Université de Montréal

Mixed effects modelling for biological systems

Par

Zhe Si Yu

Département de mathématiques et de statistique, Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de Maîtrise ès sciences

en statistique option générale

Mars 2022

Université de Montréal

Unité académique : département de mathématiques et de statistique, Faculté des arts et des sciences

---

Ce mémoire intitulé

Mixed effects modelling for biological systems

Présenté par

Zhe Si Yu

A été évalué(e) par un jury composé des personnes suivantes

Jacques Bélair

Président-rapporteur

Morgan Craig

Directrice de recherche

Philippe Gagnon

Membre du jury

Résumé

En raison des relations complexes entre les variables des systèmes biologiques, l'hétérogénéité des données biologiques pose un défi pour leur modélisation par des modèles mathématiques et statistiques. En réponse, étant conçus pour traiter des données multiniveaux et bruitées, les modèles à effets mixtes deviennent de plus en plus populaires en modélisation quantitative de systèmes biologiques. L'objectif de cette thèse est de présenter l'application de modèles à effets mixtes à différents systèmes biologiques.

Le deuxième chapitre de ce mémoire vise à déterminer la relation entre la cote de qualité du sirop d'érable, divers indicateurs de qualité couramment obtenus par les producteurs ainsi qu'un nouvel indicateur, le COLORI, et la concentration en acides aminés (AA). Pour cela, nous avons créé deux modèles à effets mixtes : le premier est un modèle ordinal qui prédit directement la cote de qualité du sirop d'érable en utilisant la transmittance, COLORI et AA ; le deuxième modèle est un modèle non linéaire qui prédit la concentration en AA en utilisant COLORI avec le pH comme approximation temporelle. Nos résultats montrent que la concentration en AA est un bon prédicteur de la qualité du sirop d'érable et que COLORI est un bon prédicteur de la concentration en AA.

Le troisième chapitre traite de l'utilisation d'un modèle de la pharmacocinétique de population (PopPK) pour décrire la dynamique de l'estradiol dans un modèle de pharmacologie quantitative des systèmes (QSP) de la différenciation des cellules mammaires en cellules myoépithéliales afin de capturer l'hétérogénéité de la population de patients. Nous avons trouvé que la composante PopPK du modèle QSP n'a pas ajoutée de grande variation dans la dynamique de patients virtuels, ce qui suggère que le modèle QSP inclut intrinsèquement l'hétérogénéité.

Dans l'ensemble, ce mémoire démontre l'application de modèles à effets mixtes au systèmes biologiques pour comprendre l'hétérogénéité des données biologiques.

Mots-clés : modèles statistiques, modèles mathématiques appliqués à la biologie, modèles à effets mixtes, hétérogénéité biologique

Abstract

Modelling biological systems with mathematical models has been a challenge due to the tendency for biological data to be heavily heterogeneous with complex relationships between the variables. Mixed effects models are an increasingly popular choice as a statistical model for biological systems since it is designed for multilevel data and noisy data. The aim of this thesis is to showcase the range of usage of mixed effects modelling for different biological systems.

The second chapter aims to determine the relationship between maple syrup quality rating and various quality indicator commonly obtained by producers as well as a new indicator, COLORI, and amino acid (AA) concentration. For this, we created two mixed effects models: the first is an ordinal model that directly predicts maple syrup quality rating using transmittance, COLORI and AA; the second model is a nonlinear model that predicts AA concentration using COLORI with pH as a time proxy. Our models show that AA concentration is a good predictor for maple syrup quality, and COLORI is a good predictor for AA concentration.

The third chapter involves using a population pharmacokinetics (PopPK) model to estimate estradiol dynamics in a quantitative systems pharmacokinetics (QSP) model for mammary cell differentiation into myoepithelial cells in order to capture population heterogeneity among patients. Our results show that the QSP model inherently includes heterogeneity in its structure since the added PopPK estradiol portion of the model does not add large variation in the estimated virtual patients.

Overall, this thesis demonstrates the application of mixed effects models in biology as a way to understand heterogeneity in biological data.

Keywords: statistical models, mathematical models in biology, mixed effects models, biological heterogeneity

Table of contents

List of tables

List of figures

List of acronyms and abbreviations

(G)L(M)M: (Generalized) linear (mixed) models

NLMEM: Nonlinear mixed effects model

(RE)ML: (Restricted) maximum likelihood

EM: Expectation-maximization

MCEM: Monte Carlo expectation-maximization

SAEM: stochastic approximation expectation-maximization

OLS: Ordinary least square

GLS: Generalized least squares

MCMC: Markov-chain Monte Carlo

LRT: Likelihood ratio test

AIC: Akaike information criterion

BIC: Bayesian information criterion

IT: Information theory

PPAQ: Producteurs et productrices acéricoles du Québec

AuNPs: Gold nanoparticles

TIR: Total internal reflection

AA: Amino acid

DTBB: Day to bud break

QSP: Quantitative systems pharmacology

(H)SC: (Hematopoietic) stem cells

MaSC: Mammary stem cells

PopPK: Population pharmacokinetics

EE: Ethinyl estradiol

IIV: Interindividual variability

PRCC: Partial rank correlation coefficient

LHS: Latin hypercube sampling

Acknowledgements

I would like to first give my deepest thanks my research director Dr. Morgan Craig for her endless support, academically or otherwise, throughout my studies. I would not be here, again academically or as a person, without her wisdom and kindness.

I would also like to greatly thank our collaborators Dr. Jean-Francois Masson, Issraa Beniani and Simon Forest for offering me a slice of their expertise. Their patience and feedback were much appreciated throughout my research process. Additionally, I greatly appreciated the generous support from Producteurs et productrices acéricole du Québec (PPAQ), who funded part of my research.

Special thanks to my fellow labmates who offered a welcoming environment that made my Master's degree that much more enjoyable. It was a great pleasure working with them all. I am especially grateful for Justin Le Sauteur-Robitaille for guiding me through our paper, and Mia Brunetti who offered her help and friendship.

Of course, I wouldn't be here without the support from my family and friends. My gratitude for them is eternal. Thank you especially to Yingke Liang who pushed me through all my lows and celebrated my highs.

Chapter 1 – Introduction

The coupled improvement of statistical analysis tools and our understanding of complex biological systems has heightened interest in the representation of the heterogeneity present in biological data. Biological systems naturally present a great amount of noise. Owing to the experimental and clinical limitations, the number of confounding factors in biological data is generally impossible to tease out, even through careful experimental design. This difficulty is particularly pronounced in observational studies. For example, a population of frogs kept in captivity and bred from the germ line would still present heterogeneity, while a wild group of frogs would experience such a large range of external factors that large noise would be inevitable. The question remains as to which statistical tools are optimal to capture these effects in statistical and/or mechanistic models.

The level of statistical complexity often depends on the quality of the available data and the type of system being described. As in all modelling, statistical models face the issue of both under and over fitting, in which one end would fail to capture critical nuance of the system while the other end results in unusable models for future data. This thesis is concerned with mixed effects models and their application to heterogeneous biological systems. From a simple extension from a linear fixed-model (e.g., linear regression) to nonlinear and general models, mixed-effect models offer the ability to capture complexity from several angles and are suitable for describing a wide variety of scenarios. This is especially important when using statistical approaches to describe biological phenomena and explains why mixed effects modelling has become a popular choice for understanding variability within biological data.

1.1 A few statistical basics

Before discussing mixed effects models in detail, it is necessary to provide several key definitions that will be used throughout this thesis. One of the most foundational applications of statistical in biology is hypothesis testing. A parametric hypothesis is an assertion about the unknown parameter θ. If all parameters exist in the parameter space $\Theta \subseteq \mathbb{R}$, then the null hypothesis is referred to as[1]

$$H_0: \theta_0 \in \Theta_0,$$

while the alternative hypothesis is referred to as

$$H_1: \theta \in \Theta_1.$$

By convention, the null hypothesis is a statement representing 'no change', meaning that the alternative hypothesis represents change. Hypothesis testing involves partitioning the sample space into disjoint sets C and C^c based on sample points $x = (x_1, x_2, \dots, x_n)$ from data such that if $x \in C$, we reject H_0 and if $x \in C^c$, we fail to reject H_0. It is important to note that we do not say that we accept H_0 or accept H_1, simply that x fails to provide sufficient evidence to reject H_0 or x does provide enough evidence to support H_1 more than H_0, respectively[1].

Before we can select a benchmark to determine the set C, we must first discuss errors that can be made when using hypothesis testing. If H_0 is incorrectly rejected in favour of H_1, a Type 1 (α) error has occurred. Conversely, if we fail to reject H_0 when it is false, then a type 2 error (commonly referred to as β, but defined in this thesis to be $\tilde{\beta}$ to not confuse it with the β in equations in Section 1.2.1) occurred. Power is then defined as

$$1 - \tilde{\beta},$$

or the probability that type 2 error did not occur. Ideally, C would be chosen such that both errors are zero, but this situation is unseen in practice[1]. Instead, the most used values used for Type 1 error and power are $\alpha = 0.05$ and $1 - \tilde{\beta} = 0.8$[2,3], chosen to minimize both errors to acceptable levels for most practices.

In practice, Type 1 errors have more (nonstatistical) consequences than type 2 errors and are the focus when performing hypothesis testing. A p-value is the probability of observing a sample outcome at least as extreme as the one already sampled under the assumption of H_0[1]. A calculated p-value is then compared to the pre-determined α-level. If the p-value is smaller than α, that is to say that the probability of obtaining a sample at most as extreme as the current sample under the assumption of H_0 is smaller than some accepted level α, then it is safe enough to reject H_0. For better or for worse, p-values are perhaps one of the most well-known statistical terms.

There are divergent philosophies when it comes to hypothesis testing and its application in practice. In contrast to classic frequentist hypothesis testing, the information theoretic (IT) approach examines several hypotheses simultaneously to identify the best model using some information criteria[4]. In the sections below, we will briefly compare and contrast these approaches, particularly with respect to data size requirements (Section 1.3.1.1) and model discrimination (Section 1.3.3).

Though we will be specifying notations throughout this thesis, we list some of the most used ones here for clarification.

- Individual **units** ($i = 1, \dots, n$): The smallest items of measurement used in the equation(s) (e.g., patients in a study of patients in different hospitals).

- If units (i) are **grouped**, the first level of grouping is $j = 1, \dots, J$. The second level of grouping (of units j) is then given by $k = 1, \dots, K$, with subsequent levels of grouping defined in text.

- Layers of groups (including the bottommost layer of units) are called **levels**. Occasionally k will be used to represent responses per unit i. If so, the notation will be written as $[k]$ for clarity. In general, indexing will go from left to right with the largest unit starting on the leftmost position e.g., $y_{ji[k]}$ is the y for the jth group, ith unit at $[k]$th time. k will never be used for both grouping and time in the same equation.

- **Response variable** $\boldsymbol{Y} = (y_1, y_2, \dots, y_n)^T$ is a vector of responses (outcome measurements) for each unit i.

- **Predictor variables** are represented by $\boldsymbol{X}$, an $n \times M$ matrix containing the measured data. $\boldsymbol{X}$ (as well as matrices defined in text) are design matrices, where the rows represent each unit i and the columns are predictor variable $m = 1, \dots, M$. The cells could be data values or indicator variables (ones and zeros).

- **Regression parameters** used for non-mixed effects models will be denoted as β_m. We will write the intercept parameter β_0 separately. For mixed effects models, fixed effects will be denoted as θ and random effects will be u_i. Other parameters may be introduced if there are more levels.

- **Covariates** in this thesis will be defined as characteristics of the units within a study that will affect the outcome of the model prediction. Common example is the body weight of a patient. It must be stressed that, in this thesis, covariate will *not* be used as a synonym for predictor variables.

We will now lay the theoretical foundation of linear and nonlinear mixed effects models (Section 1.2) before discussing their use in practice, particularly in application to biological systems (Section 1.3).

1.2 Statistical foundations of mixed effects models

A mixed effects model is a type of statistical model that contains both fixed effects and random effects. Fixed effects are model parameters that have one fixed population value, meaning that every individual within this population uses the same parameter value. Simpler models, like linear regression, solely use fixed effects as their parameters.

Generalized linear models (GLM) are defined as

$$E(\boldsymbol{Y}) = \boldsymbol{\mu}, \qquad (1.1)$$
$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \beta_0 \mathbf{1}_n + \boldsymbol{X\beta}$$

where $\boldsymbol{Y} = (y_1, y_2, \dots, y_n)^T$ is a vector of the response variable and y_i has a distribution that belongs to the exponential family, $\boldsymbol{X}$ is the $n \times m$ matrix of m predictive variables and n responses, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots \beta_m)^T$ is the vector of the (unknown) slope parameters predicted in the model, β_0 is the (unknown) intercept parameter, and $\mathbf{1}_n = (1,1, \dots,1)^T$. Though commonly taken to be a normal random variable, $\boldsymbol{Y}$ does not have to be normally distributed. The expected value of $\boldsymbol{Y}$

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T,$$

is fitted to some function $\boldsymbol{\eta} = g(\boldsymbol{\mu})$ called the linked function. The variance of $\boldsymbol{Y}$ is the variance-covariance matrix of $\boldsymbol{Y}$, $Var(\boldsymbol{Y}) = V$. The form of $\boldsymbol{\eta}$ in Eq. (1.1) is also commonly written in its matrix form $\boldsymbol{X\beta}$, where the intercept parameter β_0 is integrated in $\boldsymbol{\beta}$ and the one vector $\mathbf{1}_n$ is the first column of $\boldsymbol{X}$.

More conveniently, Eq. (1.1) can be rewritten as

$$E(\boldsymbol{Y}) = h(\beta_0 \mathbf{1}_n + \boldsymbol{X\beta}) \tag{1.2}$$

where h is the inverse link function. While not necessary, for many GLMs the link function is bijective within the correct domain, i.e., $h = g^{-1}$, which is quite convenient. Common GLMs include exponential, logistic and Poisson models.

The most common link function is simply the identity function $g(\boldsymbol{\mu}) = \boldsymbol{\mu}$. When combined with a normally distributed response variable, this gives the more familiar form of the multivariate linear model

$$\boldsymbol{Y} = \beta_0 \mathbf{1}_n + \boldsymbol{X\beta} + \epsilon, \tag{1.3}$$

where $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$ with $\epsilon_i \sim N(0, \sigma^2), i = 1, \dots, n$ the vector of residual (unobserved) random errors[1]. In Eq. (1.3), the $\boldsymbol{\beta}$ parameters represent the fixed-effects parameters, as defined above.

A simple linear model (LM) is the simplest version of a GLM where there is only one predictor and one response, and y is normally distributed. A LM can then be defined as

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{1.4}$$

In the LM in Eq. (1.4) above, there is only one predictor (x) with its one slope parameter β_1 that represents the fixed effect. For example, if one were to predict height (y) of a population children based on age (x), the regression slope β_1 determined by this model is the fixed (population) effect. GLMs like Eq. (1.3) are simple to use and easy to interpret but are not without faults. One of the limitations of a fixed effects-only model presents itself when we want to examine how an individual within a population would behave. Returning to the example of children's heights, we note that the regression slope β_1 does not quantify how an individual within the population

deviates from the population effect of age has on height. To rectify this drawback, random effects can be integrated to help capture and quantify these individual deviations.

Before we provide a more detailed overview of random effects, we will first define some common techniques used for GLMs that allow the models to better accommodate different situations. As seen in Eq. (1.3), the GLM is comprised of an intercept parameter and slope parameter(s). The intercept represents the value of the response variable if the input from the predictor variables is zero and the slopes represent changes in the response with changes to the predictor variable. In our example, the intercept β_0 would predict the average height of a newborn (age=0) whereas the slope β_1 indicates how much the child grows per year (age increases by 1). In such a simple case, there is only one intercept and one slope parameter. This means that the model in the example assumes that the population of children is homogenous, as children on average are born the same height and grow at the same rate. There are many well-explored techniques for GLM that allows more leeway. Changes to the model will include altering the format of either the intercept and/or the slope.

1.2.1 Linear mixed effects models

Variability in data is often attributed to categorical variables that separate the population into subsets. In a fixed effects model like the GLM, dummy variables would be then used to quantify categorical variables in the model in a process called dummy coding[5]. Consider the GLM in Eq. (1.3) with one continuous predictor x and one categorical predictor D, with three categories (or groups) predicting response variable y. This model would be written as

$$y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \epsilon,$$

where $d_1, d_2 = \{0,1\}$ are two of the three groups of D. Here, d_1 and d_2 are called the dummy variables and since they are binary, they are used to signify the effects of their respective group. Further, only one would be in its '1' state at a time. The third group is represented by the "base state" with $d_1, d_2 = 0$. For any categorical predictor variable D with n groups,

$$y = \beta_0 + \beta_m d_m + \epsilon,$$

where $m = 1, \ldots, M$. The choice of group used as the "base case" does not affect the model[5]. When the model is fitted, the base case is absorbed into the intercept estimate. Notice that dummy coding effectively allows the model to change its intercept value β_0 since when d_m is set as 0, its term disappears whereas when d_m is 1, the term becomes β_i, which sums with the intercept β_0. Returning to the example of age versus height in children, we can begin to improve the model by dummy coding the sexes. This would allow the model to determine whether the height of newborns (age = 0) differs between the sexes.

Dummy coding represents one extreme in the class of techniques using a categorical predictor because it estimates separate models within each level of the predictor and therefore corresponds to no pooling[6]. The other extreme is complete pooling, where the categorical variable is excluded completely and the sample is considered to be homogenous[6]. The consequence of not pooling the information for each level of a categorical predictor is poor estimation for any levels that have little information or a limited number of data points[6].

Unlike dummy coding, interaction terms are used to change the slope parameters. Consider a GLM (Eq. (1.3)) with two predictor variables (x_1, x_2) and predictor variable Y. In this case, a model that assumes x_2 affects x_1 is written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon. \tag{1.5}$$

It is convention to include the main effects, or non-interaction terms, when using interaction terms. Notice that though we stated above that it is x_2 that is affecting x_1, there is nothing in the equation itself that dictates the direction of the interaction. Instead, the third term above ($\beta_3 x_1 x_2$) simply assumes that there is some sort of two-way interaction between x_1 and x_2. The precise nature of the interaction is up for interpretation. While there is no limit to the numbers of interactions, interaction terms with more than two predictor variables become harder to justify and should be used with caution[5].

Naturally, interaction terms can be combined with dummy variables so the model can vary in both intercept and slope parameters. Consider again the categorial predictor variable D with three groups and one continuous predictor x. Eq. (1.5) then becomes

$$y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \beta_4 d_1 x + \beta_5 d_2 x + \epsilon. \quad (1.6)$$

Dummy variables and interaction terms are the main tools used in GLM fitting to accommodate for any heterogeneity in the dependent variable. Notice that while dummy variables can alter the intercept parameter β_0 there is no similar alternative to changing slope variables. Interaction terms do not perform this task. Variation among slope parameters between groups of the response variable can be achieved through a general linear model like Equation (1.3) by fitting multiple linear models, one for each group, with the same predictive variables $\boldsymbol{X}$**.** Each group is thus fitted with a linear model that shares the same model structure with the other groups but have different parameter values, including the intercept parameter. One cannot dictate which parameters are changed. Random effects are model parameters that are random variables as opposed to the fixed effects parameters (e.g., β_1). We will now briefly discuss how random effects are similar or different from the techniques used above.

1.2.2 Reasoning behind random effects

Random intercept effects can be used as an analogous method to dummy coding. The results obtained will be similar, as both seek to change the intercept parameter value. As previously mentioned in Section (1.2.1), dummy coding treats each level as a separate entity and consequently does not share information between the levels, while random intercepts generate a distribution in which it draws the effects of the levels[6]. Since the random effects all come from the same distribution, all their information contributes to the estimation of the variation and can correct estimates for levels with small sample sizes.

Random effects models offer a sort of comprise between no pooling and complete pooling coined, creatively, partial pooling[6]. We can describe a model with one categorical predictor C with m levels, one continuous predictor x and n individuals by

$$y_i \sim N\left(\beta_{0ji} + \beta_{1ji} x_i, \sigma_y^2\right), \quad (1.7)$$

$$\beta_{0j} \sim N\left(\mu_{\beta_0}, \sigma_{\beta_0}^2\right). \quad (1.8)$$

Here, the intercept parameter β_{0j} for each group j is fitted from the same (usually) normal distribution with mean μ_{β_0} and standard deviation σ_{β_0}, both of which are estimated directly from

the data. As $\sigma_{\beta_0} \to \infty$, the distribution becomes meaningless and β_{0j} once again are individually estimated as in no pooling; as $\sigma_{\beta_0} \to 0$, estimates of the standard deviation are all pulled to zero, yielding the complete pooling estimate[6]. Note that Eq. (1.7) can be generalized to more than one categorical and continuous predictor. In this case, the categorical predictors would have their own distributions, similar to Eq. (1.8).

Armed with this construction, we can now write a linear mixed effects model with random intercepts. This model is usually written as

$$y_i = \beta_{0ji} + \beta_{1ji} x_i + \epsilon_i. \tag{1.9}$$

Random slope effects are not analogous to interaction terms in the GLM discussed above. Rather, interaction terms solely describe interactions between two fixed effects and are thus irrelevant to random effects. However, we can derive some contrast between the two concepts as a bridge to better understand random slope effects. Random slope effects allow for slope variation within groups by generating a different slope parameter for each group (Figure 1.1). A simple linear mixed effects model with random slope is given by

$$y_i = \beta_{0i} + \beta_{1ji} x_i + \epsilon_i, \tag{1.10}$$

where $j = 1, \dots, J, i = 1, \dots, n$, similar to Eq. (1.7). Interaction terms do not actually change the slope parameter. Rather, in conjunction with dummy variables for categorical predictors, interaction terms create different terms for the groups in the categorial predictor, essentially resulting in similar results as random slope effects. The 5th and 6th terms in Eq. (1.6) given by $\beta_4 d_1 x$ and $\beta_5 d_2 x$ are examples of dummy variables interacting with a continuous variable. Partial pooling, as described above, can also apply to random slope effects.

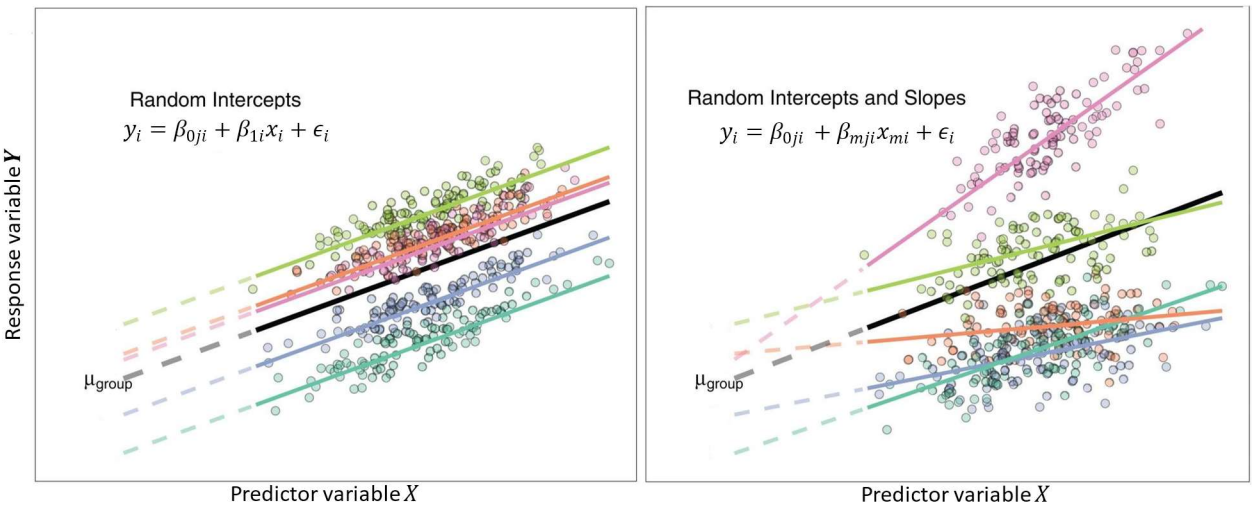

Figure 1.1. – The effects of random intercepts and slopes. eft: a linear mixed effects model fitted with random intercepts for the different groups (coloured lines; 5 total). The black line represents the global mean value of the distribution of random effects, the solid lines denote the regression line, and the dashed lines trace the regression lines back to the y intercept. Point colours correspond to the data points of each group. All the lines are parallel because they share the same slope parameter. Right: A linear mixed effects model fitted with both random intercept and slope parameters. Random slope models offer more flexibility to fit the data. Adapted with permission from Harrison et al.[7].The full linear mixed effects model with both random intercept and slope and generalized with a $m = 1, \dots, M$ predictor vector is

$$y_i = \beta_{0ji} + \beta_{mji}x_{mi} + \epsilon_i. \quad (1.11)$$

Mixed effects models use different conventions for parameters. We define the linear mixed effect model of $j = 1, \dots, J$ clusters with $i = 1, \dots n$ units as

$$y_{ij} = X_i\theta + Z_{ij}u_{ij} + \epsilon_{ij}, \qquad u_j \sim N(0, \omega^2), \qquad (1.12)$$

where X_i, Z_{ij} are the matrix elements at positions i or i, j of the design matrices of the fixed and random effects, respectively, θ is the vector of fixed effects , u_j is the vector of random effects

that are normally distributed with ω as its variance-covariance matrix, $\epsilon_i \sim N(0, \sigma^2 I_{n_i})$, and $N = \sum n_j$ is the total number of observations.

