A methodology of phenotyping ICU patients from EHR data: high-fidelity, personalized, and interpretable phenotypes estimation

Yanran Wang[1†], J.N. Stroh[2,3], George Hripcsak[4], Cecilia C. Low Wang[5], Tellen D. Bennett[2], Julia Wrobel[1], Caroline Der Nigoghossian[6], Scott Mueller[7], Jan Claassen[8], and D.J. Albers[2,3,4]

[1] *Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, 13001 East 17th Place, 3rd Floor, Mail Stop B119, Aurora, CO 80045*
[2] *Pediatrics, Section of Informatics and Data Science, University of Colorado, 13123 E. 16th Ave., B065, Aurora, CO 80045*
[3] *Department of Biomedical Engineering, University of Colorado, 12705 East Montview Boulevard, Suite 100, Aurora, CO 80045*
[4] *Biomedical Informatics, Columbia University, 622 W. 168th Street, PH20, New York, NY 10032*
[5] *Division of Endocrinology, Metabolism and Diabetes, Department of Medicine, University of Colorado School of Medicine, 12801 East 17th Avenue, 7103, Aurora, CO 80045*
[6] *Columbia University School of Nursing, 560 West 168th Street, New York, NY 10032*
[7] *Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus, 12850 East Montview Boulevard, Aurora, CO 80045*
[8] *The Neurological Institute of New York, Columbia University Irving Medical Center, 710 West 168th Street, New York, NY 10032*

Abstract

Phenotyping methods aiming to produce high-fidelity time-dependent phenotypes in a specific context with personalized interpretation are challenging, especially given the complexity of physiological systems and data quality. We present a three-stage methodological phenotyping pipeline based on a mechanistic physiological model framework applied to the glucose-insulin system of ICU patients from electronic health record data. Balancing flexibility and biological fidelity, the pipeline within the data assimilation (DA) framework is used to compute physiologically-anchored phenotypes that are high-fidelity, personalized, time-sensitive, and interpretable using data present in real-time ICU settings. We construct the computational phenotyping pipeline such that it is generalizable and we demonstrate the pipeline's accuracy and reliability through external data evaluation using electronic health record data and clinical face validation.

Keywords: phenotyping, data assimilation, physiological modeling, electronic health record, knowledge representation with machine learning, data mining

1 Introduction

Phenotype definitions and quantifications are essential to medicine both clinically and scientifically. In characterizing organisms with quantifiable properties, phenotyping has long been a leverageable conduit for health care providers and researchers to grab quantified insights of patients–and has recently surged with applications for personalizing intervention and treatment of various health conditions. The challenge is to estimate unmeasurable variables representing high-fidelity phenotypes that are physiologically-anchored,

generalizable to larger cohorts, and especially, personalized and interpretable in real time.

Because phenotypes are often defined for practical reasons in a specific context and need to be accurate and robust within that context [1] to hedge inevitable bias, phenotyping methods based on knowledge engineering tend to be work-intensive and time-consuming [2][3], and generally difficult when hampered by inadequate data. The curated data set generation process in these phenotyping methods [1] limits the construction of complex time-sensitive phenotypic models necessary for monitoring a patient's day-to-day health status progression and introduces inevitable bias, complicating generalization to larger cohorts and leading to low-fidelity phenotypes.

Computable phenotyping methods tackle current phenotype estimation problems with consistent efficiency across health care institutions via reproducible queries of data-rich electronic health record (EHR) data, based on latent or otherwise unobservable properties that may be inferred from such observations with data-driven machine learning (ML) employments [4][5][6][7][8][9][10]. State-of-the-art computable phenotyping methods, though, still suffer from data sparsity and quality issues [1][11][12][13] in EHR. Non-stationary EHR data have inherent biases in recording only a limited proportion of physiology from patients [1][12] in the health care process [14] and inconsistent quality from disparate data systems [15]. As such, the complexity of EHR data for ML-based phenotyping will be unreliable and irreproducible for knowledge extraction.

One way of coping with the complexities of EHR data and injecting our known understanding of physiology to maximize our ability to extract information from data is through a data assimilation (DA) framework [16]. Data Assimilation (DA) is a discipline and collection of computational methods that combine data with models, much like machine learning, but where the model usually (but not always) represents our mathematized knowledge of the system with, e.g., a mathematical model of physiology, the hydrodynamics of the atmosphere, etc. Essentially, DA is a Bayesian estimation procedure that uses a mathematical model to assimilate observations into posterior probability densities with either direct or variational methods [17][18][19][20]. In this setting, DA can be used to estimate model initial conditions and states that evolve when either simulated continuously in time, e.g., blood glucose, or estimated at measurement times, e.g., insulin resistance. Additionally, DA can be formulated as a filter, where the estimation and forecasts are computed sequentially in real time as data arrive, or as a smoother where an entire data set over a time window can be used to estimate states, parameters, and initial conditions. DA has been used and developed in a long list of contexts ranging from landing on the moon [21] and aerospace in general [22], to numerical weather prediction [17][19], to robotics [23], but in medicine its application is just beginning [16][24].

Within the DA framework, we incorporate external knowledge via personalizing a high-resolution physiological model on the basis of individual patient data. The mechanistic model in DA generally consists of a set of ordinary differential equations to represent a physiologic process. DA then can estimate parameters of model equations using Markov chain Monte Carlo (MCMC) or other inference techniques. The DA-based workflow reverses a typical ML phenotype computation workflow by first estimating a personalized physiological state for an individual using only their data, producing effectively a continuously defined, time-specific, and individual phenotype. It injects personalized physiological knowledge into computational phenotyping methods via model-based inference, bringing new possibilities for optimal treatment strategies, computation, and future patient state prediction.

We developed and implemented this DA-based phenotyping methodology in the context of a phenotype estimation pipeline. The methodology shown in Figure 1 with a high-level structural overview on the left has three stages, including (i) a DA-based personalized parameter characterization, (ii) an unsupervised ML-based phenotype computation process using DA output, and (iii) a phenotype external validation and analysis process. We exemplified the pipeline implementation by applying the pipeline to the glucose-insulin system of ICU patients as a use case, which can be generalized to other physiological systems. Embedded within our DA

framework is a relatively simple mechanistic physiological model that explains the ultradian oscillation of pulsatile insulin secretion. Finally, this methodology can be used to generate new phenotypic information from ICU data that can plausibly scale to high-throughput, real-time settings.

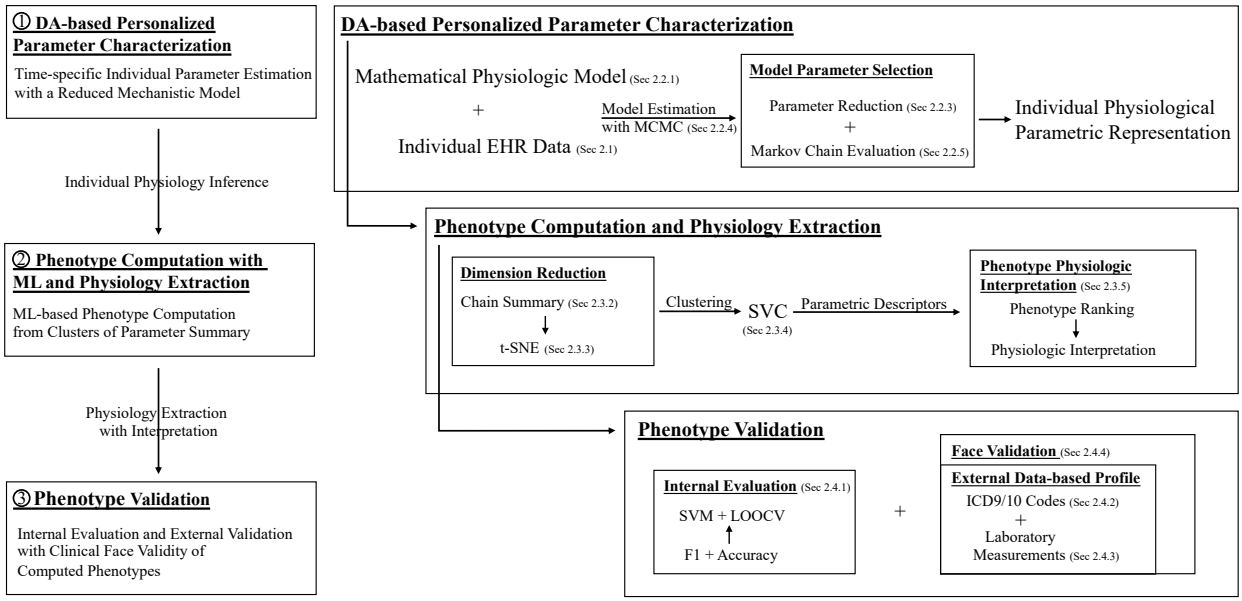


FIGURE 1: The methodological pipeline flow chart with a high-level structural overview on the left, delineated as three main stages: (i) a DA-based personalized parameter characterization, (ii) an unsupervised ML-based phenotype computation process using DA output, and (iii) a phenotype external validation and analysis process. Synchronizing with EHR data targeting the glucose-insulin system in the ICU setting, the first stage computes continuous personal phenotypes based on a mathematical mechanistic model–an a priori hypothesis of how the glucose-insulin physiological system behaves under known nutrition and exogenous insulin therapies–with parameters estimated from Bayesian inference through a nonlinear likelihood via MCMC. The second stage discretizes the continuous parametric representation of physiological states of all patients into computationally-actionable physiologically-anchored mechanistic categorical phenotypes using ML. The third stage quantitatively evaluates these categorical phenotypes for predictive power, summarizes clinically observable differences among phenotypes using ICD 9/10 codes, quantitatively validates model-based phenotypes against laboratory measurements not used in phenotype computation with statistical analysis, and checks the qualitative mapping of physiology estimated by the model to practical clinical meanings that are further externally validated through face validation by an endocrinologist.

Our DA-driven dynamical modeling phenotyping pipeline is fundamentally different from associative data-mining based digital phenotypic systems. We focus on extracting underlying physiological mechanistic phenotypes, e.g., implicitly quantifiable rate constants such as insulin secretion rate and clearance, that are different from traditional phenotypes in informatics, while estimation of these constants are bounded by chemical or biochemical laws and processes. We delineate in sequence the core contributions and superiority of our proposed method to highlight the divergence at a high level:

1. Our DA-driven approach embeds deep physiological knowledge with a mechanistic model and leverages the chemistry/biochemistry constraints in the mechanistic model structure to manage individual EHR data sparsity towards the construction of interpretable and time-specific phenotypes.

2. Our phenotyping pipeline infers individual physiology from personalized model parameter estimation and then cluster to compute new population-scale physiological phenotypic information that is validated to have potential clinical utility. Existing efforts in temporal phenotyping methods with longitudinal EHR leverages conventional ML, deep learning, and statistical analysis to identify temporal process patterns and phenotypes from patients' population strata using mining-like algorithms to cluster phenotypic variables [25][26][27][28] or graphical-based models [29].

2 Methods

The three main stages in the methodological pipeline described above require data (Sec. 2.1) and comprise an individual physiological parametric characterization process with a mechanistic model (Sec. 2.2), a phenotype computation process (Sec. 2.3), and a phenotype evaluation and validation process (Sec. 2.4). Figure 1 charts the pipeline with relevant headings of the text below.

2.1 Data

2.1.1 Data Set Construction

We use two ICU patient data sets in this work from two institutions, UCHealth Data Compass and Columbia University, both approved by the respective IRBs (Columbia University IRB AAAJ4503(M01Y06) and Colorado Multiple IRB 18-2519, respectively). To define experiment intervals in this paper for patient-level estimation from the EHR data originally identified by patient encounters, we identify two types of intervals: (i) ICU periods based on the recorded level of care, and (ii) regulated nutrition intervals based on the presence of a feeding tube with enteral nutrition (EN). Patient glycemic states estimation occurs during an ICU stay, defined by a contiguous ICU period longer than three days with EN. We also merge ICU stays when an enteral feeding tube is absent for less than 24 hours. For the remainder of this paper, we work with ICU stay ID as the identifier.

Table 1 details our two cohorts of ICU stays. We use cohort A from Columbia University (N=9) for model parameter selection (Sec. 2.2.3) and method configuration (Sec. 2.2.5) [30]. Cohort B (N=109) is used for phenotype computation (Sec. 2.3) and external validation (Sec. 2.4). We initially exclude pregnant patients from both cohorts. We identify cohort A by further excluding type 1 diabetes mellitus (T1DM) patients. To balance both the uncertainty in nutrition/insulin therapy and cohort breadth, the additional data inclusion criterion for cohort B is confining the ICU stay duration over 3 days with exogenous short-acting insulin only. Under this additional inclusion criteria, we identify cohort B (around 7% are T1DM) from a larger cohort with around 33.5% ICU stays (5157/15395) receiving exogenous insulin therapies.

We use the following longitudinal data of the two cohorts from patient EHR records: administered TPN via continuous EN, non-bolus IV, and bolus IV deliveries; exogenous insulin therapies with continuous intravenous infusions, bolus intravenous injections, and subcutaneous injections; laboratory measurements synchronized with blood glucose measurement monitoring the general endocrine system state. Glucose measurements are more frequent (every 1-2 hours) when patients receive continuous infusion insulin therapies rather than subcutaneous injections. From these patient EHRs, we also use information about patient history to aid in validating computed phenotypes (Sec. 2.4.2 & Sec. 2.4.3).

Data Summary Table			
Summary	Cohort A	Cohort B	Cohort B3
Data Resource	Manual Collection from Columbia University Irving Medical Center	Automatic Extraction from University of Colorado Health Data Compass Clinical Data Warehouse Between 2010–2019	Same as Cohort B
# Patients	9	109	45
# ICU Stays	9	109	45
Inclusion Criteria	Non-pregnancy; Non-T1DM ; ICU Level of Care	Non-pregnancy; ICU Stay Duration in [3, 60] Days; Exogenous Insulin Only; Short-acting Insulin	Same as Cohort B
Stay Length (Days) Mean±SD Median [Min, Max]	27±20 22 [13,78]	8±6 6 [3,36]	7±3 6 [3,19]
Exogenous Insulin	Continuous Intravenous Infusion; Bolus Intravenous Injection; Subcutaneous Injection (Probably Short-acting Only, But Uncertain)	Short-acting (Continuous Intravenous Infusion; Bolus Intravenous Injection; Subcutaneous Injection)	Same as Cohort B
# Insulin Delivery Mean±SD Median [Min, Max]	39±36 30 [4,125]	55±73 31 [1,387]	104±87 75 [5,387]
Nutrition	TPN	TPN, No TPN	TPN, No TPN
Race		83 White/Caucasian; 12 Black/African American; 4 Asian; 1 Multiple Race; 9 others	37 White/Caucasian; 4 Black/African American; 1 Asian; 3 others
Ethnicity		96 Non-Hispanic; 13 Hispanic	41 Non-Hispanic; 4 Hispanic
Age Mean±SD Median [Min, Max]		57±14 57 [25,90]	64±13 65 [30,87]
Gender		73 M; 36 F	30 M; 15 F

TABLE 1: The nutrition and insulin therapy and demographic summary of ICU data sets cohort A, B, and B3 used in this work, including data resources, patients counts, ICU stays counts, data set inclusion criteria, ICU stay length in days, types of exogenous insulin therapy, insulin delivery counts, types of nutrition therapy, race, ethnicity, age, and gender. Neither CGM data nor insulin measurement was included, as is the general situation in an ICU setting. Cohort A comprises of nine manually-curated non-T1DM patients from the Columbia University Irving Medical Center Neurological Intensive Care Unit. Most patients required exogenous insulin therapies delivered through continuous intravenous infusions (mostly), bolus intravenous injections, and subcutaneous injections. Cohort B is a cohort of 109 patients receiving short-acting insulin under controlled EN with only brand-name nutrition recorded from the University of Colorado Health Data Compass Clinical Data Warehouse. Short-acting insulin was typically administered as a continuous intravenous infusion (5953 in total) and a limited portion is subcutaneous injection (6 in total). In each ICU stay, only measurements after the first non-zero nutrition were used to support this paper. From cohort B, cohort B3 comprises 45 ICU stays that have sufficient data in the last three-day period. Noting here that this table summarizes the complete ICU stays in cohort B3, not only for the last three-day period. A large portion of patients in cohort B3 was possibly on hemodialysis treatment due to acute kidney failure or chronic kidney disease. Glucose measurements and nutrition insulin therapy data were used to compute the model-based phenotypes from the pipeline. Other laboratory measurements and ICD 9/10 codes are used in external validation only.

