6-23-2005

A Mechanistic Latent Variable Model for Estimating Drug Concentrations in the Male Genital Tract

Leena Choi
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, lchoi@jhsph.edu

Brian Caffo
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, bcaffo@jhsph.edu

Charles A. Rohde
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, crohde@jhsph.edu

Themba T. Ndovi
Johns Hopkins University, Division of Clinical Pharmacology, Department of Medicine, School of Medicine

Craig W. Hendrix
Johns Hopkins University, Division of Clinical Pharmacology, Department of Medicine, School of Medicine, cwhendrix@jhmi.edu

A Mechanistic Latent Variable Model for Estimating Drug Concentrations in the Male Genital Tract

Leena Choi*, Brian S. Caffo, Charles Rohde†, Themba T. Ndovi, and Craig W. Hendrix‡

June 23, 2005

Abstract

The purpose of this study is to develop statistical methodology to facilitate indirect estimation of the concentration of antiretroviral drugs and viral loads in the prostate gland and the seminal vesicle. The differences in antiretroviral drug concentrations in these organs may lead to suboptimal concentrations in one gland compared to the other. Suboptimal levels of the antiretroviral drugs will not be able to fully suppress the virus in that gland, lead to a source of sexually transmissible virus and increase the chance of selecting for drug resistant virus. This information may be useful selecting antiretroviral drug regimen that will achieve optimal concentrations in most of male genital tract glands. Using fractionally collected semen ejaculates, Lundquist (1949) measured levels of surrogate markers in each fraction that are uniquely produced by specific male accessory glands. To determine the original glandular concentrations of the surrogate markers, Lundquist solved a simultaneous series of linear equations. This method has several limitations. In particular, it does not yield a unique solution, it does not address measurement error, and it disregards inter-subject variability in the parameters. To cope with these limitations, we developed a mechanistic latent variable model based on the physiology of the male genital tract and surrogate markers. We employ a Bayesian approach and perform a sensitivity analysis with regard to the distributional assumptions on the random effects and priors. The model and Bayesian approach is validated on experimental data where the concentration of a drug should be (biologically) differentially distributed between the two glands. In this example, the Bayesian model-based conclusions are found to be robust to model specification and

*Department of Biostatistics, School of Medicine, Vanderbilt University, naturechoi@hotmail.com
†Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University
‡Division of Clinical Pharmacology, Department of Medicine, School of Medicine, Johns Hopkins University

this hierarchical approach leads to more scientifically valid conclusions than the original methodology. In particular, unlike existing methods, the proposed model based approach was not affected by a common form of outliers.
keywords: latent variable model, structural model, mechanistic model, Bayesian, Lundquist's method

1 Introduction

The purpose of this study is to develop statistical methodology to estimate the concentration of antiretroviral drugs and viral loads in the prostate gland and the seminal vesicle. Estimates of concentration in these subcompartments of the male genital tract may provide important information about the Human Immunodeficiency Virus (HIV). In particular, if a drug concentration in one of these subcompartments is low, it may provide a sanctuary for the virus, allowing it to develop a resistance to the drug. Thus, understanding the partition of drug concentration in the prostate gland and the seminal vesicle may give valuable information for the mechanism of developing a drug-resistant virus. This information can also be used to choose a drug regimen to minimize the risk of drug-resistance in HIV infected patients (Ndovi, 2005). Unfortunately, however, it is extremely invasive to measure these concentrations directly.

To evaluate drug concentration in subcompartments of the male genital tract, Ndovi (2005) and Ndovi et al. (2005) use several surrogate markers in fractionally collected ejaculates; that is, binning a ejaculate over time. They then applied Lundquist's method (1949) to estimate the concentrations. Lundquist's method involves solving a linear system of equations to estimate the amount and composition of the secretions from fractionally collected ejaculates. Though Lundquist's work is foundational, and offered the first method for estimating these concentrations in a completely non-invasive manner, it is statistically unappealing. In particular, the system of equations is not unique, often yielding extremely different solutions depending on the set of equations chosen. Moreover, the uncertainty in the estimates due to measurement error is ignored. Furthermore, the resulting estimates are individual specific, with no inherent method to integrate the information about the concentration parameters across subjects.

In this manuscript we develop a mechanistic (structural) model based on the underlying physiological theory for the male genital tract to estimate the drug concentrations in the prostate gland and the seminal vesicle. Our reliance on scientifically motivated (mechanistic) models rather than purely empirical ones is motivated by the fact that the number of subjects and observations per subject for this kind of study is small, whereas the between-subject heterogeneity is relatively large. To construct our mechanistic model, we employ a latent variable that connects

the underlying theory developed by Lundquist with biomarkers unique to accessory glands of the male genital tract. The resulting model is implemented within a Bayesian framework. Since there is little prior information and also little information to validate the latent variable distributions, we performed a sensitivity analysis with several choices of distributions for prior and random effects.

The manuscript is presented as follows. Current methods, the underlying scientific theory and the rationale for our mechanistic (structural) model, are introduced in Section 2. In Section 3, we describe a data set from an experiment investigating differential distribution of salicylic acid (the primary metabolite of aspirin) and present exploratory analyses and an empirical model. Section 4 discusses several choices of Bayesian models while Section 5 summarizes results for the salicylic acid study. A summary and discussion follows in Section 6.

2 Background

Before describing the mathematical theory, we overview the experimental process and some of the biological theory. In a split ejaculate experiment, subjects are given a device similar to a pill-box that contains several chambers. The subjects ejaculate into the box, separating the ejaculate into three to five of the bins. Therefore, the semen in the first bin should represent early-stage ejaculate. After collection, a technician takes the sequenced ejaculate and measures its volume, the concentration of the drug or virus and of biomarkers. The importance of the biomarkers will be made clear in what follows. From this information, researchers would like to estimate the concentration of the drug or virus in the prostate gland or seminal vesicle.

Lundquist (1949) was the first to develop a systematic approach to solving this inverse problem. His methods require three fundamental physiological assumptions:

Assumption 1 The ejaculate fluid is only contributed from the prostate gland and the seminal vesicle or that the contribution from other glands may be ignored with respect to the purpose of estimating the drug concentration.

Assumption 2 The semen is not mixed completely in the urethra before ejaculation.

Assumption 3 The relative contribution of the prostate gland and the seminal vesicle is time dependent. That is, the amount contributed from the prostate gland decreases and the amount contributed from the seminal vesicle increases over time as shown in Figure 1. The gray portion in each fraction originates from the prostate gland, which is decreasing over time, whereas the white portion originates from the seminal vesicle, which is increasing. For example, at the beginning of ejaculation, 100% of the semen originates from the prostate

gland and 0% of the semen from the seminal vesicle; at the end of ejaculation, 0% of the semen originates from the prostate gland and 100% from the seminal vesicle. At some time point, they assume that if 30% originates from the prostate gland, then 70% originates from the seminal vesicle.

Using these assumptions, the observed concentrations of the drug and two biochemical markers, Lundquist's method can be applied to estimate the individual-specific drug concentrations. Suppose there are three fractions of ejaculate fluid ($j = 1, 2, 3$). Let c_j be the measured drug/viral concentration and w_j be the observed total volume of ejaculate fluid in fraction j. Interest lies in estimating the drug concentrations in the prostate gland ($\theta^{(p)}$), the concentration in the seminal vesicle ($\theta^{(v)}$) and their ratio. Let P_j be the volume of prostate fluid and V_j be the volume of seminal vesicle fluid in fraction j. The measured and unmeasured quantities and the notations along with the origins of portions in each fraction are illustrated in Figure 1.

Lundquist's method estimates $\theta^{(p)}$ and $\theta^{(v)}$ by solving a system of linear equations using any two of the three fractions. For illustration, consider Fractions 1 and 2 ($j = 1, 2$). According to Assumptions 1 and 2, the observed drug concentration c_j is the result of the mixing of the two drug concentrations contributed from the prostate gland and the seminal vesicle, one from $\theta^{(p)}$ diluted by factor P_j/w_j and the other from $\theta^{(v)}$ diluted by factor V_j/w_j. That is,

$$c_1 = \theta^{(p)}\frac{P_1}{w_1} + \theta^{(v)}\frac{V_1}{w_1} \quad \text{and} \quad c_2 = \theta^{(p)}\frac{P_2}{w_2} + \theta^{(v)}\frac{V_2}{w_2}, \tag{1}$$

which yields,

$$\hat{\theta}^{(v)} = \frac{c_1 w_1 - (c_2 w_2)(\frac{P_1}{P_2})}{V_1 - V_2(\frac{P_1}{P_2})} \quad \text{and} \quad \hat{\theta}^{(p)} = \frac{c_2 - \hat{\theta}_v(\frac{V_2}{w_2})}{\frac{P_2}{w_2}}. \tag{2}$$

Notice that only the total volume, not the organ-specific volumes, P_j and V_j, are measured directly. Therefore, (2) is not identified. However, P_j and V_j can be estimated by two biochemical markers, prostate specific antigen (PSA) and the fructose. These markers are known to originate exclusively from the prostate gland and the seminal vesicle respectively.

To estimate P_j and V_j, we follow the same development as in Equation (1). The observed PSA concentration in fraction j, a_j, is the concentration of PSA in the prostate gland, A, diluted by the proportion of the ejaculate originating from the prostate P_j/w_j. Similarly, the observed fructose concentration in fraction j, b_j, is the concentration of fructose in the seminal vesicle, B, diluted by the proportion of the ejaculate originating from the seminal vesicle V_j/w_j. Thus, these relationships can be expressed as:

$$a_j = A\frac{P_j}{w_j} + 0\frac{V_j}{w_j} \quad \text{and} \quad b_j = 0\frac{P_j}{w_j} + B\frac{V_j}{w_j}. \tag{3}$$

The 0 concentrations are added to draw a parallel with Equation (1); the PSA concentration in the seminal vesicle is 0 and the fructose concentration in the prostate gland is 0. Notice that neither A nor B is known. Rearranging this equation yields

$$\frac{P_j}{w_j} = \frac{a_j}{A} \quad \text{and} \quad \frac{V_j}{w_j} = \frac{b_j}{B}. \tag{4}$$

That is, the proportional contribution of the prostate gland (seminal vesicle) to the jth fraction of the ejaculate is equal to the ratio of the observed PSA (fructose) in fraction j to the PSA (fructose) in the prostate gland (seminal vesicle). Throughout, we use P_j/w_j and V_j/w_j interchangeably with a_j/A and b_j/B respectively.