1.2.3 Nonlinear mixed effects models

Since nonlinearity is central to many biological interactions and biological noise, we need to establish a statistical model that can account for nonlinear terms. Luckily, linear mixed effects models (Eq. (1.11)) can be generalized to nonlinear mixed effects models. To do so, let $y_{i[k]}$ denote the $[k]$th measurement of the response variable under $t_{i[k]}$ $(k = 1, \dots, K_i)$ and any additional conditions u_i for the individual i. By convention, $t_{i[k]}$ usually represents time points and u_i is an empty vector or the initial value of a predictor variable (e.g., the drug concentration at time zero for individual i). We can then define predictor $x_{i[k]} = \left(t_{i[k]}, u_i\right)$, though dependence on time t_i is not always needed. We define a nonlinear mixed effect for the $[k]$th observation on the ith individual as

$$y_{i[k]} = f\left(\phi_i, x_{i[k]}\right) + \epsilon_{i[k]}, \tag{1.13}$$

where $y_{i[k]}$ is the $[k]$th response on the ith individual, f is a nonlinear function of the predictor vector $x_{i[k]}$ and a parameter vector ϕ_i of length r, and $\epsilon_{i[k]}$ is a normally distributed noise term[8,9]. The parameter vector can differ between individuals and is defined as

$$\phi_i = d(\theta, \eta_i, a_i), \tag{1.14}$$

where d is a p-dimensional function, θ is a vector of fixed effects length r, η_i is a vector of random effects for individual i of length n, and a_i is a vector of characteristics associated with the individual that does not change with time (i.e., covariates[8]). The distribution of η_i is usually independent of but conditional on a_i, implying that

$$E(\eta_i|a_i) = E(\eta_i) = 0 \text{ and } var(\eta_i|a_i) = var(\eta_i) = D,$$

where D is the covariance matrix that is the same for all i that characterises the magnitude of 'unexplained' variation. A standard assumption is $\eta_i \sim N(0, D)$, but a log-normal distribution is also common[8]. Put otherwise, η_i does not depend on a_i.

A common specific case of Eq. (1.14) is that of a linear relation between ϕ_i and its fixed and random effects i.e.,

$$\phi_i = \boldsymbol{A}_i\theta + \boldsymbol{B}_i\eta_i, \qquad (1.15)$$

where $\boldsymbol{A}_i$ is a design matrix depending on elements of a_i and $\boldsymbol{B}_i$ is a design matrix typically involving only zeros and ones, which allows some elements of β_i to have no associated random effects[8,9]. Equation (1.15) is a very typical form used for repeated measures data where the same subjects in the experiments are sampled multiple times over time.

To completely describe a nonlinear mixed effects model, we need to address inter-occasional variation which is variation between time points of a single individual. Inter-occasional variation is underdiscussed in literature, especially for linear mixed effects models, but it is relevant and widely used, particularly in the pharmaceutical sciences[10,11] (see Chapter 3). For linear mixed effects models, inter-occasional variation is included by treating time as a categorical predictor[12,13]. In the model in Eq. (1.13), we have instead that

$$E\left(y_{i[k]}\middle|u_i, \phi_i\right) = f\left(t_{i[k]}, u_i, \phi_i\right),$$

meaning that f represents how individual i changes, on average. Thus f may not capture the inter-individual variation[8].

The full definition of a nonlinear mixed effects model is

$$y_{i[k]} = f\left(t_{i[k]}, u_i, \phi_i\right) + \epsilon_{R,i[k]} + \epsilon_{M,i[k]}, \qquad (1.16)$$

where $\epsilon_{R,i[k]}$ measures the deviation of the $i[k]^{\text{th}}$ observation (out of all possible observations of $f\left(t_{i[k]}, u_i, \phi_i\right)$) from $f\left(t_{i[k]}, u_i, \phi_i\right)$, and $\epsilon_{M,i[k]}$ is potential measurement error at each time point $t_{i[k]}$. Thus, in an ideal situation there would not be $\epsilon_{M,i[k]}$[8], and it is frequently discounted when estimating nonlinear mixed effects models in practice[11,14].

Mixed effects models are often called hierarchical models, or multilevel models, since the parameters of the individual models also have their own parameters. A two-level hierarchical model has been shown previously if we consider the parameters of Eq. (1.13) (or Eq. (1.16)) to be estimated by Eq. (1.14). Conventionally, the individual level equation is called level 1 with

subsequent levels called level 2, 3 etc. until the population level[8]. For example, considering patients within Montréal, the individual patients would be modeled by the level 1 equation, the patients' hospitals by the level 2 equation, and the city by the level 3, or population level, equation. In this example, we have that each fixed effect at the lower level is estimated by Eq. (1.14), giving the three-level model

$$y_{kji} = f\left(\phi_{ji}, x_{kji}\right) + \epsilon_{kji},$$

$$\phi_{kj} = f\left(\theta_k, \eta_{kj}, a_{kj}\right),$$

$$\theta_k = g(w, u_k, b_k),$$

For each observation, y_{ijk} denotes the i^{th} measurement of the response for j^{th} individual patient and k^{th} hospital. Here, ϕ_{kj} is the parameter vector for individual patients (level 1), and θ_k is the parameter vector for hospitals (level 2). There is no limit on the number of levels in hierarchical models, and they can be extended beyond 3 levels.

Though the form of a linear mixed effects model is commonly expressed by Eq. (1.10), its form can equivalently be derived from Eqs. (1.15) and (1.16) for f as a linear equation. Further, for simplicity, linear models are usually expressed with the individual and population levels already combined. However, linear models are also hierarchical and have multiple levels, like the example of the Montréal patients discussed above.

1.3 Practical aspects of mixed effects models in biology

In the following sections, we will turn our focus to practical aspects of mixed effects models and their application to biological modelling. For most biological systems, variation within the population is too high to be realistically quantified using only fixed effects. Biological systems can also be very complex, and this complexity may be unknown or beyond the scope of the experiment. Mathematical models of biological phenomena are a well-established way to understand complexity and to quantify heterogeneity within the system[11,14]. When constructing a mathematical model to describe any biological system, it is impossible to include all the nuances and complexities that are present. However, just because some elements in the system are ignored in the model does not mean their effects disappear in the data. The residual term in any

statistical model contains variability in the data that is not explained by the model. In a perfect statistical model, the residual term only contains the natural noise that exist in all data. Realistically, the residual term becomes inflated with variability in the data that should have been described by the model[6]. As we will see in Chapter 3, sufficiently mechanistic mathematical models can recover much of the noise present in biological systems but come with drawbacks that include extensive parameterization and limits with respect to analysis.

Mixed effects modelling comes with many intrinsic benefits that either correct or augment the regression models discussed earlier. One common reason to fit a mixed effects model is to control for non-independence among data points. Data collection from complex biological systems often requires the use of repeated measurements from individuals within and across time[7,15]. Repeated measurements are not independent and violate basic regression assumptions. Simple linear regression has five core assumptions – linearity, normality, independence, homoskedasticity and no collinearity[1] – and most can be corrected by transforming variables or by fitting a more generalized regression. Non-independence, however, is considered the most serious issue and cannot be well-addressed by data analysis techniques belonging to GLMs[16]. Failure to take non-independence into account leads to an inflated Type 1 error rate. Fortunately, random effects control for non-independence by constraining non-independent units to have the same parameters[7].

As stated in Section 1.2.1, mixed effects models are partial pooling models, meaning that the parameters for each group within a level (e.g., patients in hospitals) are fitted from the same distribution. Partial pooling models share information between parameters for their estimation, which is very advantageous for sparse data and large number of grouping. While conventional regression models quickly lose inference power when the data is sparse[17], mixed effects remains robust[18].

Selecting a type of model for fitting requires choosing a model structure suitable for representing the data and its system. The specific research question and design will determine whether one focuses on the fixed effects and their interactions with no specific expectations on variances or correlations. In this case, the random effects in the model fitted for these hypotheses then

minimizes residual noise and thus maximize model power[19]. Population focus[20] studies use random intercepts and slopes to reduce Type 1 and 2 errors and the chance of overconfident estimates (i.e., unrealistically low standard error (SE))[4]. An example of a population focused study comes from Ribba et al.[21], who modeled the effects of anti-tumor drugs of tumor size using a mixed effects model and studied the variation among the units of interest (patients) captured via random effects. Their focus was the overall effect of the drug on tumor size within a population of patients with low-grade glioma (a type of brain tumor). Their model was able to describe the longitudinal mean tumor diameter of the entire duration of treatment. Their results with the same model was consistent for three different types of treatment[21].

We may also be interested in the effect within each subgroup (cluster) in the hierarchy in addition to the overall effect in the whole population. Here, the research hypothesis would concern not only the fixed effects, but the deviation of each cluster brought by their respective random effects. We will call this motivation cluster focus[20]. In a meta-analysis following the monitoring and surgical treatment of drug-resistant epilepsy by Remick et al.[22], a mixed effects model was applied to 33 studies from 25 neurosurgical centers in ten countries. Multiple levels of interest, including population of patients, nations, treatment centres, and studies, were thus included in their hierarchical model. In this case, each cluster is of interest for its effect on the treatment of drug-resistant epilepsy, classifying this study as cluster focused[22]. Put simply, population focus experiments are interested in β while cluster focused are also interested in $u_i, i = 1, \dots, n$.

The actual variation among units is also of particular interest. Rather than reduce unmodeled noise, control for non-independence, or investigate effects among each cluster, the main motivation in this case is to investigate differences between groups of interest[7]. By fitting a random intercept for the group factor, we can estimate the among-group variance for any trait of interest. Combined with the residual variance term, we can then calculate an "intra-class correlation coefficient" that measures individual repeatability in our trait. By using random effects rather than assigning a fixed effect for every group, we prevent a loss of model power from losing a degree of freedom for each group. Furthermore, the results can be better compared across studies for the same trait or across traits in the same study. Variance component analysis is a widely used tool with the potential to solve complex problems, including those frequently

seen in biomedical applications. A good example of studies of actual variation among units is seen in the paper by Kang et al., where variance component analysis shows promise as a statistical tool for large dataset analysis that replaces classic approaches such as principal component analysis[23]. The novel application of these methods was found to considerably shorten computational time. Further, it was shown that variance component analysis captured a wider range of sample structures, including both hidden relatedness and population stratification within human genome-wide association studies datasets[23].

1.3.1 Model estimation

A generalized scheme of the pipeline for applying mixed effects models is provided in Figure *1.2*. In what follows, we will describe these steps in further detail.

Figure 1.2. – Simplified steps in successfully fitting a mixed effects model. In practice, model fitting is seldomly as simple, we will follow this over-generalized schema to cover the large steps and obstacles in fitting a mixed effects model.

1.3.1.1 Data requirements

Estimating random effects can be quite a 'data hungry' process. Sample sizes are important for any study as they determine how much power its statistical tests hold (see Section 1.1). Conversely, to obtain a certain level of power, the study must achieve a corresponding sample size.

Linear mixed effect models require at least five groups for a random intercept term to generate robust estimates of variance[7]. If this condition is not fulfilled, the model may not be able to accurately predict the among-population variance or simply collapse into zero, leaving the model equivalent to an ordinary GLM. Known as a degenerate model, it is the result of the software

used for model fitting converge to a degenerate or singular covariance matrices for the variance parameters in which some linear combinations of the random effects are estimated to have no variability. In this case, if the fitted model has both random intercepts and slopes, they will have a correlation of ± 1[19]. More generally, since the number of sample points and groups needed for a good fit is dependent on the complexity of the model, overparameterization is the principal issue in this case. Overparameterization has serious detrimental consequences for inference[19]. (G)LMM solvers are usually permissive and will eventually converge to an estimate, even if the estimates correspond to degenerate or singular covariance matrices. This is often the result of convergence on the boundary in overparametrized models[19].

Furthermore, models can be unstable if sample sizes are highly variable among the groups (i.e., some groups have very sparse data). Groups with very little data will have poor predictions for their random slopes[4,7]. Generally, if level 2 has 50 clusters, the model becomes susceptible to small sample biases of both the variance components and the fixed effect standard effects[7,24]. The current rule of thumb for linear mixed effects models calls for a minimum of 30 clusters at each level of analysis[18]. As expected, statistical power increases with both an increasing number of clusters and increasing number of individuals per cluster. When the random effects variance is very low, the number of individuals per cluster has a substantial effect on power. However, when the random effects variance is not low, increasing the number of individuals has trivial returns[18,25]. For a sufficient number of clusters (>50), the proportions of singletons (clusters with a single data point) does not increase bias and only negatively affect Type 1 error rates for >50% singletons[18].

Previous studies have shown that including random slopes controls for Type 1 error rate and gives more power to detect among individual variation[7,24]. It may thus be tempting to fit the maximal number of random slopes to capitalize on this effect. However, as discussed earlier, mixed effects models are prone to suffer from small sample bias and maximizing the number of random slopes exacerbates the issue as it requires large numbers to accurately estimate variance and covariances accurately.

Sample size requirements for nonlinear mixed effects models are much harder to determine. There is little consensus on a mandatory threshold like for linear models, especially across the popular applications for nonlinear mixed effects model. General guidelines based on computational methods have been developed to suggest sample sizes based on a multitude of conditions[3,26]. For example, for nonlinear mixed effects models used in pharmacokinetics with typical factors including drug half-life, absorption rate, hypothesized differences etc., it has been suggested that as few as $n = 2$ datapoints are required per cluster[3].

Lastly, a particularly important consideration when designing experiments is the data required to observe a particular effect (see Section 1.1 for the basics of hypothesis testing). Regardless of whether a frequentist or information theoretic approach is applied, it is best practice to design a global model in which all predictors, interactions, and covariates of interest are included to investigate all possible combinations of considered factors and help select the model of best fit. While a full global model may be ideal, the data size limitations described above do not permit so many predictor terms. Instead, it is advisable to conduct *a priori* investigations of the biological mechanisms to narrow down the number of variables considered to be predictors. This holds true for most biological systems, but genetics is one notable exception since there are usually sufficient data available in genomics studies[23]. One may argue that only predictors with strong biological reasoning should be included in the model. However, in a complex or understudied biological system, previous knowledge may not provide sufficient information[14]. A general rule of thumb to start is 10:1 subject to predictor ratio. Unless there is biological incentive, it is not recommended to include second- and higher-order terms in the model[4].

1.3.1.2 Data exploration

Once the required data size has been achieved, one may or may not actually choose to do exploratory data analysis prior to fitting the model[27], depending on the model and hypothesis design of a study. For a frequentist approach with a traditional hypothesis, it is still vital to do data exploration. Zuur et al. outlines a protocol for data exploration, including testing for outliers, homogeneity, normality, zeroes, collinearity, relationships, interactions and independence[28] (Figure 1.3). They encourage expansive exploration and visualizing the results. Not only is it a step critical for data cleaning and better understanding of the data but theoretically, it can also be

used as prior knowledge to determine optimal predictors. However, advocates of the information theory (IT) approach are in principle against data exploration[4]. They suggest that predictors should only be determined from genuine previous knowledge, like pilot studies or literature[29]. As such, the IT approach does not prevent the use of data exploration, but rather calls for not letting this information influence predictor selection, thus defeating the purpose of testing for multiple hypothesis in the first place.

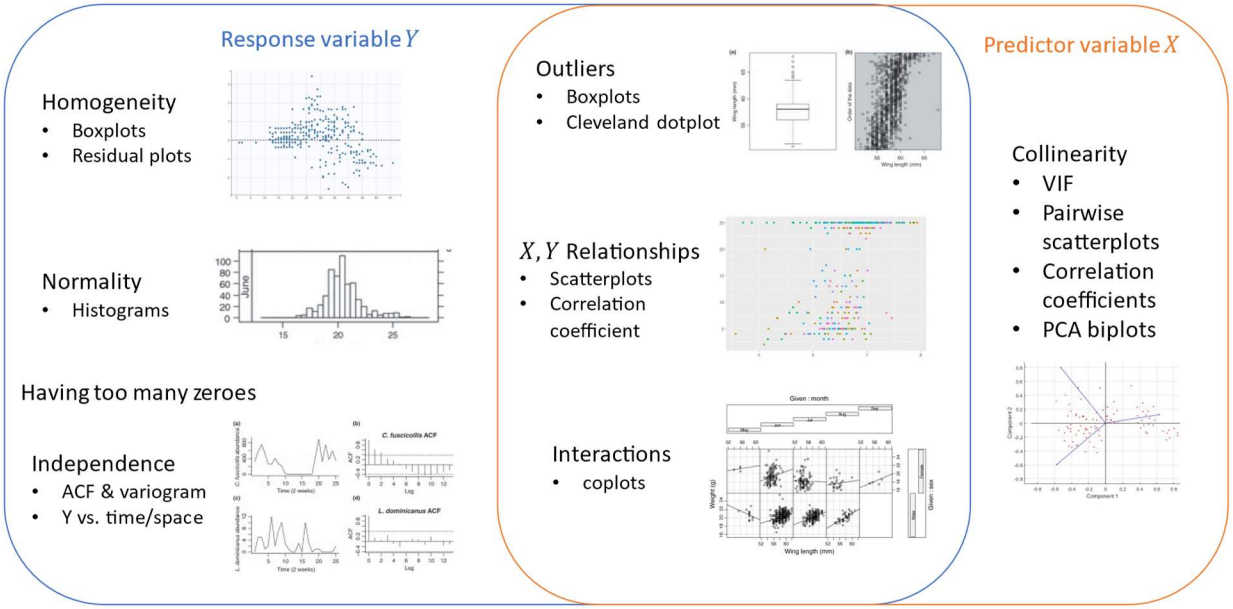

Figure 1.3. – Venn diagram of data exploration techniques for the response variable Y and predictor variable X suggested by Zuur et al.[28]. The centre of the Venn diagram provides criteria that is both relevant for X and Y, as well as between X and Y. Not all techniques recommend may be relevant, depending on the model being fitted. For example, GLM does not require normality unless a normal distribution is chosen; otherwise there is no need to check for normality. Adapted with permission from Zuur et al. [28].

In practice, it is always a good idea to explore the data. Knowing if the data have, for example, collinear predictors will influence the approach to model design and model selection. Options to transform data into linear variables may help to simplify every step along the process.

1.3.1.3 Hierarchical structure

Mixed effects models are convenient for hierarchical experimental design. Afterall, the motivation behind mixed effects modelling is repeated measures data. By the nature of mixed effects modelling, units are grouped together by some variable, which may then be grouped again by another variable. Thus, though not specific to mixed effects modelling, interactions between predictive variables are as relevant to mixed effects modelling as random parameters. A good understanding of interaction terms and how they relate to experimental design is critical to building a descriptive model for one's data. There are two main types of experimental design: crossed or nested (Figure 1.4). In a crossed design, factors of the lower level will be associated with more than one factor of the higher level. In a fully crossed design, there are observations for every combination of factors of the predictors, while in a partially crossed design there are samples for only some of the combinations. In contrast, nested design applies a hierarchy in which factors of a lower level is associated to only one factor of the higher model. Study designs can also be a combination of the two types[15].

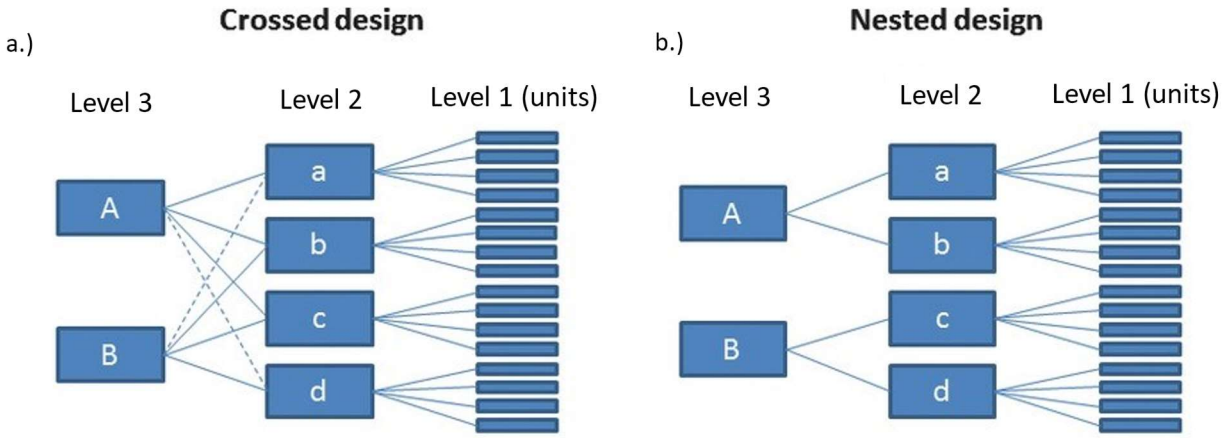

Figure 1.4. – Schematic of crossed and nested study designs. a.) For a crossed design, the groups in level 2 are associated with both groups of level 3, whether fully (solid and dashed lines) or partially (only solid lines). b.) In a nested design, the groups of level 2 are only associated with a single group in level 3 (i.e., level 2 group a and b are only associated with level 3 group A and not B). Reproduced and adapted with permission from Schielzeth and Nakagawa[15].

Though they are called design, the same concepts can also be applied to observed data. Factors can be naturally nested through the nature of the subjects; spatial factors are often naturally nested. For example, consider the study of the effects of a particular disease within major cities of states with different healthcare systems. Since each city can only exist within one state, the level of cities is nested within the level of state naturally. In contrast, drug studies are typically nested by experimental design, with the level of subjects nested within each drug type. It is also possible to design such experiments as a crossed design where each subject would receive different drug type however this may lead to too many cofounding factors, such as age or damage from previous trials.

How the model describes the design (i.e., nested, crossed, or a combination thereof) determines how variance, or noise, is accounted for in the model. Recall Section 1.2.1 where we described the interactions between any two predictive variables. Interactions between two levels are the same as interactions between predictive variables, as levels themselves are categorical predictive variables. For example, level 2 in Figure 1.4 is a categorical variable with four groups and level 3 is a categorical variable with two groups.

Between any two levels, there are four sources of variances that can potentially be estimated: the first two are the main effect variance explained by either level; the third is the interaction variance that is explained by level combinations; and the last is whether there is residual variance or noise unexplained by the levels nor their interactions. A crossed design allows us to estimate all four as long as we include an interaction term in the model[15]. That is to say that all the variation within the data would be contained in one of the four variances, whether they are explained by the main effects and interaction effect or unexplained as the residual variance. The interaction term, however, may not always be relevant. One may decide to neglect this term if it serves little biological meaning, which would lead to the interaction variance being pooled with the residual variance as unexplained noise. A nested design does not offer such luxury. By definition, the groups in the lower level are only associated with one group in the higher level in a nested design. Using Figure 1.4 as an example, the groups a and b in level 2 do not interact with group B in level 3 (and group c and d to group A), meaning that there is no information on the interactions between these groups. As a result, the variation explained by the interaction term is confounded

with the variance of level 2's main effect. Thus, it would not be possible to say whether variation of level 2 is caused by differences among level 2 groups solely or in effect of level 3 groups. In summary, in a nested design the main effect of the nested level will be pooled with the interaction effect[15]. The main effect of the nesting level (e.g., level 3) and the residual variance are not affected. It is important to reiterate here that interactions between two variables (or levels) are not estimating random slope parameters. Interactions can occur in regular GLMs but are especially relevant in mixed effects modelling due to its hierarchical nature.

