Published in final edited form as:

Stat Med. 2008 June 30; 27(14): 2697–2714. doi:10.1002/sim.3097.

A Mechanistic Latent Variable Model for Estimating Drug Concentrations in the Male Genital Tract: A Case Study in Drug Kinetics

Leena Choi^{1,*}, Brian Caffo², Charles Rohde², Themba T. Ndovi³, and Craig W. Hendrix³

¹ Department of Biostatistics, School of Medicine, Vanderbilt University

² Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University

³ Division of Clinical Pharmacology, Department of Medicine, School of Medicine, Johns Hopkins University

Summary

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. Suboptimal levels of the antiretroviral drugs may not be able to fully suppress the virus in that gland, leading to a source of sexually transmissible virus and increasing the chance of selecting for a drug resistant virus. This information may be useful for selecting an antiretroviral drug regimen that will achieve optimal concentrations in the genital tract glands. Using fractionally collected semen ejaculates, Lundquist [1] 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 does not provide population-averaged estimates after taking into account 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 on the distributional assumptions on the random effects and priors. The model and Bayesian approach are 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 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 variables; mechanistic models; structural models; Lundquist's method; Bayesian

1. Introduction

The purpose of this study is to develop statistical methodology to evaluate the distribution of antiretroviral drugs and viral loads in the prostate gland and the seminal vesicle. Several studies

---

* Correspondence to: Department of Biostatistics, School of Medicine, Vanderbilt University.

have shown that drug concentrations or viral loads may be differential across these subcompartments of the male genital tract, which may provide important information about treatment of the Human Immunodeficiency Virus (HIV) [2,3,4,5,6]. 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 [2,7,8]. 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. Unfortunately, however, it is extremely invasive to measure these concentrations directly.

To evaluate drug distribution in subcompartments of the male genital tract, Ndovi [9] and Ndovi *et al.* [10] used several surrogate markers in fractionally collected ejaculates; that is, binning an ejaculate over time. They then applied Lundquist's method [1] 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. Primary interest lies in evaluating population-averaged drug distributions over subcompartments of the male genital tract, which would facilitate the optimal choice of antiretroviral drug regimen in HIV infected populations.

In the half of a century since Lundquist's seminal article, there have been few statistical or experimental developments in this setting. In this manuscript we use modern statistical methods to compliment recently developed experimental methods. In specific, 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 generated (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 the mechanistic model, a latent variable is employed that connects the underlying theory developed by Lundquist [1] with biomarkers unique to accessory glands of the male genital tract.

The resulting model is implemented within a Bayesian framework. Since there is little information to validate the latent variable distributions, we performed a sensitivity analysis with several choices of prior distributions and models for the latent variable. In addition, several candidates of models were evaluated via simulation studies.

The manuscript is 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 how the mechanistic multilevel latent variable model is constructed based on the underlying scientific theory and discuss several choices of models. Section 4 describes a validation data set from an experiment investigating differential distribution of salicylic acid (the primary metabolite of aspirin) and implementation of our mechanistic model. Section 5 describes our model checking methods while Section 6 makes inferences based on the results. A summary and discussion follows in Section 7.

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 [11]. Therefore, the semen in the first bin should represent early-stage ejaculate. After collection, a technician takes the sequenced ejaculate and measures 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 [1] was the first to develop a systematic approach to solving this inverse problem. His method requires 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 concentrations.

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. Therefore, for example, they assume that at some point in between these extremes, 30% originates from the prostate gland; hence by Assumption 1 the remaining 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 subject-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,

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

which yields,

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

To identify (2), the organ-specific volumes, P_j and V_j , which cannot be measured directly, can be estimated by two biochemical markers, prostate specific antigen (PSA) and fructose. These markers are known to originate exclusively from the prostate gland and the seminal vesicle respectively. 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:

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

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. Rearranging

this equation yields $\frac{P_j}{w_j} = \frac{a_j}{A}$ and $\frac{V_j}{w_j} = \frac{b_j}{B}$. That is, the proportional contribution of the prostate gland (seminal vesicle) to the j th 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.

The unknowns A and B can be estimated using Assumption 1 which states that the sum of the contributions from each subcompartment is 1. Thus, $P_j/w_j + V_j/w_j = 1$ and hence $a_j/A + b_j/B = 1$. Therefore, with two observed fractions, say $j = 1$ and 2, the last equation yields the unique

solution, which is $\widehat{B} = \frac{a_1 b_2 - a_2 b_1}{a_1 - a_2}$ and $\widehat{A} = \frac{a_1}{1 - \frac{b_1}{\widehat{B}}}$ or equivalently, these estimates yield estimates, say $\widehat{P}_j$ and $\widehat{V}_j$, 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 is used ($j = 1, 2, j = 1, 3$ or $j = 2, 3$), which is a major limitation of this method. To illustrate the difference, solutions for the ratio of $\theta^{(p)}$ and $\theta^{(v)}$ using this method on validation data described in Section 4 are presented in Table I with comparisons to the results from our new approach.

