.

This model simply collects together the three single monitor models of the previous section, then extends that framework to allow for possible correlation structure across monitors via the multivariate normal distributions for the vectors of relative effects $\boldsymbol{\epsilon}_{it}$. As we now describe, $\mathbf{W}$ defines and measures the joint variability and dependence structure not only in each year, but also a modified form of the year to year *changes* in the effects.

Time series structure of random effects

The random effects vectors $\boldsymbol{\epsilon}_{it}$ are correlated over years t. We maintain the same basic AR(1) models for monitor-specific quantities, also indexing the autoregressive parameters by monitor label j. The three univariate AR(1) models are therefore combined in the vector autoregression

$$\boldsymbol{\epsilon}_{it} = \Phi \boldsymbol{\epsilon}_{i,t-1} + \boldsymbol{\omega}_{it} \tag{18}$$

where $\Phi = \text{diag}(\phi_1, \phi_2, \phi_3)$ is the diagonal matrix of monitor-specific autoregressive coefficients. The vector innovations $\boldsymbol{\omega}_{it}$ are conditionally independent over time, and normally distributed with

$$\boldsymbol{\omega}_{it} \sim N(\boldsymbol{\omega}_{it}|\mathbf{0}, \mathbf{U}) \tag{19}$$

for some variance matrix $\mathbf{U}$ to be estimated. Now the normal distribution $\boldsymbol{\epsilon}_{it} \sim N(\boldsymbol{\epsilon}_{it}|\mathbf{0}, \mathbf{W})$ is the yearly margin under this vector AR(1) model. This implies that $\mathbf{W}$ satisfies $\mathbf{W} = \Phi\mathbf{W}\Phi + \mathbf{U}$, so that correlation patterns in $\mathbf{U}$ and $\mathbf{W}$, while similar, also depend on the autoregressive parameters. In particular, for each monitor pair j, h we have variances and covariances $\mathbf{W}_{jh} = \mathbf{U}_{jh}/(1 - \phi_j \phi_h)$.

The DRG-corrected hospital-specific level parameters are now given by

$$\boldsymbol{\alpha}_{it} = \boldsymbol{\beta}_{0t} + \boldsymbol{\epsilon}_{it} \tag{20}$$

for each i, and t. These follow the centred VAR(1) model

$$\boldsymbol{\alpha}_{it} = \boldsymbol{\beta}_{0t} + \Phi(\boldsymbol{\alpha}_{i,t-1} - \boldsymbol{\beta}_{0,t-1}) + \boldsymbol{\omega}_{it} \tag{21}$$

for $t > 1$, and have yearly margins $N(\boldsymbol{\alpha}_{it}|\boldsymbol{\beta}_{0t}, \mathbf{W})$. Note again that the single monitor MODEL-2 structures are embedded here; all we have added is the cross-monitor structure through $\mathbf{W}$ (equivalently $\mathbf{U}$).

Residual components

Residual, unpredictable variations in the binomial probabilities across hospitals and years is described by the residual terms

$$\boldsymbol{\nu}_{it} \sim N(\boldsymbol{\nu}_{it}|\mathbf{0}, \mathbf{V}) \tag{22}$$

with monitor-specific variances v_1^2, v_2^2 and v_3^2 on the diagonal of the matrix $\mathbf{V}$, and now admitting cross-monitor dependencies through the covariance elements of $\mathbf{V}$. Again, we simply collect the monitor-specific residual parameters together in a vector for the multiple monitor model here, and add covariance structure that permits the estimation of cross-monitor dependencies. It might be expected that correlation patterns in $\mathbf{V}$ are similar to those in the time series random effects modelled through $\mathbf{W}$ (and/or $\mathbf{U}$). Our current model does not anticipate this, leaving $\mathbf{V}$ and $\mathbf{W}$ unrelated *a priorí*, but the framework obviously permits the assessment of potential similarities in posterior inferences.

It is important to note that the overall levels of random effects variability, and the associated overall measures of cross-monitor dependencies, are represented through the yearly marginal variance matrix $\mathbf{W} + \mathbf{V}$, i.e., the variance of the combined random effects $\boldsymbol{\epsilon}_{it} + \boldsymbol{\nu}_{it}$ in year t. In summaries of analyses below we explore posterior inferences on elements of $\mathbf{W} + \mathbf{V}$, as well as the component matrices separately.

Prior distributions

Inference is based on posterior distributions for all model parameters and random effects under essentially standard reference/uninformative priors for: (a) the annual population parameters $\boldsymbol{\beta}_{0t}$ and $\boldsymbol{\beta}_{1t}$, (b) the population residual variance matrix $\mathbf{V}$, and (c) the variance-covariance matrix $\mathbf{U}$; the prior is completed with independent uniform priors for the autoregressive parameters ϕ_j on (0,1).

Summaries of analysis of the three monitor series combined in this multivariate, multi-year model appear in the figures in Section 8.4. The formats of figures follow those of the earlier models, with additional inferences about the correlation patterns in $\mathbf{W}$ displayed as boxplots of posteriors as usual. We note the following features relative to the inferences deduced from the MODEL-2 analysis.

- There are only very minor changes evident in the posterior distributions for population parameters β_{0t} and β_{1t} across years t. See the upper rows of Figures 27, 30 and 33.

- There are similarly almost no changes evident in the posterior distributions for the variances v_j^2 of the residual noise terms ν_{ijt}. See the third row in each of Figures 27, 30 and 33; these posteriors are redisplayed in the second frame of Figure 36.

- There are similarly only very minor changes evident in the posterior distributions for random effects α_{ijt} and ϵ_{ijt} over the years for the three selected hospitals. See Figures 28 and 29 for M20, Figures 31 and 32 for M21, and Figures 34 and 35 for M22.

- Figure 36 provide summaries of the marginal posteriors for the three autoregressive parameters in Φ. These are essentially similar to those delivered in the individual MODEL-2 analyses, as might be expected.

- Figure 36 also provide summaries of the marginal posteriors for elements and functions of the variance matrices $\mathbf{W}$ and $\mathbf{V}$. First, note that the posteriors displayed for the standard deviations under both $\mathbf{W}$ and $\mathbf{V}$ are essentially as in the separate MODEL-2 analyses. The new parameters here are the correlations between the monitor effects, both in the systematic component $\mathbf{W}$ and in the residual component $\mathbf{V}$. Posterior boxplots are displayed here for the correlations in both $\mathbf{W}$ and $\mathbf{V}$, and, most importantly, in the combined variance $\mathbf{W}+\mathbf{V}$. As mentioned above, it is this latter matrix that most fully described cross-monitor structure within each year. The lower left frame of Figure 36 summarises inferences on the standard deviation and correlation elements of $\mathbf{W}+\mathbf{V}$. This indicates fairly low overall correlations, consistent with the generally low correlations exhibited by each of $\mathbf{W}$ and $\mathbf{V}$ separately. The posterior mean of $\mathbf{W}+\mathbf{V}$ is approximately

$$\mathbf{W}+\mathbf{V} = \begin{pmatrix} 0.417 & 0.044 & 0.009 \\ 0.044 & 0.271 & 0.014 \\ 0.009 & 0.014 & 0.136 \end{pmatrix}$$

which corresponds to estimated correlations as follows: between M20 and M21, 0.132; between M20 and M22, 0.036; and between M21 and M22, 0.073. These are quite low correlations, though they do support the earlier suggestion that the correlation between M20 and M21 might be higher than that between either M20 and M22, or between M21 and M22, in view of the similar care areas of M20 and M21. The posterior mean of $\mathbf{W}+\mathbf{V}$ decomposes into eigenvalues $0.43, 0.26$ and 0.13, and corresponding column eigenvector matrix

$$\begin{pmatrix} 0.961 & -0.275 & -0.015 \\ 0.273 & 0.957 & -0.097 \\ 0.041 & 0.089 & 0.995 \end{pmatrix}.$$

In order, the principal components explain roughly 52%, 32% and 16% of variation described by this estimated variance matrix, indicating that all three vary appreciably and that reduction to two or fewer seems inappropriate. The eigenvector matrix is dominated by the diagonal terms, and all three are close to unity; note that the eigenvector matrix would be the identity were the monitors uncorrelated. The first column represents an average monitor effect, dominated by M20 and M21 and essentially excluding M22; this can be viewed as a psychiatric care component alone. The second column represents a contrast between M20 and M21, again essentially excluding M22; the final column almost wholly represents M22 alone, and to the extent that the coefficients for M20 and M21 are non-ignorable, contrasts these two psychiatric care monitors with the general medical monitor M22.

Investigating conditional distributions provides additional insights into the levels and nature of dependencies. For example, we might consider the relevance of dependencies by looking at questions like *"how much might the M21 effect change if the M20 effects changes by e?"*. Various such exercises have been undertaken, and simply confirm that the levels of correlations inferred in our analyses are so low as to really limit their impact on predictive questions, especially in the context of the realistic levels of posterior uncertainties about the population parameters, the $\beta_{0t}, \epsilon_{ij,t-1}$ and so forth. Hence, this level of correlation structure between monitors, though certainly non-negligible, is relatively minor. We remark that the story may, however, be quite different with other collections of monitors.

6 Summary Inferences for Monitor M21 in 1995

To illustrate the possible additional uses of these models in exploring the variability in outcomes across the hospital system, we focus on outcomes on only Monitor M21 in the last year of the data, 1995. We do this based on the single monitor models in Section 4, noting that the results from the multiple monitor analysis are essentially similar. The analysis produces simulation-based descriptions of the joint posterior distribution for all M21/1995 parameters, namely the full set of quantities

$$\beta_0, \beta_1, \{\epsilon_i, \nu_i\}$$

for $i = 1, \ldots, I$ and where we have dropped the subscripts for monitor ($j = 2$) and year ($t = 1995$) for clarity here. So β_0 and β_1 are the hospital population level and DRG-regression coefficient in 1995, and the ϵ_i and ν_i are the systematic and residual random effects for hospital i on M21 in this year. Refer to equation (13) and suppress the t subscript. The particular analysis summarised is based on a posterior sample with Monte Carlo sample size 5,000 for all these quantities. The figures in Section 8.5 display some summaries of this posterior distribution, as follows.

Absolute outcome levels and comparisons across hospitals

For each hospital i we can easily compute the corresponding posterior sample for the actual quality outcome probability $p_i = 1/(1 + \exp(-\theta_i))$ where $\theta_i = \beta_0 + \beta_1 x_i + \epsilon_i + \nu_i$. This was done, followed by calculation of posterior medians and approximate 95% (equal tails) intervals for each p_i. These intervals, with medians marked, appear in Figure 37. The hospitals are ordered by the posterior medians so computed, and labelled by actual station numbers. The small number of hospitals with low sample sizes in M21 are identifiable here as they have wider intervals than the majority; the width of the interval reflects the spread of the marginal posterior which is a decreasing function of sample size.

If interest lies in questions about thresholds for the p_i, they can be directly addressed using these posterior distributions. The probability that any p_i exceeds or lies below any specified threshold can be immediately deduced from the posterior for p_i. A crude version of this across hospitals would involve simply drawing a vertical line at the specified thresholds on the graphs in Figure 37. Superimposed on the intervals are the values of the observed outcome proportions z_i/n_i, for comparison. Note that most of the observed outcomes are well within the interval and often very close to the median of the corresponding p_i. This partly reflects the large sample sizes n_i in most cases that lead to very influential observations and so fitted values that will be close to observed outcomes. The model is not, however, over-fitting, as is confirmed by the earlier discussions of residual structure analyses. Note, however, that the picture needs modification to more adequately represent the fit; these intervals are summary estimates of the true outcome levels p_i, and so the variability expressed by the width of the intervals does not incorporate the natural binomial variability in the outcomes around the levels. To do this we compute additional samples of binomial outcomes by drawing conditionally binomial counts $z_i^* \sim Bin(z_i^*|n_i, p_i^*)$ for each i, and where p_i^* represent the set of posterior draws for p_i. This produces the posterior predictive distribution for actual outcomes at each hospital. These can be summarised as we have summarised the p_i, in terms of interval estimates for the actual outcomes; the resulting 95% intervals with medians marked appear in Figure 38. Again, the actual data are superimposed, and it is now clear that the additional binomial variability broadens the intervals so as to very adequately cover the observed data.

This latter graph helps to immediately identify outcomes that are rather extreme compared to their posterior predictive distributions. These tend to be at hospitals with low sample sizes, and some further investigation of specific hospitals flagged in this graph, and also noted in the following graphs, may be of interest. For example, stations numbered 574, 533 and 608 are the three hospitals with observed outcome proportions most over-predicted, and cases 442, 585 and 617 are those most under-predicted. These are all cases for which the sample sizes n_i are relatively small, as is seen from Figure 42. Of the three over-predicted, the DRG-predicted proportions are very large relative to the majority of the hospitals, as is seen in Figures 43 and 44. The model naturally adapts to these lower observed outcomes for such cases, as is evident in the graph in Figure 39. This displays approximate 95% posterior intervals for the combined random effects $\epsilon_i + \nu_i$ for each hospital i. These are plotted on this logit scale in a format is similar to Figure 37 and with hospitals plotted in the same order for comparison. Here we see that these "low" hospitals have random effects lower than average, indicating that the model has indeed adapted to the extreme observations. The intervals are relatively wider for these cases, indicative of the smaller sample sizes. This adaptation is, however, constrained by the model form and also by the the high, and hence influential, values of the corresponding DRG predictor. The basic issue is that the DRG predictor for these cases seems too high, a suggestion that might be the subject for further investigation and study. Nevertheless, the fit of the model to even these very extreme cases is very good, consistent with our earlier subjective assessment of graphs of sampled data residuals.

Relative outcome levels and comparisons across hospitals

To further investigate the relative performance of hospitals we need to examine the hospital-specific departures from population levels, i.e., the combined random effects $\epsilon_i + \nu_i$ for each i. As noted above, Figure 39 displays approximate 95% posterior intervals for each of these quantities for all hospitals, with the hospitals in the order chosen for Figure 37 for comparison. There is a general increasing trend in the random effects as we go through the hospitals in this order, consistent with the ordering by outcome probabilities. The pattern is not monotonic, however; the probabilities include the effects of the DRG-based predictor, and the current graph focuses exclusively on the relative performance levels free from the regression effects. The following Figure 40 displays related posterior intervals, in the same hospital order. Here we have evaluated 95% posterior intervals for the *ranks* of the hospitals in terms of increasing levels of $\epsilon_i + \nu_i$. Evidently, the four or five hospitals with the highest *estimated* outcome probabilities have very high ranks, indicating that their *true* outcome probabilities are indeed very likely to be among the largest few across the system. Similar comments apply to those with the lowest estimated probabilities. Among the majority of the hospitals, however, there are much higher uncertainties about rankings, with posterior intervals spanning fairly wide ranges. This way of investigating relative performance should be contrasted with that based on the outcome probabilities themselves; Figure 41 displays the plot of posterior intervals for rankings of the p_i displayed in Figure 37, which gives a very different picture and resulting inferences than that based on the rankings of effects $\epsilon_i + \nu_i$. The differences are important as the two ranking graphs address very different questions: Figure 41 summarises absolute performance, impacted by patient-mix and other confounding factors; Figure 40 represents relative quality levels once these factors are accounted for through the model, and a firmer basis for assessing relative performance due to hospital-specific policies and practices.

The posterior distributions summarised here can be explored in various other ways. Subsets of hospitals can be selected for further such summary and deeper investigation, for example. Similarly, questions about changes in hospital-specific effects can be addressed by looking at interval estimates of the parameters measuring changes year to year. Much can be extracted from posterior analysis that has not been illustrated here.

7 Summary Comments and Conclusions

The key summary conclusions are as follows.

- The inferred patterns of change over time in the key population levels β_{0jt} and DRG regression coefficients β_{1jt} are essentially the same in all models explored. There are evident changes in β_{0jt} that require consideration and interpretation by VA personnel. There are differences over the years in the regression coefficient β_{1jt} too, though these changes are less marked.

- Our multiple monitor time series models isolate changes over time and dependencies among such changes in the hospital-specific random effects across the three monitors. Though dependencies across monitors exist, they are apparently quite small. There are negligible differences in inferences about the key population quantities/parameters in extending from from the single monitor `MODEL-2` framework to the linked multiple monitor `MODEL-3` framework, at least for these three monitors. With other sets of monitors or other contexts, dependence patterns may be stronger and then the `MODEL-3` framework of more interest. There will be bigger and possible meaningful differences in inferences about hospital-specific random effects in cases of very high correlations between monitors, and in cases when the sample sizes n_i are much smaller than is typical with these three monitors. On the other hand, little is lost by analysing the data in the multiple monitor framework anyhow, though the software development and computational costs are meaningful. Developing the multiple monitor framework has been the major activity on the project to date.

- A critical feature of our work has been the identification of several components of variability underlying within-year variation and across-year changes in observed quality levels. The important part of this is the partitioning the hospital-specific variation in outcome probabilities into two components: one, a partially systematic and positively dependent AR component, represented by the ϵ_{ijt}; and two, the purely unpredictable component represented by the ν_{ijt}. In each monitor separately, and with confirmation from the multi-monitor studies, the latter are very significant, and contribute between 15-30% of the total random effects variance on the logit scale. The extent of the contribution is monitor-specific, with that for the *general medical discharge* monitor M22 being significantly lower than either of the (more comparable) psychiatric monitors M20 and M21. Thus hospital-specific levels of M22 are more stable over time and hence more predictable.

- Summary graphs of posterior inferences for specific monitor:year choices can provide useful insight into the distribution of outcome probabilities across the hospital system, about relative levels of performance, and about changes over time in such levels. Issues of how to most effectively summarise and use such information require consideration. Specific hospital/monitor/year outcomes that show up as evident extremes in these analyses may require further investigation.

There are various modelling assumptions that could be explored and tested by further development of the existing framework. These include extensions to allow for non-normal error distributions for the components of random effects ϵ_{ijt} and ν_{ijt}. It is relatively straightforward to replace the normality assumptions with alternative, heavy-tailed error distributions, such as Student-T, in order to explore questions of robustness and sensitivity. On the basis of our residual analysis and exploration to date, we are somewhat skeptical that such extensions would materially impact on the analysis or substantially change our summary inferences, although this may not be the case in studies of other monitor series.

One further useful extension will be to incorporate arbitrary patterns of missing data. The current report is restricted to analyses of 152 hospitals on which we have observed outcomes on each of the three monitors in each of the eight years of study. Seven more hospitals in the data base have outcomes reported only on subsets of monitor/year pairings, so were omitted from the study. Rather routine model extensions will incorporate this and other patterns of missing data,

A related potential development involves within-year predictions for model validation. Minor changes to the analysis will allow for reanalysis with selected hospital/monitor cases removed from the data set, but with the model still including such cases in analysis. This leads to posterior distributions for the parameters at these hospitals that are updated from the priors based on information from the other hospitals, but which are not now based on the actual outcomes at the omitted hospitals. Exploration of the concordance between these true outcomes and the posterior predictions will provide out-of-sample, or cross-validatory, assessment of model adequacy.

8 Figures from all analyses

8.1 Data displays

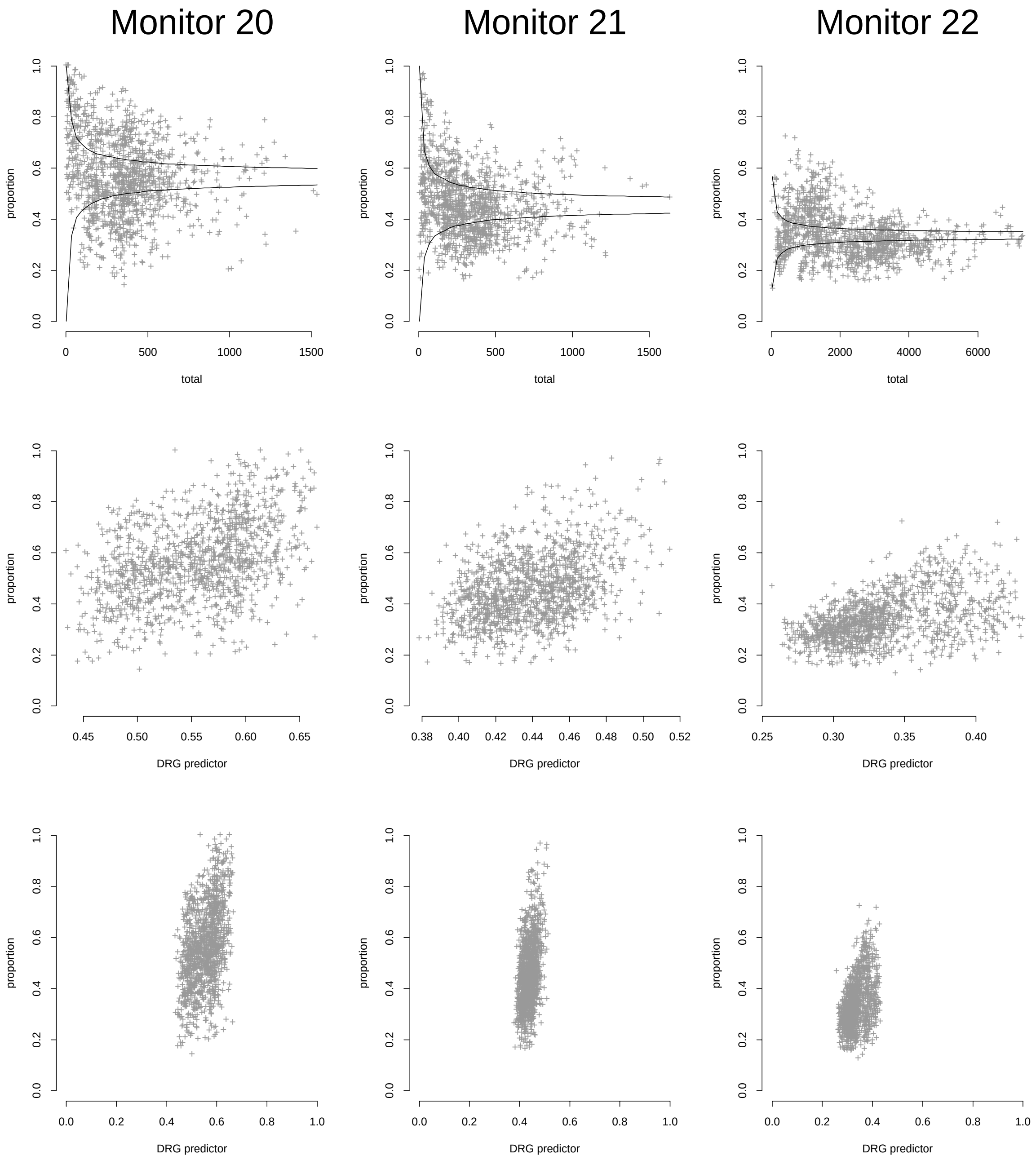
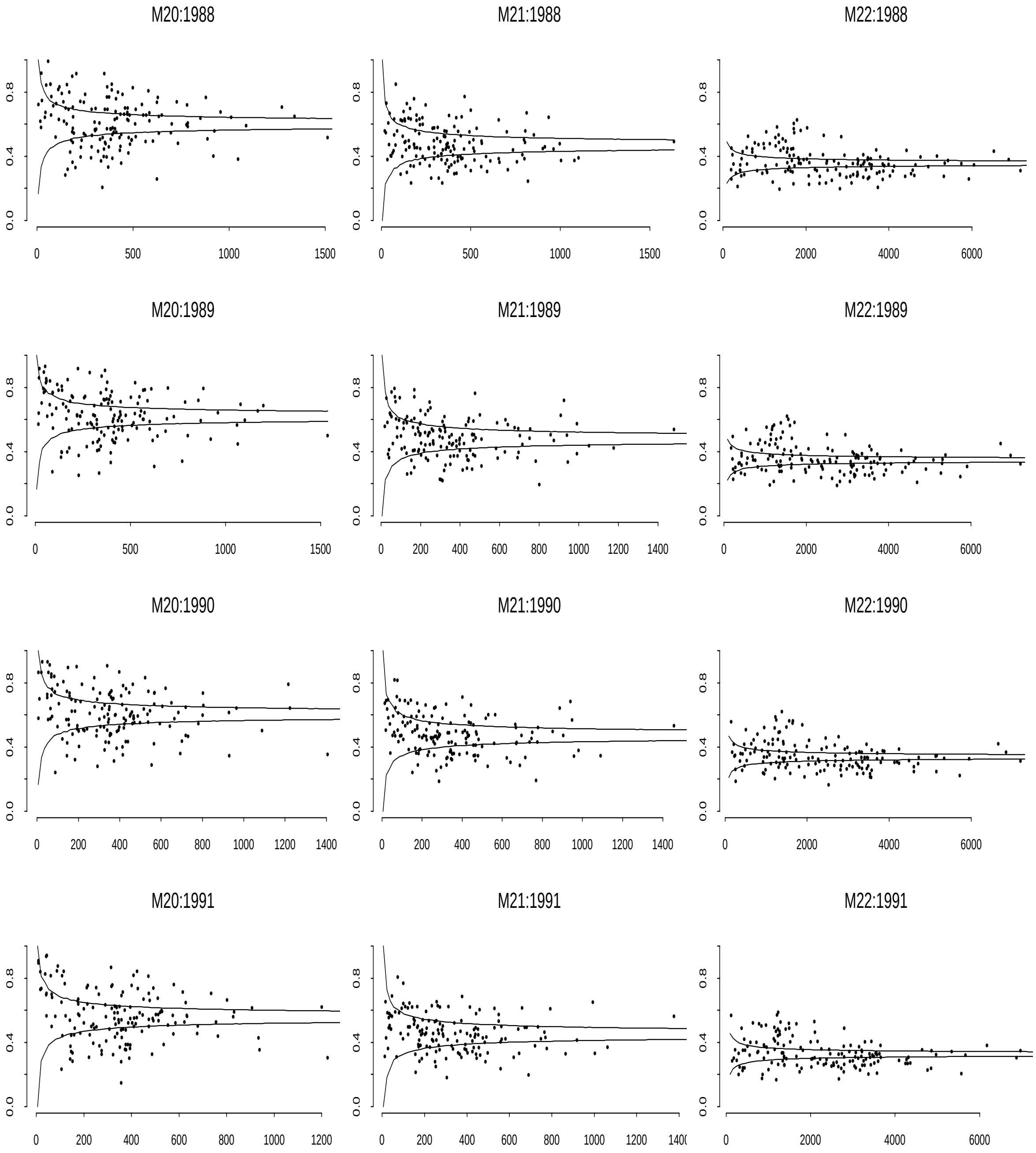


Figure 1: Observed proportions on all three monitors. All eight years of data (1988-95) appear in each frame. The upper row displays outcome proportions against corresponding total sample sizes. The superimposed lines denote pointwise 99% intervals based on assumed common binomial distributions for the outcomes in each monitor (see text for discussion). The second row of graphs displays the observed proportions versus the DRG-based predicted proportions; this display is redrawn in the lower row, but now with the DRG predictor on the probability scale.