1.3.1.4 Model generation

Once the desired predictors as selected, one can generate a 'model set' of hypotheses. This task is theoretically simple with a global model, as the model set can be all possible submodels of the global models where parameters are set to zero for different combinations. Then an information criterion, like the popular Akaike information criterion (AIC)[30], can be used to rank the model set and select the model of best fit.

There are nonetheless several potential obstacles to the use of information criteria in this way. There are two possible types of error common to all modelling processes: failure to estimate the parameters and overestimation of standard error. If it is the global model that does not converge, one should seek to resolve the complexity of the model. Interactions can be removed first, then main effects whose results in *a priori* investigations proved to be weak or biologically less relevant. This process would be repeated until the model converges. If all of the predictors are of interest and the global model cannot converge, one can break it into submodels, as complex as possible, and use these submodels as 'global models'[4]. However, this method is less desirable as it may lead to too many models for selection.

1.3.1.5 Random effects estimation

There are multiple ways to estimate the parameters of a mixed effects model, each providing varying degrees of accuracy for different situations. Mixed effects model are estimated either by maximum likelihood (ML) or restricted maximum likelihood (REML), the latter of which is used to mitigate some of the small sample bias[31]. ML estimates random and fixed effects simultaneously, which can be an issue as the random effects are latent variables in the model. Iterative

approaches such as the Expectation-Maximization (EM) algorithm are often used to integrate out the random effects in order to estimate the fixed parameters of the model[32]. Iterative processes first estimate the fixed effects, then use these estimates as known information to estimate the variance components. The process is repeated until convergence. Though the iterative nature dictates an order, estimates are still considered to be estimated simultaneously, since the action is performed in one iterative step. However, this order does pose two issues. Firstly, any variability in the fixed effect is ignored. More importantly, the degrees of freedom consumed by setting the fixed effect is unaccounted for, which heavily penalizes small sample sizes[31] and leaves the variance component underestimated.

REML tackles this issue by estimating the fixed and variance components separately. Consider a linear mixed effects model with a random intercept and slope with one predictor variable given by

$$Y_{ji} = \beta_{0i} + \beta_{1i} x_{ji} + \epsilon_{ji}.$$

The first step of REML is to obtain the ordinary least square (OLS) residuals of the model without any clustering, i.e.,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

By definition, the residual (ϵ_i) is independent of any predictor variables (i.e., there is no correlation between ϵ and x). Then, ML is calculated with the OLS residuals as the outcome instead of the original outcome variable to estimate the variance component. Essentially, the first step acts as a linear transformation of the original data, making the outcome independent of the predictors (x) but now share the same variance. When using the OLS residuals as the outcome, the fixed effects no longer need to be considered as they are zero. The effect of changing the ML output from the response variable (Y) to the OLS residuals is shown in Figure 1.5, where the top plot (x vs. Y) is transformed into the bottom plot (x vs. ϵ), removing the effects of the fixed effects but keeping the variability of the data, conditional on x. In fact, this conditional variance is identical to the variance calculated by using the original response variable as the outcome. By estimating the variance component separated from the fixed effects, REML avoids the issues related to simultaneous estimation, and result in variance components that are

appropriately higher than those estimated by ML with small samples[31]. Removing the fixed effects and estimating the variance components on its own marks the first stage of REML.

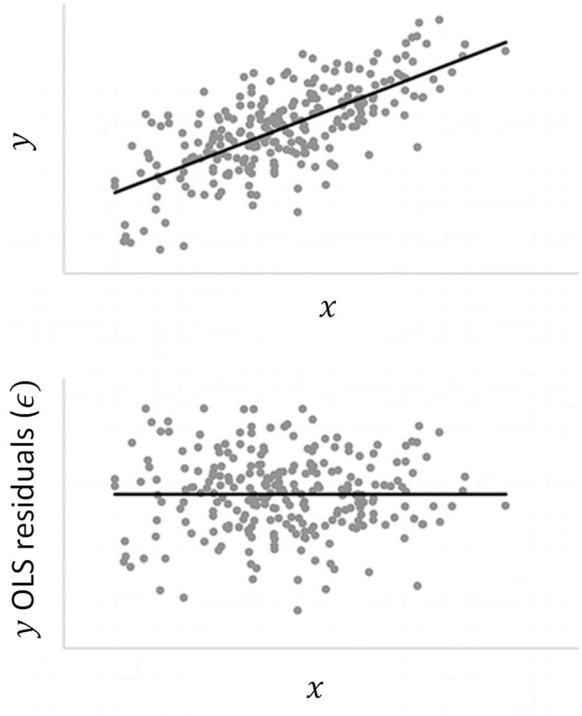

Figure 1.5. – The relationships between outcomes, predictors, and residuals. Comparison of the relation between the outcome and the predictor (top panel) and between the OLS residuals and the predictor (bottom panel). Reproduced and adapted from McNeish[31] with permission.

The next step estimates the fixed effects using generalized least squares (GLS), which is capable of accounting for clustered data. GLS is able to estimate identical fixed effect estimates to ML results provided that the covariance structure of the observations is known, which was conveniently estimated in the first stage of REML[31].

In general, REML estimates nearly identical fixed effects as ML while achieving better estimates of the variance components[31]. As REML does not completely resolve the issue of inflated Type 1 error rates for effects, the Kenward-Roger correction[33] is brought in to maintain the error rates. The full derivation of the Kenward-Roger correction is beyond the scope of this thesis, but we include the general concept of the solution. The Kenward-Roger correction is based on the earlier

Kackar-Harville correction[34], which conceptualizes the issue of inflated Type 1 errors for effects as

$$Var(\gamma) = Var^{REML}(\hat{\gamma}) + small\ sample\ bias,$$

or that the true sampling variability of the fixed effects is equal to the REML-estimated variability plus some amount of small bias incurred from violating the asymptotic assumptions. Kackar and Harville were able to estimate this small sample bias by approximating its very complex mathematical form with a Taylor series expansion (see Kackar and Harville[34] for the full derivation). Multiple iterations of improvements have been made for this technique throughout the years, eventually landing with the title Prasad-Rao-Jeske-Kackar-Harville correction[31]. In addition to using the Prasad-Rao-Jeske-Kackar-Harville correction, Kenward and Roger also corrected the term $Var^{REML}(\hat{\gamma})$. They noted that REML uses GLS to estimate the fixed effects after the variance components are estimated. GLS is used under the assumptions that the variance components are known, even though they are not. Thus, the sampling variability of the variance components are not accounted for. The Kenward-Roger corrections uses another Taylor series expansion, this time to $Var^{REML}(\hat{\gamma})$ to account for the fact that the variance components are estimated and not known. The two expansions provide a more accurate fixed effect standard error estimate[31].

1.3.1.6 Estimating nonlinear mixed effects models using Monolix

There are a few popular software choices for solving nonlinear mixed effects models (NLMEMs). Monolix is a relatively new platform designed by Lixoft that uses the stochastic approximation expectation-maximization (SAEM) algorithm which, as its name suggests, is the stochastic approximation of the EM method mentioned in Section 1.3.1.5. Monolix uses the SAEM algorithm because it has been shown to be efficient and robust to a multitude of model and data types.

EM algorithms have two steps: the expectation (E)-step that involves taking expectation over complete-data conditional distribution and the maximization (M)-step that involves complete data maximum-likelihood estimations[32,35]. The E and M steps are iterated until convergence is formed. We provide a short definition of the EM algorithm below.

Consider the nonlinear mixed effects model defined by Equations (1.13) and (1.14). Let $\eta = (\eta_1, \dots, \eta_n)$ and $z = (y, \phi)$ represent the complete data such that ϕ represents the missing data. Note that η are the unobserved random effects and y are the response variables. Let $\beta = (\theta, D, \sigma^2)$, where θ are the fixed effects, D is the variance of η and σ^2 is the variance of y, for which the ML estimate is required[35]. That is to say

$$L(\hat{\beta}|y) \geq L(\beta|y) \qquad \forall \beta \in \Omega,$$

where $L(.)$ is the likelihood function and Ω is the parameter space[35]. We define the map $Q: \Omega \times \Omega \to (-\infty, \infty)$ by

$$Q(\beta, \beta') = E\{\log p(z|\beta')|y, \beta\},$$

where p is the probability model for the observations given by the joint distribution. $Q(\beta, \beta')$ can be thought of as a local approximation to the log-likelihood of β' in a neighbourhood of β[35].

The E-step for each iteration k consists of evaluating

$$Q(\beta, \beta_{[k]}) = E\{\log p(z|\beta_{[k]})|y, \beta\}.$$

The M-step then finds the parameter that satisfies

$$\beta_{[k+1]} = \arg\max Q(\beta, \beta_{[k]}).$$

In 1990, Wei and Tanner[36] proposed the Monte Carlo EM (MCEM) algorithm that replaces the E-step of the original EM algorithm by a Monte Carlo integration. We again provide a simplified explanation of the process.

The E-step at iteration k is replaced by the simulation (S)-step[36,37]. In the S-step, we first perform $m_{[k]}$ realizations $z_{k[j]}$ $(j = 1, .., m_{[k]})$ of the missing data vector under $p(z|\beta_{[k]})$. Then, we compute the Monte Carlo approximation of $Q(\theta, \theta_{[k]})$ according to

$$\tilde{Q}_k(\beta) = m_{[k]} \sum_{j=1}^{m_{[k]}} \log f(z_{k[j]}; \beta).$$

The M-step remains unchanged[37].

The SAEM algorithm builds upon the MCEM algorithm by changing the Monte Carlo approximation $\tilde{Q}_k(\beta)$ to

$$\hat{Q}_k(\beta) = \hat{Q}_{k-1}(\beta) - \gamma_k \left(\frac{1}{m_{[k]}} \sum_{j=1}^{m_{[k]}} \log f\left(z_{k[j]}; \beta\right) - \hat{Q}_{k-1}(\beta) \right),$$

where $\{\gamma_k\}_{k \geq 1}$ is a sequence of positive step size[37].

The M step is changed in the SAEM algorithm[37]. $\hat{Q}_k(\beta)$ is maximized in the feasible set Ω, meaning that we find a $\beta_{k+1} \in \Omega$ such that

$$\hat{Q}_k(\beta_{k+1}) \geq \hat{Q}_k(\beta) \qquad \forall \beta \in \Omega.$$

The convergence of SAEM is dependent on the step size γ_k and/or the specification of $m_{[k]}$ used in the process. It is recommended that one decrease γ_k and and/or increase $m_{[k]}$ over the simulation attempts.

1.3.2 Frequentist approach to model selection

Model selection is the last but critical step for determining an optimal fitted model for inference. Like with conventional GLM, the most familiar techniques for model selection are based off predictor selection. Traditional indicators like p-value are used to determine whether a parameter in the model is statistically significant or should otherwise be removed from the model. On the other hand, information theory design, where an information criterion is used, has been rapidly gaining popularity as a superior tool for model selection[7].

Model selection using the frequentist approach to experimental design relies on using p-values. In this approach, each predictor is tested for significance and the resulting p-value determines whether it should remain in the model. Since adding or removing a predictor changes the significance of all other predictors, this approach can only compare two models that differ in a single predictor at a time.

In fixed effects models, t-tests and p-values indicate how significant a predictor is in the model. However, deciding on whether the predicted variance is significant for a mixed effects model proves to be a difficult task. Unlike fixed effects models, there is not a reliable, appropriate

denominator degrees of freedom for mixed effects models unless the study design is very simple with nicely balanced data[38]. This is due to inherent uncertainty associated with counting parameters in a model that has more than one level, as in nested or crossed design[38]. For most studies, the classical p-value offers little value in predictor evaluation. The most common methods for evaluating significance for mixed effects models are also rather ineffectual for small datasets.

For models without random slopes, p-values can be estimated using Markov-chain Monte Carlo (MCMC) sampling, which repeatedly samples from the posterior distribution of the model parameters. Unfortunately, this caveat is too large to ignore for model selection, and this method has fallen out of favour in popular mixed effects model software (e.g., R package *ime4*[39])[38]. The most used methods are the likelihood ratio test (LRT) and to use the z-distribution to obtain p-values from the Wald t-values (t-as-z). LRT is most commonly used to evaluate whether a random effect improves the fit of the model when all other parameters are held constant[38]. LRT has the advantage that it can be used for complex model structures as it also does not require an input for degrees of freedom. Alternatively, we can use the t-as-z method given the convergence of the t distribution to the z distribution as degrees of freedom increase to infinity. Since mixed effects models necessitate a robust dataset to produce accurate results, evaluating significance by testing the Wald t-values as though they are z-distributed to generate p-values is appealing. This method has the added bonus of working with t-tests and p-values. However, there is no formal guideline to determine a critical size for a dataset to justify the z-as-t method[38]. Both LRT and Wald t-values appear to be anti-conservative for smaller sample sizes, so one should be cautious if the number of units or groups are small (<40-50).

Stepwise selection using p-value is by far the most popular approach for null-hypothesis testing[7]. However, there is increasing criticism of this methodological design since stepwise selection can overestimate the effect size of significant predictors. This procedure also forces the idea that only one combination of predictors can adequately represent the data[7], which is an unreasonable assumption in regards to the complexity of biological systems. Furthermore, every significance test inflates Type 1 errors, and this builds on the shaky LRT estimation for mixed effect models.

1.3.3 Information theoretic approach to model selection

Given the limitations of p-value testing, information theory has become more popular following the rise of IT design[7]. Proponents seek to replace t test and ANOVA tables with information criterion, claiming that is it easier to compute and understand, on top of the benefits of IT design. Information criterions are used for model selection. After building a set of potential models for the data, the next step is to determine which of these models best describe the data, and in turn provide the best estimations for future data. Instead of choosing to include or exclude parameters in the model, like model selection with p-values, information criterions act as a ranking system and sort the set of models from best to worst. There are several information criteria available, each suited for different scenarios.

1.3.3.1 Akaike information criterion

First formulated in 1973[40], the Akaike information criterion (AIC) is one of the most popular information criteria. Given a family of approximating model f_ψ with unknown parameters $\psi \in \Psi$ generated from a vector of observations $y = (y_1, \dots, y_n)$ from the true underlying distribution g, AIC is defined as

$$AIC = 2k - 2\ln\big(f(y|\hat{\psi}(y)\big), \qquad (1.17)$$

where $\hat{\psi}(y)$ is the maximum likelihood estimator based on data y and $k = \dim(\psi)$ (i.e., the number of estimated parameters[41,42]).

AIC is an estimator of the Akaike information given by

$$AI = -2E_y E_{y^*} lnf(y^*|\hat{\psi}(y)),$$

which assesses a model's ability to fit future data y^* given the parameters $\hat{\psi}$ fitted through data y, assuming that the two datasets are independent. In the equation above, y and y^* are two independent datasets generated from the true underlying distribution g, and E_y and E_{y^*} are the expectations for the distribution of the respective realization, y and y^*[41]. Minimising AI is the equivalent of minimising the expectation of the distance between the true distribution and the model distribution, indicating a superior fit of the model. Since AIC is an asymptotically unbiased

estimator of AI, minimising AIC over a set of possible models can be seen as minimising the average distance of an approximating model to the underlying truth[41,42].

In practice, AIC calculates a model's trade-off between the goodness of fit, described by the log-likelihood $\ln(f(y|\hat{\psi}(y))$, and the simplicity of the model described by k, or the number of parameters used in the model. The consideration of model complexity is critical to evaluating the model of best fit. In a regression type model, the likelihood will always increase with the number of predictors simply by capturing more noise and decreasing standard error. However, increasing the number of predictors runs the risk of overfitting, so the AIC punishes the addition of predictors. The terms are multiplied by 2 in Eq. (1.17) as $-2\ln(f(y|\hat{\psi}(y))$ is known as the "deviance"[43]. AIC is used in model selection by minimising the score, since a lower score would mean that the model provides the best balance between predictive power and complexity.

AIC is a very popular criterion for model selection. Since its publication, there have been many variations created for more specific situations. One of the first corrections to the original AIC is a correction for small samples derived in 1978 given by

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1},$$

which converges to AIC as n, the number of samples, increases[44].

For mixed effects models, the interpretation of AIC is different. The formula stays the same as in Eq. (1.17), but the terms k and $\ln\left(f(y|\hat{\psi}(y)\right)$ differ depending on the research question. Recall in Section 1.2.2, we discussed the difference between population and cluster focused hypotheses. For inference that focuses on the population parameters, the likelihood function is the marginal likelihood and k is the number of fixed parameters, counting mean parameters and variance components. For inference that focuses on the clusters, the likelihood is the conditional likelihood, and $k = \rho + 1$, where ρ is the effective number of parameters for the mean model.

The likelihood of a model is $L(y|\theta, u)$, which means that the marginal maximised likelihood for population focus AIC is $L\left(y\middle|\hat{\theta}\right)$. Since in a population focus the random effects are used to capture noise and correlation between the clusters, there is usually no inference or prediction to

be done with them[20]. The marginal likelihood likewise removes the predictive properties that would otherwise be taken into consideration in calculating for the AIC. The marginal AIC ($mAIC$) is given by

$$mAIC = 2(p + q) - 2\ln L\left(y\middle|\hat{\theta}\right),$$

where p is the number of fixed effects parameter β, and q is the number of variance parameters θ of u[20,41]. However, if REML was used for estimation (see discussion above), the number of parameters is simply q. As such, mAIC based on REML cannot be used to compare models with differing fixed effects[41].

For cluster focus, u_i are the focal point and the parameters to be estimated. Thus, prediction at the cluster level is conditional on the clusters and u_i act as parameters. The relevant likelihood in this case is $L\left(y^*\middle|\hat{\theta}(y), \hat{u}\right)$, where $\hat{\theta}(y)$ is the maximum likelihood estimator of θ, and $\hat{u} = E(u|\hat{\theta}, y)$ is the empirical Bayes estimator. In addition, the conditional AI must accommodate for the fact that y and y^* should, not only share a same true distribution f, but share the same random effects ω and differ in their error terms. In practice, this translates to the conditional AIC ($cAIC$) given by

$$cAIC = 2(\rho + 1) - 2\ln L(y|\hat{\theta}(y),\ \hat{u}(y)),$$

where ρ is the effective degrees of freedom, a general term to describe degrees of freedom when it is not straightforward to count the number of parameters[45]. In the case of a mixed effect model,

$$\rho = tr\left[\begin{pmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ + \sigma^2 G(\theta)^{-1} \end{pmatrix}\begin{pmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ \end{pmatrix}\right]$$

i.e., $\rho = tr(H_1)$, where H_1 is the hat matrix mapping the observed data y into the fitted $\hat{y} = X\hat{\beta} + Z\hat{u}$ i.e. $\hat{y} = H_1 y$, and $G(\cdot)$ is the covariance matrix[41]. As such, cAIC can only be used for linear mixed model.

In general, AIC is a well-developed criterion that has many variants to accommodate and correct for different scenarios. AIC does have the tendency to favour overly complex models[43], so it is wise to be wary of overfitting when using AIC.

1.3.3.2 Bayesian information criterion

The Bayesian information criterion (BIC) is very similar to AIC in form, but its derivation is based on a Bayesian framework. The BIC formula is defined as

$$BIC = kln(n) - 2\ln L(\hat{\theta}|y),$$

where $\hat{\theta}$ is the vector of estimates obtained by maximising the likelihood $L(\hat{\theta}|y)$, n is the total number of samples, and k is the number of parameters[46]. Like AIC, model selection is done based on minimising BIC.

BIC has the advantage of being consistent, meaning it will asymptotically select the model having the correct structure[46]. If the true model is not one of the models being tested, then the BIC will choose the model that is the closest to the true model. BIC works under the assumption that the true model is within the pool of candidate models. However, a downside of this assumption is that BIC performs worse than AIC when the true model is not included in the selection[46]. BIC also punishes the addition of parameters more harshly than AIC thus favouring models that are more parsimonious. For small to moderate dataset sizes, BIC tend to outperform other popular model selection criteria and works better for low dimensional models as compared to AIC[46].

No matter the type of information criterion used, one must always determine a set of 'top models' containing all the models that are assumed to be supported by the data. This group is often determined by the change in criterion score from the best model,

$$\Delta_i = AIC_i - \min_i AIC_i,$$

where $\min_i AIC_i$ is the AIC score of the best model. Note that the equation holds for the other variations of AIC or for BIC score. The difference Δ_i estimates the information loss when using model i. A larger score equates to a less plausible model[30].

For both AIC and BIC, 0-2 difference states that the two models are essentially the same in fit[7,46]. A difference between 4 and 7 should cast some doubts on the model, while a model with a difference more than 10 has no support and can be omitted[30]. Depending on the number of

models tested and their scoring, one may choose to relax or tighten this range. Within this set of best models, it is suggested to favour smaller models, especially if using AIC, to avoid overfitting[41].

One benefit of using an information criterion instead of p-values is that information criterions can be used to compare linear and nonlinear mixed effect models as well as fixed effect regression. Since an information criterion estimates the difference between the true underlying distribution and model distribution, it is less affected by the linearity of the model and the presence of random effects[47].

1.3.3.3 Weighted criterions

Sometimes, the score difference between the top ranked models is quite small and just choosing the lowest AIC or BIC score may not select the best model in the set. A transformation of the raw score with weights can solve the issue.

For AIC, the weighted AIC is,

$$w_i(AIC) = \frac{\exp\left(\frac{1}{2}\Delta_i(AIC)\right)}{\sum_{k=1}^{K}\exp\left(\frac{1}{2}\Delta_k(AIC)\right)},$$

where ΔAIC is the difference between two models' AIC score. Note that $\sum w_i(AIC) = 1$. Weighted AIC can be interpreted as the probability that model M_i is the best model given the data and the other models. This of course means that Akaike weights are susceptible to sampling variability and may different if the dataset is changed. Now, the ratio between two Akaike weights states the probability that the numerator model is the best model within the set i.e., $r = \frac{w_2(AIC)}{w_1(AIC)}$. For example, if the ratio between two model is 1.5, then the numerator model is 1.5 times more likely than the denominator model to be the best model. Alternatively, we can normalize this evidence ratio as $r = \frac{w_2(AIC)}{w_1(AIC)+w_2(AIC)}$[48].

BIC can also be converted into its weighted version by replacing the equation for weighted AIC with BIC scores[48].

1.3.3.4 Model averaging

In biomedicine, a system's true structure is very complex[7,15], which means that multiple models can be plausible and provide multiple predictions. Model averaging is the technique of using several models to make inference. Parameter estimation can also be done through model averaging, but the practice is currently under scrutiny and not recommended[7]. While it may be tempting to average a large number of statistical models within the set, averaging the weights for all the models may cause even the best models to have a small weight. The number of models used should be first decided based on the hypothesis; a detailed question require fewer models than a wide scoped question. All redundant models should be removed at this step. Model averaging should be performed after choosing the 'best model' set, and the weights used for averaging determined by the weighted criterions. However, over-limiting the set size risks excluding the best model in the set[4]. Thus, model averaging should be used in conjunction with model selection to first select an optimal set for averaging.

A particular complication occurs if a factor of interest is excluded from the top model set. One must make the judgement to conclude that the factor poses little influence in the system or extend the model set to include at least one model that contains this factor. The latter solution may result in a bloated 'top model set' and let in overly complex models or too many other irrelevant predictors. A tentative solution is to exclude any of the complex models that nest models of a lower criterion score, though of course this causes its own issue of potentially excluding important models[4]. Unfortunately, this area remains unclear, and we will have to continue to rely on intuition for the time being.

Once this set of models is defined, there are two methods by which the estimate and error for each model's parameters are weighted. In the natural average method, the parameter estimate for each predictor is only averaged over models in which it appears and is weighted by the summed weights of these models. In the zero method, the parameter estimate is averaged over all models, where the parameter estimate in models without the predictor are replaced with a zero. The zero method essentially dilutes the effect size and errors of the parameter estimates. The decision between using either of these two techniques is based on the particularities of the research question. The zero method is convenient for determining the factors with the strongest

effects on the response variable, while the natural average method preserves the strength of each parameter so weaker factors of interest can still be studied. Using the natural average method also leads more flexibility in the final model, as the inclusion of any factor does not influence the calculation of effect size or SE[4].

1.3.4 Structure and goals of this thesis

This thesis is structured as follows.

In Chapter 2, we describe a classical approach to mixed effects models by studying maple syrup production in Québec and the development of a novel assay to assess syrup quality. This paper will be submitted for publication in the peer-reviewed journal *ACS Food Science & Technology* shortly. My contribution to this work was the conception, development, and analysis of the models, and writing and editing the manuscript.

