Generating synthetic data with a mechanism-based Critical Illness Digital Twin: Demonstration for Post Traumatic Acute Respiratory Distress Syndrome

Chase Cockrell (1), Seth Schobel-McHugh (2), Felipe Lisboa (2), Yoram Vodovotz (3) and Gary An* (1)

1. Department of Surgery, University of Vermont Larner College of Medicine
2. Department of Surgery, Uniformed Services University of the Health Sciences/Walter Reed National Military Medical Center
3. Department of Surgery, University of Pittsburgh

Keywords: Synthetic data, agent-based model, sepsis, multiple organ failure, cytokines, machine learning, adult respiratory distress syndrome, multiscale model

*Corresponding Author:
89 Beaumont Ave
Given 319D
Burlington, VT 05405
gan@med.uvm.edu

Abstract

Machine learning (ML) and Artificial Intelligence (AI) approaches are increasingly applied to predicting the development of sepsis and multiple organ failure. While there has been success in demonstrating the clinical utility of such systems in terms of affecting various outcomes, there are fundamental challenges to the ML/AI approach in terms of improving the actual predictive performance and future robustness of such systems. Given that one of the primary proposed avenues for improving algorithmic performance is the addition of molecular/biomarker/genetic features to the data used to train these systems, the overall sparsity of such available data suggests the need to generate synthetic data to aid in training, as has been the case in numerous other ML/AI tasks, such as image recognition/generation and text analysis/generation. We propose the need to generate synthetic molecular/mediator time series data coincides with the advent of the concept of medical digital twins, specifically related to interpretations of medical digital twins that hew closely to the original description and use of industrial digital twins, which involve simulating multiple individual twins from a common computational model specification. Herein we present an example of generating synthetic time series data of a panel of pro- and anti-inflammatory cytokines using the Critical Illness Digital Twin (CIDT) regarding the development of post-traumatic acute respiratory distress syndrome.

1.0 Introduction

Predicting the course of sepsis and multiple organ failure remains a challenge. More recently machine learning (ML) methods have been applied to predict the onset of sepsis, such as the TREWS (Targeted Real-time Early Warning System) algorithm, which was able to impact clinical care by improving outcomes, primarily via the earlier administration of antibiotics[1, 2]. However, despite its promise, there are known fundamental and intrinsic limitations to ML, primarily related to challenges in generalizing to new use-contexts/cases and an inevitable performance degradation over time due to *data drift* [3-6]. In short, data drift refers to an evolution or change in the factors and processes that generate the observations/data upon which the ML algorithm operates. An example of contextual data drift (i.e., accounting for differential practices, standards, or epidemiological risks at various medical centers) would be an attempt to apply a ML algorithm trained at a particular series of institutions to a new institution with a known difference in type of microbial nosocomial infections; an example of temporal data drift would be the adoption of a new care pathway, diagnostic test, or treatment algorithm. While increasing the breadth of patient data contexts (e.g. new and varied practice sites) available for model training is the most common proposed response to addressing contextual generalizability (an approach that was effectively applied in the TREWS studies), there is no currently accepted method to account for temporal data drift other than retraining. While many have consigned themselves to the perpetual need to retrain these systems (and have developed processes that will help accomplish this while minimizing significant clinical impact)[7], there is an alternative approach drawn from industrial/technological fields that can potentially enhance the efficiency and effectiveness of utilizing existing data to increase generalizable prediction. This approach is the concept of a *digital twin*, which refers to an individualized computational representation from within a specific class of objects or type of industrial process. This term is derived from industrial/material applications and, based on the original industrial definition, a digital twin consists of three components[8]:

1) A data structure for the real-world system,
2) A computational representation of some process that links data points together to form dynamics (a *simulation*), and
3) A link to the real-world system that feeds data back into the data-propagation/generation process to produce an individualized trajectory for the specific object/process being twinned.

These three components allow the digital twin to predict/forecast (in an individualized/personalized fashion) the future history of the specific twinned objects and identify specific failure points (e.g., in a hypothetical biomedical application, *when/if* the patients will develop ARDS). Further, assessment of a digital twin that follows the classical description (e.g. it can be *simulated*) can help define potential corrective actions to be taken in real-time (e.g., *how* to prevent the patient from developing ARDS). Central to this description of a digital twin within its original industrial context is that there is a *computational specification* for the class of twinned object, from which individual instances of digital twins can be created, simulated and used.

The concept of a "medical digital twin"[9-13] embodies the idealized goals of personalized and precision medicine: identifying an individual patient's risk of disease, a time-horizon in which clinical consequences can manifest, and potentially offering therapeutic measures to forestall a poor outcome; however at this early stage there are several different concepts for what actually constitutes a medical digital twin[9-13]. *For our purposes, we adopt a definition that follows closely the original definition of an industrial digital twin, specifically recognizing that inherent in the concept of industrial digital twins is the ability to simulate multiple possible future behavior(s) of the twin where individual simulations arise from a common shared computational specification/model.* This set of behaviors can then be refined by the data connection between the real world and the digital twin and used to narrow the forecasted set of possible futures for the twinned individual (in a fashion analogous to the way putative hurricane paths are presented as a rolling forecasting "cone" that is updated with new data). This need to *simulate* the digital twin points to the

recognition within the industrial application that data-centric means of linking together time series data alone (including ML) are generally insufficient to project future trajectories[14].

