Collective Variational Inference for Personalized and Generative Physiological Modeling: A Case Study on Hemorrhage Resuscitation

Ali Tivay, George C. Kramer, and Jin-Oh Hahn, *Senior Member, IEEE*

***Abstract*— *Objective:* Individual physiological experiments typically provide useful but incomplete information about a studied physiological process. As a result, inferring the unknown parameters of a physiological model from experimental data is often challenging. The objective of this paper is to propose and illustrate the efficacy of a collective variational inference (C-VI) method, intended to reconcile low-information and heterogeneous data from a collection of experiments to produce robust personalized and generative physiological models. *Methods:* To derive the C-VI method, we utilize a probabilistic graphical model to impose structure on the available physiological data, and algorithmically characterize the graphical model using variational Bayesian inference techniques. To illustrate the efficacy of the C-VI method, we apply it to a case study on the mathematical modeling of hemorrhage resuscitation. *Results:* In the context of hemorrhage resuscitation modeling, the C-VI method could reconcile heterogeneous combinations of hematocrit, cardiac output, and blood pressure data across multiple experiments to obtain (i) robust personalized models along with associated measures of uncertainty and signal quality, and (ii) a generative model capable of reproducing the physiological behavior of the population. *Conclusion:* The C-VI method facilitates the personalized and generative modeling of physiological processes in the presence of low-information and heterogeneous data. *Significance:* The resulting models provide a solid basis for the development and testing of interpretable physiological monitoring, decision-support, and closed-loop control algorithms.**

***Index Terms*— Collective Inference, Variational Inference, Personalized Medicine, Digital Twin, In Silico Clinical Trials, Virtual Patients, Hemorrhage, Fluid Resuscitation.**

I. Introduction

AUTONOMOUS physiological monitoring and medical intervention can potentially provide substantial improvements to the safety and effectiveness of medical care by making recommendations and/or performing interventions in a continuous, precise, and personalized manner [1]–[4]. In recent years, this potential has motivated a considerable body of research on the design and testing of decision-support and closed-loop control systems for medical intervention [5]–[17]. Yet, regulatory approval and widespread real-world adoption of these technologies necessitate further advancements in the state of the art in terms of patient safety, physiological interpretability, awareness of physiological context, and the ability to coordinate multiple therapeutic objectives associated with multiple physiological outputs [18], [19]. Arriving at an interpretable, context-aware, and coordinated autonomous medical care system is highly contingent upon representative mathematical models of the relevant physiological mechanisms [20]–[22]. These mathematical models, however, are usually only determined up to a set of latent (i.e., unknown) parameters that must be inferred from experimental data.

An array of versatile methods have been proposed in the literature that can be utilized to infer the unknown parameters of a mathematical model. Maximum-Likelihood Estimation (MLE) is a popular technique that can be used to find point estimates for model parameters by maximizing the likelihood of observed data [23], while Bayesian inference techniques can provide posterior beliefs about model parameters based on prior beliefs and observed data. These posterior beliefs can in turn be used to extract point estimates for model parameters and quantify parameter uncertainties [24], [25]. Obtaining posterior beliefs for model parameters is in general a non-trivial mathematical problem. However, many effective numerical solutions have been proposed for this purpose: the Markov Chain Monte Carlo (MCMC) class of algorithms such as Metropolis-Hastings and Hamiltonian-Monte-Carlo can provide high-fidelity samples from the posterior [24], [26], [27], while approaches such as Approximate Bayesian Computation can provide approximate samples from the posterior [28], [29]. In recent years, statistics and machine learning research has shown notable advances in Variational Inference (VI) techniques, through which it is possible to obtain analytical approximations to the posterior using optimization [30], [31]. VI algorithms tend to require fewer computations than MCMC, while stochastic and amortized variants of VI provide the opportunity to handle larger datasets and more complex problem formulations [32], [33].

In addition to the method of inference, the formulation of the inference problem has notable effects on the fidelity of the

Research supported by National Science Foundation CAREER Award (Grant No. 1748762), and CDMRP (Grant No. W81XWH-19-1-0322).

A. Tivay and J.O. Hahn are with the Mechanical Engineering Department, University of Maryland, College Park, MD 20742 USA (phone: 301-405-7864; fax: 301-314-9477; e-mail: jhahn12@umd.edu).

G. Kramer is with the Anesthesiology Department, University of Texas Medical Branch, Galveston, TX 77555 USA.

resulting mathematical models. For instance, in case experimental data are obtained from multiple non-identical subjects, the problem formulation must account for the existence of inter-subject variability. The Empirical Bayes framework and its closely related counterpart in Nonlinear Mixed-Effects Modeling are effective problem formulations for this purpose, where subject-level models (e.g., personalized models) are augmented with a population-level model (e.g., a model of inter-subject variability, or equivalently, an empirical prior) to represent the conditions under which the data were obtained [34]–[37]. In recent years, several interesting variants of such a hierarchical problem formulation have been proposed, especially for dynamic systems modeling and machine learning applications [33], [38]–[41]. More generally, the Probabilistic Graphical Modeling (PGM) framework provides a rich set of tools for formulating effective probabilistic dependence structures for a given problem class according to problem-specific challenges and objectives [42], [43].

Inferring the unknown parameters of a physiological model from experimental data presents a unique set of challenges that appear frequently in this area of research. Acquiring data for the purpose of hemorrhage resuscitation modeling, for example, involves applying stimuli (e.g., hemorrhage and fluid infusions) to subjects while measuring their relevant physiological variables such as hematocrit (HCT), cardiac output (CO), and mean arterial pressure (MAP) over time. Such physiological experiments tend to exhibit challenging characteristics: First, each experiment provides *low information*, in the sense that (i) stimuli only partially excite the underlying physiological dynamics, (ii) insufficient physiological measurements are available in each experiment, and (iii) the measured signals are of relatively low quality (e.g., in terms of noise and sampling rate). Second, the experiments are *heterogeneous*, in the sense that there exist (i) variations in experimental protocols (e.g., shape/timing of stimuli), (ii) variations in the availability of measured variables (e.g., HCT and CO may not be available in some subjects), and (iii) variations in subject characteristics, including the possibility of atypical responses to stimuli. These challenges, if not explicitly addressed in the course of parameter inference, tend to produce physiological models with unrealistic parameter values and/or limited predictive capability [44]–[47].

The objective of this paper is to propose and illustrate the efficacy of a *collective variational inference* (C-VI) method, intended to reconcile low-information and heterogeneous data from a collection of experiments to obtain robust personalized and generative physiological models. The *personalized* model aims to reproduce the physiological behavior of a specific subject, while the *generative* model aims to reproduce the physiological behavior of the population. To derive the C-VI method, we compose a PGM to structurally represent the scenario in which low-information and heterogeneous experiments are conducted on a collection of non-identical subjects. Given this problem formulation, obtaining personalized and generative models for a physiological process boils down to inferring the latent parameters of this PGM structure. For this purpose, we leverage recent advances in stochastic VI to obtain an algorithmic procedure that computes approximate posteriors for the PGM parameters through stochastic optimization. To illustrate the efficacy of the C-VI method, we apply it to a practically important case study on the mathematical modeling of hemodynamic responses to hemorrhage resuscitation, and compare the models produced by the C-VI method with those produced by a non-collective method based on MLE. In this context, we demonstrate that the C-VI method can reconcile heterogeneous combinations of HCT, CO, and MAP data across multiple experiments to obtain (i) robust personalized models along with associated measures of uncertainty and signal quality, and (ii) a generative model capable of reproducing the physiological behavior of the population. Finally, we discuss how the resulting models may provide basis for the development and testing of interpretable physiological monitoring, decision-support, and closed-loop control algorithms.

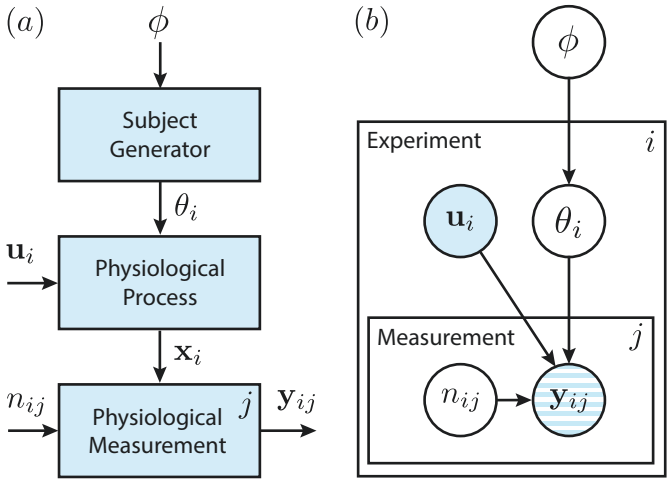

Fig. 1. Formulation of the inference problem for personalized and generative physiological modeling. (a) Schematic view of the generative model structure for one experiment conducted on one subject sample, where several physiological variables are measured. (b) Probabilistic graphical representation of the dependencies between the latent parameters (white), the measured variables (blue), and the sometimes-measured variables (striped blue) in the model structure.

II. Collective Variational Inference

In this section, we present the C-VI methodology, which is intended to enable personalized and generative modeling of physiological processes using low-information and heterogeneous data. This methodology formulates the physiological modeling problem in such a way that multiple experiments can collectively provide information to characterize a physiological system, both at the level of each subject and the population encompassing all subjects.