Despite these limitations, Lundquist's method provides the physiologic theory for a modern statistical approach. In specific, we use Lundquist's model for the male genital tract to develop a mechanistic component of a stochastic model. This component is crucial, as small sample sizes eliminate the ability to rely on purely empirical approaches.

3. Mechanistic Multilevel Latent Variable Model

In this section we develop 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 on the prior notation, let a_{ij} , b_{ij} and c_{ij} represent the concentration of PSA, fructose and the drug/virus, respectively, on subject $i = 1, \dots, I$ for fraction $j = 1, \dots, J_i$. 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 introduce a latent variable, f_{ij} , which represents V_{ij}/w_{ij} in Lundquist's Equations, the fractional contribution of the total ejaculate from the seminal vesicle for subject i . Note that, because of Assumption 1, $(1 - f_{ij})$ represents P_{ij}/w_{ij} , the fractional contribution from the prostate gland for subject i . We summarize all of this notation in the Appendix I. We assume the following model on the observed concentrations

$$\begin{aligned} a_{ij} &| A_i, f_{ij} \sim N\{A_i(1 - f_{ij}), \sigma_a^2\} \\ b_{ij} &| B_i, f_{ij} \sim N(B_i f_{ij}, \sigma_b^2) \\ c_{ij} &| a_{ij}, b_{ij}, f_{ij} \sim N\{\theta_i^{(p)}(1 - f_{ij}) + \theta_i^{(v)} f_{ij}, \sigma_c^2\}. \end{aligned}$$

The mean model in these distributions for each subject, is motivated from the main structure of Lundquist's Equations (1) and (3). Figure 2 shows a simplified anatomical diagram of the accessory reproductive organs in males along with the parameters and the latent variable f_{ij} .

On the next level, the unobserved variables A_i and B_i are assumed to follow normal distributions with means A and B and variances σ_A^2 and σ_B^2 respectively. That is,

$$A_i \sim N(A, \sigma_A^2) \text{ and } B_i \sim N(B, \sigma_B^2).$$

Notice that the latent variable, f_{ij} , is bounded between zero and one. We considered two distributions for this latent variable. First, we assumed that f_{ij} is distributed as Beta(α, β). Secondly, we considered logit models that make more direct use of the temporal information treating $\text{logit}(f_{ij})$ as a linear function of the order of the fraction, j , normalized by the total

number of fractions per subject, J_i . Specifically, $\text{logit}(f_{ij}) = \lambda_{1i} + \lambda_{2i} \frac{j}{J_i}$ where λ_{1i} and λ_{2i} are assumed to be bivariate normal distribution. Under this formulation, the Beta model is quite similar to specifying $\text{logit}(f_{ij}) = a_{ij}$ and $a_{ij} \sim \text{Normal}$. Further details on the model specification can be found in Appendix II.

It is assumed that A_i , B_i and f_{ij} are mutually independent of each other and of $\theta_i^{(p)}$ and $\theta_i^{(v)}$ since there is no physiological reason to suspect that PSA, fructose and the drug are correlated.

Being central quantities of interest, several distributional assumptions for $\theta_i^{(p)}$ and $\theta_i^{(v)}$ were investigated. Specifically, bivariate normals and bivariate log-normal distributions were considered

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

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 is available to go into the other. These effects could also cancel each other out, or one direction could dominate. With the small number of subjects in this kind of study, it is very difficult 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 functions for models can be found in Appendix II.

4. A Validation Data Set and Implementation

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 was administered to 13 healthy male volunteers [9]. This drug was chosen based on its safety and the negative logarithm of the ionization constant ($pK_a = 2.8$), of its primary metabolite, salicylic acid (SA), a weak acid. Physiologic theory predicts that the concentrations of a drug differ across body compartments based on the pK_a of the drug and the pH of the compartments. That is, drugs will move passively based on physical characteristics of the drugs, unless other mechanisms, such as an active membrane transport or protein binding are involved. According to the $pH - pK_a$ partition hypothesis, an acidic drug traveled to basic environment will be ionized and stay there, because the ionized form of the drug has difficulty passing through the membrane. Hence the concentration of an acidic drug will be higher in a gland with a basic pH .