We propose that the digital twin concept be applied to the critically ill/intensive care unit population, and in this work, we aim to create a Critical Illness Digital Twin (CIDT) that can be used to augment the development of ML-models that predict sepsis and multiple organ dysfunction by enhancing their robustness, performance, and generalizability. The CIDT integrates simulation models that embody mechanistic knowledge regarding systemic inflammation following burn injury with ML/Artificial Intelligence (AI) methods for calibration and parameterization to more comprehensively, realistically, and functionally encompass the heterogeneity seen in clinical populations; this capability is predicated on the fact that in order to create a digital twin of *any* individual we must be able to create a digital twin of *every* individual. The integration of these complementary methods of mechanistic simulation and ML analysis will enhance the robustness and generalizability of the ML/AI predictive models by generating bio-realistic synthetic data to supplement their training (e.g., generating populations of digital twins). This process is analogous to the ubiquitous use of synthetic training data in high impact ML/AI domains such as image or speech recognition[15], but cognizant of the unique challenges inherent to time series molecular/biomarker data. Specifically, to our knowledge, there are no published instances of attempts to generate synthetic multiplex cellular/molecular-level profiling time series data. It is also recognized that the simulation model that underlies the CIDT need not be comprehensive (which is also the case of industrial digital twins); rather, it needs to be sufficient to reliably reproduce the heterogeneity of behavior/output seen across a clinical population. In prior work, we have developed ML methods that achieve the above-described challenge[16, 17].

The underlying premise to this integrative approach to the CIDT is that humans all share some core fundamental biological processes that are conserved across individuals but may manifest in different trajectories based on genetic/epigenetic/environmental differences that affect the functional responsiveness of their cellular and molecular pathways; this exactly represents the role of a common computational specification in the digital twin concept. It is this variation of mechanistic responsiveness that contributes to the differences between how individuals respond to similar insults; this is exactly the process that turns the common specification model for a class of objects into a specific digital twin. Computationally, this manifests as different parameterizations of a common model structure used to represent the variation across a population (one such example of this concept in biomedical practice is the creation of simulated patient cohorts to computationally study differential drug kinetics in the field of Quantitative Systems Pharmacology[18, 19]; another is our use of in silico trials to evaluate and discover what mechanistic interventions, representing potential drugs, could be used to treat sepsis[20-23]). To reiterate: this approach is also consistent with the industrial concept of a digital twin, where the model structure represents the formal simulation specification of the targeted object/industrial process, and a specific "twin" is created by personalizing the parameterization (and updating) of that simulation model. For purposes of using the CIDT to aid in the development of a ML/AI prediction algorithm, simulating populations of CIDTs will produce a far larger and more expansive number of training trajectories than would ever be able to be collected clinically, allowing the ML/AI a fuller range of behavioral possibilities that would encompass multiple potential application/use contexts (i.e., different populations in different institutions); this process would be vital in being able to account for subpopulations that might be under-represented in existing data sets and therefore could be under-recognized/appreciated in classical ML training. As noted above, this process of synthetically creating a much larger training data set than would be possible in terms of real world data collection is analogous to the use of synthetically generated data in high success ML applications like image analysis and natural language processing.[15] An added benefit of the mechanism-based simulation component of the CIDT is the future ability to use the CIDT for therapeutic discovery and personalized treatment optimization.[21, 24]

2.0 Methods:

In this work, we underlying common computational specification model for the CIDT is the previously developed and validated Innate Immune Response Agent-based Model (IIRABM)[17, 20, 21, 25]. In this abstraction, individual cells perform actions (i.e., apoptosis, phagocytosis, or cytokine production/secretion) using rules derived from the available literature; these rules that govern cellular actions are what we term *mechanisms* for describing our models.

2.1 Data Source

The Uniform Services University/Walter Reed National Medical Military Center provided data on 199 trauma patients, 92 of which developed ARDS at some point during the course of their hospitalization, matched with 107 controls that did not develop ARDS. Data elements that were used to calibrate the model include two primary elements: 1) vital signs/laboratory observables necessary to determine SOFA score; for the respiratory compartment, this consists of the partial pressure of oxygen, complete information regarding respiratory support, and blood oxygen saturation, representing aggregate organ function; and 2) time-series blood-serum cytokine profiles consisting of IL-1β, IL-1ra, IL-6, IL-4, IL-8, IL-10, G-CSF, IFNγ, and TNFα, sampled periodically for the duration of the patient's hospitalization. To demonstrate our methodology for generating synthetic data with the CIDT, we use simulations of the IIRABM to identify parameterizations (which represent varied individuals responding to varied insults) that cannot be falsified by the available data. Using the nested GA parameter discovery and AL parameterization boundary identification noted below we determined three sets of model parameterizations: 1) patients that will certainly develop ARDS; 2) patients that certainly will not develop ARDS; and 3) patients that have a probability greater than zero, but less than one, of developing ARDS.

2.2 Calibration Procedure

The CIDT is calibrated and validated in a two-step procedure. First, Genetic Algorithms (GA) are used to discover the model parameterizations that best fit the full range of data. Next, an Artificial Neural Network (ANN) is trained using Active Learning (AL) to take model parameterizations as input and classify if they are valid (whether or not they are invalidated by available data); the AL procedure then allows for the discovery of boundaries and parameter rangers around GA-discovered parameterizations. A more detailed description of these processes are noted below.