M20:1988 M21:1988 M22:1988

M20:1989 M21:1989 M22:1989

M20:1990 M21:1990 M22:1990

M20:1991 M21:1991 M22:1991

Figure 2: Observed proportions on all three monitors for the years 1988-1991 inclusive. As in Figure 1 the outcome proportions are plotted against total sample sizes and the superimposed lines denote pointwise 99% intervals based on assumed binomial distributions for the outcomes in each monitor in each year (see text for discussion)

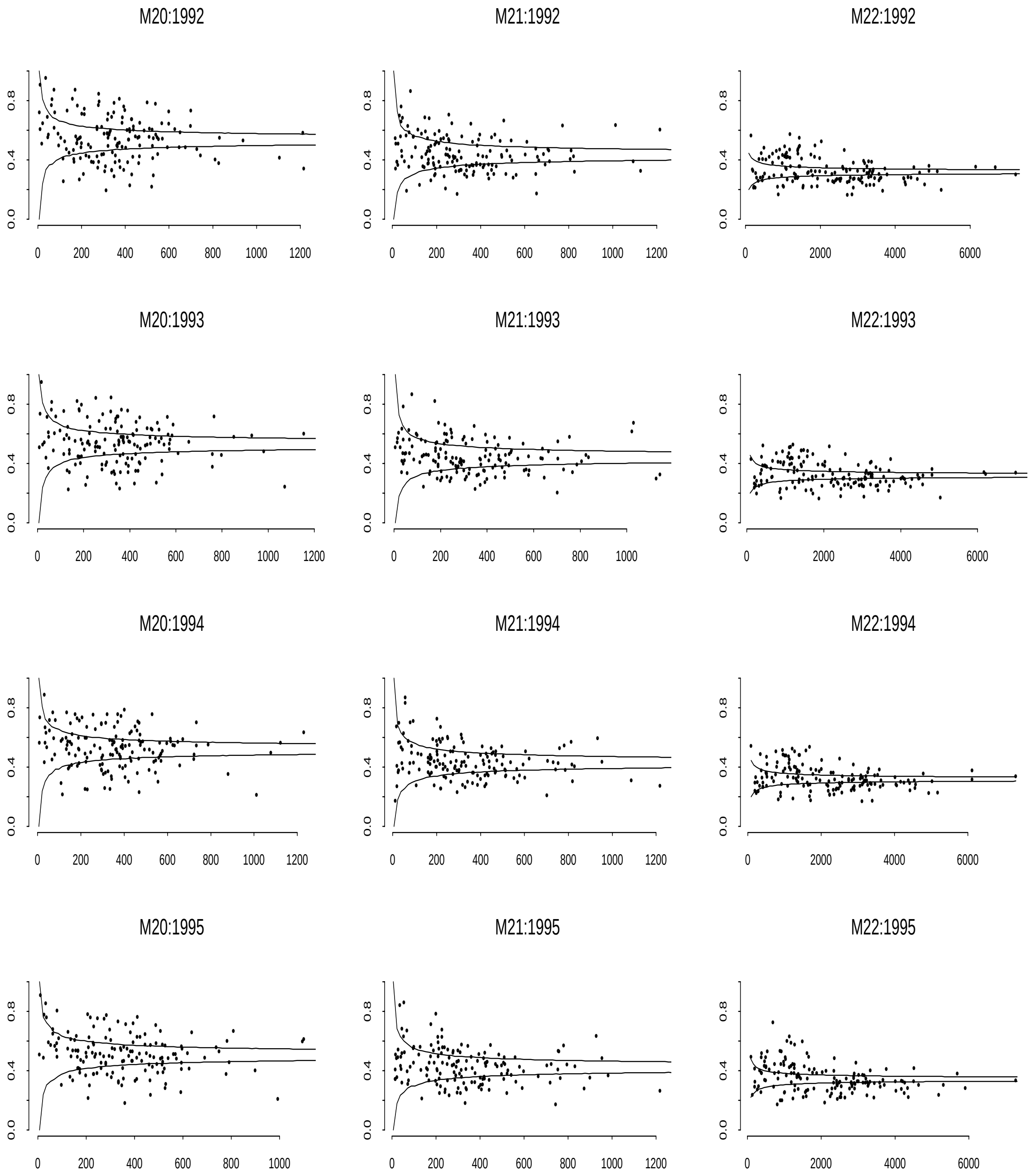


Figure 3: Data displayed as in Figure 2 but now for years 1992-1995 inclusive.

.

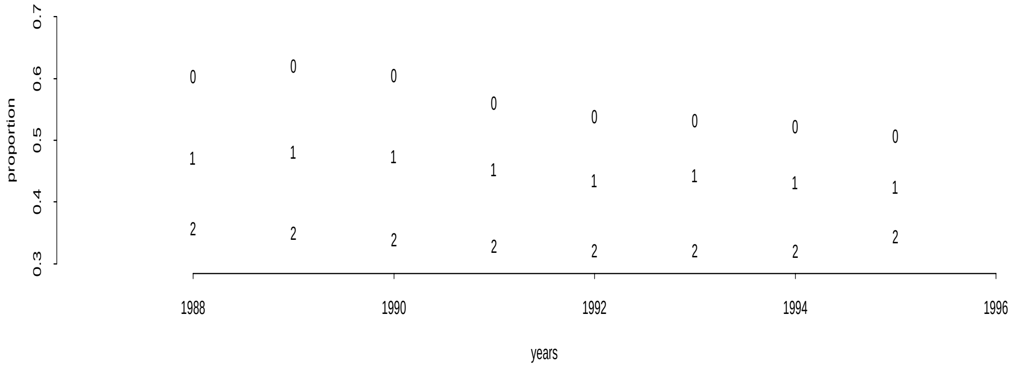


Figure 4: Raw averages of observed response proportions over the eight years, with Monitor 20 marked 0, Monitor 21 marked 1 and Monitor 22 marked 2, There is an apparent decreasing trend, most marked for Monitors 20 and 21.

8.2 `MODEL-1`: Random effects, single monitor, single year models

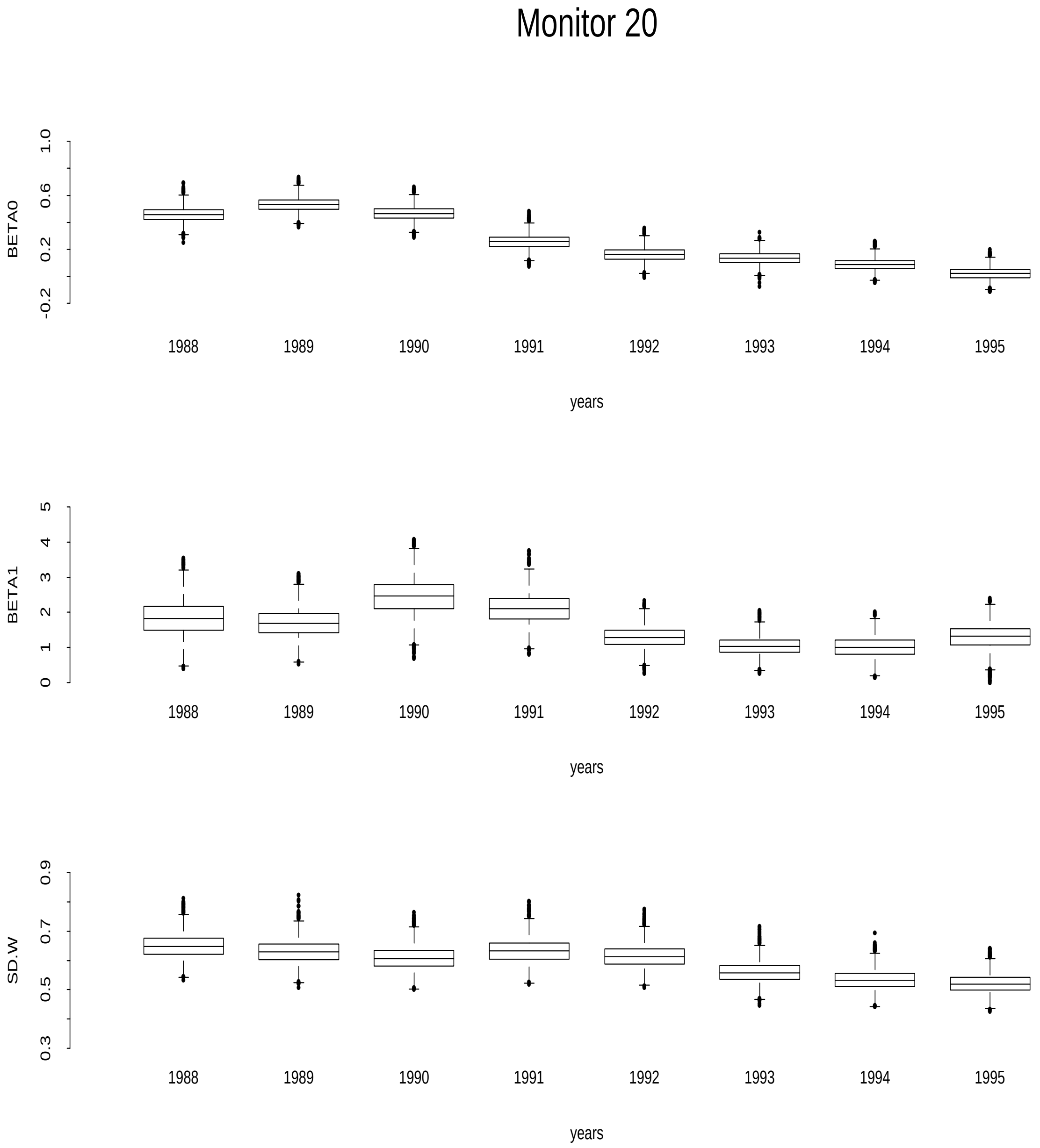


Figure 5: `MODEL-1`: Single year analyses of single Monitor 20: Summary interval estimates of β_{0t}, β_{1t} and w_t.

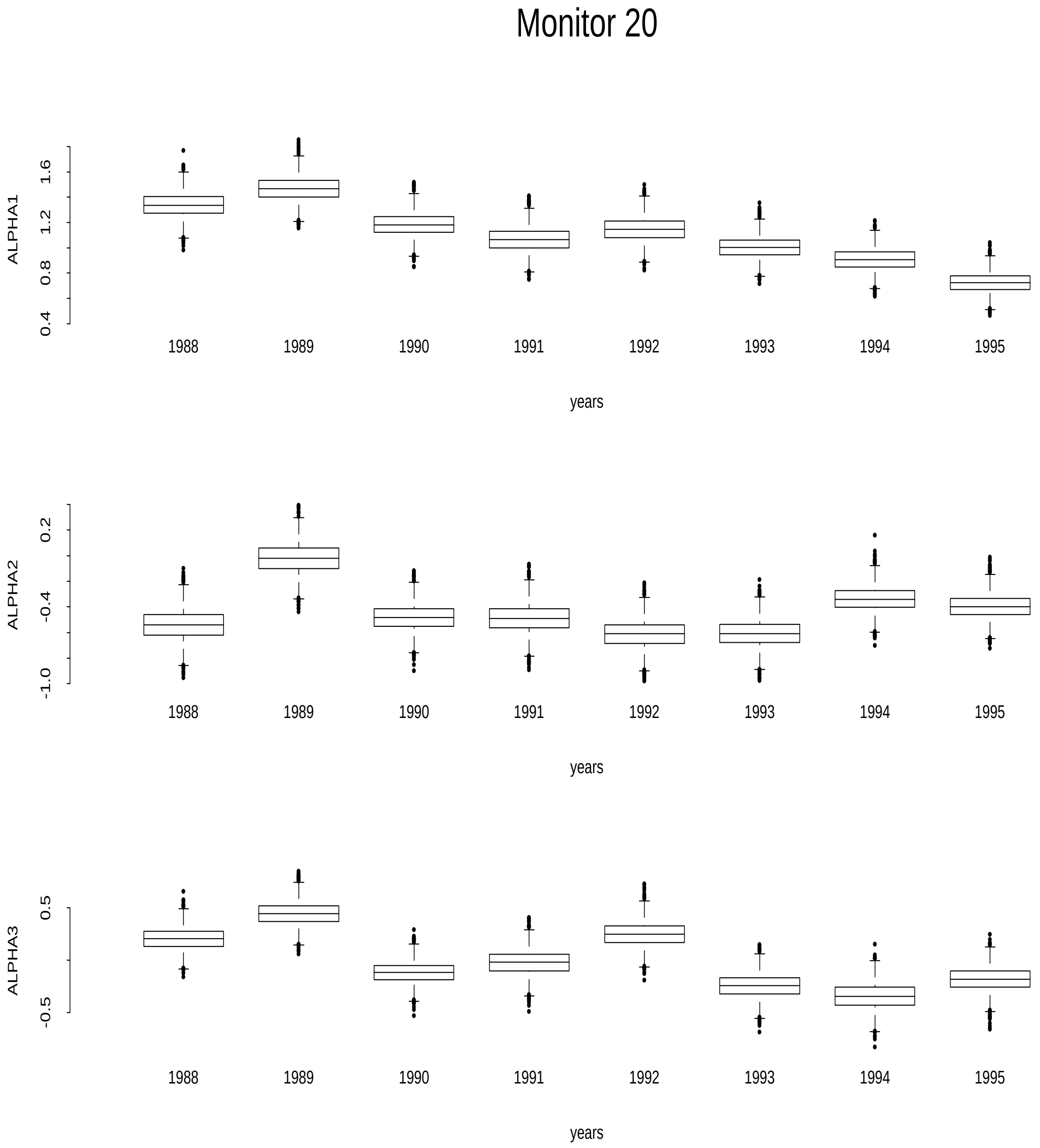


Figure 6: `MODEL-1`: Single year analyses of single Monitor 20: Summary interval estimates of hospital-specific parameters α_{it} for three hospitals.

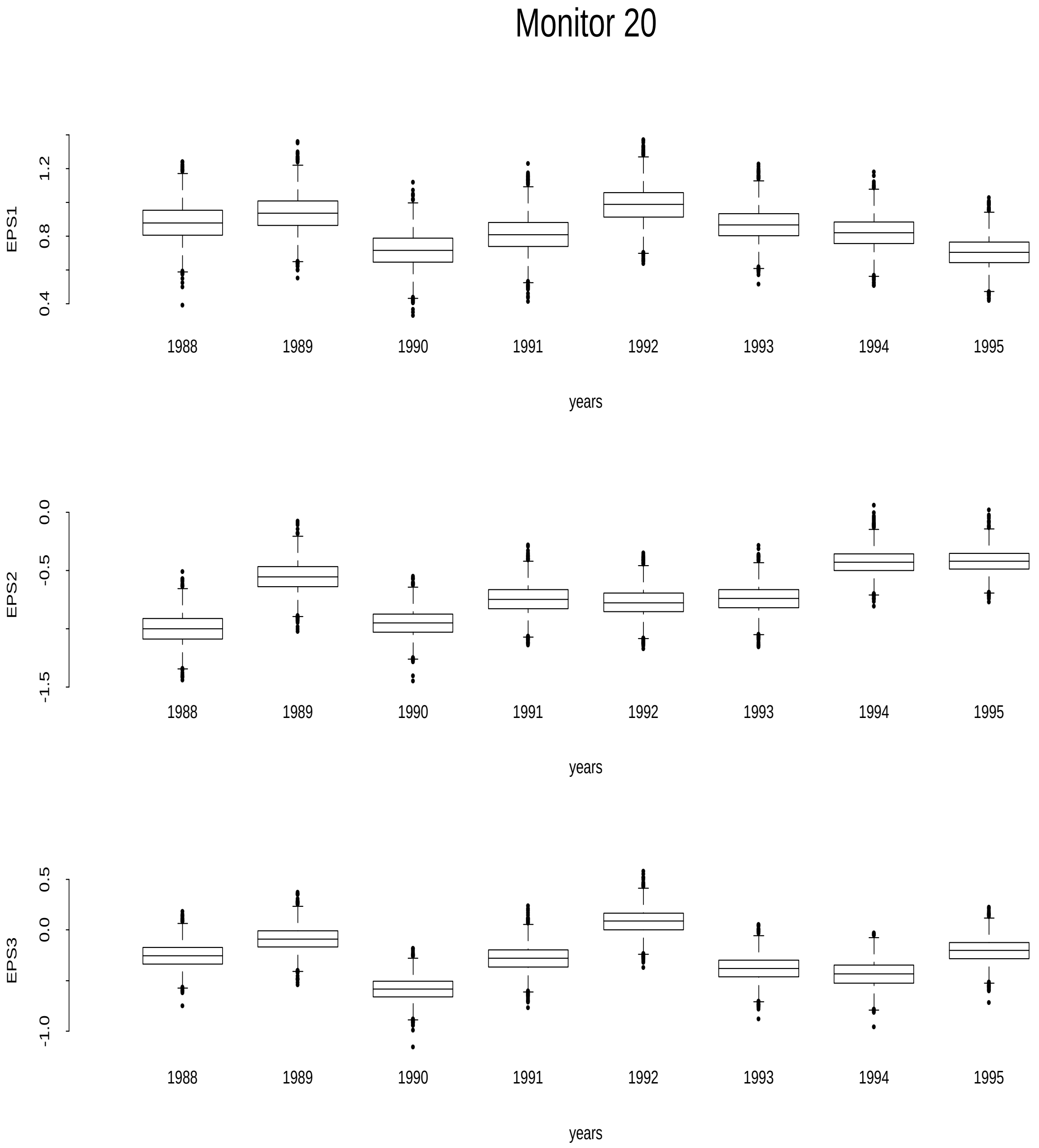


Figure 7: `MODEL-1`: Single year analyses of single Monitor 20: Summary interval estimates of hospital-specific random effects ϵ_{it} for three hospitals.

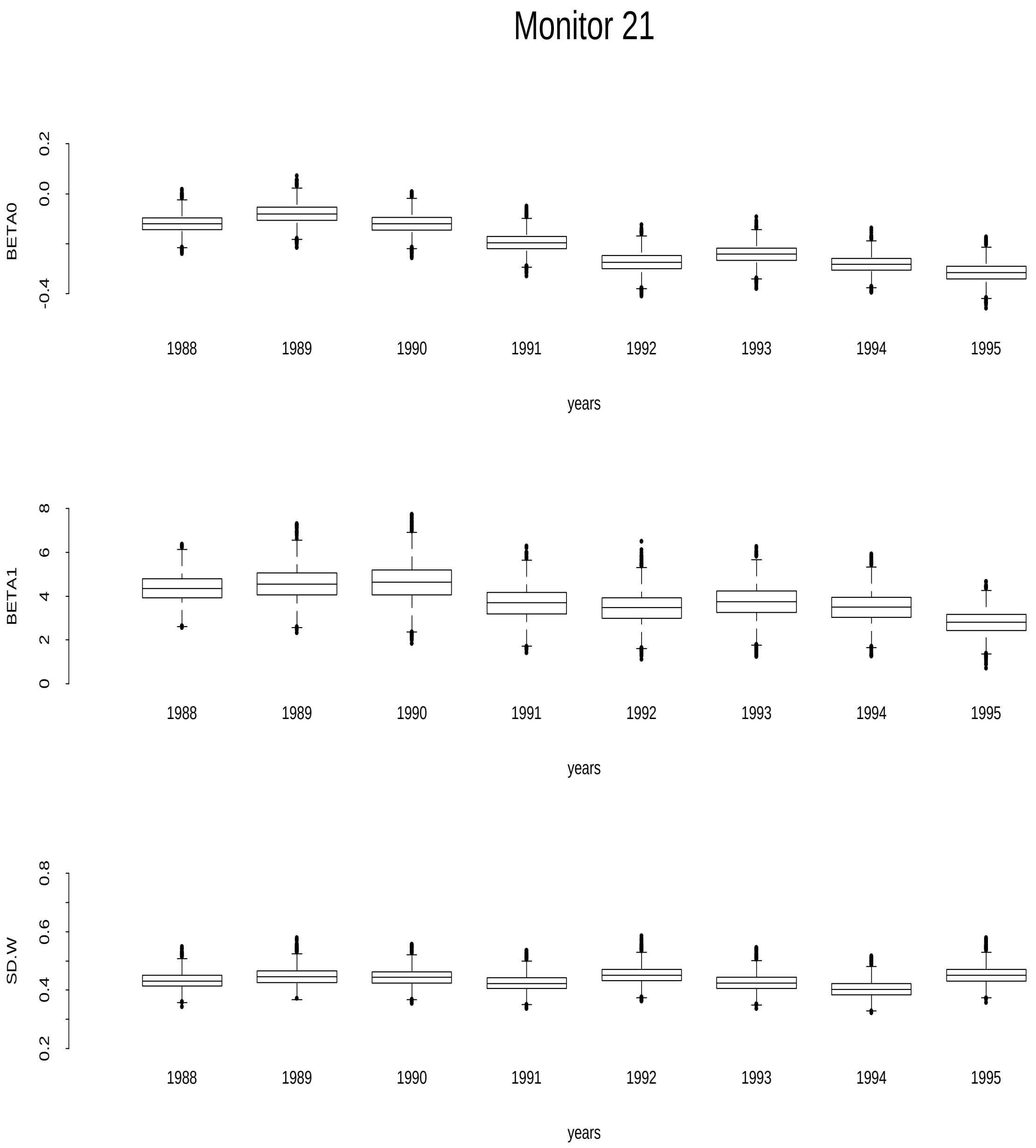


Figure 8: `MODEL-1`: Single year analyses of single Monitor 21: Summary interval estimates of β_{0t}, β_{1t} and w_t.

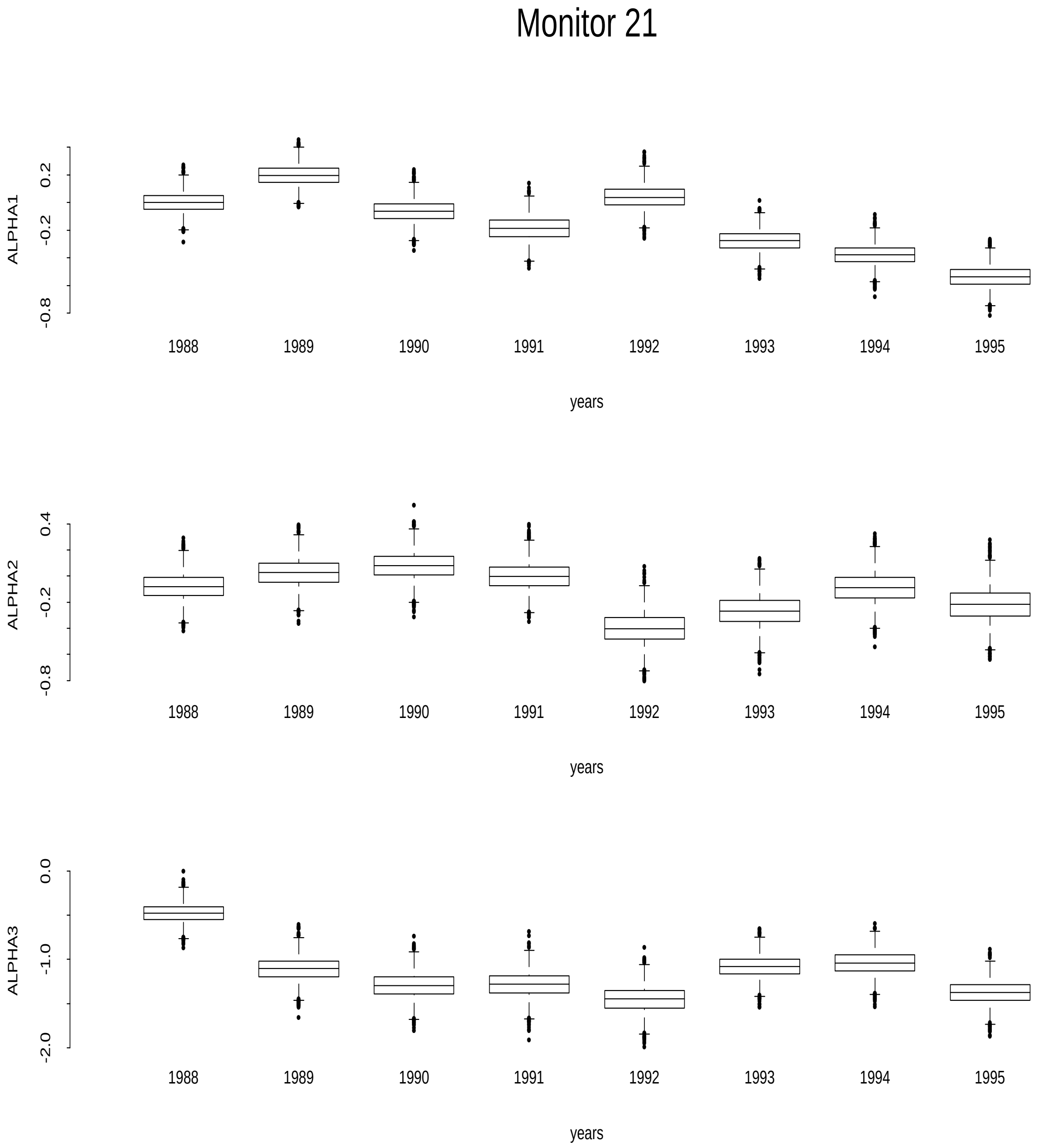


Figure 9: `MODEL-1`: Single year analyses of single Monitor 21: Summary interval estimates of hospital-specific parameters α_{it} for three hospitals.