The pH of the two organs of interest 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 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. The raw data are shown in Figure 3 along with the fitted values from the proposed mechanistic latent variable model.

There are several challenges that prevent automated fitting our mechanistic latent variable model 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} ($\text{Beta}\{\alpha, \beta\}$), is not standard in commercially available statistical packages. Third, the latent variable model with f_{ij} ($\text{Beta}\{\alpha, \beta\}$) is indexed by both the subject level, i , and the within subject level, j . To address these computational challenges, we explored maximum likelihood (ML) approximated via Gaussian quadrature implemented with our own R (www.r-project.org) program and a Bayesian framework using Markov chain Monte-Carlo (MCMC) methods implemented with WinBUGS [12].

For the Bayesian version, diffuse priors were used for the hyperparameters coincide with ML approach. To be specific, the prior distributions on $\theta^{(p)}$ and $\theta^{(v)}$ were either normal or log-normal having means of 0 and precisions (inverse variances) of 10^{-4} or 10^{-2} . Priors with and without the positivity constraints were considered. For the prior with the positivity constraints, we used a truncated normal distribution with means of 0 and precision of 10^{-4} truncated at zero. This prior was slightly informative in a sense that it incorporated the fact that the drug concentrations should be non-negative. Other than positivity constraints, we did not consider more informative priors in order to let the data be the primary arbiter of the evidence regarding differential distribution of the drug. The mean parameters, A and B , were given independent

normal priors, with means of 0 and precisions of 10^{-4} . All of the precision parameters 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, R , was used, the prior for the inverse of the correlation matrix was specified to be Wishart distribution centered at an identity matrix with 2 degrees of freedom.

We fit the validation data using the ML approach as a reference method, and compared the results with those of the Bayesian model on a subset of the models considered. ML estimates and associated standard errors were compared with the posterior means and standard deviations from the logit and the Beta models with independent normal random effects for $\theta_i^{(p)}$ and $\theta_i^{(v)}$ in Table II. The correspondence of the results for the logit model is quite good, whereas the results for the Beta model show only a slight discrepancy.

To consider this discrepancy further, notice that the Beta model requires a three to five dimensional integral per subject. Hence the slight discrepancy in the Beta model fits may be due to model numerical inaccuracies in the quadrature approximation. Computational limitations did not allow for executing larger numbers of quadrature points. Therefore, to further examine this possibility, we fit the Beta model with ten quadrature points (per dimension) using a simulated data set with three observations per subject, which is lower than the three to five observations seen in the validation data set. There, the ML and Bayesian approaches were very similar (results available on request), confirming the hypothesis that the discrepancy is likely caused by inaccuracies in the high dimensional quadrature approximation. Because of the generally good agreement between the two approaches, and the large computational burden required by the ML approach, we adopt the Bayesian method throughout.

5. Model Checking

We performed two stages of model diagnostics. In the first stage, the posterior distributions for the three fitted outcomes (PSA, fructose, and drug/viral concentrations) were compared with the corresponding three observed outcomes. This preliminary comparison, aided in producing model candidates.

Subsequently, a second exploratory tool examined goodness-of-fit for these candidates using Bayesian posterior predictive checks [13,14] through simulations. In short, we employed half-normal plots [15,16] to compare the posterior distributions of simulated data from each model to the observed data. From the posterior predictive distributions of the simulated data, we also calculated a Bayesian version of a goodness-of-fit P -value [13]. Specifically, we chose the range of each outcome as test quantities to be compared with the predictive simulations. For each test quantity, we examined the lack of fit by calculating the tail-area probability (P -value) of the test quantity, with small P -values indicating lack of fit. Further description of the method can be found in [13,14].

6. Results and Inferences

We first describe the results of the two-stage model checking described in the previous section. Especially, model comparisons focused on the Beta model versus the logit model for f_{ij} , and the independence assumption for $\theta_i^{(p)}$ and $\theta_i^{(v)}$.

Consider the first stage model checking. Figures 3 and 4 show the diagnostic plots from the Beta model where $\theta_i^{(p)}$ and $\theta_i^{(v)}$ are treated as independent. The plots for the remaining models were similar. We examine the comparison between the Beta and logit models in greater detail, since, on face value, the logit model would appear preferable. However, the opposite is true.

Compared to the Beta model, the logit model, having two or three more parameters, resulted in slightly worse fit in the diagnostic plot. This is further evidenced larger deviance information criterion (DIC) [17] for the logit model of 947 versus 919 for the Beta model. In addition, the logit model makes the biologically questionable assumption of equal increments between the fractions. Regarding the correlation between $\theta_i^{(p)}$ and $\theta_i^{(v)}$, there was no apparent difference. Thus, we employ the more parsimonious independent normal.