A. Generative Modeling for Physiological Data

In the first step toward deriving the C-VI method, we aim to formulate a generative model of the processes underlying the acquisition of low-information and heterogeneous physiological data in multiple experiments. This generative model is schematically shown in Fig. 1(a) for one representative experiment. This model is a hierarchical model consisting of three main levels. At the highest level, a subject generator

model is tasked with generating virtual subjects with varying physiological characteristics, which can be formalized as:

$$\theta_i \sim \mathcal{G}(\phi) \tag{1}$$

where $\mathcal{G}$ is the subject generator model, ϕ is the vector of latent parameters for the subject generator model, and θ_i denotes a generated parameter vector representing the physiological characteristics of a virtual subject. At the second level, a virtual physiological experiment is conducted on each virtual subject, whose response to the experiment is generated by a physiological process model $\mathcal{H}$:

$$\mathbf{x}_i(t) = \mathcal{H}(\theta_i, \mathbf{u}_i(t)) \tag{2}$$

where $\mathbf{u}_i(t)$ represents the physiological stimuli associated with the experiment on subject i, and $\mathbf{x}_i(t)$ represents the state evolution of subject i during the course of the experiment. At the third level, in each virtual experiment, one or several physiological variables (e.g. blood pressure) are measured from the state evolution $\mathbf{x}_i(t)$ and recorded as virtual data. Thus, for each measured variable j, we consider a physiological measurement model $\mathcal{M}_j$, which can be formalized as:

$$\mathbf{y}_{ij}^m = \mathcal{M}_j^m(\theta_i, \mathbf{x}_i(t)) \tag{3}$$

$$\mathbf{y}_{ij} \sim \mathcal{M}_j^o(n_{ij}, \mathbf{y}_{ij}^m) \tag{4}$$

where $\mathbf{y}_{ij}^m$ is a vector containing the model-generated outputs for the physiological variable j in subject i, $\mathbf{y}_{ij}$ is a vector containing the model-generated (and possibly noisy) virtual data for this physiological variable, and n_{ij} is a latent parameter modulating the signal quality of the virtual data. In summary, the generative model structure described by (1)-(4) is built to (i) generate a virtual subject cohort of arbitrary size, (ii) conduct virtual experiments on the generated cohort, and (iii) generate virtual datasets by compiling the results of the virtual experiments. Given this model structure, our aim is to infer the parameters ϕ, θ_i, n_{ij} using real physiological data, such that the n_{ij}'s capture the signal quality in each real experiment, the θ_i's capture the physiological characteristics of the real subjects, and the subject generator with ϕ produces virtual subjects that are representative of the population.

B. Collective Inference for Generative Modeling

The generative model presented in Section II-A imposes a hierarchical relationship between the variables of interest in the physiological modeling problem. A probabilistic graphical representation of this relationship is shown in Fig. 1(b). In this representation, the generator model parameters ϕ act as global random variables that affect all subject characteristic vectors θ_i. In addition, each subject characteristic vector θ_i together with its corresponding stimuli $\mathbf{u}_i$ act as local random variables with respect to their own experiment. Finally, the virtual data $\mathbf{y}_{ij}$ and the signal quality parameters n_{ij} act as local random variables with respect to their own experiment and measured variable. This hierarchical relationship can be formalized using the following joint density:

$$\begin{aligned} p(\phi, \boldsymbol{\theta}, \mathbf{n}, \mathbf{u}, \mathbf{y}) &= p(\phi)p(\boldsymbol{\theta}|\phi)p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{n}, \mathbf{u})p(\mathbf{n})p(\mathbf{u}) = \\ & p(\phi) \prod_i \left[p(\theta_i|\phi)p(\mathbf{u}_i)\right] \prod_{i,j} \left[p(\mathbf{y}_{ij}|\theta_i, n_{ij}, \mathbf{u}_i)p(n_{ij})\right] \end{aligned} \tag{5}$$

where the symbols $\boldsymbol{\theta}$, $\mathbf{n}$, $\mathbf{u}$, and $\mathbf{y}$ respectively denote the collection of all random variables corresponding to θ_i, n_{ij}, $\mathbf{u}_i$, and $\mathbf{y}_{ij}$. The physiological stimuli $\mathbf{u}$ and the virtual data $\mathbf{y}$ are observed random variables, while the parameters denoted by ϕ, $\boldsymbol{\theta}$, $\mathbf{n}$ are latent random variables that need to be inferred using their relationship to the observed random variables. This inference objective can be expressed in probabilistic terms as calculating the following conditional density:

$$p(\phi, \boldsymbol{\theta}, \mathbf{n}|\mathbf{u}, \mathbf{y}) = \frac{p(\phi, \boldsymbol{\theta}, \mathbf{n}, \mathbf{u}, \mathbf{y})}{p(\mathbf{u}, \mathbf{y})} \tag{6}$$

which is the *exact posterior* density, and represents the ultimate objective of inference for the purpose of personalized and generative physiological modeling in this work. Obtaining this exact posterior is tantamount to utilizing the available data in a collective manner to obtain personalized and generative physiological models along with associated measures of uncertainty and signal quality. However, computing this exact posterior is mathematically intractable and computationally expensive for many physiological applications, which motivates the derivation of an *approximate posterior* that can be computed with reasonable accuracy and computational efficiency.

C. Variational Inference for Generative Modeling

In this work, we employ a variational approach [30], [33] to finding analytical approximations to the exact posterior in (6). In this approach, a family of densities over the latent variables $q(\phi, \boldsymbol{\theta}, \mathbf{n}|\boldsymbol{\nu})$ is formulated with each member (represented by the variational parameter $\boldsymbol{\nu}$) acting as a candidate for the best approximate posterior. To develop a procedure for finding the best approximate posterior, we start from the Kullback-Leibler (KL) divergence between the exact posterior and the candidate approximate posterior:

$$D_{KL}(\boldsymbol{\nu}) = \mathbb{E}_q\big[\log q(\phi, \boldsymbol{\theta}, \mathbf{n}|\boldsymbol{\nu}) - \log p(\phi, \boldsymbol{\theta}, \mathbf{n}|\mathbf{u}, \mathbf{y})\big] \tag{7}$$

where the operator $\mathbb{E}_q$ represents expectation with respect to samples from the approximate posterior. Substituting the exact posterior equation (6) and the joint density (5) into (7), and assuming that the stimuli $\mathbf{u}$ are observed accurately, we obtain the following equation for the KL-divergence:

$$\begin{aligned} D_{KL}(\boldsymbol{\nu}) = \mathbb{E}_q\big[&\log q(\phi, \boldsymbol{\theta}, \mathbf{n}|\boldsymbol{\nu}) - \log p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{n}, \mathbf{u}) \\ &- \log p(\boldsymbol{\theta}|\phi) - \log p(\mathbf{n}) - \log p(\phi)\big] + \log p(\mathbf{y}) \end{aligned} \tag{8}$$

In this equation, two inherently hard-to-compute terms are present: (i) the $D_{KL}(\boldsymbol{\nu})$ term, i.e., the dissimilarity between the approximate and the exact posterior densities, which needs to be minimized, and (ii) the $p(\mathbf{y})$ term, i.e., the model evidence, which depends only on the model definition. Putting these two terms together, we obtain the following quantity:

$$L(\boldsymbol{\nu}) = \log p(\mathbf{y}) - D_{KL}(\boldsymbol{\nu}) \tag{9}$$

which is the *evidence lower bound* and needs to be maximized. Combining equations (8) and (9), we obtain the following expression for this objective:

$$\begin{aligned} L(\boldsymbol{\nu}) = \mathbb{E}_q\big[&\log p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{n}, \mathbf{u}) + \log p(\boldsymbol{\theta}|\phi) + \log p(\mathbf{n}) \\ &+ \log p(\phi) - \log q(\phi, \boldsymbol{\theta}, \mathbf{n}|\boldsymbol{\nu})\big] \end{aligned} \tag{10}$$

Algorithm 1 Stochastic Optimization for C-VI

Input: $\boldsymbol{\nu}_0, \beta_1, \beta_2, \alpha, \delta$ {initial guess and constants}
Output: $\boldsymbol{\nu}$ {optimized variational parameters}
  $t \leftarrow 0$
  $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu}_0$; $\mathbf{m}_{\boldsymbol{\nu}} \leftarrow 0$; $\mathbf{v}_{\boldsymbol{\nu}} \leftarrow 0$ {initialize}
  loop
    $t \leftarrow t + 1$
    $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$
    $\mathbf{g}_{\boldsymbol{\nu}} \leftarrow \nabla_{\boldsymbol{\nu}} \widetilde{L}(\boldsymbol{\epsilon}, \boldsymbol{\nu})$ {from objective}
    $\mathbf{m}_{\boldsymbol{\nu}} \leftarrow [\beta_1 \mathbf{m}_{\boldsymbol{\nu}} + (1 - \beta_1)\mathbf{g}_{\boldsymbol{\nu}}]/(1 - \beta_1^t)$
    $\mathbf{v}_{\boldsymbol{\nu}} \leftarrow [\beta_2 \mathbf{v}_{\boldsymbol{\nu}} + (1 - \beta_2)\mathbf{g}_{\boldsymbol{\nu}}^2]/(1 - \beta_2^t)$
    $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu} + \alpha[\mathbf{m}_{\boldsymbol{\nu}}/(\sqrt{\mathbf{v}_{\boldsymbol{\nu}}} + \delta)]$ {see [48]}
  end loop
  return $\boldsymbol{\nu}$ {optimized variational parameters}

Algorithm 2 Evaluation of Stochastic Objective for C-VI

Input: $\boldsymbol{\epsilon}, \boldsymbol{\nu}, \mathbf{u}, \mathbf{y}^d$ {parameters, stimuli, and data}
Output: $\widetilde{L}(\boldsymbol{\epsilon}, \boldsymbol{\nu})$
  $\mathbf{z}_\phi, \mathbf{z}_\theta, \mathbf{z}_\mathbf{n} \leftarrow f_q(\boldsymbol{\epsilon}, \boldsymbol{\nu})$ {sample posterior; see (27)}
  $p_q \leftarrow \log q(\mathbf{z}_\phi, \mathbf{z}_\theta, \mathbf{z}_\mathbf{n}|\boldsymbol{\nu})$ {see (28)}
  for all i, j **do**
    $\mathbf{x}_i \leftarrow \mathcal{H}(\mathbf{z}_{\theta_i}, \mathbf{u}_i)$ {run physiological model}
    $\mathbf{y}_{ij}^m \leftarrow \mathcal{M}_j^m(\mathbf{z}_{\theta_i}, \mathbf{x}_i)$ {get model outputs}
  end for
  $p_\mathbf{y} \leftarrow \log p(\mathbf{y}|\mathbf{z}_\theta, \mathbf{z}_\mathbf{n}, \mathbf{u})$ {compare $\mathbf{y}^d$ and $\mathbf{y}^m$; see (35)}
  $p_\theta \leftarrow \log p(\mathbf{z}_\theta|\mathbf{z}_\phi)$ {see (30)}
  $p_\phi \leftarrow \log p(\mathbf{z}_\phi)$ {prior on ϕ; see (31)}
  $p_\mathbf{n} \leftarrow \log p(\mathbf{z}_\mathbf{n})$ {prior on $\mathbf{n}$; zero for no prior}
  $\widetilde{L} \leftarrow p_\mathbf{y} + p_\theta + p_\phi + p_\mathbf{n} - p_q$
  return $\widetilde{L}$ {objective value}

which represents the main objective for the physiological modeling problem addressed in this work.