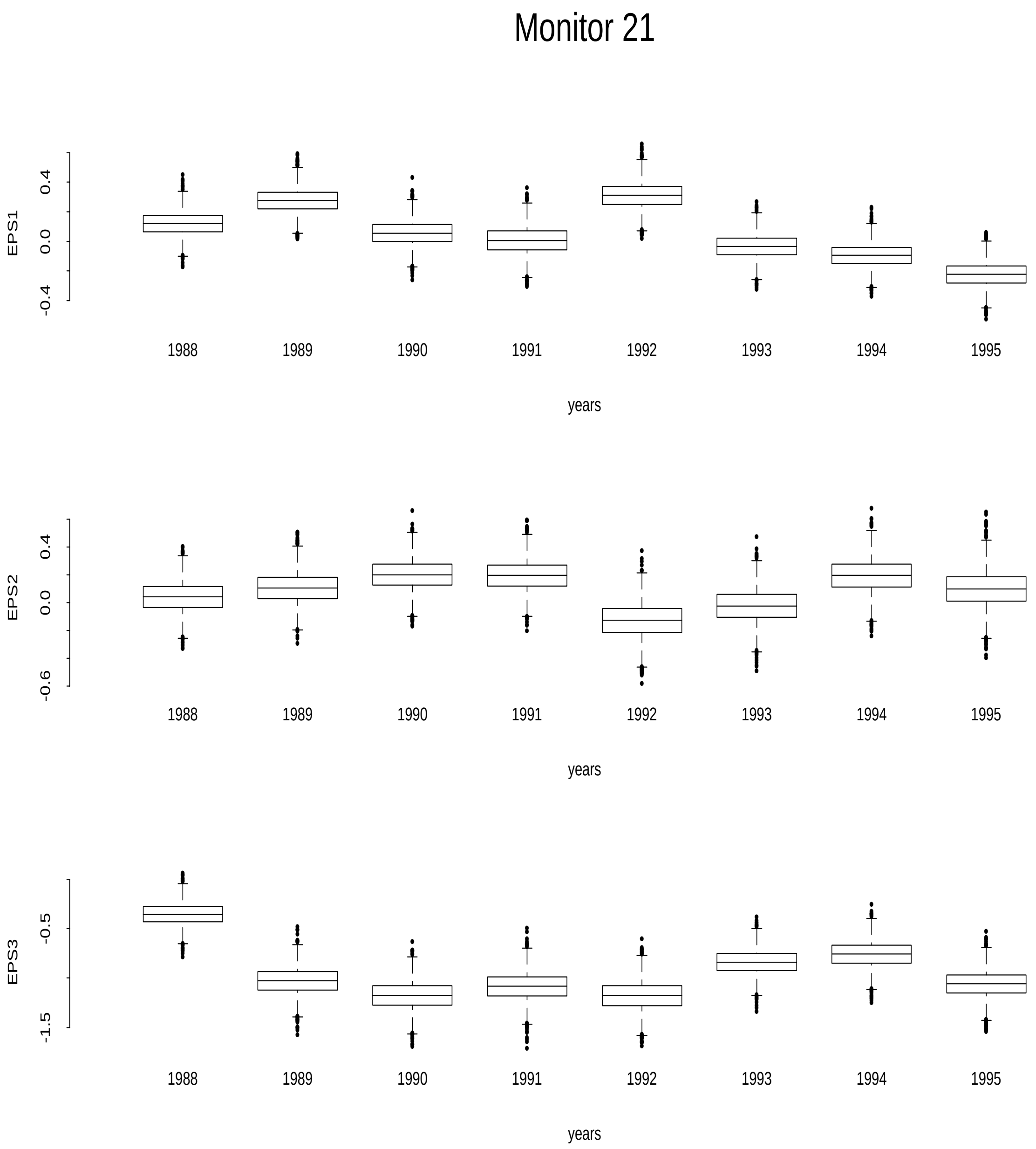


Figure 10: MODEL-1: Single year analyses of single Monitor 21: Summary interval estimates of hospital-specific random effects ϵ_{it} for three hospitals.

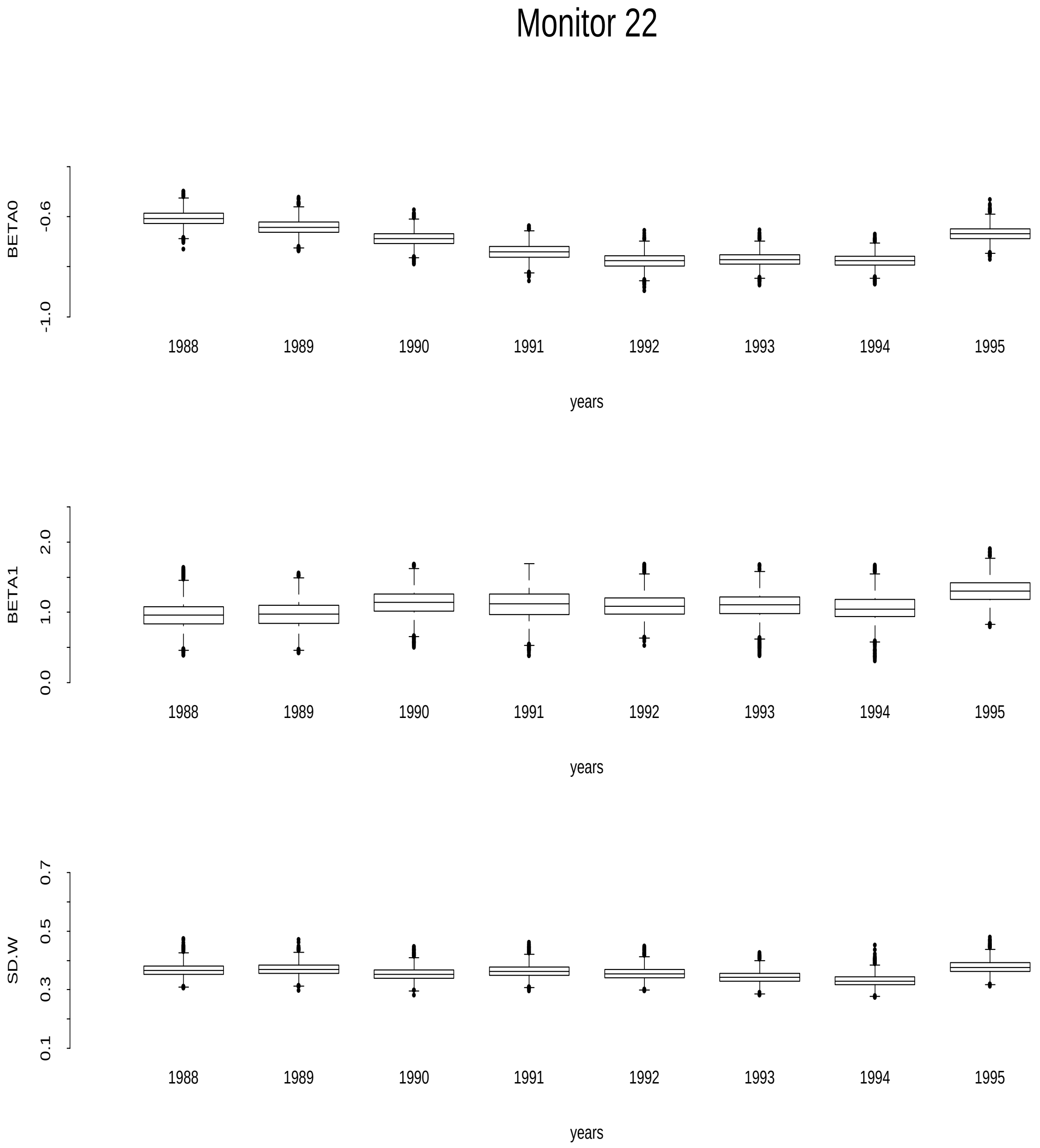


Figure 11: `MODEL-1`: Single year analyses of single Monitor 22: Summary interval estimates of β_{0t}, β_{1t} and w_t.

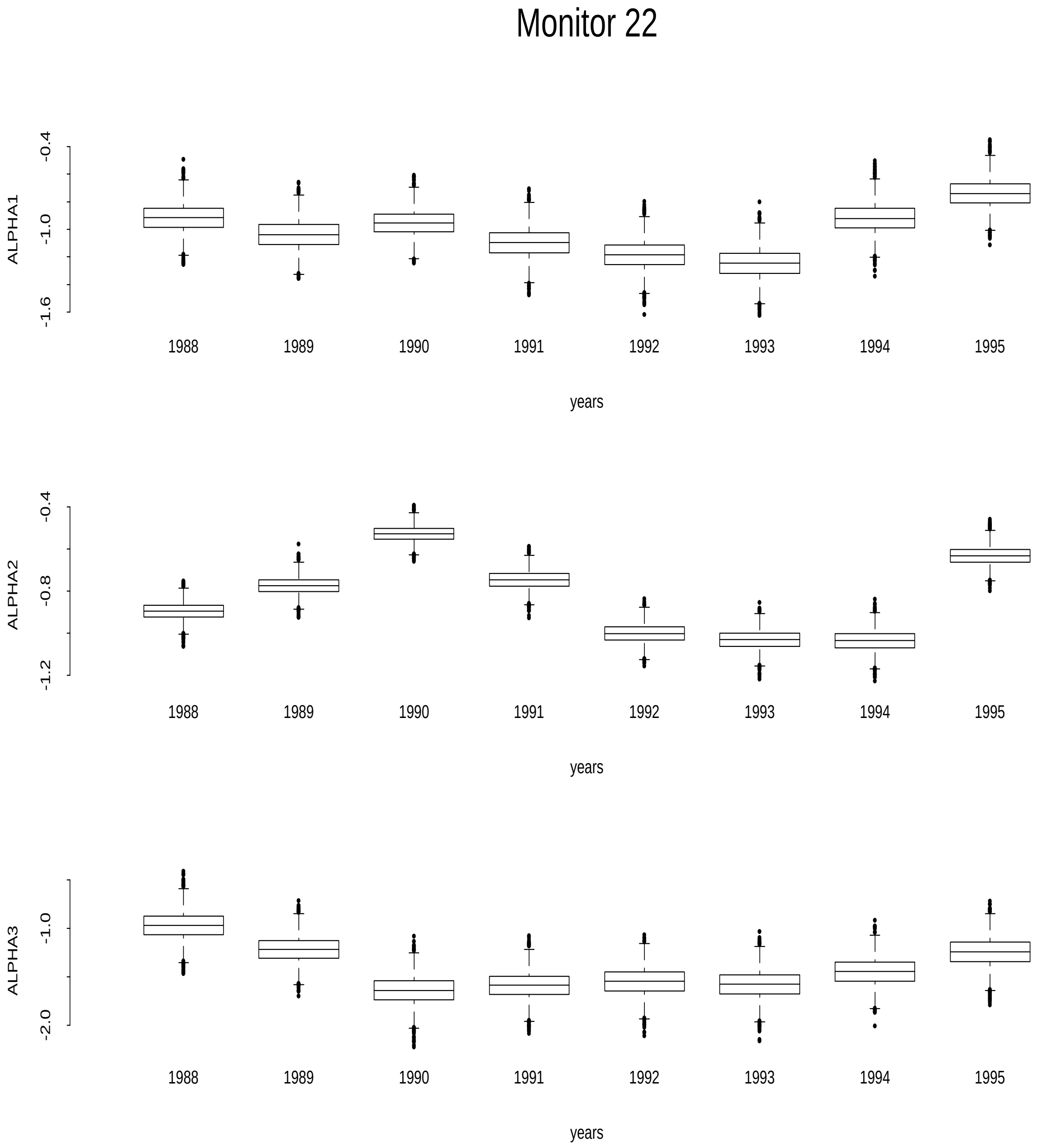


Figure 12: `MODEL-1`: Single year analyses of single Monitor 22: Summary interval estimates of hospital-specific parameters α_{it} for three hospitals.

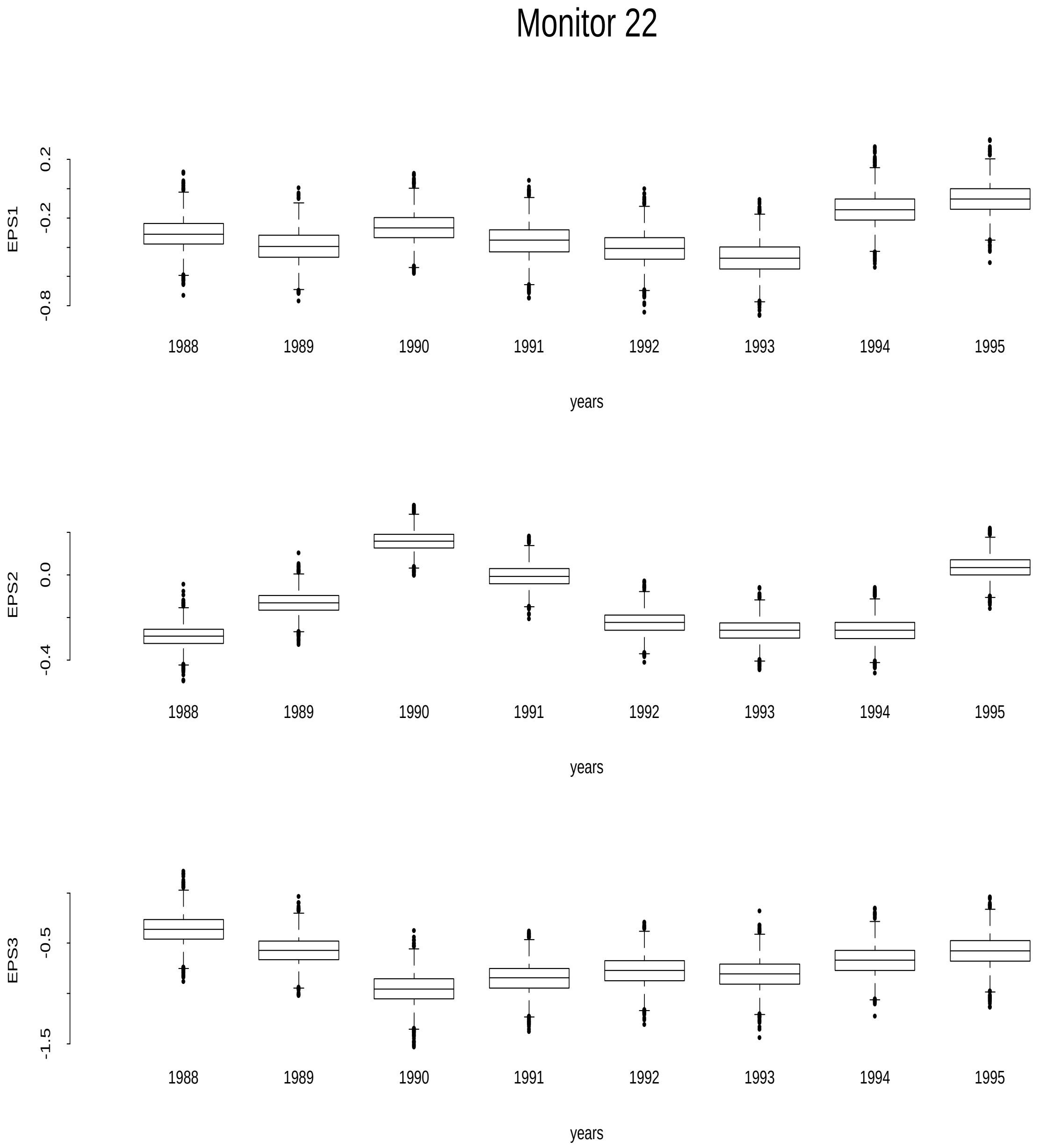


Figure 13: MODEL-1: Single year analyses of single Monitor 22: Summary interval estimates of hospital-specific random effects ϵ_{it} for three hospitals.

8.3 `MODEL-2`: Random effects time series models for single monitor series

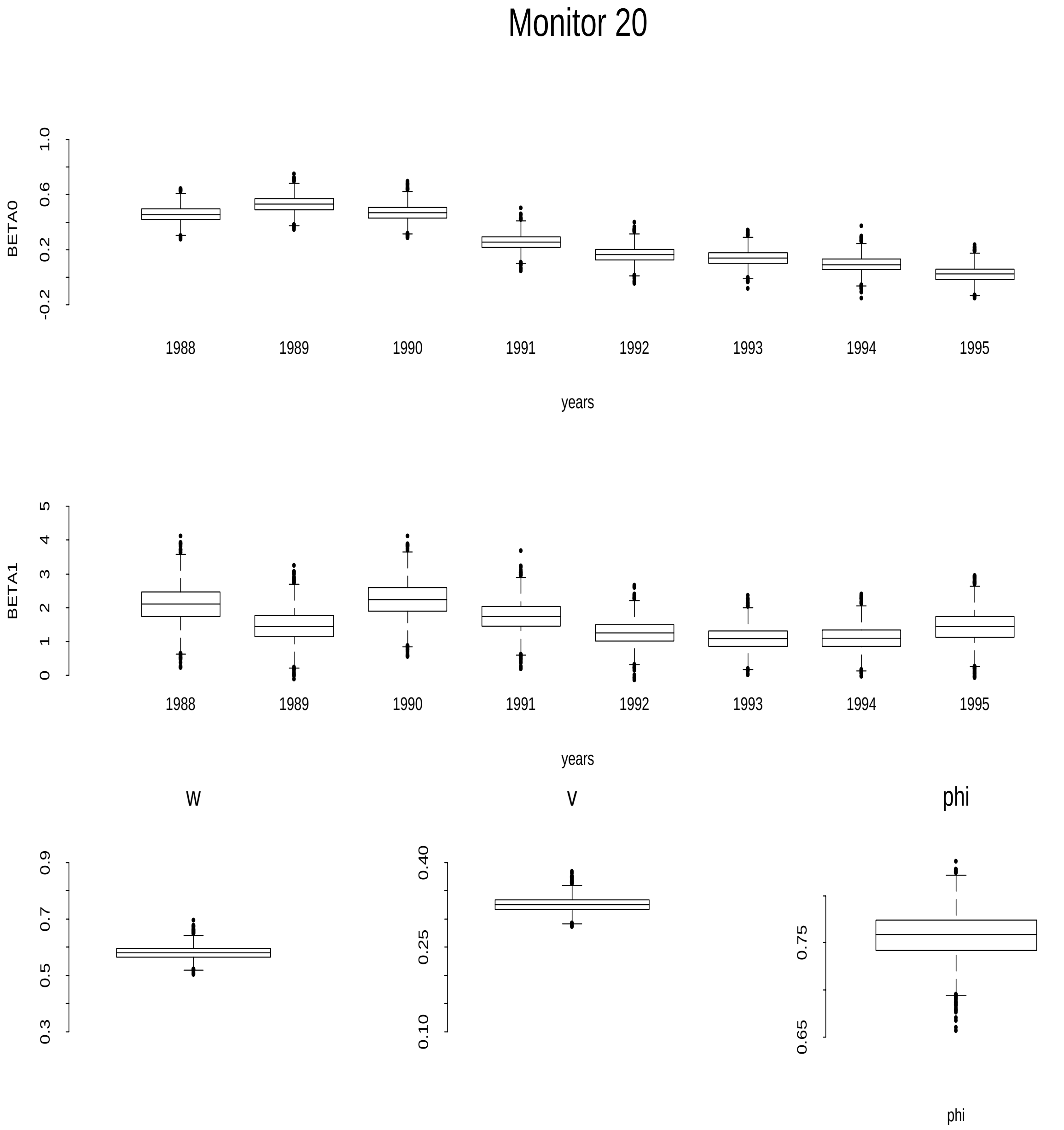


Figure 14: `MODEL-2`: Multi-year analysis of single Monitor 20: Summary interval estimates of β_{0t}, β_{1t}, standard deviations w, v and correlation parameter ϕ.

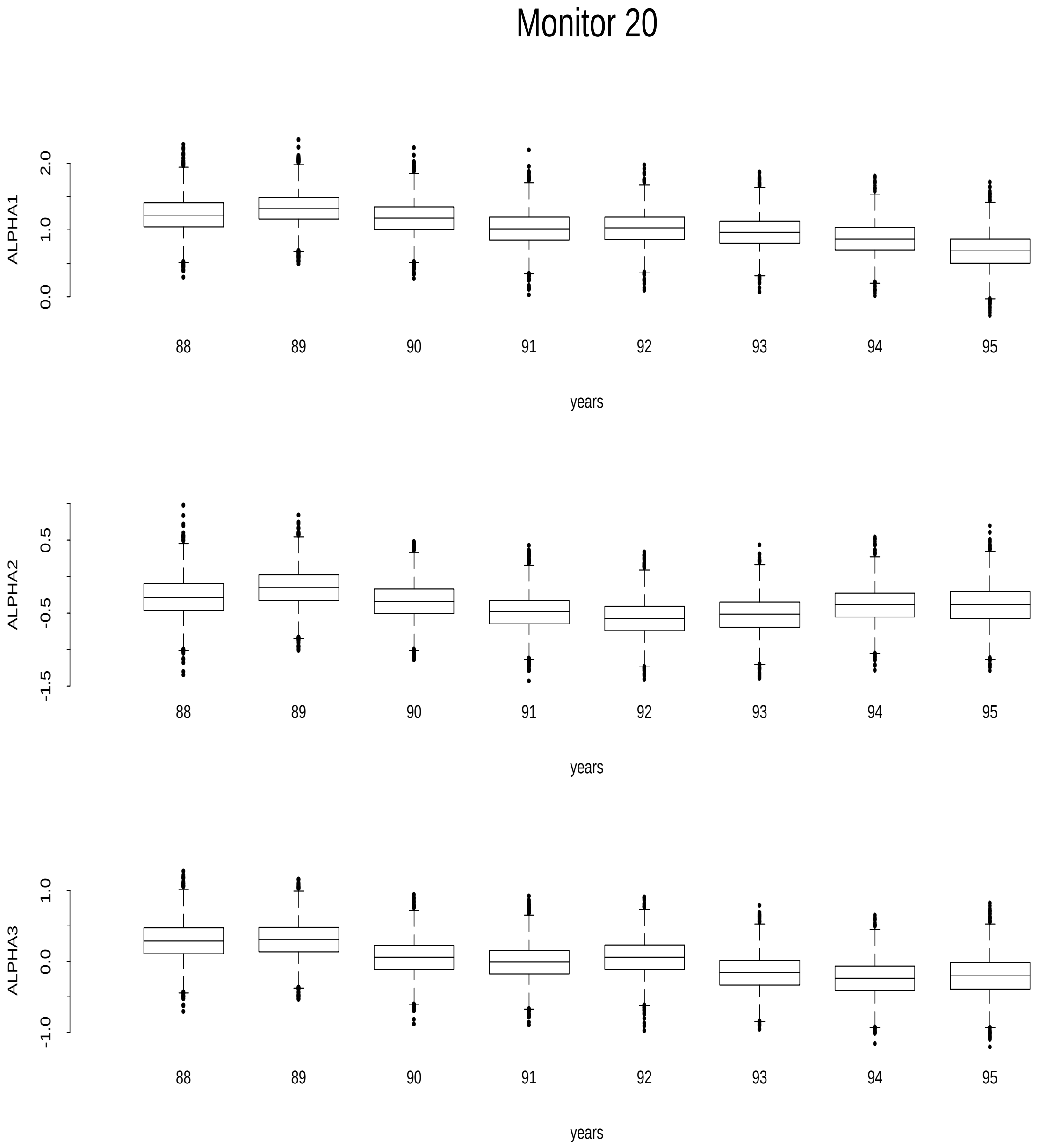


Figure 15: `MODEL-2`: Multi-year analysis of single Monitor 20: Summary interval estimates of hospital-specific parameters α_{it} for three hospitals.

