INVITED PAPER

Machine Learning and Decision Support in Critical Care

This paper discusses the issues of compartmentalization, corruption, and complexity involved in collection and preprocessing of critical care data.

By Alistair E. W. Johnson, Mohammad M. Ghassemi, Shamim Nemati, Katherine E. Niehaus, David A. Clifton, and Gari D. Clifford, *Senior Member IEEE*

ABSTRACT | Clinical data management systems typically provide caregiver teams with useful information, derived from large, sometimes highly heterogeneous, data sources that are often changing dynamically. Over the last decade there has been a significant surge in interest in using these data sources, from simply reusing the standard clinical databases for event prediction or decision support, to including dynamic and patient-specific information into clinical monitoring and prediction problems. However, in most cases, commercial clinical databases have been designed to document clinical activity for reporting, liability, and billing reasons, rather than for developing new algorithms. With increasing excitement surrounding "secondary use of medical records" and "Big Data" analytics, it is important to understand the limitations of current databases and what needs to change in order to enter an era of "precision medicine." This review article covers many of the issues involved in the collection and preprocessing of critical care data. The three challenges in critical care are considered: compartmentalization, corruption, and complexity. A range of applications addressing these issues are covered, including the modernization of static acuity scoring; online patient tracking; personalized prediction and risk assessment; artifact detection; state estimation; and incorporation of multimodal data sources such as genomic and free text data.

KEYWORDS | Critical care; feature extraction; machine learning; signal processing

I. INTRODUCTION

The intensive care unit (ICU) treats acutely ill patients in need of radical, life saving treatments. ICUs have evolved from the notion that specialized units used for close monitoring and treatment of patients could improve outcomes; many predecessors of the modern ICU were established in the late 1950s to provide respiratory support during a polio epidemic [1]. ICUs frequently have a high number of staff compared to other hospital departments, and studies have shown reduced incidence of mortality, lower hospital length of stay, and fewer illness complications [2], [3], corroborating the efficacy of the intensive monitoring approach. However, real world constraints restrict the number of nurses and doctors attending to the patients in the ICU [4]. ICUs cost $81.7 billion in the US, accounting for 13.4% of hospital costs and 4.1% of national health expenditures [5]. Between 2000 and 2005, the number of hospital beds in the United States shrank by 4.2%, but the number of critical care beds increased by 6.5% with occupancy increasing by 4.5%.

The ubiquitous monitoring of ICU patients has generated a wealth of data which presents many opportunities but also great challenges. In principle, the majority of the information required to optimally diagnose, treat and discharge a patient are present in modern ICU databases. This information is present in a plethora of formats including lab results, clinical observations, imaging scans, free text notes, genome sequences, continuous waveforms and more. The acquisition, analysis, interpretation, and presentation of this data in a clinically relevant and usable format is the premier challenge of data analysis in critical care [6].

In this review, we highlight how machine learning has been used to address these challenges. In particular,

Manuscript received May 26, 2015; revised October 7, 2015; accepted November 16, 2015. Date of current version January 19, 2016.
A. E. W. Johnson and **M. M. Ghassemi** are with the Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Boston, MA 02139 USA.
S. Nemati is with the Department of Biomedical Informatics, Emory University, Atlanta, GA 30322 USA.
K. E. Niehaus and **D. Clifton** are with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK.
G. D. Clifford is with the Department of Biomedical Informatics, Emory University, Atlanta, GA 30322 USA, and also with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30322 USA (e-mail: gari@gatech.edu).
Digital Object Identifier: 10.1109/JPROC.2015.2501978

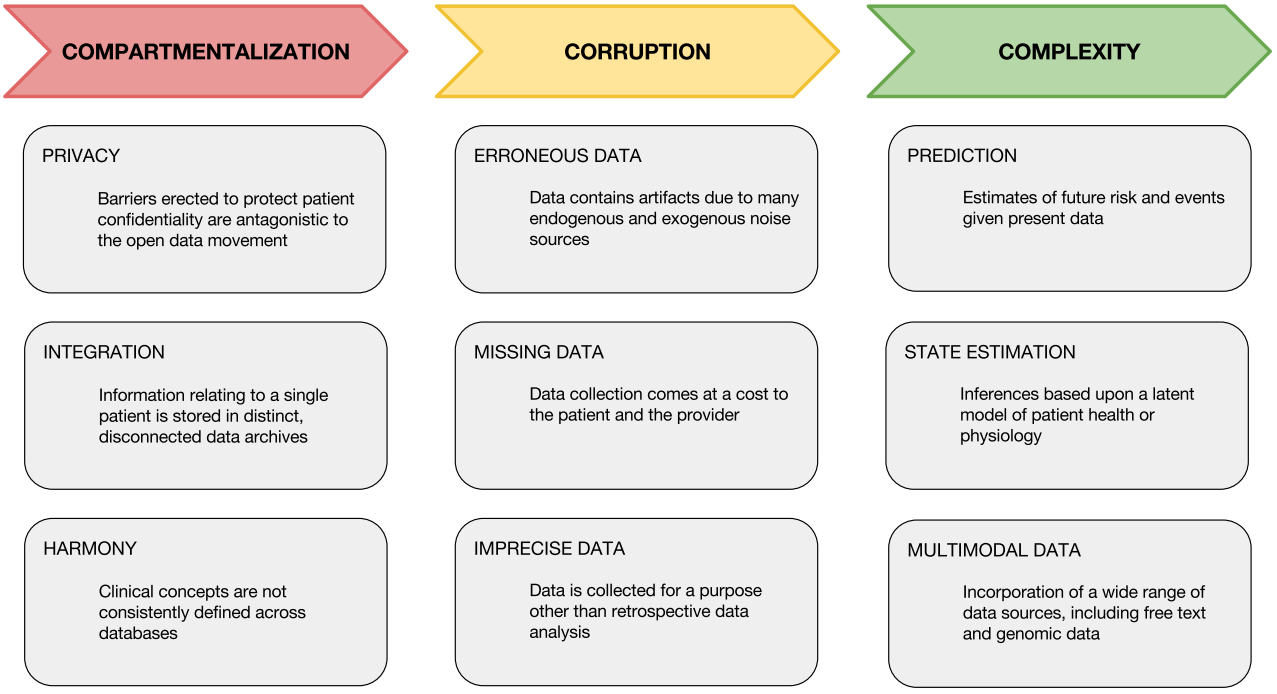

Fig. 1. ***Overview of the primary challenges in critical care. The three challenges that are presented to researchers in this field are discussed in turn: the compartmentalization of the data, which results in disparate data sets that are difficult to acquire and interrogate; the corruption of the data during collection, which necessitates nontrivial corrective work; and the complexity inherent in the systems monitored.***