2.2.1 Genetic Algorithms for Parameter Discovery

The GA[26-28] is a population-based metaheuristic optimization algorithm that is inspired by biological evolution. We have previously utilized GA both to develop personalized therapeutic strategies[21] and to calibrate a model of the human innate immune response, the IIRABM, to a burn data set, simultaneously discovering novel rules regions associated model parameterizations which led to biologically plausible output[17]. Ultimately, the goal of the GA is to evolve the model's rule set such that the solution replicates not only what is seen experimentally, but also its associated variance. The form of the fitness function we will use is:

$$F = \sum_{i,t} k_i D_{i,t},$$

where

$$D_{i,t} = \left|\max\left(d_{exp,i,t}\right) - \max\left(d_{comp,i,t}\right)\right| + \left|\min\left(d_{exp,i,t}\right) - \min\left(d_{comp,i,t}\right)\right|$$

and where $\max\left(d_{exp,i,t}\right)$ represents the maximum value (from all subjects at that time point) for the experimental measurement of data element i at time t; $\max\left(d_{comp,i,t}\right)$ represents the maximum value of

computational model output (from all stochastic replicate simulations) for data element i at time t; and the constants k_i allow for varying weights to be placed on individual data elements (i.e., we would not expect the raw difference in a protein concentration to be on the same scale as differential cell populations over time). By summing the differences between the maximum and minimum values of the computational model output and experimentally collected data, this instantiation of the GA seeks to find model parameterizations that generate a range of data which most closely matches the range of data experimentally and does so by operating on the model parameterization such that, when instantiated in the model, lead it to minimize the above fitness function. We note that initially, we will assign equal weights to each cytokine concentration, vital sign, or laboratory observable to be matched. As the process continues, weights may be adjusted to bias the algorithm to better fit certain aspects of the data, e.g., if the GA is slow to converge to an adequate solution that explains IL10 dynamics, this data element would be assigned a greater weight to increase its contribution to model training.

We note we are employing GA in a non-standard fashion, where rather than seeking a specific optimal parameterization of the model we are using the process of convergence of the GA to identify an ensemble of valid parameterizations. To construct this ensemble, we add an ensemble retainment criterion to the GA procedure, the rational for which is as follows: we recognize that any putative parameterization which generates cytokine trajectories that always lie within the cilnically observed range cannot be invalidated by the data, and are therefore biologically plausible; thus, these parameterizations should be retained for inclusion into the final ensemble. As the goal of the fitness function is to obtain maximum coverage over the clinical data range, some of these viable parameterizations may otherwise be lost as the population evolves.

2.2.2 Active Learning to Determine Boundaries of GA-Derived Candidate Parameterizations

Active learning (AL) is a sub-field of machine learning (ML) which focusses on finding the optimal selection of training data to be used to train a ML or statistical model.[29] AL can be used for classification[30, 31] or regression[32, 33]. AL is an ideal (in terms of total cost) technique for modeling problems in which there is a large amount of unlabeled data and manually labelling that data is expensive. In these circumstances (specifically the costly data labelling) *AL provides the most generalizable and accurate model for the cheapest cost.* The problem of combinatorial complexity in the selection of model parameters is well-established in the computational/biological modeling communities[34-38]. In previous work[25], we utilized high-performance computing to demonstrate the need for comprehensive "data coverage" among possible model states as well as the importance of internal parameter variation (as compared to model structure) to capture the full range of biological heterogeneity seen clinically. AL[16] rendered this problem computationally tractable by reducing the number of simulations require to characterize the search space by ~95% when compared to a brute-force exploration. Additionally, we demonstrated that more complex models with a larger number of variables may expect further improvements in efficiency[16].

The output of the AL workflow is an artificial neural network (ANN) classifier, for which the classes are biologically plausible (cannot be invalidated by the data) or not. After the application of the GA procedure, we will have an ensemble of biologically plausible model parameterizations. Each of these parameterizations represents a single point in a high-dimensional space. *AL will allow us to determine the biologically plausible parameter ranges around those single points in a computationally efficient manner.* We will begin the AL procedure by defining initial search boundaries, and we will discretize the search space lying within those boundaries, creating a set of putative model parameterizations with unknown classification. Each parameterization contained within the discretized search space will then be fed into the trained ANN, and have its associated class predicted. We then posit the existence of some function,

$$y = f(\vec{x}), x \in \chi \subset \mathbb{R}^n, y \in \mathbb{R},$$

which accepts as input a model parameterization and predicts the associated class, and that this function can be approximated given input data from the training set:

$$D_{train} = \{x_j^t, f(x_j^t)\}, \text{ for } j = 1, \dots, n,$$

where x_j^t represents a single parameterization used to train the ANN. The ANN model uses a binary cross-entropy [39] loss function, in which the loss is given by:

$$L = -\sum_{i=1}^{2} y_i \log(\hat{y}_i),$$

where y_i is the ground truth value and $\hat{y}_i$ is the ANN-approximated score. To generate the initial training set, 100 putative parameterizations will be selected from the discretized set. The IIRABM simulation then uses these putative parameterizations runs a fixed number of stochastic replicates of the input points to determine class membership. This information is then used to train the ANN classifier. The algorithm then ranks the remaining unlabeled parameterizations by class-membership uncertainty:

$$\{x_{i+1}\} = \min_x \left(0.5 - P_i(y|x)\right).$$

Those parameterizations whose class (biologically plausible or not) are most uncertain in the current ANN classifier model are then selected for labeling and the process repeats until a stopping criterion is reached.