D. Interpretation of the Objective and Special Cases

The objective shown in (10) is based on an expectation with respect to samples from the approximate posterior $q(\phi, \boldsymbol{\theta}, \mathbf{n}|\boldsymbol{\nu})$, the shape of which can be modulated through the variational parameters $\boldsymbol{\nu}$. The term $\log p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{n}, \mathbf{u})$ is a log-likelihood term that promotes similarity between observed physiological data and the model-generated virtual data for every experiment. The term $\log p(\boldsymbol{\theta}|\phi)$ promotes personalized models that are likely under the subject generator, and a subject generator that is likely to generate the personalized models. The terms $\log p(\phi)$ and $\log p(\mathbf{n})$ represent prior densities that can encode prior knowledge about the subject generator and the physiological measurement model parameters. Finally, the term $\log q(\phi, \boldsymbol{\theta}, \mathbf{n}|\boldsymbol{\nu})$ promotes a diffuse approximate posterior, acting as a mechanism for uncertainty quantification. As a result of this formulation, special cases of the inference problem can be obtained by removing a subset of terms from the objective. Removing $\log q(\phi, \boldsymbol{\theta}, \mathbf{n}|\boldsymbol{\nu})$ promotes a concentrated approximate posterior, resulting in a point estimation problem over the unknown parameters ϕ, $\boldsymbol{\theta}$ and $\mathbf{n}$. Removing the rest of the terms except for $\log p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{n}, \mathbf{u})$ results in a non-collective MLE problem over $\boldsymbol{\theta}$ and $\mathbf{n}$, promoting separate estimation of model parameters for every subject.

E. Stochastic Optimization Algorithm

As presented in Section II-C, performing inference for the purpose of personalized and generative physiological modeling boils down to maximizing the evidence lower bound objective shown in (10) over the variational parameters. In the absence of further algorithm engineering, this maximization can turn into a prohibitively expensive computational task due to the presence of the expectation operator $\mathbb{E}_q$. In this work, we employ an approach based on stochastic gradients of the objective [32], [33] to perform the desired maximization in a computationally feasible manner. For this purpose, we assume a function f_q that takes as its input the variational parameters $\boldsymbol{\nu}$, and a sample $\boldsymbol{\epsilon}$ from the standard normal distribution (of appropriate dimension), and produces as its output samples from the approximate posterior $q(\phi, \boldsymbol{\theta}, \mathbf{n}|\boldsymbol{\nu})$:

$$[\mathbf{z}_\phi; \mathbf{z}_\theta; \mathbf{z}_\mathbf{n}] = f_q(\boldsymbol{\epsilon}, \boldsymbol{\nu}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I) \tag{11}$$

where $\mathbf{z}_\phi$ is a sampled subject generator model parameter, $\mathbf{z}_\theta$ denotes sampled subject characteristic parameters, and $\mathbf{z}_\mathbf{n}$ denotes sampled signal quality parameters. Substituting (11) into (10), and taking the gradient of both sides yields an equation of the form:

$$\nabla_{\boldsymbol{\nu}} L(\boldsymbol{\nu}) = \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,I)}\left[\nabla_{\boldsymbol{\nu}} \widetilde{L}(\boldsymbol{\epsilon}, \boldsymbol{\nu})\right] \tag{12}$$

where the operator $\mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,I)}$ denotes expectation with respect to samples from the standard normal distribution, and the gradient operator $\nabla_{\boldsymbol{\nu}}$ has been moved inside the expectation since the expectation operator no longer depends on $\boldsymbol{\nu}$. The term inside the expectation $\nabla_{\boldsymbol{\nu}} \widetilde{L}(\boldsymbol{\epsilon}, \boldsymbol{\nu})$ is a stochastic gradient of the objective, which can be written in expanded form as:

$$\begin{aligned}\nabla_{\boldsymbol{\nu}} \widetilde{L}(\boldsymbol{\epsilon}, \boldsymbol{\nu}) = \nabla_{\boldsymbol{\nu}}\big[&\log p(\mathbf{y}|\mathbf{z}_\theta, \mathbf{z}_\mathbf{n}, \mathbf{u}) + \log p(\mathbf{z}_\theta|\mathbf{z}_\phi) \\ &+ \log p(\mathbf{z}_\mathbf{n}) + \log p(\mathbf{z}_\phi) - \log q(\mathbf{z}_\phi, \mathbf{z}_\theta, \mathbf{z}_\mathbf{n}|\boldsymbol{\nu})\big]\end{aligned} \tag{13}$$

According to (12), the stochastic gradient $\nabla_{\boldsymbol{\nu}} \widetilde{L}(\boldsymbol{\epsilon}, \boldsymbol{\nu})$ is an unbiased noisy sample from the actual gradient $\nabla_{\boldsymbol{\nu}} L(\boldsymbol{\nu})$. Therefore, this stochastic gradient can be used along with a stochastic optimization algorithm to maximize the objective $L(\boldsymbol{\nu})$. Stochastic optimization with unbiased gradients has been shown to exhibit favorable convergence properties in theory and practice [48]–[50]. In addition, although convergence results may apply only locally to non-convex objectives, the randomized nature of stochastic optimization has been shown to facilitate escapes from local extrema, resulting in state-of-the art solutions in many practical applications [51].

The stochastic optimization procedure used in this work is shown in Algorithm 1. This procedure operates by sampling the stochastic gradient of the objective in each iteration, and producing corresponding increments to the variational parameters through adaptive moment estimation [48]. Within each iteration, the stochastic objective is computed according to Algorithm 2 based on (11)-(13). This procedure computes the objective by considering a sample from the approximate

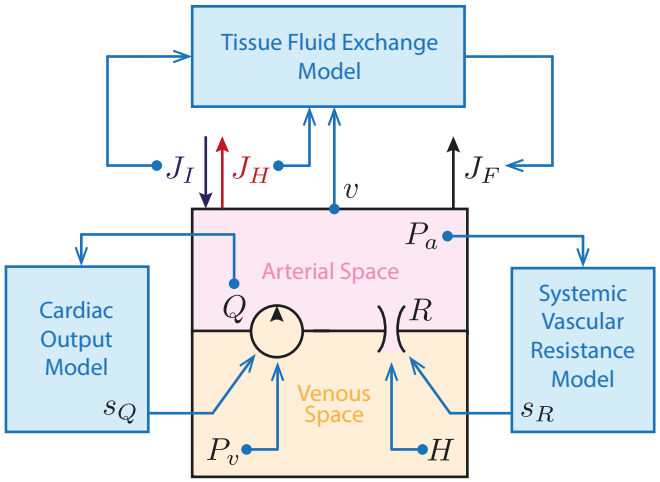

Fig. 2. A schematic representation of the hemorrhage resuscitation model. The model consists of four main components: (i) the blood circulation model, (ii) the tissue fluid exchange model, (iii) the systemic vascular resistance model, and (iv) the cardiac output model.

posterior, and evaluating the consistency of the sample both with respect to the available data and the structure of the generative model. In this way, the proposed iterative procedure searches for a generative model that is consistent with the available physiological data as a whole, and additionally produces signal quality estimates, personalized models, and uncertainty quantification as byproducts.

III. The Hemorrhage Resuscitation Model

In this work, we use a hemorrhage resuscitation modeling case study to illustrate the efficacy of the C-VI method in the context of personalized and generative physiological modeling. In this modeling problem, the aim is to obtain interpretable mathematical models that can reproduce and predict the hemodynamic effects of hemorrhage and fluid resuscitation, both in specific subjects and in a given population. In addition to illustrating the efficacy of C-VI, the knowledge encoded in these models may be leveraged and extended to complement and further advance the state of the art in the development and testing of physiological monitoring [52], [53], decision support, and closed-loop control algorithms [10], [18] for hemorrhage resuscitation. In this section, we build upon our previous work [44], [54], [55] to derive a mathematical model of the main physiological phenomena in hemorrhage resuscitation, and specify the stimuli, states, and model parameters for this problem. A schematic representation of this model is shown in Fig. 2. The model consists of four main components that are described in the following subsections.