In second-stage diagnostics, we compared 95% credible intervals for the posterior predictive distribution of the simulated outcomes to that of the corresponding observed outcomes for the subset of models. Assuming the model selected above, we next considered the impact of the prior and random effect distributions by comparing a diffuse and a truncated diffuse priors assuming normality and log-normality for $\theta^{(v)}$ and $\theta^{(p)}$. Figure 5 shows an example goodness-of-fit plot. All of the posterior predictive means of PSA and fructose at those observed counterparts lie on the reference identity line and the credible intervals are reasonably narrow given the setting, suggesting that all models predict PSA and fructose very well. In addition, the posterior predictive means of SA from all models agreed well with the observed SA. Moreover, there was virtually no difference in the prediction of SA between the model employing a normal with a diffuse prior and the normal with the truncated diffuse prior. However, the posterior predictive distribution of SA from the log normal model presented larger variability than that of the normal models, especially for large values of SA. This would suggest defaulting to the normal model; however it has the potentially unappealing property of allowing for negative values. Table III shows the Bayesian P -value based on the test quantities. None of the Bayesian P -values show a clear discrepancy between the posterior predictive distribution of the three simulated outcomes and the corresponding observed outcomes. As these results showed only minor discrepancies between the models in the terms of fit, we report final results for all of them.

Figure 6 shows the posterior distributions, posterior means and 95% credible intervals for the parameters of interest, $(\theta^{(p)}, \theta^{(v)})$, in the left panel, and the corresponding ratio, $\theta^{(p)}/\theta^{(v)}$, in the right panel. The top, middle and bottom panels correspond to the normal model with diffuse prior, the normal model with the truncated diffuse prior and the log-normal model with diffuse prior.

The posterior distributions for $\theta^{(v)}$ are very similar across all models, whereas those for $\theta^{(p)}$ appear to be more sensitive to the modelling assumptions. The posterior distribution for $\theta^{(p)}$ from the log-normal model is shifted toward zero, compared to those of the two normal models. Of note, the log-normal model allows for a high skewness and small positive values. Thus, the posterior distribution of $\theta^{(p)}$ tends to concentrate toward the small positive values. In contrast, the normal model with the truncated normal prior is shifted toward the larger positive values. Based on the $pH - pK_a$ partition hypothesis, we expect that the drug concentration in the prostate gland $\theta^{(p)}$ should be small compared to that of the seminal vesical $\theta^{(v)}$. Thus, the distribution of $\theta^{(p)}$ is likely to be positively skewed and hence the log-normal model might be more appropriate for $\theta^{(p)}$. However, with only thirteen subjects in the data set, it is difficult for the data to choose one over the other.

Regardless, the posterior distributions for $\theta^{(p)}$ and $\theta^{(v)}$ are well separated for all models via 95% credible intervals. Also, the distribution for the ratio of $\theta^{(p)}$ to $\theta^{(v)}$ is mostly below one and the 95% credible interval does not include one for all models. The posterior means of the ratio from all models are between .14 and .26, suggesting that the concentration of SA in the seminal vesicle is about four to seven times higher than that in the prostate gland. The subject-specific estimates for the ratio of $\theta^{(p)}$ and $\theta^{(v)}$ are presented in Table I along with comparisons to the results from Lundquist's method. Although some of the subject-specific estimates from Lundquist's method are not reasonable, their median is close to the estimate from our new

approach, adding further support to the differential distribution hypothesis. Thus, the hypothesis of differential distribution of SA in the two subcompartments is supported by the data. We remind the readers 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 is basic. Despite the slight differences across models, it is important to emphasize that the posterior distributions of $\theta^{(p)}$ and $\theta^{(v)}$ are not sensitive to the distributional assumptions for the random effects and priors to change the substantive conclusions.

7. Summary and Discussion

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 ML and 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.* [11]. 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. We emphasize the importance of taking into account the variability in estimating nuisance concentration parameters. Some subjects 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 from Lundquist's method. The working physiological theory is that the ejaculate becomes too mixed in the urethra for such subjects to differentiate early and late-stage ejaculates. Current practice tends to rely on ad hoc methods, such as subjectively discarding data with perceived low information. A principle 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.

In general, the number of subjects in the data for this kind of study is small and the heterogeneity is large. Therefore, the added information from mechanistic modelling is useful for making valid inferences, as a completely empirical approach is not feasible. As it is difficult to check the distributional assumptions with such a small number of subjects, we performed sensitivity analysis using several choices of distributions for random effects and priors, and examined goodness-of-fit using model checking techniques. We found that the estimates were insensitive to these assumptions and did not change the substantive conclusions.