3.0 Results:

In Figures 1 and 2, we demonstrate the preliminary generation of synthetic data for the lung compartment of the CIDT. In Figure 1, Panel A, colored points are the clinically collected data, with red points representing patients that developed ARDS sometime between days 6 and 10 and blue points representing patients that did not. The dotted lines (shown only in Panel A, TNFα for clarity) indicate representative clinical trajectories, which are sparse and typically only contain 3-4 measurements. The non-ARDS clinical trajectory, for example, only contains measurements at t=0, 3, and 7 days post-hospitalization, leaving significant uncertainty as to the patient state in intermediate times. This feature is present for nearly all mediators and patients in the currently available data. The shading indicates the boundaries of the model behavior space for the parameterizations that generate ARDS (red) or not (blue), with significant overlap between the two spaces. In Panel B, we show actual simulated trajectories for blood-serum concentration of TNFα, noting the similar, but much richer, dynamics to the clinical trajectories. The two patients whose trajectories are represented by the red lines ultimately develop ARDS, while the patient represented by the blue line does not. In Figure 2, we show the clinical data, simulation boundaries, and simulated trajectories for blood-serum GCSF and IL-10 in Panels A and B. In Panel C, we show the oxygen deficit trajectories, an inverse-measure of patient health, for 500 patients each from the parameterizations of the model that develop ARDS (red) and those that to not (blue), with SOFA scores correlated to the clinical data, using the Berlin definition of ARDS,[40] in which mild ARDS corresponds to a Respiratory SOFA score of 2, moderate ARDS corresponds to a Respiratory SOFA score of 3, and severe ARDS corresponds to a Respiratory SOFA score of 4. It should also be noted that different clinical patients may only manifest some of the mediator variations that can distinguish between ARDS and non-ARDS; therefore, this

heterogeneity of response can only be identified across a population of parameterized simulations that can generate the entire set of mediators in concert.