we posit that data analysis in critical care faces challenges in three broad categories: compartmentalization, corruption, and complexity. Critical care data has historically been compartmentalized, with many distinct measurements of patient health being stored separately, even within the same institution. These data warehouses have been likened to silos, and the integration of data across these silos is a crucial first step before any insight can be gleaned. In the United States, integrating the Medicare and Medicaid records is necessary because Medicare does not pay for nursing home services, and only by connecting these databases can costs associated with both acute and long-term care be ascertained [7]. National critical care audits have been established in many other countries including the United Kingdom, Australia, and Canada, but these databases frequently require manual entry by a skilled worker at each individual institution, rather than the automatic synchronization which is feasible with modern technology. The second challenge is the corruption of data collected during critical care. Researchers must address a multitude of sources of data corruption including sensor drop off, artifacts related to treatment interventions, and incomplete measurements. Johnson *et al.* [8] demonstrated that removal of outliers during preprocessing of data prior to development of a mortality prediction model was as important, or even more important, than the use of nonlinear machine learning classifiers capable of capturing higher order interactions. Finally, and perhaps most self-evident, is the complexity inherent to critical care. ICUs provide technologically advanced life saving treatments that aim to both recover and maintain a healthy state in a very intricate and multifaceted system: the human body. The high level of monitoring in the ICU provides a unique opportunity for machine learning to provide new insights and has stimulated research into novel methods for this purpose.

This review provides an overview of each of these challenges and presents techniques from the field of machine learning that have been used to address them. We also discuss the future directions of research necessary to advance the field of data analytics in critical care. Fig. 1 provides a diagram outlining the paper and briefly describing the topics covered. It illustrates how this paper is organized along the lines of the three key challenges (the three data "C's") in the field: compartmentalization, corruption, and complexity.

II. CHALLENGE 1: COMPARTMENTALIZATION

There are a multitude of measurements possible to quantify the current state of a patient. These measurements

range from laboratory measurements performed on blood samples, real-time monitoring devices quantifying vital signs, billing codes for health care visits, procedure codes for services provided within health care environments, and more. For patients admitted to the ICU, the data volume is even higher as devices continuously monitor and provide information about the patient's state. However, due to a variety of factors, all data relating to a patient's health is rarely integrated into a single system. In fact, data collected at the same institution is frequently compartmentalized. The reasons for this phenomenon are primarily as follows: the private nature of the data, the technical difficulty in integrating heterogeneous sources of data into a single location, and the challenge of harmonizing of data to facilitate its analysis.

A. Privacy

Fundamental to the analysis of any data related to human subjects is respect of the private nature of the data. In 1996, the U.S. Congress passed the Health Insurance Portability and Accountability Act (HIPAA) [9] which mandated confidential handling of protected health information (PHI). The National Health Service (NHS) in the United Kingdom outlined similar regulations regarding the safe keeping of PHI [10]. These acts, and their respective counterparts in different countries, are crucial for protecting the subjects of health research. While openly available computer programs and data are highly desirable to ensure the reproducibility of science [11], the private nature of the data prohibits this approach with any PHI. Data protection is achieved by health care institutions through the use of encryption protocols, access restricted systems, and strict regulations regarding the breadth and quantity of patient data which can be archived.

Inevitably, these systems have erected barriers for research using human subjects. In a survey by Ness *et al.* [12], 67.8% of respondents said that HIPAA made research more difficult (level 4 to 5 on a Likert scale), and the proportion of institutional review board applications in which the privacy rule was detrimental was significantly higher than the number of applications where the rule was beneficial.

Enabling the use of health data can be done in two formats: restricted access and altered data [13]. Restricted access entails sharing the data with a subset of approved researchers, usually at some cost and only allowing for data storage in well secured restricted locations. The second method, altered data, involves removing some aspect of the data to allow for its more general release. This could involve removing PHI from the data set (release of data in this manner is allowed for under HIPAA safe harbor or, less frequently, the expert determination rule [14]), providing high level statistics of the data, or grouping subsets of individuals together. Selecting the optimal balance between providing useful statistical data from data and ensuring the privacy of individuals—so-called "statistical disclosure control"—has been a heavily researched area [15].

Automated de-identification of free-text medical records is often the initial barrier to the analysis. Neamatullah *et al.* developed a software package which used lexical lookup tables, regular expressions, and simple heuristics to deidentify free-text medical records from PHI including doctors' names and years of dates. The investigators reported a precision and recall of 0.749 and 0.967, respectively, with a fallout value of 0.002 on a test corpus [16].

The Integrating Biology and the Bedside (i2b2) project is a successful application of both methods: data is stored locally at each institution with PHI, and researchers can query for aggregate summaries of the data without access to individual level information [17]. i2b2 has also provided open access to various medical notes to encourage research in natural language processing to deidentify medical records, among other tasks. Building on this is the concept of differential privacy, where the probability of data output is almost equally likely to have been drawn from all nearly identical input data, which consequently guarantees that all outputs are insensitive to any individual's data [18]. Research has extended this concept into the unique setting of health care data and evaluated the utility of data after being anonymized using differential privacy; this may be a useful tool for future release of critical care data [19].

A notable success in the release of data in critical care is the PhysioBank component of PhysioNet [20], and in particular the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database [21], [22]. PhysioNet is a resource for openly available physiologic signals, many of which are collected during a patient's stay in critical care. MIMIC-II is a large openly available clinical database which provides deidentified patient records for over 30 000 patients admitted to the Beth Israel Deaconess Medical Center in Boston, MA, USA. The data is provided to researchers after certification of completion of a human subjects training course and the signing of a data use agreement. The database is a great step toward removing barriers between researchers and real-world data necessary to validate their work. MIMIC-III has recently been released, which includes more patients and additional information regarding their individual stays (e.g., additional discharge information).

B. Integration

There are over 200 000 medical devices registered by the U.S. Food and Drug Administration [23]. Yet there is a scarcity of interoperability among these devices. Monitoring patients in the ICU generates large volumes of data, but these data cannot be thought of as comprising one entity. Devices to measure various aspects of patient

health have been developed independently and organically. One of the first treatments provided by ICUs was respiratory support [1], and ventilators, which initially only provided positive pressure through gas or pneumatic driven processes, can now electronically control volume and pressure while recording many other parameters. The ECG is one of the most frequently used measurement devices, but the data available can vary greatly: almost all devices calculate and record heart rate, but others automatically determine rhythm, ST elevation, or QT interval. Oxygen saturation devices began to be routinely used in the ICUs in the 1980s, most providing a measure of blood oxygen saturation, but some also providing heart rate. With just these few examples, it becomes clear that the integration of information from various devices into a single data management system is nontrivial, requiring well-defined standards for transferred packets of data, interoperability of devices, and cooperation among competitive device manufacturers. Unfortunately, there has been a lack of standardization among clinical devices [24]. The consequence of the lack of standardization and interoperability is a heterogeneous landscape of databases and record systems which can only be integrated with a great deal of labor.