Assumption 1 states that the sum of the contributions from each subcompartment is 1. Thus, $P_j/w_j + V_j/w_j = 1$ and hence

$$a_j/A + b_j/B = 1. \tag{5}$$

Therefore, with two observed fractions, say 1 and 2, Equation (5) yields two equations with two unknowns. The solution is

$$\hat{B} = \frac{a_1 b_2 - a_2 b_1}{a_1 - a_2} \quad \text{and} \quad \hat{A} = \frac{a_1}{1 - \frac{b_1}{\hat{B}}}.$$

Of course, by the equivalence given in (4), these estimates yield estimates, say $\hat{P}_j$ and $\hat{V}_j$, which can be used in (2).

Notice that the resulting estimates for A, B, $\theta^{(p)}$ and $\theta^{(v)}$ depend on which pair of fractions are used ($j = 1, 2$, $j = 1, 3$ or $j = 2, 3$). Furthermore, directly solving these equations as such ignores the boundary constraint that all of the concentrations must be positive. Though the latter problem could be addressed by linear programming, the former is more intractable. To illustrate, Table 1 shows examples of the variability in the estimates for a data set that will be introduced in Section 3. There is substantial variability in the estimates depending on the fractions chosen.

One method to mitigate this problem, at least with regard to A and B, is as follows. Recall that, as shown above, $a_j = A - \frac{A}{B} b_j$. Therefore, Ndovi et al. (2005) estimate A as the intercept of a linear regression fit of a_j (the response) on b_j (the regressor). The estimate of B can be obtained as the negative of the estimated intercept *divided by the estimated slope.* It is informative to consider the biological justification for this linear regression. Imagine that there is a fraction, say Fraction 0, which is collected immediately prior to ejaculation (which is, of course, not attainable). Then, because there is no contribution from the seminal vesicle in this fraction, it should contain no fructose. The PSA concentration of this fraction, under our linearity assumptions, is indeed the PSA concentration in the prostate gland. The same exercise

can be done on a conceptual final fraction of ejaculate (just after the completion of ejaculation) containing only fructose and no PSA, having originated from only the seminal vesicle.

The decision to regress a_j on b_j is arbitrary. Estimates can be obtained in the same way using the reversed regression relationship, $b_j = B - \frac{B}{A}a_j$. In this case, the estimate of B is the estimated intercept while the estimate of A is the negative estimated intercept divided by the slope. However, we recommend a strategy by which the two intercepts are used to estimate A and B. The intercept from regressing a_j on b_j yields the estimate for A while regressing b_j on a_j yields the estimate for B.

The impact of which regression strategy is chosen is illustrated in Table 2 using the data in Section 3. The absolute values of $\hat{A}$ and $\hat{B}$ obtained by the negative intercept divided by the slope are consistently larger than $\hat{A}$ and $\hat{B}$ estimated by the regression intercepts. This discrepancy is due to measurement errors in the independent variables, which causes the estimated slopes to be attenuated (Fuller, 1987). Hence, the resulting estimates are inflated, as they are divided by the attenuated slopes. If the (a_j, b_j) pairs lie exactly on a line, all of the methods agree, including using only two of the fractions.

Before continuing, we summarize the principal statistical limitations in existing methodologies that we are attempting to address. Using only two of the fractions to estimate $\theta^{(p)}$ and $\theta^{(v)}$ disregards a substantial amount of information, and also eliminates any information regarding within-subject (measurement) error. Note that experimentally reducing the number of fractions collected to only two would yield unique solutions. However, this is not recommended because of large measurement errors and also this would aggregate important temporal information. Also, current practice only yields point estimates with no immediate method for assessing uncertainty at any stage. In particular, measurement error and the error in estimating A and B are ignored. Moreover, all estimation methods are individual-specific. There is also no immediate method for combining information across subjects.

In addition to these limitations, the biological assumptions may not hold. For example, consider Figure 2 which displays PSA concentration (a_j) by fructose (b_j) for the data in Section 3. In instances such as Subject 18, the PSA and fructose concentrations are very similar over all fractions of the ejaculate. This suggests that the semen was completely mixed in the urethra before ejaculation. That is, the temporal information was lost. In this instance, there is no information to estimate drug concentrations using Lundquist's method or Ndovi et al. (2005)'s modified version of Lundquist's method. Moreover, any attempt to combine estimates across subjects that disregards the error in estimating A and B would be thwarted by this subject's outlying estimate.

3 A Validation Data Set

A preliminary study was conducted to develop and test methodology to later study how antiretroviral drugs and viral load distribute differentially to the prostate gland and the seminal vesicle. aspirin (ASA) was administered to 13 healthy male volunteers (Ndovi, 2005). This drug was chosen based on the pKa, 2.8, of its primary metabolite, salicylic acid (SA), a weak acid. The pH of the two organs is differential. The seminal vesicle is basic $(pH = 7.8)$ while the prostate gland is acidic $(pH = 6.6)$. Thus, the experimental hypothesis is that SA, according to the $pH - pKa$ partition hypothesis, will be found at higher concentrations in the seminal vesicle than the prostate. Because the direction of the differential concentration of this drug is known, this experiment is useful to debug and validate both experimental and statistical methods.

Recall that Equation (4) equates the ratio of the observed PSA in fraction j and the PSA in the prostate gland (a_j/A) to the proportion of the ejaculate originating from the prostate (P_j/w_j). Also it equates the ratio of the observed fructose in fraction j and the fructose in the seminal vesicle (b_j/B) to the proportion of ejaculate originating from the seminal vesicle (V_j/w_j). Therefore, under Lundquist's Equation (1), it is informative to plot the concentration of SA (c_j) on both a_j/A and b_j/B (Figures 3 and 4). Here A and B were estimated as described in the previous section as the intercepts of separate linear regression fits. Notice that the range of $a_j/\hat{A}$ and $b_j/\hat{B}$ lies in $(0, 1)$, except for Subject 18.

Figure 3 illustrates that the concentration of SA decreases as the fraction of contribution from the prostate gland increases. Similarly, Figure 4 illustrates that the SA concentration increases as the fraction of contribution from the seminal vesicle increases. This implies that the SA concentration in the seminal vesicle is likely to be higher than that in the prostate gland, supporting the hypothesis of differential distribution for this drug.

These exploratory figures agree with Lundquist's Equations (1) and (4). Furthermore, there appears to be a negative relationship between the contribution from the prostate gland and the seminal vesicle; that is, Assumption 3 appears to be reasonable. However, Assumption 2 may not be valid for subjects such as Subject 18. All pairs of measurements for Subject 18 are very similar, implying that the semen of this subject is nearly completely mixed in the urethra before ejaculation. Hence, this individual displays no information for estimating A and B. Notice also that $\hat{B}$ is a small negative value (which is theoretically impossible) and hence the range of $b_j/\hat{B}$ is far from $(0, 1)$.

3.1 Two-stage regression

Before taking on more aggressive modelling, we consider a two-stage model fit to the data. We use the results of these fits to guide model building. The first stage of this approach estimates A and B using the intercepts of linear regression fits of a_i on b_i and b_i on a_i, as discussed in Section 2. Taking the intercepts obtained by reversing the regression relationships was motivated by Winsor (1946). This approach eliminates the attenuation seen in Table 2 when only one regression equation is employed. Also, it is more convenient, as there is no need to decide which variable to treat as the response and which to treat as the predictor.

To obtain estimates of $\theta^{(p)}$ and $\theta^{(v)}$, we treat Lundquist's equation as a regression relationship with mean model

$$E[c_j] = \theta^{(p)} \frac{a_j}{\hat{A}} + \theta^{(v)} \frac{b_j}{\hat{B}}.$$

That is, we regress c_j on $a_j/\hat{A}$ and $b_j/\hat{B}$ through the origin. The results of the two-stage regression fits are given in Table 3.

Notice that, unlike existing methods, this two-stage approach yields a unique solution. Also, no negative values are recorded, despite the lack of boundary constraints. Therefore, we view this approach as an improvement over Lundquist's original method and currently implemented modifications (Ndovi et al., 2005). However, as illustrated in Table 3, it is important to account for the error in estimating A and B. If the error in estimating A and B is disregarded, then observations with very high variance, such as 18 and 19, would dominate any attempt to combine the estimates of $\theta^{(p)}$, $\theta^{(v)}$ or their ratio across subjects. These inadequacies could potentially be addressed within this two-stage framework, via bootstrapping strategies, for example. However, because of the small number of subjects and multilevel structure in the data, we prefer a more model based approach.

4 A Multilevel Latent Variable Model

In this section we create a model that accounts for between-subject heterogeneity in the parameters and employ a mechanistic latent variable to account for variation in the unknown fraction. Here, we distinguish between random effects, which we view as a tool for modelling population level heterogeneity in parameters, and the mechanistic latent variable, which we view as an underlying biological construct. Of course, this distinction applies to interpretation; mathematically, the two kinds of unobserved variables are handled identically.

Expanding the prior notation, let a_{ij}, b_{ij} and c_{ij} represent the concentration of PSA, fructose and the drug/virus, respectively, on subject $i = 1, \ldots, I$ for fraction $j = 1, \ldots, J_i$. Let A_i be

the PSA concentration in the prostate gland and B_i be the fructose concentration in the seminal vesicle for subject i. Similarly, let $\theta_i^{(p)}$ be the drug concentration in the prostate gland and $\theta_i^{(v)}$ be the drug concentration in the seminal vesicle for subject i. We assume the following model on the observed concentrations

$$\begin{aligned} a_{ij} &\mid A_i, f_{ij} \sim \mathrm{N}\left\{A_i(1-f_{ij}), \tau_a^{-1}\right\} \\ b_{ij} &\mid B_i, f_{ij} \sim \mathrm{N}\left\{B_i f_{ij}, \tau_b^{-1}\right\} \\ c_{ij} &\mid a_{ij}, b_{ij}, f_{ij} \sim \mathrm{N}\left\{\theta_i^{(p)}(1-f_{ij}) + \theta_i^{(v)} f_{ij}, \tau_c^{-1}\right\} \end{aligned}$$

The mean model in these distributions for each subject, are exactly Lundquist's Equations (1) and (3). The latent variable f_{ij} represents V_{ij}/w_{ij}, the fractional contribution of the total ejaculate from the seminal vesicle for subject i. Because of Assumption 1, $(1-f_{ij})$ represents P_{ij}/w_{ij}, the fractional contribution from the prostate gland for subject i. Figure 5 shows a simplified anatomical diagram of the accessory reproductive organs in males along with the parameters and the latent variable f_j.