In Chapter 3, we explore the combination of mechanistic mathematical models and a nonlinear mixed effects model applied to describe the physiological interactions of estrogen in mammary stem cell differentiation. Here the focus is on the degree of noise captured within mechanistic models, in contrast to pharmacokinetic models that are estimated using NLMEM but without consideration of the underlying physiological system. This paper is published in AIMS Mathematics (Le Sauteur-Robitaille, Yu, and Craig, Impact of estrogen population pharmacokinetics on a QSP model of mammary stem cell differentiation into myoepithelial cells. AIMS Mathematics, 6(10), 10861-10880. 2021. 6(10), 10861-10880. 2021.). My contribution to this chapter was the adaptation, analysis, and use of a previously published nonlinear mixed effects model into the quantitative systems pharmacology model we developed to generate a virtual population to explore the effects of the nonlinear mixed effects model on the physiological model, and manuscript writing.

Finally, in Chapter 4, we provide perspectives on the use of mixed effects models in biological applications.

References

1 Rohatgi, V. K. S. A. K. M. E. *An introduction to probability theory and statistics*. (2015).

2 Jones, S. R., Carley, S. & Harrison, M. An introduction to power and sample size estimation. *Emerg Med J* **20**, 453-458, doi:10.1136/emj.20.5.453 (2003).

3 Kang, D., Schwartz, J. B. & Verotta, D. A sample size computation method for non-linear mixed effects models with applications to pharmacokinetics models. *Statistics in Medicine* **23**, 2551-2566, doi:https://doi.org/10.1002/sim.1695 (2004).

4 Grueber, C. E., Nakagawa, S., Laws, R. J. & Jamieson, I. G. Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology* **24**, 699-711, doi:https://doi.org/10.1111/j.1420-9101.2010.02210.x (2011).

5 Suits, D. B. Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association* **52**, 548-551, doi:10.1080/01621459.1957.10501412 (1957).

6 in *Data Analysis Using Regression and Multilevel/Hierarchical Models Analytical Methods for Social Research* (eds Andrew Gelman & Jennifer Hill) 251-278 (Cambridge University Press, 2006).

7 Harrison, X. A. *et al.* A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **6**, e4794-e4794, doi:10.7717/peerj.4794 (2018).

8 Davidian, M. & Giltinan, D. M. Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics* **8**, 387, doi:10.1198/1085711032697 (2003).

9 Lindstrom, M. J. & Bates, D. M. Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics* **46**, 673-687, doi:10.2307/2532087 (1990).

10 Brown, H. & Prescott, R. *Applied mixed models in medicine*. (John Wiley & Sons, 2015).

11 Craig, M., González-Sales, M., Li, J. & Nekka, F. in *Mathematical Sciences with Multidisciplinary Applications.* (ed Bourama Toni) 91-112 (Springer International Publishing).

12 Cleasby, I. R., Nakagawa, S. & Schielzeth, H. Quantifying the predictability of behaviour: statistical approaches for the study of between-individual variation in the within-

individual variance. *Methods in Ecology and Evolution* **6**, 27-37, doi:https://doi.org/10.1111/2041-210X.12281 (2015).

13 Hruschka, D. J., Kohrt, B. A. & Worthman, C. M. Estimating between- and within-individual variation in cortisol levels using multilevel models. *Psychoneuroendocrinology* **30**, 698-714, doi:https://doi.org/10.1016/j.psyneuen.2005.03.002 (2005).

14 Bon, C. *et al.* Mathematical modeling and simulation in animal health. Part III: Using nonlinear mixed-effects to characterize and quantify variability in drug pharmacokinetics. *Journal of Veterinary Pharmacology and Therapeutics* **41**, 171-183, doi:https://doi.org/10.1111/jvp.12473 (2018).

15 Schielzeth, H. & Nakagawa, S. Nested by design: model fitting and interpretation in a mixed model era. *Methods in Ecology and Evolution* **4**, 14-24, doi:https://doi.org/10.1111/j.2041-210x.2012.00251.x (2013).

16 Brauer, M. & Curtin, J. J. Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychol Methods* **23**, 389-411, doi:10.1037/met0000159 (2018).

17 Martens, H. A. & Dardenne, P. Validation and verification of regression in small data sets. *Chemometrics and Intelligent Laboratory Systems* **44**, 99-121, doi:https://doi.org/10.1016/S0169-7439(98)00167-1 (1998).

18 Bell, B. A., Ferron, J. M. & Kromrey, J. D. Cluster Size in Multilevel Models: The Impact of Sparse Data Structures on Point and Interval Estimates in Two-Level Models. (2008).

19 Bates, D., Kliegl, R., Vasishth, S. & Baayen, H. Parsimonious mixed models. *arXiv preprint arXiv:1506.04967* (2015).

20 Vaida, F. & Blanchard, S. Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351-370, doi:10.1093/biomet/92.2.351 (2005).

21 Ribba, B. *et al.* A Tumor Growth Inhibition Model for Low-Grade Glioma Treated with Chemotherapy or Radiotherapy. *Clinical Cancer Research* **18**, 5071, doi:10.1158/1078-0432.CCR-12-0084 (2012).

22 Remick, M., Ibrahim, G. M., Mansouri, A. & Abel, T. J. Cross-national disparities contribute to heterogeneity in patient outcomes following invasive monitoring: A hierarchical mixed-effects analysis. *Epilepsia* **61**, e116-e123, doi:https://doi.org/10.1111/epi.16647 (2020).

23 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-354, doi:10.1038/ng.548 (2010).

24 Harrison, X. A. A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution. *PeerJ* **3**, e1114-e1114, doi:10.7717/peerj.1114 (2015).

25 Austin, P. C. & Leckie, G. The effect of number of clusters and cluster size on statistical power and Type I error rates when testing random effects variance components in multilevel linear and logistic regression models. *Journal of Statistical Computation and Simulation* **88**, 3151-3163, doi:10.1080/00949655.2018.1504945 (2018).

26 McNeish, D. M. & Harring, J. R. Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics - Simulation and Computation* **46**, 855-869, doi:10.1080/03610918.2014.983648 (2017).

27 Zuur, A. F. & Ieno, E. N. A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution* **7**, 636-645, doi:https://doi.org/10.1111/2041-210X.12577 (2016).

28 Zuur, A. F., Ieno, E. N. & Elphick, C. S. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* **1**, 3-14, doi:https://doi.org/10.1111/j.2041-210X.2009.00001.x (2010).

29 Burnham, K. P. Model selection and multimodel inference. *A practical information-theoretic approach* (1998).

30 Portet, S. A primer on model selection using the Akaike Information Criterion. *Infect Dis Model* **5**, 111-128, doi:10.1016/j.idm.2019.12.010 (2020).

31 McNeish, D. Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction. *Multivariate Behavioral Research* **52**, 661-670, doi:10.1080/00273171.2017.1344538 (2017).

32 Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1-38 (1977).

33 Kenward, M. G. & Roger, J. H. Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics* **53**, 983-997, doi:10.2307/2533558 (1997).

34 Kackar, R. N. & Harville, D. A. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* **79**, 853-862 (1984).

35 Walker, S. An EM Algorithm for Nonlinear Random Effects Models. *Biometrics* **52**, 934-944, doi:10.2307/2533054 (1996).

36 Greg, C. G. W. & Tanner, M. A. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association* **85**, 699-704, doi:10.2307/2290005 (1990).

37 Bernard, D., Marc, L. & Eric, M. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics* **27**, 94-128, doi:10.1214/aos/1018031103 (1999).

38 Luke, S. G. Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods* **49**, 1494-1502, doi:10.3758/s13428-016-0809-y (2017).

39 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1 - 48, doi:10.18637/jss.v067.i01 (2015).

40 Akaike, H. in *Selected Papers of Hirotugu Akaike* (eds Emanuel Parzen, Kunio Tanabe, & Genshiro Kitagawa) 199-213 (Springer New York, 1998).

41 Greven, S. & Kneib, T. On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97**, 773-789 (2010).

42 Benjamin, S., Thomas, K., Clara-Sophie van, W. & Sonja, G. A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electronic Journal of Statistics* **8**, 201-225, doi:10.1214/14-EJS881 (2014).

43 Burnham, K. P., Anderson, D. R. & Huyvaert, K. P. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology* **65**, 23-35, doi:10.1007/s00265-010-1029-6 (2011).

44 Sugiura, N. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-theory and Methods* **7**, 13-26 (1978).

45 Overholser, R. & Xu, R. Effective degrees of freedom and its application to conditional AIC for linear mixed-effects models with correlated error structures. *Journal of Multivariate Analysis* **132**, 160-170, doi:https://doi.org/10.1016/j.jmva.2014.08.004 (2014).

46 Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *WIREs Computational Statistics* **4**, 199-203, doi:https://doi.org/10.1002/wics.199 (2012).

47 Buatois, S., Ueckert, S., Frey, N., Retout, S. & Mentré, F. Comparison of Model Averaging and Model Selection in Dose Finding Trials Analyzed by Nonlinear Mixed Effect Models. *The AAPS Journal* **20**, 56, doi:10.1208/s12248-018-0205-x (2018).

48 Wagenmakers, E.-J. & Farrell, S. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review* **11**, 192-196, doi:10.3758/BF03206482 (2004).

Chapter 2 – Predicting maple syrup quality using mixed effects models

Zhe Si Yu[1], Issraa Beniani[2], Juan Manuel Montiel Leon[2], Simon Forest[2,3], Jean-François Masson[2], Morgan Craig[1,4*]

[1]Department of Mathematics and Statistics, Université de Montréal, Montréal, Canada

[2]Department of Chemistry, Université de Montréal, Montréal, Canada

[3] Producteur et productrices acéricole du Québec (PPAQ), Longueuil, Canada

[4]Sainte-Justine University Hospital Research Centre, Montréal, Canada

*Corresponding author

Keywords: Mixed effects models, maple syrup quality, COLORI, mathematical modelling

Abstract: Maple syrup is produced by boiling sap from the sugar maple tree and is mostly comprised of sugar and water. The organic molecules that give maple syrup its distinctive taste can also give off-flavour if the syrup is produced too late in the season, rendering the syrup less marketable and wasting product and effort. Though certain quantitative measures exist to help producers judge when to finish the production season, most rely on heuristic measures. To help producers predict the end of the season, we developed two mixed effects models (ordinal and nonlinear) that integrate several such quantitative measures. These include transmittance, pH, and COLORI, a colorimetric test we developed previously that uses gold nanoparticles aggregation to discriminate maple syrup flavour profiles. Our results suggest that these models can help guide the maple syrup production season by predicting syrup quality either outright (ordinal) or by use of total amino acid concentrations (nonlinear). These models are therefore important tools for producers to help mitigate financial and person-time losses at the end of the season.

2.1 Introduction

Maple syrup is produced by boiling sap from *Acer saccharum* (sugar maple tree). The province of Quebec, Canada is the largest producer of maple syrup across the world, accounting for 72% of global maple syrup production in 2020[1]. In 2020, Quebec producers contributed to around 175 million pounds (79.4 million kilograms) of maple syrup, which translates to a net worth of more than 505 million dollars[1].

While maple syrup is mostly comprised of sugar (66%) and water (33%), it is the last 1% of organic molecules, including phenols, pyrazines, minerals, organic acids and amino acids, that give maple syrup its distinctive taste[2]. The process of producing the syrup is demanding and heavily dependent on seasonal variations. The Day to Bud Break (DTBB), which marks the number of days until sugar maple trees start budding, is one of the most important determinants to syrup quality. This is because the maple sap harvest occurs throughout the snow melt days in the beginning of the spring season but ends abruptly with the emergence of buds. While syrup quality increases up to zero DTBB (the day to bud break), it decreases after this day, rendering syrups produced after this point subject to potential losses. Between the start of the production, tapping the maple trees for sap, to the final production of syrup, several days pass, thus if not watchful, could translate to a waste of efforts and product. The quality of the syrup is graded at the end of the season by both the ACER Inspection Division and the "Producteurs et productrices acéricoles du Quebec" (PPAQ)[3]. Together, they assign a grade to the syrup that determines to whom it can be sold and its market price. Every barrel of maple syrup produced in the province of Quebec is tested. The highest quality syrups are sold for retail markets (i.e., consumer markets), whereas the grades that have not passed a predetermined quality threshold (assuming the quality is not so low that it must be thrown out) are sold for industrial markets, resulting in an important economic loss for the producers[3]. Because of the delay between syrup production and syrup grading, the few days of delays can result in potential losses for the producers.

There are currently some quantitative measurements available to aid producers in the determination of syrup quality throughout the production process. These include transmittance (colour determination), presence of lead, and °Brix (quantification of the sugar content by refractometry)[3]. These tests can be used in addition to the trained technicians hired by the PPAQ

who grade the final product to profile syrup quality. However, despite these indicators, producers still heavily rely on their intuition and experience to judge when to stop production. Furthermore, because every producer is in a different geographical location, the DTBB is different for each producer. For example, if they are further north, they have a later DTBB than someone in the south of the province, because they often experience spring later in the year. In other words, the quality of the syrup for any batch during the season relies on the producer's experience but is also some kind of guessing game until the syrup is evaluated.

One of the largest barriers to more accurate testing of the minute differences in organic molecule content is the absence of a convenient protocol. In some cases, fluorescence spectroscopy is used to evaluate the fluorescent organic molecules in the maple syrup to classify the product into normal or off-flavour categories. However, this process is not readily accessible to producers, as it requires specialised equipment and training to conduct[2].

Colorimetric tests are an alternative method that have the benefit of being available to untrained users, but they also face the issue of being unpractical for the rapid classification of on-site testing of the nearly 300,000 barrels of syrup produced annually. Furthermore, colorimetric tests using classical dyes are unsuited for detecting the complex changes in flavour profiles of maple syrup. To address these limitations, Forest et al.[2] created a non-specific colorimetric test, coined COLORI, that instead uses gold nanoparticles aggregation to discriminate maple syrup flavour profiles.

The COLORI test uses gold nanoparticles (AuNPs) which are tiny gold particles the size between 1 to 100 nanometres (nm). These particles do not have to be spherical but can instead be a variety of sizes and shapes, including nanostars and nanoraspberries (such as those used by Forest et al.[2]). AuNPs have several distinct qualities that make them suitable for a multitude of tasks and they make excellent biosensors.

Surface plasmon resonance is one of the more desirable traits of AuNPs[4]. Surface plasmon resonance (SPR) is a phenomenon that has attracted increased interest in recent years. The theory behind the mechanism is beyond the scope of this paper, but we will provide a simplified description.

Incident light, or incoming light from light sources towards a material, will interact with a material by refracting, reflecting, or absorbing. In refraction, the light ray changes directions entering and exiting the medium whereas the light ray rebounds at the same angle as the angle of the incident light ray when reflecting[5]. Different media will interact with light differently depending on their quality, which is described by the refractive index. Formally defined as the ratio of the velocity of light in a vacuum to the velocity of light in a material[6], a higher value indicates that the medium has a smaller refractive angle, or in other words bends towards the perpendicular lines to the surface. When light travels from medium with a relatively higher refractive index to one of a lower refractive index, the ray of light tends to reflect as opposed to refract[5]. When light does not cross the boundary and is entirely reflected within the medium, a total internal reflection (TIR) occurs. During the occurrence of TIR at the boundary between two nonabsorbing media, the fully reflected light ray leaks some of its electrical field into the medium with the lower refractive index. This electrical field is known as the evanescent field and its amplitude exponentially decays away from the interface. In SPR, the evanescent wave (from the evanescent field) excites electrons within the metal layer of a metal and dielectric interface, yielding surface plasmons. Surface plasmons are electromagnetic surface waves that propagate parallel to the interface. The wave will change in intensity when one of its media changes in property in some way. In AuNPs, the SPR is the result from the collective oscillation of the electrons across the nanoparticle and is extra convenient as the resonance conditions is satisfied at visible wavelengths[4]. The COLORI test developed by Forest et al. takes advantage of this trait further by studying the effects of the aggregation of two or more AuNPs, which would cause a change in the observed resonance[2]. Depending on the presence of varying amino acids in the syrup, the changes in resonance, and thus observed colours, can indicate the quality of the syrup.

In a test of 1818 validated syrup samples from the 2018 harvest, COLORI was shown to detect off-flavour maple syrup from normal syrup with 98% accuracy[2]. It was established that normal flavour profiles led to strongly aggregated AuNPs and an absorbance higher than 520 nm, while off-flavour syrups showed little to no aggregation and less than 625 nm absorbance. The corresponding colours of these scenarios are red and blue, respectively. However, the quality of

maple syrup is a gradient and COLORI returns uncertain results if the flavour profile is between normal and off-flavours.

There exists a large range of quality between normal and off-flavours of maple syrup that cannot be readily predicted by COLORI. Within this range of quality are several categories that indicate broad flavour notes of the maple syrup (Table 2.1). While the categories OK and "Crochet" are considered suited for commercial sale directly to consumers, the categories that indicate flavour defects (VR1 to VR5) as to PPAQ quality grading standards, are only sold for industrial uses. As such, maple syrups with flavour defects are submitted to a market price reduction. VR5, for instance, with its buddy flavour profile, receives a particularly high price cut.

Quality Category	Description	Prominent Note
OK	No detectable off-flavours	N/A
Crochet	Slight defect; still acceptable	Any slight off-flavours
VR1	Defect of natural origins	Flavours of wood, sap, over caramelization, burnt etc.
VR2	Microbiological defect	Mold, fermentation
VR4	Defects of artificial tastes	Metal, soured, acid, smoke, antifoam
VR5	Defect due to budding	Specific bad budding taste

Table 2.1. – Summary of the quality ratings for maple syrup flavours. Adapted from the PPAQ[3].

Current methods for effectively differentiating syrup quality, like COLORI, still have the issue of being difficult to use in the field. As one can imagine, working with gold nanoparticles requires specialized training. Furthermore, COLORI is not very accurate for quality categories between the outer ranges. There is thus a need to investigate the relationship between maple syrup quality and its quantitative measurements using mathematical models.

To establish this relationship, we used mixed effects models, that better represent experiments with repeated measures, to predict how COLORI varies as a function of transmittance, pH, and the amino acid (AA) concentration, which has been shown before to be a good indicator of syrup quality. Experiments designed using repeated measures allocate multiple treatments throughout

time to the subjects. Strictly speaking, repeated measures are not an ideal experimental strategy, but they are largely unavoidable in practice. By using mixed effects models, we can capitalize on the more realistic setting of a repeated measure experimental design. Because repeated measures are often used for longitudinal studies, time as a variable is built into the structure of a mixed effects model. However, we do not know the DTBB for any given year nor producer, implying that we cannot use a true time variable. Thus, we established an ordinal mixed effects model that uses transmittance, COLORI and AA as input and quality as output, and a nonlinear mixed effects model using some of the quantitative measurements, like pH and COLORI, as inputs to predict the amino acid (AA) concentration. Our results suggest that these statistical models can be used in parallel to COLORI to prospectively predict maple syrup quality, with potential economic benefits to producers.

2.2 Methods

2.2.1 Ordinal mixed effects model

We first created a general linear mixed effects model with an ordinal link function (i.e., an ordinal mixed effects model) to predict the PPAQ rating grades of maple syrup quality. Let $j = 1, \dots, J$ denote each producer, $i = 1, \dots, n_j$ be the samples of any given producer j, and Y_{ij} the ordinal response for maple syrup quality associated with each producer i and sample j. Let $c = \{1, 2, 3, 4\}$ denote the ordered maple syrup quality ordinal response (i.e., VR5, VR1, "crochet" and "aucun", respectively). We defined the cumulative probabilities of the quality ratings of the ordinal outcome Y as

$$P_{ijc} = P(Y_{ij} \leq c) = \sum_{k=1}^{c} p_{ijk},$$

where p_{ijk} denotes the individual quality rating probabilities. The ordinal mixed effects model given in terms of the cumulative logits $\gamma_{ijc}, c = 1, 2, 3$ is

$$\gamma_{ijc} = \log\left(\frac{P_{ijc}}{1 - P_{ijc}}\right) = \gamma_c - [X_{ij}\theta_c + Z_{ij}u_{jc}], \tag{2.1}$$

where X_{ij}, Z_{ij} are the elements of design matrices of the fixed and random effects, respectively, θ_c is the fixed effects for ordinal response c , u_{jc} is the vector of random effects for ordinal response c that are normally distributed with ω as its variance-covariance matrix, $\epsilon_j \sim N(0, \sigma^2 I_{n_j})$, γ_c is the threshold value for c, and $N = \sum n_j$ is the total number of observations. We did not include any covariates in this model.

Parameters for the model in Equation (2.1) was fit using the 'Ordinal' library and function 'clmm' in R version 4.1.2

2.2.2 Nonlinear mixed effects model

To further investigate the relationship between COLORI and AA concentration, we altered the classic nonlinear mixed effects model (NLMEM) by using a proxy variable (pH) to estimate the time variable. This was to avoid using the DTBB, which is not known during the season and thus cannot be used as an independent model variable. Instead, the nonlinear mixed effects model contains three nested equations of pH, COLORI and AA, the structure of which were each determined through data exploration. The model is given by

$$\tau = k_{pH1} pH + k_{pH2} e^{pH} + c_{pH}, \tag{2.2}$$

$$COLORI = \frac{k_{colori} \tau^h}{a^h + \tau^h}, \tag{2.3}$$

$$AA = \frac{k_{AA}}{COLORI} + c_{AA}. \tag{2.4}$$

Unconventionally, we replaced the time variable in Equation (2.2) by the dummy time variable τ through the sum of a linear term of pH, with slope term k_{pH1}, an exponential term $\exp(pH)$ with coefficient k_{pH2}, and a constant term c_{pH}. τ is used as in input to predict $COLORI$ in Equation (2.3) through a Hill function, where h is the Hill coefficient and $k_{colori} = 25$ as $COLORI$ ranged

between 0 and 25 in the experimental setup. Finally, AA was predicted through an inverse function of COLORI with a coefficient $\mathrm{k_{AA}}$ and constant $\mathrm{c_{AA}}$ in Eq. (2.4).

Parameters in Eq. (2.2) were estimated using random effects to account for noise when using pH as an estimate for time. In Eq. (2.3), random effects were used to instead account for any differences between producers, including different scales or dates of DTBB. The parameters in Eq. (2.4) were all assumed to be fixed effects parameters. This is because COLORI testing measures the amino acid content of maple syrup samples and was assumed to not differ according to the producers of the syrup. The model in Eqs. (2.2)-(2.4) reflect the relationship between COLORI testing results and AA content by having population parameters for all producers. In Eq. (2.3), parameters a and $\mathrm{k_{pH}}$ were assumed to have correlated random effects. The NLMEM in Equations (2.2)-(2.4) was fitted using Monolix 2021R1.

2.3 Results

2.3.1 Ordinal mixed effects model suitable for quality prediction

Firstly, we wanted to evaluate the relationships between the predictive variables (Table 2.2) and maple syrup quality. We found that quality increased with increases in transmittance and COLORI, while it decreased with increases in AA concentrations (see Table 2.3 for a list of parameter estimates of the model in Equation (2.1), including the fixed and random effects and the threshold coefficients). Further, our parameter estimates showed that the coefficients for COLORI and AA were both highly significant. Conversely, though transmittance was found to have an effect, it was not determined to be significant. We nonetheless included transmittance in the model as it significantly lowered the Akaike Information Criterion (AIC[7]) score from 456.85 to 288.82 from the model with just COLORI and AA as inputs. Other variables, namely pH and °BRIX, were not found to be significant nor did they significantly lower the AIC. Thus, to reduce the risk of overfitting with additional predictor variables, they were excluded from the final model.

Variable	Unit	Description

pH	-	Regular scale of pH limited to relevant range for maple syrup (5 to 9)
°BRIX	°Bx	1 °Bx equals 1 gram of sucrose in 100 grams of solution (sugar content)
Transmittance	%	Fraction of incident light at 560nm which is transmitted through the syrup
COLORI	# of drops	Number of drops of sap necessary to change the AuNP solution (1 to 25)
Amino Acid (AA)	$\mu g/mL$	Total amino acid concentration in maple syrup

Table 2.2. – Summary table of the variables relevant to maple syrup.

Figure 2.1 and Figure 2.2 show the 1D and 2D decision boundaries of the model (Eq. (2.1)), respectively. Note that variables not included in each of the graphs were considered to be zero. For any given value of the variables (e.g., 40% transmittance, zero COLORI and AA as shown in Figure 2.1a), the probability of each quality rating is indicated by the line plots. The model then predicts the rating with the highest probability (i.e., VR1). Figure 2.2b shows the 2D extension of the same idea, where for any given transmittance and COLORI value (zero AA), the model prediction is indicated by the rating plane with the highest probability. Though we cannot display the analogous 4D plot with all three variables, the logic follows the same path as the previous two examples.