The United States has recently passed the Health Information Technology for Economic and Clinical Health (HITETCH) act, enforcing interoperability among various systems and partly addressing this issue. The consequences of this have been immediately apparent in the uptake of electronic health records (EHRs): in 2008, the number of U.S. hospitals with EHRs was 9.4%, while in 2014, it had grown to 75.5% [25]. Furthermore, over 95% of these EHRs were certified, indicating that they possessed a required minimum level of interoperability. Black and Payne [26] proposed a system for defining the quality of a database, though their concepts of coverage and accuracy do not sufficiently summarize the utility of a database, due to an equal weighting of the various components [27]. Cooke and Iwashyna [27] provide an excellent approach for selecting an existing database to address a proposed research question. The authors highlight the advantage of integrating, or linking, two data sets, providing an example where Iwashyna *et al.* [28] study quality of life among severe sepsis survivors by using an already-established link between the Health Retirement Study and Medicare files for patients admitted to ICUs. Finney *et al.* developed a data linkage scheme that allowed their hospital trust to link data from distinct databases using various identifiers with 99.8% positive predictivity [29].

Cooke and Iwashyna [27] conclude with a poignant statement—that the major barrier for optimal care for all critically ill patients is a lack of an integrated openly available data warehouse—even though this is a feasible goal. The MIMIC database has demonstrated that integration of data from disparate sources of the hospital is possible even when it requires integration of distinct databases for provider order entries, laboratory measurements, echocardiogram notes, discharge summaries, clinical observations, and mortality outcomes [21]. Furthermore, the large multicenter eICU database, collected from units which take advantage of Philips Healthcare's telemetry services, has successfully integrated data from hundreds of hospitals across the continental United States [30].

C. Harmony

The integration of databases, while in itself a monumental and difficult task, provides no guarantees of a usable data set. The reason for this is the lack of data *harmony*, where a concept in one database is not linked with a concept in the other database, or the definition of concepts in one database is not congruent with the linked concept in another. An ontology is a systematic categorization of concepts, and matching ontologies is one of the largest challenges to overcome when integrating two databases. The APACHE IV mortality prediction system utilizes 114 admission diagnostic categories, and the difficulty in mapping a given ICU's diagnosis ontology to these categories has been listed as one of the major barriers to its clinical acceptance [31], [32]. Many coding schemes have been devised that aim to standardize ontologies across databases to facilitate harmonizing of their respective contents. The International Classification of Diseases (ICD) aimed to standardized all possible disease categories for patients [33], though variation in coding practice has been highlighted as a potential source of error [34]. As these codes are frequently retrospectively assigned by trained human coders reading patient notes, there is a great opportunity for natural language processing techniques to automate and improve the current work flow. The 2007 Computational Medicine Challenge provided a corpus of de-identified radiology reports and gave participants the task of assigning two codes from a set of 45 ICD-9 codes [35]. The highest performing participants used medically informed features in combination with machine learning classifiers such as C4.5. SNOMED-CT is another coding system [36] which has been shown to cover 93% of clinical concepts in a problem list [37]. Another coding system is LOINC [38], which was originally purposed for laboratory measurements but has since been extended to other clinical concepts. In fact, the growing number of distinct ontologies, many of which overlapping in purpose, has led researchers to create a database of ontologies [39]. As mentioned, the concept of interoperability has become a major area of interest due to recent U.S. legislation changes which penalize hospitals without EHRs and stipulate requirements for their communication [25]. Yet harmony among these EHRs has yet to be achieved [40]. While other disciplines have benefited from the use of

machine learning on large data sets, the lack of harmony among EHRs in critical care has stymied applications.

III. CHALLENGE 2: CORRUPTION

Once data has been merged, linked, and stored in a single unified location, it is necessary to evaluate the data using some measure of quality. While preprocessing the data is a common step in many machine learning applications, it becomes critical in the medical environment because the data is collected with the intention of enhancing patient care, not to facilitate analysis. A prominent example of this phenomenon is the use of free-text comments to highlight spurious readings: a high potassium measurement can be explained by a comment stating that the sample has been hemolyzed and is not an accurate reflection of the patient's health, and while this comment is trivial for a care giver to parse, it complicates retrospective analysis. Discerning true measurements from noisy observations, the hallmark of processing so-called "dirty" data, is nontrivial and many pioneers in the field have created elegant solutions to these problems. Data corruption in this review has been classified into three variants: erroneous data, occurring when a value is not an accurate reflection of the true measurement; missing data, occurring when data is unavailable for a parameter of interest; and imprecise data, occurring when surrogate labels are provided instead of the desired concept label. Note that we have made a distinction between *erroneous* data, which have been modified by an aberrant phenomenon to no longer reflect the truth, and *imprecise* data, in which the data collected is accurate but does not explicitly capture the concept of interest (e.g., an ICD-9 code relating to diabetes is not identical to a diagnosis of diabetes).

A. Erroneous Data

As the removal of untrustworthy data is an important step in the training and testing of any predictive model, there is a justifiable need for algorithms that can identify artifactual data or utilize an inherent confidence measure to inform the user of questionable data. In [41], Nouira *et al.* note that many methods have been proposed for the task of outlier rejection in time-series analysis in the intensive care unit, including autoregression, integration, moving average (ARIMA) models [42], Bayesian forecasting [43], and a variety of robust signal estimators [44]. Three broad categories in which there can be erroneous data are explored here: waveforms, observations, and data fusion. These categories have been chosen as the type of data determines the types of artifacts possible, and consequently the various methods used to rectify the data. Waveform data continuously recorded from sensors is susceptible to high-frequency artifacts associated with patient movement or clinical care. Periodic clinical measurements can be contaminated by data collection and coding practices (e.g., monitors recording missing heart rates as 0). The last category is less data specific than the previous categories, and highlights methods that take advantage of the redundant information streams in the ICU to extract data that is robust against artifacts. As these methods can be equally applied to either waveforms or observations, they have been discussed independently.

An example of data corruption, which resulted in a false alarm in the ICU, is given in Fig. 2.

1) Waveforms: A comprehensive review of artifact detection techniques in critical care is given by Nizami *et al.* [45]. The review highlights the complexity of artifact detection and removal: algorithms must be shown to generalize across units, manufacturers and varying patient demographics. Most algorithms utilize a signal quality index (SQI) which assesses how physiologically reasonable a signal is, excluding the data if it appeared invalid. Overall, the authors conclude that most existing algorithms were developed in an *ad hoc* manner, lacked proper validation, were rarely evaluated in real time, and usually not implemented in clinical practice. The authors also noted that the proprietary nature of many monitors creates an unknown element when analyzing derived signals from these monitors (e.g., unknown filters are used to process the signal prior to acquisition). This ambiguity complicates reproducibility in research and prevents algorithms developed on data acquired from one manufacturer being extended to another. Nizami *et al.* [45] also noted that a paucity of the commercially implemented signal quality indices were evaluated in the literature.