A. Blood Circulation Model

In order to model the effects of hemorrhage and fluid resuscitation on the physiological state of a subject, a blood circulation model must be formulated that accounts for the volume and composition of blood in relevant circulatory spaces. A macro-state realization of such a model is described by the following differential equations:

$$\dot{v}_a = Q - (P_a - P_v)/R - J_H - J_F \tag{14}$$
$$\dot{v}_v = -Q + (P_a - P_v)/R + J_I \tag{15}$$
$$\dot{v}_r = -J_H H \tag{16}$$

where the states v_a and v_v respectively denote the arterial and venous blood volumes, and v_r denotes the total red blood cell volume. The changes in these volumes are driven by several flow-rate terms. The term Q denotes CO, which is the flow-rate of blood pumped by the heart. The term $(P_a - P_v)/R$ is the flow-rate of blood moving through the vascular system, where P_a is the MAP, P_v is the central venous pressure (CVP), and R is the systemic vascular resistance (SVR). The terms J_H and J_I respectively denote the flow-rates of hemorrhage and fluid resuscitation, while J_F is the net rate of fluid exchange with the tissue space. Finally, the term $J_H H$ is the flow-rate for red blood cell loss due to hemorrhage, where $H = v_r/(v_a + v_v)$ denotes the blood HCT. Prior to any perturbation, the system described by (14)-(16) is set up to be in equilibrium, with baseline arterial, venous, and red blood cell volumes at v_{a0}, v_{v0}, and v_{r0}, respectively, and:

$$Q_0 = (P_{a0} - P_{v0})/R_0 \tag{17}$$

where Q_0 is the baseline CO, P_{a0} and P_{v0} denote baseline MAP and CVP, and R_0 is the baseline SVR. Changes in MAP and CVP are modeled to depend on changes in arterial and venous blood volume as follows:

$$P_a = P_{a0} + K_a(v_a - v_{a0}) \tag{18}$$
$$P_v = P_{v0} + K_v(v_v - v_{v0}) \tag{19}$$

where K_a is the arterial elastance, and K_v is the venous elastance. The model described by equations (14)-(19) is determined except for the responses of J_F, R, and Q, which are addressed in the the next three subsections.

B. Tissue Fluid Exchange Model

Circulating blood interacts with the fluid in the surrounding tissues (i.e., interstitial fluid) through lymphatic and microvascular exchange systems. By virtue of this interaction, the body can regulate the blood volume in the event of external perturbations such as hemorrhage and fluid resuscitation. This is achieved through shifting excess fluid from the blood to the tissue space, or compensating for a dearth of blood by drawing fluid from the tissue space into the blood [54], [56]. The net rate of fluid exchange with the tissue space J_F is thus an important quantity to model. For this purpose, we use a blood volume controller formulation of the form:

$$J_F = K_p(v - v_0 - r_F) \tag{20}$$

where $v = v_a + v_v$ is the total blood volume, $v_0 = v_{a0} + v_{v0}$ is the baseline total blood volume, K_p is the proportional gain of the controller, and r_F is the reference signal for the controller. The reference signal r_F determines the final value of blood volume change long after a blood volume perturbation (i.e.,

hemorrhage and/or fluid resuscitation), and is defined based on the history of blood volume perturbations as follows:

$$\dot{r}_F = \frac{1}{1+\alpha_I} J_I - \frac{1}{1+\alpha_H} J_H \tag{21}$$

where the parameter α_I determines the fraction of the resuscitation fluid that will remain in the blood after the exchange of fluid with the tissue space, and the parameter α_H determines the fraction of hemorrhaged blood that will remain uncompensated after the exchange of fluid with the tissue space (see [54], [57] for more details on this modeling approach).

C. Systemic Vascular Resistance Model

The SVR (R) is the resistance of the vascular system to blood flow. In hemorrhage resuscitation scenarios, the SVR is affected by two main mechanisms: (i) the constriction and dilation of the blood vessels, and (ii) the viscosity of the blood moving through the blood vessels. The control mechanisms in the body (e.g., the baro-reflex mechanism) modulate SVR through vasoconstriction and vasodilation in order to maintain adequate MAP and tissue perfusion [58]. In addition, changes in blood HCT cause changes in blood viscosity, which in turn disturb the SVR [59]. To model these mechanisms, we formulate the SVR in the following form:

$$R = R_0 + K_h(H - H_0) + s_R \tag{22}$$

where $H_0 = v_{r0}/(v_{a0}+v_{v0})$ is the baseline blood HCT, K_h is a parameter representing the sensitivity of the SVR to changes in HCT, and s_R is a state representing the amount of SVR change prompted by the control mechanisms in the body. The state equation for s_R is therefore formulated as follows:

$$\dot{s}_R = -\frac{1}{\tau_R} s_R - \frac{K_R}{\tau_R}(P_a - P_{a0}) \tag{23}$$

where K_R is the controller gain, and τ_R is the time constant of the control system. Equations (22) and (23) together describe a control system whose objective is to maintain adequate MAP by changing SVR, while the SVR is disturbed by viscosity changes resulting from variations in blood HCT.

D. Cardiac Output Model

The CO (Q) is the flow-rate of blood pumped into circulation by the heart. In hemorrhage resuscitation scenarios, the changes in CO stem from two main mechanisms. First, perturbations in CVP directly affect the right atrial pressure and subsequently the left ventricular preload. According to the Frank-Starling law, a higher preload results in higher cardiac muscle tension which in turn results in a more forceful stroke and higher CO. Second, the control mechanisms in the body modulate the heart rate and cardiac contractility in order to maintain adequate CO [58]. To model these mechanisms, we formulate the CO equation in the following form:

$$Q = Q_0 + \beta_v(P_v - P_{v0}) + s_Q \tag{24}$$

where β_v is a parameter representing the sensitivity of the CO to changes in CVP, and s_Q is a state representing the amount of CO change prompted by the control mechanisms in the body. The state equation for s_Q is formulated as follows:

$$\dot{s}_Q = -K_Q(Q - Q_0) \tag{25}$$

where K_Q is the controller gain. Equations (24) and (25) together describe a control system whose objective is to maintain adequate CO through changing s_Q, while the CO is disturbed by changes in CVP.

E. Physiological Model Summary

The presented hemorrhage resuscitation model corresponds to the physiological process model $\mathcal{H}$ defined in (2). For subject i, the physiological stimuli can be summarized as $\mathbf{u}_i(t) = \{J_I(t), J_H(t)\}_i$, and the physiological characteristics of the subject can be summarized as:

$$\theta_i = \big[\, v_0 \;\; H_0 \;\; Q_0 \;\; P_{a0} \;\; K_v \;\; K_a/K_v \;\; K_p \;\; \alpha_I \;\; \alpha_H \;\; K_h \;\; \tau_R \;\; K_R \;\; \beta_v \;\; K_Q \,\big]_i \tag{26}$$

Given θ_i, the rest of the physiological parameters in the model are determined as follows: the baseline arterial and venous blood volumes are nominally set to $v_{a0} = 0.3v_0$, and $v_{v0} = 0.7v_0$; the baseline CVP is set to a nominal value P_{v0} from measured data; and the initial SVR is calculated from $R_0 = (P_{a0} - P_{v0})/Q_0$. Given these parameters, the state evolution of the physiological system $\mathbf{x}_i(t)$ is obtained by numerically solving the differential equations described by (14)-(25).

IV. Methods

In this section, we present the details of applying the C-VI method presented in Section II to the hemorrhage resuscitation modeling problem presented in Section III. In addition, we present the details pertaining to the available physiological data and the methods used for data analysis to illustrate the efficacy of C-VI in the context of personalized and generative physiological modeling, especially in the presence of low-information and heterogeneous data. Further details follow.

A. The Approximate Posterior

As presented in Sections II-C and II-E, variational inference is performed by leveraging a family of densities that act as candidates for the best approximate posterior. In this work, we employ a family of diagonal Gaussian densities for this purpose, which can be written in function form as:

$$[\mathbf{z}_\phi; \mathbf{z}_\theta; \mathbf{z}_\mathbf{n}] = \boldsymbol{\nu}_\mu + \mathrm{diag}(\boldsymbol{\nu}_\sigma)\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I) \tag{27}$$

where $\boldsymbol{\nu}_\mu = [\boldsymbol{\nu}_{\mu:\phi}; \boldsymbol{\nu}_{\mu:\theta}; \boldsymbol{\nu}_{\mu:\mathbf{n}}]$ is the mean vector of the approximate posterior, which represents most-likely values for the model parameters (ϕ, θ, and $\mathbf{n}$), and $\boldsymbol{\nu}_\sigma = [\boldsymbol{\nu}_{\sigma:\phi}; \boldsymbol{\nu}_{\sigma:\theta}; \boldsymbol{\nu}_{\sigma:\mathbf{n}}]$ is the standard deviation vector of the approximate posterior, which represents the uncertainty associated with the model parameters. Thus, the variational parameters for this choice of approximate posterior can be summarized as $\boldsymbol{\nu} = \{\boldsymbol{\nu}_\mu, \boldsymbol{\nu}_\sigma\}$. Given this formulation, the approximate posterior density associated with a sample generated by (27) can be computed from:

$$\log q(\mathbf{z}_\phi, \mathbf{z}_\theta, \mathbf{z}_\mathbf{n}|\boldsymbol{\nu}) = \sum_k [-\tfrac{1}{2}\boldsymbol{\epsilon}_k^2 - \log(\boldsymbol{\nu}_\sigma)_k - \tfrac{1}{2}\log(2\pi)] \tag{28}$$

where the sum $\sum_k$ is computed over the elements of the vectors $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}_\sigma$. The density in (28) is used in Algorithm 2 as part of the stochastic objective.

B. The Subject Generator Model

As presented in Section II-A, the proposed modeling scheme includes a subject generator model $\mathcal{G}(\phi)$. In this work, we employ a full-covariance Gaussian generator for this purpose, which can be written in function form as:

$$\theta_i = \phi_\mu + \phi_L \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I) \tag{29}$$

where ϕ_μ is the mean vector of the subject generator, which represents the most typical subject, and ϕ_L is a lower-triangular matrix denoting the Cholesky decomposition of the covariance matrix of the subject generator. The covariance matrix itself can be computed from $\phi_\Sigma = \phi_L \phi_L^T$. Thus, the subject generator model parameters can be summarized as $\phi = \{\phi_\mu, \phi_L\}$. Given this formulation, the subject generator density associated with a sample from the approximate posterior (as in (27)) can be computed from the following equation:

$$\begin{aligned}\log p(\mathbf{z}_\theta|\mathbf{z}_\phi) = \textstyle\sum_i [-\frac{1}{2}(\mathbf{z}_{\theta_i} - \mathbf{z}_{\phi_\mu})^T \mathbf{z}_{\phi_\Sigma}^{-1}(\mathbf{z}_{\theta_i} - \mathbf{z}_{\phi_\mu}) \\ -\tfrac{1}{2}\log(|\mathbf{z}_{\phi_\Sigma}|) - \tfrac{d_\theta}{2}\log(2\pi)]\end{aligned} \tag{30}$$

where $\mathbf{z}_{\theta_i}$ is the sample associated with the physiological model parameters for subject i, $\mathbf{z}_{\phi_\mu}$ is the sample associated with the mean vector of the subject generator, $\mathbf{z}_{\phi_\Sigma}$ is the sample associated with the covariance matrix of the subject generator, and d_θ is the dimension of physiological model parameters (and also the dimension of $\mathbf{z}_{\theta_i}$ and $\mathbf{z}_{\phi_\mu}$). Equation (30) is used in Algorithm 2 as part of the stochastic objective.