The unobserved variables A_i and B_i are assumed to follow normal distributions with means A and B (respectively) and precisions τ_A and τ_B (respectively). That is,

$$A_i \sim \mathrm{N}\{A, \tau_A^{-1}\} \quad \text{and} \quad B_i \sim \mathrm{N}\{B, \tau_B^{-1}\}.$$

Because the latent variable, f_{ij}, is bounded by $(0,1)$, it is naturally assumed to be Beta(α, β). Here we assume that the A_i, B_i and f_{ij} are mutually independent of each other and of $\theta_i^{(p)}$ and $\theta_i^{(v)}$, because there is no physiological reason to suspect that PSA, fructose and the drug are correlated.

Because they are the central quantities of interest, we investigated several choices of distributions for $\theta_i^{(p)}$ and $\theta_i^{(v)}$. We considered both bivariate normal or bivariate log-normal distributions, with and without positivity constraints,

$$\begin{pmatrix} \theta_i^{(p)} \\ \theta_i^{(v)} \end{pmatrix} \sim \mathrm{BVN}\left\{ \begin{pmatrix} \theta^{(p)} \\ \theta^{(v)} \end{pmatrix}, \Omega \right\} \text{ or } \begin{pmatrix} \log\theta_i^{(p)} \\ \log\theta_i^{(v)} \end{pmatrix} \sim \mathrm{BVN}\left\{ \begin{pmatrix} \log\theta^{(p)} \\ \log\theta^{(v)} \end{pmatrix}, \Omega \right\}$$

Recall that $\theta_i^{(p)}$ and $\theta_i^{(v)}$ both represent the concentration of the drug under study. Therefore, it is important to consider the possibility that they are correlated, in either direction. A positive correlation might occur because the drug concentrations in both subcompartments depend on the overall drug concentration in the blood; the higher the drug concentration in the blood, the higher are the drug concentrations in most of organs in the body. On the other hand, the concentrations might be negatively correlated because if more of the drug goes into one compartment, less of

the drug is available to go into the other. Depending on the drug, these effects could also cancel each other out or one direction could dominate. We note that with only 13 subjects, there is almost no evidence in the data to verify this assumption. Therefore, for both the normal and log-normal models, we consider instances where the correlation is present or forced to 0.

The likelihood and notational model specification can be found in Appendix A. Although the models may be categorized within the framework of hierarchical linear models, there are several challenges that prevent fitting via commercially available statistical packages. First, there are multivariate outcomes, all of which involve the latent variable and random effects. Second, the distribution of the latent variable, f_{ij} (Beta$\{\alpha, \beta\}$), is non-conjugate. Third, the latent variable is indexed by both the subject level, i, and the within subject level, j. To address these computational challenges, we explored both maximum likelihood (ML) via the Monte-Carlo Expectation and Maximization (MCEM) algorithm and Bayesian modelling via Markov chain Monte-Carlo (MCMC). We found that the MCEM/ML solution was less convenient, with no applicable software, requiring extensive reprogramming with minor model respecifications. The Bayesian modelling/MCMC approach proved to be more flexible, easier to implement and allowed us to consider a wider variety of models. Overall, the two approaches provided similar results on a subset of the models.

4.1 Inference based on a Bayesian modelling

We fit the model under a Bayesian framework using MCMC methods implemented with WinBUGS (Spiegelhalter et al., 1996). We used diffuse priors on the hyperparameters. In particular, the prior distributions on $\theta^{(p)}$ and $\theta^{(v)}$ were either diffuse normal or diffuse log-normal having means of 0 and precisions of 10^{-4}. We considered both priors with and without the positivity constraints. The mean parameters, A and B, had independent diffuse normal priors, again with means of 0 and precisions of 10^{-4}. All of the precisions parameters, τ_a, τ_b, τ_A and τ_B, were given gamma priors with shapes and scales of 10^{-3}. The same prior was used for the precision parameters for $\theta_i^{(p)}$ and $\theta_i^{(v)}$ when they were treated as uncorrelated. When an unstructured correlation matrix was used, Ω, it was given a Wishart distribution prior centered at an identity matrix with 2 degrees of freedom.

5 Analysis of the Salicylic Acid Data

Figures 6 and 7 show the posterior distributions, posterior means and 95% credible intervals for the parameters of interest: $\theta^{(p)}$, $\theta^{(v)}$ and $\theta^{(p)}/\theta^{(v)}$. The left panels show the posterior distributions for

the cases where $\theta_i^{(p)}$ and $\theta_i^{(v)}$ are treated as correlated, while the right panels show the posterior distributions for the cases where $\theta_i^{(p)}$ and $\theta_i^{(v)}$ are treated as uncorrelated. The top, middle and bottom panels correspond to the normal, normal with positivity constraints and log-normal distributions. Table 4 summarizes the model specification and their associated layout in Figures 6 and 7.

The rational for the choices of models is as follows. To select candidate models, we plotted jittered values of $\theta_i^{(p)}$ and $\theta_i^{(v)}$ estimated using the two-stage regression approach (see Figure 8). Even though it is impossible to check the normality assumptions on $\theta_i^{(p)}$ and $\theta_i^{(v)}$, this figure suggests that the marginal distributions are better represented by the truncated normal or perhaps log-normal distributions because of the skewness. This is especially true for $\theta_i^{(p)}$, whose distribution appears to be highly skewed.

Figure 6 shows that the posterior distributions of $\theta^{(p)}$ for the log-normal models are shifted toward 0. Those of the normal model are positioned in the middle of the other two models with the most spread while those of the truncated normal models are shifted toward large positive values. On the other hand, the posterior distribution for $\theta^{(v)}$ is similar across model specifications. That is, the posterior distribution of $\theta^{(p)}$ appears to be more sensitive to model choice. The empirical two-stage regression fits shed some light on this behavior. The $\hat{\theta}_i^{(v)}$ appear "better behaved" than the $\hat{\theta}_i^{(p)}$ in the sense that they are only mildly skewed without a cluster of values near 0. In contrast, $\hat{\theta}_i^{(p)}$ is highly skewed, with a high proportion of very small positive values. The log-normal model allows for a high skewness with very small positive values, whereas the truncated normal models force the distributions toward larger positive values, explaining the shift in the mean.

Figure 6 indicates that the posterior distributions of $\theta^{(p)}$ and $\theta^{(v)}$ become slightly closer together for most models when they are allowed to be correlated, implying that the negative correlation seems to dominate. That is, the more of the drug in one subcompartment, the less of the drug in the other. Despite this slight shrinkage, it is important to emphasize that the posterior distributions of $\theta^{(p)}$ and $\theta^{(v)}$ are not sensitive enough to the distributional assumptions for the random effects and priors to change the substantive conclusions.

Overall, Figure 6 shows that the posterior distributions for $\theta^{(p)}$ and $\theta^{(v)}$ are well separated for all models via 95% credible intervals. Figure 7 focuses on the ratio of $\theta^{(p)}$ and $\theta^{(v)}$, the parameter of central interest. This distribution is mostly below 1 and the 95% credible interval does not include 1 for all models. The posterior means of the ratio appears to be around .3, suggesting that the concentration of SA in the seminal vesicle is about three times as much as that in the prostate gland. Thus, the hypothesis of differential distribution of SA to two subcompartments is supported by the data. We remind the reader that this conclusion is scientifically well founded,

as the acidic characteristics of SA makes it more likely to be distributed in the seminal vesicle which has basic characteristics.

Our model is useful to estimate population-averaged estimates of the drug concentrations in two subcompartments and the ratio of drug concentrations, which cannot be directly estimated using Lundquist's method. The proposed Bayesian methodology results in an appropriately weighted population-averaged estimate to examine the hypothesis of the differential distribution of SA. In addition, Figures 6 and 7 suggest that the estimates of $\theta^{(p)}$, $\theta^{(v)}$ and $\frac{\theta^{(p)}}{\theta^{(v)}}$ are robust to the distributional assumptions for random effects and priors in drawing the conclusion.

6 Summary and Discussions

We developed a mechanistic model based on physiological theory for fractionally collected ejaculates. To the best of our knowledge, this is the first rigorous statistical development and first application of Bayesian methodology in this scientific area. Our model is motivated by the structural theory originally developed by Lundquist and later refined by Ndovi et al. (2005). We formalized their modelling approach within a hierarchical framework. This framework included random effects to describe the population level variation in the subject-specific concentration parameters. Moreover, we utilized a latent variable to represent an unobserved biological construct.

The method was validated on experimental data where the direction of the differential distribution of the drug of interest (salicylic acid) in the prostate gland and seminal vesicle is expected. This proof-of-concept data set both validated the experimental technique as well as the Bayesian inference laid out in this manuscript. It is important to emphasize that current methodology fails to distinguish differential concentration. These results were found to be robust to different distributional assumptions.

Finally, we emphasize the importance of taking into account the variability in estimating nuisance concentration parameters. As mentioned, subjects such as Subject 18 offer little or no information to estimate A_i and B_i, the subject-specific PSA concentration in the prostate gland and fructose concentration in the seminal vesicle. The working physiological theory is that the ejaculate becomes too mixed in the urethra for such subjects to differentiate early and late-stage ejaculate. Current practice tends to rely on ad hoc methods, such as subjectively discarding data with perceived low information. A principal benefit of the proposed Bayesian approach is that the uncertainty in estimating these quantities is accounted for within a unified framework. As a result, posterior estimates of the drug concentration are appropriately shrunk towards an overall mean. Figure 9 illustrates this by depicting the empirical estimates for the ratio of SA concentrations in the two compartments based on the two-stage regression approach and the

resulting Bayesian posterior means. It was necessary to present results on the log-scale because the empirical estimate for Subject 18 was so large. This estimate is not only large, but contrary to the evidence presented by the remaining subjects. The Bayesian approach appropriately shrinks this observation towards the overall mean.