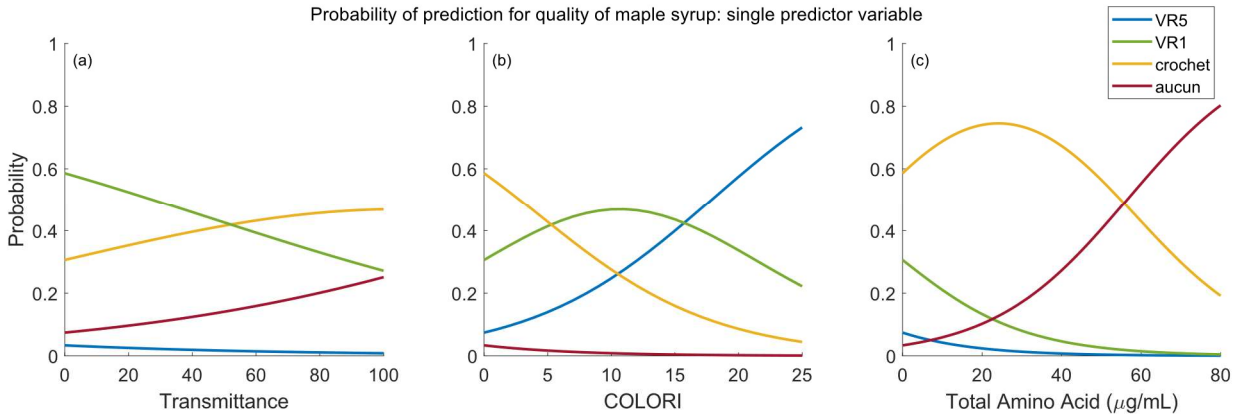

Figure 2.1. – Prediction probability of maple syrup quality for single predictor variables. The quality category with the highest probability for any given value given a) transmittance, b)

COLORI, or c) total amino acid concentration (AA) as predictor. The decision boundary is determined by where the lines of the quality category with the highest probability cross.

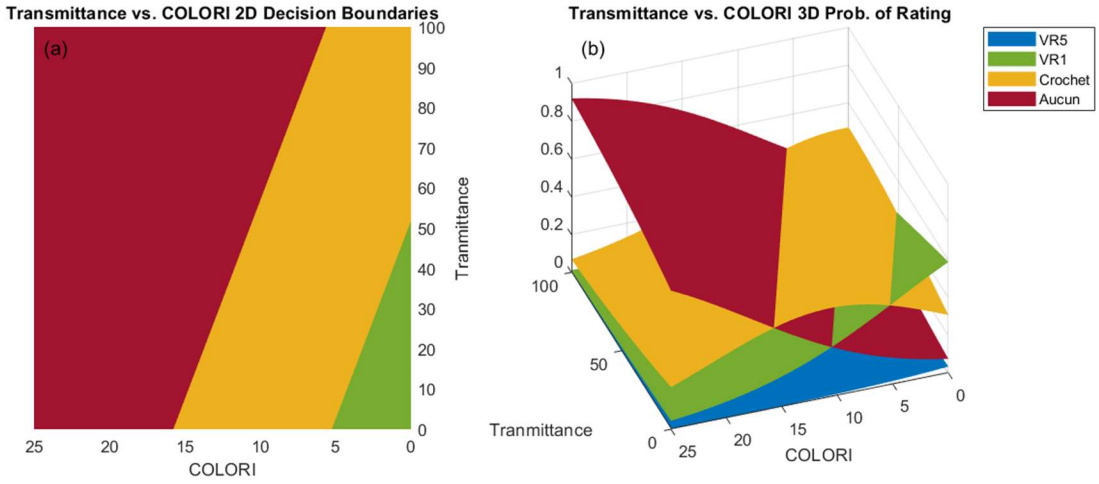

Figure 2.2. – Example of decision boundaries with two variables. a) Decision boundaries for Transmittance vs. COLORI when Totol Amino Acid (AA) is set to 0. Panel a) is a top-down view of the b) probabilily planes for each of the quality ratings. The intersection of the planes marks the decision boundaries where the plane with the highest probability is the model outcome for any give transmittance and COLORI value (when AA is set to zero).

The intraclass correlation (ICC)[8] for within producers was calculated according to,

$$ICC = \frac{var(u_{ID})}{var(u_{ID}) + \left(\frac{\pi^2}{3}\right)} = 0.204.$$

The ICC of 0.204 indicated that about 20% of the variability was explained by between producer differences. In our model, observations were not heavily dependent on being clustered by producer IDs but were not negligible either.

Fixed effects	**Estimate**	**Standard Error (SE)**	**p-value**
$\boldsymbol{\theta_{transmittance}}$	0.0143	0.0139	0.30
$\boldsymbol{\theta_{COLORI}}$	0.141	0.0333	$2.2 \times 10\text{^}(-5)$
$\boldsymbol{\theta_{AA}}$	-0.0560	0.0195	0.0022
Threshold coefficients			

VR5|VR1	-3.36	1.25
VR1|crochet	0.487	1.01
Crochet|aucun	2.52	1.06
Random effects	**Variance**	
u_{ID}	0.844	

Table 2.3. – Ordinal model estimates, standard errors, and p-values.

2.3.2 COLORI is a good continuous predictor of AA concentration

Amino acid concentration plays a vital role in determining the flavour profile of maple syrup. The ordinal model uses both COLORI and AA concentration as a predictor, which can be considered redundant, as the COLORI score is directly linked to AA concentrations. Furthermore, determining the AA concentration is a laborious process. We thus extended our results by using AA concentration as the response variable.

Figure 2.3 shows the observations and the model prediction between pH and COLORI for each of the 16 producers selected in the study. We found the relevant range of pH values for maple syrup to be between 5 and 9. Further, our estimates showed that Equation (2.3) plateaued at 0 and 25 COLORI for most producers. This may be a consequence of our experimental protocol, as explained in the Discussion.

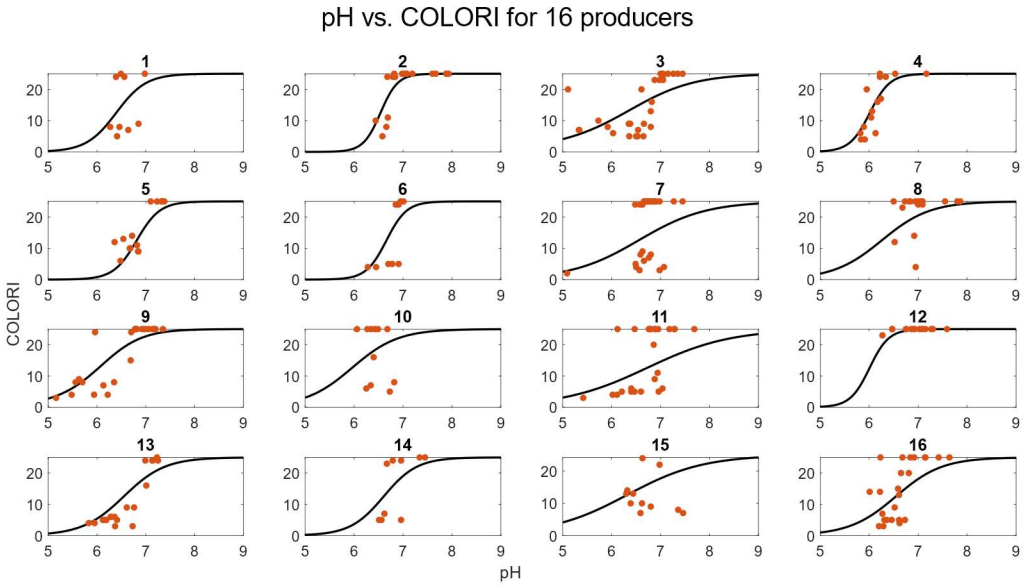

Figure 2.3. – Observations and model predictions of pH vs. COLORI for all 16 producers. The range of pH was chosen between 5 to 9, which is the functional range of maple syrup pH described in the data. Black lines: model predictions; orange circles: data.

Predicted values of COLORI were then used as inputs to predict total AA concentrations (Figure 2.4). We found the model's predictive range to correspond to the whole range of COLORI for most of the producers. However, Producer 3 did not have an output for COLORI values under 3, and Producer 11 did not have any 0 output.

The accuracy of predictions COLORI predictions using pH were found to vary greatly between producers. In some cases, like that of Producer 4 (Figure 2.3), the model predictions and observed measurements were extremely close, sometimes directly overlapping on the prediction curve, whereas we found large discrepancies between predictions and actual data for others, like Producer 11. For Producer 3, we found outliers for pH values of 5, as this producer had an uncharacteristically high COLORI correspondences. These discrepancies may have skewed the model estimation. We observed two main trends in the data: some producers' observations had more slope that was more easily fit to the sigmoid curve of the model (Eq. (2.3)), while sharper drops were observed amongst other producers and these were not as easily predicted by the model.

Despite these two classes of observational trends, the fitted curves for COLORI vs. AA Total (Figure 2.4) were found to correspond well to the data from each producer. Aside from one point of producer 3 which has a very high AA value (~150 $\mu g/mL$), we found no other visual outlier points. The model for all producers reached a maximum AA concentration of 50 $\mu g/mL$, and only producer 5 showed a large deviance between the data points and the model prediction. Interestingly, producers that had deviances between the data points and fitted curves for pH vs. COLORI were not necessarily predicted to have similar results for COLORI vs. AA Total.

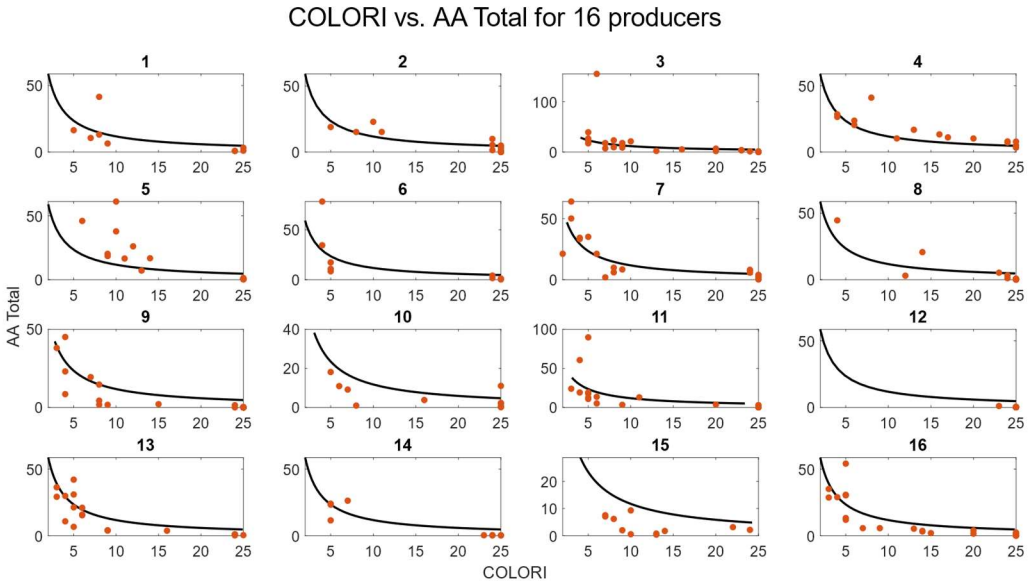

Figure 2.4. – Observations and model predictions of COLORI vs. AA for all 16 producers. The range of COLORI is 0 to 25. The range of Total Amino Acids (AA) concentration ($\mu g/mL$) differs between producers, but for most producers it ranged between 0 to 50.

The population estimation results are shown in Figure 2.5. The fit for pH vs. COLORI (Figure 2.5a) and COLORI vs. AA Total (Figure 2.5b) were found to be visually similar to their corresponding individual graphs (Figure 2.3 and Figure 2.4, respectively). Figure 2.5a shows more clearly that the model predicted values undershot lower pH values and overshot observations at higher pH values. This issue was exacerbated for high COLORI values because the highest COLORI values were assumed to be 25 in the experimental assay and many data points reached this upper limit. We also found bunching at the upper limits of COLORI (Figure 2.5b), though the issue was not as

pronounced due to the nature of the hyperbolic curve having little variability for higher COLORI values. Nonetheless, at these higher COLORI values, the AA concentration was so low that we do not anticipate it affecting the quality determination and thus this does not likely have a meaningful impact on the real-world use of the model.

The probable effective predictive range was found to be between COLORI values of 3 to 20. When predicting AA total from pH directly, we found that the predicted curve did not pass through the majority of points for pH values around 6.5 (Figure 2.5c), which is consistent with previous results (Figure 2.5a) over similar pH range. The few data points that have pH values less than 6 seems to have a heavy effect on the model estimation (long tail).

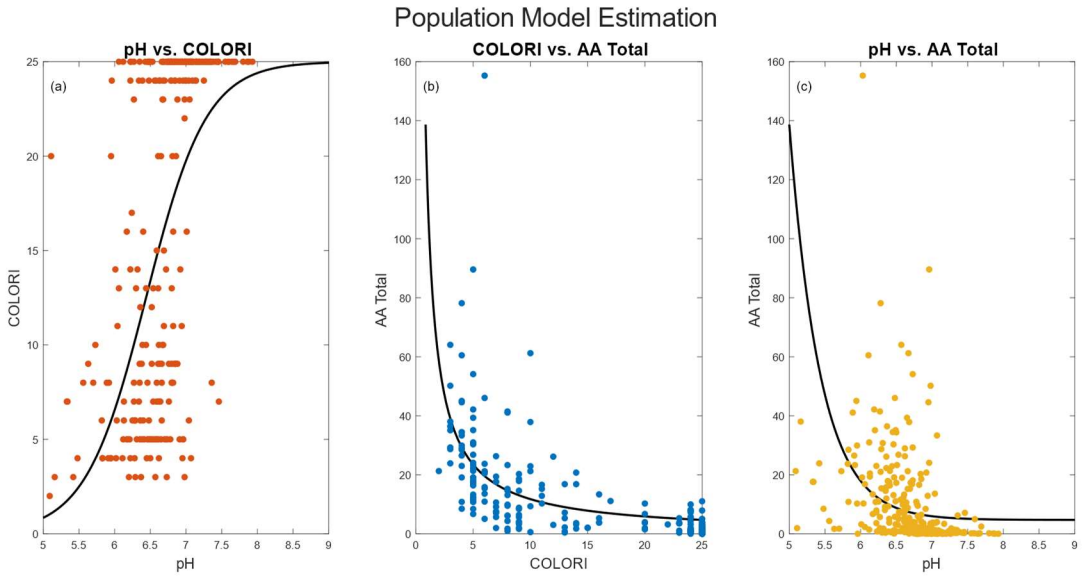

Figure 2.5. – Final population NLMEM model. a.) pH predicting COLORI, b.) COLORI predicting Total Amino Acid (AA) and c.) pH predicting total AA.

2.4 Discussion

Maple syrup production is a laborious and lengthy process that is heavily dependent on the conditions of the yearly sap harvest. Despite the many methods established to facilitate estimates of the quality of maple syrup throughout the season, there is still little understanding of the relationship between these variables and syrup quality. Differences between yearly harvest and even between producers make it difficult to gauge the outcome of any given year's production before the syrup is tasted and judged by professional tasters.

The study of Forest et al.[2] showed the important relationship between the presence of amino acids and the quality of maple syrup, providing COLORI as a new variable. While COLORI was accurate when discerning off-flavour syrup to normal syrup, there was uncertainty in its prediction of syrup quality between the two extremes.

To address this shortcoming, here we investigated these relationships using two mathematical models, one with maple syrup quality as the output and the other to predict total amino acid concentration. Both models are mixed effects model, ordinal linear and nonlinear, respectively, that vary between producers. The goal of both models was to determine the relationship between the variables created from tests that are already used for maple syrup grading (°BRIX, pH, Transmittance), amino acids (COLORI and AA) and syrup quality. We did not use °BRIX as a variable for either model because it is heavily correlated to pH and transmittance, and we found the inclusion of these latter variables to give better results. Further, the addition of °BRIX did risked collinearly and overfitting.

In the ordinal model (Eq. (2.1)) COLORI and AA were found to be statistically significant predictors, and while transmittance was not determined to be a statistically significant input, it lowered the model's AIC score substantially. Thus, the final model used transmittance, COLORI, and AA to predict syrup grading (OK, crochet, VR1, and VR5).

To provide continuous estimates of syrup quality throughout the season, we also established a nonlinear model (Equations (2.2)-(2.4)) that used pH as a proxy for time, and COLORI to predict AA. While the predictive power of pH was less optimal, our model nonetheless had satisfactory results when using COLORI to predict AA concentration. Amino acid concentration is highly indicative of maple syrup quality since it dictates the actual flavor of the syrup (separating maple syrup from bottom-shelved, common corn syrup)[2]. The COLORI test is accurate when differentiating between off-flavours and normal syrup but is unclear about the ratings in between. Our models provide more predictive value to the COLORI test by treating it as a continuous scale rather than the dichotomous scale when used on its own.

One of the challenges faced in model construction is determining outliers (if any). As mentioned in Section 2.3.2, there were a non-negligible number of data points with pH of less than 6. These

points have enough weight to shift the model estimation away from most data points. However, we were hesitant to remove any data points first due to data sparsity. Further, there were enough data points that have pH of less than 6 for them to be classified as outliers without substantial evidence from the maple syrup production process, indicating that pH less than 6 are outliers. For future work, it would be worthwhile to weigh the trade-off of removing points with pH below 6 and model accuracy.

Another limitation of our approach is the lack of validating data. In Quebec, the 2021 harvest produced very different results where the vast majority of syrup contained little to no amino acids. As such, the data was not suitable for validation. The mixed effects nature of our models also made cross-validation tricky to perform. Using some data points from each producer as a training set and the rest as validation would thin our already sparse dataset. However, if we used some producers for model fitting and others for validation, we could not validate the individual results we fitted for the producers in the training group. In the future, we plan to validate our models to more data as they are collected.

Mathematical models provide an accessible bridge for people in different fields to use quantitative measurements in a way that would not have been previously available to them. This work creates a foundation for an improved understanding of the factors that contribute to high quality maple products and facilitates future testing protocols. Importantly, our model will allow maple syrup producers to use the highly accurate COLORI test to better predict the quality of the syrup they are producing, ultimately reducing waste of product, effort, money, and time.

Acknowledgements: ZSY, SF, and JFM were funded by a Natural Science and Engineering Research Council of Canada Alliance grant in collaboration with the Producteurs et productrices acéricoles du Québec and the Consortium de recherche et innovations en bioprocédés industriels au Québec. ZSY and MC were funded by Natural Science and Engineering Research Council of Canada Discovery Grant RGPIN-2018-04546. We thank the 16 producers who participated in our study.

Author contributions: SF, IB, and JMML performed the experimental assays, supervised by JFM. ZSY and MC conceptualized the models. ZSY performed the model fitting, and wrote, and edited the manuscript. MC supervised, wrote, and edited the manuscript.

Declaration of interests: This work was funded in part by the Producteurs et productrices acéricoles du Québec and the Consortium de recherche et innovations en bioprocédés industriels au Québec, both of whom had no impact on the experimental design or results.

References

1 Statistical Overview of the Canadian Maple Industry 2020. 16 (Crops and Horticulture and Division Agriculture and Agri-Food Canada 2021).

2 Forest, S., Théorêt, T., Coutu, J. & Masson, J.-F. A high-throughput plasmonic tongue using an aggregation assay and nonspecific interactions: classification of taste profiles in maple syrup. *Analytical Methods* **12**, 2460-2468, doi:10.1039/C9AY01942A (2020).

3 Quality Control of Our Bulk Maple Syrup. (2021).

4 Saha, K., Agasti, S. S., Kim, C., Li, X. & Rotello, V. M. Gold Nanoparticles in Chemical and Biological Sensing. *Chemical Reviews* **112**, 2739-2779, doi:10.1021/cr2001178 (2012).

5 Bakhtiar, R. Surface Plasmon Resonance Spectroscopy: A Versatile Technique in a Biochemist's Toolbox. *Journal of Chemical Education* **90**, 203-209, doi:10.1021/ed200549g (2013).

6 Aghamollaei, H. *et al.* in *Materials for Biomedical Engineering* (eds Valentina Grumezescu & Alexandru Mihai Grumezescu) 507-554 (Elsevier, 2019).

7 Bozdogan, H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345-370, doi:10.1007/BF02294361 (1987).

8 Hedeker, D., Demirtas, H. & Mermelstein, R. J. A mixed ordinal location scale model for analysis of Ecological Momentary Assessment (EMA) data. *Stat Interface* **2**, 391-401, doi:10.4310/sii.2009.v2.n4.a1 (2009).

Chapter 3 – Impact of estrogen population pharmacokinetics on a QSP model of mammary stem cell differentiation into myoepithelial cells

Justin Le Sauteur-Robitaille[1,2], Zhe Si Yu[1], and Morgan Craig[1,2]*

[1]Département de mathématiques et de statistique, Université de Montréal

[2]Sainte-Justine University Hospital Research Centre

* **Correspondence:** morgan.craig@umontreal.ca; Tel: +1-514-343-7471;
Fax: +1- 514 343-5700.

Abstract: Stem cell differentiation cascades are critical components of healthy tissue maintenance. Dysregulation in these systems can lead to serious diseases, including cancer. Myoepithelial mammary cells are produced from differentiated mammary stem cells in processes regulated, in part, by estrogen signalling and concentrations. This signalling cascade therefore has particular implications for breast cancer plasticity. To quantify and predict the production of mammary myoepithelial cell production by estrogen, we developed a mechanistic, quantitative systems pharmacology (QSP) model that includes the explicit characterization of free and unbound estrogen concentrations in circulation. Linking this model to a previously developed population pharmacokinetics model for ethinyl estradiol, a synthetic form of estrogen included in oral contraceptives, we predicted the effects of estrogen on myoepithelial cell development. Interestingly, pharmacokinetic intraindividual variability alone did not significantly impact on our model's predictions, suggesting that combinations of physiological and pharmacokinetic variability drive heterogeneity in mechanistic QSP models. Our model is one component of an improved understanding of mammary myoepithelial cell production and development, and our results support the call for mechanistically constructed systems models for disease and pharmaceutical modelling.

Keywords: mammary myoepithelial cells, mammary stem cells, quantitative systems pharmacology, estrogen, pharmacokinetic variability, breast cancer

3.1 Introduction

Stem cell (SC) differentiation cascades are critical components of cellular and tissue maintenance [1]. SCs are pluripotent, largely quiescent, populations that repopulate terminally differentiated cells to maintain the latter's basal concentrations. This continuous and progressive differentiation scheme is regulated through cytokine and hormonal signals that control the proliferation and maturation of progenitor cell populations. One of the best understood SC systems is the hematopoietic compartment, which has been extensively studied since the 1960s[2,3]. Hematopoietic stem cells (HSCs) sit at the apex of the blood system and produce some 100 billion blood cells per day[4] in a remarkably regulated hierarchical (or quasi-hierarchical) structure. In the hematopoietic system, and other SC compartments, perturbations to this finely controlled signalling cascade are responsible for a number of diseases, including complex diseases[5]. For example, myelosuppression (lack of myeloid cells) due to cancer chemotherapy is a well-known consequence of cytotoxic anti-cancer treatments[6,7]. Thus, many mathematical models have been conceived to better understand the biology of myeloid cell production from HSCs to predict how patients will be affected by drug therapy[6,8–10].

Given the (intended or unintentional) influence of drugs on signalling networks, it is unsurprising that pharmaceutical scientists have long been interested in using mathematical modelling during preclinical and post-market planning to assess a drug's action at the target site and to forecast potential adverse events[11]. Historically, the pharmaceutical sciences have been concerned with empirical, mixed effects models that can distinguish sources of variability within data, but which are limited in their ability to establish mechanistic relationships at the heart of drug effects[12]. In response, quantitative systems pharmacology (QSP) has emerged as a discipline concentrated on developing physiological systems models that, when linked to pharmacokinetics/pharmacodynamics (PK/PD), provide a holistic understanding of how drugs work[13]. In turn, the broader adoption of QSP has illustrated how cytokines regulate hematopoietic production and how best to schedule exogenous cytokines during

chemotherapy[6,14–16], implement optimal dosing strategies for new cancer immunotherapies[17–21], and individualized drug schedules for children with attention deficit hyperactivity disorder[22], to name a few. There is therefore an increasing interest in applying QSP techniques to a broad set of unanswered physiological questions to provide a quantitative understanding of regulation within SC systems.