Signal quality is frequently an important quantity for real-time alerting systems currently utilized in clinical practice. In a real-time alerting system, the aim is to detect a sudden change in the patient state (e.g., transition from normal sinus rhythm to life threatening arrhythmia) and subsequently alert the clinical staff to this event. As discussed by Nouira *et al.* [41], these change points are often life threatening, and ICU alarm systems were developed to alert the clinical staff with a minimal delay so as to not compromise patient care. Unfortunately, many sources of noise in the ICU are transient and imitate these change points. This problem is further exacerbated by the simplicity of rules behind most ICU alarm systems, often utilizing simple magnitude thresholds to indicate a change of state [46], [47].

In order to evaluate the level of noise or conversely the signal quality, Li and Clifford proposed a series of techniques for pulsatile signals based on a fusion of different "simple" features [48], [49]. These features can be classified into three general categories, given their nature. The first category is based on the agreement of two independent beat detectors with different noise sensitivities. Both detectors are run simultaneously on the ECG

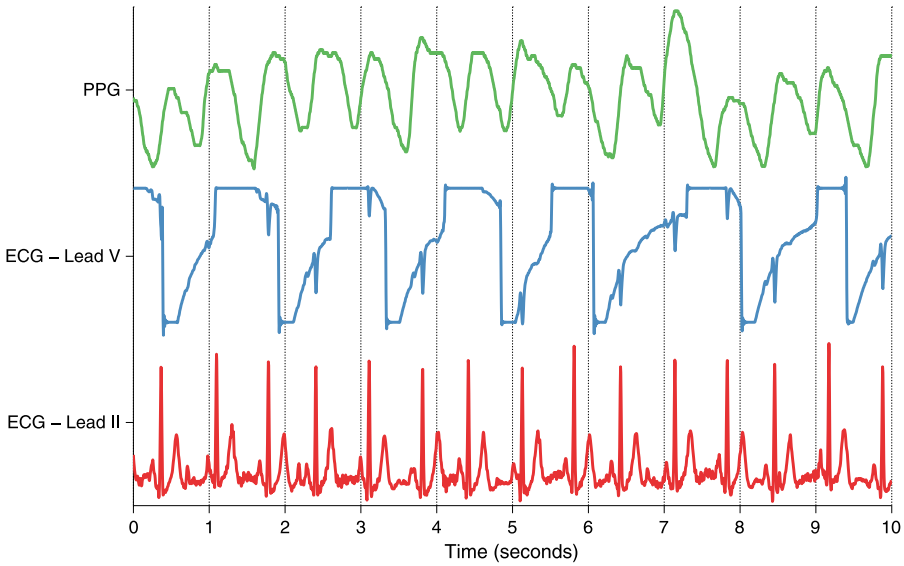

Fig. 2. ***Example of a false alarm which incorrectly asserted the patient was in asystole. The signals shown are the photoplethysmogram (PPG, top in green), the electrocardiogram lead V (ECG, middle in blue), and the electrocardiogram lead II (ECG, bottom in red). The alarm likely triggered univariately on ECG lead V. At least two methods reviewed in this section could have prevented this false alarm: the use of signal quality on lead V or a multimodal data fusion approach which incorporated ECG lead II, the PPG, or both.***

signals, the first one being based on the detection of the ECG peak's energy [50], [51] and the second being based on the length transform [52]. Since the length transform is much more sensitive to noise than the energy detector, the level of agreement between the two detectors tends to be proportional to the level of signal quality. Other SQIs were also proposed, including features based on the power spectral density, statistical moments, and "flat line" detectors. In general, it appears that the extraction of SQIs, followed by their fusion in a machine learning framework, has had success in the literature. Behar *et al.* [53] utilized a support vector machine (SVM) [54] to directly estimate signal quality of ECG leads (achieving 95% accuracy across a variety of heart rhythms), while Li *et al.* [55] suppressed false arrhythmia alarms using SQIs and a relevance vector machine (RVM) [56] and achieved false alarm suppression rates between 17.0% for extreme bradycardia and 93.3% for asystole. Both Li *et al.* [55] and Behar *et al.* [53] highlighted the impact of rhythm type on signal quality, noting that SQIs must be tailored to a variety of arrhythmias and calling for more labeled training data to facilitate this task. More recently, Morgado *et al.* [57] estimated the cross correlation across a 12-lead ECG in combination with machine learning classifiers CART [58], C4.5 [59], RIPPER [60], and a SVM [54] to achieve an accuracy of up to 92.7% and an AUROC of up to 0.925 for the task of signal quality estimation. This method is similar to the Riemannian "potato" [61], which also uses the covariance matrix of a set of simultaneous leads to estimate signal quality. The averaging of data across time periods has also been shown to improve robustness to noise. Tsien *et al.* [62] employed decision tree induction classifiers to classify a variety of artifacts from carbon dioxide, blood pressure, heart rate, and oxygen saturation trends, showing that models developed from one minute aggregations of second by second data were more accurate than those built on second by second data.

Low signal quality has a large impact on alarm systems currently in place in ICUs. Most manufacturers are conservative with alarm thresholds and tune algorithms to be extremely sensitive, resulting in a false alarm rate of up to 95% [63]. This in turn has resulted in "alarm fatigue," which creates an unsafe patient environment due to desensitization of caregivers—life threatening events can potentially be missed [64], [65]. Zong *et al.* [66] proposed a fuzzy logic approach to accept or reject alarms on the arterial blood pressure waveform. The algorithm maintains a running average of various physiologic measurements derived from the waveform and suppresses an alarm if one of these components is not physiologically plausible (e.g., a systolic blood pressure above 300). Additional measures of signal quality were based on comparison of the current measurements to a running average.

The recent PhysioNet/Computing in Cardiology Challenge 2015 provided a public database of 750 training and 500 test alarms to stimulate research into the area of false alarm reduction [67]. Participants in the Challenge were given samples of ICU patient waveforms that were identified by the bedside monitor as falling into one of five rhythms: asystole, extreme bradycardia,

extreme tachycardia, ventricular tachycardia and ventricular fibrillation, or flutter. All submitted methods involved a form of signal quality estimation: Plesinger *et al.* [68] used physiologic thresholds on extracted features including heart rate and blood pressure, Antink *et al.* [69] used autocorrelation and a linear discriminant analysis classifier, and Fallet *et al.* [70] used mathematical morphology to provide additional robustness to noise in the underlying signal. Winning competitors were able to suppress 88% of the false alarms with a corresponding 8% true alarm suppression rate. This true alarm suppression rate dropped to 1% (with a suppression of 80% of the false alarms) when the algorithm was given an extra 30 seconds for rhythm classification. For a more detailed review of the specific issues around time-series data collection and signal processing, we refer the reader to previous work in the literature [71].