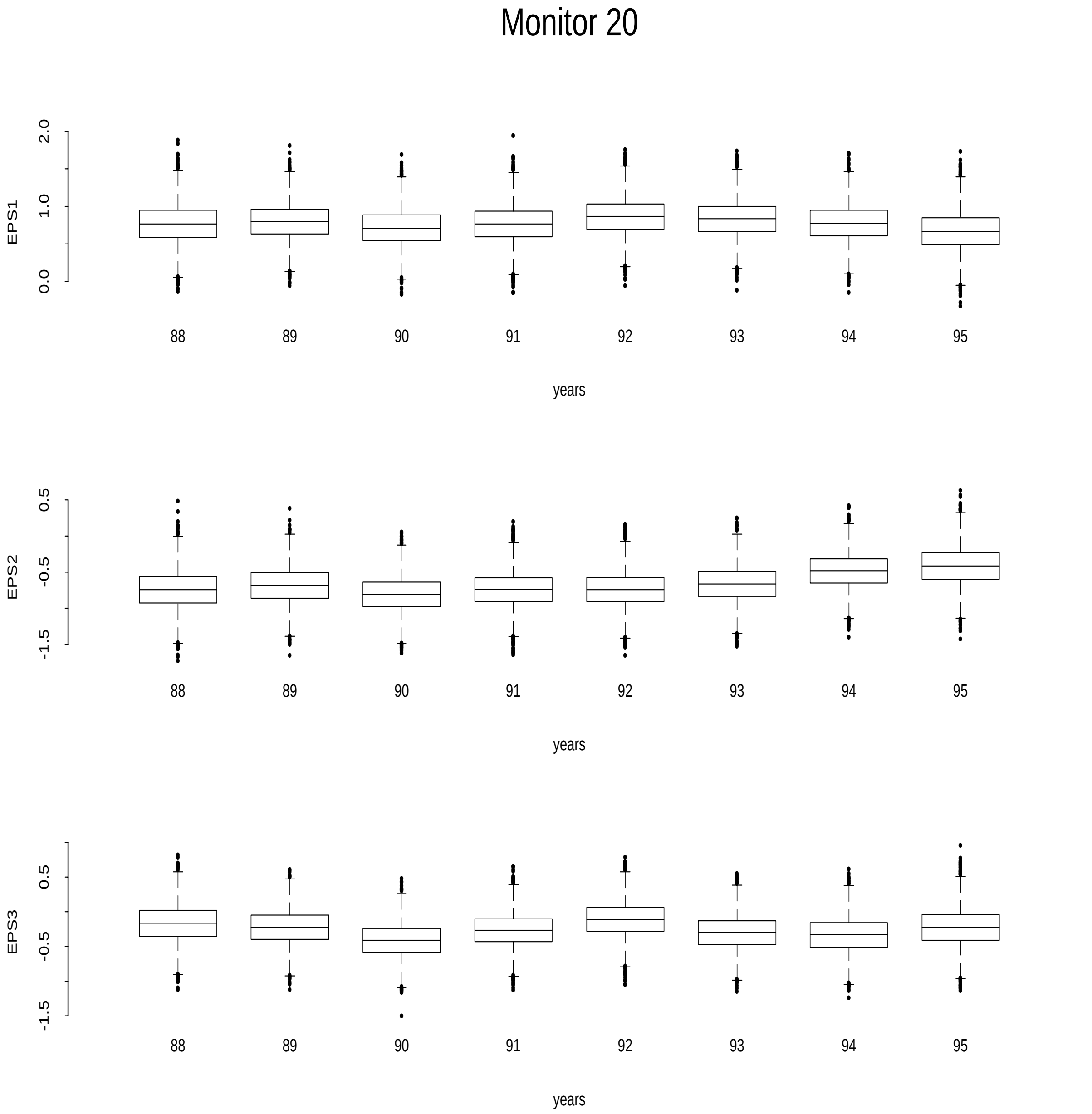


Figure 16: `MODEL-2`: Multi-year analysis of single Monitor 20: Summary interval estimates of hospital-specific random effects ϵ_{it} for three hospitals.

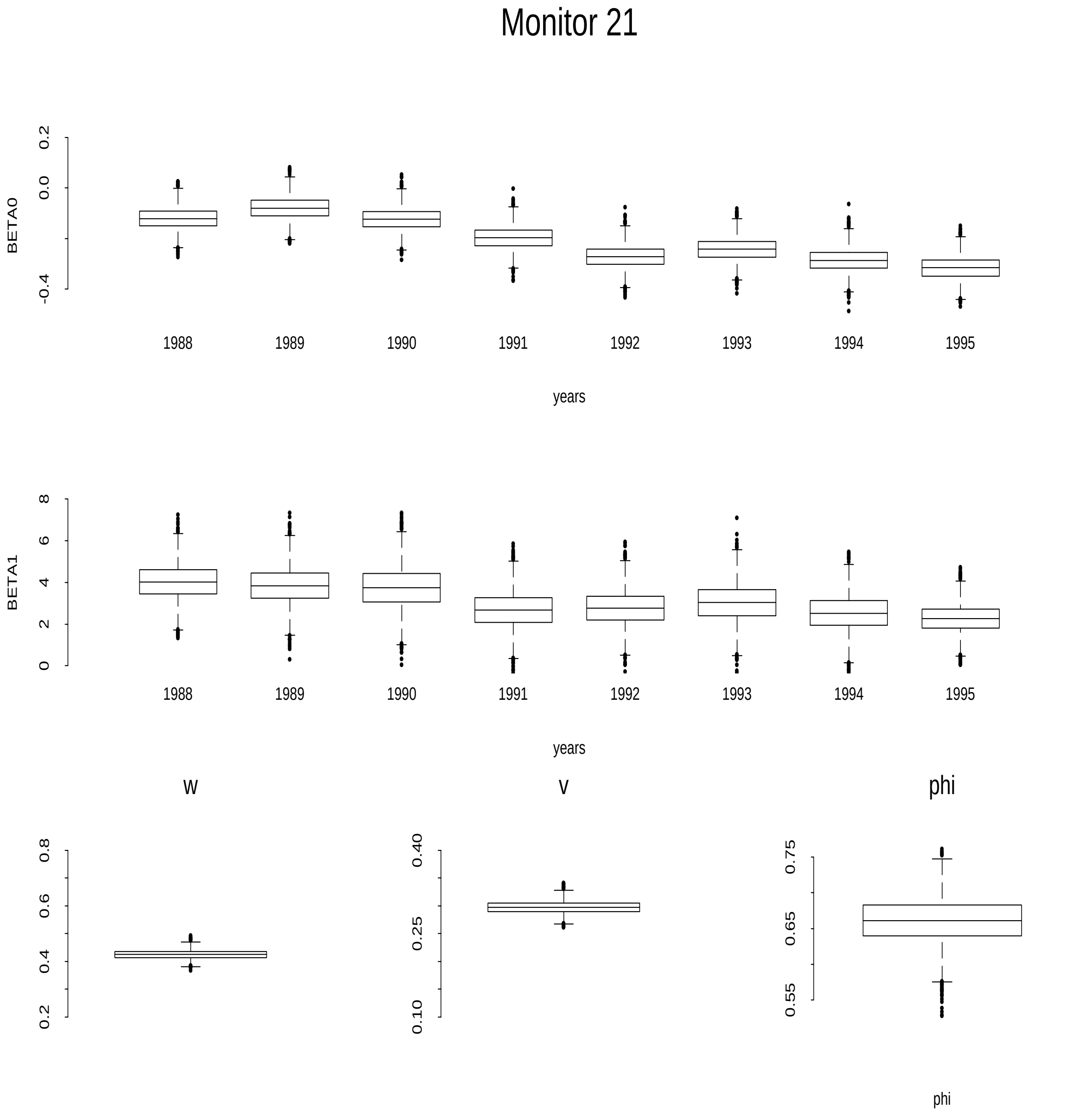


Figure 17: `MODEL-2`: Multi-year analysis of single Monitor 21: Summary interval estimates of β_{0t}, β_{1t}, standard deviations w, v and correlation parameter ϕ.

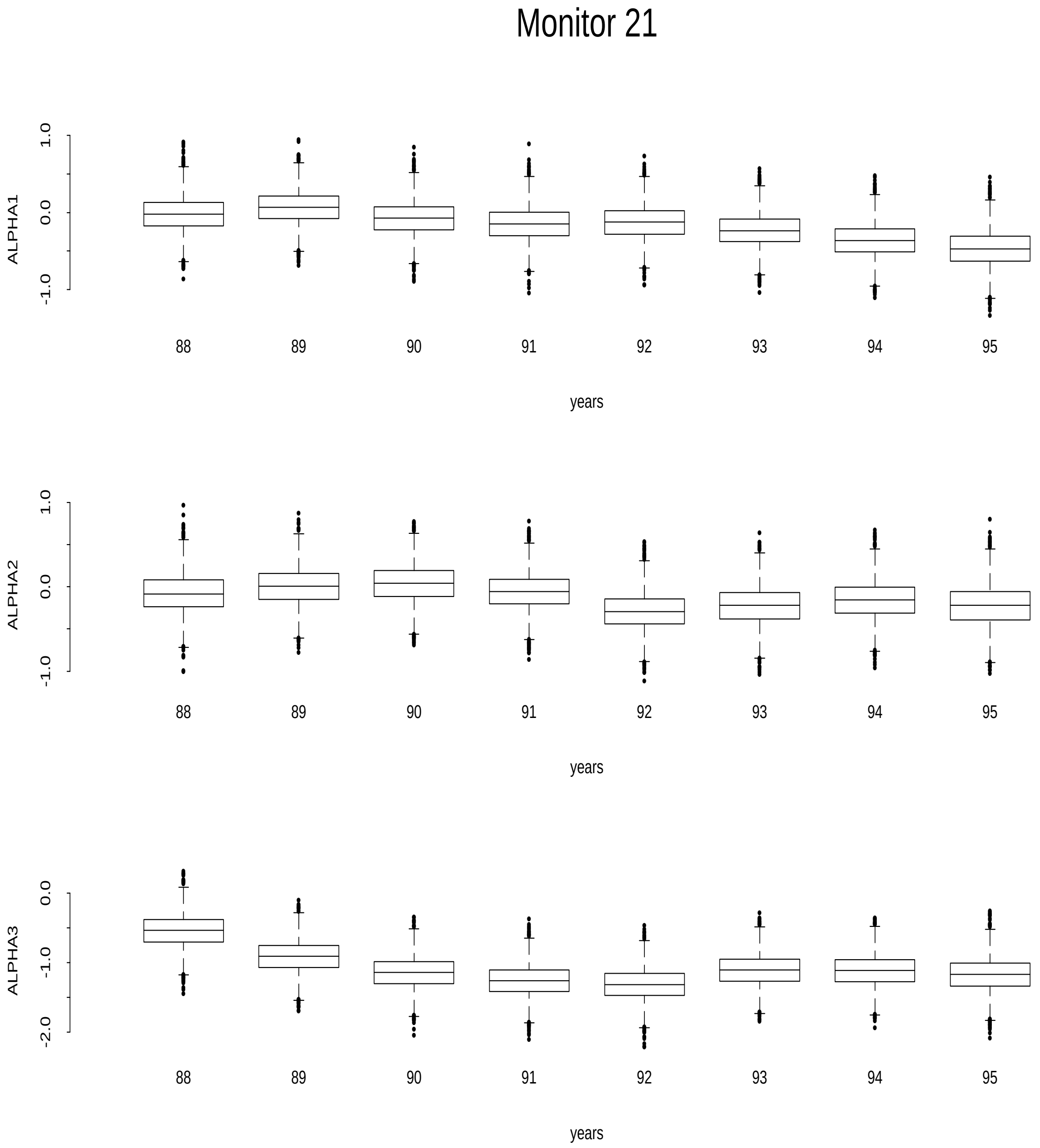


Figure 18: `MODEL-2`: Multi-year analysis of single Monitor 21: Summary interval estimates of hospital-specific parameters α_{it} for three hospitals.

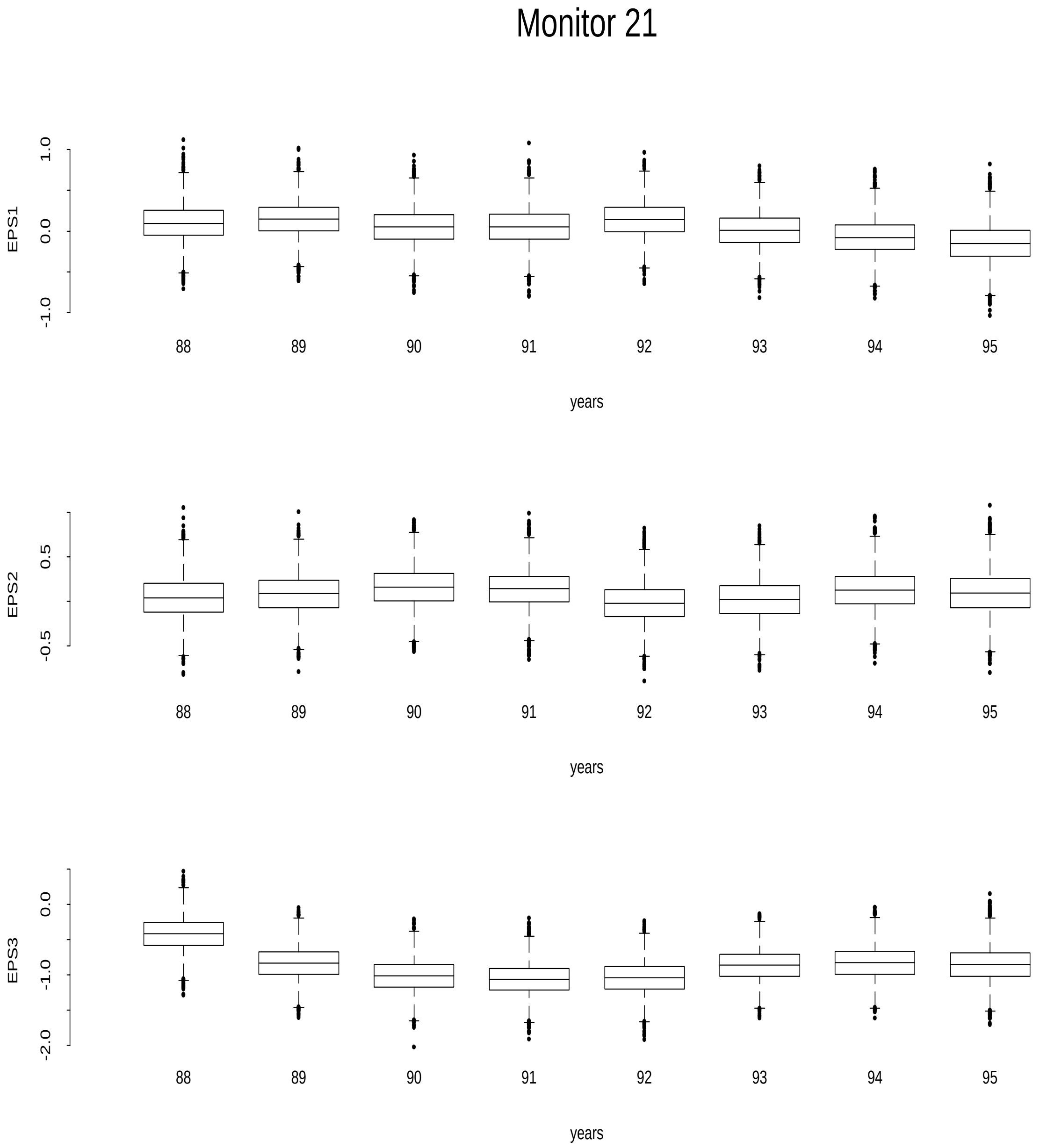


Figure 19: MODEL-2: Multi-year analysis of single Monitor 21: Summary interval estimates of hospital-specific random effects ϵ_{it} for three hospitals.

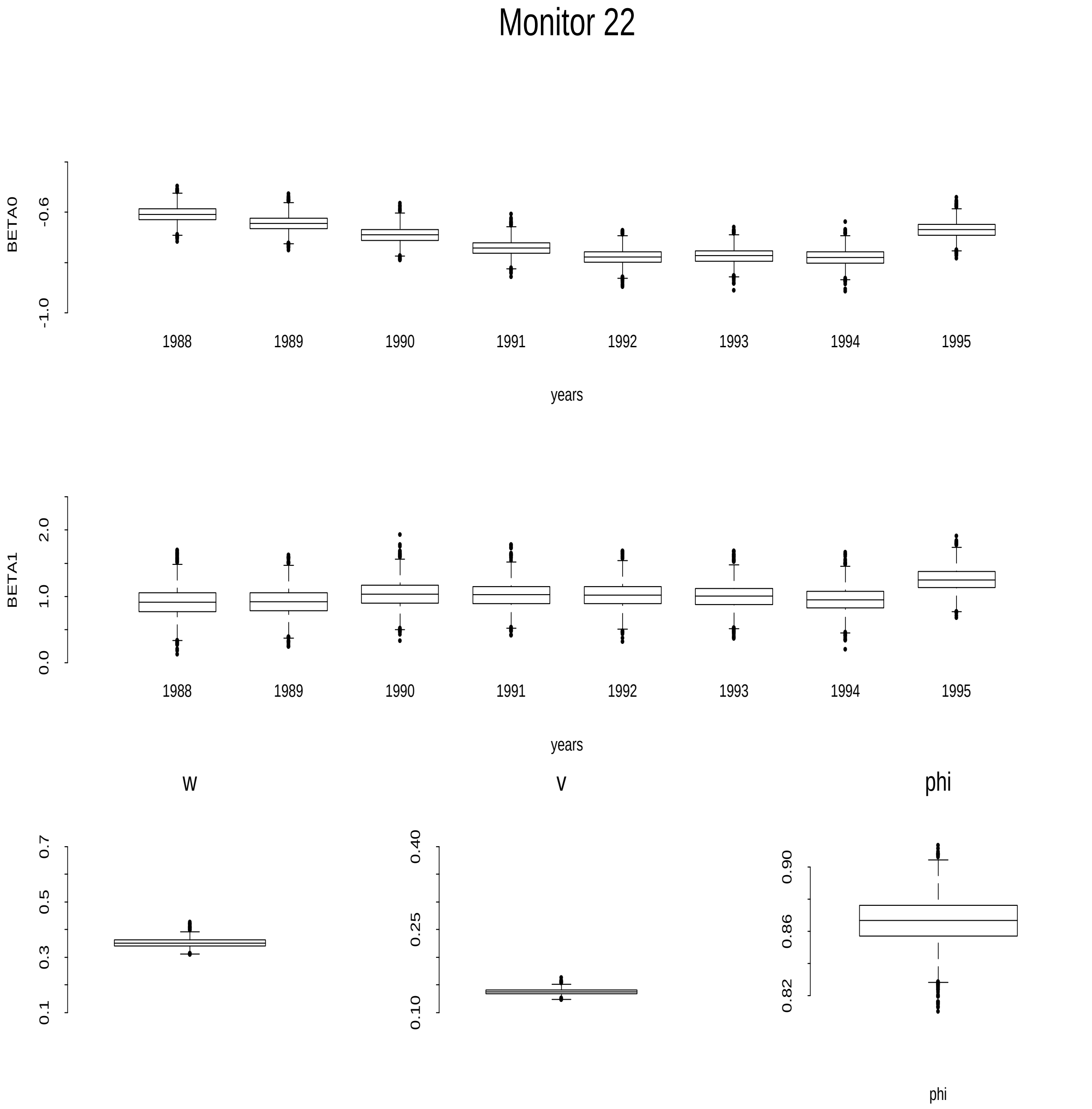


Figure 20: `MODEL-2`: Multi-year analysis of single Monitor 22: Summary interval estimates of β_{0t}, β_{1t}, standard deviations w, v and correlation parameter ϕ.

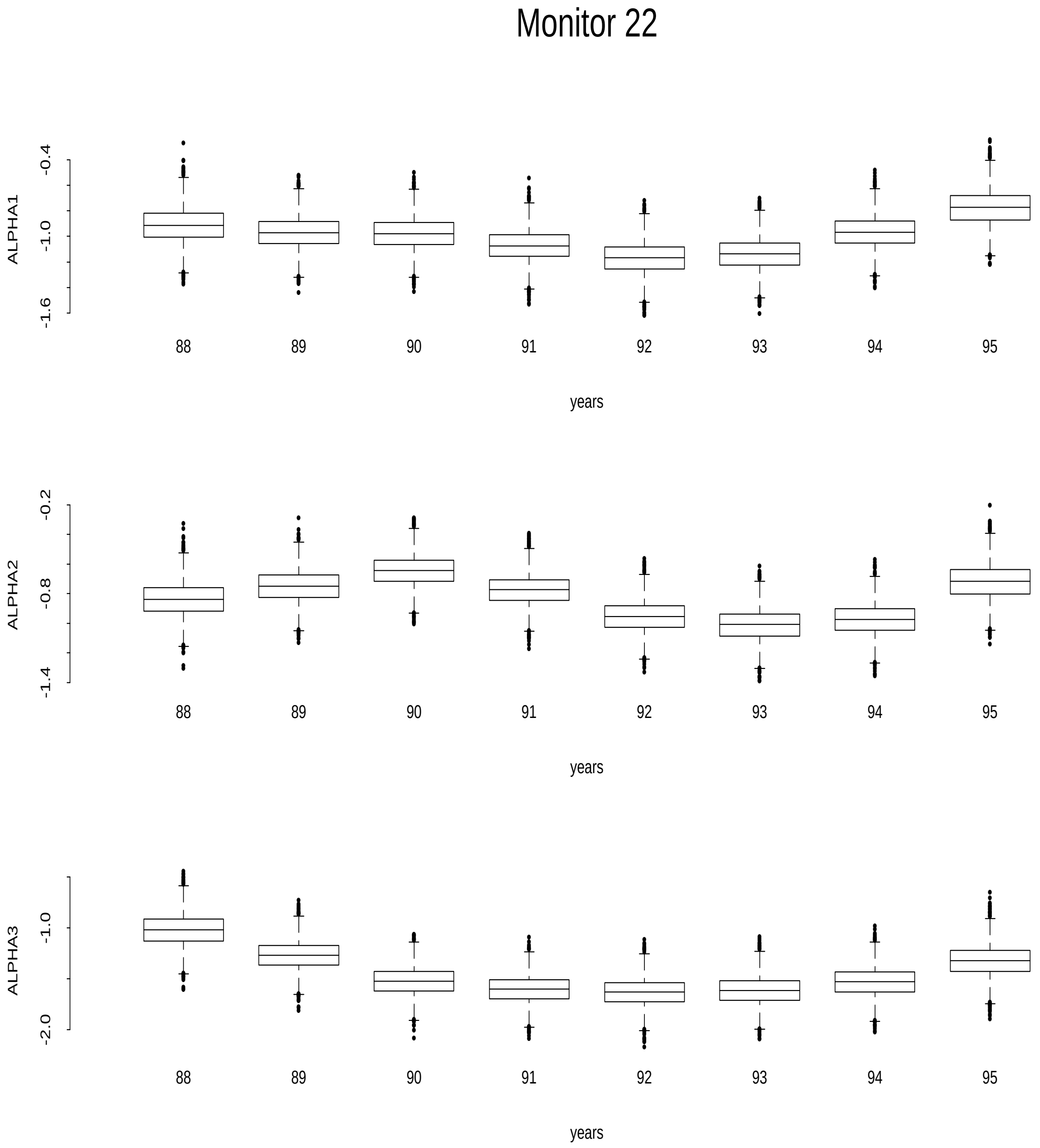


Figure 21: `MODEL-2`: Multi-year analysis of single Monitor 22: Summary interval estimates of hospital-specific parameters α_{it} for three hospitals.

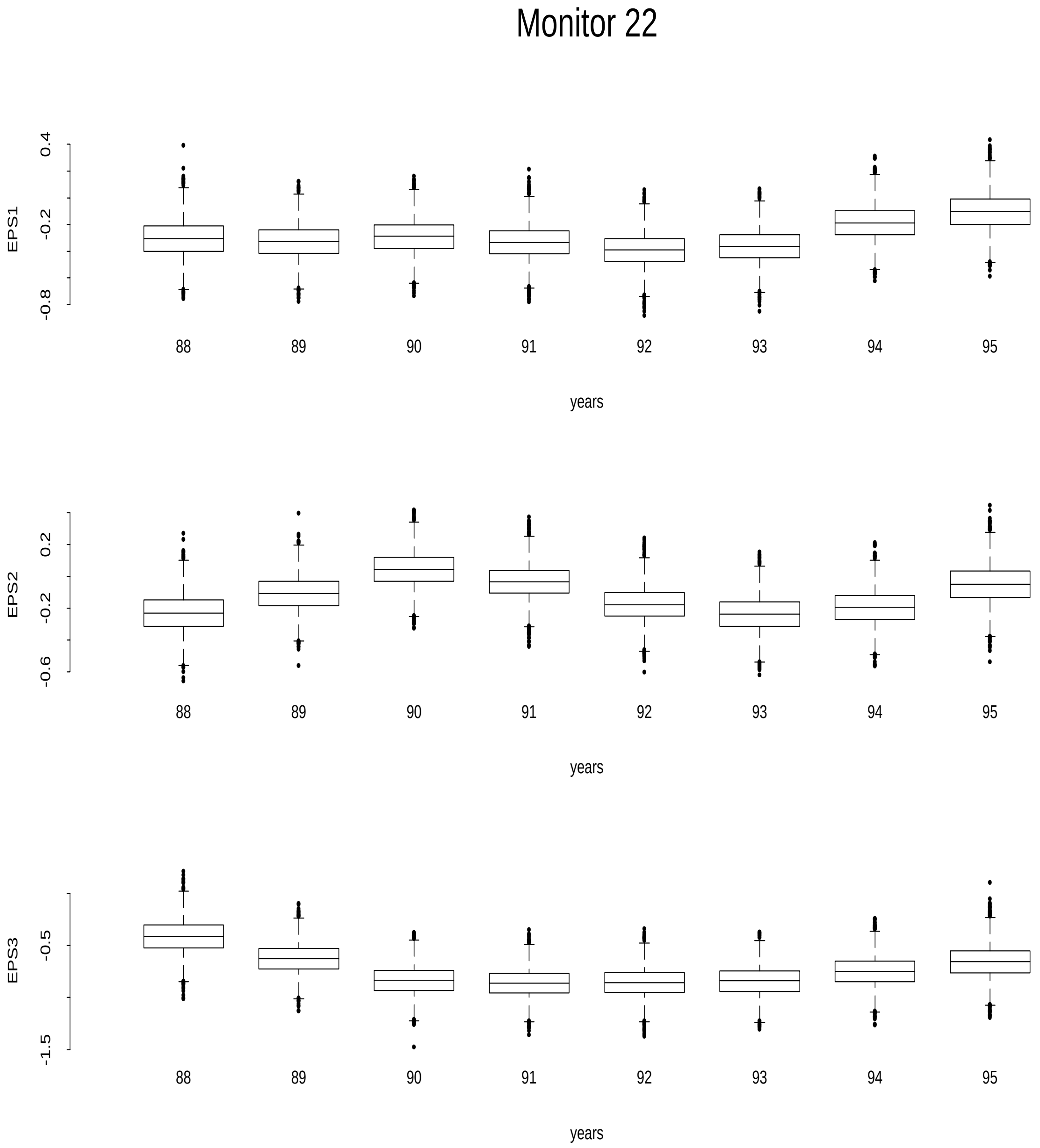


Figure 22: `MODEL-2`: Multi-year analysis of single Monitor 22: Summary interval estimates of hospital-specific random effects ϵ_{it} for three hospitals.