Estrogen is a broadly acting, primary female sex hormone that is also implicated in the regulation of mammary cells. Within the mammary differentiation program, estrogen signalling plays a role in the maintenance of the mammary stem cell (MaSC) population, differentiation into the various types of terminally differentiated mammary cells (myoepithelial, ductal, and alveolar)[23,24], and the downstream processes that produce each of these cell types. Mammary epithelium is a bilayer made up of an inner luminal layer and an outer basal myoepithelial layer[25]. A better understanding of the normal mammary cell differentiation program by signalling molecules inducing MaSC differentiation critical to provide a comprehensive view of normal regulation of the system, and interrogate how dysfunctionality caused by diseases such as breast cancer, the most common cancer and second leading cause of death amongst Canadian women[26], disrupts healthy regulatory processes and vice versa. Breast cancers are distinguished by the expression of certain receptors on mammary cells (i.e. estrogen, progesterone, and HER2). Breast cancer heterogeneity is driven by the plasticity of breast cancer cells that transition from epithelial-like (less aggressive, more treatable) to mesenchymal-like (more aggressive, more “stem”-like)[27]. This dedifferentiation (i.e. the change from more terminal to more stem-like cells) complicates therapy, is characteristic of more aggressive cancers, and is regulated by hormones, including estrogen, progesterone, and prolactin[28]. However, stem and progenitor cells carry more proliferation potential than fully mature cells and breast cancer stem-like cells proliferation can be reduced by forcing tumour cells to differentiate through the use of hormones[29–32]. The use of QSP modeling can therefore help uncover the interactions regulating the dynamic evolution of mammary stem cells to provide a fundamental understanding that can subsequently be interpreted experimentally and clinically.

In this work, we leveraged our previous model of hematopoietic production regulated by cytokines[7] to establish a model of mammary stem cell differentiation into myoepithelial cells

through estrogen signalling. Linking the model with the population PKs of ethinyl estradiol allowed us to establish how heterogeneity in PKs affects mammary cell production. Our results suggest that our QSP modelling approach inherently incorporates pharmacokinetic variability in its construction, in line with our previous results in other stem cell differentiation cascades. These results underline the use of QSP models in normal and pathological systems to identify mechanisms of disease and provide guidance for drug administration and scheduling.

3.2 Methods

3.2.1 Mechanistic model of myoepithelial differentiation cascade regulated by estrogen

We constructed a physiological, QSP model describing the differentiation of mammary stem cells transitioning towards mature myoepithelial cells through estrogen regulation (**Figure 3.1**). Our model is based upon the classic G0 model of stem cells, originally developed to explain hematopoietic stem cell division[33] wherein stem cells are largely dormant, dividing to either self-renew or differentiate. In the mammary stem cell system, the emerging picture is that MaSCs largely lay dormant until puberty, with further lineage-restricted expansion during pregnancy[34]. Estrogen acts in a paracrine fashion to regulate differentiation of MaSCs into epithelial cells and certain ER+ luminal cells[23,34,35]. As a better understanding of mammary cell development is still unfolding, we adapted the stem cell models in[7,33] to study myoepithelial production from MaSCs.

Let $Q(t)$ be the concentration of mammary stem cells at time t, $M(t)$ the concentration of myoepithelial cells, $E_F(t)$ the concentration of free estrogen, and $E_b(t)$ the concentration of bound estrogen. Both free and bound estrogen concentrations were explicitly included as we have previously found that cytokine concentrations are far from quasi-equilibrium at homeostasis in the hematopoietic system[7].

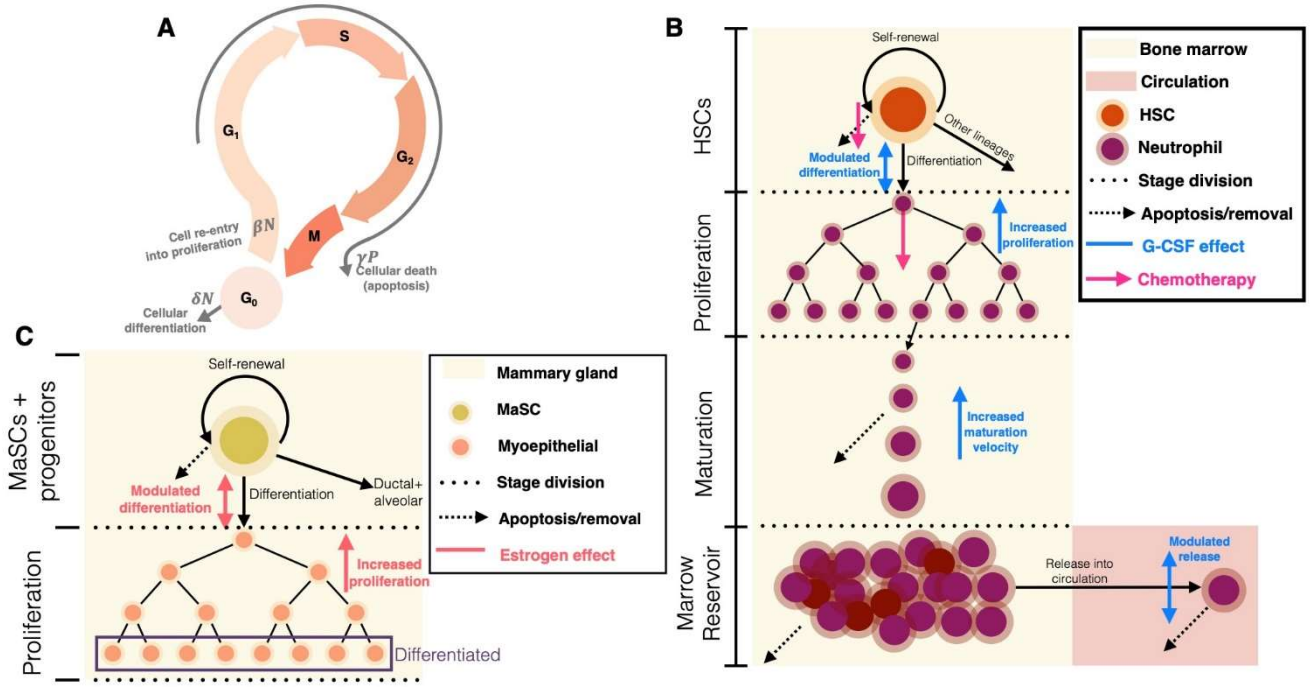

Figure 3.1. – Modelling stem cell systems. A) Schematic representation of the Mackey G0 stem cell model [33]. Labels indicate the passage through the cell cycle (G_0, G_1, S, G_2, and M). Reproduced with permission from [36]. B) Schematic representation of the Craig et al. granulopoiesis model [7], with hematopoietic stem cell model based on the model in A. Hematopoietic stem cells self-renew or differentiate into neutrophils or other lineages. After differentiation into the neutrophil lineage, they undergo a period of exponential expansion before maturing. Mature neutrophils reside in the bone marrow reservoir before egressing out into circulation. These processes are modulated by the cytokine granulocyte colony-stimulating factor (G-CSF), as indicated. Reproduced under Creative Commons license from [12]. C) Schematic of the mammary stem cell model developed here. Based on A and B, MaSCs self-renew to maintain population numbers, and differentiate into alveolar, ductal, or myoepithelial lines. After differentiating, cells undergo a period of exponential expansion, regulated by estrogen concentrations to become myoepithelial cells.

Stem cell compartments maintain their populations through self-renewal via cell division. We therefore considered MaSCs that had divided τ_Q days ago to re-enter the compartment, and modelled the effective amplification resulting from self-renewing mitosis using

$$A_Q(t) = A_Q^* = 2e^{-\gamma_Q \tau_Q}, \quad (3.1)$$

where τ_Q is the time for cell division, γ_Q is the apoptotic rate during mitosis, and the factor of 2 accounts for the formation of two daughter cells from a single MaSC (symmetric cell division). The rate at which MaSCs enter the mitotic phase is given by

$$\beta(Q) = f_Q \frac{\theta_2^{s_2}}{\theta_2^{s_2} + Q^{s_2}}, \quad (3.2)$$

where f_Q represents the maximal rate of self-renewal, θ_2 is the concentration of SCs eliciting 50% of the maximal rate, and s_2 is a Hill coefficient that regulates the slope of the stimulatory effects curve.

Decreases in the concentration of MaSCs arise in two ways: by cells entering mitosis at rate $\beta(Q)$ or through their exit from the compartment to begin differentiation toward mature cells. We modelled the second scenario as dependent on the concentration of free estrogen by

$$\kappa(E_f) = \kappa^* + (\kappa^* - \kappa^{min}) \left[\frac{E_f^{s_1} - (E_f^*)^{s_1}}{E_f^{s_1} + (E_f^*)^{s_1}} \right]. \quad (3.3)$$

Here κ^* represents the homeostatic rate of differentiation, κ^{min} the minimal rate of differentiation (included to ensure that differentiation continues in absence of estrogen, which is known to occur through signalling by other hormones including progesterone and prolactin), E_f^* is the homeostatic concentration of free estrogen, and s_1 is the Hill coefficient regulating the slope of the effect curve. In this model we kept the assumption of Craig et al.[7], where κ^* is found at the middle point between its maximum and minimum values. It is therefore defined as $1/2(\kappa^{max} + \kappa^{min})$ which implies the supremum of $\kappa(E_f)$ is $\kappa^{max} = 2\kappa^* - \kappa^{min}$. We chose to use a Hill function to model both the rate of differentiation in Eq. (3.1) and the rate of proliferation (see Eq. (3.6) below) given that there are limits to the physiological capacity of cells to divide. This

implies that both of these rates will saturate at high concentrations of stimulation. Further, since estrogen is but one of a host of hormones acting in this system, the minimal concentration takes into account that there are overlapping functions for these signalling molecules and that differentiation/proliferation continues in absence of estrogen[34].

We discounted differentiation into alveolar and ductal cells and focused solely on mammary myoepithelial cell production. Thus, after differentiating, MaSCs become myoepithelial progenitor cells and undergo a period of exponential expansion lasting τ_M days. To quantify the concentration of cells produced during this proliferative period, we introduced a second amplification term specific to the myoepithelial population defined by

$$A_M\,(\mathrm{t}) = \exp\left[\int_{t-\tau_M}^{t} \eta_P\left(E_f(s)\right) ds\right]. \tag{3.4}$$

This amplification term accounts for proliferation occurring at rate η_P, regulated by free estrogen E_f, during which cells undergo mitosis for τ_M days. To facilitate computational implementation, we differentiated Eq. (3.4) using Leibniz's rule to obtain the following differential equation

$$\frac{\mathrm{dA_M(t)}}{\mathrm{dt}} = \mathrm{A_M(t)}\left[\eta_\mathrm{P}\left(\mathrm{E}_f(\mathrm{t})\right) - \eta_\mathrm{P}\left(\mathrm{E}_f(\mathrm{t}-\tau_\mathrm{M})\right)\right], \tag{3.5}$$

as in previous work[7]. The rate of mitosis was described by

$$\eta_P\,\left(E_f(\mathrm{t})\right) = \eta_P^* + \left(\eta_P^* - \eta_P^{min}\right)\frac{b_P}{E_f^*}\left(\frac{E_f(t) - E_f^*}{E_f(t) - b_P}\right), \tag{3.6}$$

where η_P^* is the homeostatic rate of proliferation, η_P^{min} is the minimal rate of proliferation, and b_P is the half-maximal concentration of estrogen. Mature myoepithelial cells were modelled to apoptose at rate γ_M.

Taken together, the complete model of the mammary myoepithelial cell differentiation program regulated by estrogen is described by

$$\frac{dQ}{dt} = A_Q(t)\beta\left(Q\left(t-\tau_Q\right)\right)\mathrm{Q}\left(t-\tau_Q\right) - \left(\kappa\left(E_f(t)\right) + \beta\left(Q(t)\right)\right)\mathrm{Q}(t), \quad (3.7)$$

$$\frac{dM}{dt} = A_M(t)\kappa\left(E_f(t-\tau_M)\right)\mathrm{Q}(t-\tau_M) - \gamma_M \mathrm{M}(t). \quad (3.8)$$

3.2.2 Population pharmacokinetic and QSP model of estrogen

Pharmacokinetic (PK) models describe drug concentrations over time, and are generally built on "compartments" (sections of the body) in which the drug readily distributes and can therefore be summarized as having a single concentration at any given time. Population PK (PopPK) models aim to characterize variations in PK parameters within a given population using mixed-effect models, which are statistical models that contain both fixed (mean) and random (interindividual) effects. From data, a structural model (the compartmental PK model) is established by finding the best estimates for the population parameter value (fixed effect) with reference to the variation within the sample population (random) effects. Parameters are then represented as an exponentially distributed random variable by a singular fixed effect (usually referred to as θ) with noise parameters denoting the variability between individual subjects and the population value. These random effects are generally assumed to be normal or log-normal with zero mean and variance Ω. Parameters may also have covariate effects – additional variables that account for some of the variation of said parameter, and are particularly relevant in pharmaceutical models as drug kinetics are known to be determined by a variety of biological factors often modulated by specific patient characteristics e.g., age, body weight, habits, etc.

Thus, to account for the effects of interindividual variability (IIV) in estrogen pharmacokinetics, we leveraged a previous PopPK model[37] that characterized the pharmacokinetics of ethinyl estradiol (EE), an estrogen medication which, alongside drospirenone, is one of the active compounds in combined oral contraceptives. In the Reif et al. study[38], n = 1109 healthy young

women received EE 20 µg and drospirenone 3 mg as a combined oral contraceptive. Patients were randomized to receive three different cyclical regimens

1. Flexible (n = 671): one tablet per day for a flexible number of cycles (between 3 to 13 cycles). The minimum duration of active treatment was 24 days ('mandatory phase'). Afterwards, the cycle could continue up to 120 days or until the subject experienced 3 consecutive days of breakthrough bleeding or spotting ('flexible phase');
2. Conventional (n = 224): subjects received one dosage per day for 13 cycles. Each cycle comprised 24 days of active hormonal intake followed with 4 days of placebo tablets;
3. Fixed (n = 214): subjects received one dose per day for 3 cycles; each cycle comprised 120 days of active hormone intake followed by 4 days of placebo tablet;

during which four blood samples were collected from all subjects in the study, two during week 3, and two during week 27. EE concentrations were determined by immunoassay with a lower limit of quantification of 5.0 pg/ml. Data on age, BMI, body weight (BW), and cigarette and alcohol consumption were collected at both weeks, and data was then fit by means of nonlinear mixed effects modeling using the NONMEM software package to construct the compartmental PK model.

In the analysis by Reif et al., the pharmacokinetics of estradiol were characterized by a three-compartment population PK model with linear elimination[37] given by

$$\begin{aligned}
\frac{dA_1}{dt} &= -\left(\frac{CL}{V_1} + \frac{Q_2}{V_1} + \frac{Q_3}{V_1}\right) A_1 + \frac{Q_2}{V_2} A_2 + \frac{Q_3}{V_3} A_3, \\
\frac{dA_2}{dt} &= -\frac{Q_2}{V_2} A_2 + \frac{Q_2}{V_1} A_1, \\
\frac{dA_3}{dt} &= -\frac{Q_3}{V_3} A_3 + \frac{Q_3}{V_1} A_1,
\end{aligned}$$

where A_1, A_2, A_3 represent the amount of EE in the central (1) and two peripheral (2, 3) compartments. CL denotes clearance (in units L/h), Q_2 the intercompartmental clearance between compartments 1 and 2 (in units L/h), Q_3 the intercompartmental clearance between compartments 1 and 3 (in units L/h), V_1 the central volume of distribution (in units L), V_2 the

volume of distribution in compartment 2 (in units L), and V_3 is the volume of distribution in compartment 3 (in units L). Note that intercompartmental transit can also be represented by rate constants $k_{ij} = \frac{Q_{ij}}{V_i}$, where Q_{ij} is the intercompartmental clearance between compartments *i* and *j* and V_i is the volume of distribution in compartment *i.* Reif et al. also found covariate relationships on the bioavailability F defined by

$$F = F_{week3}(1 + F_{week27}OCA),$$

where $OCA = \begin{cases} 0 & if\ week\ 3 \\ 1 & else \end{cases}$,

and

$$CL = TVCL \exp(ETA_{CL})\,(CO2)(CO1)$$

for clearance, where $CO1 = 1 + CL_{AGE}(\log(AGE) - \log(24))$ and $CO2 = (1 + CL_{BW}(BW - 62))$ are covariates of age and body weight (kg) on clearance. Parameter values of the PopPK model are provided in **Table 3.1**.

Parameter (units)	Interpretation	Estimate	%RSE
Fixed Effects			
		ϑ	
$TVCL/F$ (L/h)	Oral clearance	25.3	1.24
V_1/F (L)	Apparent volume of central compartment	23.9	13.6
V_2/F (L)	Apparent volume of second compartment	1,330	3.62
V_3/F (L)	Apparent volume of third compartment	23.9	-
Q_2/F (L/h)	Intercompartmental clearance to Compartment 2	52.9	7.01
Q_3/F (L/h)	Intercompartmental clearance to Compartment 3	8.49	34.3

k_a (1/h)	Absorption rate constant	0.295	6.98
F_{week3}	Relative bioavailability in week 3	1	-
DF_{week27} (%)	Difference in relative bioavailability in week 27 to week 3	8.15	11.0
$ALAG$ (h)	Lag time	0.353	2.78
Fixed covariate effects			
CL_{age} (%/ln(year))	Influence of age on clearance	20.8	29.1
CL_{BW} (%/kg)	Influence of body weight on clearance	0.591	20.1
Interindividual Variability			
		%CV	
IIV_{CL}	IIV of clearance	33.4	2.65
Residual Error			
Proportional error	Proportional residual error	24.4	1.38

Table 3.1. – Population pharmacokinetic parameters from Reif et al.[38]*.*

A downside of compartmental PK models is their lack of connection to known physiology. To improve physiological realism, we integrated the Reif et al. PopPK model[38] into our previous mechanistic PK framework to further track the concentrations of free and bound estrogen in the plasma[7]. This integrated model is given by

$$\frac{dE_f}{dt} = E_{prod} - k_{ren}\mathrm{E}_f(t) - k_b\big((Q+M)V - E_b(t)\big)E_f(t)^{Pow} + k_u E_b(t) - (k_{12}+k_{13})E_f + k_{21}E_2 + k_{31}E_3, \tag{3.9}$$

$$\frac{dE_b}{dt} = -k_{int}E_b(t) + k_b\big((Q+M)V - E_b(t)\big)E_f(t)^{Pow} - k_u E_b(t), \tag{3.10}$$

$$\frac{dE_2}{dt} = k_{12}E_f - k_{21}E_2, \tag{3.11}$$

$$\frac{dE_3}{dt} = k_{13}E_f - k_{31}E_3, \tag{3.12}$$

where E_f and E_b are free and bound estrogen concentrations, E_{prod} is the rate of endogenous estrogen production, k_b and k_u are the respective binding and unbinding rates, k_{int} is the internalization rate of bound cytokine, k_{ren} is the elimination rate, pow is a stoichiometric constant relating the number of estrogen molecules per receptor, V is a scaling factor given by

$$V = \hat{p}E_{MW}K10^n,$$

with $\hat{p}$ a constant relating the stoichiometry between estrogen and its receptor, K the number of estrogen receptors on a cell's surface, and 10^n is a factor correcting for cellular units. In the equation above, E_{MW} represents the molecular weight of estrogen which was calculated by dividing its molar mass (MM) by Avogadro's number. The remaining terms in Eqs. (3.3), (3.5), and (3.6) correspond to a reconfiguration of the PopPK model above to be expressed in concentrations in the central (E_f), second (E_2), and third (E_3) compartments, with $k_{12} = Q_2/V_1$, $k_{21} = Q_2/V_2$, $k_{13} = Q_3/V_1$, $k_{31} = Q_3/V_3$, and $k_{ren} = CL/V_1$, with CL representing linear (renal) clearance.

3.2.3 Parameter estimation

Since our model was constructed from physiological mechanisms, a majority of the QSP model's parameters were estimated directly from the literature. In absence of direct information about mammary stem cells, we made an assumption of a parallel between MaSCs and HSCs and leveraged our previous work in Craig et al.[7]. Estrogen PopPK parameters were obtained directly from the estimates in Supplementary Table 1 in Reif et al.[38] (**Table 3.1**). Any remaining parameters were calculated to ensure homeostasis in our model as follows. We calculate$d\ \theta_2$ from the equation for the self-renewal of mammary stem cells $\beta\big(Q(t)\big)$ by

$$\theta_2 = \sqrt[S_2]{\frac{Q^{*S_2}\beta^*}{f_Q - \beta^*}} = \frac{Q^*\beta^{*\frac{1}{S_2}}}{\left(f_Q - \beta^*\right)^{\frac{1}{S_2}}}. \tag{3.13}$$

Setting Q^* to its estimated homeostatic value (**Table 3.2**), we calculated the homeostatic rate of differentiation as

$$\kappa^* = \beta^*\left(A_Q^* - 1\right). \tag{3.14}$$

Similarly, the homeostatic amplification factor for myoepithelial cells was calculated from Eq. (3.8*J*) by

$$A_M^* = \frac{M^*\gamma_M}{\kappa^* Q^*}. \tag{3.15}$$

Here, A_M^* quantifies the effective amplification of progenitor cells after differentiation given by the ratio of the rate cells leave the progenitor compartment (through $M^*\gamma_M$) to the rate cells enter (i.e. $\kappa^* Q^*$) at homeostasis. Assuming there are 10 divisions between stem cells and terminally differentiated myoepithelial cells[39], the number of proliferating cells is calculated explicitly as $\kappa^* Q^* 2^{10}$, thus

$$\eta_P{}^* = \frac{(A_M^* - 1)}{(10)(2^{10})}. \tag{3.16}$$

The homeostatic concentration of bound estrogen (E_b^*) was calculated directly from Eq. (3.10) using

$$E_b^* = \frac{(k_{12} + k_{13})E_f^* - k_{21}E_2^* - k_{31}E_3^* - E_{prod} + k_{ren}E_f^* + k_b(Q^* + M^*)V\left(E_f^*\right)^{Pow}}{k_u + k_b\left(E_f^*\right)^{Pow}}, \tag{3.17}$$

and the rate of estrogen production from Eq. (3.9)**Error! Reference source not found.**

$$E_{prod} = k_{ren}E_f^* + k_b\left((Q^* + M^*)V - E_b^*\right)\left(E_f^*\right)^{Pow} - k_u E_b^* + (k_{12} + k_{13})E_f^* - k_{21}E_2^* - k_{31}E_3^*. \tag{3.18}$$

Lastly, the homeostatic concentration of estrogen in the second and third compartments, E_2^* and E_3^*, were estimated by setting Eqs. (3.11) and (3.12) to 0, given that E_f^* has been previously estimated [40].

$$E_2^* = \frac{\mathrm{k_{12}E_f^*}}{\mathrm{k_{21}}}, \tag{3.19}$$

$$E_3^* = \frac{k_{13}E_f^*}{k_{31}}. \tag{3.20}$$

3.2.4 Model simulations

Model simulations were carried out using the *ddesd* function in Matlab 2020a. Unless otherwise noted, initial values of all variables were set to their homeostatic values, indicated by superscript *.

3.2.5 Sensitivity analysis

To investigate the effects of variability in parameters on model predictions, we used partial rank correlation coefficient (PRCC) analysis to assess uncertainty and sensitivity on the multi-dimensional parameter space[41]. For each investigated parameter, we generated a uniform distribution around the mean (or fixed) parameter value. This set of parameter values was then included to create a Latin hypercube sampling (LHS) scheme. We next simulated and recorded the relevant outputs of the model (specifically $min(Q(t))$ and $max(M(t))$) using the samples in the LHS scheme. Finally, we calculated the PRCC to establish correlations between parameters and the model outputs. These correlations were also assessed for statistical significance, given an α-level of 0.05. For this analysis, we adapted the approach and code from Marino et al.[42].

3.3 Results

1.1. Estimated and calculated parameter values of myoepithelial cell production by mammary stem cells regulated by estrogen

Physiological and mechanistic PK parameter values were estimated from literature sources or through calculations to ensure homeostasis in the model (see Methods). **Table 3.2** and **Table 3.3** summarize all parameter values in these model components.