The full-covariance subject generator in (29) is a relatively expressive model that may capture inter-subject variabilities in the form of a covariance matrix ϕ_Σ. However, effective characterization of this covariance matrix from data is contingent on the availability of a sufficient number of subjects in the dataset. In many physiological modeling applications, only a limited number of subjects are available for experimentation, which may in turn result in an "over-fitted" covariance matrix. To create a balance between generator model complexity and subject availability, we utilize regularization on the generator model parameters according to the following function:

$$\log p(\mathbf{z}_\phi) = -\lambda \|\mathbf{z}_{\phi_\Sigma}\|_* \tag{31}$$

where $\|.\|_*$ denotes the nuclear norm of the matrix, and λ is a scalar hyper-parameter. The nuclear norm $\|\mathbf{z}_{\phi_\Sigma}\|_*$ promotes a compressed subject generator in the sense of the sum of singular values for its covariance matrix, while the hyper-parameter λ can be used to modulate the rate of compression. Equation (31) is used in Algorithm 2 as part of the stochastic objective. The hyper-parameter λ is selected using a method conceptually similar to the L-curve approach [60]: λ is increased from zero while the likelihood in (35) (which represents the goodness of fit to the available data) is evaluated. λ is chosen as the value at which the likelihood starts to exhibit large deterioration. Such a choice of λ can achieve an adequate balance between generator complexity and subject availability [44], [60].

C. The Physiological Measurement Model

As presented in Section II-A, the proposed modeling approach includes physiological measurement models $\mathcal{M}_j$ intended to represent the measurement processes used to obtain a real dataset. For the hemorrhage resuscitation problem, we consider three potential types of measurement: HCT, CO, and MAP. We model each of these measurements as the output of a process that observes the model outputs H, Q, and P_a, while the observations are corrupted by additive white Gaussian noise. This can be formulated as:

$$\mathbf{y}_{i,H} = \{H(t) + n_{i,H}.\epsilon \mid t \in T_{i,H},\ \epsilon \sim \mathcal{N}(0,1)\} \tag{32}$$

$$\mathbf{y}_{i,Q} = \{Q(t) + n_{i,Q}.\epsilon \mid t \in T_{i,Q},\ \epsilon \sim \mathcal{N}(0,1)\} \tag{33}$$

$$\mathbf{y}_{i,P_a} = \{P_a(t) + n_{i,P_a}.\epsilon \mid t \in T_{i,P_a},\ \epsilon \sim \mathcal{N}(0,1)\} \tag{34}$$

where $\mathbf{y}_{i,H}$, $\mathbf{y}_{i,Q}$, and $\mathbf{y}_{i,P_a}$ respectively denote the model-generated virtual data for HCT, CO, and MAP in subject i. The sets $T_{i,H}$, $T_{i,Q}$, T_{i,P_a} contain the time points at which observations are made for subject i, and the latent parameters $n_{i,H}$, $n_{i,Q}$, n_{i,P_a} denote the standard deviations of the Gaussian noises corrupting the observations of HCT, CO, and MAP in subject i. For inference purposes, the likelihood associated with the physiological data can be obtained from:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{z}_\theta, \mathbf{z}_\mathbf{n}, \mathbf{u}) = \\ \textstyle\sum_i \sum_j \big[-\frac{1}{2\mathbf{z}_{n_{ij}}^2}(\mathbf{y}_{ij}^m - \mathbf{y}_{ij}^d)^T(\mathbf{y}_{ij}^m - \mathbf{y}_{ij}^d) \\ - d_{ij}\log(\mathbf{z}_{n_{ij}}) - \tfrac{d_{ij}}{2}\log(2\pi)\big]\end{aligned} \tag{35}$$

where $j \in \{H, Q, P_a\}$ is and index for the measured variables in the hemorrhage resuscitation problem, $\mathbf{y}_{ij}^m$ represents the (uncorrupted) physiological model outputs, $\mathbf{y}_{ij}^d$ denotes real data, $\mathbf{z}_{n_{ij}}$ is an approximate posterior sample associated with the signal quality parameter n_{ij}, and d_{ij} denotes the length of the data vector $\mathbf{y}_{ij}^d$. The likelihood (35) is substituted into its place in Algorithm 2 to complete the formulation of the stochastic objective. Furthermore, in this work, we do not assume further prior knowledge of the noise, which can be reflected in Algorithm 2 by setting $\log p(\mathbf{z}_\mathbf{n})$ to zero.

D. The Physiological Data

The physiological data used in this work are derived from a series of hemorrhage resuscitation experiments conducted on sheep subjects ($N = 23$) in an array of previous work [61]–[63]. In each experiment, the animal is subjected to a large initial hemorrhage and two subsequent smaller hemorrhages. To counter the physiological effects of hemorrhage over time, the subject is resuscitated using Ringer's Lactate infusions. The infusions are performed according to pre-determined closed-loop control laws designed to restore and regulate MAP. See Fig. 3 for hemorrhage and resuscitation profiles received by two example subjects. During each experiment, the HCT, CO, and MAP responses of the subject are measured and recorded as data. These measurements are performed at $\sim$ 5-minute intervals over the course of 180 minutes. These animal experiments are useful for physiological modeling purposes, as (i) they can provide informative physiological measurements that are not commonly available in clinical settings, and (ii) the exact timing and amount of hemorrhages and fluid infusions applied to each subject are known [44], [54], [57].

E. Data Analysis

In the case study on hemorrhage resuscitation modeling, our aim is to illustrate the efficacy of the C-VI method in inferring personalized and generative physiological models from low-information and heterogeneous data. For this purpose, we consider four inference scenarios: (i) C-VI given full data, (ii) C-VI given partial data, (iii) non-collective MLE given full data, and (iv) non-collective MLE given partial data. The full data scenarios were created by presenting all available data to the methods, which allowed us to obtain highly-informed physiological models. The partial data scenarios were created by random exclusion of measured variables from data, which allowed us to evaluate the methods against low-information and heterogeneous data. Further details follow.

C-VI versus Non-Collective MLE: To perform C-VI, we used the (partially or fully) available HCT, CO, and MAP data as inputs to the stochastic objective shown in Algorithm 2 and maximized this objective using Algorithm 1 to obtain the optimized variational parameters $\boldsymbol{\nu}$. According to the approximate posterior formulation in (27), the most-likely personalized physiological model parameters can be obtained from $\boldsymbol{\nu}_{\mu:\boldsymbol{\theta}}$, while their associated uncertainties can be obtained from $\boldsymbol{\nu}_{\sigma:\boldsymbol{\theta}}$. In addition, the most-likely subject generator model parameters can be obtained from $\boldsymbol{\nu}_{\mu:\phi}$, and additional virtual subjects can be generated using the generator in (29) with ϕ set to $\boldsymbol{\nu}_{\mu:\phi}$.

To perform non-collective MLE, we modified the stochastic objective shown in Algorithm 2 by setting the non-likelihood terms (i.e., p_q, $p_{\boldsymbol{\theta}}$, p_ϕ, and $p_{\mathbf{n}}$) to zero. We maximized this objective using Algorithm 1, and obtained the estimates for the personalized physiological model parameters from $\boldsymbol{\nu}_{\mu:\boldsymbol{\theta}}$. To generate virtual subjects in this case, we calculated the mean and standard deviation of the estimated personalized values for each physiological parameter, and generated additional virtual subjects by sampling from a Gaussian distribution with the same mean and standard deviation.

Full Data, Partial Data, and Method Evaluation: To study and compare the effects of information scarcity and data heterogeneity on the fidelity of the personalized models and the generated virtual subjects, we performed inference in two data availability cases. In the full data case, we used the available HCT, CO, and MAP data across all subjects to perform C-VI and non-collective MLE. In the partial data case, we randomly excluded two out of the three measured variables (HCT, CO, and MAP) for each subject. The resulting partial dataset thus consisted of 8 subjects with only HCT measurements, 8 subjects with only CO measurements, and 7 subjects with only MAP measurements. We used this partial dataset to perform inference in both C-VI and non-collective MLE cases. To evaluate the fidelity of the personalized physiological models, we computed the mean-absolute error of the model predictions with respect to the excluded data for each subject. To determine significance in difference between C-VI and non-collective MLE cases, we utilized the Wilcoxon signed-rank test. To evaluate the fidelity of the subject generator models, we computed the likelihood of the excluded data in the generated virtual subjects, which can be written as:

$$p(\mathbf{y}_E|\mathbf{n},\mathbf{u}) = \mathbb{E}_{\theta_i\sim\mathcal{G}}\left[p(\mathbf{y}_E|\theta_i,\mathbf{n},\mathbf{u})\right] \tag{36}$$

where $\mathbf{y}_E$ denotes the space of excluded data, $p(\mathbf{y}_E|\theta_i,\mathbf{n},\mathbf{u})$ denotes the likelihood of the excluded data given a generated subject θ_i, and $p(\mathbf{y}_E|\mathbf{n},\mathbf{u})$ denotes the likelihood of the excluded data under the subject generator $\mathcal{G}$.