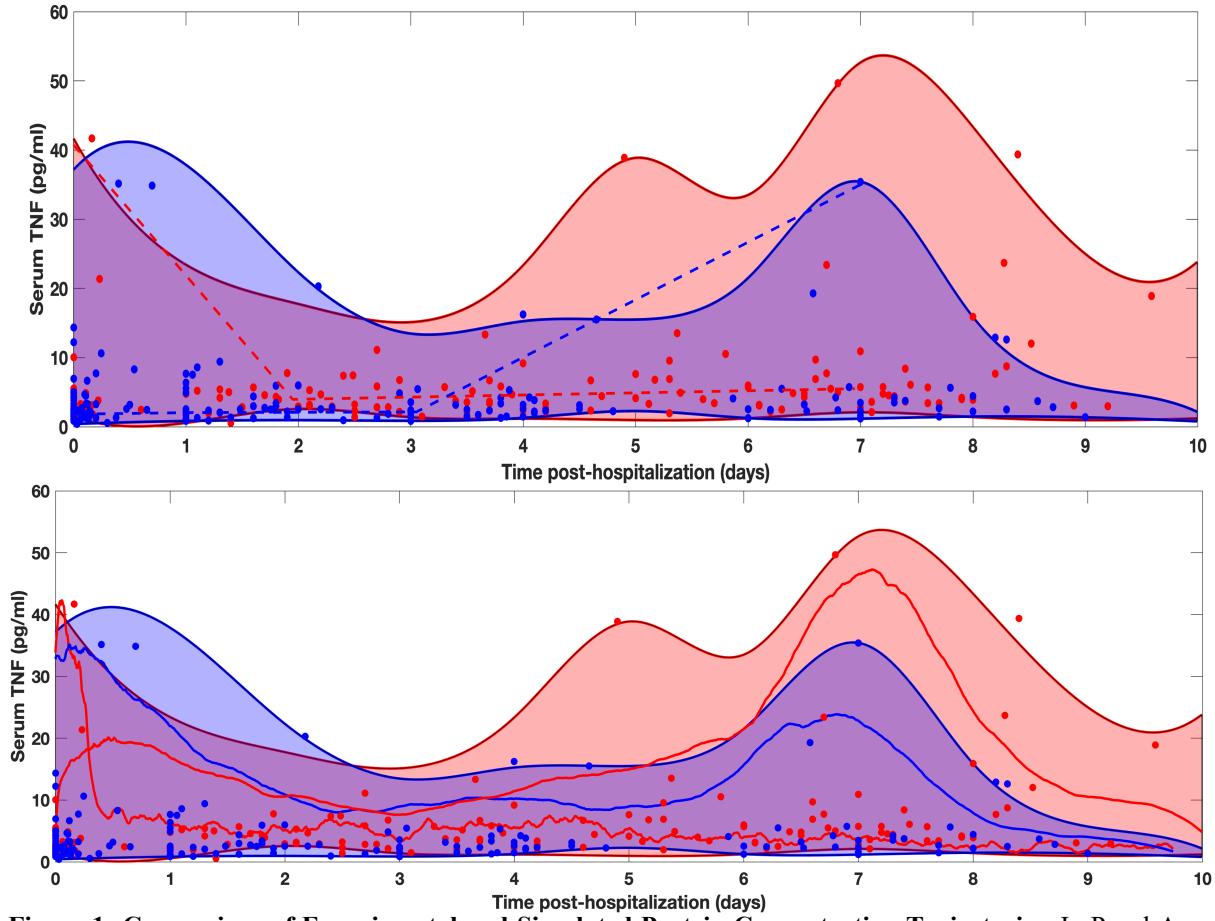

Figure 1: Comparison of Experimental and Simulated Protein Concentration Trajectories. In Panel A, we show both simulated and clinically collected blood-serum cytokine measurements for TNFα. Distinct points are the clinically collected data, with red points representing patients that developed ARDS sometime between days 7 and blue points representing patients that did not. The dotted lines indicate representative clinical trajectories. The shading indicates the boundaries of the model trajectory space for the parameterizations that generate ARDS (red) or not (blue). Note significant overlap between the two spaces given the overlap of data points. However, the key point is that differential parameterizations are able to identify clear regions that are unique to each group. In Panel B, the three solid lines (two red lines eventually developing ARDS and the blue line not) show actual simulated trajectories of TNFα blood-serum concentrations generated by the lung compartment of the CIDT.

4.0 Discussion:

While sepsis prediction algorithms, such as TREWS, have been demonstrated to improve outcome, presumed in the case of TREWS to be due primarily to the earlier administration of antibiotics, there is ongoing interest in being able to augment the physiological features used by these early warning systems by adding more detailed, biologically informative features that would help discriminate between the disease trajectories of individuals. This desire is consistent with the general trend towards more precise and personalized therapy. With respect to forecasting the potential progression to ARDS, numerous ML predictive algorithms that operate on data from the electronic health record [41, 42] (also see Ref [43] for