A Likelihood and model notation

A.1 Likelihood

$$
\begin{aligned}
&L(\theta^{(p)}, \theta^{(v)}, \Omega, D|\boldsymbol{y_1}, \boldsymbol{y_2}, \boldsymbol{y_3}) \\
&= \int \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_{\boldsymbol{y_{ij}}|\boldsymbol{u}}(\boldsymbol{y_{ij}}|\boldsymbol{u}, \Omega) f_{\boldsymbol{u}}(\boldsymbol{u}|\theta^{(p)}, \theta^{(v)}, D) d\boldsymbol{u} \\
&= \int \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_{\boldsymbol{y_{ij}}|\boldsymbol{u}}(y_{1ij}, y_{2ij}, y_{3ij}|A_i, B_i, \theta_i^{(p)}, \theta_i^{(v)}, f_{ij}, \Omega) f_{\boldsymbol{u}}(f_{ij}, A_i, B_i, \theta_i^{(p)}, \theta_i^{(v)}|\theta^{(p)}, \theta^{(v)}, D)\ d\boldsymbol{u} \\
&= \int\int\int\int\int \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_{y_{1ij}}(y_{1ij}|A_i, f_{ij}, \sigma_1^2) f_{y_{2ij}}(y_{2ij}|B_i, f_{ij}, \sigma_2^2) \\
&\quad \times f_{y_{3ij}}(y_{3ij}|\theta_i^{(p)}, \theta_i^{(v)}, f_{ij}, \sigma_3^2) f_{f_{ij}}(f_{ij}|\alpha, \beta) f_{A_i}(A_i|A, \sigma_A^2) f_{B_i}(B_i|B, \sigma_B^2) \\
&\quad \times f_{\theta_i^{(p)}, \theta_i^{(v)}}(\theta_i^{(p)}, \theta_i^{(v)}|\theta^{(p)}, \theta^{(v)}, \sigma_p^2, \sigma_v^2)\ df_{ij}\ dA_i\ dB_i\ d\theta_i^{(p)}\ d\theta_i^{(v)} \\
&= \int \prod_{i=1}^{m} f_{\boldsymbol{y_{1i}}}(\boldsymbol{y_{1i}}|A, \boldsymbol{f_i}, \sigma_A^2, \sigma_1^2) f_{\boldsymbol{y_{2i}}}(\boldsymbol{y_{2i}}|B, \boldsymbol{f_i}, \sigma_B^2, \sigma_2^2) f_{\boldsymbol{y_{3i}}}(\boldsymbol{y_{3i}}|\theta^{(p)}, \theta^{(v)}, \boldsymbol{f_i}, \sigma_p^2, \sigma_v^2, \sigma_3^2) \\
&\quad \times f_{\boldsymbol{f_i}}(\boldsymbol{f_i}|\alpha, \beta)\ d\boldsymbol{f_i}
\end{aligned} \tag{6}
$$

where

$$
\begin{aligned}
\boldsymbol{y_{ij}} &= (y_{1ij}, y_{2ij}, y_{3ij}), \quad \boldsymbol{u} = (f_{ij}, A_i, B_i, \theta_i^{(p)}, \theta_i^{(v)}), \\
\Omega &= (\sigma_1^2, \sigma_2^2, \sigma_3^2), \qquad D = (\alpha, \beta, A, B, \sigma_A^2, \sigma_B^2, \sigma_p^2, \sigma_v^2, \rho),
\end{aligned}
$$

and

$$
\begin{aligned}
&f_{\boldsymbol{y_{1i}}}(\boldsymbol{y_{1i}}|A, \boldsymbol{f_i}, \sigma_A^2, \sigma_1^2) \\
&= (2\pi)^{-n_i/2}|\Sigma_{1i}|^{-1/2}\exp\Big\{-\frac{1}{2}\big[\boldsymbol{y_{1i}} - A(1-\boldsymbol{f_i})\big]^t\Sigma_{1i}^{-1}\big[\boldsymbol{y_{1i}} - A(1-\boldsymbol{f_i})\big]\Big\} \\
&f_{\boldsymbol{y_{2i}}}(\boldsymbol{y_{2i}}|B, \boldsymbol{f_i}, \sigma_B^2, \sigma_2^2) \\
&= (2\pi)^{-n_i/2}|\Sigma_{2i}|^{-1/2}\exp\Big\{-\frac{1}{2}\big[\boldsymbol{y_{2i}} - B\boldsymbol{f_i}\big]^t\Sigma_{2i}^{-1}\big[\boldsymbol{y_{2i}} - B\boldsymbol{f_i}\big]\Big\} \\
&f_{\boldsymbol{y_{3i}}}(\boldsymbol{y_{3i}}|\theta^{(p)}, \theta^{(v)}, \boldsymbol{f_i}, \sigma_p^2, \sigma_v^2, \sigma_3^2) \\
&= (2\pi)^{-n_i/2}|\Sigma_{3i}|^{-1/2} \\
&\quad\times\exp\Big\{-\frac{1}{2}\big[\boldsymbol{y_{3i}} - \theta^{(p)}(1-\boldsymbol{f_i}) - \theta^{(v)}\boldsymbol{f_i}\big]^t\Sigma_{3i}^{-1}\big[\boldsymbol{y_{3i}} - \theta^{(p)}(1-\boldsymbol{f_i}) - \theta^{(v)}\boldsymbol{f_i}\big]\Big\} \\
&f_{\boldsymbol{f_i}}(\boldsymbol{f_i}|\alpha,\beta) = \prod_{j=1}^{n_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}f_{ij}^{\alpha-1}(1-f_{ij})^{\beta-1}
\end{aligned}
$$

where

$$
\begin{aligned}
&\Sigma_{1i} = \sigma_1^2 I_{n_i} + \sigma_A^2(1-\boldsymbol{f_i})(1-\boldsymbol{f_i})^t, \\
&\Sigma_{2i} = \sigma_2^2 I_{n_i} + \sigma_B^2\boldsymbol{f_i}\boldsymbol{f_i}^t, \text{ and} \\
&\Sigma_{3i} = \sigma_3^2 I_{n_i} + \sigma_p^2(1-\boldsymbol{f_i})(1-\boldsymbol{f_i})^t + \sigma_v^2\boldsymbol{f_i}\boldsymbol{f_i}^t \text{ for Model 1 and} \\
&\Sigma_{3i} = \sigma_3^2 I_{n_i} + \sigma_p^2(1-\boldsymbol{f_i})(1-\boldsymbol{f_i})^t + \sigma_v^2\boldsymbol{f_i}\boldsymbol{f_i}^t + \rho\sigma_p\sigma_v(1-\boldsymbol{f_i})\boldsymbol{f_i}^t \\
&\qquad + \rho\sigma_p\sigma_v\boldsymbol{f_i}(1-\boldsymbol{f_i})^t \text{ for Model 2.}
\end{aligned}
$$

The inverse and determinant of covariance matrices are following:

$$
\begin{aligned}
&\Sigma_{1i}^{-1} = \frac{1}{\sigma_1^2}I_{n_i} - \frac{(\sigma_A^2/\sigma_1^4)(1-\boldsymbol{f_i})(1-\boldsymbol{f_i})^t}{1+(\sigma_A^2/\sigma_1^2)(1-\boldsymbol{f_i})^t(1-\boldsymbol{f_i})} \\
&\Sigma_{2i}^{-1} = \frac{1}{\sigma_2^2}I_{n_i} - \frac{(\sigma_B^2/\sigma_2^4)\boldsymbol{f_i}\boldsymbol{f_i}^t}{1+(\sigma_B^2/\sigma_2^2)\boldsymbol{f_i}^t\boldsymbol{f_i}} \\
&\Sigma_{3i}^{-1} = V^{-1} - \frac{(\sigma_v^2)V^{-1}\boldsymbol{f_i}\boldsymbol{f_i}^tV^{-1}}{1+\sigma_v^2\boldsymbol{f_i}^tV^{-1}\boldsymbol{f_i}} \\
&\qquad \text{where } V^{-1} = \frac{1}{\sigma_3^2}I_{n_i} - \frac{(\sigma_p^2/\sigma_3^4)(1-\boldsymbol{f_i})(1-\boldsymbol{f_i})^t}{1+(\sigma_p^2/\sigma_3^2)(1-\boldsymbol{f_i})^t(1-\boldsymbol{f_i})} \text{ for Model 1} \\
&|\Sigma_{1i}| = (\sigma_1^2)^{(n_i-1)}(\sigma_1^2 + \sigma_A^2(1-\boldsymbol{f_i})^t(1-\boldsymbol{f_i})) \\
&|\Sigma_{2i}| = (\sigma_2^2)^{(n_i-1)}(\sigma_2^2 + \sigma_B^2\boldsymbol{f_i}^t\boldsymbol{f_i}).
\end{aligned}
$$

A.2 Model notation

1. The distributions of the observed data a_{ij}, b_{ij} and c_{ij} are:

$$\begin{aligned} a_{ij} &\sim \mathsf{N}\{A_i(1-f_{ij}),\ \tau_a^{-1}\}, \\ b_{ij} &\sim \mathsf{N}\{B_i f_{ij},\ \tau_b^{-1}\}, \\ c_{ij} &\sim \mathsf{N}\{\theta_i^{(p)}(1-f_{ij}) + \theta_i^{(v)} f_{ij},\ \tau_c^{-1}\} \end{aligned}$$

   where τ_a, τ_b and τ_c are precision parameters which are the inverse of variances.

2. The distributions of random effects and the latent variable

   - the distributions for the random effects and the latent variable are:

$$A_i \sim \mathsf{N}\{A, \tau_A^{-1}\},\ \ B_i \sim \mathsf{N}\{B, \tau_B^{-1}\} \ \text{ and } \ f_{ij} \sim \mathsf{Beta}\{\alpha, \beta\},$$

     where τ_A and τ_B are precision parameters.