2) Observations: The framework for quality assessment and artifact removal is much more established for high-resolution physiologic waveforms as compared to lower resolution clinical measurements contained in an electronic data management system (referred to here as "observations"). For such less granular information, a commonly employed technique for handling artifacts is the use of domain knowledge to remove (or disallow on input) physiologically implausible values [31], [72]. Certain measurements intrinsically lend themselves to this approach: oxygen saturation values cannot go above 100%, biochemical concentrations have known reference ranges, vital signs have implausible ranges, etc. However, the domain knowledge approach of outlier rejection has limitations. Certain variables, especially those that have logarithmic distributions, with orders of magnitude between plausible values, are not easily processed using domain knowledge. Furthermore, due to the primary use of the data for clinical care, and not retrospective modeling, these errors are often not easily corrected at the source of the data collection. Other statistical rules of thumb are commonly employed in place of domain knowledge (e.g., the removal of extreme percentiles, sometimes referred to as "Winsorization") [73], [74].

Fialho *et al.* [75] classified outliers as data that were further than 1.5 times the interquartile range away from either the 25th or 75th percentile (for normally distributed data, this is approximately 2.7 standard deviations and 99.3% of the distribution resides within these limits). The authors replaced these outliers using the previous value in time, frequently referred to as sample and hold, and predicted fluid response using disease specific models. They were able to achieve AUROCs 0.04 higher than general purpose models. Johnson *et al.* demonstrated that a regularized logistic regression with no preprocessing (AUROC of 0.832) was inferior to a RF (AUROC of 0.841), but use of either domain knowledge based thresholds or an automatic method for outlier rejection resulted in the logistic regression model outperforming the RF (AUROC of 0.848 versus 0.843). They also demonstrate equivalent performance between rejection methods using automatic outliers and those relying upon domain knowledge. In their discussion of the challenge of applying knowledge-based methods, they highlight the problems of cross-institution differences in unit of measurement, labor intensity, and the lack of known thresholds for heavy tailed distributions (as noted earlier). An example of the difficulty in the identification of outliers is given in Fig. 3, where the respiratory rates are implausible but may represent true respiratory distress.

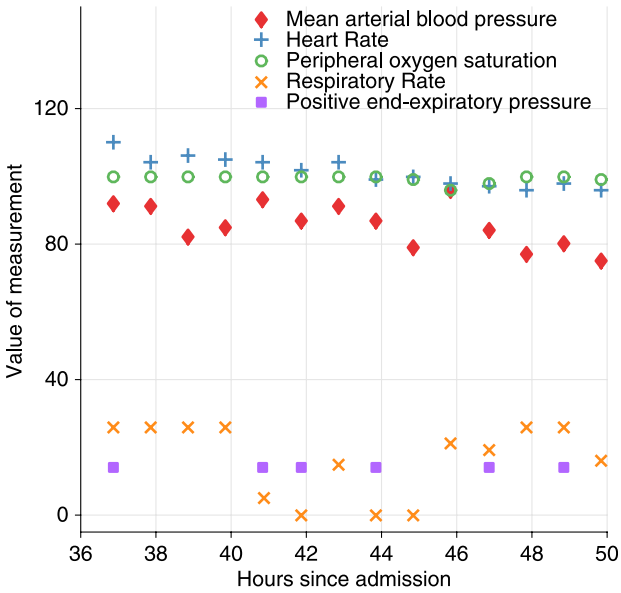

Fig. 3. ***Example of low, sometimes zero respiratory rates. As a sustained breathing rate of zero for hours is incompatible with life, the data here may represent: 1) undersampling of true respiratory distress with intermittent apnea; 2) erroneous data corresponding to sensor fault; or 3) manually entered data intended to represent poor physiologic state.***

Aleks *et al.* [76] considered the problem of modeling arterial-line blood pressure sensors, which are subject to frequent data artifacts and frequently cause false alarms in the ICU. They utilized a dynamic Bayesian network to model the sensor artifacts in a generative manner and reported an artifact classification performance on par with the experienced physician's. As pointed out by the authors, the problem of artifact detection is complicated by the fact that (depending on the bedside monitor brand and data archiving and streaming protocols) the sensor data are often averaged over fixed intervals, whereas the events causing data artifacts may occur at any time and often have durations significantly shorter than the data collection interval. Factorial switching linear dynamical systems (FSLDS) have been used to switch between latent modes representing stable physiology, known artifact types, and unknown noise types [77]. In particular,

the authors' use of the "X-factor," a single latent mode that captures both unknown artifact and novel physiology, gave the model additional flexibility to classify uncertain signals as abnormal, rather than forcing a decision between classifications.

Recent extensions to the FSLDS model [78] utilize a supervised framework to create a discriminative model (as opposed to a generative model) to first classify the sensor data as belonging to one of several clinical/sensor factors (e.g., blood sampling via arterial line, suction, sensor detachment, etc.) followed by inferring the underlying physiological state of the patient conditioned on each factor. This approach allows for incorporation of a richer set of features for patient state estimation and was shown to perform better for certain classes of artifact. However, the learning algorithm relies on availability of labeled data to provide a training data set for learning various artifacts and clinical states.

Finally, we note that incorrect values are often physiologically plausible, particularly as the source monitors are designed to provide data within such ranges in the first place. Brutal filters such as sample and hold are often employed by the manufacturers (because persistence is a good estimate of physiology in the short term, and many monitors have been designed to present the best estimate "right now"). However, when using parameters derived from bedside monitors, or "clinically validated parameters," there is a danger that significant bias and variance is introduced into the estimate, and that clinically relevant events can be missed for long periods of time. Hug *et al.* [79] demonstrated that by rederiving blood pressures from the raw arterial blood pressure waveform, and using stringently validated signal quality indices to remove erroneous data, it is possible to see that clinical teams miss significant episodes of transient hypotension (leading to subsequent sepsis, which in turn is connected to higher mortality rates) for an average of four hours. This is an example of how, by rolling back to the original waveform data, significant extra clinical information can be extracted.

Of course, this leads to the enormous issue of labeling data (for developing quality indices and predictive algorithms). In practice, labeling of clinical data is often expensive, labor intensive, and consensus is difficult to obtain due to variations in clinical practice, interobserver variability, human biases, and incomplete capturing of clinical context in the EHR. However, recent advances in clinical data crowdsourcing may mitigate the problem of obtaining labeling consensus [80], [81].

As we have noted, some progress has been made in developing signal quality indices, but the vast majority of signals in the ICU lack any confidence levels. In many cases, the manufacturers of ICU medical equipment themselves generate such confidence or quality indices, but these are rarely shared (and if provided, the information is usually only displayed in the form of a traffic light system on a monitor). There is a need to open up such algorithms and require manufacturers to routinely report the confidence levels in their parameter estimates.

3) Data Fusion: The high level of monitoring in the ICU provides ample opportunity for methods that can fuse estimates of a given physiologic parameter from multiple sources to provide a single measurement, with high confidence in its veracity. One commonly encountered example is the estimation of heart rate, which is essential in many applications, such as the identification of extreme bradycardia or tachycardia. Such conditions frequently require immediate intervention. Since the ECG generally comprises a series of large amplitude spikes corresponding to each beat, heart rate can be estimated by event or "beat" detection algorithms [82]. Although beat detection has been well explored over the last four decades, good beat detection algorithms can still be easily confused by the high level of noise encountered in challenging recording environments. In order to increase the robustness of the heart rate extraction, fusing the estimations from different ECG channels can be highly beneficial.