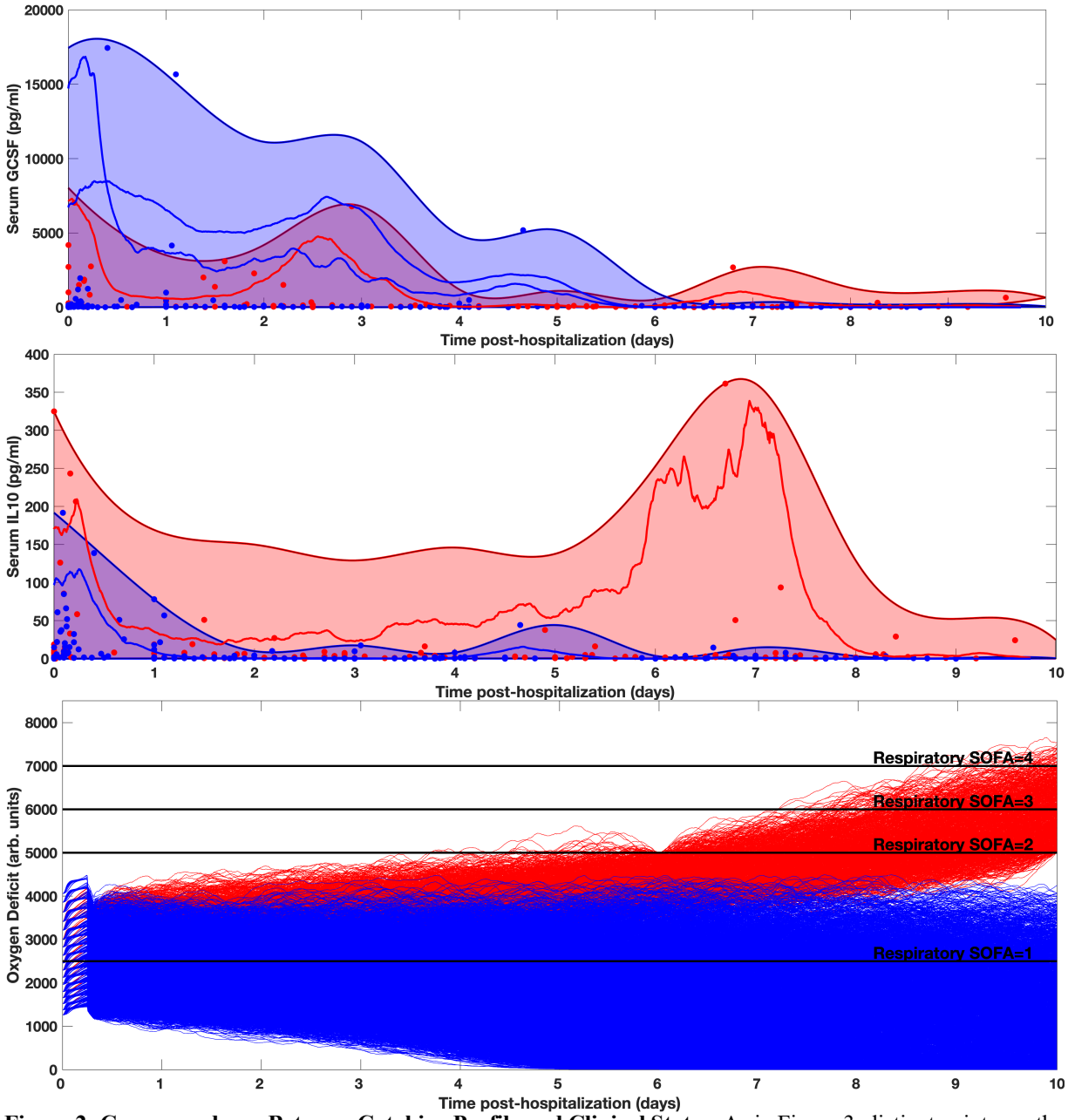

Figure 2: Correspondence Between Cytokine Profile and Clinical State. As in Figure 3, distinct points are the clinically collected data, with red points representing patients that developed ARDS sometime between days 7 and blue points representing patients that did not. The shading indicates the boundaries of the model behavior space for the parameterizations that generate ARDS (red) or not (blue). Simulated trajectories are shown for GCSF (Panel A) and IL10 (Panel B). In Panel C, we show the oxygen-deficit trajectories of simulated patients and their corresponding respiratory SOFA scores as they develop ARDS (increasing oxygen deficit) or progress towards successful healing of the lung injury.

a review) have been augmented with radiographic [44-46] and respiratory/ventilator waveform [47-49] data. As in other fields, such as precision oncology, there is also interest in being able to extend the feature set to include biomarkers and genetic/epigenetic information (see Ref [50] for a review). It is in this latter group that the ability to generate synthetic time series data becomes important using a mechanism-based Digital Twin. The reason for this is the Curse of Dimensionality [51] as related to the logistical challenges of collecting sufficient samples in a clinical setting: in short, the addition of numerous features/variables to

a predictive data set leads to a combinatorial explosion of the number of samples needed to identify/infer a relationship between those features and the desired goal. Since the explicit goal of a clinical forecasting/prediction algorithm is to define the risk probability for an *individual* patient/trajectory, representing and generating these potential trajectories requires deconstruction of the aggregating assumptions used for creating statistical models and in standard calibration/parametrization methods for dynamic computational models. The focus on simulating individual trajectories is inherent to the concept of digital twins, which is why we believe that the implementation of this concept in a biomedical context using mechanism-based simulation models as the underlying digital twin specification presents a possible solution to the Curse of Dimensionality, and, by extension, the challenge of dealing with temporal data drift for the implementation and maintenance of AL/ML predictive algorithms (since the wider space of possibility, and therefore generalizability, is captured by the creation of synthetic data). Inherent to this application of mechanism-based computer simulations is the calibration phase, where instead of fitting to a mean of highly variable and often overlapping data points, the simulated trajectories produce a space that encompasses the full range of the available data (see Figures 1 and 2); we present a particular workflow that accomplishes this goal [16, 17] and hope that others will develop analogous approaches. While there are clear challenges to this approach in the biomedical arena, with many overlaps with the underlying issues in translating the industrial digital twin concept to biomedicine (specifically with respect to the development of "trustworthy" mechanism-based simulation specifications [52]), this use of medical digital twins for the purpose of generating synthetic time series for the training of ML/AI forecasting algorithms represents a path forward to what are otherwise intractable limitations regarding the utility of pure ML/AI systems to forecast sepsis and multiple organ failure.

5.0 Summary

Synthetic data is proving to be invaluable in the training of modern ML/AI systems. There are particular challenges to generating synthetic time series data of molecular features of biological systems; these challenges include but are not limited to data sparsity, the Curse of Dimensionality and over-fitting. We propose that these challenges can be overcome using sufficiently complex mechanism-based models calibrated in a fashion that encompasses all available training data by generating distinct system trajectories. The further development of these models and methods can provide a means of generating synthetic time series data that can increase the robustness and generalizability of ML/AI systems applied to clinical scenarios where molecular/biomarker/gene expression time series data is of interest for the prediction/forecasting of disease progression and the discovery/testing of novel drug targets.