   - the distributions for $\theta_i^{(p)}$ and $\theta_i^{(v)}$
     - uncorrelated $\theta_i^{(p)}$ and $\theta_i^{(v)}$

$$\begin{aligned} \theta_i^{(p)} \sim \mathsf{N}\{\theta^{(p)}, \tau_p^{-1}\} \ \text{ and } \ \theta_i^{(v)} \sim \mathsf{N}\{\theta^{(v)}, \tau_v^{-1}\}, \ \text{ or} \\ \log\theta_i^{(p)} \sim \mathsf{N}\{\log\theta^{(p)}, \tau_p^{-1}\} \ \text{ and } \ \log\theta_i^{(v)} \sim \mathsf{N}\{\log\theta^{(v)}, \tau_v^{-1}\}, \end{aligned}$$

       where τ_p and τ_v are the precision parameters. We put constraints on the range of $\theta_i^{(p)}$ and $\theta_i^{(v)}$ when considering the normal distribution model so that $\theta_i^{(p)} > 0$ and $\theta_i^{(v)} > 0$.

     - correlated $\theta_i^{(p)}$ and $\theta_i^{(v)}$

$$\begin{aligned} \begin{pmatrix} \theta_i^{(p)} \\ \theta_i^{(v)} \end{pmatrix} \sim \mathsf{BVN}\Big\{ \begin{pmatrix} \theta^{(p)} \\ \theta^{(v)} \end{pmatrix}, \Omega \Big\}, \ \text{ or} \\ \begin{pmatrix} \log\theta_i^{(p)} \\ \log\theta_i^{(v)} \end{pmatrix} \sim \mathsf{BVN}\Big\{ \begin{pmatrix} \log\theta^{(p)} \\ \log\theta^{(v)} \end{pmatrix}, \Omega \Big\} \end{aligned}$$

       where Ω is the precision matrix. Again we positively constrain on $\theta_i^{(p)}$ and $\theta_i^{(v)}$ when considering the bivariate normal distribution model.

3. The prior distributions

- the priors commonly used in all models are:

$$\alpha \sim \text{Gamma}\{0.001, 0.001\}, \quad \beta \sim \text{Gamma}\{0.001, 0.001\},$$
$$A \sim \text{N}\{0, 10,000\}, \quad B \sim \text{N}\{0, 10,000\}, \quad \text{and}$$
$$\tau_a, \tau_b, \tau_c, \tau_A \quad \text{and} \quad \tau_B \sim \text{Gamma}\{0.001, 0.001\}$$

- the prior distributions on the precision parameters for uncorrelated $\theta_i^{(p)}$ and $\theta_i^{(v)}$ models are:

$$\tau_p \sim \text{Gamma}\{0.001, 0.001\} \quad \text{and} \quad \tau_v \sim \text{Gamma}\{0.001, 0.001\}$$

- the prior distribution on the precision matrix Ω in correlated $\theta_i^{(p)}$ and $\theta_i^{(v)}$ models is:

$$\Omega \sim \text{Wishart}\Big\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \rho \Big\},$$

  where $\rho = 2$ is degrees of freedom.

- the prior distributions on $\theta^{(p)}$ and $\theta^{(v)}$ are:

$$\theta^{(p)} \sim \text{N}\{0, 10,000\} \quad \text{and} \quad \theta^{(v)} \sim \text{N}\{0, 10,000\}$$

  with the constraint $I(\theta^{(p)} > 0, \theta^{(v)} > 0)$ depending on models (see Table 4).

- the prior distributions on $\log \theta^{(p)}$ and $\log \theta^{(v)}$ are:

$$\log \theta^{(p)} \sim \text{N}\{0, 10,000\} \quad \text{and} \quad \log \theta^{(v)} \sim \text{N}\{0, 10,000\}$$

  with the constraint $I(\log \theta^{(p)} > 0, \log \theta^{(v)} > 0)$.

References

Fuller, W. A. *Measurement Error Models*. Wiley (1987).

Lundquist, F. "Aspects of the biochemistry of human semen." *Acta Physiologica Scandinavica*, 19:1–95 (1949).

Ndovi, T., Parsons, T., and Hendrix, C. "Estimating seminal vesicle (SV) and prostate (PR) gland concentrations of drugs using linear regression of gland-specific biochemical markers in split ejaculate fractions." In *American Society for Clinical Pharmacology and Therapeutics Annual Meeting, Orlando, March 2005*, Abstract PI–46 (2005).

Ndovi, T. T. "Compartmental kinetics of antiretroviral drugs (ARVs) in the human male genital tract." Ph.D. thesis, The Johns Hopkins University, Baltimore (2005).

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.5*. Cambridge, UK: MRC Biostatistics Unit (1996).

Winsor, C. P. "Which regression?" *Biometrics*, 2:101–109 (1946).

	ID	Fractions 1 & 2	Fractions 1 & 3	Fractions 2 & 3	Mean	Median
A	8	37.8	34.5	4.3	25.6	34.5
	9	141.3	123.8	12.8	92.6	123.8
	11	24.6	18.3	1.9	14.9	18.3
B	8	15.1	19.5	30.6	21.7	19.5
	9	13.1	14.9	26.5	18.2	14.9
	11	12.5	14.2	15.6	14.1	14.2
$\theta^{(p)}$	8	-8.1	1.7	26.4	6.7	1.7
	9	-8.6	4.7	26.3	7.5	4.7
	11	21.4	3.4	-5.2	6.5	3.4
$\theta^{(v)}$	8	36.6	16.3	-34.3	6.2	16.3
	9	66.7	62.4	34.3	54.5	62.4
	11	55.3	70.3	81.3	69	70.3
$\theta^{(p)}/\theta^{(v)}$	8	-0.22	0.1	-0.77	-0.3	-0.22
	9	-0.13	0.07	0.77	0.24	0.07
	11	0.39	0.05	-0.06	0.12	0.05

Table 1: The estimate of A, B, $\theta^{(p)}$, $\theta^{(v)}$ and $\theta^{(p)}/\theta^{(v)}$ using Lundquist's method for some subjects who have 3 fractions.

ID	Estimates of A		Estimates of B	
	regress a_j on b_j (Itc)	regress b_j on a_j $(-\frac{Itc}{Slp})$	regress b_j on a_j (Itc)	regress a_j on b_j $(-\frac{Itc}{Slp})$
2	53.9	54.8	27	27.4
4	75.5	85.6	28.4	30.9
8	33.7	35.4	17.5	18
9	125.8	130.4	14	14.2
11	17.9	19.3	13.6	13.9
12	123.1	132.9	28.1	29.5
13	84	84.6	7.2	7.2
14	149.5	152.7	13.7	14.3
15	97.8	102.4	7.4	7.7
16	25.5	26.8	28.7	32.1
17	126.5	144.1	23.3	25.1
18	25.4	0.1	0	-6
19	79.9	84.9	20.1	25.4

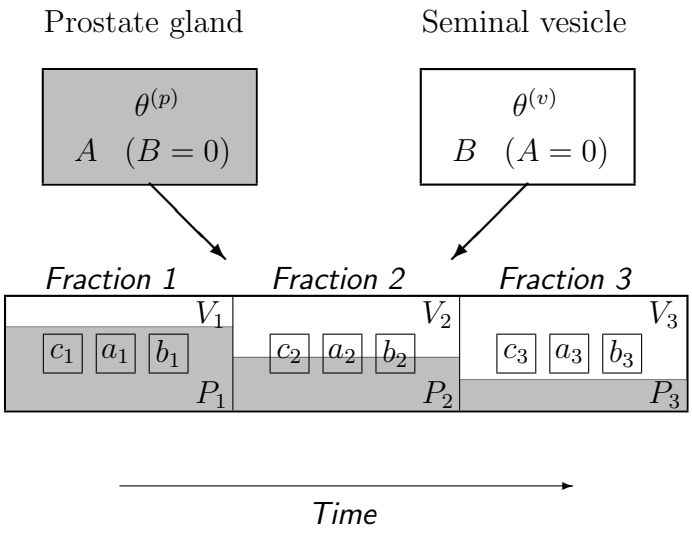
Table 2: The estimates of A and B using different regressions. Here Itc refers to the estimated intercept of the regression model and Slp refers to the estimated slope.

ID	A	B	$\theta^{(p)}$	$\theta^{(v)}$	$\theta^{(p)}/\theta^{(v)}$
2	53.9	27.0	6.6	75.1	0.09
4	75.5	28.4	38.8	24.5	1.58
8	33.7	17.5	0.5	23.7	0.02
9	125.8	14.0	0.4	64.0	0.01
11	17.9	13.6	5.2	65.3	0.08
12	123.1	28.1	63.5	111.7	0.57
13	84.0	7.2	40.6	89.6	0.45
14	149.5	13.7	1.2	125.7	0.01
15	97.8	7.4	27.2	66.5	0.41
16	25.5	28.7	33.9	223.0	0.15
17	126.5	23.3	81.6	27.9	2.93
18	25.4	0.0	174.6	0.4	421.57
19	79.9	20.1	39.6	28.6	1.39

Table 3: The estimates of A, B, $\theta^{(p)}$, $\theta^{(v)}$ and $\theta^{(p)}/\theta^{(v)}$ using the two-stage regression approach where A and B are estimated using both regressions.