Several methods have been proposed in order to improve the estimation of other physiological parameters from noisy measurements. Among the different approaches, the most obvious solutions consist in, again, aggregating the estimated values on each channel (for those parameters estimated from physiological signals collected through multiple measurement channels). For example, Jakob *et al.* [83] demonstrated that a median filter was useful for removing a large proportion (41%–98%) of artifacts from blood pressure signals in post-operative cardiac patients. Yang *et al.* [84] described a technique based on an hybrid median approach where the median of a single channel is combined with median values from other channels. The resulting estimate will be accurate when no more than half the channels are corrupted, or when artifacts span less than half the width of the median window. Techniques based on signal quality assessment, a topic which has been extensively covered in the previous section, have also been successfully applied to fuse estimates of physiologic parameters from multiple signals [85]–[88].

While the median is a robust method of fusing multiple sources of data, a variety of tractable approaches to data fusion have also been applied. The Kalman filter (KF), a state space approach, is naturally suited for the processing of time series that frequently have artifacts [89]. KFs treat measurements, such as heart rate, as noisy observations of an underlying state (e.g., "true" heart rate), and update the state only if the confidence in the current observation is high, conditioned on the previous observation. New observations with high "innovation" are more likely to be artifacts, and these are consequently down weighted in the calculation of the state.

KFs can be seen as a natural evolution of the hybrid median approach within a well defined paradigm. KFs offer the advantage of incorporating knowledge about the dynamics of the underlying signal, even in situations of great uncertainty in the observations. KF methods can identify trends and abrupt changes in the underlying (or latent) state without a large computational cost [90]–[92]. An approach initially proposed by Tarassenko and Townsend [93] used the KF innovation to weight heart rate derived from multiple channels. Li and Clifford [48] extended this method to include signal quality in the state updates and fusion step, thereby ensuring that low quality data and artifacts are deweighted in the estimate of the physiological parameters.

Bayesian fusion has also recently been proposed to fuse estimates of heart rate [94], [95]. These methods treat each sensor as an independent measurement of heart rate and apply Bayes' rule to estimate the current state given the current and previous observations. Oster *et al.* [96] applied a switching KF for beat classification, allowing automatic selection of beat type from multiple "modes," which were simultaneously evaluated. Furthermore, in a similar manner to the approach presented above [77], the method contains an extra mode unrelated to beat type, the "X-factor," which facilitates classifying unrecognized signals as unknown. The use of an unknown class is a form of uncertainty: if the algorithm cannot be sure of a heart beat type, it is not forced to choose and can instead default to an uncertain classification. Incorporating uncertainty in medical practice has been highlighted as one of the most important components of quality improvement [97], and this should be acknowledged in models intended for use in clinical practice.

B. Missing Data

Missing data is common and difficult aspect of data collection and analysis and has been heavily researched to date [98]. Yet, clinical care infrequently acknowledges the challenges associated with the phenomenon. Vesin *et al.* [99] found that out of 44 published clinical studies, 16 did not make any mention of missing data. Worse still, only two out of 44 studies (less than 5%) acknowledged the importance of missing data and explicitly described the methods they addressed it with. There are three types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data is MCAR when the mechanism causing its absence is completely random, for example, if a laboratory machine breaks down and is unable to supply measurements for a patient. In this case, imputation of values will result in unbiased estimates. Data is MAR if the missingness mechanism is unrelated to the value of the variable. An example of data MAR would be subsequent troponin values: while an initial value may be useful in diagnosis of MI subsequent values may not be of interest and consequently would be MAR. Finally, the most difficult mechanism occurs when data is MNAR and whether the data is missing or not depends on the value of the measurement. This may be the most common mechanism of missing data as many measurements are not performed if the clinician suspects them to be normal and provide no prognostic benefit. It is worth emphasizing however that these concepts are best considered as assumptions made during an analysis, rather than properties of the data, and an analysis is not invalidated solely for making an assumption regarding the mechanism behind the missingness which may not entirely reflect reality [100].

Many methods either remove missing cases with too many missing values or impute plausible values in their place. Shah *et al.* [101] used an iterative approach incorporating singular value decomposition to impute missing data under the assumption that data were MAR. Waljee *et al.* [102] compared missing value imputation methods and demonstrated that a RF based missing value imputation method performs best in their simulation study using data which was MAR. Kim *et al.* [103] use principal component analysis in combination with EM to estimate the value of missing data from physiologic time series.

Mean imputation remains one of the most common methods of missing data handling [104], and does not appear to degrade performance of various prediction systems in critical care greatly even though it assumes data is MAR [31], [72], [105]. Nevertheless, missing value imputation tends to bias the uncertainty in subsequent model estimates downward [106]. In the 1970s, Dempster *et al.* [107] published an algorithm for performing expectation–maximization (EM) with missing data, and this represented a fundamental shift of thought among statisticians from removing missing data as a nuisance toward averaging over the uncertainty caused by missing data [106]. This paradigm shift has slowly begun to occur in critical care, though most studies have yet to acknowledge the impact of missing data [99]. Multiple imputation, a technique which involves repeatedly imputing plausible values for missing data and averaging over many instances of imputation [108], [109], has received wide praise among the medical literature but has yet to gain traction in the critical care literature [99], though this is changing [110]. Gaussian processes (GPs) have been proposed as well as a principled method for handling missing data [111]. An example of a GP inferring data is given in Fig. 4.

Lasko [112] used a nonstationary GP regression approach to explicitly estimate the time-varying volatility of latent functions to describe four laboratory values: uric acid (UA), thyroid stimulating hormone (TSH), creatinine (Cr), and LDL cholesterol (LDL). Lasko estimated that these clinical laboratory tests were undersampled on average by 190% (as judged by the variables' information rate) but oversampled only by 27%. While GPs are a

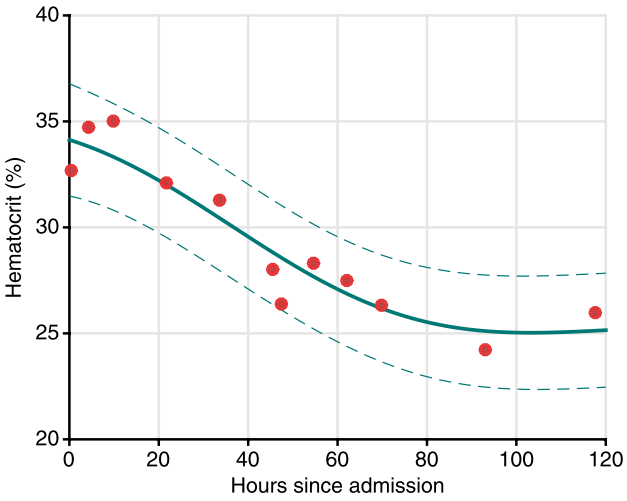

Fig. 4. *Example of a GP regression inferring the value of missing data on an unevenly sampled time series of hematocrit values. The raw values are plotted as red circles against the mean of the GP (solid green line) and the 95% confidence intervals (dashed green lines).*

theoretically appealing method due to their ability to handle missing data, their use has yet to become widespread.