2.1.2 Time Window Partitioning to Balance Non-stationarity versus Data Sparsity

The word stationarity in our phenotypic context means that a patient's underlying physiological state is not changing or moving. In fact, all patients are non-stationary, because a patient's physiological state–the phenotype–changes very fast in ICU as a result of both internally cooperative mechanism in human body systems and externally therapeutic interventions from health care providers.

Extracting a phenotype with DA, though, assumes a stationary patient state. Under the DA framework, forecasts of fast-moving variables, e.g., glucose, are made assuming the slow-moving physiological states defined by model parameters, e.g., insulin secretion, is stationary. At the same time, computing phenotype via DA estimation of slow-moving parameters becomes more accurate as the amount of data for computation is increased. Our temporal computable phenotyping method, though, is still limited by data sparsity issue self-imposed by the individual dynamical modeling approach that requires joint data of glucose measurements, administration of insulin, and nutrition for one individual patient.

To balance non-stationarity effects with data sparsity, we estimate the phenotypes on a time window that is the longest possible to give robust inference to maximize stationarity and also the shortest possible with sufficient data to produce an acceptably low phenotype computation error, conceptually managing a version of the variance-bias trade-off. This balance can also be described as temporal resolution vs. sufficient statistical representation. From cohort B we pick a sub-cohort in the window of the last three days (cohort B3) as our focus to establish the clinical context of our model-derived phenotypes for two main reasons. Firstly, this cohort is closer to stationarity with a well-defined end point that simply aligns/registers patients by enteral feeding tubes removal time based on either healthy or dying conditions. Secondly, this cohort has a duration of three days in the end with sufficient data required by DA to balance non-stationarity effects with data sparsity. Specifically, we have 45 ICU stays associated with 45 patients in cohort B3 in the last three-day period. Other sub-cohorts with a shorter time window could be selected as patients are more dynamic in the early or middle of ICU stay than at the end. For completeness we also partition ICU stays from cohort B into two other sub-cohorts that have sufficient data in the window of the first (cohort B1) and middle (cohort B2) three days. We start with 109 patients/ICU stays in cohort B but end with a different number of ICU stays for these sub-cohorts due to model estimation criterion and data availability.

2.2 Physiological Model Parameter Estimation to Characterize Individual ICU Stay

Here we will introduce the data assimilation (DA) computational framework that includes the physiological model and the model estimation methods that characterize patient physiological state as individual continuous model parameter estimates under qualification criteria.

2.2.1 Mechanistic Model Representing Our Understanding of Physiological Functions

A mechanistic model reflects essential features or properties of reality relevant to the study at hand. To this end, we use the parsimonious Ultradian model [31] to extract mechanistic knowledge from observed physiology to gain deep physiological insight. This relatively simple and widely validated mechanistic model provides an unambiguous quantitative representation of the ultradian glycemic oscillation when patients are receiving enteral nutrition [32]. Though only approximate, this model characterizes the essential biological fidelity at a multi-day scale. This allows for estimation of physiology that is not directly measured or measurable but can sometimes be implicitly observed by clinicians; for example, the model estimation with a patient's data can characterize insulin secretion. A schematic diagram of this Ultradian model is shown in Figure 2, with all parameters interposed in all channels between the three main pools. A reader interested in technical details of the model mathematical formulas is referred to the Appendix (Sec. A.1).

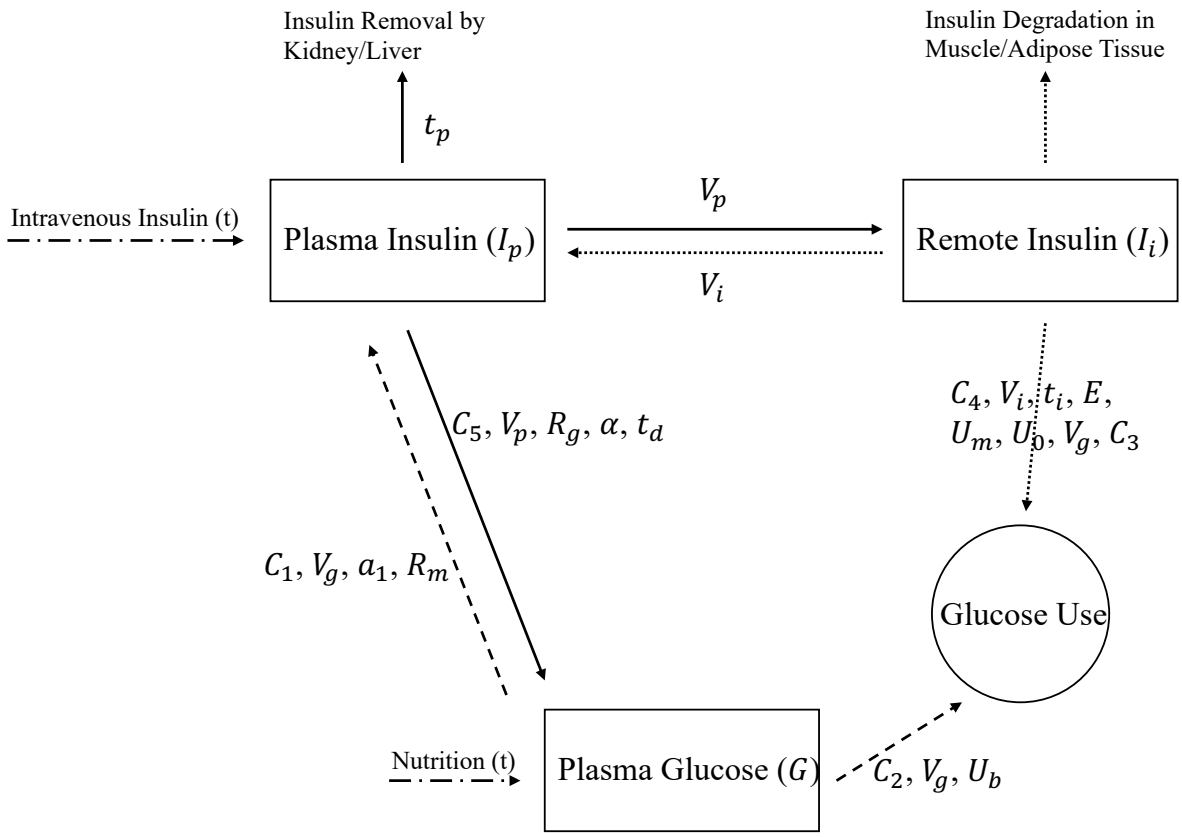


FIGURE 2: The Ultradian model diagram. The glucose-insulin system consists of three main pools: the plasma insulin, the remote insulin in the intercellular space, and the plasma glucose. This non-autonomous Ultradian model is represented by a system of ODEs with 6 states and 21 parameters. Exponential constants a_1 and C_1 affect insulin secretion; t_p is the time constant for plasma insulin degradation by the kidney and liver; R_g is the linear constant affecting insulin-dependent glucose production rate; C_3 is the linear constant affecting remote insulin-dependent glucose removal rate in an implicit delayed process.

2.2.2 Solving Inverse Problems with Data Assimilation (DA)

Often DA is the case of inverse problems, or the process of estimating variables believed to be causal, as represented by a mathematical model of the underlying mechanics, from data. Here we use DA within the inverse problems context in the smoother formulation. In this paper, we posit one model, the nonlinear mechanistic Ultradian model, to represent our understanding of glucose-insulin physiology to limit the analysis for a robust solution such that we incorporate external glycemic knowledge via estimating personalized glycemic states with individual data. Working in the DA smoothing context, we optimize model parameters by minimizing the prediction cost in glucose measurements in L_2 norm (least squares minimization) under the assumption of stationary physiological states represented by constant model parameters within a short time. We enforce the stationarity assumption by smoothing over a minimum time window length for which we have sufficient data to estimate model parameters. The minimization is computed with Markov Chain Monte Carlo (MCMC) Method.

2.2.3 Model Parameter Selection to Maximize Parameter Identifiability

For clarification purpose, herein we note that model selection includes both the process of picking a mechanistic model and model parameter selection that is the process of identifying and estimating a reduced list of model parameters from the model we chose. In Sec. 2.2.1, we pick the Ultradian model to represent the endocrine glucose-insulin physiological process, as it is currently the best model to explain the oscillation of pulsatile insulin secretion at a multi-day scale. Our phenotyping pipeline extracts underlying physiological mechanistic phenotypes through estimating model parameters. However, the task of estimating all 21 parameters in the Ultradian model [31] will lead to identifiability issues, e.g., no guarantee of the existence and uniqueness of a global minimum of forecasting error as the model solution, especially when data is sparse [33] to maximize stationarity. In order to overcome this challenge, we identify and estimate only a subset of model parameters for model inversion to get a unique inference solution at the expense of physiologic fidelity, namely the model parameter selection. We select Ultradian model parameters from three perspectives: the Houlihan method, our ability to estimate patients from a perspective of bias-variance trade off with our understanding of the mechanistic structure of the model, and practical clinical needs. The scalable optimization algorithm, parameter Houlihan [34], reduces the model to a selection of parameter nominees. Noting that the Houlihan method is a rank-order algorithm that perturbs and picks the most influential model parameter/sets that affects the mean and variation of glucose level. The ML-based Houlihan method is also a combination of generalized Lasso in L1 regularization, L2 regularization, and principal component analysis (PCA). Therefore, this method is necessary but not sufficient as it might pick parameters that are not meaningful/interesting to our research and clinicians. Besides using the Houlihan method, we also isolate the model parameter selection process to our non-clinical understanding of how each parameter impacts the evolution of fast-moving variables based on the Ultradian model ODEs, managing a balance between fidelity and forecasting error to maximize identifiability and accuracy. This model reduction step is also where clinicians can make decisions to inject expert knowledge by choosing clinically interesting parameters despite the suggestions given by the Houlihan method.

Using cohort A, we adopt a triplet of parameters (a_1, t_p, R_g) identified by the Houlihan method as the reduced model that minimizes mean squared error (MSE). A reader interested in technical details surrounding the model parameter selection with the Houlihan method is referred to the Appendix (Sec. A.2). This reduced model keeps the correspondence of model parameters with actual physiological system components because it is related to the hepatic-renal-insulin subsystem. The adopted parameter triplet is associated with three important endocrine physiology descriptors that are implicitly quantifiable or unmeasurable/clinically unmeasurable: insulin insensitivity (R_g), pancreatic insulin secretion (a_1), and insulin degradation via liver and kidney (t_p). Because the model selection process is based on personalized estimation with individual data, the calibration experiments above conducted on the small cohort A can transfer to cohort B. As such, we adopt and estimate the same parameters triplet in cohort B for the rest of this paper.

2.2.4 Markov Chain Monte Carlo (MCMC) Methods for Model Parameter Estimation

Given the mechanistic equation of the physiological model, we use Markov Chain Monte Carlo (MCMC) with multiple Markov chains to estimate model parameters with uncertainty quantification (UQ) that reflects clinical reality, where each chain comprises the sampling iterations of MCMC parameter estimates. Herein we adopt a Metropolis-Hastings-within-Gibbs MCMC method, which is the Metropolis-Hastings algorithm incorporating a Gibbs sampler that updates parameters one at a time [35]. This method, while less efficient than other MCMC methods, excels in cases where the full conditional probability cannot be determined analytically [36][37][38], which is the case when estimating parameters of a complex physiological model. For each ICU stay, we generate 10 Markov chains with different initial conditions, each with 20K iterations.

For informatics researchers who are unfamiliar with the MCMC method, we note here that a Markov chain in the current context is a stochastic process with continuous parametric physiological state space indexed by

sampling time. We characterize each individual ICU stay as 10 Markov chains of model parameter estimates, where each Markov chain is a sampled time series with 20K points.

2.2.5 Markov Chain Evaluation of Estimation Accuracy and Convergence

Theoretically, MCMC methods guarantee Markov chains iterations to asymptotically arrive at a global minimum. In practice, these asymptotic properties and ergodicity [39] hinge on infinite computing resources that are impractical for several reasons. Under the computing power limitation, the non-linear nature of MCMC without convexity prevents us from guaranteeing a unique solution to represent a global error minimizer for model parameter estimation. Additionally, chains may fail to converge due to using a misspecified reduced model [34][36]. For these reasons, it is prudent to screen chains for both prediction accuracy and convergence, especially when using Metropolis-Hastings-within-Gibbs method to identify global minima for highly correlated variables–the estimated model parameters in the current context. Therefore, we construct a set of objective quality criteria to select from multiple Markov chains that have plausibly converged to global minima.

Within multiple chains for a single ICU stay, we assess two types of chain convergence to determine inclusion: intra-chain and across-chain convergence. Intra-chain convergence is the convergence of a single Markov chain to a stationary distribution. This convergence is assessed via the Geweke statistic comparing the initial and terminal fractions of Markov chain iterations [40] with spectral density estimation to detect periodical behaviors. In nature, the beginning and the middle iterations in each Markov chain are abundant in variation of 'drifts' as the chain typically starts somewhere far from the truth. This is how Geweke statistic differentiates each Markov chain based on convergence status at the end of iterations. In fact, Geweke statistic takes advantage of the 'drifts' between the end and the beginning iterations to identify converged Markov chains that stayed around equilibrium towards the end without 'drifts' among all three parameters simultaneously. Across-chain convergence is more complex. Markov chains independent by nature may converge to different solutions, indicating coexisting phenotypes for a single ICU stay; this case is distinct from those Markov chains that fail to converge. The Gelman-Rubin (G-R) statistic measures whether all Markov chains within a single ICU stay converge to a common solution, indicated as across-chain convergence. Chains burn-in–the beginning iterations to be thrown away–for G-R statistic is set to be the first 80 percent of the total iterations [41].

When implementing the pipeline, we choose for analysis all Markov chains that not only have converged based on Geweke statistic, but also have low mean squared error (MSE) with solutions close to the optimal MSE (+/- 2SD) of any chain to ensure parameter estimates represent the data. Across-chain convergence based on G-R statistic within a single ICU stay to a common solution was used for diagnosis only, as small G-R statistic does not exist under practical consideration without a well-constructed scenario, and model flexibility permitted non-unique inference solutions. As a default configuration in the implementation, the Geweke fractions are identified as the initial 3% and terminal 50% of iterates, and chains significant in Geweke statistic ($\alpha = 0.06$) are considered to have converged. Further technical description of the Geweke fractions identification is provided in Appendix (Sec. A.3). Under these qualification criteria, we will compress all selected Markov chains estimated for individual ICU stay to form individual parametric phenotype summaries and further cluster a cohort of summaries into phenotypes.