V. Results and Discussion

In this section, we present the results of applying the C-VI method presented in Section II to the hemorrhage resuscitation modeling problem presented in Section III, and discuss the effectiveness of C-VI in the context of personalized and generative physiological modeling. Further details follow.

A. Personalized Physiological Modeling

Fig. 3 shows the personalized hemorrhage resuscitation model responses in two representative subjects. The model responses in the rest of the subjects are provided in the supplementary material. These models were obtained by presenting the full dataset to the C-VI method. For each subject in Fig. 3, the most-likely physiological models (shown as solid blue lines) reproduced the trends in measured data for HCT, CO, and MAP. Among the N=23 subjects in the dataset, these physiological models reproduced the trends in measured data with an average mean-absolute error of 0.51 % for HCT, 0.36 L/min for CO, and 7.77 mmHg for MAP. Furthermore, sampling the approximate posterior yielded other likely physiological models that also reproduced the trends in measured data. This is shown in Fig. 3 using shaded blue areas representing the 2σ interval for the likely physiological model responses. In addition, for each subject, the measurement models captured the unexplained variations in measured data for HCT, CO, and MAP. This is shown in Fig. 3 using dashed blue lines representing the 2σ interval for the measurement model responses. Overall, these results suggest that (i) the proposed hemorrhage resuscitation model can adequately reproduce the trends in the physiological data despite its relatively simple structure, and (ii) the C-VI method can infer personalized physiological models that capture the trends in measured data as well as personalized measurement models that capture the extent of unexplained variations in measured data.

Fig. 4 shows the 2σ intervals for the personalized physiological model parameters, which are plotted against the 2σ intervals for the parameters generated by the subject generator. The full set of parameter intervals is provided in the supplementary material. These intervals were obtained by presenting the full dataset to the C-VI method. Naturally, the size of the personalized intervals varied across different physiological model parameters. For some parameters (e.g., the pair shown on the right hand side of Fig. 4), the personalized intervals were smaller, and traveled further into the parameter space. For others (e.g., the pair shown on the left hand side of Fig. 4), the personalized intervals were larger, and resided closer to each other in the parameter space. This phenomenon is a consequence of the C-VI formulation. In case the data provides strong information about a parameter, the likelihood

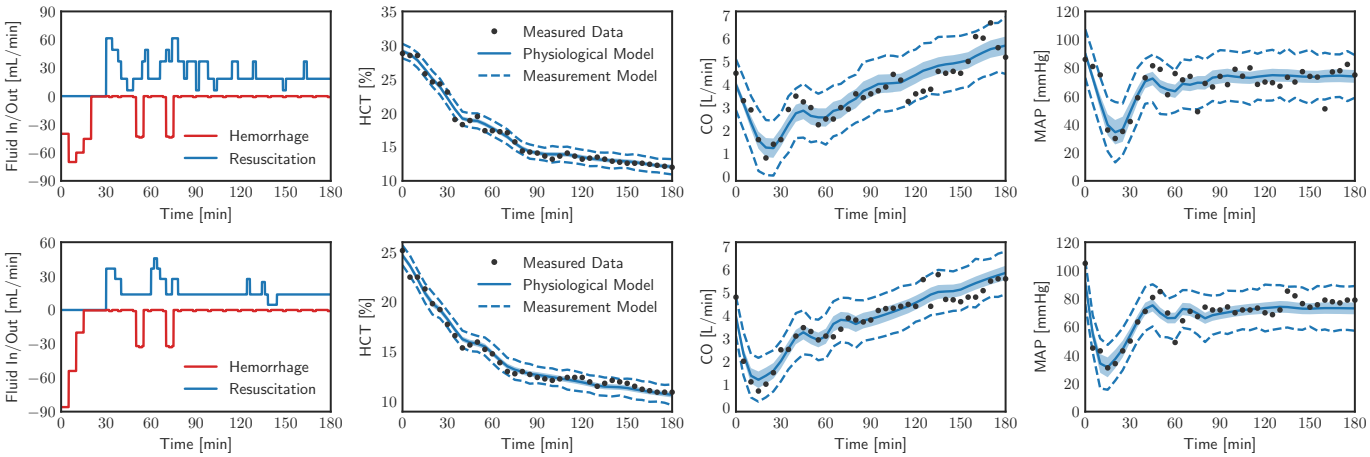

Fig. 3. Personalized physiological model and measurement model responses to hemorrhage and fluid resuscitation for two example subjects. Each row corresponds to a subject. The first column shows the stimuli received by each subject, and the second to fourth columns respectively show hematocrit (HCT), cardiac output (CO), and mean arterial pressure (MAP) responses for each subject against measured physiological data. Shaded blue areas show the $\mathbf{2}\boldsymbol{\sigma}$ intervals associated with personalized physiological model responses, and dashed blue lines show the $\mathbf{2}\boldsymbol{\sigma}$ intervals associated with the measurements generated by the measurement model.

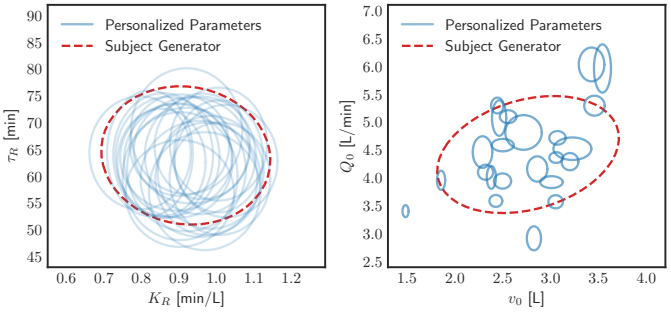

Fig. 4. Personalized physiological model parameters versus the subject generator model for two example parameter pairs associated with weak information (left) and strong information (right) availability. Blue ellipses show $\mathbf{2}\boldsymbol{\sigma}$ intervals for personalized physiological model parameters, and the red ellipse shows the $\mathbf{2}\boldsymbol{\sigma}$ interval associated with the subjects generated by the subject generator model.

term $p(\mathbf{y}|\mathbf{z_\theta}, \mathbf{z_n}, \mathbf{u})$ in (13) dominates the objective, causing the personalized intervals to shrink, move away toward their appropriate values, and alter the shape of the subject generator so that it can reproduce these parameter variations. Conversely, if the data provides weak information about a parameter, the $p(\mathbf{z_\theta}|\mathbf{z}_\phi)$, $p(\mathbf{z}_\phi)$, and $q(\mathbf{z}_\phi, \mathbf{z_\theta}, \mathbf{z_n}|\boldsymbol{\nu})$ terms in (13) dominate the objective, causing the personalized intervals to expand and the subject generator to shrink, gathering the personalized invervals together. Arguably, this provides opportunity for the personalized intervals to share their weak information in order to reach a consensus about appropriate values for the parameter. Overall, these results suggest that the C-VI method can utilize data across multiple experiments to (i) find personalized physiological model parameters for the subjects in a dataset and provide a measure of confidence about their values, and (ii) provide the opportunity to collectively determine plausible values for weakly-informed parameters by aggregating information across different experiments.

Table I shows the mean-absolute prediction errors for the personalized physiological models inferred from partial data. The partial dataset was constructed by randomly excluding

TABLE I

MEAN-ABSOLUTE PREDICTION ERRORS FOR PERSONALIZED MODELS INFERRED FROM PARTIAL DATA [MEDIAN (INTER-QUARTILE RANGE)]

	HCT [%]	CO [L/min]	MAP [mmHg]
Non-Collective MLE	5.10 (6.00)	1.88 (1.65)	35.2 (31.0)
C-VI	3.27 (3.93)	1.04 (1.04)*	9.98 (3.65)*

* Significant with respect to non-collective MLE ($p < 0.05$).

two out of three measured variables (HCT, CO, and MAP) from each subject, simulating a case of low-information and heterogeneous data. The prediction errors (associated with excluded data) were computed against the excluded data for both C-VI and non-collective MLE methods (see Section IV-E for details). As presented in Table I, using C-VI on partial data resulted in lower HCT prediction error, and significantly lower CO and MAP prediction errors when compared to using the non-collective MLE method. This advantage can be attributed to the collective formulation of C-VI, in which the variables form an interconnected tree-like structure (as in Fig. 1(b)). As a result, a loss of observation on some variables (in this case, a random subset of $\mathbf{y}_{ij}$'s) is partially counteracted by the information coming from other observed variables (in this case, information from observed $\mathbf{y}_{ij}$'s travels up to θ_i's, ϕ, and back down to unobserved $\mathbf{y}_{ij}$'s), giving the personalized models superior prediction performance in the face of low-information and heterogeneous data. Overall, these results suggest that the C-VI method can reconcile data from a collection of experiments to produce superior (more predictive) personalized models when compared to a non-collective MLE method, especially when only low-information and heterogeneous data are available.