The statistical and experimental methods were 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, as the differential distribution was clearly evidenced, and was robust to modelling assumptions.

Appendix: I. Summary of notation

For subject $i = 1, \dots, I$ in fraction $j = 1, \dots, J_i$,

a_{ij} : observed PSA concentration

b_{ij} : observed fructose concentration

c_{ij} : observed drug concentration (or viral load)

For subject $i = 1, \dots, I$,

A_i : PSA concentration in the prostate gland (unknown)

B_i : fructose concentration in the seminal vesicle (unknown)

$\theta_i^{(p)}$: drug concentration in the prostate gland (unknown)

$\theta_i^{(v)}$: drug concentration in the seminal vesicle (unknown)

f_{ij} : V_{ij}/w_{ij} in Lundquist's Equations (unknown)

$(1 - f_{ij})$: P_{ij}/w_{ij} in Lundquist's Equations (unknown)

Appendix: II. Likelihood

Let denote $\mathbf{a} = (a_{11}, \dots, a_{ij}, \dots, a_{IJ})$, $\mathbf{b} = (b_{11}, \dots, b_{ij}, \dots, b_{IJ})$, $\mathbf{c} = (c_{11}, \dots, c_{ij}, \dots, c_{IJ})$, and $\mathbf{f}_i = (f_{i1}, \dots, f_{iJ})$.

$$\begin{aligned}
 & L(\theta^{(p)}, \theta^{(v)}, \Omega, D | \mathbf{a}, \mathbf{b}, \mathbf{c}) \\
 = & \prod_{i=1}^m \int \prod_{j=1}^{n_i} f_{a_{ij}, b_{ij}, c_{ij}} | u(a_{ij}, b_{ij}, c_{ij} | u, \Omega) f_u(u | \theta^{(p)}, \theta^{(v)}, D) du \\
 = & \prod_{i=1}^m \int \prod_{j=1}^{n_i} f_{a_{ij}, b_{ij}, c_{ij}} | u(a_{ij}, b_{ij}, c_{ij} | A_i, B_i, \theta_i^{(p)}, \theta_i^{(v)}, f_{ij}, \Omega) f_u(f_{ij}, A_i, B_i, \theta_i^{(p)}, \theta_i^{(v)} | \theta^{(p)}, \theta^{(v)}, D) du \\
 = & \prod_{i=1}^m \int \int \int \int \prod_{j=1}^{n_i} f_{a_{ij}}(a_{ij} | A_i, f_{ij}, \sigma_a^2) f_{b_{ij}}(b_{ij} | B_i, f_{ij}, \sigma_b^2) \\
 & \times f_{c_{ij}}(c_{ij} | \theta_i^{(p)}, \theta_i^{(v)}, f_{ij}, \sigma_c^2) f_{f_{ij}}(f_{ij} | \alpha, \beta) f_{A_i}(A_i | A, \sigma_A^2) f_{B_i}(B_i | B, \sigma_B^2) \\
 & \times f_{\theta_i^{(p)}, \theta_i^{(v)}}(\theta_i^{(p)}, \theta_i^{(v)} | \theta^{(p)}, \theta^{(v)}, \sigma_p^2, \sigma_v^2) df_{ij} dA_i dB_i d\theta_i^{(p)} d\theta_i^{(v)} \\
 = & \prod_{i=1}^m \int f_{a_i}(a_i | A, f_i, \sigma_A^2) f_{b_i}(b_i | B, f_i, \sigma_B^2) f_{c_i}(c_i | \theta^{(p)}, \theta^{(v)}, f_i, \sigma_p^2, \sigma_v^2, \sigma_c^2) \\
 & \times f_{f_i}(f_i) df_i
 \end{aligned} \tag{4}$$

where $u = (f_{ij}, A_i, B_i, \theta_i^{(p)}, \theta_i^{(v)})$, $\Omega = (\sigma_a^2, \sigma_b^2, \sigma_c^2)$, $D = (A, B, \sigma_A^2, \sigma_B^2, \sigma_p^2, \sigma_v^2, \rho)$, and