2.3 Phenotype Computation and Physiological Knowledge Extraction from Parameter Estimates

In Sec. 2.2 we presented how the patient's physiological state is encoded into continuous model parameters qualified by data-optimized MCMC estimates. These estimates represent continuous time, individual phenotypes or phenotypic information. To compute discrete phenotypes that represent population-scale

categories from these estimates, we describe in this subsection a three-step process for extracting computationally-actionable mechanistic categorical phenotypes.

2.3.1 Representing Parameter Estimates as Empirical Distribution Functions (EDF) Defined by Markov Chains

We keep only the last 20% sampling iterations of each Markov chain with high probability region to represent physiology. In this way, we only kept the variation around the equilibrium of each Markov chain, which excluded the variation of the 'drifts' in the full chain. This process also generates an empirical distribution function (EDF) of Markov chain to represent parameter estimation of individually modeled physiology. Each ICU stay is represented by parameter EDFs of selected Markov chains that we will use to extract individual parametric phenotype summaries.

2.3.2 Summarizing Parameter Empirical Distribution Functions (EDF) for Clustering

To find shared phenotypic information among different patients, we cluster a cohort of N patients into different groups from individual EDFs that estimates patient physiology. From the previous step, now each patient has a dimension of (counts of selected Markov chains) * 4K iterations of each chain * 3 parameters, which poses a problem for later cohort clustering as each data point in the 4K iterations can be identified as a unique cluster, leading to render too many/few clusters in a cohort. Clustering of these individual EDFs to compute phenotypes necessitates dimension reduction, especially when N is large. Therefore, we summarize each individual EDF as a parametric 5D vector with five statistical indices–the position and the spread of the EDF–that can appropriately compress the function for each parameter separately: mean, standard deviation (std), interquartile range (IQR), naive standard error of the mean, and the time-series standard error based on estimation of spectral density at 0. Each patient now has a dimension of (counts of selected chains) * 5 indexes * 3 parameters. We pick these five indices because each parameter has its own encoded physiologic functionalities and we want to fully capture all properties–the 'shape'–of different parameter EDFs that continuously represent the functionalities. For example, median is a good summary statistic only when the EDF is not 'flat'/'uniform', so only using the median is not comprehensive enough to include every possible scenario.

2.3.3 Dimension Reduction with t-Distributed Stochastic Neighbor Embedding (t-SNE)

One current challenge still existing in mostly clustering methods is determining the number of clusters. It is a common and standard approach to reduce dimension before clustering [42] to help identifying the number of clusters to improve later clustering efficiency, e.g., RNA clustering methods work better by calculating distance on a lower dimension [43]. Moreover, tuning hyper parameters for clustering methods is easier when we can visualize the data points in high dimensions.

The t-Distributed Stochastic Neighbor Embedding (t-SNE) method is a nonlinear unsupervised technique for visualizing and projecting high-dimensional data into a low-dimensional space while retaining distributional features such as local similarity and global structure [44][45]. Compared with multidimensional scaling (MDS) methods, t-SNE is also a nonlinear dimension reduction technique while it does not involve eigen-analysis that MDS is usually based on. We use the t-SNE method before clustering patients to serve two purposes here: (i) to further reduce the dimension of individual EDF summaries as the next clustering step maps points to an even higher dimension, and (ii) to geometrically represent the cohort in 2D based on an abstract notion of the distance between summaries. Here we use t-SNE mainly to reduce dimension from 5D–for each parameter in the (a_1, t_p, R_g) triplet–to 2D to improve later clustering efficiency, as the t-SNE method is proven to preserve clusters in the original space [45]. This dimension reduction step keeps one

dimension unchanged, meaning that the total dimension changes from (counts of selected chains) * 15D to (counts of selected chains) * 2D t-SNE vector. The other supplemental goal of using t-SNE is to visualize the reduced dimension as the t-SNE space is generally in 2D or 3D. Although t-SNE in 3D could potentially help reduce cluster intersections by identifying more patient clusters with less reduced dimensions, we choose 2D over 3D because (i) the t-SNE 2D coordinates of summaries given proper tunable parameters retain the properties necessary for cluster analysis [46], and (ii) a visualization 'rules of thumb' [47] that involves considering the added information from third spatial dimension that has costs of depth cues including stereoscopic disparity and occlusion. Noting here that the t-SNE 2D axes are arbitrary and meaningless.

2.3.4 Phenotype Computation by Clustering t-SNE Coordinates with Support Vector Clustering (SVC)

Phenotypes in a cohort of ICU stays are computed by clustering nearby 2D t-SNE coordinates. Following the t-SNE 2D projection, we use a subsequent clustering method support vector clustering (SVC), for grouping stays and patients. This unsupervised learning kernel algorithm identifies regions of nearby t-SNE coordinates–possibly belonging to different patients–most efficiently bounded by a set of small circles through mapping to a feature space in high dimensions [48]. Noting here that this SVC step does not change the dimension. Then each cluster identified by SVC determines a discretized categorical phenotype with a characterizable continuous parametric representation of physiological states. This feature identification is implemented numerically in Matlab [49][50] using the statistical pattern recognition toolbox (stprtool) [51].

To generate phenotypes for representing self-consistent patients, we then identify coherent groups of patients from SVC-identified clusters. We define that SVC identifies coherent clusters when t-SNE coordinates representing the same ICU stay at the same time are clustered into only one phenotype. We do not always identify coherent clusters from SVC, though, as phenotypes are not necessarily mutually exclusive. A patient may be associated with multiple SVC-identified clusters and further exist in more than one phenotype at the end of the pipeline, which complicates potential actionable support for clinical decision-making.

2.3.5 Extracting Physiological Definitions of Computed Phenotypes from Model Parameter Descriptors

We extract the physiologic meanings of the clusters in the following four steps to understand the SVC-identified clusters on the discretized parameter space without prior group information of physiology. First, we generate parametric representations of population-scale phenotypes by concatenating the empirical distribution functions (EDFs) within each SVC-identified cluster. Second, for each phenotype we extract the physiologic meanings using the unique physiologic descriptors of the parameter triplet (a_1, t_p, R_g) from the Ultradian model. In the third step, to address the issues of SVC identifying incoherent clusters of patients, we identify ICU stays associated with multiple SVC clusters as auxiliary groups and separate these ICU stays from previous SVC-identified groups. To produce self-consistent phenotypes, we then expand the SVC-identified groups with these auxiliary groups and drop groups with too few patients considering statistical power. Fourth, we delineate phenotypic patterns of model-derived coherent phenotypes using a customized, relative ranking. Specifically, we tabulate the model-quantified physiology by ranking each coherent phenotype relative to an average ICU phenotype. Each categorical population-scale coherent phenotype with physiology extracted is computationally-actionable and potentially viable for clinical mapping. Noting here that we only analyze and evaluate the phenotypes computed in the last three-day period, although we also compute the phenotypes in the first and middle three days period as well for completeness and to show readers what phenotypes exist at different time periods.

2.4 Validating Model-derived Phenotypes Against Laboratory Measurements, ICD 9/10 Data, and Clinical Notes

Model-based phenotypes are potentially clinically valuable when they can be mapped to practical clinical meanings. To provide potential clinical decision support as guidance for ontology, we quantitatively evaluate the phenotype computation process with statistical analysis and externally validate the model-based phenotypes both quantitatively and qualitatively against external data not used to estimate the model.

2.4.1 Statistical Evaluation of Phenotypes by Their Predictive Power

We access the predictive power of phenotypes computation to quantitatively evaluate the reliability of model-derived phenotypes. We calculate prediction performance using the support vector machine (SVM) algorithm that classifies features of each EDF labeled with identified phenotypes. The SVM method normally requires specifying a few configuration parameters. We use leave-one-out-cross-validation (LOOCV) to tune these parameters for robust prediction, where the selected SVM parameters reach the highest average accuracy for multi-class phenotype labels prediction. Based on this SVM configuration to predict initial phenotypes, we report the F1-score and average accuracy to conclude the phenotype computation predictive power.

2.4.2 Summarizing Phenotypic Difference Using ICD 9/10 Codes

To find evidence of the clinically observed differences between phenotypes, we analyze the correspondence between the extracted physiology of model-derived phenotypes and specific ICD 9/10 codes directly related to the organ functionalities encoded by patient parameters.

2.4.3 External Validation of Phenotype Distinctiveness by Statistical Testing Against Laboratory Measurements Excluded From Phenotype Computation

We statistically test the phenotypes consistency with external data of laboratory measurements to validate model-derived phenotypes against observed patient physiology. From EHR data we investigate a list of laboratory values that indicate problems with organs whose functionality are encoded in model parameters. Then we extract listed laboratory values from patients associated with each model-derived phenotype. We statistically test laboratory mean values among all phenotypes with a Welch two-sample t-test to conclude quantitative difference between model-derived phenotypes.

2.4.4 External Validation of Phenotypes with Clinical Face Validity

In general, face validation is a sanity check based on experts' opinions that provides a crude confidence quantification of the sensibility and relevance of the modeling results, including evaluation of methods and conclusions.

Evaluation and validation of computed phenotypes in this section are qualitative. To further externally validate that the phenotypes have well-resolved ICU stays correctly categorized, we use face validation to confirm whether the phenotypes accurately represent the patient from an endocrinologist's opinion based on qualitative chart review of patient charts. We check the computer-to-clinical phenotypes mapping at the model parameters level instead of at the level of phenotypes, meaning that we qualitatively evaluate the physiology estimated by the model used to create the phenotypes. Note that the qualitatively evaluated physiology is not or cannot be directly measured, but can be inferred indirectly by clinicians. Hence, we ask the clinician through a blind face validation to categorize each of the three physiology descriptors of the parameter triplet (a_1, t_p, R_g)

into three levels from low, medium, and high for each patient. Recall that for this parameter triplet, phenotypes are associated with three descriptors of the modeled kidney-liver-pancreas subsystem: insulin insensitivity (R_g), pancreatic insulin secretion (a_1), and insulin degradation via liver and kidney (t_p).

3 Results

3.1 Physiological Model Parameter Estimation to Characterize Individual ICU Stay

Within the DA framework, we used empirical distribution functions of selected Markov chains under qualification criteria to parametrically represent individual physiological states of patients associated with ICU stays in a cohort with acceptable accuracy. Recall that we estimated each patient using 10 Markov chains of the parameter triplet (a_1, t_p, R_g) with 20K iterations in each chain. At this step, the dimension of each Markov chain was 20K * 3 and we had a total dimension of 10 * 20K * 3 for each patient. Then we used Geweke statistic and MSE as criteria to select qualified chains out of 10 chains. Figure 3 illustrates the Markov chains selection (cf. Sec. 2.2.5) applied to one ICU stay from cohort A as an example. There was consistent MSE among the chains, particularly for those with low MSE. Also, given the narrow bandwidth, every chain had at least one convergent parameter (indicated in blue) based on fraction-identified Geweke statistics. For this ICU stay, we selected the seven chains indicated in black outline.

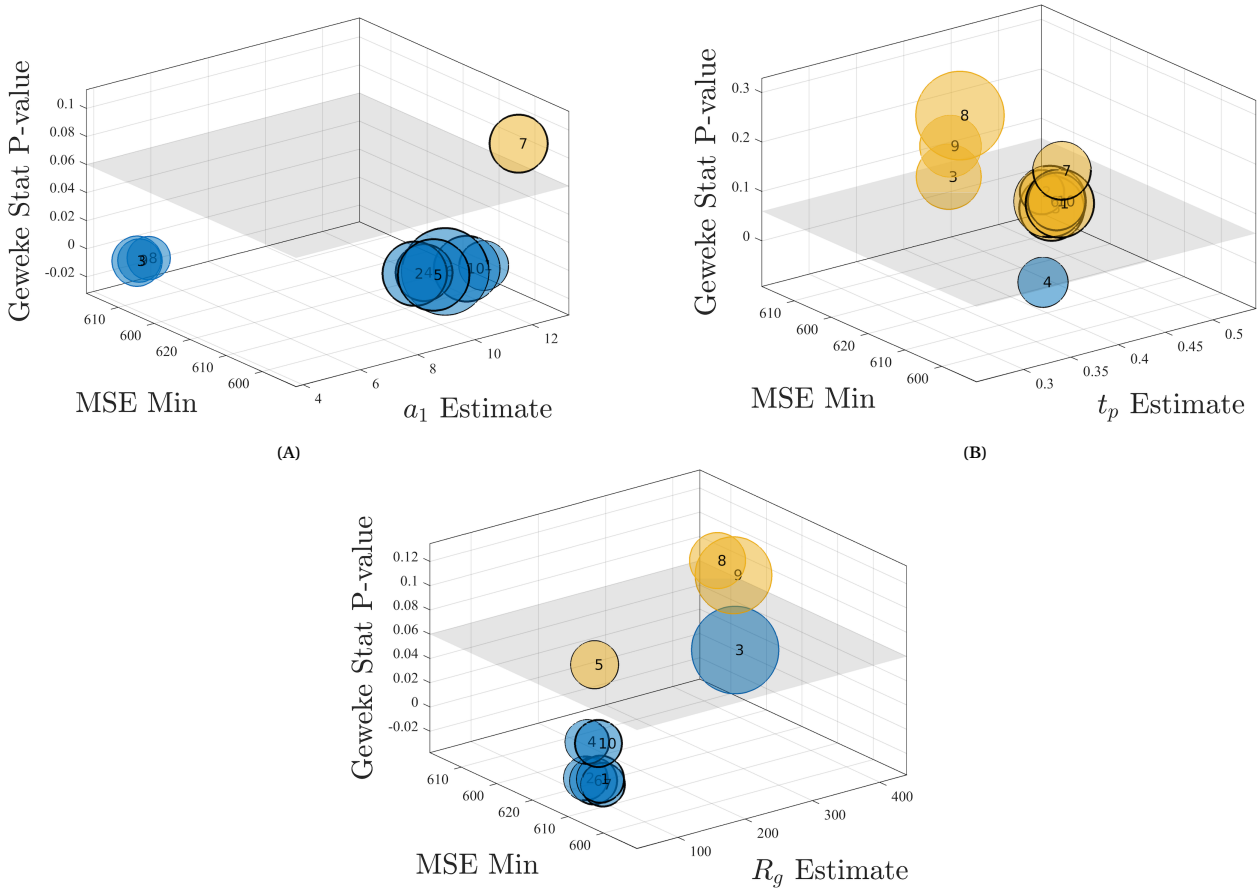


FIGURE 3: Markov chains selection to parametrically represent a patient. Colored bubble charts of Markov chains selection for all 10 chains of one ICU stay from cohort A, with axes of parameter estimate, MSE, and Geweke statistics p-value, for each model parameter **(A)** a_1 , **(B)** t_p, and **(C)** R_g separately. Each bubble representing a single Markov chain was labeled with chain number. Bubbles were drawn with radius corresponding to Markov chain IQR and colored blue to indicate Geweke convergence of that parameter. Markov chains with low MSE and intra-chain convergence of at least one parameter were selected. The seven Markov chains meeting these criteria were outlined in black and were selected to represent this ICU stay.