B. Generative Physiological Modeling

Fig. 5 shows the HCT, CO, and MAP responses of the virtual subjects generated by two subject generators, obtained from applying C-VI to full and partial datasets. In the first row,

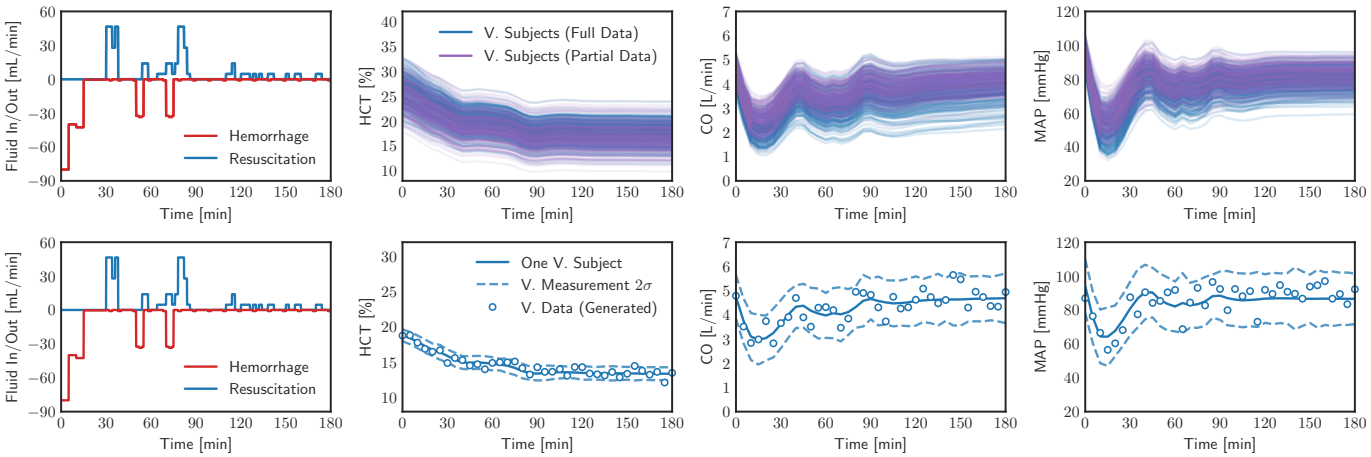

Fig. 5. Generative physiological model responses to hemorrhage and fluid resuscitation: First row shows the physiological model responses of 100 virtual subjects generated by each subject generator model. Second row shows physiological and measurement responses of one example virtual subject. The first column shows the stimuli received by the subjects, and the second to fourth columns respectively show hematocrit (HCT), cardiac output (CO), and mean arterial pressure (MAP) responses.

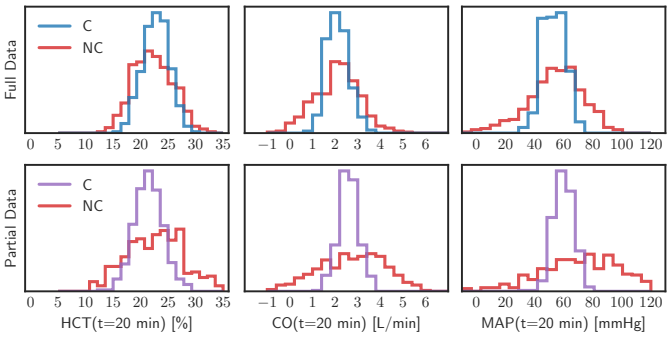

Fig. 6. Histogram plots for post-hemorrhage (at 20 min in Fig. 5) hematocrit (HCT), cardiac output (CO), and mean arterial pressure (MAP) in generated virtual subjects. First row corresponds to generation results given full data, while second row corresponds to generation results given partial data. Blue and purple histograms correspond to the collective (C) approach to generation while red histograms correspond to a non-collective (NC) approach to generation.

TABLE II

NEGATIVE LOG-LIKELIHOOD OF THE EXCLUDED DATA UNDER SUBJECT GENERATORS INFERRED FROM PARTIAL DATA (LOWER IS BETTER)

	Negative Log-Likelihood $[-\log p(\mathbf{y}_E \vert \mathbf{n}, \mathbf{u})]$
Non-Collective MLE	2.09×10^{10}
C-VI	6.06×10^{2}

we show the physiological responses of one hundred virtual subjects to an example hemorrhage resuscitation profile. In the second row, we show an instance of virtual data produced from a virtual subject in response to the same hemorrhage resuscitation profile. From visual inspection, the virtual subjects exhibited a reasonable range of behavior for HCT, CO and MAP, and did not show any objectively unrealistic behavior (e.g., responses that are out of a physiologically meaningful range). In addition, the generated virtual data appeared reasonable and visually similar to data acquired from a hemorrhage resuscitation experiment. In the following paragraph, we aim to analyze and discuss these observations more concretely.

Fig. 6 shows histogram plots of post-hemorrhage (at 20 min in Fig. 5) HCT, CO, and MAP responses in the generated virtual subjects. The first row in Fig. 6 corresponds to virtual subject generation given full data, while the second row corresponds to virtual subject generation given partial data. Blue and purple histograms correspond to generation using the results of C-VI, while red histograms correspond to generation using the results of non-collective MLE (see Section IV-E for details). With full data availability, the non-collective MLE approach resulted in scattered (and occasionally unrealistic) virtual subject responses especially for CO and MAP. With partial data availability, the non-collective MLE approach produced even more scattered responses for HCT, CO, and MAP, including many objectively unrealistic responses (e.g., MAP's above 100 mmHg, and CO's above 5 L/min post-hemorrhage). In contrast, the C-VI approach produced a more realistic range of post-hemorrhage behavior in the case of full data availability, and also retained its generation performance in the case of partial data availability. Furthermore, in the case of partial data availability, the negative log-likelihood of the excluded data (Table II) was lower for the C-VI approach, indicating superior quality for the virtual subjects generated by this method. This can be attributed to the formulation of the C-VI method, which is built to simultaneously infer both the personalized parameters (θ_i's and n_{ij}'s) and the generator parameters (ϕ) in a consistent manner based on all available measured data. In contrast, in the non-collective MLE approach, we generated virtual subjects by mimicking parametric variations across separately inferred personalized parameters, which are sensitive to a lack of information in the data. Overall, these results suggest that the C-VI method can reconcile low-information and heterogeneous data from multiple experiments to obtain a robust generative model of the studied physiological process. This generative model can in turn be used to create new virtual subjects not already in the population in the form of data, but distributed according to the population of real subjects in the dataset.

C. Potential Applications

As discussed, the C-VI method can be used to reconcile low-information and heterogeneous data from a collection of experiments to obtain robust *personalized* and *generative* physiological models. In the following paragraphs, we discuss several potential applications of these results.

In Silico Clinical Trials: In silico clinical trial is an emerging field of research concerned with the use of qualified mathematical/computational models and simulations to perform clinical trials [64]–[67]. As discussed in our case study, C-VI produces a generative physiological model that can generate cohorts of virtual subjects, compute their dynamic responses to given stimuli, and produce virtual data by mimicking specific measurement processes. In addition, C-VI produces personalized physiological models that can serve as "digital twins" of specific subjects in a dataset. Together, these physiological models provide a rich set of representations that may be used in conjunction with recent in silico clinical trial methodologies [64]–[66] to build credible systems and methodologies for systematic testing of physiological monitoring, decision-support, and closed-loop control algorithms.

Monitoring and Control with Limited Measurements: As presented in Introduction, a lack of sufficient information in experimental data is a long-standing challenge in physiological modeling. This challenge is expected to be even more pronounced in clinical settings, where invasive and/or expensive measurements (e.g., HCT and CO) are available only occasionally. This presents a potential use case for the C-VI method in clinical settings. Arguably, in a procedure similar to the partial data results presented in this work, the C-VI method can be used to reconcile limited clinical measurements with past information (e.g., measurements from past patients and/or from past lab experiments) to derive high-quality models of new patients. These results may in turn be leveraged to build model-based algorithms that can monitor and control the physiological states of the patient using limited clinical measurements.

D. Limitations

The presented study has limitations. First, the inference problem addressed is in general a non-convex problem. While stochastic optimization, as employed, is known to provide robust solutions for such problems in practice [51], theoretical guarantees are limited to local optimality [48]–[50]. Second, C-VI is built to produce a subject generator that follows the distribution of the real subjects in the dataset. Therefore, in case the dataset consists of homogeneous subjects and/or subjects biased toward a specific application, this tendency will likely be reflected in the generated virtual subjects. Thus, the resulting virtual subjects are most suited to contexts that are close to that of the original dataset. In our case study, the data were collected in the context of closed-loop control for hemorrhage resuscitation. Therefore, the resulting physiological models may be primarily suited to the design and in silico testing of hemorrhage resuscitation algorithms. In this regard, we expect that additional data gathering and physiological modeling efforts may be needed in order to extend the utility of the presented physiological models to applications beyond hemorrhage resuscitation.

Conclusion

In this paper, we proposed the C-VI method to facilitate the personalized and generative modeling of physiological systems given low-information and heterogeneous data. To illustrate the effectiveness of the C-VI method, we applied it to a practically important case study on modeling the hemodynamic effects of hemorrhage and fluid resuscitation. In the context of this case study, we demonstrated that the C-VI method can reconcile heterogeneous combinations of HCT, CO, and MAP data across multiple experiments to produce robust personalized and generative physiological models. In addition, we demonstrated that the C-VI method produces superior (more predictive) physiological models when compared to a non-collective MLE method, especially when only low-information and heterogeneous data are available. Future efforts should be devoted to the study of approaches that incorporate the collective inference perspective into the design and testing of interpretable physiological monitoring, decision support, and closed-loop control algorithms.

References

[1] E. Brogi *et al.*, "Clinical performance and safety of closed-loop systems: a systematic review and meta-analysis of randomized controlled trials," *Anesthesia & Analgesia*, vol. 124, no. 2, pp. 446–455, 2017.

[2] S. Coeckelenbergh *et al.*, "Automated systems for perioperative goal-directed hemodynamic therapy," *Journal of Anesthesia*, vol. 34, no. 1, pp. 104–114, 2020.

[3] A. Weisman *et al.*, "Effect of artificial pancreas systems on glycaemic control in patients with type 1 diabetes: a systematic review and meta-analysis of outpatient randomised controlled trials," *The Lancet Diabetes & Endocrinology*, vol. 5, no. 7, pp. 501–512, 2017.

[4] L. Pasin *et al.*, "Closed-loop delivery systems versus manually controlled administration of total iv anesthesia: a meta-analysis of randomized clinical trials," *Anesthesia & Analgesia*, vol. 124, no. 2, pp. 456–464, 2017.

[5] C. Zaouter *et al.*, "Autonomous systems in anesthesia: Where do we stand in 2020? a narrative review," *Anesthesia & Analgesia*, vol. 130, no. 5, pp. 1120–1132, 2020.