Funding: The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views, opinions or policies of Uniformed Services University of the Health Sciences (USUHS), The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., the Department of Defense (DoD), the Departments of the Army, Navy, or Air Force. Mention of trade names, commercial products, or organizations does not imply endorsement by the US Government. Research activities leading to the development of this study were funded by the Department of Defense's Defense Health Program, Joint Program Committee 6/Combat Casualty Care (USUHS HT9404-13- 1-0032 and HU0001-15-2-0001). This study was approved by Institutional Review Boards at the Walter Reed National Military Medical Center in compliance with all Federal regulations governing the protection of human subjects. This work was also supported in part by the National Institutes of Health Award UO1EB025825. This research is also sponsored in part by the Defense Advanced Research Projects Agency (DARPA) through Cooperative Agreement D20AC00002 awarded by the U.S. Department of the Interior (DOI), Interior Business Center. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nature medicine. 2022;28(7):1455-60.
2. Henry KE, Adams R, Parent C, Soleimani H, Sridharan A, Johnson L, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nature medicine. 2022;28(7):1447-54.
3. Nelson K, Corbin G, Anania M, Kovacs M, Tobias J, Blowers M, editors. Evaluating model drift in machine learning algorithms. 2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA); 2015: IEEE.
4. Rahmani K, Thapa R, Tsou P, Chetty SC, Barnes G, Lam C, et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. medRxiv. 2022.
5. Baier L, Jöhren F, Seebacher S, editors. Challenges in the Deployment and Operation of Machine Learning in Practice. ECIS; 2019.
6. Ackerman S, Farchi E, Raz O, Zalmanovici M, Dube P. Detection of data drift and outliers affecting machine learning model performance over time. arXiv preprint arXiv:201209258. 2020.
7. Festor P, Jia Y, Gordon AC, Faisal AA, Habli I, Komorowski M. Assuring the safety of AI-based clinical decision support systems: a case study of the AI Clinician for sepsis treatment. BMJ health & care informatics. 2022;29(1).
8. Grieves MW. Virtually intelligent product systems: digital and physical twins. 2019.
9. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine. 2019;25(1):44-56.
10. Masison J, Beezley J, Mei Y, Ribeiro HAL, Knapp AC, Sordo Vieira L, et al. A modular computational framework for medical digital twins. Proceedings of the National Academy of Sciences. 2021;118(20):e2024287118.
11. Björnsson B, Borrebaeck C, Elander N, Gasslander T, Gawel DR, Gustafsson M, et al. Digital twins to personalize medicine. Genome medicine. 2020;12(1):1-4.
12. Laubenbacher R, Sluka JP, Glazier JA. Using digital twins in viral infection. Science. 2021;371(6534):1105-6.
13. An G, Cockrell C. Drug Development Digital Twins for Drug Discovery, Testing and Repurposing: A Schema for Requirements and Development. Frontiers in systems biology. 2022;2. Epub June 2, 2022. doi: 10.3389/fsysb.2022.928387.
14. Wright L, Davidson S. How to tell the difference between a model and a digital twin. Advanced Modeling and Simulation in Engineering Sciences. 2020;7(1):1-13.
15. Nikolenko SI. Synthetic data for deep learning: Springer; 2021.
16. Cockrell C, Ozik J, Collier N, An G. Nested active learning for efficient model contextualization and parameterization: pathway to generating simulated populations using multi-scale computational models. Simulation. 2021;97(4):287-96. Epub 20201214. doi: 10.1177/0037549720975075. PubMed PMID: 34744189; PubMed Central PMCID: PMCPMC8570577.
17. Cockrell C, An G. Using Genetic Algorithms to reproduce the heterogeneity of clinical data through model refinement and rule discovery in a high-dimensional agent-based model of systemic inflammation. Frontiers in Physiology. 2021;12:726.

18. Allen R, Rieger TR, Musante CJ. Efficient generation and selection of virtual populations in quantitative systems pharmacology models. CPT: pharmacometrics & systems pharmacology. 2016;5(3):140-6.
19. Knight-Schrijver V, Chelliah V, Cucurull-Sanchez L, Le Novère N. The promises of quantitative systems pharmacology modelling for drug development. Computational and structural biotechnology journal. 2016;14:363-70.
20. An G. In silico experiments of existing and hypothetical cytokine-directed clinical trials using agent-based modeling. Critical care medicine. 2004;32(10):2050-60.
21. Cockrell RC, An G. Examining the controllability of sepsis using genetic algorithms on an agent-based model of systemic inflammation. PLoS Comput Biol. 2018;14(2):e1005876. Epub 2018/02/16. doi: 10.1371/journal.pcbi.1005876. PubMed PMID: 29447154; PubMed Central PMCID: PMCPMC5813897.
22. Brown D, Namas RA, Almahmoud K, Zaaqoq A, Sarkar J, Barclay DA, et al. Trauma in silico: Individual-specific mathematical models and virtual clinical populations. Science translational medicine. 2015;7(285):285ra61-ra61.
23. Li NY, Verdolini K, Clermont G, Mi Q, Rubinstein EN, Hebda PA, et al. A patient-specific in silico model of inflammation and healing tested in acute vocal fold injury. PloS one. 2008;3(7):e2789.
24. Petersen BK, Yang JC, Grathwohl WS, Cockrell C, Santiago C, An G, et al. Deep Reinforcement Learning and Simulation as a Path Toward Precision Medicine. J Comput Biol. 2019;26(6):597-604. doi: 10.1089/cmb.2018.0168. PubMed PMID: WOS:000477087900008.
25. Cockrell C, An G. Sepsis reconsidered: Identifying novel metrics for behavioral landscape characterization with a high-performance computing implementation of an agent-based model. J Theor Biol. 2017;430:157-68. Epub 20170718. doi: 10.1016/j.jtbi.2017.07.016. PubMed PMID: 28728997; PubMed Central PMCID: PMCPMC5635265.
26. Fonseca CM, Fleming PJ, editors. Genetic Algorithms for Multiobjective Optimization: FormulationDiscussion and Generalization. Icga; 1993: Citeseer.
27. Goldberg DE, Holland JH. Genetic algorithms and machine learning. Machine learning. 1988;3(2):95-9.
28. Haupt RL, Haupt SE. Practical genetic algorithms: John Wiley & Sons; 2004.
29. Cohn DA, Ghahramani Z, Jordan MI. Active learning with statistical models. J Artif Intell Res. 1996;4:129-45. doi: DOI 10.1613/jair.295. PubMed PMID: WOS:A1996UD02900001.
30. Brinker K. On active learning in multi-label classification. From Data and Information Analysis to Knowledge Engineering: Springer; 2006. p. 206-13.
31. Huang S-J, Jin R, Zhou Z-H, editors. Active learning by querying informative and representative examples. Advances in neural information processing systems; 2010.
32. Schein AI, Ungar LH. Active learning for logistic regression: an evaluation. Machine Learning. 2007;68(3):235-65. doi: 10.1007/s10994-007-5019-5. PubMed PMID: WOS:000248983300002.
33. Tsymbalov E, Panov M, Shapeev A, editors. Dropout-Based Active Learning for Regression. International Conference on Analysis of Images, Social Networks and Texts; 2018: Springer.
34. Karp RM. On the computational complexity of combinatorial problems. Networks. 1975;5(1):45-68.