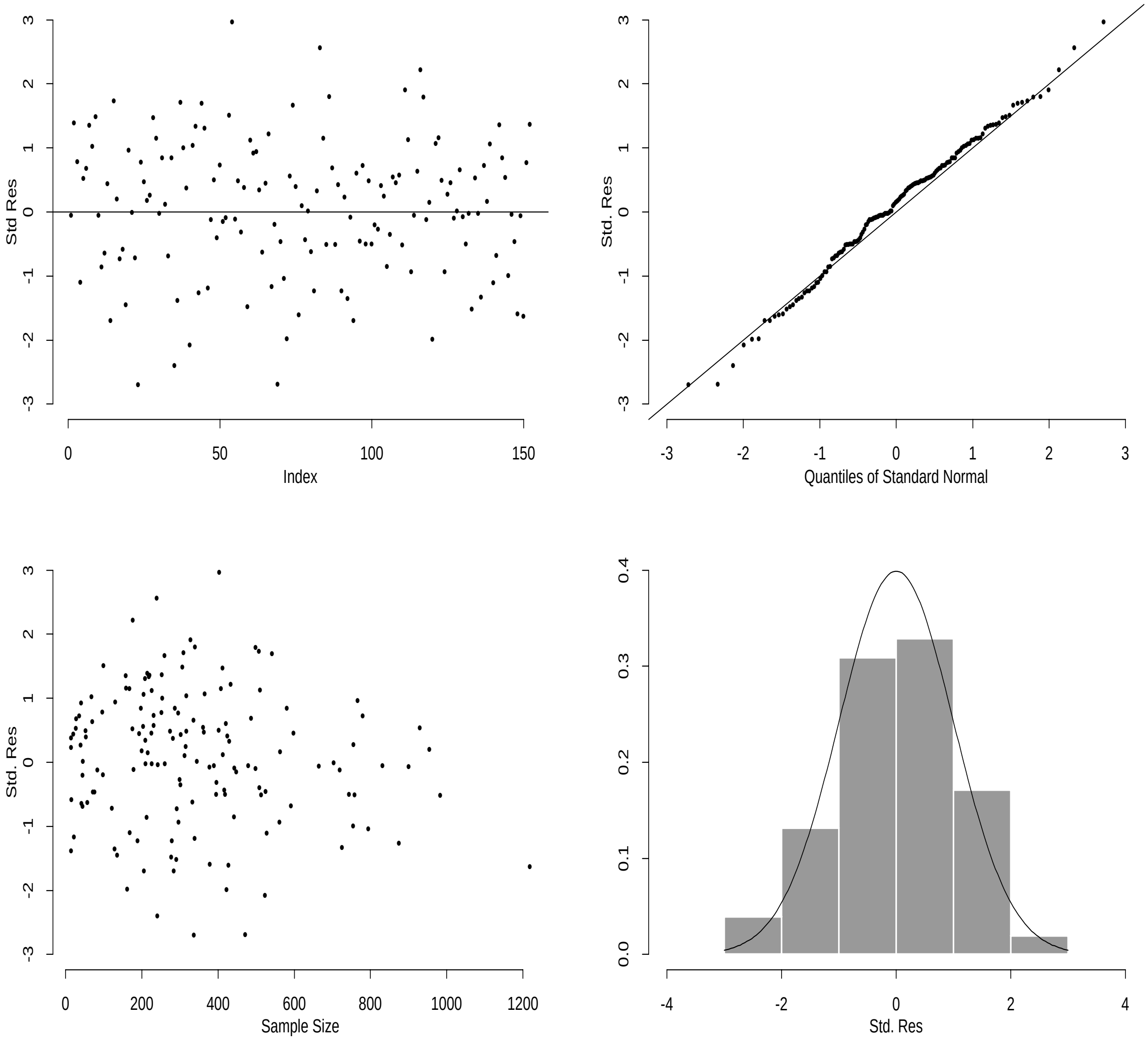


Figure 23: `MODEL-2`: Summaries of a randomly chosen draw from the posterior distribution of the observation residuals e_{it} for all hospitals i on M21 in year $t = 1995$.

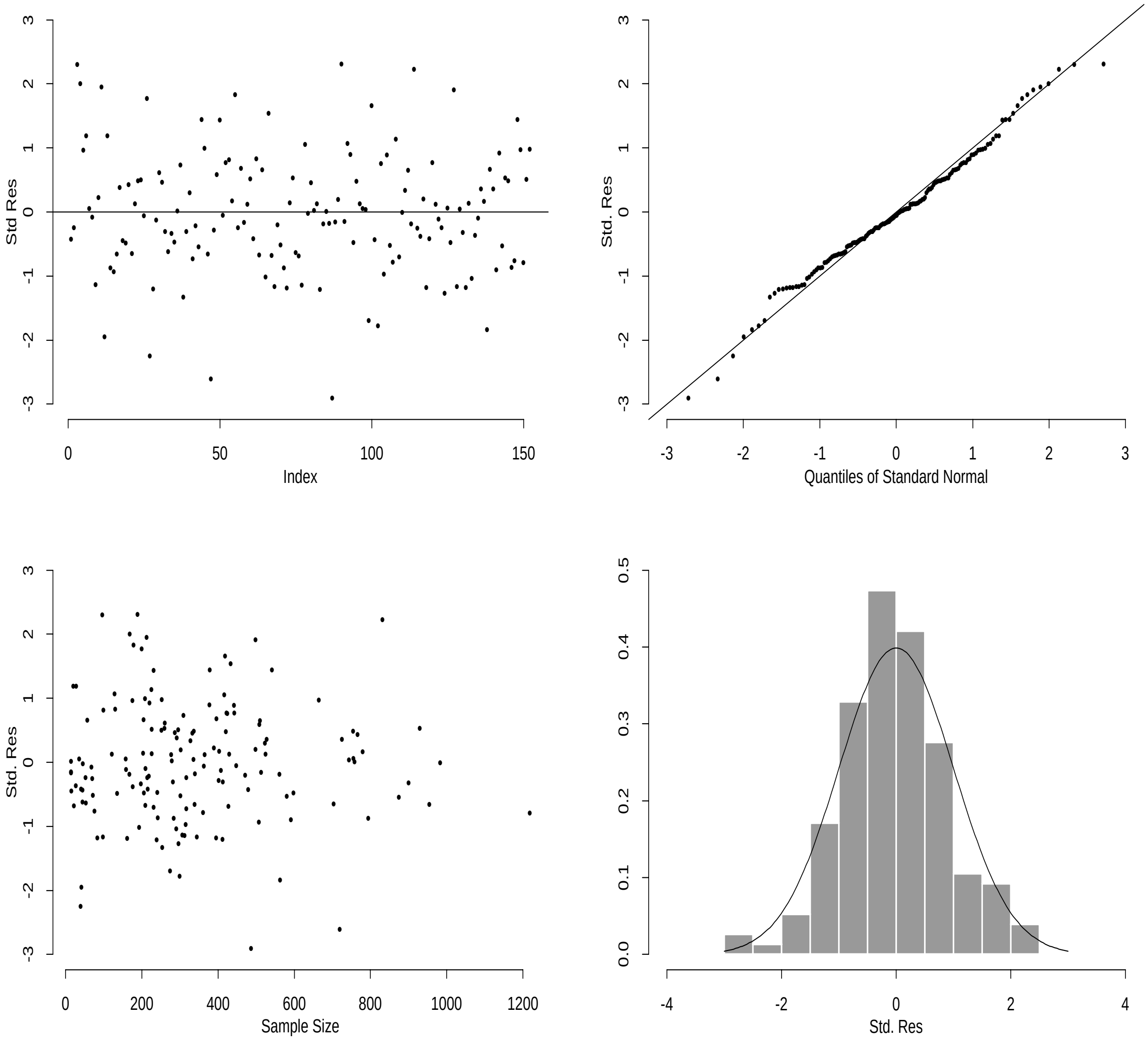


Figure 24: MODEL-2: Summaries of a second randomly chosen draw from the posterior distribution of the observation residuals e_{it} for all hospitals i on M21 in year $t = 1995$.

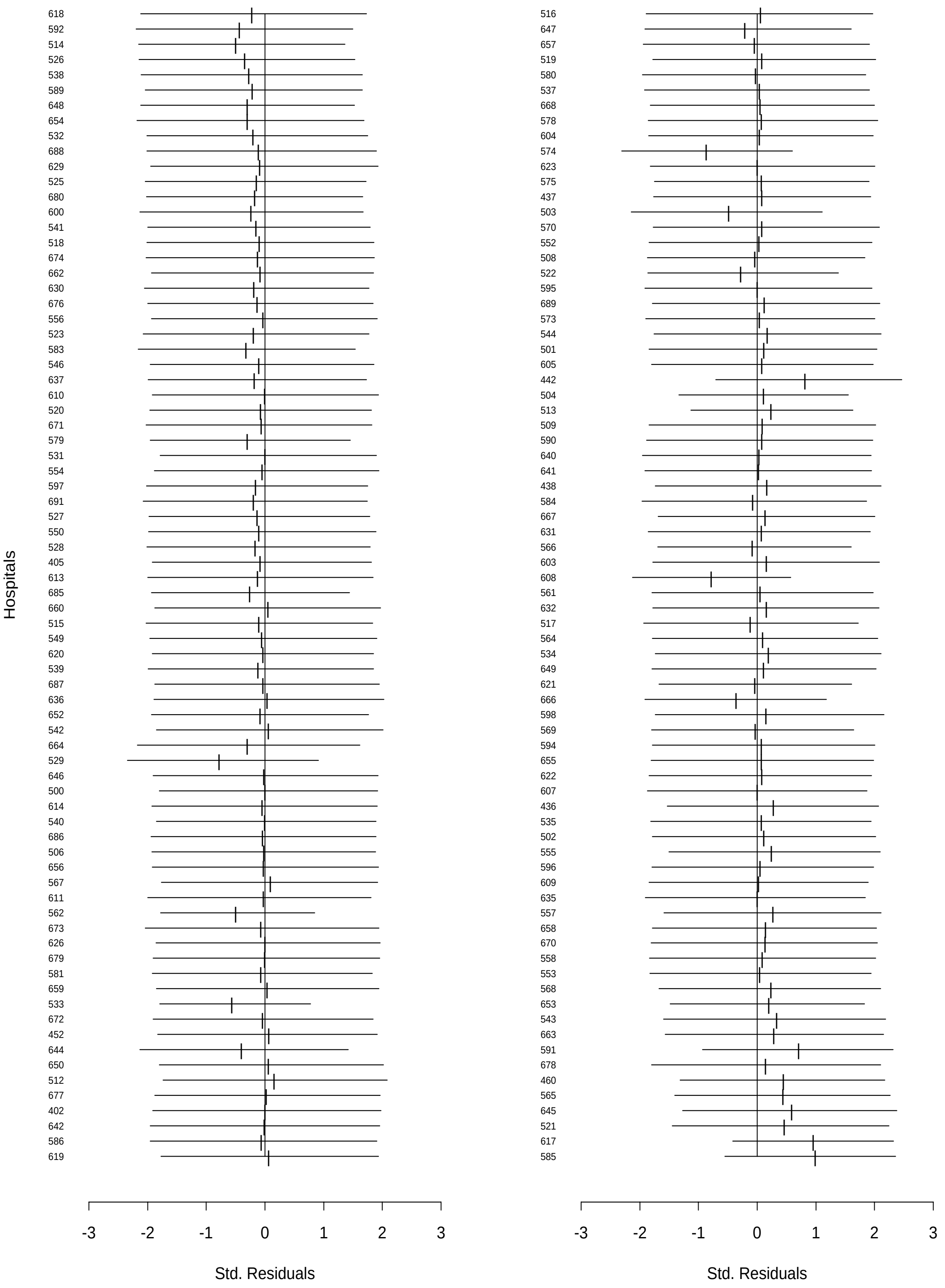


Figure 25: `MODEL-2`: Posterior 95% intervals, with means marked, for the observation residuals e_{it} for all hospitals i on M21 in year $t = 1995$. The hospitals are ordered according to the posterior medians of the underlying outcome probabilities p_{it}.

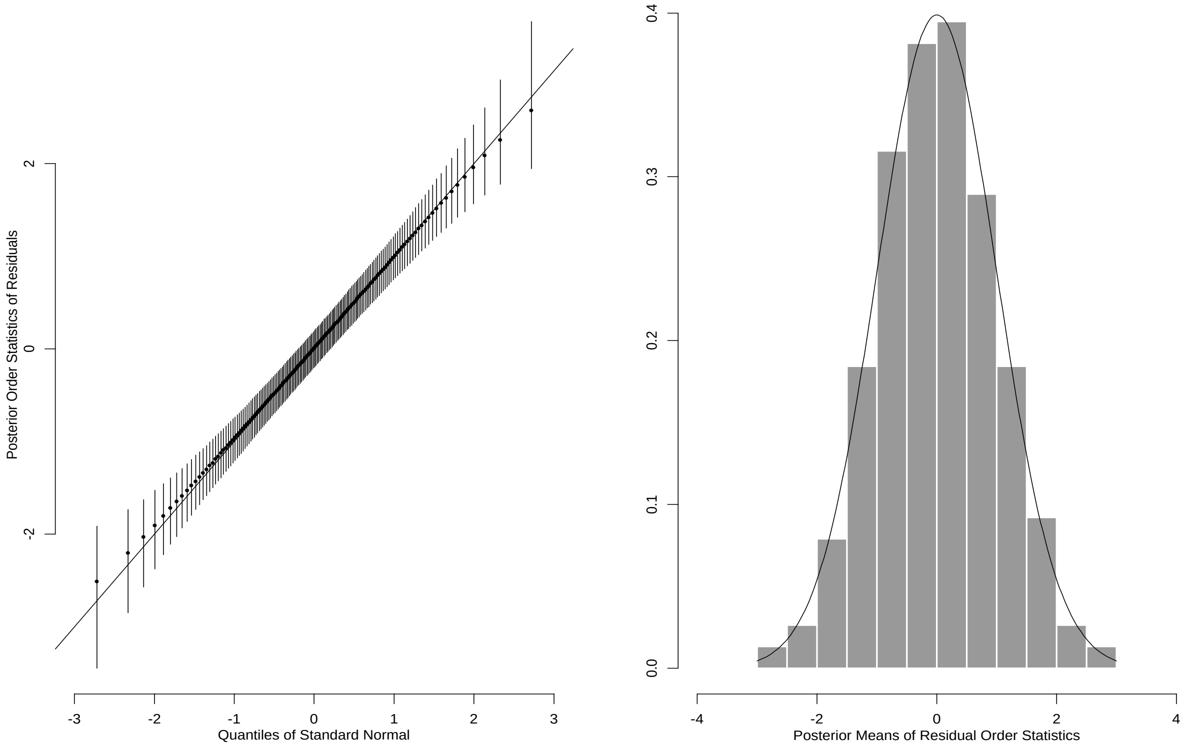


Figure 26: `MODEL-2`: Normal quantile plot, with approximate 95% intervals shown, and histogram of the posterior means of the *ordered* observation residuals across all hospitals i on M21 in year $t = 1995$.

8.4 `MODEL-3`: Random effects multiple monitor time series models

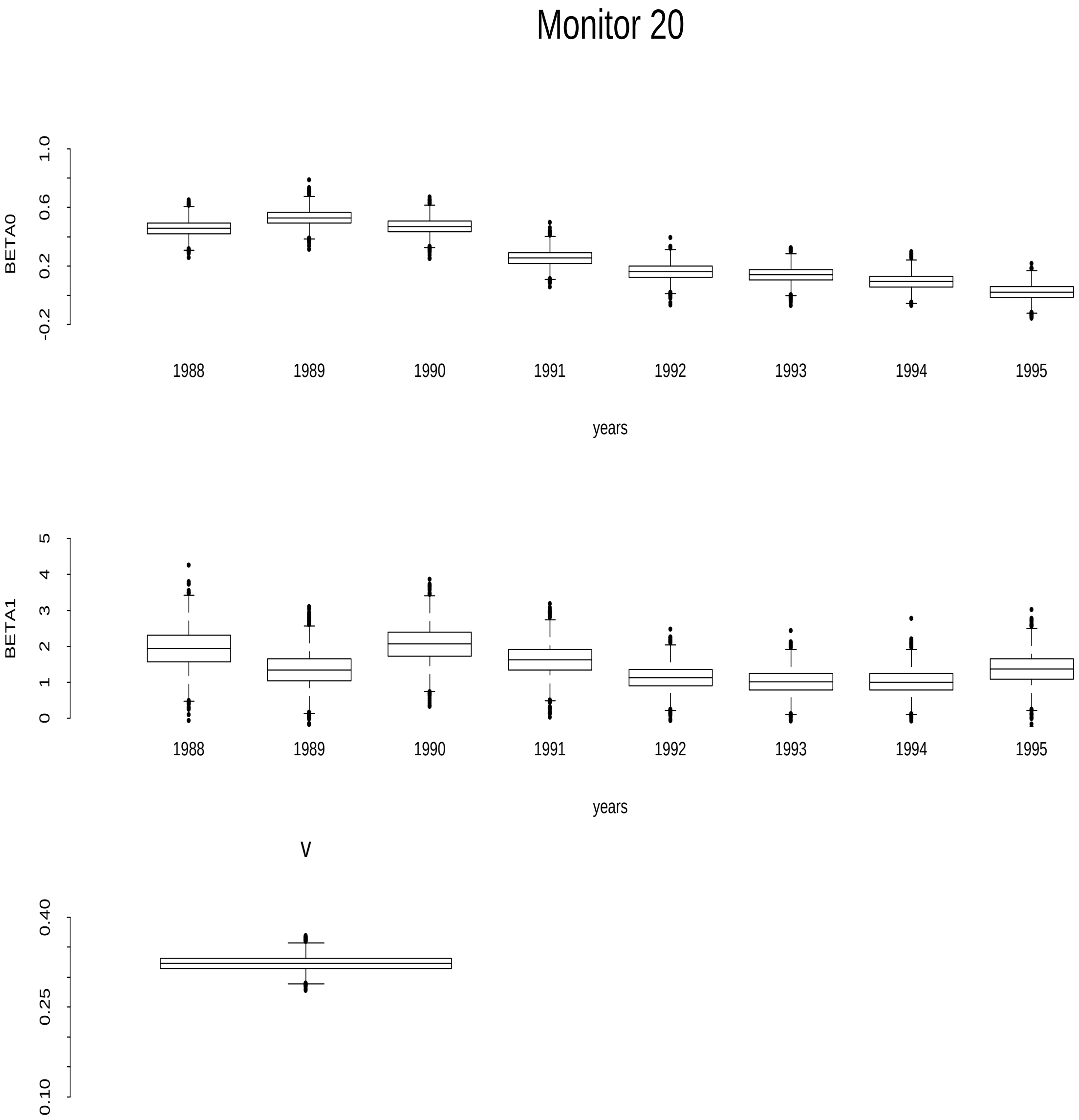


Figure 27: `MODEL-3`: Monitor 20: Summary interval estimates of β_{01t}, β_{11t} and standard deviation v_1.

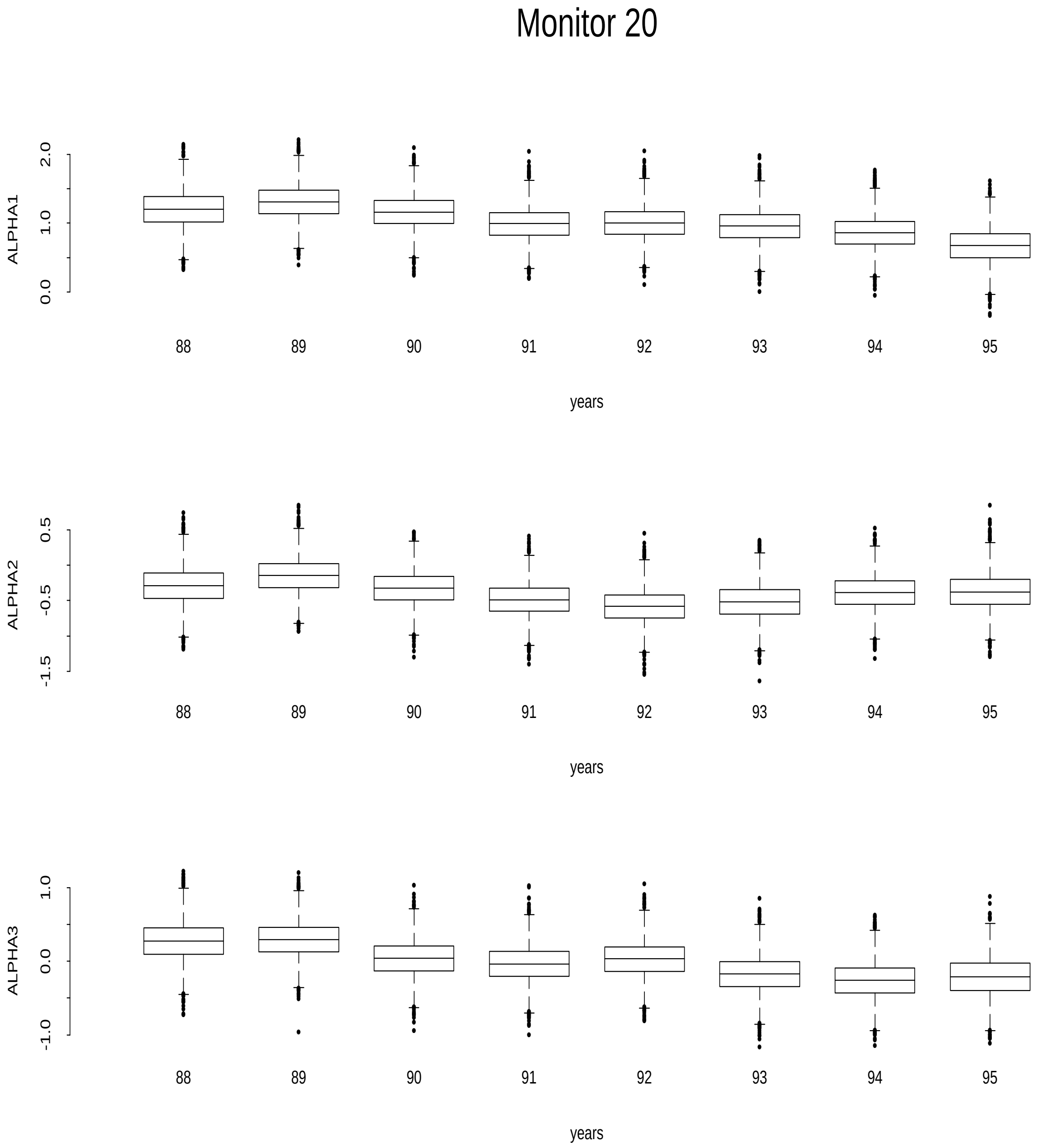


Figure 28: MODEL-3: Monitor 20: Summary interval estimates of hospital-specific parameters α_{i1t} for three hospitals.

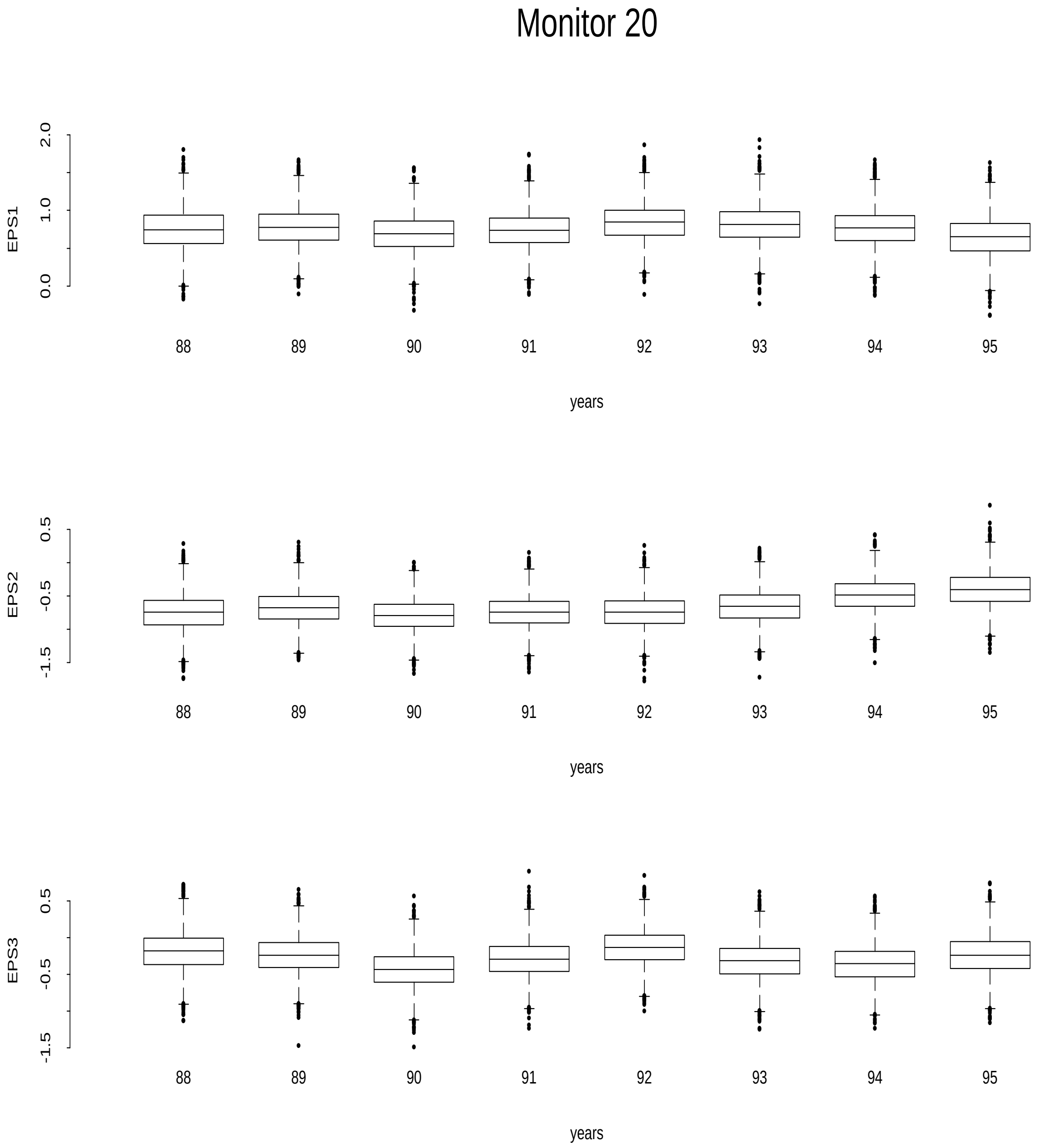


Figure 29: `MODEL-3`: Monitor 20: Summary interval estimates of hospital-specific random effects ϵ_{i1t} for three hospitals.