3.2 Phenotype Computation and Physiological Knowledge Extraction from Parameter Estimates

Note here that all previous steps in the results used the smaller cohort A. Starting from this section unless otherwise noted, all steps in the pipeline were conducted on cohort B3 to better illustrate the phenotype computation and validation process.

3.2.1 Summarizing Parameter Empirical Distribution Functions for Clustering

The parameter estimates of individually modeled physiology in cohort B3 were represented by 330 EDFs of qualified Markov chains. To compute phenotypes from these estimates using clustering methods, we first compressed each EDF into five statistical indexes. This step facilitates the clustering by compressing the EDF while keeping most of the characteristics of each EDF. In this summary step, we transformed 330 EDFs representing cohort B3 into 330 vectors in 5D, which also reduced the total dimension from 330 * 4K * 3 to 330 * 5 * 3.

3.2.2 Dimension Reduction with t-Distributed Stochastic Neighbor Embedding (t-SNE)

Using t-SNE, we reduced the dimensionality of 330 5D vectors into 2D/3D coordinates while leaving the total counts of vectors unchanged (cf. Sec. 2.3.3). For better visualization, we used t-SNE to reduce the data to two dimensions, which further reduced the total dimension from 330 * 5 * 3 to 330 * 2D as 330 t-SNE 2D vectors. We selected L-Infinity (Chebyshev) metric in t-SNE to distinguish scales in different parameter summaries, because it measures the maximum distance between boundaries in a single direction and produced the lowest loss of around 0.14 in Kullback-Leibler divergence for the embedding data. The t-SNE requires specifying the perplexity parameter with a normal range from 10 to 50, which refers to the effective number of local neighbors for each summarized chain and defines the balance between local and global structure via controlling the local standard deviation of a Gaussian distribution. Given the size of cohort B3, we set 30 as the perplexity parameter. Figure 4 (inset) shows the raw, un-clustered t-SNE result with a clear structural separation in the lower left and top left area, indicating at least three possible phenotypes in cohort B3.

3.2.3 Phenotypes Computation by Clustering t-SNE Coordinates with Support Vector Clustering (SVC)

We used SVC to group ICU stays of patients based on nearby t-SNE 2D coordinates. The population-scale phenotypes were found by applying SVC (cf. Sec. 2.3.4) to 330 t-SNE coordinates without changing the dimension. SVC requires a few initial methodological choices. We chose the radial basis function (RBF) kernel and adjusted the smoothness and the fraction of outliers allowed in a cluster for coefficients iteratively until visually sensible to detect more plausible clusters in the cohort. As a manual scaling approach, we selected 4 as the kernel radius and 0.6 as the boundary penalization. Figure 4 shows five SVC-computed phenotypes in cohort B3, with each phenotype having a unique centroid in the 2D space, indicating five potential phenotypes. A similar process yields three and four clusters for cohort B1 and B2, respectively. In fact, as a potent internal validation, this clustering process was clustering parametric representation of ICU stays without knowing how patients were partitioned in the EHR data.

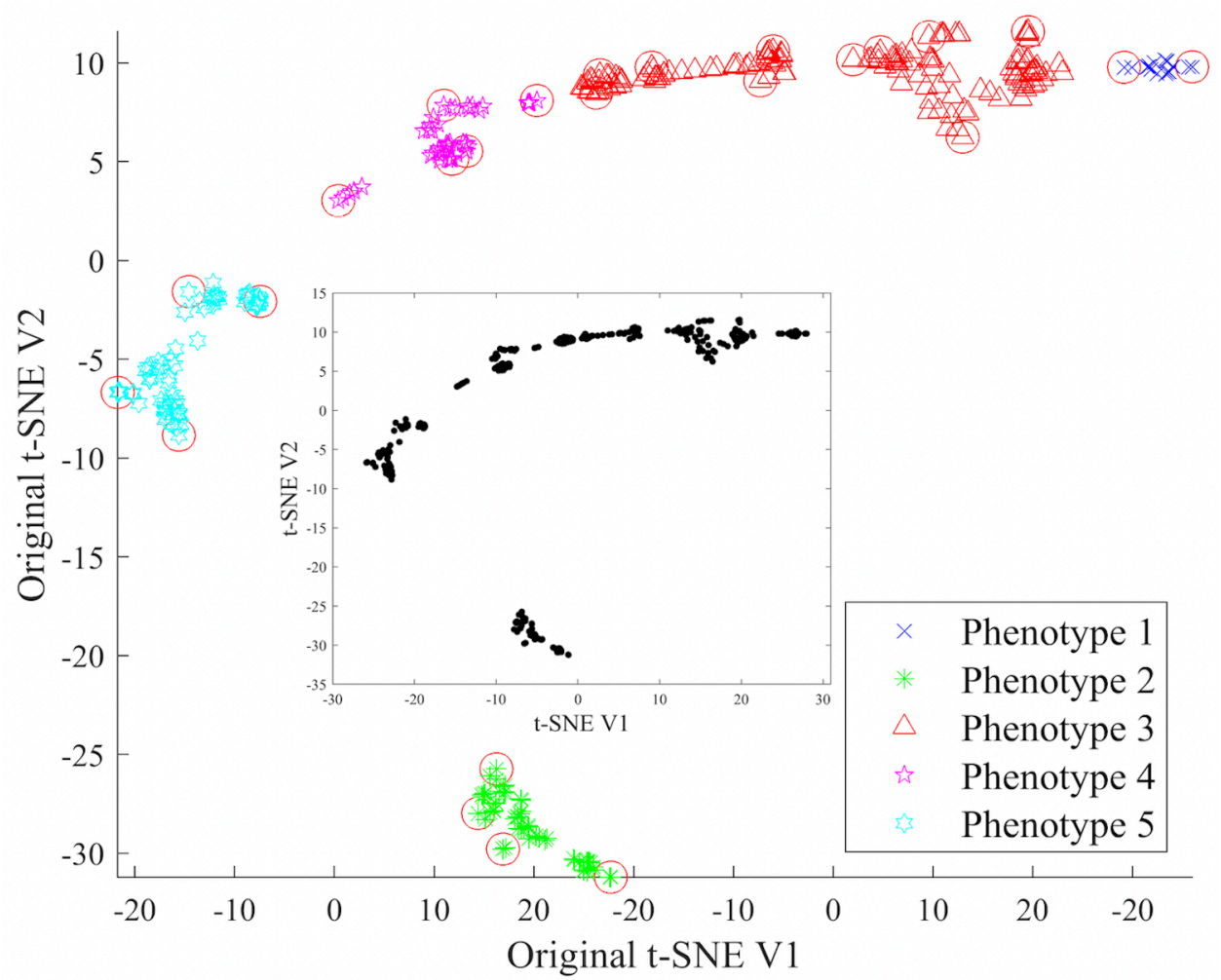


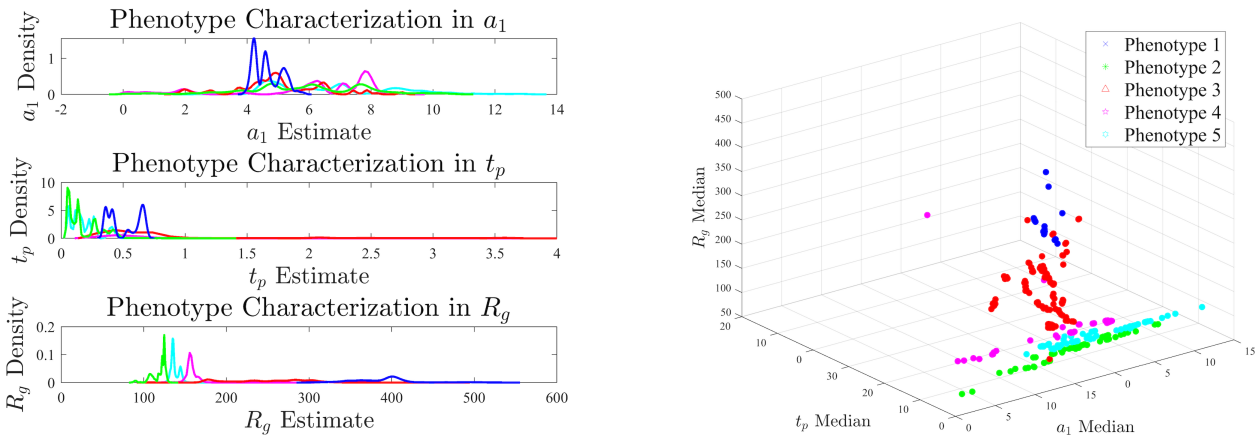
FIGURE 4: **Step 1 (inset)**: Raw and un-clustered t-SNE dimension reduction of selected chains summary of 45 ICU stays in cohort B3 with specified methodological choices. **Step 2**: Colored phenotype computation with SVC that clusters t-SNE coordinates of chains summaries based on SVC methodological choices. Circled coordinates are support vectors mapped to the surface of the high dimensional feature space sphere. Five clusters were identified indicating five potential phenotypes.

3.2.4 Extracting Physiologic Definitions of Computed Phenotypes from Model Parameter Descriptors

We reconstructed physiology from the SVC-identified clusters in four steps to map the population-scale phenotypes to clinical interpretations. Firstly, we mapped individual t-SNE coordinates in SVC-identified clusters back to the original parameter estimates of (a_1, t_p, R_g) represented by empirical distribution functions (EDFs) of Markov chains. This step means that EDFs associated with t-SNE coordinates within each SVC-identified cluster were concatenated to generate a parametric representation of each phenotype. In the second step, we extracted the physiologic meanings of each phenotype using the unique descriptors of the estimated parameter triplet (a_1, t_p, R_g). Recall that for this parameter triplet, computed phenotypes are associated with insulin insensitivity (R_g), pancreatic insulin secretion (a_1), and insulin degradation via liver and kidney (t_p).

Figure 5 shows the parametric representation of five distinct phenotypes identified from SVC in cohort B3 in 2D and 3D median space based on these two steps. These five parameter densities from 330 individual EDFs

provide a basis for identifying unique physiological traits present in the ICU stays associated with them. As one important observation from figure 5, these five different distributions of phenotypes are converged, meaning that we found five different phenotypes. These five phenotypes are also physiologically distinctive, e.g., the phenotype in blue has distinctively lower modes and variation compared to the cluster in pink.



(A) Phenotype Cluster in 2D Density

(B) Phenotype Cluster in 3D Median

FIGURE 5: Parametric characterization of phenotypes from SVC in cohort B3 with density plots in **(A)** 2D and **(B)** 3D median space. Clusters coloring is the same as Figure 4. Both density plots in 2D and 3D median space show five different patterns–the phenotypes–over the entire parametric space and thus reveal five different phenotypes. For example, the phenotype in blue is interpreted as the phenotype with patient ICU stays that have a_1, the exponential constant for insulin secretion inverse rate, roughly between 4 and 5; t_p, the time constant for plasma insulin degradation inverse rate by the kidney and liver, between 0.25 and 0.7 min; and R_g, the insulin-dependent glucose production rate or insulin insensitivity, between 300 and 550 mg/min.

Thirdly, we identified and separated coherent phenotypes from current SVC-identified clusters to represent self-consistent phenotypes (cf. Sec. 2.3.4). This step is necessary because t-SNE coordinates associated with the same ICU stay at the same time can be clustered into more than one phenotypes, which complicates further phenotypic interpretation at the patient level. For example, we previously identified five phenotypes from SVC clusters in cohort B3. We augmented these five phenotypes with five auxiliary phenotypes that were associated with multiple SVC clusters at the ICU-stay level. Then five out of these ten phenotypes are associated with fewer than three ICU stays and we dropped them considering statistical power. We ended with five self-consistent phenotypes that have well-resolved ICU stays. These five phenotypes are also self-consistent, meaning that the patients inside the phenotypes had uniformity in terms of parameterized physiology.

In the fourth step, we discretized the continuous phenotypic information with parameter densities to map population-scale phenotypes to clinical interpretations. We used a customized ranking relative to an average ICU phenotype to delineate phenotypic patterns. Then we arrived at a phenotyping ranking summary table through ranking each categorical phenotype for each parameter with elucidation based on parametric descriptors.

Table 2 tabulates the phenotyping ranking summary as a result of the last step to extract physiology of computed phenotypes. This table delineates coherent phenotypes identified in cohort B1, B2, and B3, along with relative discretized qualities of their parameter triplet (a_1, t_p, R_g) based on the customized ranking, respectively. Phenotypes computed from each cohort are 'relative' instead of 'absolute', meaning that they are relative to a period-specific level. Even if the phenotype descriptors look the 'same', e.g., G7 and G9, they are not the same phenotypes as they were computed from different periods. For example, phenotype G8, identified in cohort B3 (Table 2 bottom), is interpreted as having patients with insulin secretion rate relatively slightly below the mean, insulin degradation rate relatively much higher than average, and relatively slightly decreased insulin

Phenotype	Patients	ICU Stays	a_1	t_p	R_g
G1	4	4	↓	↓↓↓	↑↑
G2	11	11	↓	↓↓↓	↑↑↑
G3	36	36	↑	↑	↓
G1 & G2	4	4	↑	↑↑↑ (∗)	↑↑
G1 & G3	3	3	↓	↓↓	↑
G2 & G3	4	4	↓	↑	↑↑↑

(A) Cohort B1

Phenotype	Patients	ICU Stays	a_1	t_p	R_g
G4	12	12	↑	↓↓↓	↓↓
G5	7	7	↓	↓	↑↑↑
G6	15	15	↓	↑↑	↑
G7	16	16	↑	↓↓	↓
G5 & G6	4	4	↓	↑↑	↑↑↑ (∗)
G6 & G7	4	4	↓	↑↑↑ (∗)	↑

(B) Cohort B2

Phenotype	Patients	ICU Stays	a_1	t_p	R_g
G8	9	9	↑	↓↓↓	↓
G9	6	6	↑	↓↓	↓
G10	17	17	↓	↑↑↑	↑
G11	9	9	↑	↓↓↓	↓↓
G12	4	4	↓	↓↓	↑↑↑

(C) Cohort B3

TABLE 2: Phenotype delineation of consistent phenotypes identified in cohort **(A)** B1, **(B)** B2, and **(C)** B3. Five phenotypes were identified in cohort B3, six in B1, and six in B2. The delineation tables have columns of consistent phenotype, number of patients, number of ICU stays in the phenotype, and parameter ranking of the parameters triplet (a_1, t_p, R_g): a_1, the exponential constant for insulin secretion inverse rate; t_p, the time constant for plasma insulin degradation inverse rate; and R_g, the insulin insensitivity. Within each cohort, the customized ranking method compares each phenotype parameters triplet estimation average to the posterior average across all phenotypes–the midpoint of all parameter estimates. We qualitatively discretized phenotype parameter estimation into six bins centered at the midpoint. The number of upward and downward arrows indicates a ranking of how far from this midpoint the phenotype average is, ranging from 0 to double the midpoint value. Phenotype average that falls outside of either end of the range has a complementary star sign (*). We named auxiliary phenotypes as the incorporation of SVC clusters names they were previously associated with and then separated from. For example, auxiliary phenotype 'G1 & G2' in cohort B1 was separated from previously SVC-identified clusters G1 and G2.

insensitivity. Based on this table, similarly in cohort B1, B2, and B3, we observed coherent ICU stays within each phenotype that was computed without prior knowledge of the patients. We also used this table to show that phenotypes identified in the first three days are likely not to show up in the last three days, as patient health status change drastically during the entire ICU stay. Observation of the 'same' phenotypes across three periods indicates that patients' sickness condition is not improved. More complete technical descriptions of the ranking method are provided in Appendix (Sec. A.4).