35. Sneddon MW, Faeder JR, Emonet T. Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. Nat Methods. 2011;8(2):177-83. Epub 20101226. doi: 10.1038/nmeth.1546. PubMed PMID: 21186362.
36. Hopfield JJ, Tank DW. "Neural" computation of decisions in optimization problems. Biol Cybern. 1985;52(3):141-52. Epub 1985/01/01. doi: 10.1007/BF00339943. PubMed PMID: 4027280.
37. Edwards R, Glass L. Combinatorial explosion in model gene networks. Chaos. 2000;10(3):691-704. Epub 2003/06/05. doi: 10.1063/1.1286997. PubMed PMID: 12779419.
38. Neumann F, Witt C. Combinatorial optimization and computational complexity. Bioinspired Computation in Combinatorial Optimization: Springer; 2010. p. 9-19.
39. De Boer P-T, Kroese DP, Mannor S, Rubinstein RY. A tutorial on the cross-entropy method. Annals of operations research. 2005;134(1):19-67.
40. Fanelli V, Vlachou A, Ghannadian S, Simonetti U, Slutsky AS, Zhang H. Acute respiratory distress syndrome: new definition, current and future therapeutic options. Journal of thoracic disease. 2013;5(3):326.
41. Le S, Pellegrini E, Green-Saxena A, Summers C, Hoffman J, Calvert J, et al. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). J Crit Care. 2020;60:96-102. Epub 20200724. doi: 10.1016/j.jcrc.2020.07.019. PubMed PMID: 32777759.
42. Singhal L, Garg Y, Yang P, Tabaie A, Wong AI, Mohammed A, et al. eARDS: A multi-center validation of an interpretable machine learning algorithm of early onset Acute Respiratory Distress Syndrome (ARDS) among critically ill adults with COVID-19. PloS one. 2021;16(9):e0257056.
43. Wong A-KI, Cheung PC, Kamaleswaran R, Martin GS, Holder AL. Machine learning methods to predict acute respiratory failure and acute respiratory distress syndrome. Frontiers in big Data. 2020;3:579774.
44. Solti I, Cooke CR, Xia F, Wurfel MM, editors. Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches. 2009 IEEE international conference on bioinformatics and biomedicine workshop; 2009: IEEE.
45. GGillies R, Kinahan P, Hricak H. Radiomics: Images Are More than Pictures. They Are Data Radiology. 2016;278(2):563-77.
46. Chen Y, Wang Y, Zhang Y, Zhang N, Zhao S, Zeng H, et al. A quantitative and radiomics approach to monitoring ARDS in COVID-19 patients based on chest CT: a retrospective cohort study. International journal of medical sciences. 2020;17(12):1773.
47. Rehm GB, Woo SH, Chen XL, Kuhn BT, Cortes-Puch I, Anderson NR, et al. Leveraging IoTs and machine learning for patient diagnosis and ventilation management in the intensive care unit. IEEE Pervasive Computing. 2020;19(3):68-78.
48. Sottile PD, Albers D, Higgins C, Mckeehan J, Moss MM. The association between ventilator dyssynchrony, delivered tidal volume, and sedation using a novel automated ventilator dyssynchrony detection algorithm. Critical care medicine. 2018;46(2):e151.
49. Adams JY, Lieng MK, Kuhn BT, Rehm GB, Guo EC, Taylor SL, et al. Development and validation of a multi-algorithm analytic platform to detect off-target mechanical ventilation. Scientific reports. 2017;7(1):1-11.

50. Bhattarai S, Gupta A, Ali E, Ali M, Riad M, Adhikari P, et al. Can Big Data and Machine Learning Improve Our Understanding of Acute Respiratory Distress Syndrome? Cureus. 2021;13(2).
51. Chen L. Curse of Dimensionality. In: Liu L, ÖZsu MT, editors. Encyclopedia of Database Systems. Boston, MA: Springer US; 2009. p. 545-6.
52. An G. Specialty Grand Challenge: What it will take to cross the Valley of Death: Translational Systems Biology,"True" Precision Medicine, Medical Digital Twins, Artificial Intelligence and In Silico Clinical trials. Frontiers in Systems Biology. 2022:5. Epub April 28, 2022. doi: 10.3389/fsysb.2022.901159