$$\begin{aligned}
 & f_{a_i}(a_i | A, f_i, \sigma_A^2) \\
 = & (2\pi)^{-n_i/2} |\Sigma_{a_i}|^{-1/2} \exp \left\{ -\frac{1}{2} [a_i - A(1 - f_i)]^t \Sigma_{a_i}^{-1} [a_i - A(1 - f_i)] \right\} \\
 & f_{b_i}(b_i | B, f_i, \sigma_B^2) \\
 = & (2\pi)^{-n_i/2} |\Sigma_{b_i}|^{-1/2} \exp \left\{ -\frac{1}{2} [b_i - Bf_i]^t \Sigma_{b_i}^{-1} [b_i - Bf_i] \right\} \\
 & f_{c_i}(c_i | \theta^{(p)}, \theta^{(v)}, f_i, \sigma_p^2, \sigma_v^2, \sigma_c^2) \\
 = & (2\pi)^{-n_i/2} |\Sigma_{c_i}|^{-1/2} \\
 & \times \exp \left\{ -\frac{1}{2} [c_i - \theta^{(p)}(1 - f_i) - \theta^{(v)} f_i]^t \Sigma_{c_i}^{-1} [c_i - \theta^{(p)}(1 - f_i) - \theta^{(v)} f_i] \right\} \\
 f_{f_i}(f_i): & (1) f_{f_i}(f_i) = f_{f_i}(f_i | \alpha, \beta) = \prod_{j=1}^{n_i} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} f_{ij}^{\alpha-1} (1 - f_{ij})^{\beta-1} \\
 & (2) f_{f_{ij}}(f_{ij}) \sim \text{Bernoulli}(f_{ij}), \quad \text{logit}(f_{ij}) = \lambda_{1i} + \lambda_{2i} \frac{j}{J_i} \\
 & \begin{pmatrix} \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} \sim \text{BVN} \left\{ \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}, \Sigma, \Sigma \right\}, \Sigma = \begin{pmatrix} \sigma_{\lambda_1} & \delta_{\lambda_1, \lambda_2} \\ \delta_{\lambda_1, \lambda_2} & \sigma_{\lambda_2} \end{pmatrix}
 \end{aligned}$$

where

$$\begin{aligned}
\Sigma_{a_i} &= \sigma_a^2 I_{n_i} + \sigma_a^2 (1 - f_i)(1 - f_i)^t, \\
\Sigma_{b_i} &= \sigma_b^2 I_{n_i} + \sigma_b^2 f_i f_i^t, \text{ and} \\
\Sigma_{c_i} &= \sigma_c^2 I_{n_i} + \sigma_p^2 (1 - f_i)(1 - f_i)^t + \sigma_v^2 f_i f_i^t \text{ for independent normal model} \\
\Sigma_{c_i} &= \sigma_c^2 I_{n_i} + \sigma_p^2 (1 - f_i)(1 - f_i)^t + \sigma_v^2 f_i f_i^t + \rho \sigma_p \sigma_v (1 - f_i) f_i^t \\
&\quad + \rho \sigma_p \sigma_v f_i (1 - f_i)^t \text{ for bivariate normal model.}
\end{aligned}$$

The inverse and determinant of covariance matrices are following:

$$\begin{aligned}
\Sigma_{a_i}^{-1} &= \frac{1}{\sigma_a^2} I_{n_i} - \frac{(\sigma_a^2 / \sigma_a^4)(1 - f_i)(1 - f_i)^t}{1 + (\sigma_a^2 / \sigma_a^4)(1 - f_i)^t(1 - f_i)} \\
\Sigma_{b_i}^{-1} &= \frac{1}{\sigma_b^2} I_{n_i} - \frac{(\sigma_b^2 / \sigma_b^4) f_i f_i^t}{1 + (\sigma_b^2 / \sigma_b^4) f_i^t f_i} \\
\Sigma_{c_i}^{-1} &= V^{-1} - \frac{(\sigma_p^2) V^{-1} f_i f_i^t V^{-1}}{1 + \sigma_v^2 f_i^t V^{-1} f_i}
\end{aligned}$$

$$\text{where } V^{-1} = \frac{1}{\sigma_c^2} I_{n_i} - \frac{(\sigma_p^2 / \sigma_c^4)(1 - f_i)(1 - f_i)^t}{1 + (\sigma_p^2 / \sigma_c^4)(1 - f_i)^t(1 - f_i)} \text{ for independent normal model}$$

$$\begin{aligned}
|\Sigma_{a_i}| &= (\sigma_a^2)^{(n_i-1)} (\sigma_a^2 + \sigma_a^2 (1 - f_i)^t (1 - f_i)) \\
|\Sigma_{b_i}| &= (\sigma_b^2)^{(n_i-1)} (\sigma_b^2 + \sigma_b^2 f_i^t f_i).
\end{aligned}$$

The likelihoods for log-normal $\theta_i^{(p)}$ and $\theta_i^{(v)}$ models are straightforward.