Parameter	Value	Units	Definition	Reference

τ_Q	24	hours	Length of MaSC cell cycle	[43]
τ_M	10	days	Duration of proliferative phase	[44]
γ_Q	0.1	1/day	MaSC apoptotic rate during mitosis	[45]
A_Q^*	1.8097	--	Amplification factor	[7]
β^*	0.043	1/day	Stem cell rate of self-renewal	[7]
f_Q	8	1/day	Maximal stem cell renewal	[7]
θ_2	0.0809	10^6 cells/kg	Self-renewal half-effect concentration of MaSCs	Calculations from [7]
s_2	2	--	MaSC self-renewal Hill coefficient	[7]
s_1	1.5	--	MaSC differentiation Hill coefficient	[7]
κ^*	0.0348	1/day	Homeostatic rate of differentiation	Calculated (Eq. (14))
κ^{min}	0.0174	1/day	Minimal rate of differentiation	[46]
Q^*	1.1	10^6 cells/kg	Homeostatic concentration of MaSC	[7]
A_M^*	27697	--	Myoepithelial proliferation amplification factor	Calculated (Eq. (15))
$\eta_P{}^*$	2.7047	1/day	Homeostatic rate of proliferation of myoepithelial progenitors	Calculated (Eq. (16))

$\eta_P{}^{min}$	1.3523	1/day	Minimal rate of proliferation of myoepithelial progenitors	[47]
b_p	0.022868	ng/ml	Half-maximal concentration of proliferation of myoepithelial progenitors	[7]
γ_M	0.113	1/day	Apoptotic rate of myoepithelial cells	[48]
K	4973	--	Number of estrogen receptors on MaSC/myoepithelial cell surface	[49]
V	0.0283	(ng/ml)/(10^9cell/kg)	Conversion factor	Calculated
M^*	9.387	10^9 cells/kg	Homeostatic concentration of myoepithelial cells	[50]

Table 3.2. – Physiological model parameter values.

Parameter	**Value**	**Units**	**Definition**	**Reference**
E_{prod}	3.2776	1/day	Rate of endogenous estrogen production	Calculated (Eq. (18))
k_u	8.64	1/day	Rate of estrogen unbinding	[51]
k_b	23.53536	ng/ml/day	Rate of estrogen binding	[52]
k_{int}	43.2	1/day	Rate of estrogen internalization	[52]
k_{ren}	35.9335	1/day	Rate of renal clearance	Calculated
k_{12}	53.1213	1/day	Transit rate from 1st compartment to 2nd	[38]

k_{21}	0.9546	1/day	Transit rate from 2nd compartment to 1st	[38]
k_{13}	8.5255	1/day	Transit rate from 1st compartment to 3rd	[38]
k_{31}	8.5255	1/day	Transit rate from 3rd compartment to 1st	[38]
E_f^*	0.080	ng/ml	Homeostatic concentration of free estrogen	[40]
E_b^*	0.0093	ng/ml	Homeostatic concentration of bound estrogen	Calculated (Eq. (17))
E_2^*	4.4519	ng/ml	Homeostatic estrogen concentration in 2nd compartment	Calculated (Eq. (19))
E_3^*	0.08	ng/ml	Homeostatic estrogen concentration in 3rd compartment	Calculated (Eq. (20))

Table 3.3. – Additional estrogen pharmacokinetic parameter values.

3.3.1 Effects of repeated oral estrogen administration on myoepithelial cell production for an average individual

To provide a rationale for the use of estrogen as a means to induce myoepithelial differentiation of mammary stem cells, we first investigated the homeostatic regulation of the differentiation cascade. For this, we discounted interindividual variability, set all PK parameters to their fixed (θ) values (see **Methods**), and administered a single dose of 20μg of ethinyl estradiol (as in [38]). This single dose caused a small rise in myoepithelial cell counts, peaking around 10 days after administration, despite estrogen concentrations in most compartments returning to the homeostatic concentrations within 1-2 days (**Figure 3.2**).

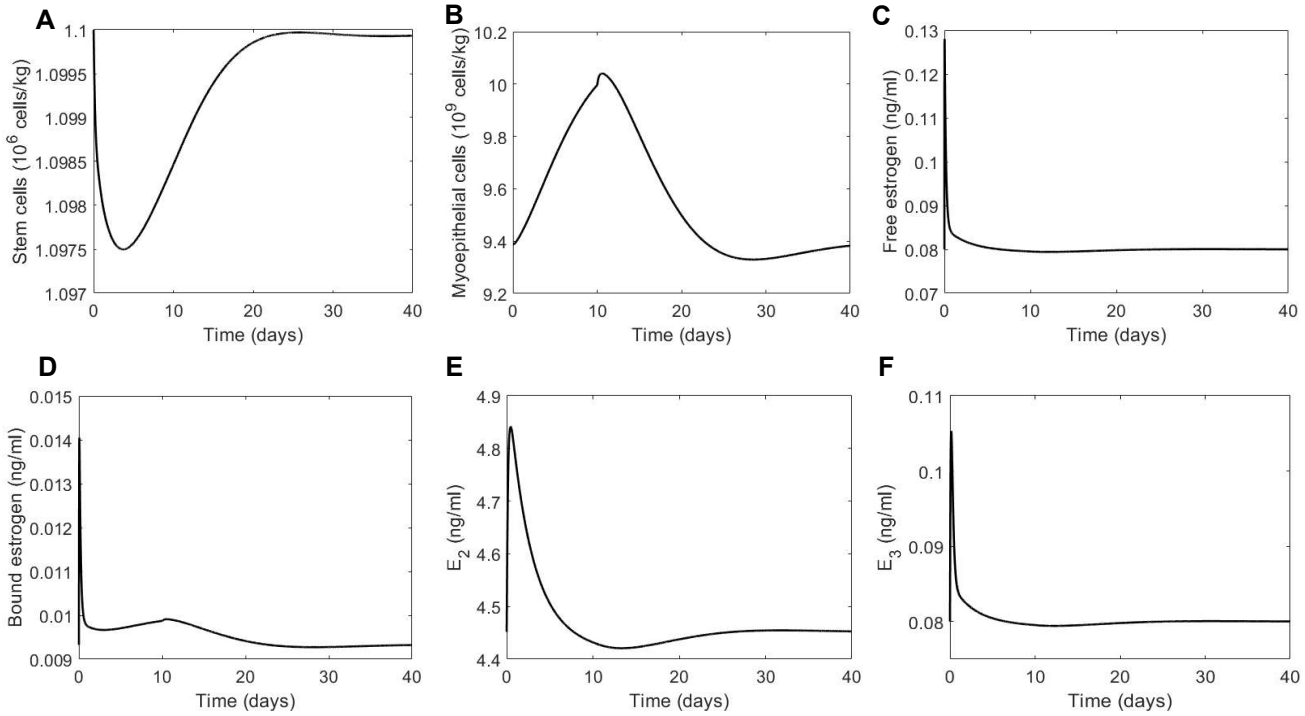

Figure 3.2. – Single dose of oral estrogen weakly induces mammary stem cell differentiation into myoepithelial cells. Effects of a single 20μg dose of ethinyl estradiol on A) mammary stem cells, B) myoepithelial cells, C) free estrogen, D) bound estrogen, E) estrogen concentrations in the second compartment, and F) estrogen concentrations in the third compartment in an average individual.

These results suggest that a single 20µg dose of estrogen is not enough to sufficiently stimulate myoepithelial cell production, as myoepithelial cell concentrations peaked at only 1.06-times their homeostatic values. Therefore, we next interrogated the effects of exogenous, oral estrogen administration by simulating daily doses of 20 µg of ethinyl estradiol (as in [38]). As expected, the concentrations of free estrogen in the central and peripheral compartments rapidly attained steady-state during the dosing period (**Figure 3.3C-F**). The rise in free estrogen concentrations induced the continued differentiation of MaSC (**Figure 3.3A**), during which mammary myoepithelial cell concentrations nearly doubled, again peaking around 10 days after administration.

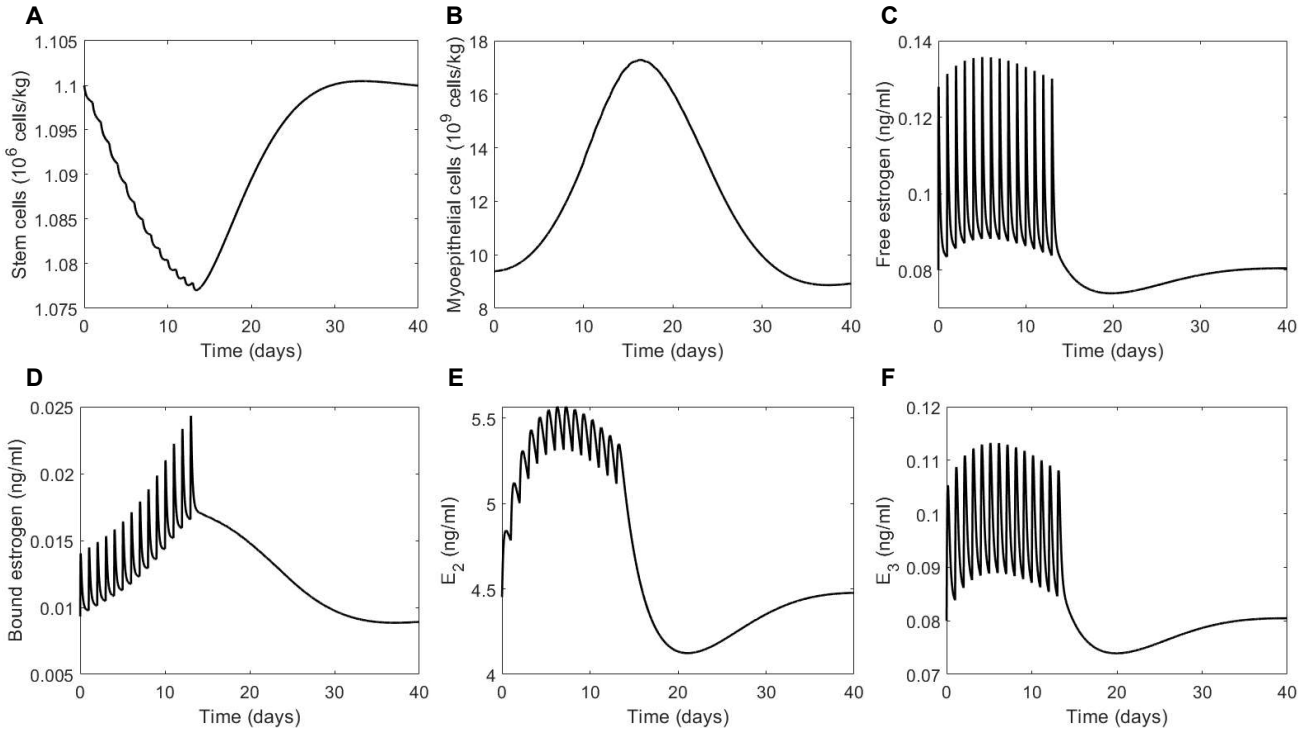

Figure 3.3. – Multiple administrations of oral estrogen stimulate mammary myoepithelial production. Effects of a 14 repeated 20μg doses of ethinyl estradiol on A) mammary stem cells, B) myoepithelial cells, C) free estrogen, D) bound estrogen, E) estrogen concentrations in the second compartment, and F) estrogen concentrations in the third compartment in an average individual.

3.3.2 Effects of pharmacokinetic variability on mammary stem cell differentiation

Up to this point, we have explored the effects of estrogen stimulation on mammary stem cell differentiation for a single average individual. However, given the variability estimated in our adapted PopPK model[38], we next sought to understand how population heterogeneity affects a QSP model of MaSC differentiation. We have previously shown that interindividual variability has little to no effect on pharmacokinetic and pharmacodynamic variability in a similar model of hematopoietic differentiation, likely due to the inherent heterogeneity captured in mechanistic QSP models[53]. Therefore, to study the effects of interindividual variability in the context of mammary stem cell differentiation, we generated 300 virtual patients by sampling from a normal

distribution of body weights (BWs) centered around the mean BW of 63kg reported in Reif et al.[38], with a standard deviation of 8.6kg informed by two recent surveys carried out by the NCD Risk Factor Collaboration[54,55]. We chose to generate virtual patients using body weight as it was found to be a covariate of clearance in the original PopPK analysis[38]. The effects of IIV on clearance (also estimated by Reif et al.) were also investigated and found to be consistent with the results of the body weight analysis. Normality in the generated body weights was confirmed using Shapiro-Wilk and Shapiro-Francia normality tests (p-value of 0.5565 at the 5% significance level, see **Figure 3.4**)[56].

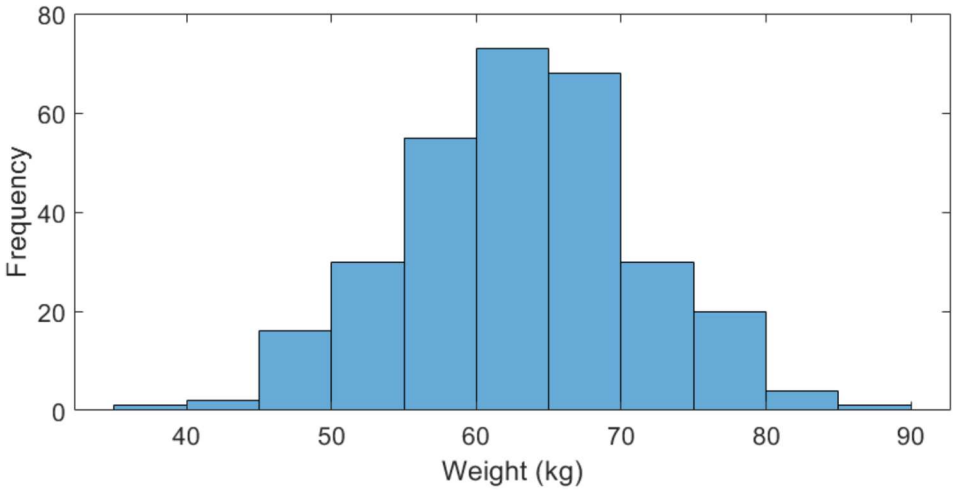

Figure 3.4. – Distribution of body weights in generated virtual population. Virtual patients were generated by sampling from a normal distribution with mean body weight as reported in Reif et al.[38] and standard deviation informed by population sampling[54,55].

We next again simulated a single oral dose for each virtual patient to quantify the degree to which pharmacokinetic variability affected our model's predictions. Across both cellular populations and all estrogen compartments, we observed no effect of pharmacokinetic variability on our predictions (**Figure 3.5**).

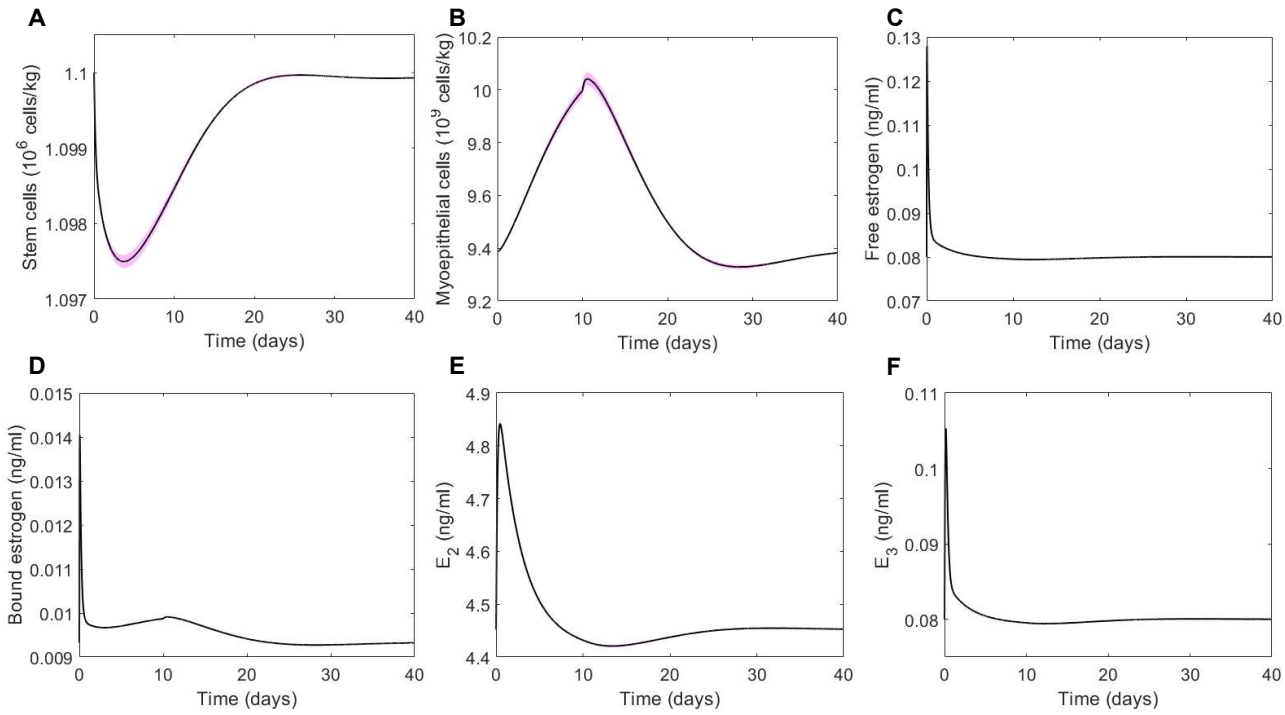

Figure 3.5. – Pharmacokinetic variability has little to no impact on QSP model predictions in cohort of 300 virtual patients after single administration of estrogen. Effects of a single 20μg dose of ethinyl estradiol on A) mammary stem cells, B) myoepithelial cells, C) free estrogen, D) bound estrogen, E) estrogen concentrations in the second compartment, and F) estrogen concentrations in the third compartment in a virtual cohort of 300 patients. Virtual patients were generated by sampling from a distribution of bodyweight, a covariate for clearance in the PopPK model[37] (see **Methods**). Black solid lines: mean model prediction in cohort of 300 virtual patients; shaded pink regions: range of responses in virtual patient cohort.

These results were further recapitulated when we considered repeated daily doses in the same virtual population. This additional heterogeneity had very little discernable impact on either the MaSCs or the myoepithelial cells (**Figure 3.6A** and **Figure 3.6B**), and virtually no impact in any of the estrogen compartments (**Figure 3.6C-F**). This is notable given that the bound estrogen concentrations drive all the pharmacodynamic effects in our model and, despite observing no IIV impact on bound estrogen concentrations, there are nonetheless observable (but small) amounts of heterogeneity in the MaSC and myoepithelial populations after repeated administrations.

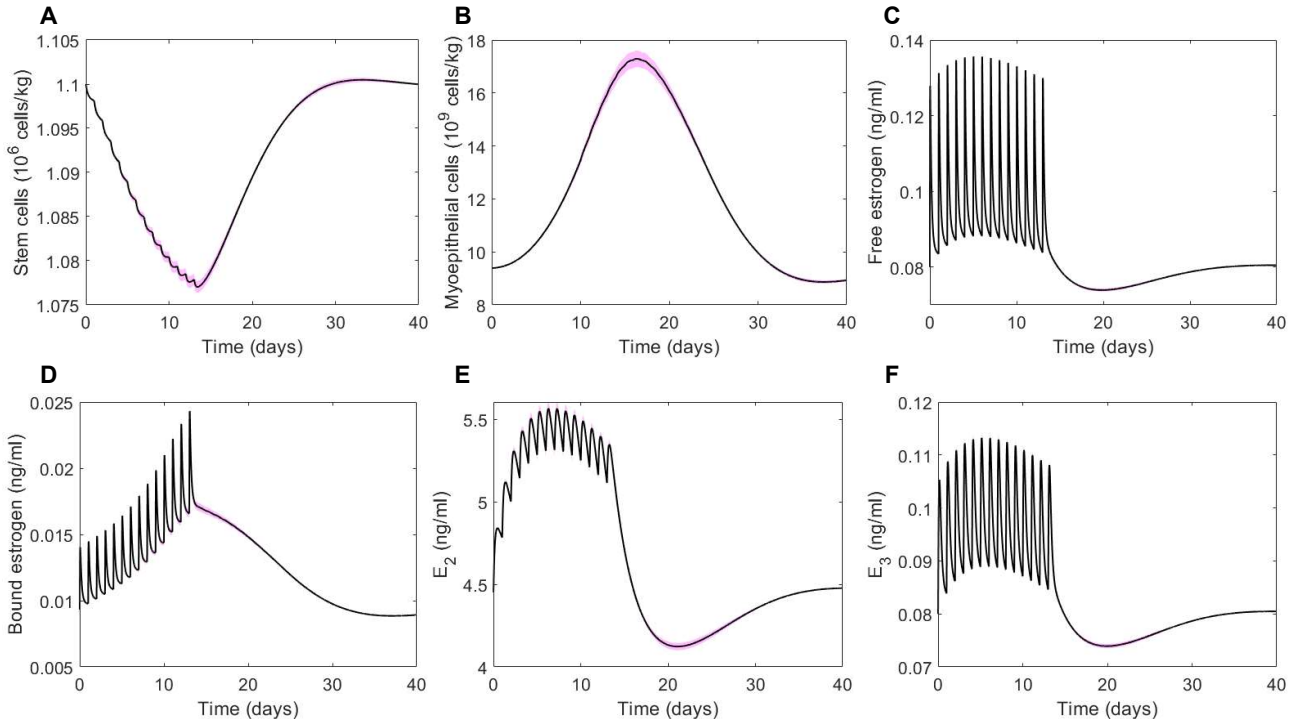

Figure 3.6. – Pharmacokinetic variability has little to no impact on QSP model predictions in cohort of 300 virtual patients after repeated daily doses of estrogen. Effects of 14 repeated 20μg doses of ethinyl estradiol on A) mammary stem cells, B) myoepithelial cells, C) free estrogen, D) bound estrogen, E) estrogen concentrations in the second compartment, and F) estrogen concentrations in the third compartment in a virtual cohort of 300 patients. Virtual patients were generated by sampling from a distribution of bodyweight, a covariate for clearance in the PopPK model[37] (see Methods). Black solid lines: mean model prediction in cohort of 300 virtual patients; shaded pink regions: range of responses in virtual patient cohort.

3.3.3 Sensitivity analysis

Sensitivity of model predictions to parameter variability was assessed using partial rank correlation coefficients (PRCC), as described in the Methods. To perform the PRCC analysis, we included several of our model parameters to assess their effect on the stem cell and myoepithelial cell populations after the introduction of a single dose of estradiol. Physiological parameters included were the time delays (τ_Q and τ_M) and death rates (γ_Q and γ_M) for both cell populations, in addition to β^* and b_p to assess effects related to mitosis and proliferation.

Further, since our virtual population was generated from body weights that impact on the rate of renal clearance, we also investigated the effects of variability in k_{ren} to assess the degree to which our model is sensitive to this important parameter.

We generated our LHS scheme by creating uniform distributions around mean estimated parameter values (Table 3.1-**6**), with maximum and minimum values for the distributions set to be the mean value $\pm$25%. With a sample size of 100, we ran the PRCC analysis for the effect on the minimum stem cell concentration ($min\ Q(t)$) and the maximum myoepithelial cell concentration ($max\ M(t)$).

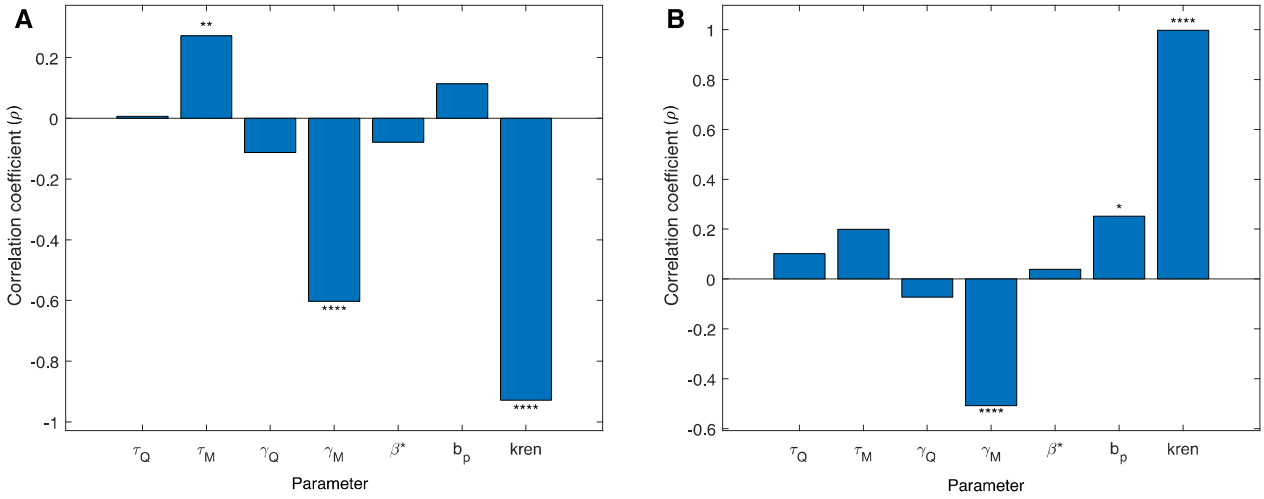

Figure 3.7. – Individual correlation coefficient for each parameter with its p-value significance from the PRCC analysis. The sensitivity of model predictions to parameter changes was assessed using partial rank correlation coefficient analysis and Latin hypercube sampling. Correlation coefficients between each parameter and A) the minimum stem cell concentration and B) the maximum myoepithelial cell concentration after a single dose of estradiol. * indicates a p-value smaller or equal to 0.05, ** p-value smaller or equal to 0.01 but greater than 0.001, and **** a p-value smaller or equal to 0.0001 (calculated at a level of significance of $\alpha = 0.05$).