		Left Panel	Right Panel
Distribution of $\theta_i^{(p)}$ and $\theta_i^{(v)}$	Top	BVN	normal
Prior of $\theta^{(p)}$ and $\theta^{(v)}$	Panel	diffuse	diffuse
Distribution of $\theta_i^{(p)}$ and $\theta_i^{(v)}$	Middle	BVN w/ pos. constraint	normal w/ pos. constraint
Prior of $\theta^{(p)}$ and $\theta^{(v)}$	Panel	diffuse w/ pos. constraint	diffuse w/ pos. constraint
Distribution of $\theta_i^{(p)}$ and $\theta_i^{(v)}$	Bottom	bivariate log-normal	log-normal
Prior of $\log \theta^{(p)}$ and $\log \theta^{(v)}$	Panel	diffuse w/ pos. constraint	diffuse w/ pos. constraint

Table 4: A summary of model specifications and layout corresponding to Figures 6 and 7. The position in the matrix of models above corresponds to the row and column for the figures.



c_j : observed drug concentration (or viral load) in fraction j, $j = 1, 2, 3$
a_j : observed PSA concentration in fraction j
b_j : observed fructose concentration in fraction j
w_j: total volume of ejaculate fluid in fraction j (measured)
P_j: volume of prostate fluid in fraction j (unknown)
V_j: volume of seminal vesicle fluid in fraction j (unknown)
A: PSA concentration in the prostate gland (unknown)
B: Fructose concentration in the seminal vesicle (unknown)
$\theta^{(p)}$: drug concentration in the prostate gland (unknown)
$\theta^{(v)}$: drug concentration in the seminal vesicle (unknown)

Figure 1: Illustration of Lundquist's equations. The gray portion in each fraction originates from the prostate gland, whereas the white portion originates from the seminal vesicle. The a_j originates from A diluted by factor of P_j/w_j (% of the gray portion) and likewise the b_j originates from B diluted by factor of V_j/w_j (% of the white portion) . The observed drug concentration c_j is the mixture of two drug concentrations contributed from the prostate gland and the seminal vesicle: one originates from $\theta^{(p)}$ diluted by factor P_j/w_j and the other originates from $\theta^{(v)}$ diluted by factor V_j/w_j.

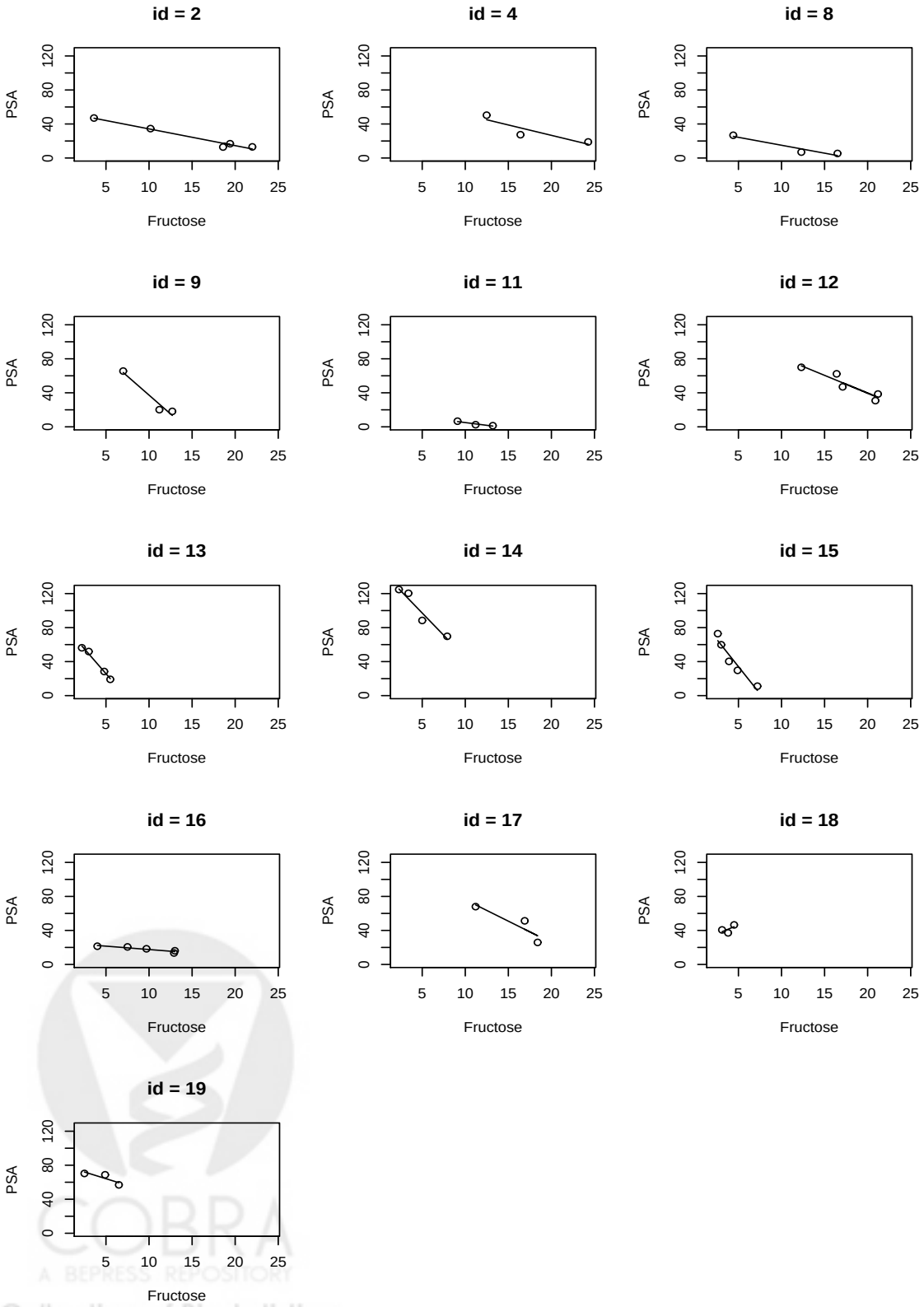


Figure 2: Scatter plot and fitted regression line for PSA on fructose by individual.

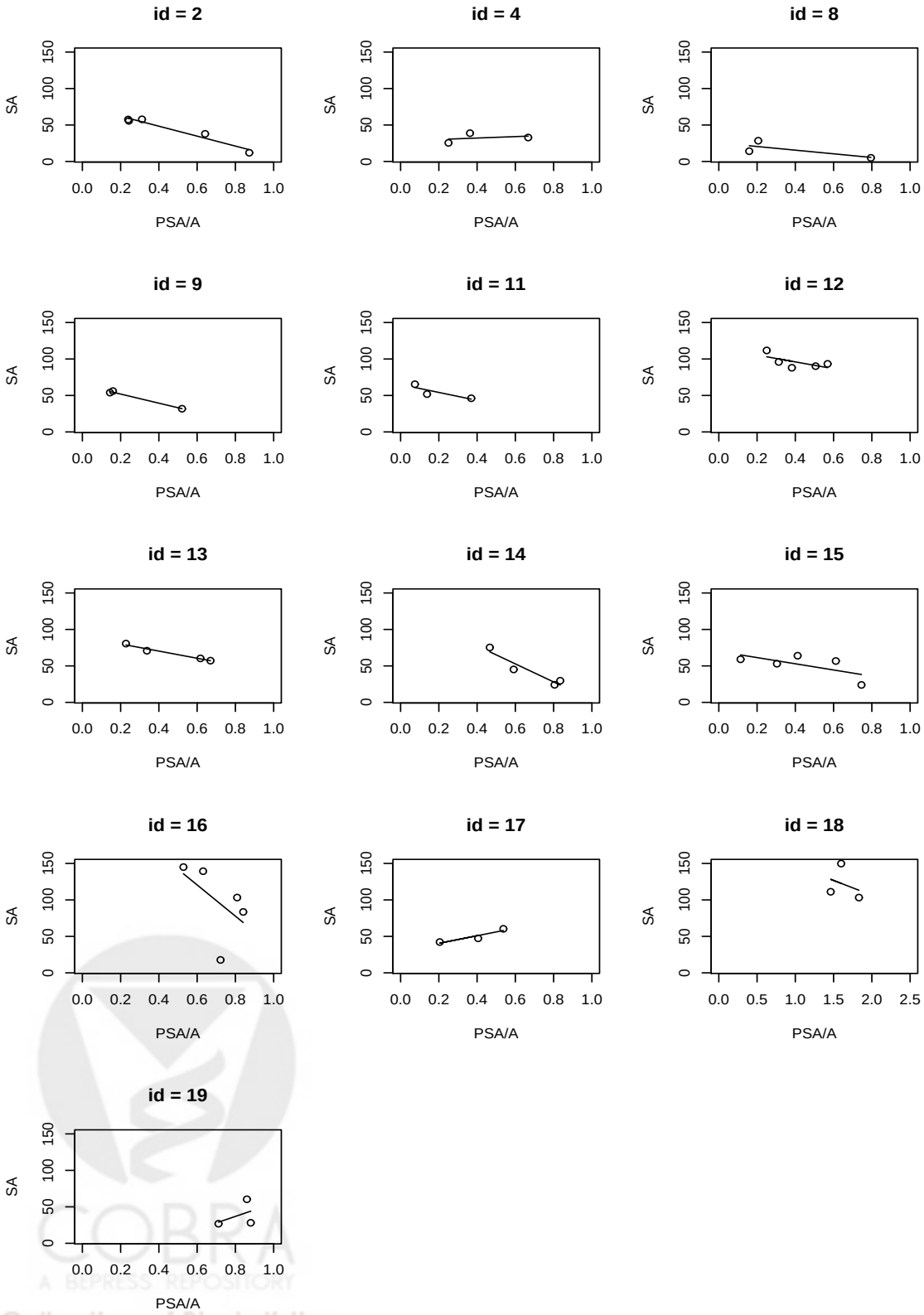


Figure 3: Scatter plot and fitted regression line for SA on PSA/$\hat{A}$ by individual.