References

1. Lundquist F. Aspects of the biochemistry of human semen. *Acta Physiologica Scandinavica* 1949;19:1–95.
2. Ghosn J, Chaix ML, Peytavin G, Rey E, Bresson JL, Goujard C, Katlama C, Viard JP, Treluyer JM, Rouzioux C. Penetration of enfuvirtide, tenofovir, efavirenz, and protease inhibitors in the genital tract of HIV-1-infected men. *AIDS* 2004;18(14):1958–1961. [PubMed: 15353984]
3. Cruciani M, Liuzzi G, Chirianni A, Audagnotto S, Bonora S, Di Biagio A, Sinicco A, Bassetti M, Gatti G. Penetration of didanosine in semen of HIV-1-infected men. *Journal of Antimicrobial Chemotherapy* 2006;57(6):1244–1247. [PubMed: 16556633]
4. Taylor S, Back DJ, Drake SM, Workman J, Reynolds H, Gibbons SE, White DJ, Pillay D. Antiretroviral drug concentrations in semen of HIV-infected men: differential penetration of indinavir, zidovudine and saquinavir. *Journal of Antimicrobial Chemotherapy* 2001;48(3):351–354. [PubMed: 11532998]
5. Reddy YS, Gotzkowsky SK, Eron JJ, Kim JY, Fiske WD, Fiscus SA, Petch L, Cohen MS, Kashuba AD. Pharmacokinetic and pharmacodynamic investigation of efavirenz in the semen and blood of human immunodeficiency virus type 1-infected men. *Journal of Infectious Diseases* 2002;186(9):1339–43. [PubMed: 12402205]
6. van Praag RM, Repping S, de Vries JW, Lange JM, Hoetelmans RM, Prins JM. Pharmacokinetic profiles of nevirapine and indinavir in various fractions of seminal plasma. *Antimicrobial Agents and Chemotherapy* 2001;45(10):2902–2907. [PubMed: 11557488]
7. Solas C, Lefeuvre A, Halfon P, Chadapaud S, Hittinger G, Lacarelle B. Discrepancies between protease inhibitor concentrations and viral load in reservoirs and sanctuary sites in human immunodeficiency virus-infected patients. *Antimicrobial Agents and Chemotherapy* 2003;47(1):238–243. [PubMed: 12499197]

8. Kepler TB, Perelson AS. Drug concentration heterogeneity facilitates the evolution of drug resistance. *Proceedings of the National Academy of Sciences of the United States of America* 1998;95(20):11514–11519. [PubMed: 9751697]
9. Ndovi, TT. Compartmental kinetics of antiretroviral drugs (ARVs) in the human male genital tract. The Johns Hopkins University; Baltimore: 2005.
10. Ndovi TT, Choi L, Caffo B, Parsons T, Baker S, Zhao M, Rohde C, Hendrix CW. Quantitative assessment of seminal vesicle and prostate drug concentrations by use of a noninvasive method. *Clinical Pharmacology and Therapeutics* 2006;80(2):146–158. [PubMed: 16890576]
11. Ndovi TT, Parsons T, Choi L, Caffo B, Rohde C, Hendrix CW. A new method to estimate quantitatively seminal vesicle and prostate gland contributions to ejaculate. *British Journal of Clinical Pharmacology* 2007;63(4):404–420. [PubMed: 17076697]
12. Spiegelhalter, DJ.; Thomas, A.; Best, NG.; Gilks, WR. BUGS: Bayesian inference Using Gibbs Sampling, Version 0.5. MRC Biostatistics Unit; Cambridge, UK: 1996.
13. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. Chapman & Hall; London: 1995.
14. Gelman A. Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics* 2004;13(4):755–779.
15. Demétrio CGB, Hinde JP. Half-normal plots and overdispersion. *GLIM Newsletter* 1997;27:19–26.
16. Choi L, Dominici F, Zeger SL, Ouyang P. Estimating treatment efficacy over time: a logistic regression model for binary longitudinal outcomes. *Statistics in Medicine* 2005;24(18):2789–2805. [PubMed: 16134133]
17. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2002;64(4):583–616.

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
 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.

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 3.

The observed outcomes (PSA, fructose, and SA concentrations) for each subject vs. the posterior means of the fitted outcomes from the Beta model for f_{ij} and independent normal for $\theta^{(p)}$ and $\theta^{(v)}$ (with an identity line for reference).

Beta f_{ij} and independent normal for $\theta^{(p)}$ and $\theta^{(v)}$ with diffuse prior

Figure 4.

95% credible intervals of the posterior distribution for the fitted outcomes (PSA, fructose, and SA concentrations) versus the corresponding observed outcomes (dots: posterior means; gray bars: 2.5th - 97.5th percentiles; solid line: the line of agreement). The results from other candidate models are similar and not reported.