As shown in **Figure 3.7**, renal clearance was found to be the most sensitive parameter in our model, strongly impacting on both minimal MaSC and maximal mature cell concentrations. This analysis therefore further supports the use of body weight to generate the virtual cohort, as variations in bodyweight directly affect k_{ren}.

3.4 Discussion

Mammary stem cell plasticity is a contributing factor to the development of aggressive breast cancer subtypes. Given the frequency of estrogen receptor positive breast cancer, a quantitative understanding of the regulation of the mammary cell differentiation program controlled by estrogen is needed to provide insights into how this regulation can be therapeutically targeted. In this work, we created a QSP model to study both mammary stem cell and myoepithelial cell populations and their regulation by estrogen signalling. Our model takes into account physiologically relevant pharmacokinetics by explicitly accounting for both unbound and bound estrogen concentrations in addition to empirically determined population pharmacokinetics.

By investigating the effects of single and repeated oral ethinyl estradiol doses in an average patient, we demonstrated the small effect of a single dose on both cell populations included in our model. However, in the two-week regimen of daily doses, we observed a doubling in the concentration of mammary myoepithelial cells, suggesting that repeated administrations are required to sustainably induce MaSC differentiation into this cell type. Since the process of dedifferentiation relies on hormones such as estrogen and could therefore be affected by an increase in estrogen concentrations, it has been proposed that forcing cancer cells to differentiate can ultimately reduce cancer aggressivity and proliferation capacities. Thus, our results support continued work investigating the regulation of the mammary stem cell niche.
Importantly, we also observed that PK variability in clearance provided little to no impact on our QSP model by generating a population of 300 virtual patients based on the integrated population pharmacokinetic model. In this cohort, both the single dose and repeated administrations over 14 days did not show meaningful variation in cell population and even less variation in the estrogen compartments. This suggests that sources of empirically determined PK variability may be accounted for within QSP models constructed from physiological first principles. Indeed, PK variability was shown to insignificantly affect our model predictions. As our PRCC analysis showed, the effect of renal clearance was found to be most significantly correlated with changes in model output. However, other physiological parameters were also found to be sensitive, and future studies should explore combinations of physiological parameters to assess their effects on driving

heterogeneity in our model. Interestingly, we have previously shown similar results when integrating PopPK/PD estimates into our similar model of neutrophil production[53]. There, estimated pharmacodynamic variability had a much larger impact on heterogeneous model outcomes than interindividual variability estimated through PopPK analyses, as in the present study. It should also be noted that we did not explore the effects of age or interoccasion variability, which was found in Reif et al.[38] to be related to heterogeneity in clearance.

As noted, mammary stem cells can additionally differentiate into alveolar and ductal cell types, and the MaSC differentiation program is regulated by other hormones including progesterone and prolactin. Future studies will incorporate these considerations to increase the physiological realism and applicability of our system, and focus on experimental validation. Nonetheless, our approach establishes a theoretical framework with which we can explore how dysregulated mammary stem cell signalling can lead to malignancy and, as such, is an important element of preclinical and translational studies as part of a precision medicine in breast cancer.

Acknowledgments: JLSR was supported by ISM and CAMBAM graduate fellowships. All authors were supported by NSERC Discovery Grant RGPIN-2018-04546 (MC).

Conflict of interest: The authors have no conflicts of interest to declare.

References

1. Sieck, G. C. Physiology in perspective: stem cells and regenerative physiology. *Physiology* **33**, 14–15 (2018).

2. Mackey, M. C. Cell kinetic status of hematopoietic stem cells. *Cell Proliferation* **34**, 71–83 (2001).

3. Till, J. & McCullogh, E. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiation Research* **14**, 213–222 (1961).

4. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).

5. Krieger, M. S. *et al.* A blueprint for identifying phenotypes and drug targets in complex disorders with empirical dynamics. *Patterns* **1**, (2020).

6. Craig, M. *et al.* Neutrophil dynamics during concurrent chemotherapy and G-CSF administration: mathematical modelling guides dose optimisation to minimise neutropenia. *Journal of Theoretical Biology* **385**, 77–89 (2015).

7. Craig, M., Humphries, A. R. & Mackey, M. C. A mathematical model of granulopoiesis incorporating the negative feedback dynamics and kinetics of G-CSF/neutrophil binding and internalisation. *Bulletin of Mathematical Biology* **78**, 2304–2357 (2016).

8. Friberg, L. E. & Karlsson, M. O. Mechanistic models for myelosuppression. *Investigational New Drugs* **21**, 183–194 (2003).

9. Scholz, M., Engel, C. & Loeffler, M. Modelling human granulopoiesis under polychemotherapy with G-CSF support. *Journal of Mathematical Biology* **50**, 397–439 (2005).

10. Vainstein, V. *et al.* The complex effect of granulocyte colony-stimulating factor on human granulopoiesis analyzed by a new physiologically-based mathematical model. *Journal of Theoretical Biology* **235**, 311–327 (2005).

11. Alfonso, S., Jenner, A. L. & Craig, M. Translational approaches to treating dynamical diseases through in silico clinical trials. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **30**, 123128 (2020).

12. Craig, M. Towards quantitative systems pharmacology models of chemotherapy-induced neutropenia. *CPT: Pharmacometrics and Systems Pharmacology* **6**, 293–304 (2017).

13. Sorger, P. *et al. Quantitative and systems pharmacology in the post-genomic era: new approaches to discovering drugs and understanding therapeutic mechanisms. An NIH white paper by the QSP workshop group -- October 2011*. *An NIH White Paper by the QSP Workshop Group* (2011).

14. Friberg, L. E. & Karlsson, M. O. Mechanistic models for myelosuppression. *Investigational New Drugs* **21**, 183–194 (2003).

15. Scholz, M., Engel, C. & Loeffler, M. Modelling human granulopoiesis under polychemotherapy with G-CSF support. *Journal of Mathematical Biology* **50**, 397–439 (2005).

16. Vainstein, V. *et al.* The complex effect of granulocyte colony-stimulating factor on human granulopoiesis analyzed by a new physiologically-based mathematical model. *Journal of Theoretical Biology* **235**, 311–327 (2005).

17. Jenner, A. L., Cassidy, T., Belaid, K., Bourgeois-Daigneault, M.-C. & Craig, M. In silico trials predict that combination strategies for enhancing vesicular stomatitis oncolytic virus are determined by tumor aggressivity. *Journal for ImmunoTherapy of Cancer* **9**, e001387 (2021).

18. Cassidy, T. & Craig, M. Determinants of combination GM-CSF immunotherapy and oncolytic virotherapy success identified through in silico treatment personalization. *PLOS Computational Biology* **15**, e1007495 (2019).

19. Wang, H. *et al.* In silico simulation of a clinical trial with anti-CTLA-4 and anti-PD-L1 immunotherapies in metastatic breast cancer using a systems pharmacology model. *Royal Society Open Science* **6**, 190366 (2020).

20. Ma, H. *et al.* Combination therapy with T cell engager and PD-L1 blockade enhances the antitumor potency of T cells as predicted by a QSP model. *Journal for ImmunoTherapy of Cancer* **8**, e001141 (2020).

21. Milberg, O. *et al.* A QSP Model for Predicting Clinical Responses to Monotherapy, Combination and Sequential Therapy Following CTLA-4, PD-1, and PD-L1 Checkpoint Blockade. *Scientific Reports* **9**, 11286 (2019).

22. Soufsaf, S., Robaey, P., Bonnefois, G., Nekka, F. & Li, J. A quantitative comparison approach for methylphenidate drug regimens in attention-deficit/hyperactivity disorder treatment. *Journal of Child and Adolescent Psychopharmacology* **29**, 220–234 (2019).

23. Visvader, J. E. & Stingl, J. Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes & Development* **28**, 1143–1158 (2014).

24. Mehta, R. G. *et al.* Differential Roles of ERα and ERβ in Normal and Neoplastic Development in the Mouse Mammary Gland. *PLoS ONE* **9**, e113175 (2014).

25. Romagnoli, M. *et al.* Deciphering the Mammary Stem Cell Niche: A Role for Laminin-Binding Integrins. *Stem Cell Reports* **12**, 831–844 (2019).

26. Canadian Cancer Statistics Advisory Committee. *Canadian Cancer Statistics 2018*. (2018).

27. Fedele, M., Cerchia, L. & Chiappetta, G. The epithelial-to-mesenchymal transition in breast cancer: Focus on basal-like carcinomas. *Cancers (Basel)* **9**, 1–19 (2017).

28. López-Ozuna, V. M., Hachim, I. Y., Hachim, M. Y., Lebrun, J. J. & Ali, S. Prolactin modulates TNBC aggressive phenotype limiting tumorigenesis. *Endocr Relat Cancer* **26**, 321–337 (2019).

29. Dontu, G., El-Ashry, D. & Wicha, M. S. Breast cancer, stem/progenitor cells and the estrogen receptor. *Trends in Endocrinology and Metabolism* **15**, 193–197 (2004).

30. Liu, F. *et al.* Prolactin/Jak2 directs apical/basal polarization and luminal linage maturation of mammary epithelial cells through regulation of the Erk1/2 pathway. *Stem Cell Research* **15**, 376–383 (2015).

31. Shams, A. *et al.* Prolactin receptor-driven combined luminal and epithelial differentiation in breast cancer restricts plasticity, stemness, tumorigenesis and metastasis. *Oncogenesis* **Accepted**, (2020).

32. Lopez-Ozuna, V. M., Hachim, I. Y., Hachim, M. Y., Lebrun, J. J. & Ali, S. Prolactin Pro-Differentiation Pathway in Triple Negative Breast Cancer: Impact on Prognosis and Potential Therapy. *Scientific Reports* **6**, 1–13 (2016).

33. Mackey, M. C. Unified hypothesis for the origin of aplastic anemia and periodic hematopoiesis. *Blood* **51**, 941–956 (1978).

34. Mallepell, S., Krust, A., Chambon, P. & Brisken, C. Paracrine signaling through the epithelial estrogen receptor α is required for proliferation and morphogenesis in the mammary gland. *Proceedings of the National Academy of Sciences* **103**, 2196–2201 (2006).

35. Cristea, S. & Polyak, K. Dissecting the mammary gland one cell at a time. *Nature Communications* **9**, 2473 (2018).

36. Brunetti, M., Mackey, M. C. & Craig, M. Understanding Normal and Pathological Hematopoietic Stem Cell Biology Using Mathematical Modelling. *Current Stem Cell Reports* (2021) doi:10.1007/s40778-021-00191-9.

37. Reif, S., Snelder, N. & Blode, H. Characterisation of the pharmacokinetics of ethinylestradiol and drospirenone in extended-cycle regimens: population pharmacokinetic analysis from a randomised Phase III study. *The Journal of Family Planning and Reproductive Health Care* **39**, e1–e13 (2013).

38. Reif, S., Snelder, N. & Blode, H. Characterisation of the pharmacokinetics of ethinylestradiol and drospirenone in extended-cycle regimens: population

pharmacokinetic analysis from a randomised Phase III study. *The Journal of Family Planning and Reproductive Health Care* **39**, e1–e13 (2013).

39. Huang, S. Gene expression profiling, genetic networks, and cellular states: An integrating concept for tumorigenesis and drug discovery. *Journal of Molecular Medicine* **77**, 469–480 (1999).

40. Becker, J. B. *et al. Sex Differences in the Brain: From Genes to Behavior*. (Oxford University Press, 2007).

41. Marino, S., Hogue, I. B., Ray, C. J. & Kirschner, D. E. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology* **254**, 178–196 (2008).

42. Marino, S., Hogue, I. B., Ray, C. J. & Kirschner, D. E. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology* **254**, 178–196 (2008).

43. Sutherland, R. L., Hall, R. E., Pang, G. Y. N., Musgrove, E. A. & Clarke, C. L. Effect of Medroxyprogesterone Acetate on Proliferation and Cell Cycle Kinetics of Human Mammary Carcinoma Cells. *Cancer Research* **48**, 5084–5091 (1988).

44. Clarke, R. B. *et al.* A putative human breast stem cell population is enriched for steroid receptor-positive cells. *Developmental Biology* **277**, 443–456 (2005).

45. Chang, C. H. *et al.* Mammary Stem Cells and Tumor-Initiating Cells Are More Resistant to Apoptosis and Exhibit Increased DNA Repair Activity in Response to DNA Damage. *Stem Cell Reports* **5**, 378–391 (2015).

46. Huang, J. *et al.* Downregulation of estrogen receptor and modulation of growth of breast cancer cell lines mediated by paracrine stromal cell signals. *Breast Cancer Research and Treatment* **161**, 229–243 (2017).

47. Huang, J. *et al.* Downregulation of estrogen receptor and modulation of growth of breast cancer cell lines mediated by paracrine stromal cell signals. *Breast Cancer Research and Treatment* **161**, 229–243 (2017).

48. Humphreys, R. C. *et al.* Apoptosis in the terminal endbud of the murine mammary gland: A mechanism of ductal morphogenesis. *Development* **122**, 4013–4022 (1996).

49. Joseph, R. *et al.* Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. *Molecular Systems Biology* **6**, 1–13 (2010).

50. Eirew, P. *et al.* A method for quantifying normal human mammary epithelial stem cells with in vivo regenerative ability. *Nature Medicine* **14**, 1384–1389 (2008).

51. Tyson, J. J. *et al.* Dynamic modelling of oestrogen signalling and cell fate in breast cancer cells. *Nature Reviews Cancer* vol. 11 523–532 (2011).

52. Tyson, J. J. *et al.* Dynamic modelling of oestrogen signalling and cell fate in breast cancer cells. *Nature Reviews Cancer* vol. 11 523–532 (2011).

53. Craig, M., González-Sales, M., Li, J. & Nekka, F. Impact of pharmacokinetic variability on a mechanistic physiological pharmacokinetic/pharmacodynamic model: a case study of neutrophil development, PM00104, and filgrastim. in *Mathematical Sciences with Multidisciplinary Applications* (ed. Toni, B.) vol. 157 91–112 (Springer Science + Business Media, 2016).

54. NCD Risk Factor Collaboration. Trends in adult body-mass index in 200 countries from 1975 to 2014: A pooled analysis of 1698 population-based measurement studies with 19.2 million participants. *The Lancet* **387**, 1377–1396 (2016).

55. NCD Risk Factor Collaboration. A century of trends in adult human height. *Elife* **5**, 1–29 (2016).

56. BenSaïda, A. Shapiro-Wilk and Shapiro-Francia normality tests. MATLAB Central File Exchange (2021).

Chapter 4 – Discussion

The application of mixed effects models continues to grow in relevance for understanding biological systems[1]. This thesis demonstrates the use of mixed effects modelling for two drastically different situations. The second chapter featured the use of mixed effects within a linear model with cumulative linkage in addition to a nonlinear mixed effects model to examine two very similar situations from slightly different angles, while the third concerned the more classical scenario of a population pharmacokinetics model integrated within a physiological quantitative systems pharmacology model (albeit in a non-standard way). In each of the three models studied within this thesis, mixed effects modelling was used primarily to capture inter-individual variation rather than an interest in the variation itself.

Linear mixed effects modelling is by far the most used approach when comparing linear and nonlinear mixed effects models[1,2]. This model is similar enough to classical linear regression, making it relatively easier to learn and apply. The general preference towards linear models also stems from their ease of interpretation and the facility of visualization[2]. Further, there is also more and powerful software available for linear mixed effects modelling; the large software support base also helps to guide new users[3,4].

In Chapter 2, the ordinal model described by Eq. (2.1) is simply an extension to the more classic linear mixed effects model introduced in Eq. (1.11). By using a cumulative link function, we simply extended the common linear mixed effects model to an ordinal response variable. We applied this model to estimate a mixed intercept to account for variability among maple syrup producers and used mixed effects to capture the differences between producers caused by uncontrolled variables such as the latitude of the producer's orchard. Our results showed that COLORI and total amino acid concentrations are strong predictors for syrup quality. It also showed that transmittance provides valuable information as a predictor variable even though it itself is not a great predictor. This result is unexpected and therefore warrants further investigation. The model is more accessible to producers during any given season and provides a greater range in quality prediction than a COLORI test.

The second model in Chapter 2 (Eqs. (2.2)-(2.4)) is a nonlinear mixed effects model with a modification to the time variable. This modelling choice was made as we could not use time as an independent variable as the maple syrup production season varies year to year and there is no way to determining the end date of a season during the season. The nonlinear nature of the model allowed us to use another explanatory variable (pH) as a proxy for time. We again accounted for variability among producers by estimating parameters for the parameters in the time proxy and COLORI as mixed effects parameters. Equations (2.2)-(2.4)were built based on *a priori* data exploration. Our results showed that COLORI is, unsurprisingly, an excellent predictor for amino acid concentration. pH serves as a decent time proxy. Since pH tests are easy and accessible to the producers, this model can provide an immediate estimation of the amino acid concentration with minimum effort. With COLORI results, good estimations of amino acid concentration can be fitted.

Both models from the second chapter of this thesis did not have any prior framework, especially the nonlinear model in Eqs. (2.2)-(2.4), so it was important for us to follow the protocol described in the Figure *1.2* when constructing these models. We were able to use the method of starting with a global model combining all explanatory variables to fit every variable combination to find the best AIC score (see Eq. (1.17)). Due to the linearity of the model, we were also able to use variable significance as an indicator for model validation. In contrast, the nonlinear model (Eqs. (2.2)-(2.4)) was more reliant on data exploration since both the structure of the model and the relationship between variables were unknown. For this, correlation plots between all variables were explored, under the requirement that COLORI be included in the model. Different curves were fitted for COLORI versus total amino acid concentration to find a representative equation. For all other variables, we fit different curves between pairs of variables to determine which variable(s) were suitable as the time proxy and which were highly correlated with COLORI. While the outcomes of both models were similar, each model required more attention to different portions in the model building process.

The use of mixed effects model is also highly prevalent in pharmacokinetic modelling in the form of population pharmacokinetics (PopPK) models that rely on nonlinear mixed effects models[5,6]. In fact, it could be posited that the PopPK models had a hand in establishing some of the

theoretical basis of NLMEMs as they developed concurrently. During drug development and after a drug reaches the market, PopPK models are used to analyse information from a patient pool that is reflective of a target population. By estimating inter-individual variability, PopPK models tell us which parameters vary significantly within the population, indicating that there may be a need to administer drugs differently to certain populations. Covariates within the PopPK model allow is to further characterize subgroups of the population, for example elder or female patients (or both), that may have otherwise been overlooked. Mixed effects modelling also allow more sparse data to be used due to its 'partial pooling' nature (see Section 1.2.2) which is advantageous when data collecting is difficult, a scenario that is more common than not[7].

We explored the use of NLMEM in PopPK models in Chapter 3 to study the effects of estradiol on myoepithelial mammary cell production. Unlike the two models built in Chapter 2, PopPK models are well defined in their structure. The estradiol PopPK model is a three compartment PK model with one central compartment and two peripheral compartments. The central compartment represents any physiological space in which there is easy access to estradiol, usually the plasma in which circulating estradiol can quickly move around the body. The peripheral compartments are physiological spaces, like fat or organs, to which estradiol can enter and become inaccessible[8]. The central compartment portion of the estradiol PopPK model was integrated into the main structural model equal for free estrogen. By including the PopPK model, the dynamics of estradiol are represented more realistically. This model does not assume that any absorbed estradiol is constantly available in the blood, ready to be bound to receptors. The addition of PK variability made little to no impact on the QSP model's predictions suggesting that, though PopPK models account for inter-individual variability, some of this heterogeneity may not be due necessarily to physiological noise but instead to measurement errors or other sources of noise. Further, the integration of PK models within physiologically constructed quantitative systems pharmacology (QSP) models may not need to rely heavily on PopPK analyses, as QSP models inherently account for physiological variability driving heterogenous drug outcomes.

One of the main issues we faced throughout this thesis was the lack of validation data. Like all mathematical models, mixed effects models can only be as good as the data on which they are built. Mixed effects models have the advantage over regression by being able to share

information among the groups via partial pooling. In general, mixed effects models can be built upon less data. Even so, there is a limit to how little the training dataset can be before model estimation quality deteriorates, or worse, not converge at all. For our maple syrup study described in Chapter 2, we could not afford to split the data into a training and testing set without risking the fit of the model. Luckily, we did not face such issues for our breast cancer model. Because the PopPK model for estradiol was previously built and the rest of the QSP model was structural (i.e., constructed based off physiological knowledge), all of the available data was used as validation data. However, we did face a lack of data to provide estimates of the physiological parameters in the QSP model, and this may affect our results.

A rather unique issue we had in Chapter 2 was the time variable, or rather lack thereof, in the models of maple syrup production. NLMEMs are designed for repeated measures which, by definition, involve a time variable. In our breast cancer model in Chapter 3, the beginning of time is marked by the administration of ethinyl estradiol. The simulations were then performed for a pre-determined set of time that was relevant to the dynamics we were studying. Maple syrup production seasons do not have such well defined beginnings nor ends. The season is dictated by its end date, the day of bud break, which is unknown until after the season's completion. The start and progression of the season is heavily dependent on the orchard's location and the annual weather patterns. Thus, we did not have access to a time variable that did not require either the user to guess the time value or to wait after the season ends for the exact time scale of that season. Both of these scenarios would defeat the purpose of the modelling exercise, as they are retrospective instead of prospective. Our solution to this issue was to replace the time variable with a proxy variable. However, there are more sophisticated methods that should be explored in the future.

In this thesis, we explored the use of mixed effects modelling in a classic case, pharmacokinetics, and a more novel case, maple syrup quality. Though quite different, the fact that both studies can be tackled by the same model type showcases the versatility of NLMEM. Our results highlight that biological systems often have intrinsic, uncontrollable variability to which NLMEMs excels at capturing. Thus, as experimental protocols in biomedicine continue to generate vast amount of data, the popularity of NLMEMs in biological systems will only increase, with important

consequences to our understanding of the system under study and individual- and population-level effects on this system.

References

1 Harrison, X. A. *et al.* A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **6**, e4794-e4794, doi:10.7717/peerj.4794 (2018).

2 Brauer, M. & Curtin, J. J. Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychol Methods* **23**, 389-411, doi:10.1037/met0000159 (2018).

3 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1 - 48, doi:10.18637/jss.v067.i01 (2015).

4 Luke, S. G. Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods* **49**, 1494-1502, doi:10.3758/s13428-016-0809-y (2017).

5 Craig, M., González-Sales, M., Li, J. & Nekka, F. in *Mathematical Sciences with Multidisciplinary Applications.* (ed Bourama Toni) 91-112 (Springer International Publishing).

6 Bon, C. *et al.* Mathematical modeling and simulation in animal health. Part III: Using nonlinear mixed-effects to characterize and quantify variability in drug pharmacokinetics. *Journal of Veterinary Pharmacology and Therapeutics* **41**, 171-183, doi:https://doi.org/10.1111/jvp.12473 (2018).

7 Kang, D., Schwartz, J. B. & Verotta, D. A sample size computation method for non-linear mixed effects models with applications to pharmacokinetics models. *Statistics in Medicine* **23**, 2551-2566, doi:https://doi.org/10.1002/sim.1695 (2004).

8 Zou, H., Banerjee, P., Leung, S. S. Y. & Yan, X. Application of Pharmacokinetic-Pharmacodynamic Modeling in Drug Delivery: Development and Challenges. *Frontiers in Pharmacology* **11**, doi:10.3389/fphar.2020.00997 (2020).