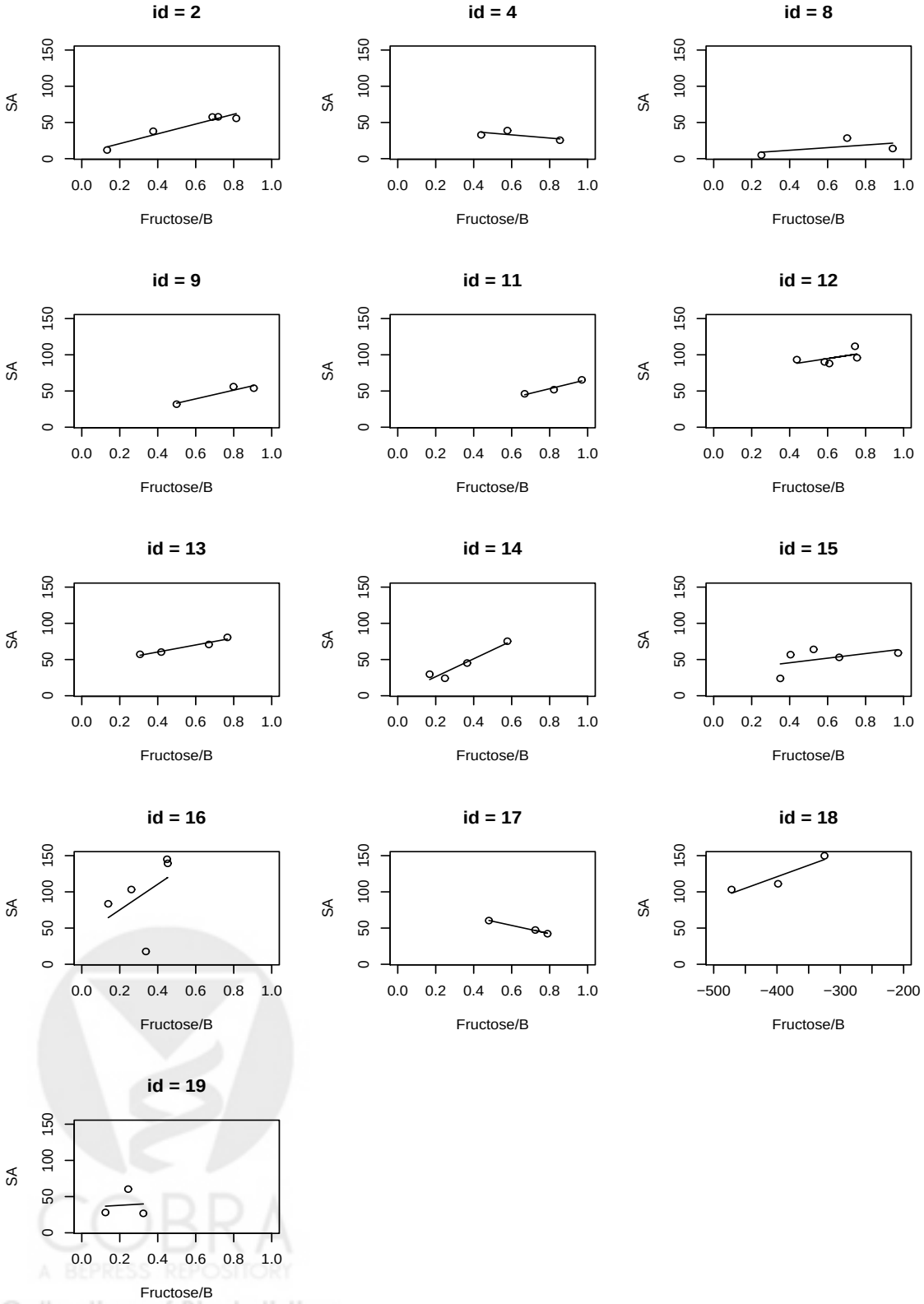


Figure 4: Scatter plot and fitted regression line for SA on fructose/$\hat{B}$ by individual.

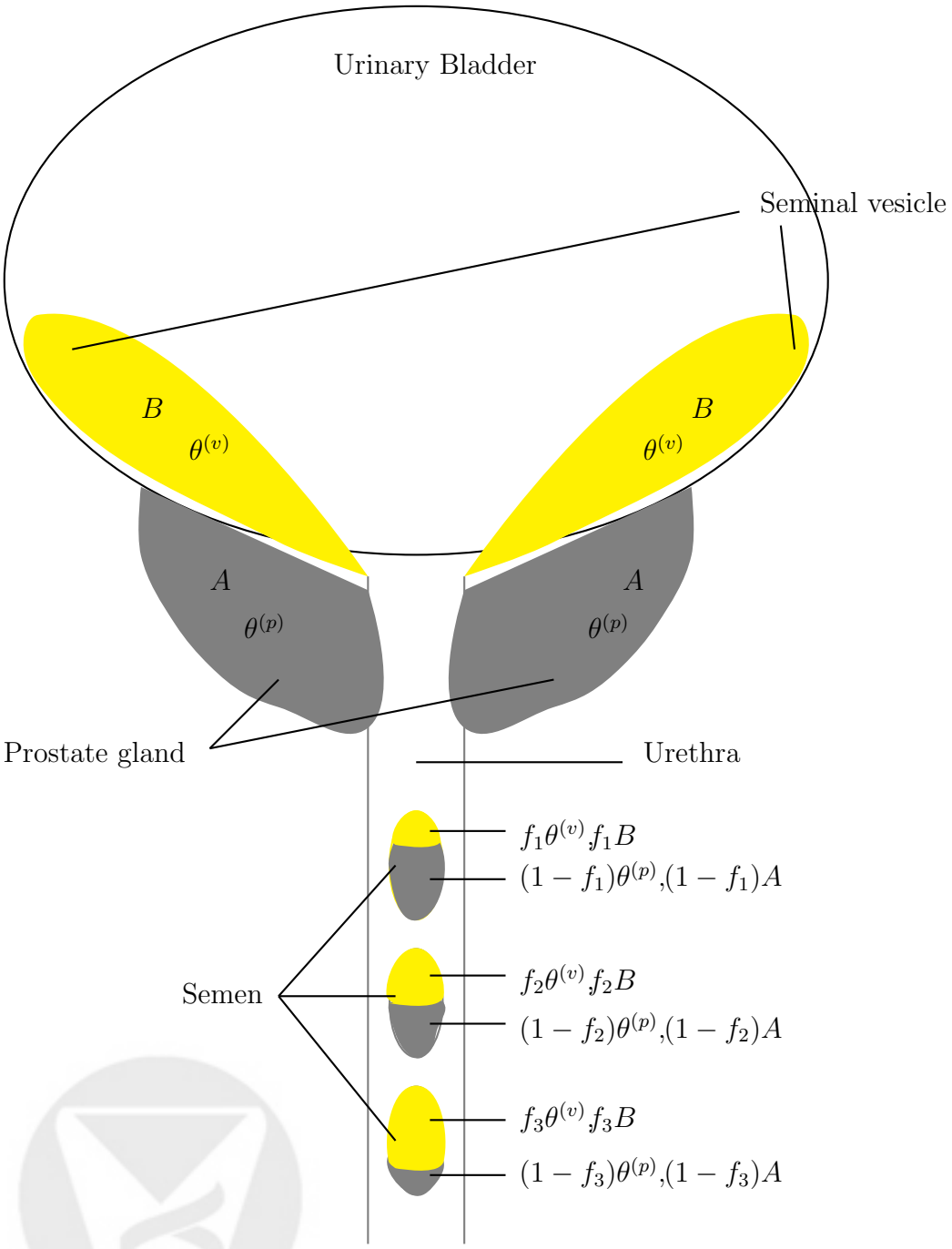


Figure 5: Simplified anatomical diagram of accessory reproductive organs in males. A latent variable f_j, $j = 1, 2, 3$ represents the fraction of contribution from the seminal vesicle in semen. The fraction of contribution from the prostate gland in semen is represented by $1 - f_j$. The parameter A represents the PSA concentration in the prostate gland while the parameter B represents the fructose concentration in the seminal vesicle. The parameter $\theta^{(p)}$ represents the drug concentration in the prostate gland while $\theta^{(v)}$ represents the drug concentration in the seminal vesicle.

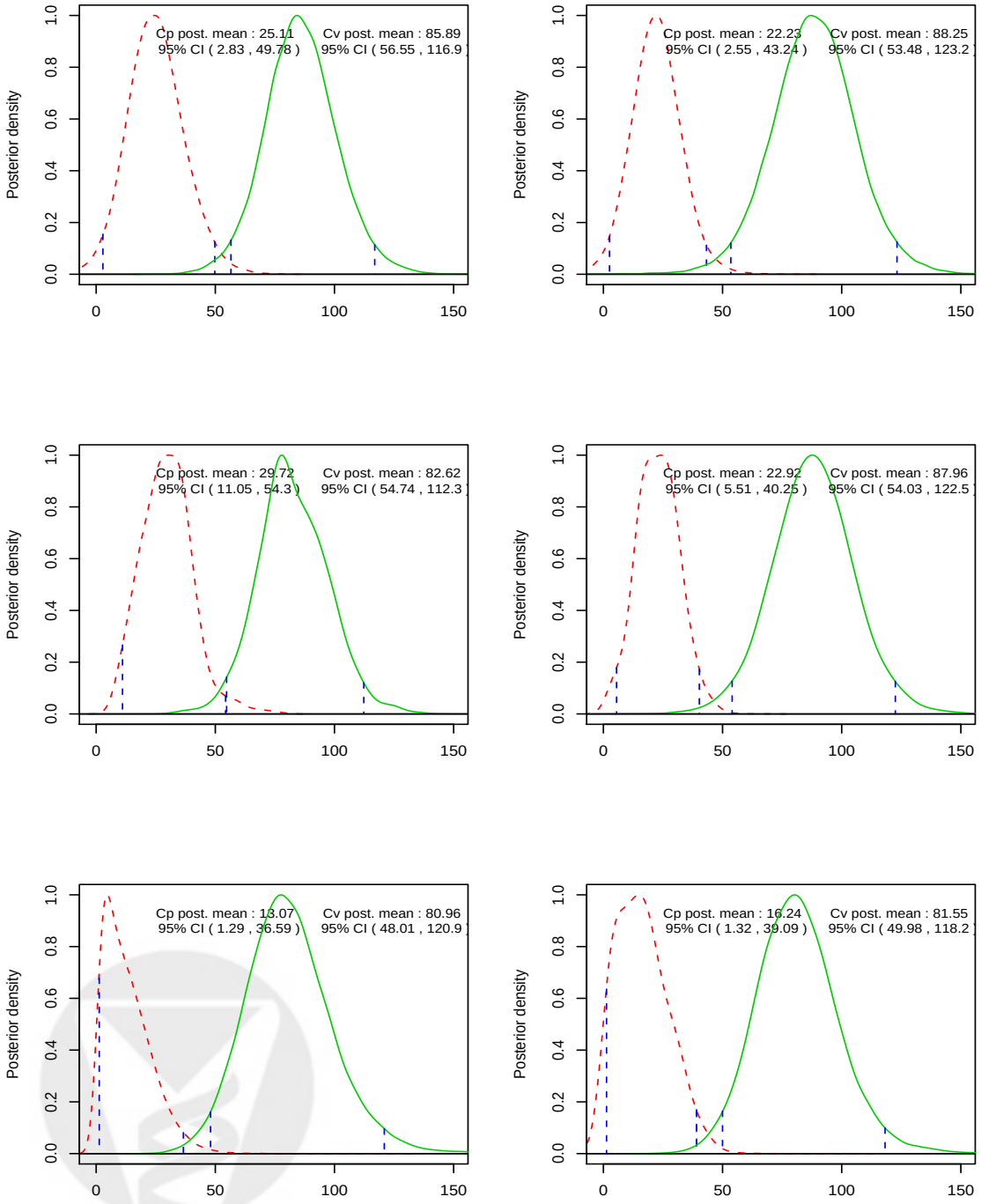


Figure 6: The posterior distributions of $\theta^{(p)}$ (labelled Cp in the plot) and $\theta^{(v)}$ (labelled Cv in the plot) using the models specified in Table 4.