C. Imprecise Data

Supervised learning is a large area of machine learning that involves learning a mapping between data and an output label; learning this mapping requires a collected set of training data with known labels. Unfortunately, as labels collected in critical care databases are usually recorded for purposes other than retrospective data analysis, it can be difficult to define a true "ground truth." Frequently only surrogate annotations are available, which capture only some component of the label of interest. A further complication is the fuzzy nature of most classification tasks of interest. For example, the definition of sepsis has evolved over time, and patients who were once classified using a dichotomous diagnosis are now thought to reside within a spectrum of the disorder [113]. Even mortality, a relatively robust outcome used in many prediction tasks, is primarily used as a surrogate to quantify patient severity of illness. ICD-9 codes are frequently used to define patient diagnosis, but the use of ICD-9 codes for billing purposes has detrimentally affected the accuracy of the codes: since they are used to maximize costs, they do not necessarily best reflect patient etiology [7]. The use of ICD-9 codes as labels in supervised learning is further complicated by the fact that the codes are susceptible to coding practice changes, and patients with the same disease profile may be assigned different codes [114].

An approach used by Halpern *et al.* to derive labels from the noisy codes available in the EHR is through the use of "anchors" in place of accurate labels [115]. The authors define a feature, such as the appearance of an ICD-9 code in discharge documentation, as an anchor if and only if it is always positive when the label of interest is positive. For example, the use of insulin therapy would be an effective anchor for diabetes. A set of anchors is used to create a data set of only positive cases, and a classifier learned using this subset of data can be generalized to apply to all positive and negative cases [116]. Through the use of a "human-in-the-loop" framework, Halpern *et al.* demonstrate that a subset of anchors can be defined which facilitate large scale unsupervised classification (since humans are required to label a subset of the data, this process is frequently referred to as semi-supervised learning).

Another common source of ground truth annotations against which an algorithm or treatment is evaluated is through manual labels provided by clinical experts. However, significant intraobserver and interobserver variability and various human biases limit accuracy [117]. Even in the case of a well-described and explored field such as electrocardiography, inter-rater disagreements in ECG diagnoses and labels can be as high as 20%–40% [118]. This may be due to intrinsic difficulties in interpreting the signals that are linked to the level of training or experience of the annotators [119]. Disagreements may be exacerbated by significant noise contamination due to motion artifacts, electrode contact noise, and baseline drift [120]. Moreover, the temporal window to which a label applies is often arbitrary and undefined, resulting in labels being applied to transient segments of data which fall either partially into two or more classes, or perhaps none.

Historically, inter- and intra-rater disagreements have often been ignored, and the errors associated in noisy labels have not been associated with performance measurements of classifiers. Even in cases where consensus or voting procedures have been applied, there is a risk of significant bias in the labeling. However, there have been several principled approaches which have attempted to address the issue of bias and variance in weighted voting strategies. Dawid and Skene [121] first proposed a model to probabilistically combine multiple expert annotations in an application to identify patients fit for general anaesthesia. In brief, the model learns a precision for each annotator which represents the accuracy of their annotations compared to the consensus. The estimated ground truth is calculated as a weighted sum of each annotators' label, using their precision as the weight. One of the major strengths of the approach is the ability of the EM algorithm to handle missing annotations [107]. Raykar *et al.* [122], [123] extended the algorithm to jointly model the ground truth and a regression model. Zhu *et al.* [124] demonstrated that the

inclusion of contextual features, such as heart rate and signal quality, ensured that the estimated ground truth in a QT interval labeling task was always as accurate as the best human annotator without any knowledge of which annotator performed best. Welinder and Perona [125] proposed a similar model in a Bayesian framework, again estimating the precision (or inverse variance) associated with each annotator's labels. Annotator bias was incorporated into the same model for binary classification tasks by Welinder *et al.* [126]. Zhu *et al.* [127] outlined a fully Bayesian description of the model, which is capable of estimating both the precision of an annotator and their bias for continuous labels. Crowdsourcing of medical labels may be an important component in future machine learning research as it facilitates creation of large annotated databases and provides better estimates of ground truth for studies employing two or more domain experts for labeling.

IV. CHALLENGE 3: COMPLEXITY

Having addressed the issues around data collection and validation, the final challenge is at the core of this review: machine learning of complex data. Machine learning is simultaneously the most exciting task and the most challenging issue in critical care data analytics. The high volume of data, which frequently overwhelms care providers [128], provides ample opportunity for computerized algorithms. The research covered in this article has been grouped as follows: models that aim to predict an outcome (prediction), inferences about a latent state using measurements (state estimation), and models that analyze multiple types of data regarding a patient, including physiology or free text notes (multimodal data).

A. Prediction

1) Mortality Prediction: One of the first applications of (supervised) machine learning in critical care, and indeed one of the most readily obvious applications in a unit with such severely ill patients, is the prediction of patient mortality. Prediction of patient outcomes, either time based (30 day mortality) or event based (in-hospital mortality), has been highlighted as a key component in the efficient and optimal delivery of ICU care [129]. The first model aimed at predicting severity of illness of a general ICU population was the Acute Physiology, Age, and Chronic Health Evaluation (APACHE) system [130]. The APACHE system was originally created by a panel of experts who collectively assigned higher scores for increasing physiologic abnormality. Over time, data driven analysis was incorporated into the creation of the APACHE systems to provide better models with higher performance. APACHE II simplified APACHE I by using correlation between each feature and outcome to reduce the number of features from 34 to 12 [131]. APACHE III was the first generation to utilize multivariate logistic regression to estimate the weights for each component of the model [132]. Finally, APACHE IV, the latest generation, used step-wise feature selection techniques to select a subset of covariates in the model. The steady progression of the APACHE system towards increasing reliance on data for each subsequent generation has been echoed by other mortality prediction systems, including the Simplified Acute Severity Score (SAPS) [72], [105], [133], [134] and the Mortality Prediction Model (MPM) models [135]–[137]. Recent work has shown that the combination of feature selection techniques (in this case, a genetic algorithm) with non-convex optimization can result in a parsimonious feature set, which provides equivalent performance to previous higher dimensional severity scores [138].