3.2.5 Phenotype Contextual Interpretation with Laboratory Measurements and ICD 9/10 Codes

To better match with clinical language for potentially guiding clinical ontology, we generated a hierarchical visualization with phenotypic description delineation that incorporates physiologic interpretation with data-based profiles of patients. We first investigated a list of laboratory measurements and ICD 9/10 codes that can best reflect organ functionalities encoded by patient parameters. The selected list is included in Appendix (Sec. A.5). Our attention focused on observations that serve specifically as indicators of liver, pancreas, and kidney functions. Combining these selected clinical observations of patients and extracted physiology from population-scale categorical phenotypes, we described the time-specific health status of patients in each identified phenotype. The description was performed separately for the first three-day and last three-day

period to highlight the change in patient health status between these two periods. In a hierarchical visualization we delineated all model-identified phenotypes by each discretized estimated physiologic functionality at each hierarchy in sequence, and prescribed the synthesized phenotypic description to patients associated with each phenotype at the last hierarchy.

Figure 6 presents the five phenotypes identified in cohort B3 in a hierarchical visualization starting with model-derived phenotypes on the left and combined phenotypic description on the right. This delineation process demonstrates how the new information from estimated physiology is synthesized into a phenotypic picture of a patient from implementing this pipeline, and how the hierarchical visualization can be potentially helpful to differentiate patients in a clinical context.

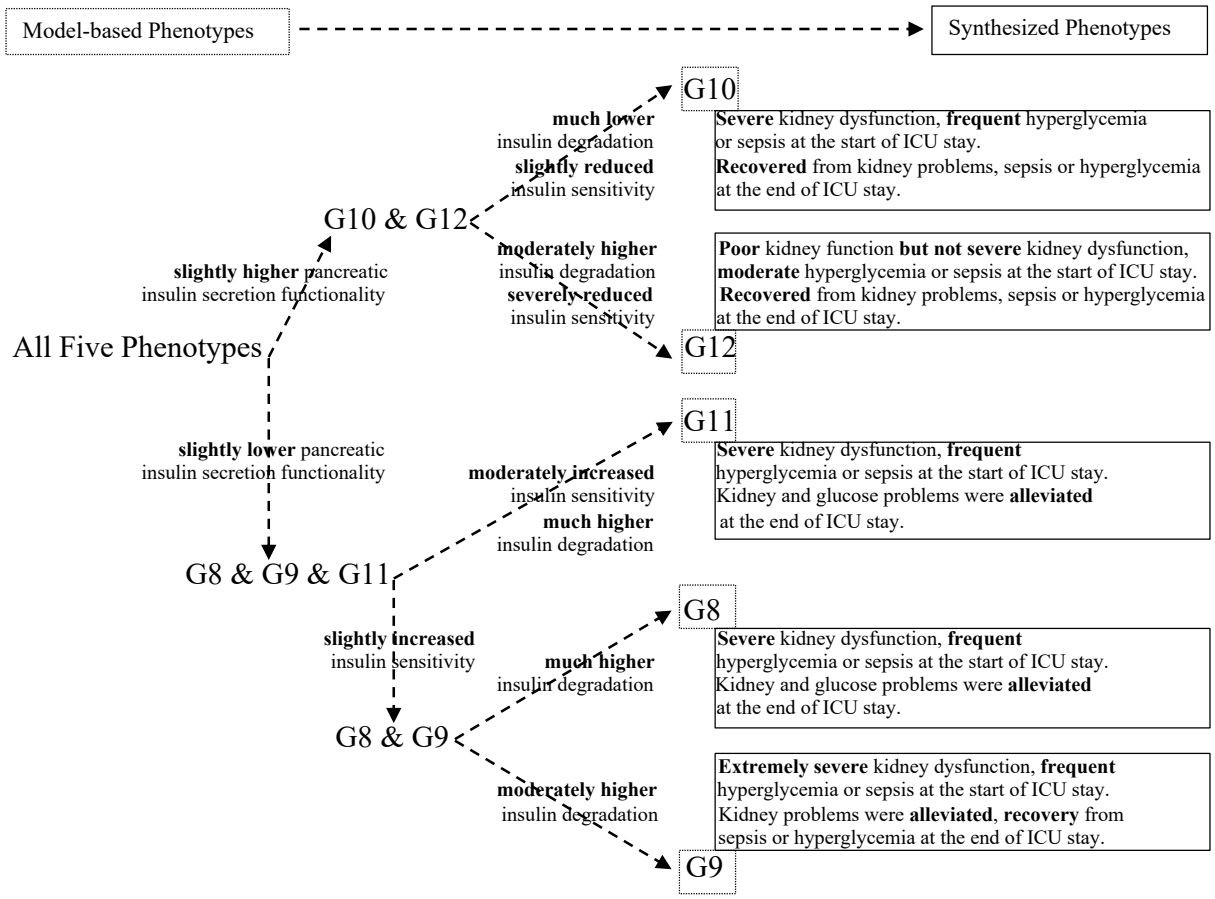


FIGURE 6: Phenotypes delineation of model-derived phenotypes and corresponding synthesized phenotypic interpretations. This figure is a hierarchical visualization of the phenotypes to guide delineating five phenotypes identified in cohort B3. Starting from left, we separate model-based phenotypes at each hierarchy, and ends with the synthesized interpretations on the right with consistency. Comparing the level of pancreatic insulin secretion functionality at the first hierarchy, phenotypes G8, G9, and G11 can be differentiated from G10 and G12. Then comparing the level of insulin degradation at the second hierarchy, G10 can be differentiated from G12. Similarly, comparing the level of insulin insensitivity at the second hierarchy, G11 can be differentiated from G8 and G9, followed by comparing the level of insulin degradation at the third hierarchy to differentiate G8 from G9.

3.3 Validating Model-derived Phenotypes Against Laboratory Measurements, ICD 9/10 Data, and Clinical Notes

In the previous section we discretized individual DA-based continuous parameter estimates of physiological states for ICU stays in cohort B3 into five coherent categorical phenotypes with physiologic interpretations. However, these phenotypes remain in physiological space without an explicit clinical context. To establish the clinical context of the model-derived phenotypes, we focused on those five phenotypes identified in cohort B3–the last three-day period of cohort B–because patient states are more stable approaching feeding tube removal. We then evaluated these model-based phenotypes and validated them with extracted physiological interpretations against EHR data of laboratory measurements and ICD 9/10 codes, both quantitatively and qualitatively. Note that these data were not used previously in the DA phenotyping pipeline and because of this, we can use this contextualization as a validation of the DA phenotypes.

3.3.1 Statistical Evaluation of Phenotypes by Their Predictive Power

As we used SVC to compute phenotypes as clusters of t-SNE coordinates compressed from empirical distribution functions (EDFs) of Markov chains (cf. Sec. 3.2.3), we chose to use support vector machine (SVM), which shares a similar mechanism with SVC, to access phenotype computation predictive power with statistical evaluation. To compute the initial phenotypes we picked a radial kernel for SVM to classify features of 330 EDFs labeled with phenotypes that we identified as G8-G12 (cf. Sec. 3.2.4). Features of EDFs were summarized as mean and standard deviation and we used kernel parameters tuned by LOOCV to quantitatively evaluate prediction performance from these features. The prediction macro F1-score is 88.3%, the weighted F1-score is 89.3%, and the average accuracy is 89.4%. We concluded that the phenotypes computation is reliable with an acceptable predictive power.

3.3.2 Summarizing Phenotypic Difference Using ICD 9/10 Codes

From EHR data we selected five ICD 9/10 codes (cf. Appendix A.5) that indicate problems with organs whose functionality are encoded in model parameters. We calculated and compared percentage counts of selected ICD 9/10 codes for patients within each model-derived phenotype to find clinical evidence that supports difference between phenotypes. The patients physiology difference with regard to clinical observation of organ functionalities revealed by these ICD 9/10 codes is in general consistent with the findings of five model-derived phenotypes, which also supports the phenotypes delineation illustrated in the hierarchical visualization in Sec. 3.2.5.

3.3.3 External Validation of Phenotype Distinctiveness by Statistical Testing Against Laboratory Measurements Excluded From Phenotype Computation

To validate the distinctiveness of model-derived categorical phenotypes with extracted physiology (cf. Sec. 3.2.4) against observed patient physiology not used in phenotype computation, we analyzed the data-based profiles based on seven selected laboratory measurements of patients in each phenotype. We investigated a list of laboratory measurements (cf. Appendix A.5) not used in phenotype computation and that clinicians refer to as dysfunction indicators of certain organs whose functionality are relatable to the extracted physiology from phenotypes. Table 3 contains a summary of patients selected laboratory values in the five phenotypes identified in cohort B3. Through this table we found evidence of the effective separation between model-based phenotypes from data-based profiles (horizontal comparison). For example, in the last three-day period, we found high counts of creatinine (CREAT) test for phenotype G10, which relates to monitoring kinetic filtering functionality in the absence of dialysis treatments. We also observed frequent lipase (LIPA) tests only for phenotype G11 in the last three-day period, which suggests distinctive pancreatic disorders. In the table were

also records of elevated levels of aspartate aminotransferase (AST) tests indicating potential hepatic issues for phenotype G8 and G10. Also in the laboratory-wise vertical comparison, this table shows that patient's health status change consistently over time in the same phenotype. This consistency of patients physiology is revealed by laboratory measurements that are clinically different throughout the stay, from the first enteral nutrition after ICU admission (the initial three-day period) to feeding tube removal (the last three-day period).

Moving forward from the crude summary of laboratory values presented in Table 3, we also statistically tested laboratory mean values among all phenotypes with a Welch two-sample t-test to check inter-phenotype clinical differences. The statistical analysis description is incorporated into the synthesized qualitative phenotypic descriptions in the next section.

Data-based Profile with Laboratory Measurements						
Laboratory (Last Three & First Three Days)		G8(N=9)	G9(N=6)	G10(N=17)	G11(N=9)	G12(N=4)
AMY	Last	1	NA	3	1	NA
Counts		15.0		15.7±3.1	39.0	
Mean±SD				15.0 [13.0,19.0]		
Median [Min, Max]	First	NA	1	3	NA	NA
			22	15.3±2.5		
				15.0 [13.0,18.0]		
AST	Last	27	10	27	12	6
Counts		698.9±908.6	33.6±17.8	1262.0±3652.0	148.8±155.9	39.2±24.2
Mean±SD		446.0	27.0 [19.0,75.0]	47.0	73.5	25.5 [21.0,78.0]
Median [Min, Max]		[10.0,4240.0]		[20.0,15920.0]	[25.0,415.0]	
	First	28	8	32	16	5
		1118.5±1160.3	89.8±141.1	1076.0±3373.0	117.1±121.1	41.8±26.0
		815.5	21.0	47.0	84.5	25.0 [21.0,78.0]
		[15.0,4240.0]	[12.0,420.0]	[22.0,15920.0]	[13.0,415.0]	
LIPA	Last	NA	NA	NA	16	NA
Counts					66.5±76.9	
Mean±SD					44.0	
Median [Min, Max]					[23.0,338.0]	
	First	1	1	NA	19	NA
		9	15		77.8±110.5	
					17.0	
					[10.0,338.0]	
BUN	Last	NA	2	3	1	1
Counts			51.5±17.7	16.3±11.4	33	40
Mean±SD			51.5 [39.0,64.0]	13.0 [7.0,29.0]		
Median [Min, Max]	First	3	4	6	5	3
		21.0±7.0	30.3±11.1	27.8±12.9	17.0±11.7	29.7±17.7
		18.0 [16.0,29.0]	27.5 [20.0,46.0]	26.5 [13.0,47.0]	17.0 [6.0,36.0]	21.0 [18.0,50.0]
CREAT	Last	79	50	119	58	23
Counts		2.3±1.1	2.2±1.0	1.1±0.5	1.3±0.8	2.3±0.8
Mean±SD		2.4 [0.7,4.4]	1.9 [1.2,4.3]	1.0 [0.3,3.4]	1.1 [0.5,3.8]	2.6 [0.6,3.0]
Median [Min, Max]	First	81	53	125	69	23
		1.8±0.8	1.6±0.8	1.1±0.4	1.3±0.8	2.3±0.8
		1.4 [0.8,4.0]	1.4 [1.0,4.3]	1.1 [0.6,2.7]	1.2 [0.5,3.8]	2.6 [0.6,3.0]

TABLE 3: Data-based phenotype profile with amylase (AMY), aspartate aminotransferase (AST), lipase (LIPA), blood urea nitrogen (BUN), and creatinine (CREAT) measured in the last three days, and in the first three days of the five model-based consistent phenotypes in cohort B3. This table has columns of phenotypes with patients counts and rows of laboratory measurements summary, including measurements counts, mean±SD, and median [min, max]. Quantities not measured for each phenotype were noted as NA. Interestingly, laboratory measurements related to liver/kidney function, including alanine aminotransferase, albumin, and bilirubin, are not present in any phenotype identified in the last three days, such that they are not included into the data-based profile of the first three-day period.

3.3.4 Synthesizing Qualitative Phenotypic Descriptions Using Laboratory Measurements and ICD 9/10 Codes

To qualitatively summarize the concordance and discordance between model-derived phenotypes and data-based profiles, we generated qualitative phenotypic interpretations by synthesizing extracted physiology of model-identified phenotypes with ICD 9/10 codes percentage counts analysis (cf. Sec. 3.3.2) and statistical testing of selected laboratory measurements (cf. Sec. 3.3.3). Compared to the hierarchical visualization of phenotypes delineation (cf. Sec. 3.2.5), the qualitative phenotypic interpretations are more amendable to clinical confirmation because of the phenotypic concordance analysis with external observational data. The synthesis process also defines a map from the unmeasurable model-based phenotypes to the measurable data-based profile, which aids in the analysis and construction below.

Table 4 uses phenotype G8 as an example to illustrate the synthesis of each qualitative phenotype description in cohort B3. Note here that cohort B3 reported no obese patients and few with diabetes per ICD 9/10 codes. Also, as in a typical ICU environment, critically ill patients primarily receive intravenous infusion of insulin. Therefore, the average ICU patient is an ICU patient who would only be given short-acting insulin therapy, and we use this average ICU patient as the baseline to compare relative qualitative interpretations among phenotypes.