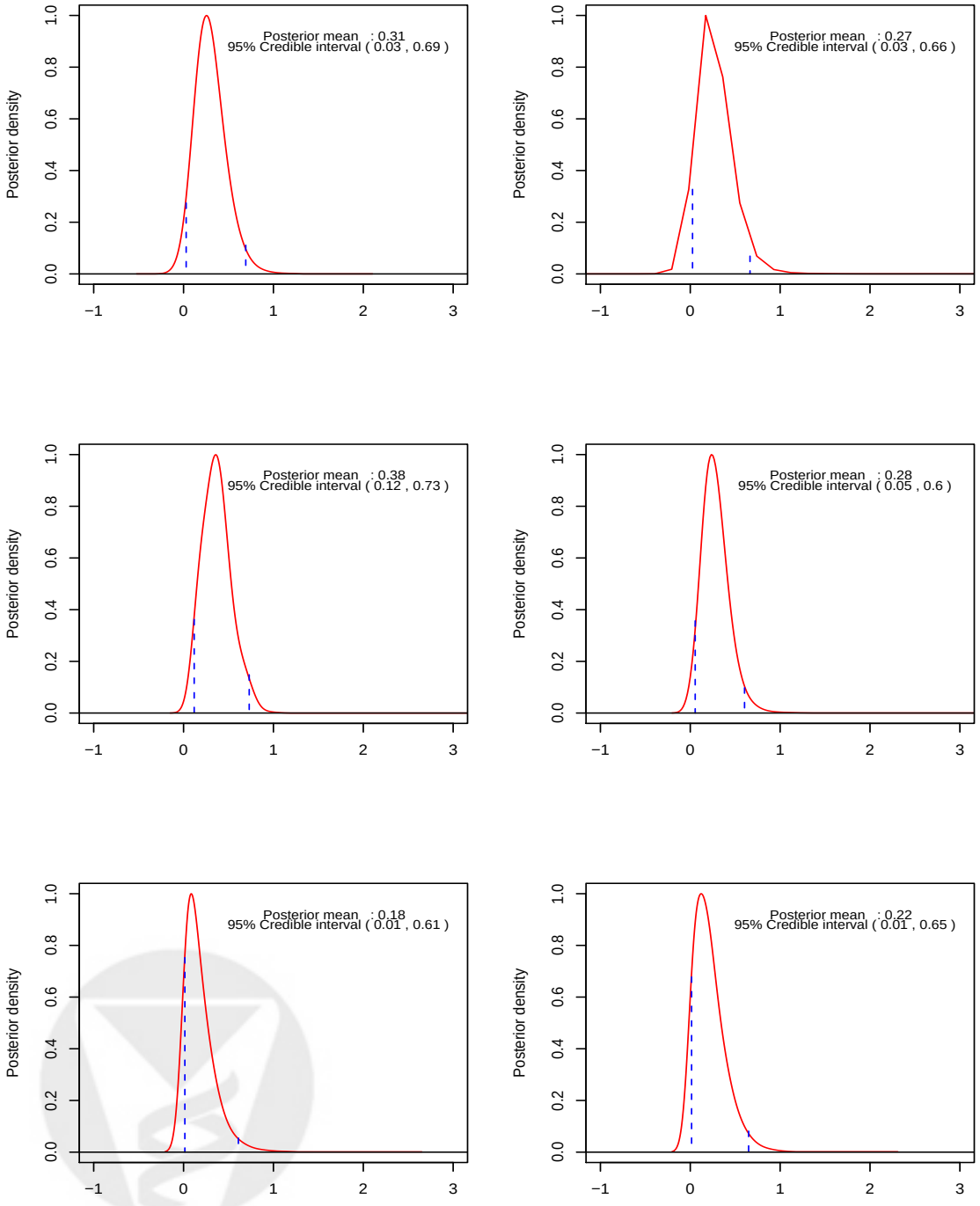


Figure 7: The posterior distributions of $\theta^{(p)}/\theta^{(v)}$ using the models specified in Table 4.

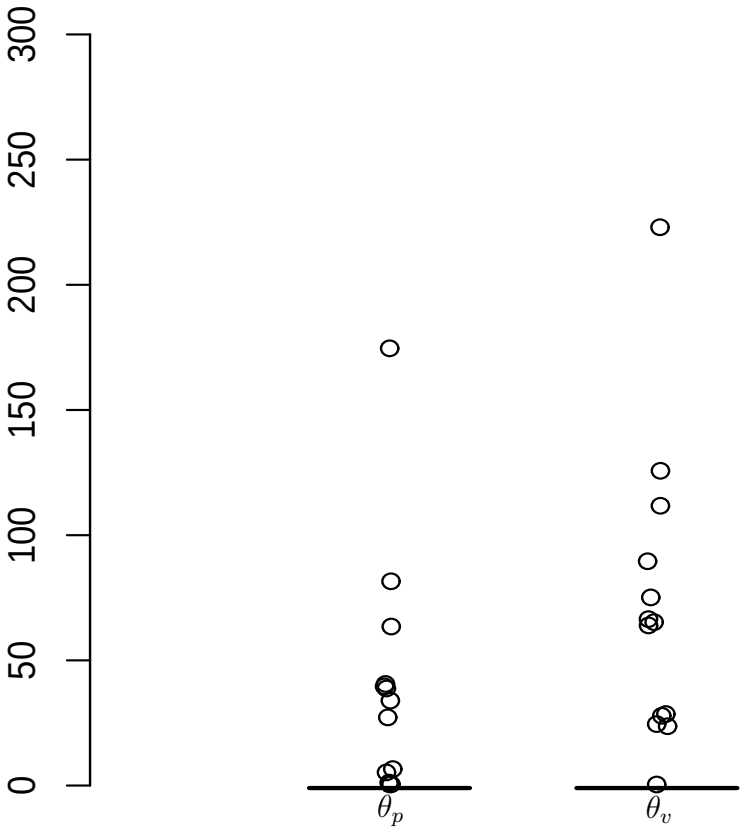


Figure 8: Jittered plot of $\hat{\theta}_p$ and $\hat{\theta}_v$ estimated by the two-stage regression approach for 13 subjects.

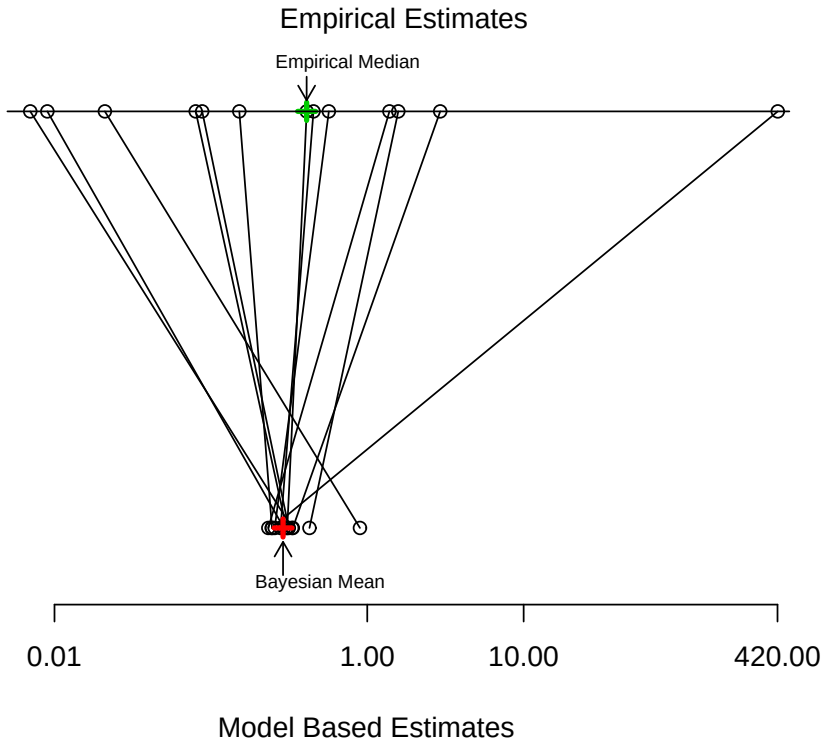


Figure 9: The comparison of the estimated ratio of $\theta^{(p)}$ and $\theta^{(v)}$. The dots in the upper line present the empirical estimates of the ratio of $\hat{\theta}_p$ and $\hat{\theta}_v$ for each subject using the from the two-stage regression approach and the dots in the bottom line present the posterior mean of the ratio of $\theta^{(p)}$ and $\theta^{(v)}$ for each subject using the normal correlated $\theta_i^{(p)}$ and $\theta_i^{(v)}$ model without constraint (left top model in Table 4). The cross and arrow in the upper line presents the median of the empirical estimates of the ratio whereas the cross and arrow in the bottom line presents the posterior mean of the ratio of $\theta^{(p)}$ and $\theta^{(v)}$ for population.