Figure 5.

95% credible intervals of the posterior predictive distribution for the simulated outcomes versus the corresponding observed outcomes (dots: posterior means; gray bars: 2.5th - 97.5th percentiles; solid line: the line of agreement). The results from other candidate models are similar and not reported.

Figure 6.

The posterior distributions of $\theta^{(p)}$ and $\theta^{(v)}$ (left panel), and the posterior distribution of the ratio of $\theta^{(p)}$ and $\theta^{(v)}$ (right panel). MCMC draws from the posterior distribution are presented using tick marks under the curves.

Comparison between Lundquist's method and Bayesian approach for the estimates of $\theta^{(p)}/\theta^{(v)}$. For Lundquist's method, the estimates from pairs of the first 3 fractions (fractions 1 & 2, 1 & 3 and 2 & 3) are presented as examples. The presented mean and median columns were calculated using all pairs of fractions; the subjects having more than 3 fractions have an * next to their identification numbers. The columns mean** and median** are the posterior subject-specific means and medians from Bayesian approach.

Table 1

ID	fractions	Lundquist Method			Bayesian Approach		
		posterior			posterior		
		1 & 2	1 & 3	2 & 3	mean	median	mean**
01*	-0.02	0.02	0.08	0.07	0.05	0.33	0.30
02	0.31	1.93	-6.14	-1.3	0.31	0.77	0.50
03	-0.22	0.1	-0.77	-0.3	-0.22	1.18	0.70
04	-0.13	0.07	0.77	0.24	0.07	0.38	0.34
05	0.39	0.05	-0.06	0.12	0.05	0.32	0.28
06*	1.58	0.9	0.2	0.31	0.2	0.18	0.16
07*	0.5	0.54	0.19	0.26	0.19	0.24	0.22
08*	-0.32	0.2	-0.08	-0.04	-0.03	0.27	0.22
09*	-0.92	-0.5	0.13	0.17	0.16	0.32	0.29
10*	0.1	-0.34	-0.24	-0.08	-0.03	0.13	0.12
11	2.85	11.67	0.2	4.91	2.85	0.39	0.34
12	-1.06	0.54	1.72	0.4	0.54	0.13	0.12
13	0	1.34	-0.48	0.29	0	0.05	0.28
Overall	0.24	1.27	-0.34	0.39	0.32	0.26**	0.24**

Table II
A comparison between a maximum likelihood approach with 20 quadrature points for the logit model and 6 for the beta model) and the Bayesian approach. An * denotes the posterior mean; while a ** denotes the posterior standard deviation.

	Logit f_{ij} Model			Beta f_{ij} Model		
	MLE		Bayesian	MLE		Bayesian
	estimates	s.e.	mean* s.d.**	estimates	s.e.	mean* s.d.**
$\theta^{(p)}$	28.04	10.51	28.83 9.66	$\theta^{(p)}$	10.60	21.41 9.96
$\theta^{(v)}$	85.22	15.16	83.23 16.84	$\theta^{(v)}$	16.19	88.43 17.83
A	84.60	14.00	84.88 13.65	A	14.29	84.42 12.12
B	17.35	2.25	16.91 1.99	B	2.59	17.76 2.36
σ_A	46.52	12.43	44.77 11.69	σ_A	9.88	47.42 11.35
σ_B	7.47	1.63	8.02 1.91	σ_B	1.86	8.74 2.01
σ_p	7.81	16.79	4.89 8.78	σ_p	11.60	3.75 6.81
σ_v	44.22	11.41	51.33 15.54	σ_v	12.90	53.82 15.18
σ_a	7.06	1.19	6.06 0.92	σ_a	2.22	3.82 1.42
σ_b	1.24	0.23	1.01 0.18	σ_b	0.44	1.03 0.34
σ_c	19.72	2.40	20.49 2.70	σ_c	2.34	18.59 2.79
λ_1	-1.43	0.25	-2.69 0.28	α	1.02	2.89 0.72
λ_2	2.72	0.38	4.88 0.33	β	0.81	2.42 0.57
$\sigma_{\lambda 1}$	0.00	0.21	2.85 1.60			
$\sigma_{\lambda 2}$	1.23	0.44	7.63 3.81			
$\delta_{\lambda 1, \lambda 2}$	-0.90	0.22	-4.01 2.29			

Table III

Bayesian P -values based on test quantities (the ranges) for the three outcomes (PSA, fructose and SA concentrations).

	PSA	fructose	SA
Normal with diffuse prior	0.530	0.269	0.287
Normal with truncated diffuse prior	0.544	0.305	0.335
Log normal with diffuse prior	0.525	0.290	0.250