Qualitative Interpretations of Phenotype G8	
Steps	Descriptions
Step 1: Model-based Phenotypic Interpretation	Patients in phenotype G8, compared to an average ICU patient, have: a slightly decreased pancreatic insulin secretion, much higher insulin degradation assuming no dialysis, and a slightly decreased insulin insensitivity. This interpretation is based on physiologic interpretation indicating the slightly higher (↑) insulin secretion rate, much lower (↓↓↓) insulin degradation rate, and slightly lower (↓) insulin-dependent glucose production rate. Compared to the nominal model parameter values, patients in G8 have a lower pancreatic insulin secretion functionality, higher insulin degradation assuming no dialysis, and a lower insulin insensitivity.
Step 2: Data-based Profile with Laboratory Measurements	Using the first three days of laboratory measurements of patients in phenotype G8, we observe significantly higher AST values compared to G9 ($p<0.001$), G11 ($p<0.001$), and G12 ($p<0.001$). This indicates higher level of dysfunction in liver and kidney in phenotype G8 compared to all other phenotypes except G10. We also observe that patients in phenotype G8 are missing laboratory measurements of AMY. Using the last three days of laboratory measurements of patients in phenotype G8, we observe significantly higher AST values compared to G9 ($p<0.001$), G11 ($p=0.004868$), and G12 ($p<0.001$). This indicates different liver functionality in phenotype G8 compared to all other phenotypes. We also observe that patients in phenotype G8 are missing laboratory measurements of LIPA and BUN.
Step 3: Data-based Profile with ICD 9/10 Data	Based on ICD 9/10, patients associated with ICU stays in phenotype G8 were diagnosed as having percentage counts of hyperglycemia or sepsis 67% of the time during the complete stay, and 11% in the last three days. Patients associated with ICU stays in phenotype G8 were also diagnosed as having percentage counts of dialysis or renal/kidney failure 78% of the time during the complete stay, and 22% in the last three days.
Step 4: Concordance Between Model-based Phenotype and Data-based Profile	Only through ICD 9/10 data can we know that phenotype G8 had severe kidney dysfunction and a high incidence rate of hyperglycemic events or sepsis at the ICU admission. Both model and laboratory measurements indicate liver dysfunction in phenotype G8 in the last three days. Both laboratory measurements and ICD 9/10 data indicate patients in phenotype G8 managed to alleviate kidney dysfunction in the last three days. Only through ICD 9/10 data can we know that patients in phenotype G8 managed to alleviate hyperglycemia and even sepsis approaching ICU discharge.
Step 5: Synthesized Phenotypic Description	Patients in phenotype G8 had a slightly decreased pancreatic insulin secretion, much higher insulin degradation assuming no dialysis, and a slightly decreased insulin insensitivity in the last three days. They had liver dysfunction, severe kidney dysfunction, and a high incidence rate of hyperglycemic events or sepsis at the ICU admission, but managed to alleviate kidney dysfunction and hyperglycemia, even sepsis approaching the ICU discharge.

TABLE 4: Steps in generating synthesized descriptions for phenotype G8. We summarized the concordance and discordance between model-based phenotypes and data-based profiles, organically combining the interpretations of both model-based phenotypes and data-based profiles. The average ICU patient here represents the average health status of an ICU patient who received short-acting insulin therapy during their complete stay. Missing laboratory measurements were noted. For each ICD 9/10 code, the percentage count was dividing the code counts in each phenotype by the total number of distinct ICU stays, respectively.

3.3.5 External Validation of Phenotypes with Clinical Face Validity

Evaluation and validation of computed phenotypes in this section are qualitative. We used face validity [52] to check methodological pipeline's clinical consistency based on an endocrinologist's review of patient notes in the last three-day period. We opted to validate at the level of the physiology descriptors: the parameter triplet (a_1, t_p, R_g) (cf. Sec. 3.2.4). We picked 20 members from the resolved set of 45 ICU stays in cohort B3 (cf. Sec. 3.2.4) and extracted their parameter estimates represented by empirical distribution functions (EDFs) of Markov chains. Similar to the customized ranking method in section 3.2.4 and Appendix (Sec. A.4), we discretized individual parameter ranges into three levels on the cohort range, and then prescribed parameter likelihoods as weights for the discretized levels based on the EDFs. Using this weighted basis, we statistically defined the similarity of expert response to model-based phenotypes expectation based on one minus a simple ordinal distance metric.

Our model-based phenotype parametric estimation qualitatively agreed with associated physiological functionalities identified by the blind chart review, which shows that the model-derived phenotypes are reasonable. The concordance–measured by the mean of 20 responses–of each parameter are: 83% ± 27% for a_1 (insulin secretion), 66% ± 35% for R_g (insulin insensitivity), and 52% ± 44% for t_p (insulin degradation), respectively. However, the physiological model-based estimates, and thus the concordance between the model and clinician may be confounded by interventions and therapies in ICU that are not estimable from the model. For example, the dialysis treatment due to acute kidney disease confounds our estimation of insulin degradation, leading to a lower concordance of model parameter t_p. Table 5 summarizes the potential confounding factors that affect model estimation of insulin clearance, combining nurses notes during patient ICU stay and clinician notes during face validation, which is further discussed in conclusion Sec 4.4.

Clinical Confounding Factors in Face Validation (FV)				
FV Patient	Clinician Perspective	Diagnosis	Therapy/Procedure (FV period)	Therapy/Procedure (complete ICU stay)
1	passed away on the last day after withdrawal of support			
2	passed away on the last day after withdrawal of support	other acute kidney failure†		
3	passed away on the last day after withdrawal of support			placement and removal of hemodialysis access non-tunneled double lumen dialysis catheter
4		CKD (chronic kidney disease) stage 2†, renal insufficiency*		placement and removal of hemodialysis access non-tunneled double lumen dialysis catheter, heart transplant*
5	CVVH	renal failure acute (HC code)†		
6	previous lung transplants, CVVH			
7	T1DM	acute kidney failure unspecified†		
8			placement and removal of hemodialysis access non-tunneled double lumen dialysis catheter	
9	from low blood pressure to then improved renal function			
10	heart transplant			
11	RRT hemodialysis			
12	previous heart transplants			
13	worsening renal function			
14		medical renal disease*		renal ultrasound*
15		chronic renal failure*		

TABLE 5: Confounding factors of insulin clearance from clinician perspective and ICU notes that may affect face validity concordance. 15 of 20 picked ICU stays either had hemodialysis treatment or continuous veno-venous hemofiltration (CVVH), or had kidney failure related conditions indicating the necessity of dialysis treatment during the ICU stay. Diagnoses based on ICD 9/10 codes have dagger superscripts †. The asterisk superscripts * indicate diagnoses, therapies, and procedures identified from free text narratives.

3.4 Additional Insights: Impact of Health Care Process on Phenotype Accuracy

To learn the impact of health care process bias due to, e.g., hemodialysis or insulin interventions, on our phenotype computation, we alter confounding factors that perturb physiological functionalities, e.g., insulin production, in the pipeline. Specifically, because the models we estimate only include physiology, their inability to differentiate intervention effects from physiology can reduce phenotype accuracy. To test this, we can remove patients with T1DM–eight in cohort B, seven in cohort B3, and two in the face validation cohort–and consider the impact on the pipeline. We perform this natural experiment of perturbing the pipeline by removing patients with T1DM because these patients have an abnormal pancreas that does not produce insulin. The mechanistic Ultradian model cannot differentiate the intervention from a patient's glycemic state because the model, given the sparse nature of the data, has no way of discerning insulin deficiency in patients with T1DM since a dense time-varying response to nutrition and insulin administration is not observed. We applied the first and second stages of the pipeline to the updated cohort B3 excluding patients with T1DM patients and computed updated phenotypes. The result was four consistent phenotypes with one changed phenotype using the same pipeline configuration and an improvement in pipeline predictive power and phenotype clinical concordance.

Table 6 top two colored subtables detail patients indexed without ranking and columned by phenotypes in each cohort. We used colors to track each patient updated labeling from phenotypes G8-G12 (cohort B3) to four phenotypes (cohort B3 excluding T1DM), providing additional insights into our hypothesis of machine robustness and consistency. Excluding T1DM patients from the pipeline reorganizes phenotypes definition by still keeping the original phenotypes G11, G12, and the majority of patients in G10. The only one phenotype that appears when the T1DM patients are removed is the combination of all patients previously labeled as phenotypes G8 and G9 and a portion labeled as G10. This phenotype 'G13' shares characteristics of G8 and G9 in terms of interpreted alleviated kidney dysfunction and similar levels of insulin secretion and resistance, which can be explicitly observed in the hierarchical phenotypes delineation in Figure 6. The splitting of G10 among two phenotypes is related to liver problems and kidney disease indicated from higher levels of aspartate aminotransferase, blood urea nitrogen, and creatinine in patients clustered into 'G13'. We included the updated visualization of patients projection via t-SNE and SVC in Appendix (Sec. A.6).

Removing some of the health care process bias also improved the pipeline quantitatively internal evaluation performance and qualitative phenotype mapping concordance. Table 6 bottom table details the before and after comparison of phenotype-specific predictive power. The pipeline internal evaluation manifested improvement by a margin in terms of average accuracy (from 89.39% to 91.97%), macro F1 score (from 88.30% to 91.59%), and weighted F1 score (from 89.26% to 91.58%). The other improvements of phenotypes sensibility and clinical relevance see a boost in the concordance between extracted physiology and practical clinical meanings in face validation for all three physiology descriptors: from 83%±27% to 87%±26% for a_1 (insulin secretion), from 66% ±35% to 71%±33% for R_g (insulin insensitivity), and from 52%±44% to 58%±43% for t_p (insulin degradation), respectively.

G8	G9	G10	G11	G12
1	10	16	33	42
2	11	17	34	43
3	12	18	35	44
4	13	19	36	45
5	14	20	37	
6	15	21	38	
7		22	39	
8		23	40	
9		24	41	
		25		
		26		
		27		
		28		
		29		
		30		
		31		
		32		

(A) Phenotypes in Cohort B3

$\widetilde{G11}$	$\widetilde{G10}$	G13	$\widetilde{G12}$
35	17	11	42
36	18	1	43
37	22	2	44
38	23	19	45
39	25	4	
40	26	20	
41	27	12	
	28	5	
	31	24	
	32	6	
		7	
		30	
		46	
		13	
		14	
		47	
		15	

(B) Phenotypes in Cohort B3 Excluding T1DM

	Cohort B3					Cohort B3 (Excluding T1DM)			
Predictive Power	G8	G9	G10	G11	G12	$\widetilde{G11}$	$\widetilde{G10}$	G13	$\widetilde{G12}$
Precision (%)	85.71	93.02	91.16	82.14	100	80.43	97.3	90.78	100
Recall (%)	80.6	83.33	99.26	82.14	87.5	80.43	94.74	92.09	100
F1 Score (%)	83.08	87.91	95.04	82.14	93.33	80.43	96	91.43	100

(C) Phenotype-specific Statistical Internal Evaluation Comparison

TABLE 6: The impact of health care process bias on phenotypes accuracy in terms of machine sensitivity to phenotypes computation and predictive power. Tracking of indexed patients updated labeling from **(A)** phenotypes including T1DM patients to **(B)** phenotypes excluding T1DM patients in cohort B3. The color scheme is consistent with Figures 4 and 5 to indicate each patient's association with phenotypes G8-G12 that include seven T1DM patients highlighted in yellow. Patients in grey are those from coherent phenotypes identified in the current cohort but dropped from the other cohort as we considered them to be associated with SVC-identified auxiliary incoherent phenotypes below a sufficient sample size (cf. Sec. 2.3.5). **(C)** The phenotype-specific predictive power comparison shows improvement of mapping concordance across all three performance metrics: precision, recall, and F1 score.

4 Conclusion

4.1 Key Results and Related Contribution to Current Phenotyping Methods

4.1.1 Summary: The Three-stage Methodological Pipeline

In this paper we developed a novel DA-based three-stage methodological pipeline that can compute personalized, physiologically-anchored, time-specific, and interpretable phenotypes from EHR data. We applied the pipeline to ICU patients using an established mechanistic model–the Ultradian model–and MCMC estimation to infer unmeasurable endocrine physiology using blood glucose measurements and administered nutrition and insulin. Much of the phenotype computation was performed on a cohort of 109 patients with a focus on the last three-day period of enteral nutrition in the ICU and with the further restriction to patients who were given either no or short-acting insulin. We reduced the model for inversion based on objective criteria to avoid technical complications that arise from non-unique inference solutions due to sparse and non-stationary data, balancing model flexibility and fidelity. Then individual patient physiologic states were continuously represented by qualified probability distribution estimates of the reduced model parameters. Through this representation process we inferred and created new physiological information that we then clustered into population-scale phenotypes that were subsequently discretized and evaluated.

The phenotype computation was performed using unsupervised ML methods including t-SNE for dimensionality reduction and SVC for clustering. The computation of phenotyping discretized the continuous parametric estimates of the ICU stays in the cohort into physiologically-anchored mechanistic categorical phenotypes that are coherent defined by mutually exclusive SVC-identified clusters. We found six coherent phenotypes in the first three-day, six in the middle three-day, and five in the last three-day period. We interpreted these phenotypes by reconstructing their physiology from three unmeasurable endocrine physiological descriptors that define the identified phenotypes, where these descriptors are relevant aspects of patient kidney, liver, and pancreas functionalities that are clinically important but unobservable. Based on these new phenotypic interpretations, we also generated a hierarchy visualization to assist delineation of phenotypes for clinical mapping.

The phenotype validation is imperative in establishing the potential clinical utility of the computed phenotypes. Clinically observed inter-phenotype difference in external data including selected laboratory measurements and ICD 9/10 codes were consistent with the five model-based phenotypes identified in the cohort of the last three-day period. We further used face validity to check the methodological pipeline's clinical consistency at the level of three model physiology descriptors. We obtained good results of concordance from parameter-wise face validation of phenotype characteristics with concordance measured over 50% on average for all three descriptors, although work remains to clarify potential confounding factors in order to establish clinical translatability.

4.1.2 Pathway for Automating and Embedding the Pipeline into a Health System

This pipeline system is already largely automated, although several methodological parts still need manual tuning and optimization. This pipeline also uses proprietary programming environment Matlab that is not typically available inside EHR data system. Similar to ubiquitous clinical decisions that can never be fully automated, necessary manual procedures injected in the pipeline would be completed by medical providers as potential users. Nevertheless, these manual procedures provide decision makers the freedom to project expert knowledge in deciding the model complexity, making adjustments to default method configurations, and identifying a plausible number of phenotypes. The mechanical computational parts of this system also can easily scale to high-throughput in real-time settings, as we can increase the cohort size or simply use more CPUs for higher fidelity cluster results based on the automation. In tandem with the systematic automation, we also inadvertently touch on the generalizability of the methodological pipeline, as we computed phenotypes

for two ICU patient data sets with shared methodological choices of the pipeline. Therefore, the estimated model, being anchored to generalizable physiological knowledge, was proven to be robust to leverage its generalizable information across two hospitals.

4.2 Methodological Choices

We made several preliminary methodological choices, including model section, parameter estimation techniques, dimension reduction approaches, and clustering methods, and our methodology can presumably be sensitive to these methodological choices. For example, we estimated model parameters using MCMC to accurately quantify parameter uncertainty. We also combined t-SNE and SVC to compress parameter estimates, reduce dimensions, and further clustering to compute population-scale categorical phenotypes.

4.2.1 Physiological Model Choice: Fidelity and Flexibility

A model reduces nature to only several elements found to be adequate to explain a particular phenomenon. The mechanistic physiological model is an approximation and hypothesis-based representation of the actual system. Even if we have the right and available data, different models provide different explanations to resolve the same physiological process, which can be differently inaccurate or wrong. For example, the time scales variation between models–from seconds to minutes or even hours–in representing a physiological process.

The physiological model complexity has an impact on what phenotypes are possible, computable, and with different resolutions. On one hand, a very simple model may fail to represent the EHR data hence harms the glycemic states prediction precision. At the other extreme, a model with higher complexity will require high resolution data to estimate that normally do not exist in clinical settings. In this case, model accuracy may be even more vexing due to model identifiability issues producing many seemingly equivalent solutions. Given the perplexing mechanism of insulin oscillations overall with low-fidelity sparse data, predicting physiological states with robust estimation requires a model representing the system with a balanced physiologic fidelity and flexibility. Therefore, we selected a subsystem of the mechanistic Ultradian model.