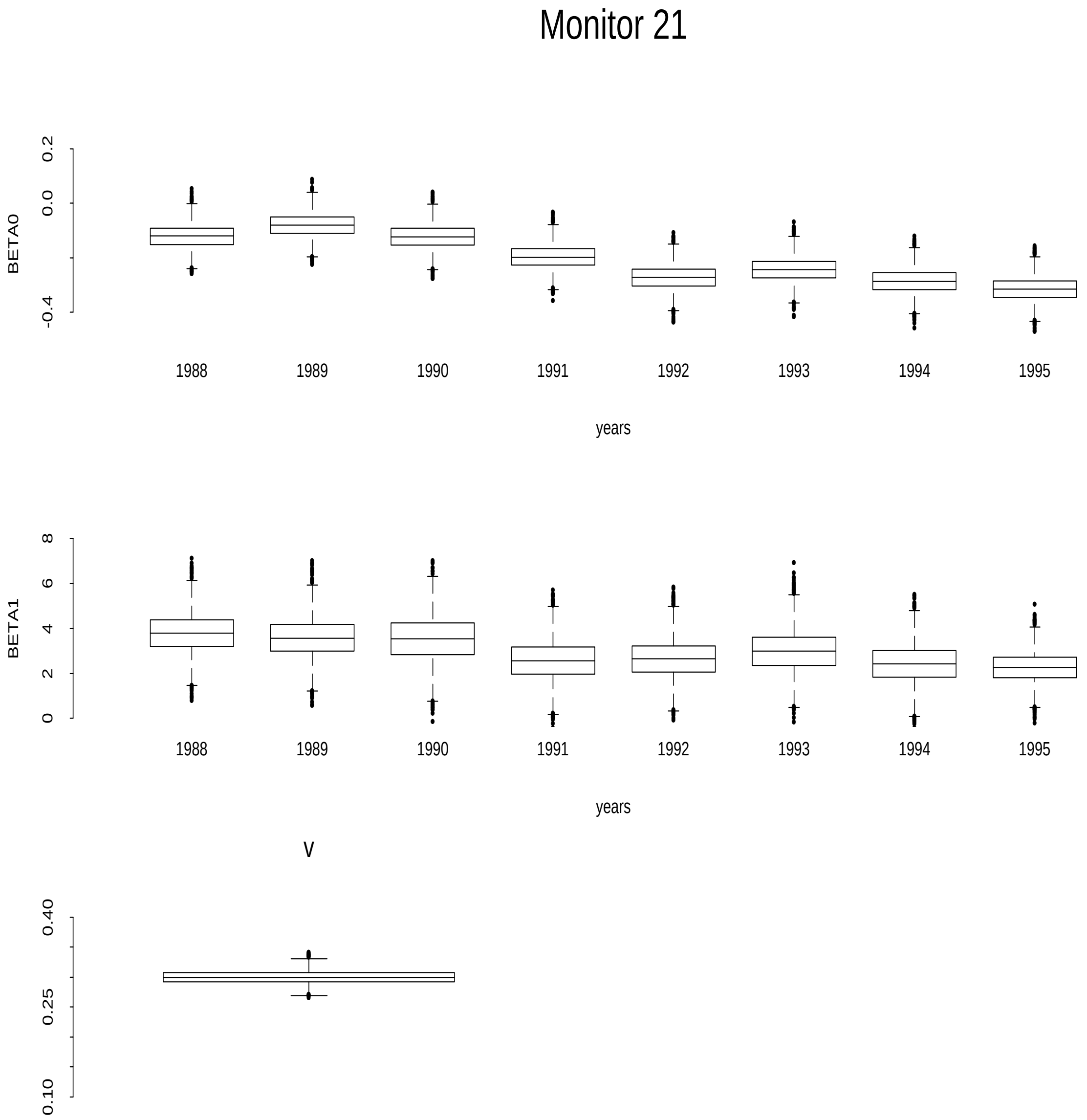


Figure 30: `MODEL-3`: Monitor 21: Summary interval estimates of β_{02t}, β_{12t} and standard deviation v_2.

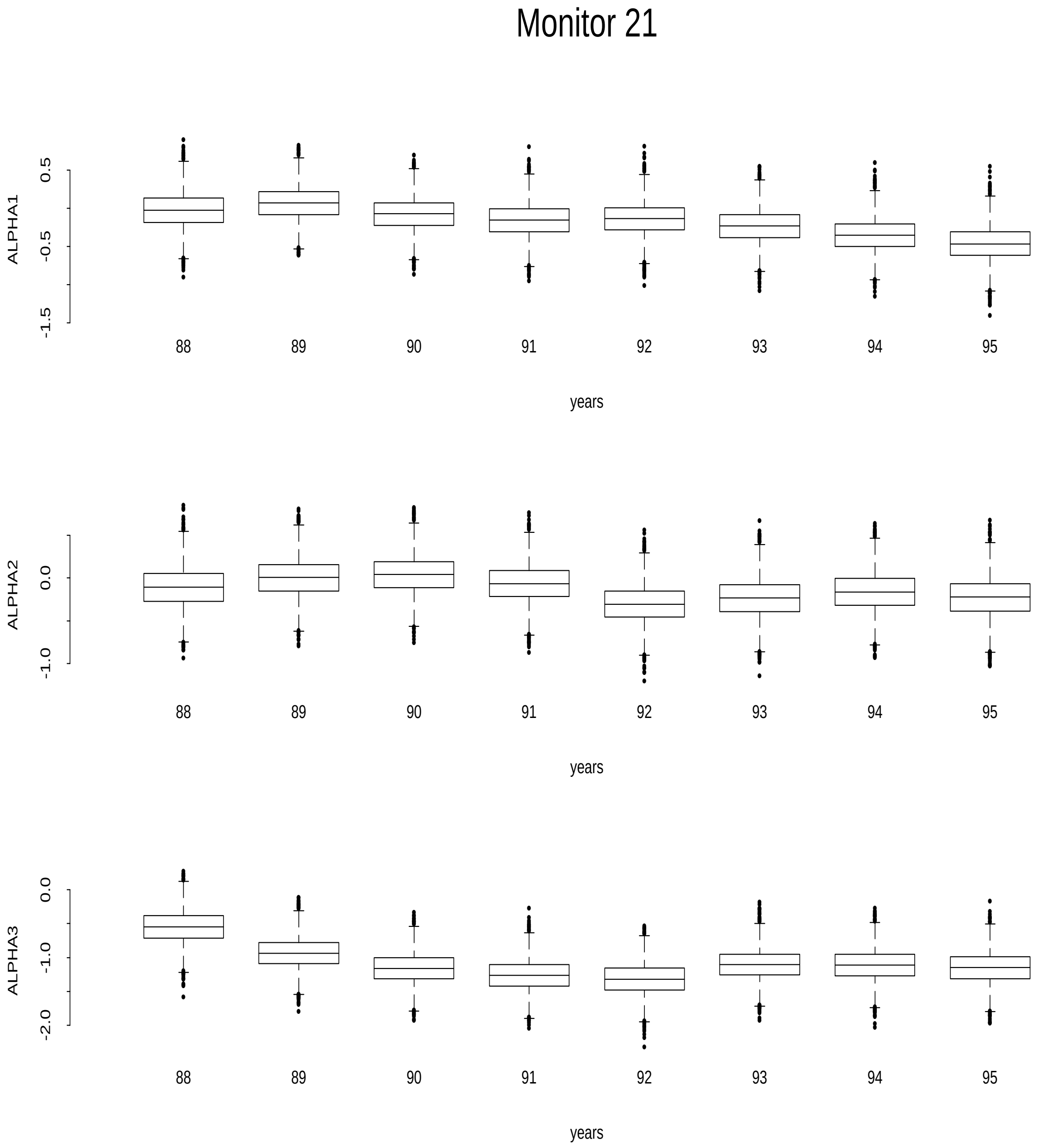


Figure 31: `MODEL-3`: Monitor 21: Summary interval estimates of hospital-specific parameters α_{i2t} for three hospitals.

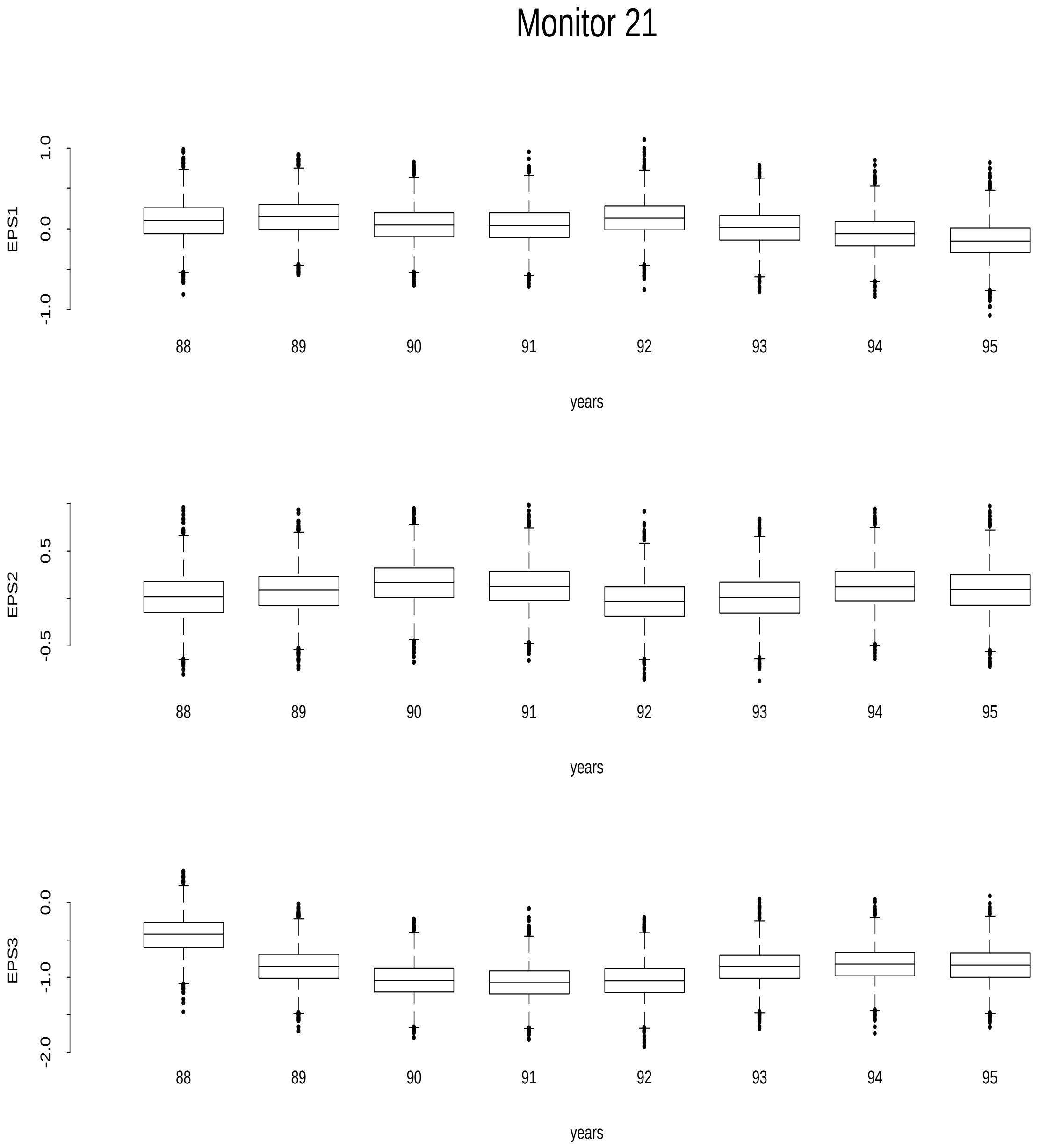


Figure 32: `MODEL-3`: Monitor 21: Summary interval estimates of hospital-specific random effects ϵ_{i2t} for three hospitals.

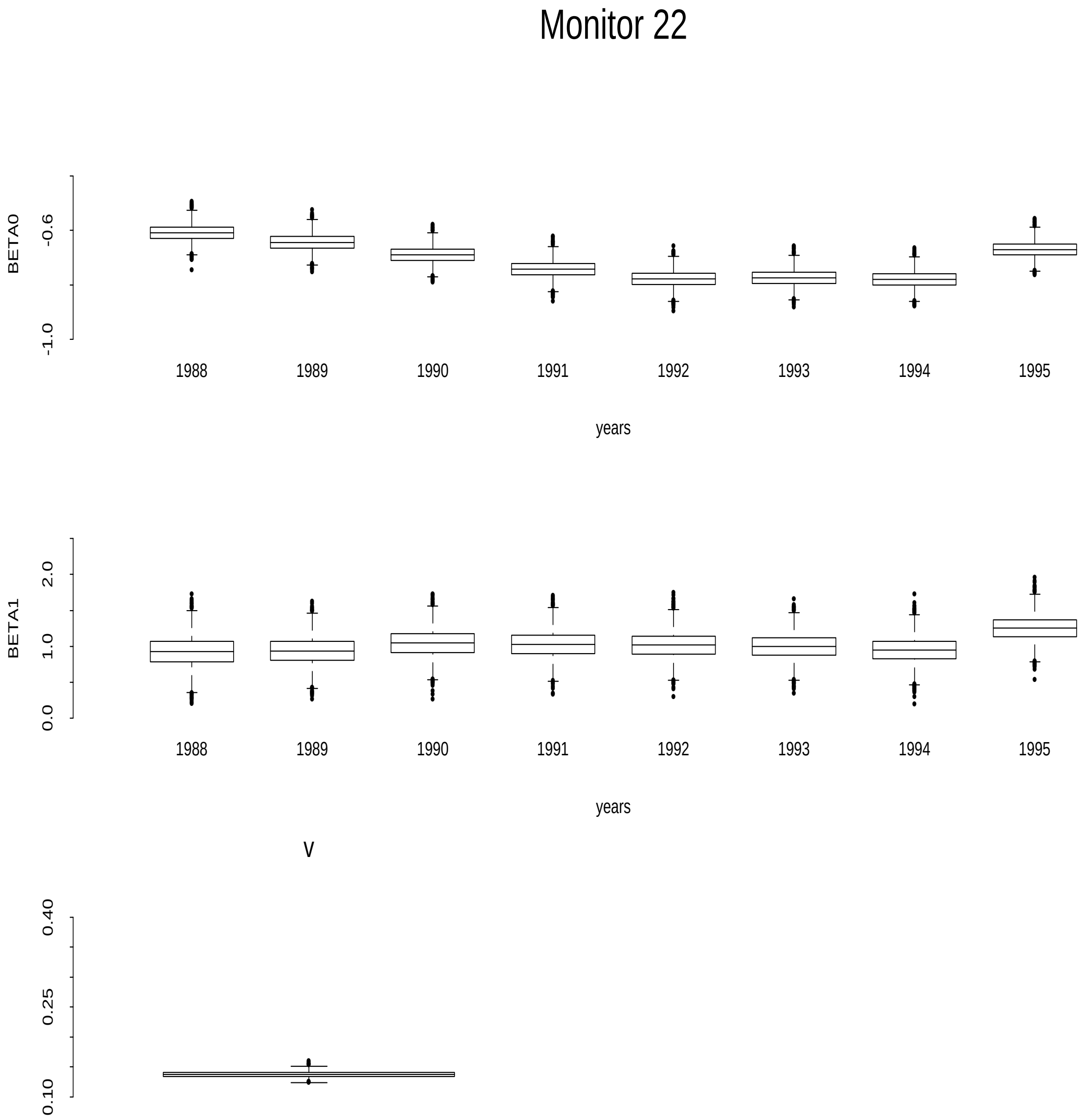


Figure 33: MODEL-3: Monitor 22: Summary interval estimates of β_{03t}, β_{13t} and standard deviation v_3.

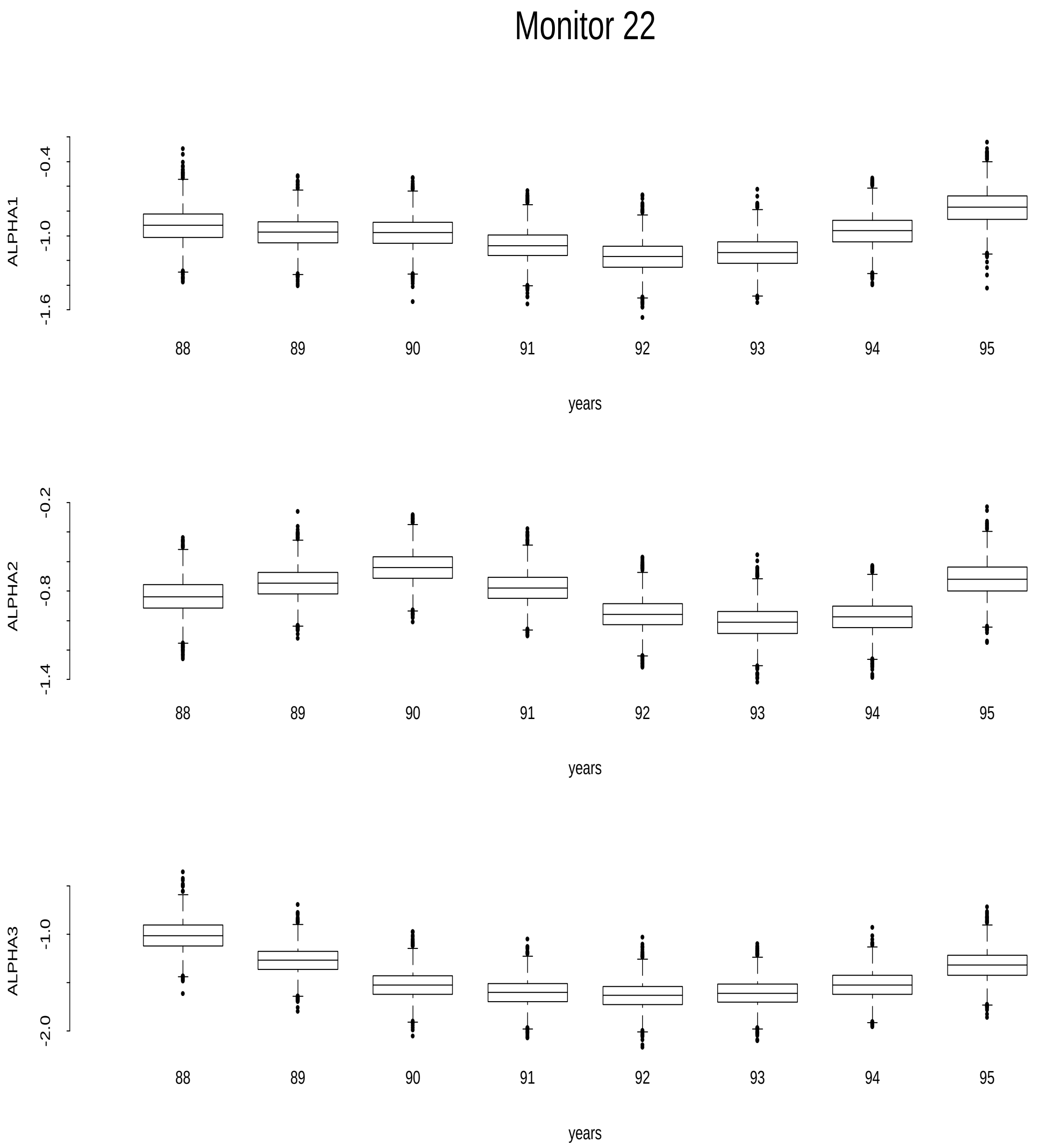


Figure 34: `MODEL-3`: Summary interval estimates of hospital-specific parameters α_{i3t} for three hospitals.

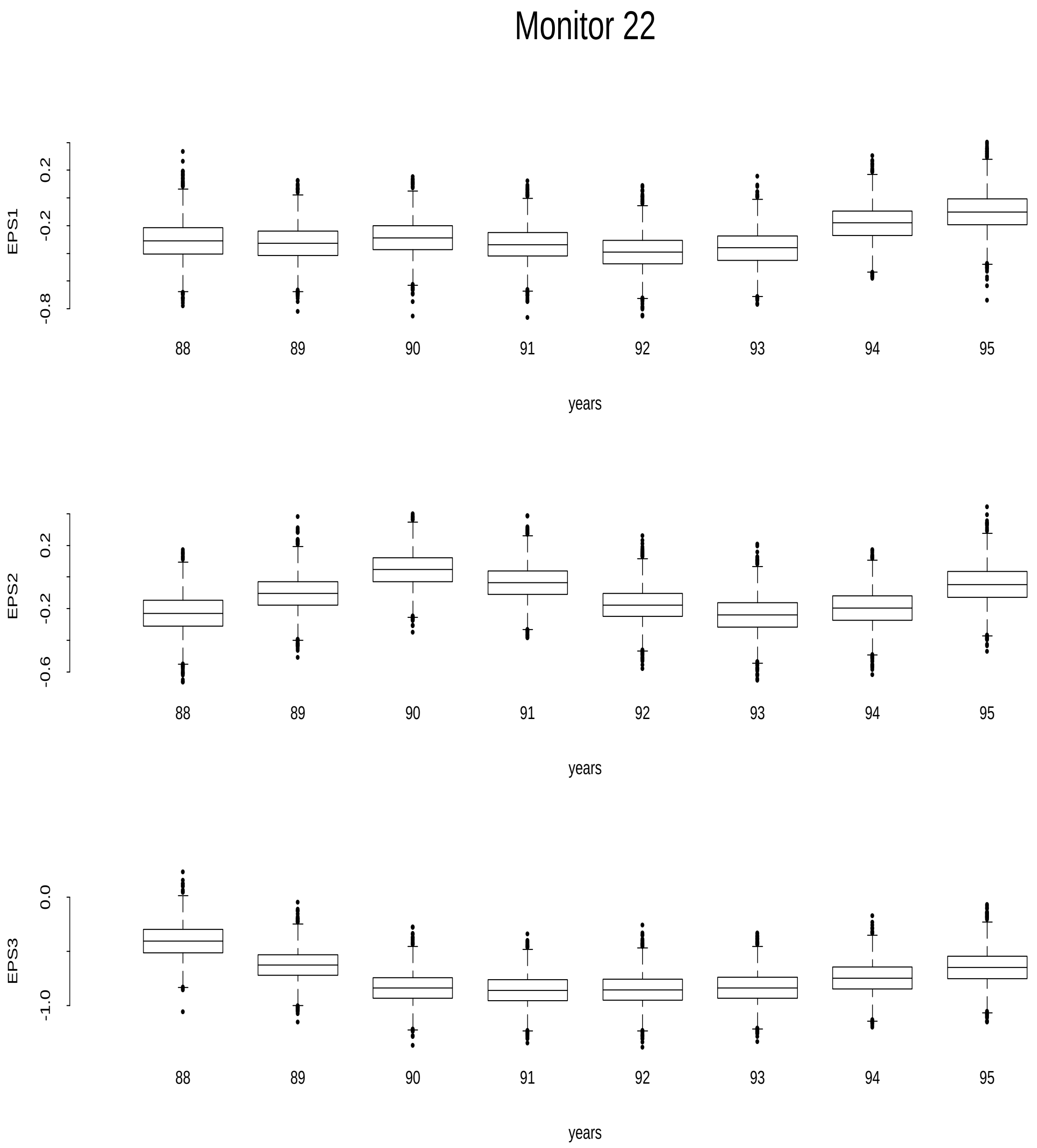


Figure 35: `MODEL-3`: Summary interval estimates of hospital-specific random effects ϵ_{i3t} for three hospitals.