[6] K. van Heusden *et al.*, "Robust MISO control of propofol-remifentanil anesthesia guided by the NeuroSENSE monitor," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 5, pp. 1758–1770, 2017.

[7] M. Yousefi *et al.*, "Falsified model-invariant safety-preserving control with application to closed-loop anesthesia," *IEEE Transactions on Control Systems Technology*, 2018.

[8] G. Hundeshagen *et al.*, "Closed-loop and decision-assist guided fluid therapy of human hemorrhage," *Critical Care Medicine*, vol. 45, no. 10, p. e1068, 2017.

[9] N. Libert *et al.*, "Performance of closed-loop resuscitation of haemorrhagic shock with fluid alone or in combination with norepinephrine: an experimental study," *Annals of Intensive Care*, vol. 8, no. 1, p. 89, 2018.

[10] X. Jin *et al.*, "Development and in silico evaluation of a model-based closed-loop fluid resuscitation control algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 7, pp. 1905–1914, 2018.

[11] A. Joosten *et al.*, "Feasibility of fully automated hypnosis, analgesia, and fluid management using 2 independent closed-loop systems during major vascular surgery: a pilot study," *Anesthesia & Analgesia*, vol. 128, no. 6, pp. e88–e92, 2019.

[12] J. Salinas *et al.*, "Computerized decision support system improves fluid resuscitation following severe burns: an original study," *Critical Care Medicine*, vol. 39, no. 9, pp. 2031–2038, 2011.

[13] R. Cartotto *et al.*, "Burn state of the science: fluid resuscitation," *Journal of Burn Care & Research*, vol. 38, no. 3, pp. e596–e604, 2017.

[14] A. Joosten *et al.*, "Automated titration of vasopressor infusion using a closed-loop controller: In vivo feasibility study using a swine model," *Anesthesiology: The Journal of the American Society of Anesthesiologists*, vol. 130, no. 3, pp. 394–403, 2019.

[15] F. H. El-Khatib *et al.*, "A bihormonal closed-loop artificial pancreas for type 1 diabetes," *Science Translational Medicine*, vol. 2, no. 27, pp. 27ra27–27ra27, 2010.

[16] F. J. Doyle *et al.*, "Closed-loop artificial pancreas systems: engineering the algorithms," *Diabetes Care*, vol. 37, no. 5, pp. 1191–1197, 2014.

[17] C. K. Boughton and R. Hovorka, "Advances in artificial pancreas systems," *Science Translational Medicine*, vol. 11, no. 484, 2019.

[18] B. Parvinian *et al.*, "Regulatory considerations for physiological closed-loop controlled medical devices used for automated critical care: food and drug administration workshop discussion topics," *Anesthesia and Analgesia*, vol. 126, no. 6, p. 1916, 2018.

[19] B. Parvinian *et al.*, "Credibility evidence for computational patient models used in the development of physiological closed-loop controlled devices for critical care medicine," *Frontiers in Physiology*, vol. 10, p. 220, 2019.

[20] O. Wolkenhauer, "Why model?" *Frontiers in physiology*, vol. 5, p. 21, 2014.

[21] R. E. Baker *et al.*, "Mechanistic models versus machine learning, a fight worth fighting for the biological community?" *Biology Letters*, vol. 14, no. 5, p. 20170660, 2018.

[22] I. Tavassoly *et al.*, "Systems biology primer: the basic methods and approaches," *Essays in Biochemistry*, vol. 62, no. 4, pp. 487–500, 2018.

[23] R. J. Rossi, *Mathematical statistics: an introduction to likelihood based inference*. John Wiley & Sons, 2018.

[24] A. Gelman *et al.*, *Bayesian data analysis*. CRC press, 2013.

[25] D. J. Albers *et al.*, "Mechanistic machine learning: how data assimilation leverages physiologic knowledge using bayesian inference to forecast the future, infer the present, and phenotype," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1392–1401, 2018.

[26] S. Chib and E. Greenberg, "Understanding the metropolis-hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[27] M. Girolami and B. Calderhead, "Riemann manifold langevin and hamiltonian monte carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 2, pp. 123–214, 2011.

[28] T. Toni *et al.*, "Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems," *Journal of the Royal Society Interface*, vol. 6, no. 31, pp. 187–202, 2009.

[29] M. A. Beaumont, "Approximate bayesian computation," *Annual Review of Statistics and Its Application*, vol. 6, pp. 379–403, 2019.

[30] D. M. Blei *et al.*, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[31] C. Zhang *et al.*, "Advances in variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 2008–2026, 2018.

[32] M. D. Hoffman *et al.*, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[33] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[34] C.-H. Zhang, "Compound decision theory and empirical bayes methods," *The Annals of Statistics*, vol. 31, no. 2, pp. 379–390, 2003.

[35] B. Efron, "Bayes, oracle bayes and empirical bayes," *Statistical Science*, vol. 34, no. 2, pp. 177–201, 2019.

[36] M. Davidian and D. M. Giltinan, "Nonlinear models for repeated measurement data: an overview and update," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 8, no. 4, p. 387, 2003.

[37] Y. Wang, "Derivation of various NONMEM estimation methods," *Journal of Pharmacokinetics and pharmacodynamics*, vol. 34, no. 5, pp. 575–593, 2007.

[38] W. Pan *et al.*, "Identification of nonlinear state-space systems from heterogeneous datasets," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 2, pp. 737–747, 2017.

[39] W. R. Jacobs *et al.*, "Sparse bayesian nonlinear system identification using variational inference," *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4172–4187, 2018.

[40] G. Roeder *et al.*, "Efficient amortised bayesian inference for hierarchical and nonlinear dynamical systems," in *International Conference on Machine Learning*, 2019, pp. 4445–4455.

[41] A. Nazabal *et al.*, "Handling incomplete heterogeneous data using VAEs," *Pattern Recognition*, p. 107501, 2020.

[42] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[43] M. J. Johnson *et al.*, "Structured VAEs: Composing probabilistic graphical models and variational autoencoders," *arXiv preprint arXiv:1603.06277*, vol. 2, p. 2016, 2016.

[44] A. Tivay *et al.*, "Practical use of regularization in individualizing a mathematical model of cardiovascular hemodynamics using scarce data," *Frontiers in Physiology*, vol. 11, 2020.

[45] M. K. Transtrum *et al.*, "Why are nonlinear fits to data so challenging?" *Physical review letters*, vol. 104, no. 6, p. 060201, 2010.

[46] M. K. Transtrum *et al.*, "Perspective: Sloppiness and emergent theories in physics, biology, and beyond," *The Journal of chemical physics*, vol. 143, no. 1, p. 07B201_1, 2015.

[47] H. H. Mattingly *et al.*, "Maximizing the information learned from finite data selects a simple model," *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1760–1765, 2018.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[49] J. Duchi *et al.*, "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research*, vol. 12, no. 7, 2011.

[50] A. Défossez *et al.*, "A simple convergence proof of adam and adagrad," *arXiv preprint arXiv:2003.02395*, 2020.

[51] B. Kleinberg *et al.*, "An alternative view: When does SGD escape local minima?" in *International Conference on Machine Learning*. PMLR, 2018, pp. 2698–2707.

[52] V. A. Convertino and N. J. Koons, "The compensatory reserve: potential for accurate individualized goal-directed whole blood resuscitation," *Transfusion*, vol. 60, pp. S150–S157, 2020.

[53] V. A. Convertino *et al.*, "Validating clinical threshold values for a dashboard view of the compensatory reserve measurement for hemorrhage detection," *Journal of Trauma and Acute Care Surgery*, vol. 89, no. 2S, pp. S169–S174, 2020.

[54] R. Bighamian *et al.*, "Control-oriented physiological modeling of hemodynamic responses to blood volume perturbation," *Control Engineering Practice*, vol. 73, pp. 149–160, 2018.

[55] A. Tivay *et al.*, "Virtual patient generation using physiological models through a compressed latent parameterization," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 1335–1340.

[56] R. G. Hahn and D. S. Warner, "Volume kinetics for infusion fluids," *The Journal of the American Society of Anesthesiologists*, vol. 113, no. 2, pp. 470–481, 2010.

[57] R. Bighamian *et al.*, "A lumped-parameter subject-specific model of blood volume response to fluid infusion," *Frontiers in Physiology*, vol. 7, p. 390, 2016.

[58] J. E. Hall and M. E. Hall, *Guyton and Hall textbook of medical physiology*. Elsevier Health Sciences, 2020.

[59] Y. Çınar *et al.*, "Effect of hematocrit on blood pressure via hyperviscosity," *American journal of hypertension*, vol. 12, no. 7, pp. 739–743, 1999.

[60] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems," *SIAM journal on scientific computing*, vol. 14, no. 6, pp. 1487–1503, 1993.

[61] A. D. Rafie *et al.*, "Hypotensive resuscitation of multiple hemorrhages using crystalloid and colloids," *Shock*, vol. 22, no. 3, pp. 262–269, 2004.

[62] S. U. Vaid *et al.*, "Normotensive and hypotensive closed-loop resuscitation using 3.0% nacl to treat multiple hemorrhages in sheep," *Critical Care Medicine*, vol. 34, no. 4, pp. 1185–1192, 2006.

[63] N. R. Marques *et al.*, "Automated closed-loop resuscitation of multiple hemorrhages: a comparison between fuzzy logic and decision table controllers in a sheep model," *Disaster and Military Medicine*, vol. 3, no. 1, pp. 1–10, 2017.

[64] M. Viceconti *et al.*, "Credibility of in silico trial technologies—a theoretical framing," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 4–13, 2019.

[65] M. Viceconti *et al.*, "In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products," *Methods*, 2020.

[66] O. Inan *et al.*, "Digitizing clinical trials," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[67] S. Niederer *et al.*, "Creation and application of virtual patient cohorts of heart models," *Philosophical Transactions of the Royal Society A*, vol. 378, no. 2173, p. 20190558, 2020.