In this paper, we still had both data and model limitations in estimating the Ultradian model to compute phenotypes. We obtained from an ICU environment continuous nutrition data preferred by the relatively simple Ultradian model that is not designed for an ICU setting. We also had very sparse data as we further confined the cohort by only picking patients on short-acting insulin therapies to accommodate the Ultradian model with simplified pharmacodynamics. Nevertheless, as other data sources become available, e.g., continuous glucose monitors, other more complex models may be more advantageous. We were also not exhaustive in our use of different models, albeit we suspected that in the ICU context one important requirement is the model's ability to exhibit glycemic oscillations.

4.2.2 Alternative Methodological Choice Example in Dimension Reduction of Parameter Estimates: t-SNE vs PCA

In this paper we used t-SNE to reduce the dimensionality of compressed empirical distribution functions (EDFs) of Markov chains to 2D coordinates. An alternative approach of dimension reduction is to represent data in other coordinates. Preferred over t-SNE combined with distribution compression, variation within single EDF can be mostly preserved using PCA. Dimension reduction of parameter estimates using PCA, though, is inadequate for the purpose of phenotyping. Multicollinearity in single EDF due to model parameter synergy necessitates combining all EDFs associated with the same patient, which complicates cohesive phenotype computation that we expect to be independent of patient label.

4.3 Reasoning Behind Identifying Non-mutually Exclusive Discretized Categorical Phenotypes

In the phenotype computation process when we used SVC to cluster t-SNE coordinates, we found that the clusters are not always coherent, meaning that a patient may be associated with more than one SVC-identified cluster. There are two possible reasons for this to happen, even when we use the correct methodology and methods. One reason is that an ICU stay can pertain to more than one phenotype that co-exists sharing partial physiological features in common. For example, at the early stages of an ICU stay, patient glycemic status can be highly non-stationary due to critical illness treatments that disturb renal functionality, and hence affect insulin degradation. The other reason is because we implemented this methodology to compute phenotypes that represent physiology in some sense as a continuum–a continuous variable or function, allowing for phenotypic characteristics to overlap. Therefore, the difference between phenotypes may not be sharp, although the methodology indeed computes the right phenotypes. The non-mutually exclusive phenotypes are a result of balancing both biological fidelity and model flexibility, where we maximize the likelihood of having a unique model solution–using a flexible model–at the expense of methodologically clustering parameter estimates of a single ICU stay into multiple phenotypes.

Incorrect methodological choices in computing phenotypes can produce two other means of identifying non-mutually exclusive phenotypes. One possibility is that the chosen mechanistic model might not fully represent the physiology, e.g., the Ultradian we used to extract physiology may be simpler than required. The second possibility is an inappropriate MCMC method choice under finite computing time and resources, which leads to errors in parameter estimates due to a lack of convergence in the Markov chains to a unique equilibrium.

4.4 A Deeper Dive into Face Validity

We established the clinical context of the model-derived phenotypes by extracting underlying model-estimated physiology for phenotypic interpretations. Under this contextualization we then validated the interpreted phenotypes by mapping them to clinically observed data profiles. Measuring the consistency between model-based phenotypes and clinical diagnosis at the phenotype level through face validation, however, can be problematic. The clinician might have a relatively low level of confidence in categorizing the patients, as a patient might only satisfy certain aspects of the descriptions. Therefore, we disentangled this complex phenotyping problem by validating parameter estimates qualitatively rather than phenotypes.

The validation process at the parameter level is also complicated by language translation issues. The discordance in delineation and quantification between knowledge domains, e.g., our current definition of an average ICU phenotype is different from a typical clinician mindset, implies an unaddressable domain-specific gap that is difficult to detect and also a less related topic in this paper. This discordance also complicates the long-term object of this work to guide clinical decision support or to be implemented into health systems for clinicians as users. This problem Moreover, the disconnect between the language of physiology and language of clinical diagnosis, e.g., the mapping between physiology and SNOMED CT can lead to error and bias. For now this translation problem is very complex but might be treatable through the framework of human-computer-interaction (HCI) for future clinical decision support. Under the HCI framework, related work on this translation problem can motivate research in cognitive science, behavioral science in biomedical informatics, etc., that mainly hinge on users' intention of usage [53].

Here we computed the face validity concordance at parameter level. The average concordance of each parameter are: 83% ± 27% for a_1 (insulin secretion), 66% ± 35% for R_g (insulin insensitivity), and 52% ± 44% for t_p (insulin degradation), respectively. The decreasing concordance with increasing uncertainty in sequence for parameter a_1 (83% ± 27%) to t_p (52% ± 44%) was probably due to an increasing amount of health care interventions and potential confounding factors. For example, hemodialysis treatments confound our

estimation of insulin degradation. Among the 20 patients investigated during face validation, 15 of them either had hemodialysis treatment or continuous veno-venous hemofiltration treatment, or had kidney failure related conditions that would require dialysis treatment during the ICU stay. We also hypothesized that the insulin dosage in EHR data enabled us to identify model parameter R_g that relates to insulin insensitivity.

5 Acknowledgments

The work is supported by US National Institutes of Health (NIH)/National Library of Medicine # 5R01LM012734.

A Appendix

A.1 The Ultradian Model Mathematical Formulas

The mechanistic Ultradian model has six ordinary differential equations with three physiologic states (I_p is plasma insulin, I_i is remote insulin, G is plasma glucose) and a three stage filter (h_1, h_2, h_3) that represents how plasma insulin affects glucose.

$$\frac{dI_p}{dt} = \frac{R_m}{1+exp(\frac{-G}{V_g C_1}+a_1)} - (\frac{I_p}{V_p} - \frac{I_i}{V_i})E - \frac{I_p}{t_p} \tag{1}$$

$$\frac{dI_i}{dt} = (\frac{I_p}{V_p} - \frac{I_i}{V_i})E - \frac{I_i}{t_i} \tag{2}$$

$$t_d\frac{dh_1}{dt} = I_p - h_1 \tag{3}$$

$$t_d\frac{dh_2}{dt} = h_1 - h_2 \tag{4}$$

$$t_d\frac{dh_3}{dt} = h_2 - h_3 \tag{5}$$

$$\frac{dG}{dt} = \frac{R_g}{1+exp[(\alpha(\frac{h_3}{C_5 V_p}-1))]} - U_b[1-exp(\frac{-G}{C_2 V_g})] - \frac{1}{C_3 V_g}(U_0 + \frac{U_m - U_0}{1+(\frac{1}{C_4}(\frac{1}{V_i}+\frac{1}{Et_i})I_i)^{-\beta}})G + I_G(t) \tag{6}$$

$$\tag{7}$$

A.2 The Houlihan Algorithm for Model Parameter Selection

The Houlihan algorithm uses a specific combination of feature metrics and influence function as a selection method to rank order single parameter and parameter pairs/triplet [34]. Under the Houlihan approach, parameters a_1, t_p, and R_g were found to have the highest-rank individual influences on model fidelity, while parameter pairs of (a_1, C_1), (R_g, C_3) and triplet of (a_1, C_1, t_p), (a_1, C_1, C_3) were also ranked high. The method also justified reducing model parametrization to only a few core parameters set nominees: (a_1, t_p, R_g, C_1), (a_1, t_p, R_g, C_3) and (a_1, t_p, R_g) became our model parameter selection nominees, noting that both (a_1, t_p, R_g) and (C_1, t_p, R_g) correspond to representing the kidney-liver-pancreas subsystem. According to Sturis' Ultradian model with the physiology encoded in parameters [31], other parameters besides a_1, t_p, R_g, C_1, and C_3 are likewise physiologically interpretable, but in this paper we focused only on these parameters with the highest influence.

A.3 Geweke Statistic Fractions Identification

We used Geweke statistic as one of the objective quality criteria to screen Markov chains for intra-chain convergence under limited computing resources. The convergence in Geweke statistic is assessed by comparing the initial and terminal fractions of a Markov chain. Since we normally start our iterations somewhere not close to the true value, for a coherent chain we expect to see distinctive difference between the terminal and initial portion of iterations, which indicates some degree of convergence approaching the tail of iterations. So we only kept chains with significant p-values in Geweke statistic under the significance level α=0.06. Although Geweke suggested using 10% and 50% as the initial and terminal fractions, it is not a burden grid searching for the best in a careful tuning exploration. We explored in Figure A.1 different values for the initial fraction while fixing the terminal fraction to be a relatively large ratio, as we would like to only keep chains that converge and also stay stable at the tail. We identified the best Geweke fractions as 3% for the initial and 50% for the terminal iterates.

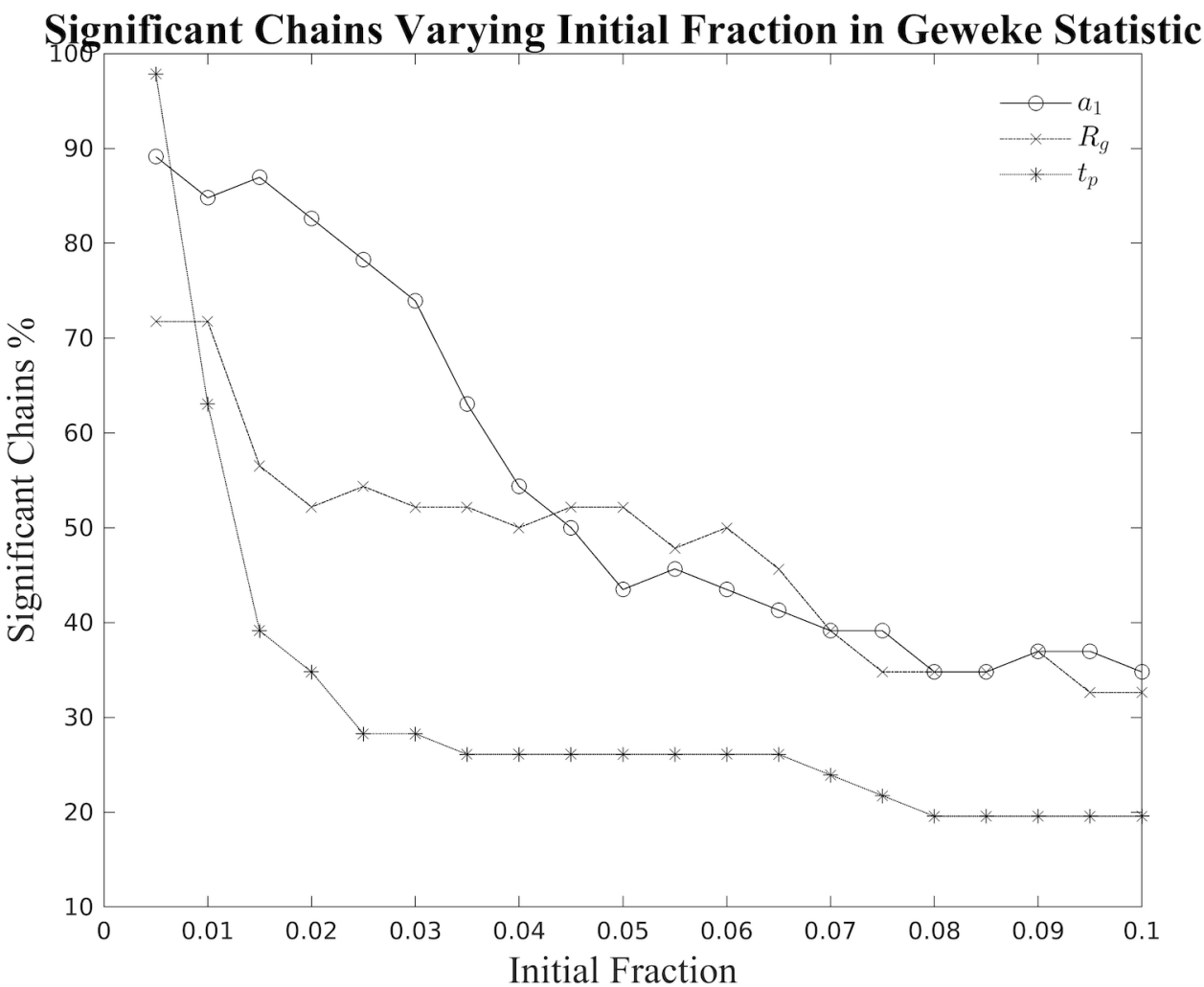


FIGURE A.1: Fraction identification of Geweke Statistic. Using cohort A data, we ran 3 Markov chains for the parameter triplet (a_1, R_g, T_p) then counted percentage of converged chains for each parameter using Geweke statistic with the initial fraction ranging from 0.05 to 0.1 and the terminal fraction fixed at 0.5. We chose the initial fraction as the first smallest value that decreased the convergence percentage for all parameters, hence we identify the initial fraction as 0.03. With this initial fraction, 47 out of 90 chains converged in Geweke statistic. Chains selection optima from cohort A were adopted for cohort B.

Selected Laboratory Measurements in EHR for External Validation	
Laboratory Measurements	Related Physiological Functionality
Amylase (AMY)	Pancreatic enzyme production for digestion (an indicator of insulin secretion)
Aspartate aminotransferase (AST)	Liver problems
Lipase (LIPA)	Pancreatitis and other pancreatic disorders
Creatinine (CREAT)	Kidney filtering capability in general (an indicator of insulin degradation under a non-dialysis condition)
Blood urea nitrogen (BUN)	Kidney filtering capability in general (an indicator of insulin degradation under a non-dialysis condition)
Alanine aminotransferase (ALT)	Signs of jaundice indicating liver/kidney dysfunction
Albumin (ALB)	Part of a comprehensive metabolic panel of liver and kidney
Bilirubin (BIL)	Liver damage or disease

TABLE A.1: Selected laboratory measurements as external data and their related physiological functionalities encoded in model parameter triplet (a_1, t_p, R_g).

A.4 Customized Ranking of Model Parameters for Phenotype Interpretation

Based on the parameter estimation of the parameters triplet (a_1, t_p, R_g), the customized ranking method compares each group's posterior mean to the posterior mean across all groups identified in a specific window, e.g., the last three days of the complete ICU stay. Then setting the posterior mean across all identified groups as the midpoint, group mean that falls in (0,1/3*midpoint) was ranked as 'much lower' and noted as '↓↓↓'; (1/3*midpoint, 2/3*midpoint) was 'moderately lower' and '↓↓'; (2/3*midpoint, midpoint) was 'slightly lower' and '↓'; (midpoint, 4/3*midpoint) was 'slightly higher' and '↑'; (4/3*midpoint, 5/3*midpoint) was 'moderately higher' and '↑↑'; (5/3*midpoint, 2*midpoint) was 'much higher' and '↑↑↑'.

A.5 List of ICD 9/10 Codes and Laboratory Measurements for External Validation

The list of ICD codes we picked to indicate clinically observable evidence includes keywords of 'hyperglycemia', 'sepsis or septicemia', 'renal failure or kidney failure', 'renal', and 'dialysis'. We tabulated in Table A.1 the selected laboratory measurements that are related to pancreas, kidney, and liver functions in general for external validation. Noting here that we only selected laboratory measurements not used in phenotype computation to externally validate the computed phenotypes. For example, we did not select glucose laboratory measurements to validate the phenotype clinical consistency in terms of hyperglycemia, because we used glucose in the phenotype computation as required by the model estimation.

A.6 Updated Phenotype Projection Excluding T1DM Patients

Referring to Section 3.2.3, this section presents in Figure A.2 the updated visualization of patients projection in the 2D t-SNE space and the further clustering of these non-T1DM patients into mechanistically distinctive phenotypes with SVC.