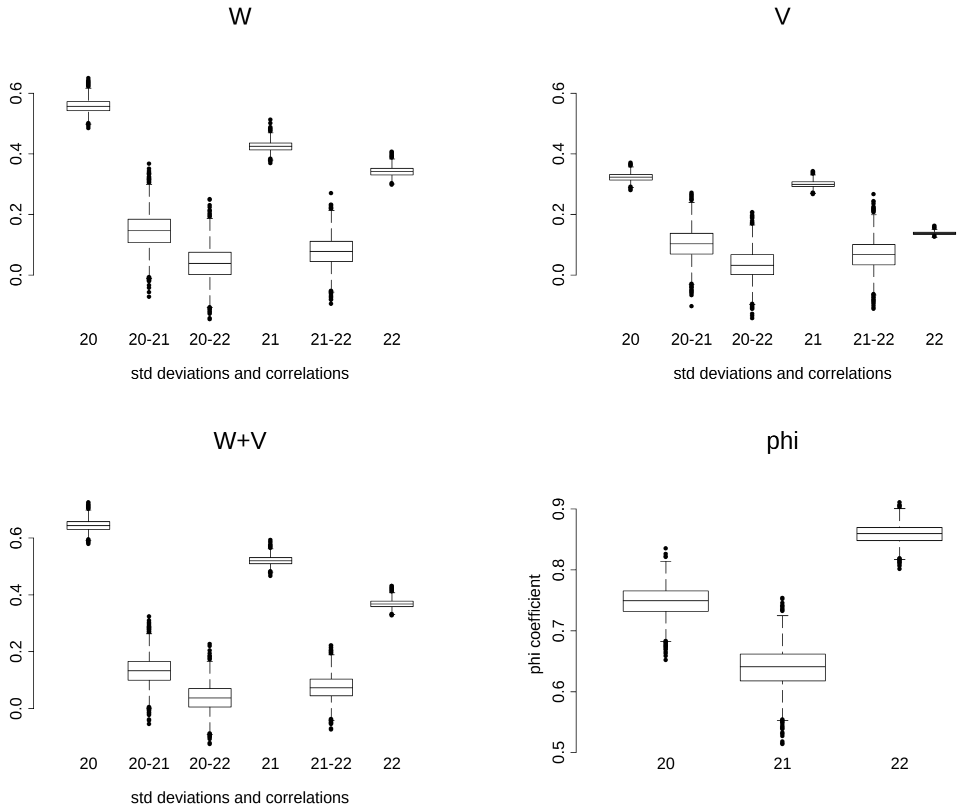


Figure 36: `MODEL-3`: Multi-monitor time series model for all three monitors: Summary interval estimates of population parameters, including standard deviations and correlations of the variance matrice $\mathbf{W}, \mathbf{V}$ and the overall $\mathbf{W}+\mathbf{V}$, and the three AR coefficients ϕ_j.

8.5 Summary inferences for M21 in 1995

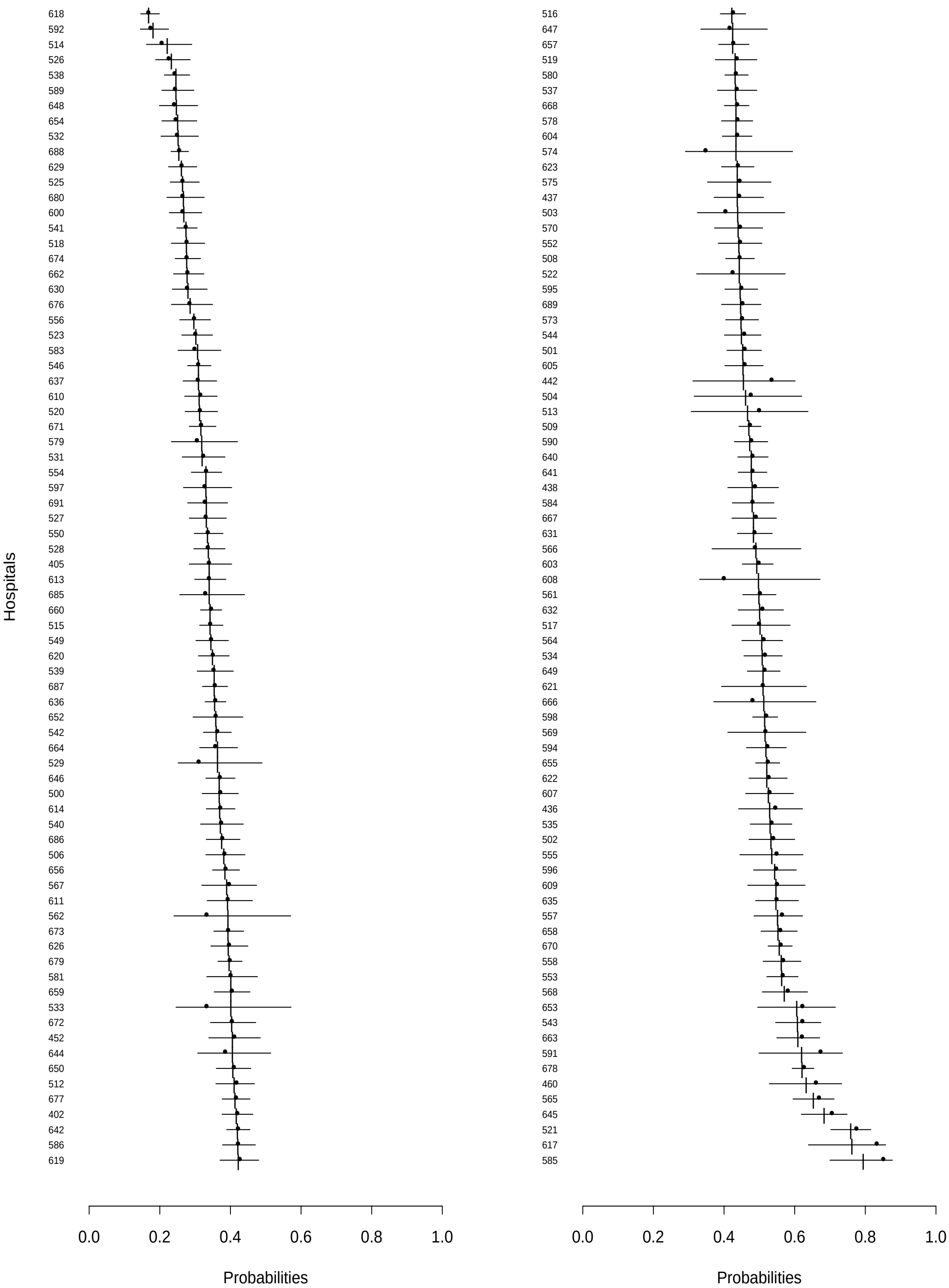


Figure 37: Posterior 95% intervals for outcome probabilities p_i on M21 in 1995 for all hospitals, ordered by posterior medians. Station number identifies hospitals and observed proportions are marked.

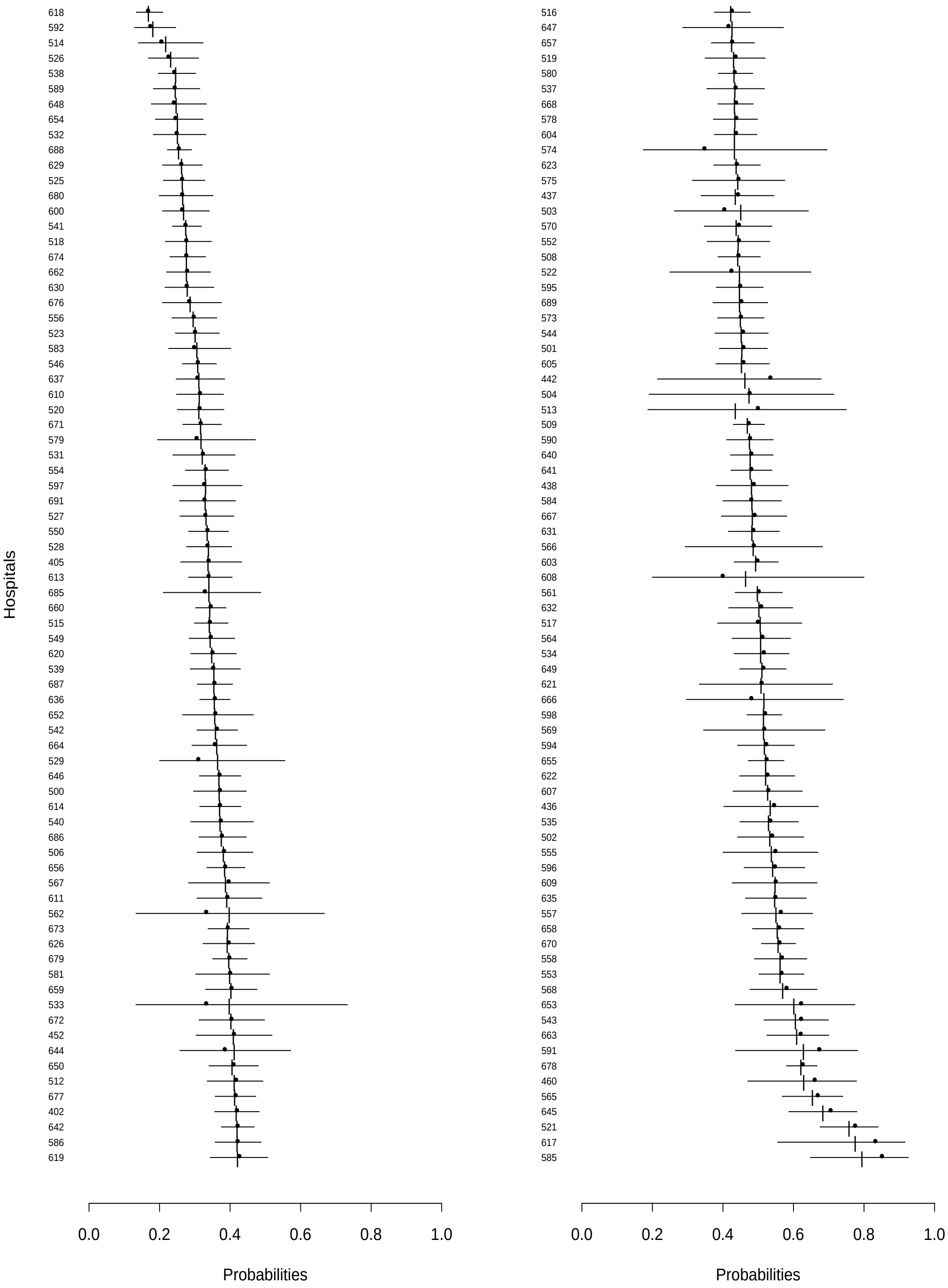


Figure 38: Posterior 95% intervals for actual outcome proportions on M21 in 1995 for all hospitals. Station number identifies hospitals and actually observed proportions are marked.

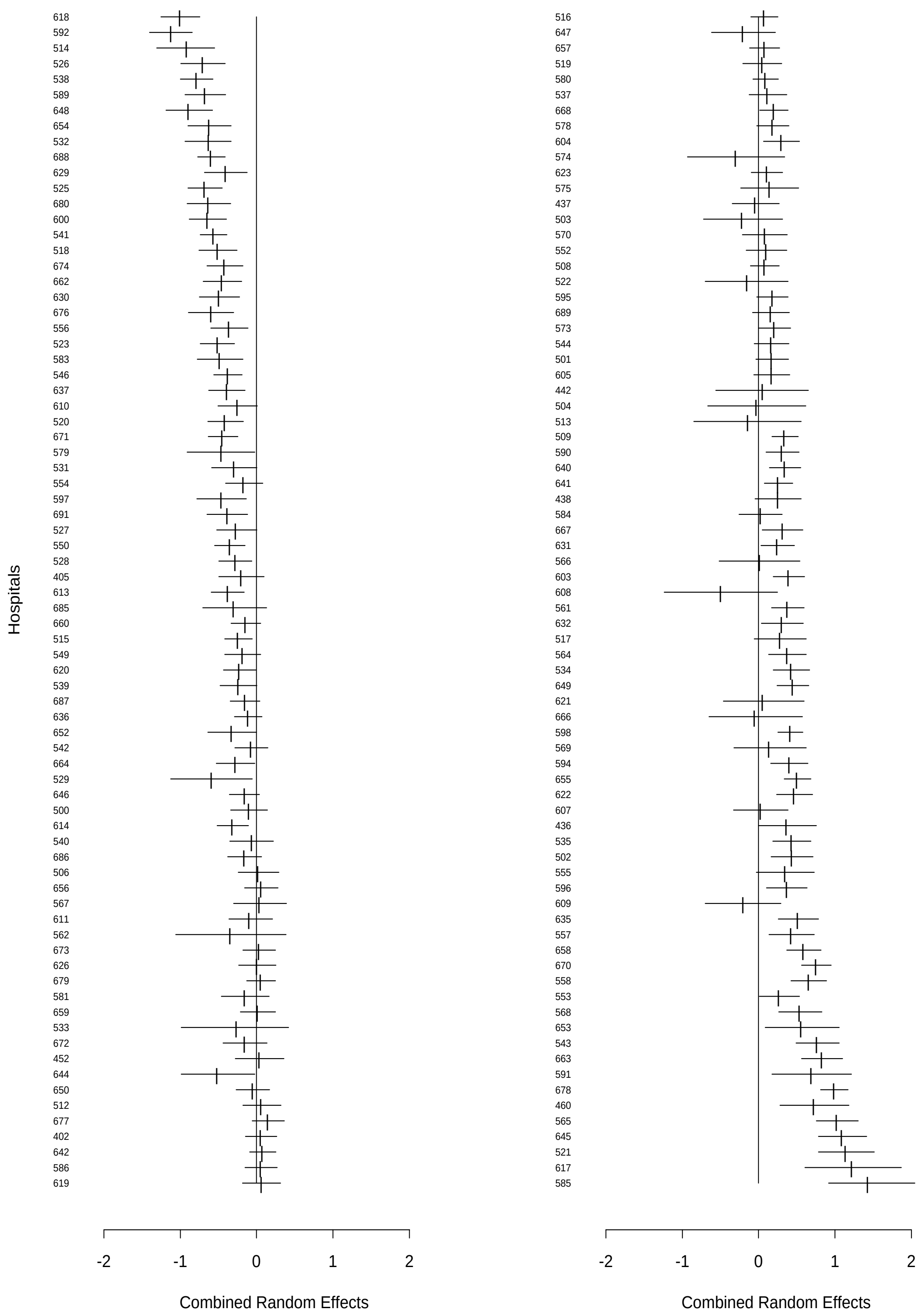


Figure 39: Posterior 95% intervals for combined random effects $\epsilon_i + \nu_i$ on M21 in 1995 over all hospitals.

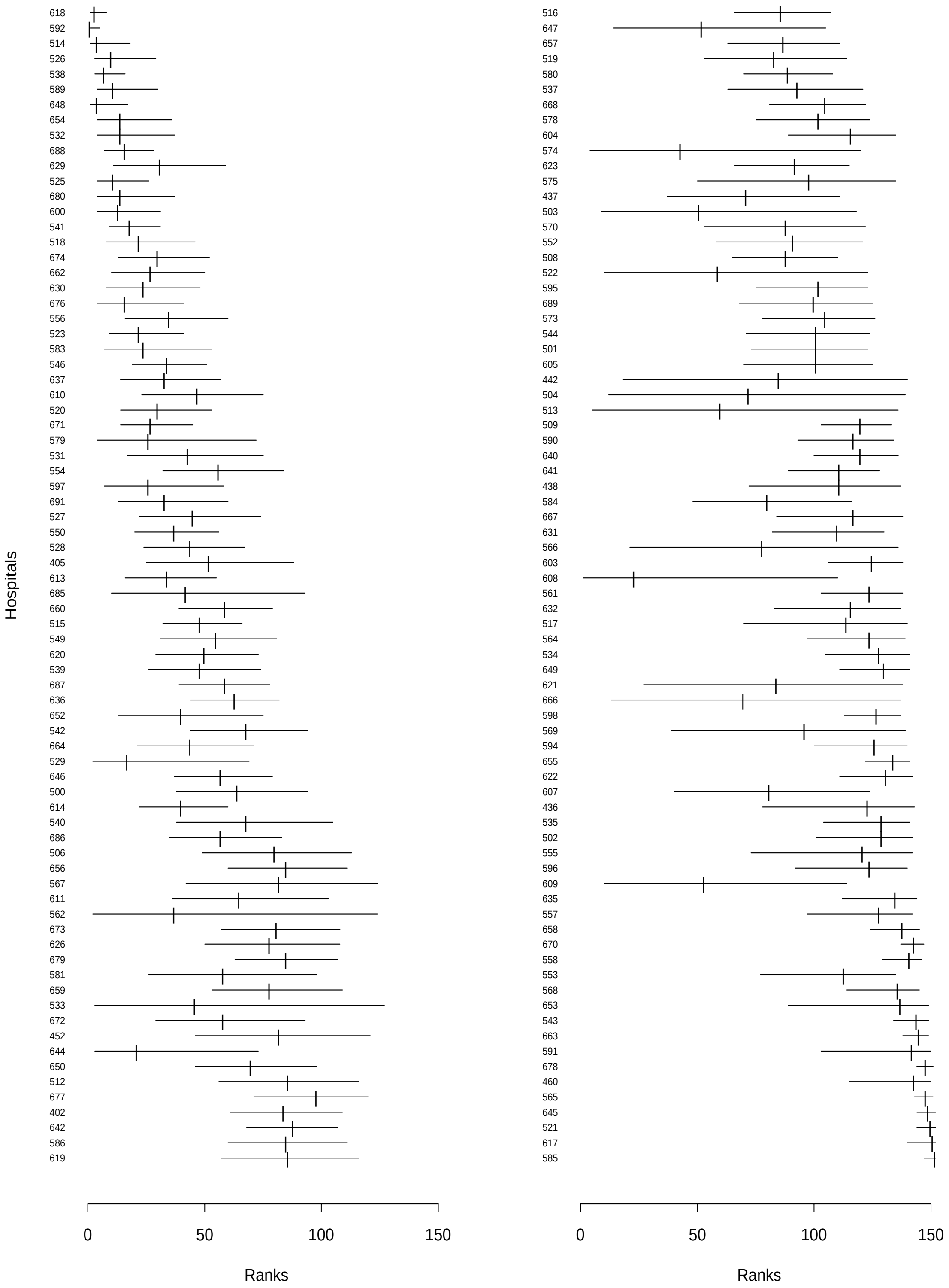


Figure 40: Posterior 95% intervals for ranks of hospitals in terms of combined random effects $\epsilon_i + \nu_i$ on M21 in 1995.

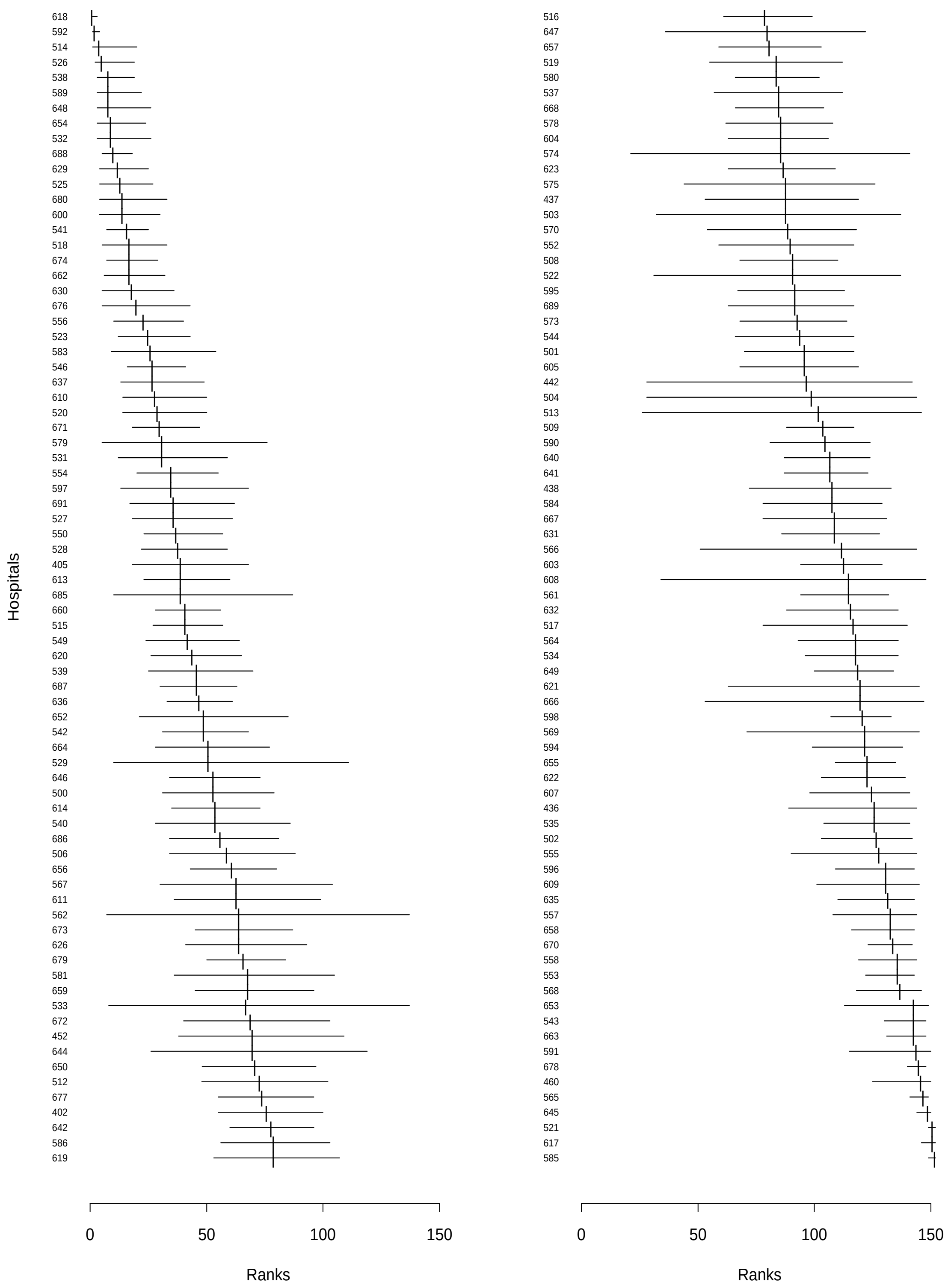


Figure 41: Posterior 95% intervals for ranks of hospitals in terms of outcome probabilities p_i on M21 in 1995.

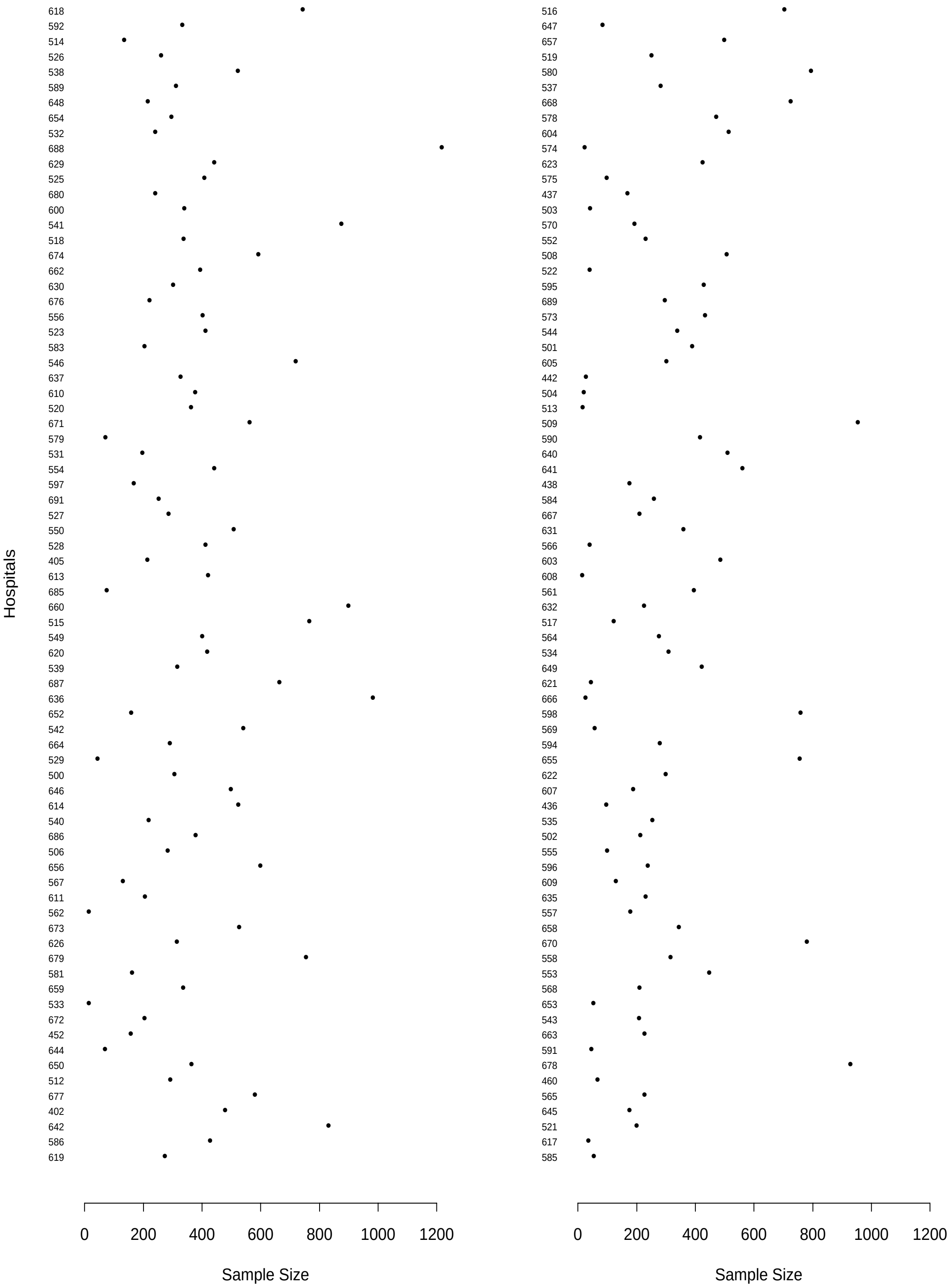


Figure 42: Observed sample sizes n_i on M21 in 1995 over all hospitals.