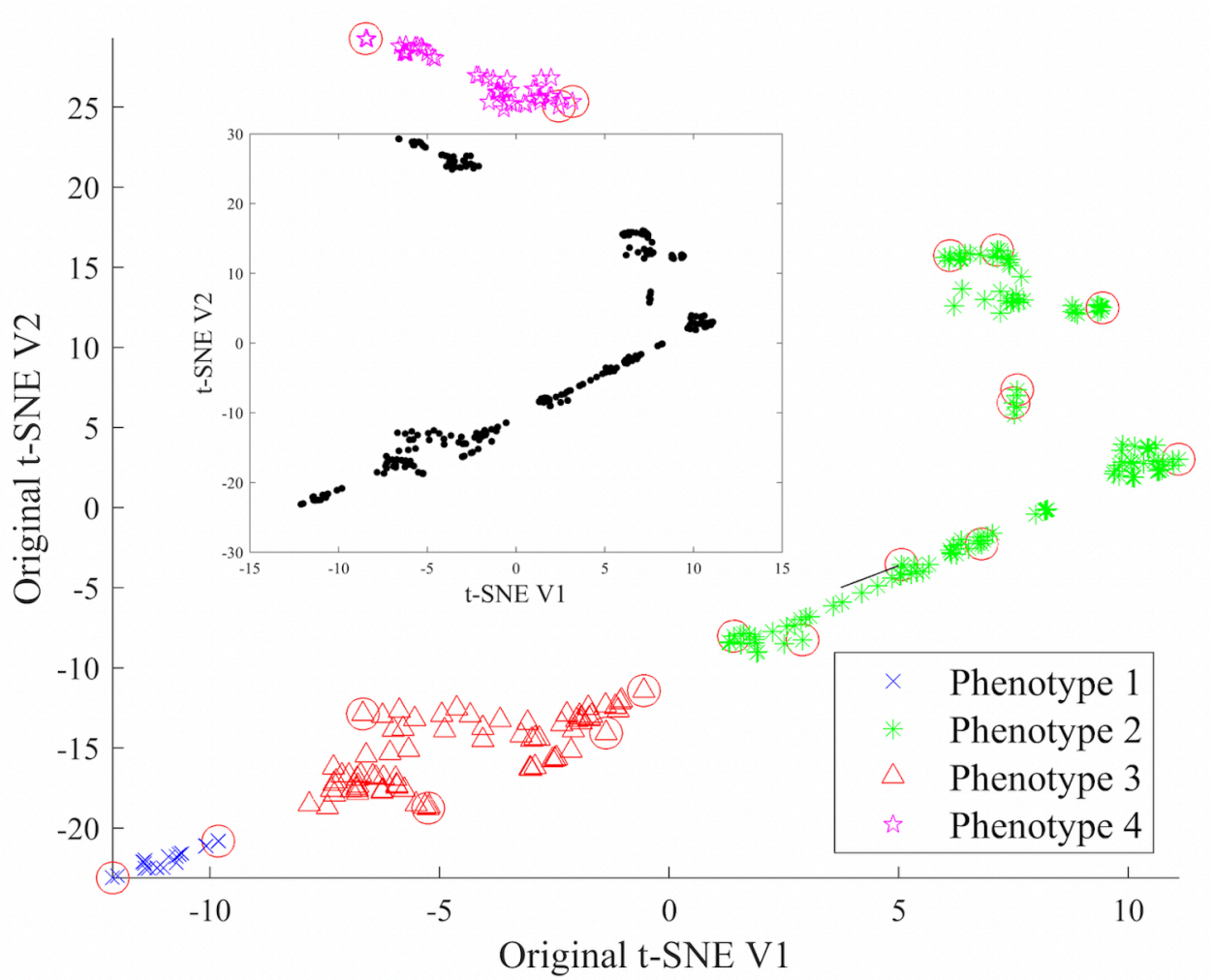


FIGURE A.2: **Step 1 (inset)**: Raw and un-clustered t-SNE dimension reduction of selected chains summary of 38 ICU stays in cohort B3 excluding T1DM using the same methodological choices. **Step 2**: Phenotype computation with SVC that clusters t-SNE coordinates of chains summaries. Circled coordinates are support vectors that lie on the feature space sphere boundaries. Four clusters were identified indicating four potential phenotypes.

References

[1] G. Hripcsak and D. J. Albers. "Next-generation phenotyping of electronic health records". In: *Journal of the American Medical Informatics Association* 20.1 (2013). OA status: bronze, pp. 117–121. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2012-001145. URL: https://academic.oup.com/jamia/article-pdf/20/1/117/9516924/20-1-117.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3555337/pdf/amiajnl-2012-001145.pdf.

[2] Catherine A Mccarty et al. "The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies". In: *BMC Medical Genomics* 4.1 (2011), p. 13. ISSN: 1755-8794. DOI: 10.1186/1755-8794-4-13. URL: https://bmcmedgenomics.biomedcentral.com/track/pdf/10.1186/1755-8794-4-13.

[3] A. N. Kho et al. "Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium". In: *Science Translational Medicine* 3.79 (2011). OA status: green_accepted, 79re1–79re1. ISSN: 1946-6234. DOI: 10.1126/scitranslmed.3001807. URL: http://europepmc.org/articles/pmc3690272?pdf=render.

[4] Syed Hasib Akhter Faruqui et al. "Development of a Deep Learning Model for Dynamic Forecasting of Blood Glucose Level for Type 2 Diabetes Mellitus: Secondary Analysis of a Randomized Controlled Trial". In: *JMIR mHealth and uHealth* 7.11 (2019), e14452. ISSN: 2291-5222. DOI: 10.2196/14452.

[5] David Zeevi et al. "Personalized Nutrition by Prediction of Glycemic Responses". In: *Cell* 163.5 (2015). OA status: bronze, pp. 1079–1094. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.11.001.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022. ISSN: 1532-4435. URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com.

[7] Suchi Saria, Daphne Koller, and Anna Penn. "Learning individual and population level traits from clinical temporal data". In: *Proceedings of Neural Information Processing Systems*. Citeseer, pp. 1–9.

[8] Rachel L Richesson et al. "A comparison of phenotype definitions for diabetes mellitus". In: *Journal of the American Medical Informatics Association* 20.e2 (2013). OA status: green_published, e319–e326. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2013-001952. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3861928.

[9] Rimma Pivovarov. *Electronic health record summarization over heterogeneous and irregularly sampled clinical data*. Columbia University, 2016. ISBN: 1339300788. DOI: 10.7916/D89W0F6V.

[10] D. J. Albers et al. "Dynamical Phenotyping: Using Temporal Analysis of Clinically Collected Physiologic Data to Stratify Populations". In: *PLoS ONE* 9.6 (2014). OA status: green_published, e96443. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0096443. URL: http://europepmc.org/articles/pmc4059642?pdf=render.

[11] Sarah Collins Rossetti et al. "Healthcare Process Modeling to Phenotype Clinician Behaviors for Exploiting the Signal Gain of Clinical Expertise (HPM-ExpertSignals): Development and evaluation of a conceptual framework". In: *Journal of the American Medical Informatics Association* 28.6 (2021), pp. 1242–1251. ISSN: 1527-974X. DOI: 10.1093/jamia/ocab006.

[12] Denis Agniel, Isaac S Kohane, and Griffin M Weber. "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study". In: *Bmj* 361 (2018). ISSN: 0959-8138. DOI: 10.1136/bmj.k1479.

[13] Rimma Pivovarov et al. "Identifying and mitigating biases in EHR laboratory tests". In: *Journal of biomedical informatics* 51 (2014), pp. 24–34. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2014.03.016.

[14] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. "Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data". In: *PLoS ONE* 8.6 (2013). OA status: green_published, e66341. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0066341. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3691199.

[15] Martin R Cowie et al. "Electronic health records to facilitate clinical research". In: *Clinical Research in Cardiology* 106.1 (2017), pp. 1–9. ISSN: 1861-0692. DOI: 10.1007/s00392-016-1025-6.

[16] David 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". In: *Journal of the American Medical Informatics Association* 25.10 (2018), pp. 1392–1401. ISSN: 1067-5027. DOI: 10.1093/jamia/ocy106.

[17] Mark Asch, Marc Bocquet, and Maëlle Nodet. *Data assimilation: methods, algorithms, and applications*. SIAM, 2016. ISBN: 1611974534. DOI: 10.1137/1.9781611974546.

[18] Sebastian Reich and Colin Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015. ISBN: 1107069394. DOI: 10.1017/CBO9781107706804.

[19] Geir Evensen. *Data assimilation: the ensemble Kalman filter*. Vol. 2. Springer, 2009. DOI: 10.1007/978-3-642-03711-5.

[20] Kody Law, Andrew Stuart, and Kostas Zygalakis. "Data assimilation". In: *Cham, Switzerland: Springer* 214 (2015). DOI: 10.1007/978-3-319-20325-6.

[21] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1 (1960), pp. 35–45. ISSN: 0021-9223. DOI: 10.1115/1.3662552.

[22] Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman filter: Particle filters for tracking applications*. Artech house, 2003. ISBN: 1580538517.

[23] Simon J Julier and Jeffrey K Uhlmann. "Unscented filtering and nonlinear estimation". In: *Proceedings of the IEEE* 92.3 (2004), pp. 401–422. ISSN: 0018-9219. DOI: 10.1109/JPROC.2003.823141.

[24] Sven Zenker, Jonathan Rubin, and Gilles Clermont. "From inverse problems in mathematical physiology to quantitative differential diagnoses". In: *PLoS computational biology* 3.11 (2007), e204. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1007155.

[25] Arianna Dagliati et al. "A process mining pipeline to characterise COVID-19 patients' trajectories and identify relevant temporal phenotypes from EHR data". In: *Frontiers in Public Health* (2022), p. 1331. ISSN: 2296-2565.

[26] Hossein Estiri, Zachary H Strasser, and Shawn N Murphy. "High-throughput phenotyping with temporal sequences". In: *Journal of the American Medical Informatics Association* 28.4 (2021), pp. 772–781. ISSN: 1527-974X.

[27] Weilin Meng et al. "Temporal phenotyping by mining healthcare data to derive lines of therapy for cancer". In: *Journal of Biomedical Informatics* 100 (2019), p. 103335. ISSN: 1532-0464.

[28] Juan Zhao et al. "Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: Cardiovascular disease case study". In: *Journal of biomedical informatics* 98 (2019), p. 103270. ISSN: 1532-0464.

[29] Sebastian Mate et al. "A method for the graphical modeling of relative temporal constraints". In: *Journal of biomedical informatics* 100 (2019), p. 103314. ISSN: 1532-0464.

[30] Karel GM Moons et al. "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration". In: *Annals of internal medicine* 162.1 (2015), W1–W73. ISSN: 0003-4819. DOI: 10.1002/bjs.9736.

[31] J. Sturis et al. "Computer model for mechanisms underlying ultradian oscillations of insulin and glucose". In: *American Journal of Physiology-Endocrinology and Metabolism* 260.5 (1991), E801–E809. DOI: 10.1152/ajpendo.1991.260.5.E801. URL: https://journals.physiology.org/doi/abs/10.1152/ajpendo.1991.260.5.E801.

[32] C. Simon and G. Brandenberger. "Ultradian Oscillations of Insulin Secretion in Humans". In: *Diabetes* 51.Supplement 1 (2002), S258–S261. ISSN: 0012-1797. DOI: 10.2337/diabetes.51.2007.s258.

[33] Oana-Teodora Chis, Julio R. Banga, and Eva Balsa-Canto. "Structural Identifiability of Systems Biology Models: A Critical Comparison of Methods". In: *PLoS ONE* 6.11 (2011), e27755. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027755.

[34] David J. Albers et al. "The parameter Houlihan: A solution to high-throughput identifiability indeterminacy for brutally ill-posed problems". In: *Mathematical Biosciences* 316 (2019). OA status: green_accepted, p. 108242. ISSN: 0025-5564. DOI: 10.1016/j.mbs.2019.108242. URL: https://authors.library.caltech.edu/98225/1/1902.01978.pdf.

[35] S. L. Cotter et al. "MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster". In: *Statistical Science* 28.3 (2013), pp. 424–446. ISSN: 0883-4237. DOI: 10.1214/13-sts421.

[36] Gelman A et al. *Bayesian Data Analysis*. 3rd. Boca Raton, FL: CRC Press, 2014. DOI: 10.1201/9780429258411.

[37] Steve Brooks et al. *Handbook of markov chain monte carlo*. CRC press, 2011. ISBN: 1420079425. DOI: 10.1111/biom.12099.

[38] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012. ISBN: 144713267X. DOI: 10.1007/978-1-4471-3267-7.

[39] Gareth O. Roberts and Jeffrey S. Rosenthal. "General state space Markov chains and MCMC algorithms". In: *Probability Surveys* 1.none (2004), pp. 20–71. ISSN: 1549-5787. DOI: 10.1214/154957804100000024.

[40] John F Geweke. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Report. Federal Reserve Bank of Minneapolis, 1991.

[41] Stephen P. Brooks and Andrew Gelman. "General Methods for Monitoring Convergence of Iterative Simulations". In: *Journal of Computational and Graphical Statistics* 7.4 (1998), pp. 434–455. ISSN: 1061-8600. DOI: 10.1080/10618600.1998.10474787.

[42] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[43] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. "Challenges in unsupervised clustering of single-cell RNA-seq data". In: *Nature Reviews Genetics* 20.5 (2019), pp. 273–282. ISSN: 1471-0056.

[44] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.11 (2008). ISSN: 1532-4435. URL: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl.

[45] George C Linderman and Stefan Steinerberger. "Clustering with t-SNE, provably". In: *SIAM Journal on Mathematics of Data Science* 1.2 (2019), pp. 313–332. ISSN: 2577-0187. DOI: 10.1137/18M1216134.

[46] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. "How to use t-SNE effectively". In: *Distill* 1.10 (2016), e2. ISSN: 2476-0757. DOI: 10.23915/distill.00002. URL: http://distill.pub/2016/misread-tsne.

[47] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014. ISBN: 1498759718.

[48] Asa Ben-Hur et al. "Support vector clustering". In: *Journal of machine learning research* 2.Dec (2001), pp. 125–137. URL: https://www.jmlr.org/papers/volume2/horn01a/horn01a.pdf.

[49] Jaewook Lee and Daewon Lee. "An improved cluster labeling method for support vector clustering". In: *IEEE Transactions on pattern analysis and machine intelligence* 27.3 (2005), pp. 461–464. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2005.47.

[50] Jaewook Lee and Daewon Lee. "Dynamic characterization of cluster structures for robust and inductive support vector clustering". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.11 (2006), pp. 1869–1874. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2006.225.

[51] Michail I Schlesinger and Václav Hlavác. *Ten lectures on statistical and structural pattern recognition*. Vol. 24. Springer Science and Business Media, 2013. ISBN: 9401732175. DOI: 10.1007/978-94-017-3217-8.

[52] Stephen P Turner. "The concept of face validity". In: *Quality and Quantity* 13.1 (1979), pp. 85–90. ISSN: 1573-7845. DOI: 10.1007/BF00222826.

[53] Vimla L Patel, David R Kaufman, and Thomas Kannampallil. "Human-Computer Interaction, Usability, and Workflow". In: *Biomedical Informatics*. Springer, 2021, pp. 153–175. DOI: 10.1007/978-3-030-58721-5_5.