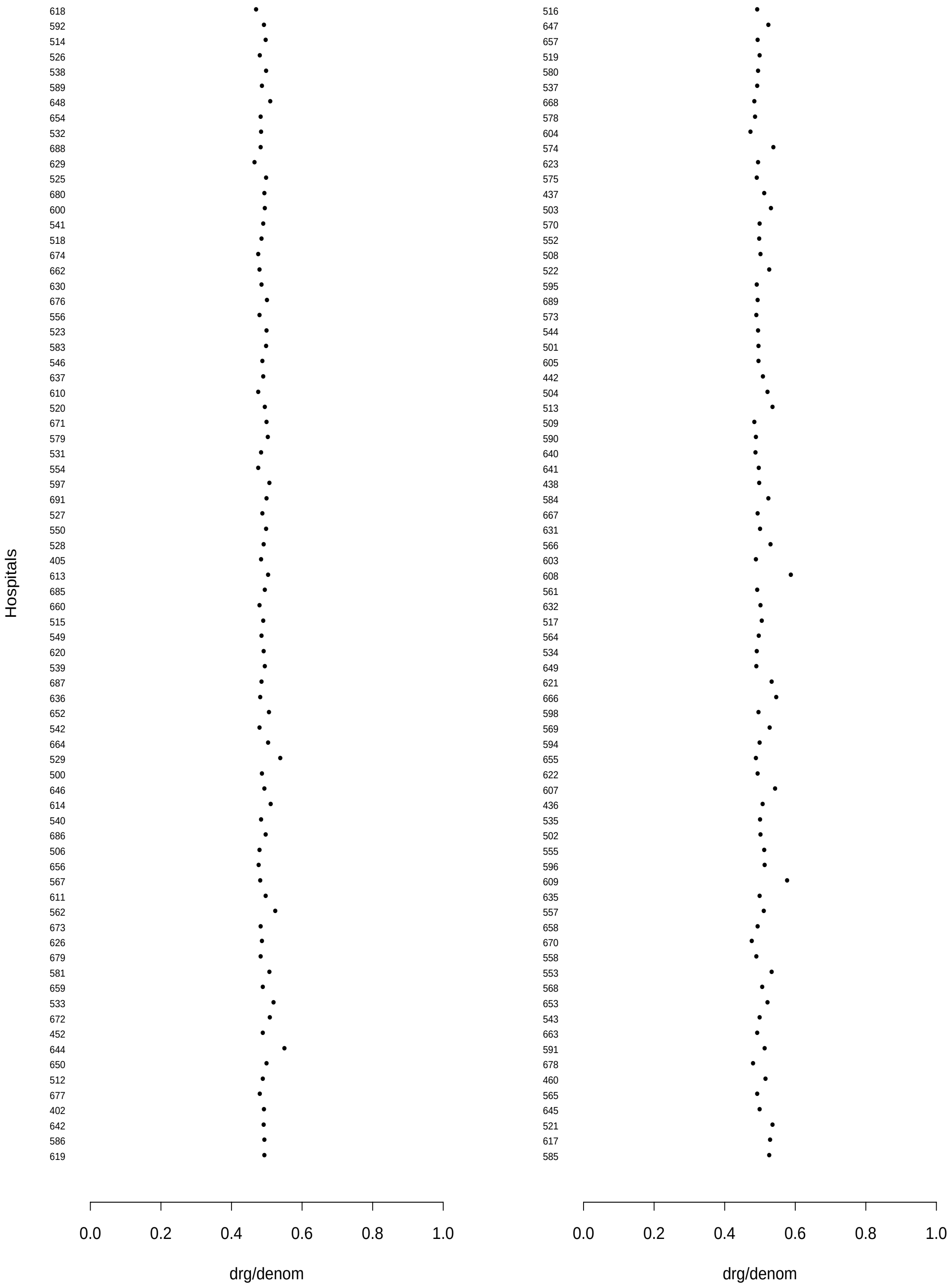


Figure 43: Observed value of DRG predicted proportion on M21 in 1995 over all hospitals.

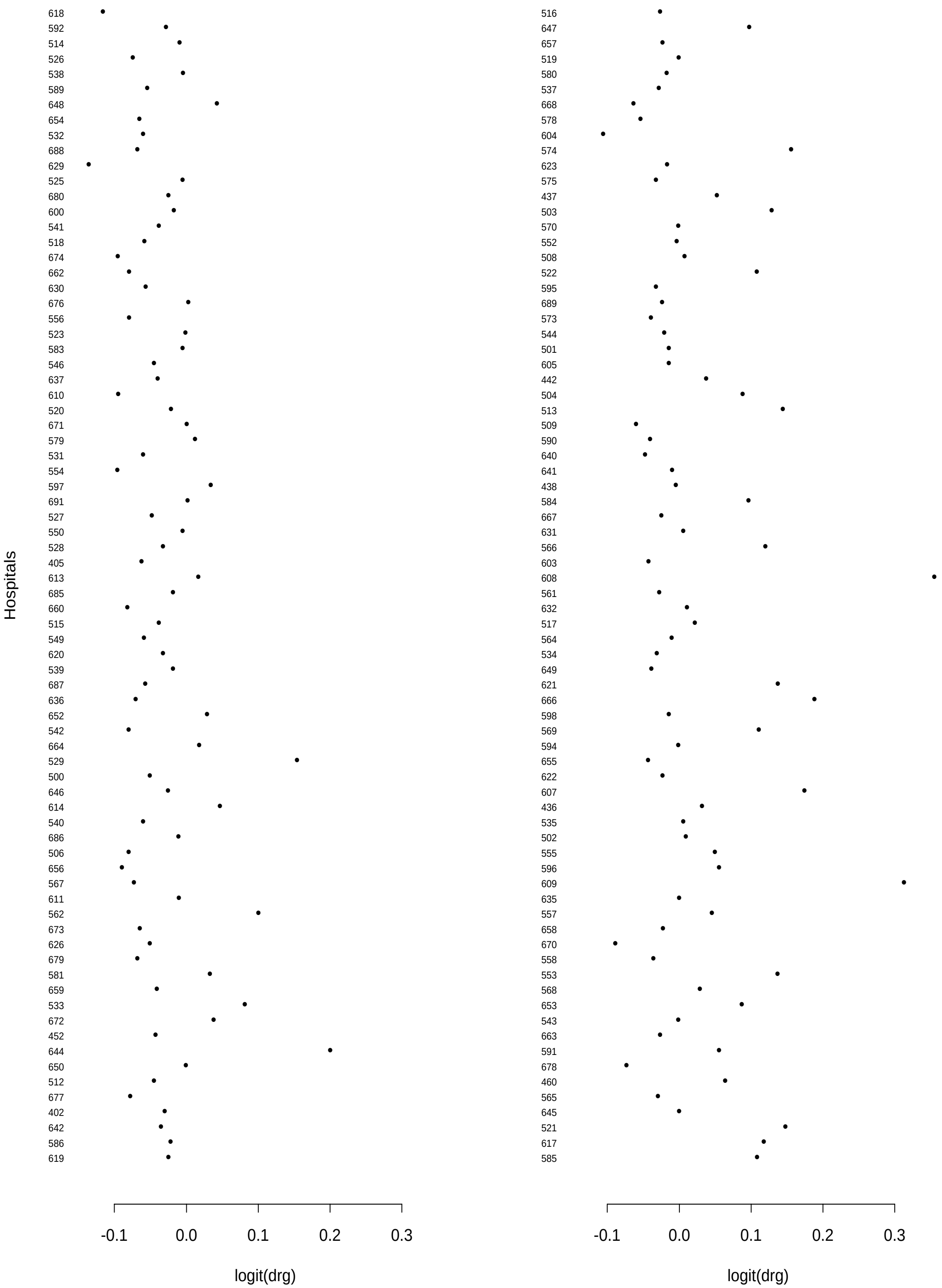


Figure 44: Observed value of DRG predicted proportions, after logit transformation, on M21 in 1995 over all hospitals.

9 Methodology and Computation

General comments on technology and software

The fitting of the binomial/logit random effects time series models, both single monitor and multiple monitors, involves new statistical methodology and computational techniques. It is beyond our scope here to develop complete details, but we expect that future work will include developing these models further and with a view to applications in other areas. The models presented share much of the structure and conceptual basis of models recently published in Cargnoni, Müller and West (1997). Components of the computational methods that have been developed in this context are novel and will apply to other hierarchical generalised linear models with time series structure, and be of relevance to researchers in many socio-economic areas.

All posterior computations are performed using Markov chain Monte Carlo simulation methods, the standard for implementation of Bayesian inference in other than very simple models. Our analyses are implemented in Fortran programs developed specifically for these models and this project and including both multiple and single monitor models. Reported analyses in the exploratory `MODEL-1` framework were independently performed using the Bayesian software package BUGS (Spiegelhalter et al 1995) as well as our custom programs. Repeating analyses using different software provides opportunity to assess the relative accuracy of the numerical approximations to posterior distributions. Our reported summaries of analyses of the single monitor series have been so validated. The Fortran software constructed for the project is portable and does not rely on either public domain or commercial libraries. Post-processing of the outputs of analyses to produce appropriate graphical and numerical summaries are currently implemented in functions in S-Plus.

Technical development of posterior analyses: `MODEL-3` framework

In each model class we perform analysis using customised Markov chain Monte Carlo (MCMC) simulation methods to evaluate, explore and summarise posterior distributions. We describe some technical details here for the most general `MODEL-3` class. Details for the other models are routine simplifications. MCMC involves iteratively re-simulating sets of parameters and random effects from collections of conditional posterior distributions in a Gibbs sampling format, and also using simulated values from so-called proposal distributions when, in some cases, Metropolis-Hastings methods are used. Our methods are similar to those developed and proven in Cargnoni, Müller and West (1997). All summarised results are based on over 100,000 simulations of posteriors, which are generated following 10,000 "burn-in" simulations that are discarded. Of the total 100,000, we subsample a set of 5,000 spaced 20 apart so as to break correlations and lead to essentially independent samples for summary analysis.

For the MODEL-3 class, the relevant unknown parameters are $\{\boldsymbol{\beta}_{0t}, \boldsymbol{\beta}_{1t}, \mathbf{V}, \mathbf{U}, \Phi, \boldsymbol{\alpha}_{i,t}, \boldsymbol{\mu}_{it}\}$. For any subset ξ of these quantities we will write ξ^- for the remaining variables together with the full data set $\mathbf{Z}$. The Gibbs/Metropolis-Hastings framework then involves iteratively resampling from exact conditional posterior distributions $p(\xi|\xi^-)$ for some subsets of parameters, and from proposal distributions that approximate conditionals for other parameters, the latter then subject to accept/reject tests to ensure the accepted samples are from the true posterior. In turn we sequence through

- $\boldsymbol{\beta}_{0t}$ given $\boldsymbol{\beta}_{0t}^-$
- $\boldsymbol{\beta}_{1t}$ given $\boldsymbol{\beta}_{1t}^-$
- $\mathbf{V}$ given $\mathbf{V}^-$
- $\mathbf{U}$ given $\mathbf{U}^-$
- Φ given Φ^-
- $(\boldsymbol{\alpha}_{it}, \boldsymbol{\mu}_{it})$ given $(\boldsymbol{\alpha}_{it}, \boldsymbol{\mu}_{it})^-$.

The relevant conditional posteriors and simulation structures are now detailed for each of these steps.

Sampling $\boldsymbol{\beta}_{0t}|\boldsymbol{\beta}_{0t}^-$

The $\boldsymbol{\beta}_{0t}$ are conditionally independent with full conditional posterior distributions, under reference priors, given by

$$p(\boldsymbol{\beta}_{01}|\boldsymbol{\beta}_{01}^-) \propto \prod_{i=1}^{I} p(\boldsymbol{\alpha}_{i1}|\boldsymbol{\beta}_{01})$$

and, for $t > 1$,

$$p(\boldsymbol{\beta}_{0t}|\boldsymbol{\beta}_{0t}^-) \propto \prod_{i=1}^{I} p(\boldsymbol{\alpha}_{it}|\boldsymbol{\alpha}_{i,t-1}, \boldsymbol{\beta}_{0t}, \boldsymbol{\beta}_{0,t-1}).$$

We have multivariate normal posteriors

$$\boldsymbol{\beta}_{01}|\boldsymbol{\beta}_{01}^- \sim N(\boldsymbol{\beta}_{01}|\mathbf{b}_{01}, \mathbf{W}/I) \quad \text{and} \quad \boldsymbol{\beta}_{0t}|\boldsymbol{\beta}_{0t}^- \sim N(\boldsymbol{\beta}_{0t}|\mathbf{b}_{0t}, \mathbf{U}/I),$$

where

$$\mathbf{b}_{01} = \sum_{i=1}^{I} \boldsymbol{\alpha}_{i1}/I$$

and, for $t > 1$,

$$\mathbf{b}_{0t} = \sum_{i=1}^{I} \{\boldsymbol{\alpha}_{it} - \Phi(\boldsymbol{\alpha}_{i,t-1} - \boldsymbol{\beta}_{0,t-1})\}/I.$$

Sampling $\boldsymbol{\beta}_{1t}|\boldsymbol{\beta}_{1t}^{-}$

Under reference priors, the $\boldsymbol{\beta}_{1t}$ vectors are conditionally independent over t, with full conditional posterior distributions $N(\boldsymbol{\beta}_{1t}|\mathbf{b}_{1t}, \mathbf{B}_{1t})$ for each t, and where

$$\boldsymbol{\beta}_{1t} = \mathbf{B}_{1t}\sum_{i=1}^{I}\mathbf{X}_{it}'\mathbf{V}^{-1}(\boldsymbol{\mu}_{it} - \boldsymbol{\alpha}_{it}) \quad \text{and} \quad \mathbf{B}_{1t}^{-1} = \sum_{i=1}^{I}\mathbf{X}_{it}'\mathbf{V}^{-1}\mathbf{X}_{it}.$$

Sampling $\mathbf{V}|\mathbf{V}^{-}$

Assuming a reference prior for the variance-covariance matrix $\mathbf{V}$, conditional posterior is the inverse Wishart distribution $Wi(\mathbf{V}^{-1}|8I, \mathbf{H})$ where

$$\mathbf{H} = \sum_{i=1}^{I}\sum_{t=1}^{8}\boldsymbol{\nu}_{it}\boldsymbol{\nu}_{it}'.$$

Sampling $\mathbf{U}|\mathbf{U}^{-}$

Assuming a reference prior for the variance-covariance matrix $\mathbf{U}$, the conditional posterior is given, in terms of the inverse $\mathbf{U}^{-1}$, by

$$\begin{aligned} p(\mathbf{U}^{-1}|\{\boldsymbol{\epsilon}_{it}\}, \Phi) &\propto p(\mathbf{U}^{-1})p(\{\boldsymbol{\epsilon}_{it}\}|\Phi, \mathbf{U}) \\ &\propto p(\mathbf{U}^{-1})\prod_{i=1}^{I}p(\boldsymbol{\epsilon}_{i1}|\Phi, \mathbf{U})\prod_{t=2}^{8}p(\boldsymbol{\epsilon}_{it}|\boldsymbol{\epsilon}_{i,t-1}, \Phi, \mathbf{U}) \\ &\propto a(\mathbf{U})Wi(\mathbf{U}^{-1}|7I, \mathbf{G}) \end{aligned}$$

with

$$\mathbf{G} = \sum_{i=1}^{I}\sum_{t=2}^{8}(\boldsymbol{\epsilon}_{it} - \Phi\boldsymbol{\epsilon}_{i,t-1})(\boldsymbol{\epsilon}_{it} - \Phi\boldsymbol{\epsilon}_{i,t-1})'$$

and

$$a(\mathbf{U}) = |\mathbf{W}|^{-I/2}\exp(-\text{trace}(\mathbf{W}^{-1}\mathbf{A})/2)$$

where $\mathbf{A} = \sum_{i=1}^{I}\boldsymbol{\epsilon}_{i1}\boldsymbol{\epsilon}_{i1}'$ and $\mathbf{W} = \Phi\mathbf{W}\Phi + \mathbf{U}$. We use the inverse Wishart distribution as a proposal distribution in the Metropolis-Hastings algorithm. That is, given a “current” value of $\mathbf{U}$ and corresponding $\mathbf{W}$, we sample a “candidate” value $\mathbf{U}^*$ from the inverse Wishart distribution, and accept it with probability

$$\min\{1, a(\mathbf{U}^*)/a(\mathbf{U})\}$$

where $\mathbf{W}^* = \Phi\mathbf{W}^*\Phi + \mathbf{U}^*$.

Sampling $\Phi|\Phi^-$

Conditional on Φ^- the posterior for Φ depends only on the random effects $\boldsymbol{\epsilon}_{it}$ and $\mathbf{U}$ via

$$\begin{aligned} p(\Phi|\{\boldsymbol{\epsilon}_{it}\},\mathbf{U}) &\propto p(\Phi)p(\{\boldsymbol{\epsilon}_{it}\}|\Phi,\mathbf{U}) \\ &\propto p(\Phi)\prod_{i=1}^{I} p(\boldsymbol{\epsilon}_{i1}|\Phi)\prod_{t=2}^{8} p(\boldsymbol{\epsilon}_{it}|\boldsymbol{\epsilon}_{i,t-1},\Phi) \\ &\propto p(\Phi)\prod_{i=1}^{I} N(\boldsymbol{\epsilon}_{i1}|\mathbf{0},\mathbf{W})\prod_{t=2}^{8} N(\boldsymbol{\epsilon}_{it}|\Phi\boldsymbol{\epsilon}_{i,t-1},\mathbf{U}) \end{aligned}$$

where $\mathbf{W} = \Phi\mathbf{W}\Phi + \mathbf{U}$ is easily evaluated as a function of Φ and $\mathbf{U}$. Write $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)'$ for the diagonal of Φ, and $\mathbf{E} = \text{diag}(\boldsymbol{\epsilon}_{i,t-1})$. Then the conditional posterior may be written as proportional to

$$p(\Phi)c(\Phi)N(\boldsymbol{\phi}|\mathbf{f},\mathbf{F})$$

where

$$\mathbf{f} = \mathbf{F}\sum_{i=1}^{I}\sum_{t=2}^{8}\mathbf{E}'\mathbf{U}^{-1}\boldsymbol{\epsilon}_{it} \quad \text{and} \quad \mathbf{F}^{-1} = \sum_{i=1}^{I}\sum_{t=2}^{8}\mathbf{E}'\mathbf{U}^{-1}\mathbf{E},$$

and

$$c(\Phi) = |\mathbf{W}|^{-I/2}\text{exp}(-\text{trace}(\mathbf{W}^{-1}\mathbf{A})/2)$$

where $\mathbf{A} = \sum_{i=1}^{I}\boldsymbol{\epsilon}_{i1}\boldsymbol{\epsilon}_{i1}'$ and $\mathbf{W} = \Phi\mathbf{W}\Phi + \mathbf{U}$. Under independent uniform priors for the ϕ_j, the full conditional posterior distribution for Φ is the above multivariate normal form truncated to the $(0, 1)$ regions in each dimension, and then multiplied by the factor $c(\Phi)$. This may be sampled in several ways; we use a Metropolis Hastings algorithm that takes the truncated multivariate normal component as a proposal distribution. That is, given a "current" value of $\boldsymbol{\phi}$, with corresponding matrices Φ and $\mathbf{W}$, we sample a "candidate" vector $\boldsymbol{\phi}^*$ from this truncated normal, compute the corresponding diagonal matrix Φ^* and variance matrix $\mathbf{W}^*$ such that $\mathbf{W}^* = \Phi^*\mathbf{W}^*\Phi^* + \mathbf{U}$, then accept this new $\boldsymbol{\phi}^*$ vector with probability

$$\min\{1, c(\Phi^*)/c(\Phi)\}.$$

Sampling $\boldsymbol{\alpha}_{it}|\boldsymbol{\alpha}_{it}^-$

Simulations are based on proposal distributions derived from the normal-logit approximations to the data model. Write

$$\tilde{\mathbf{y}}_{it} = \mathbf{y}_{it} - \mathbf{X}_{it}\boldsymbol{\beta}_{1t} = \boldsymbol{\alpha}_{it} + \boldsymbol{\eta}_{it}$$

where $\mathbf{y}_{it}$ is the vector of logit transforms of the observed outcome proportions. Under the model structure and assumptions, $\boldsymbol{\eta}_{it} \sim N(\boldsymbol{\eta}_{it}|\mathbf{0},\mathbf{V}+\mathbf{S}_{it})$ where

$$\mathbf{S}_{it} = \text{diag}(s_{i1t}, s_{i2t}, s_{i3t})$$

is the diagonal matrix of approximate data variances in the normal-logit model. Combined with the model equations

$$\boldsymbol{\alpha}_{it} \quad = \quad \boldsymbol{\beta}_{0t} + \Phi(\boldsymbol{\alpha}_{i,t-1} - \boldsymbol{\beta}_{0,t-1}) + \boldsymbol{\omega}_{it}$$

and the initial version for the $\boldsymbol{\alpha}_{i1}$, this gives us a multivariate dynamic linear model with known variance matrices and state vector sequence $\boldsymbol{\alpha}_{it}$. Standard results for DLMs now apply, as in West and Harrison (1997). To sample from the full conditional posterior distribution we implement multivariate versions of the forward-filtering, backwards-sampling algorithm detailed in West and Harrison (1997, chapter 15).

Sampling $\boldsymbol{\mu}_{it}|\boldsymbol{\mu}_{it}^{-}$

Given the just-sampled values of all $\boldsymbol{\alpha}_{it}$, we again use the normal-logit data model to generate candidate values of the $\boldsymbol{\mu}_{it}$ that are then tested for acceptance based on the exact conditional posteriors. As above, the approximate normal-logit data model serves to provide a very useful candidate generating model, as follows. For each i and t the exact conditional posteriors are

$$p(\boldsymbol{\mu}_{it}|\boldsymbol{\mu}_{it}^{-}, \mathbf{z}_{it}) \propto p(\mathbf{z}_{it}|\mathbf{n}_{it}, \boldsymbol{\mu}_{it})p(\boldsymbol{\mu}_{it}|\boldsymbol{\mu}_{it}^{-})$$

where the likelihood function $p(\mathbf{z}_{it}|\mathbf{n}_{it}, \boldsymbol{\mu}_{it})$ is the product of the three binomial-logit functions, and the conditional prior $p(\boldsymbol{\mu}_{it}|\boldsymbol{\mu}_{it}^{-})$ is the trivariate normal

$$\boldsymbol{\mu}_{it}|\boldsymbol{\mu}_{it}^{-} \sim N(\boldsymbol{\mu}_{it}|\boldsymbol{\alpha}_{it} + \mathbf{X}_{it}\boldsymbol{\beta}_{1t}, \mathbf{V}).$$

Again using the normal-logit approximation to the binomial data models, we obtain the approximate, and conditionally independent, normal posteriors

$$\boldsymbol{\mu}_{it}|\mathbf{y}_{it} \approx N(\boldsymbol{\mu}_{it}|\mathbf{m}_{it}, \mathbf{Q}_{it})$$

where

$$\mathbf{Q}_{it} = (\mathbf{V}^{-1} + \mathbf{S}_{it}^{-1})^{-1} \quad \text{and} \quad \mathbf{m}_{it} = \mathbf{Q}_{it}(\mathbf{V}^{-1}(\boldsymbol{\alpha}_{it} + \mathbf{X}_{it}\boldsymbol{\beta}_{1t}) + \mathbf{S}_{it}^{-1}\mathbf{y}_{it})$$

for each i and t.

We use these latter normal distributions as proposal distributions: generate candidate $\boldsymbol{\mu}_{it}$ values from each of this set of approximate posteriors, and accept/reject them according to a Metropolis-Hastings test. It is easily seen that this is a simple test, based on the ratio of the exact binomial to the approximate normal-logit likelihood functions. Specifically, if $\boldsymbol{\mu}_{it}$ is the current, "old" value of $\boldsymbol{\mu}_{it}$ from the previous MCMC iteration, a new value $\boldsymbol{\mu}_{it}^{*}$ from the probing distribution $N(\boldsymbol{\mu}_{it}|\mathbf{m}_{it}, \mathbf{Q}_{it})$ is accepted with probability

$$\min\{1, a(\boldsymbol{\mu}_{it}^{*})/a(\boldsymbol{\mu}_{it})\}$$

where $a(\cdot)$ is the ratio

$$a(\boldsymbol{\mu}_{it}) = p(\mathbf{z}_{it}|\mathbf{n}_{it}, \boldsymbol{\mu}_{it})/N(\mathbf{y}_{it}|\boldsymbol{\mu}_{it}, \mathbf{S}_{it}),$$

i.e., the ratio of the product of the three exact binomial likelihood components to the product of the three approximate normal-logit components. Note the very close similarities of this method to that in multinomial time series modelling in Cargnoni, Müller and West (1997).

Acknowledgements

The project was initiated by Ted Stefos of the VA Management Science Group, Bedford MA, and our work has been performed in consultation with Jim Burgess, Cindy Christiansen, Carl Morris and Ted Stefos. Our models represent multivariate time series extensions of hierarchical binomial regression models studied by other authors, and in particular build on the foundational contributions of Cindy Christiansen and Carl Morris in this area and applied to VA quality monitors.

References

Burgess, J.F., Christiansen, C.L., Michalak, S.E., and Morris, C.N. (1996) Risk adjustment and economic incentives in identifying extremes using hierarchical models: A profiling application using hospital monitors, *Manuscript,* Management Science Group, U S Department of Veterans Affairs, Bedford MA.

Cargnoni, C., Müller, P., and West, M. (1997) Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models, *Journal of the American Statistical Association,* **92**, 640-647.

Christiansen, C.L., and Morris, C.N. (1997) Hierarchical Poisson regression modeling, *Journal of the American Statistical Association,* **92**, 618-632.

Spiegelhalter, D.J,. Thomas, A., Best, N.G. and Gilks, W.R. (1995) *BUGS: Bayesian Inference using Gibbs Sampling, Version 0.50,* Cambridge: Medical Research Council Biostatistics Unit.

West, M., and Harrison, P.J. (1997) *Bayesian Forecasting and Dynamic Models,* (2nd Edn.), New York: Springer Verlag.