While none of the aforementioned models attained the calibration necessary to be utilized on a patient to patient basis, they have paved the way for more sophisticated machine learning methods to predict mortality and other outcomes of interest. Dybowski *et al.* [139] developed an artificial neural network (ANN) model optimized using a genetic algorithm for the purposes of mortality prediction. They demonstrated that neural networks had the flexibility to model complex patient physiology, and that this non-linear technique improved upon a logistic regression (LR) model with only linear terms. While in retrospect the study had limited power (due to the low training set size of 168 patients and large number of parameters in the neural network), it nevertheless demonstrated that the advances in machine learning could be translated into clinical practice. Clermont *et al.* later directly compared LR and ANN models [140]. When isolating the ANN's ability to model variable interactions, they showed no difference in discrimination between the LR and ANN models (AUROC of 0.848 for both). However, when allowing the ANN to directly model the relationship between the variable and the outcome, the ANN's AUROC increased to 0.857. They further demonstrated that the capability of the ANN to predict patient mortality was greatly reduced for sample sizes below 800 patients. Wong and Young similarly found a gain in discrimination from ANN models as compared to LR models (0.84 versus 0.83) [141].

The PhysioNet/Computing in Cardiology 2012 Challenge [142] aimed to stimulate research in patient specific mortality prediction systems. The primary evaluation metric, the minimum of the sensitivity (Se) and positive predictivity (PPV), was chosen to encourage algorithms to optimally classify patients who eventually died in the hospital (true positives). The best performing method, a tree based classifier with surrogate importance learned for missing data, achieved a score of 53.53%, indicating that it correctly classified half of the patients who eventually died [143]. Similar performance was achieved by set of SVMs, which were

combined in a final regression step, acting as a bias correction and recalibration stage (minimum Se/PPV of 53.52%) [144]. This was a vast improvement over the (recalibrated) severity score SAPS I [133], which only achieved a score of 31.25% [142]. In a study using the openly available MIMIC-II database [20], Pirracchio *et al.* developed 12 models and an aggregate model which fused the outputs of the prior 12 (the so-called "super learner") [145]. Again, gains in performance were similar to before, with the AUROC of a regression model (0.84) increasing with the use of a more flexible model such as a random forest (0.88).

Clearly the use of regression models for prediction has been a boon for critical care, but more complicated models seem to provide little benefit in this area. One possible explanation is the exclusive use of aggregate features over large temporal windows, such as the lowest value over 24 h. Indeed, the incorporation of features derived from patient time series is a promising and challenging task. The concept of entropy, or the amount of disorder in the signal, can be calculated in a multitude of ways; the optimal quantification of this concept as a feature in predictive models continues to be an open area of research [146].

Saria *et al.* provide an example of how features derived from shorter-range time frames can be used in ICU prediction, in this case for preterm infants [147]. The authors used vital signs (HR, respiratory rate, and oxygen saturation) from 138 preterm infants to create a predictive risk score for severe comorbidities. They first preprocessed the time-series data to obtain the mean and variance of both long-term and short-term trends. The resulting summary features were then modeled using long-tailed distributions, and patient log-odds ratios used to train a LR classifier to distinguish between low- and high-morbidity infants. The resulting scoring system attained an AUROC of 0.92 for predicting high morbidity, in comparison to alternative available risk scores, which had AUROCs in the range of 0.70–0.85.

Imhoff *et al.* [42] discuss the application of time-series analysis in the ICU for monitoring lab variables and prediction of individual patient response to therapeutic interventions, in the context of monitoring of blood pressure lactate after liver resections and acute respiratory distress syndrome.

2) Medication Dosing: Another important predictive question encountered in the ICU is that of medication dosing. A recent study by Ghassemi *et al.* [148] highlighted that the misdosing of medications in the ICU is both problematic and preventable. Their paper showed that up to two-thirds of patients at the study institution received a non-optimal initial dose of heparin and that the problem persisted regardless of the initial dose, due to the highly personal and complex factors that affect the dose–response relationship. They utilized a joint LR model and routinely collected clinical variables (e.g., race, ICU type, gender, age, and sequential organ failure assessment) to estimate a personalized initial dose of heparin. Their model had improved performance compared to a model based on weight alone (increase in volume under surface, a multiclass version of the AUC measure, of 0.06).

Ghassemi *et al.* extended their work to consider the problem of learning an optimal medication dosing policy individualized to a patent's phenotype and evolving clinical state. [149]. They describe a method for dose estimation similar to [148], but estimate optimal model parameters for each patient using a weighted combination of the incoming data from the individual and available data from a population of similar patients. They demonstrated an average improvement in AUC of 0.25, 0.19, and 0.25 for the classification of subtherapeutic, therapeutic, and supratherapeutic patients, respectively, and an average improvement in AUC between their personalized and a nonpersonalized model of greater than 0.05 for all three therapeutic states.

Recently, Nemati and Adams proposed a deep reinforcement learning approach to sequential optimization of medications in the ICU [150]. Their technique aimed to learn latent factors in routinely collected clinical time series, which can be directly optimized to assist in sequential adjustment of heparin dosage. They utilized a discriminative HMM for state estimation, followed by function-approximation approach to Q-learning to learn an optimal medication dosing policy. They showed that end-to-end training of the discriminative HMM and the Q-network yielded a dosing policy superior to the hospital protocol. In fact, while the expected reward over all dosing trajectories in their cohort was negative, patients whose administered heparin trajectory most closely followed the reinforcement learning agent's policy could on average expect a positive reward (that is, spending the majority of their time within the therapeutic range).

In another example, many ICU patients experience hyperglycemia in the ICU, even if not diabetic. To predict future insulin requirements, Nachimuthu *et al.* used an expert-informed Bayesian network structure, with the values of its parameters determined using expectation maximization (to accommodate missing data) [151].

B. State Estimation

Even with the vast resources available in modern intensive care, there remain many parameters that cannot be directly measured in the ICU. For example, while many clinicians are primarily interested in evaluating cardiac output, no thoroughly validated device for its measurement is available, and various models or approximations must be utilized for its estimation. In this instance, cardiac output can be considered as a latent state, from which we measure noisy observations. In general, many aspects of patient health are not directly

measurable, but can be inferred through the use of state space approaches.

1) Time-Series-Based Estimation of Physiological States: Application of KFs in critical care has a long history extending beyond the artifact detection approaches discussed earlier. For instance, in the early 1980s, Smith *et al.* [152] applied a KF to the time-series data from a group of kidney transplant patients, where they were able to show that in some patients, algorithmic detection of kidney rejection preceded that of experienced clinicians.

Another method for incorporating temporal information into disease prognosis is through dynamic Bayesian networks (DBNs), which are extensions of probabilistic graphical models to allow modeling of temporal data. The nodes of a DBN correspond to the random variables of interest, edges indicate the relationship between these random variables, and additional edges model the time dependency. DBNs have the desirable property that they allow for interpretation of the interactions between different variables, which is not the case for "black box" methods such as SVMs and the traditional ANNs. Gather *et al.* [153] pioneered the application of DBNs to model the conditional dependence structure of physiological variables. DBNs have been applied to the problem of parsing continuous waveforms collected at the bedside of an adult or neonatal patient for clinically significant events [154]. van der Heijden *et al.* used a DBN to model variables such as sputum volume, temperature, and blood oxygen saturation for patients with chronic obstructive pulmonary disease in order to predict exacerbation events [155].

Lehman *et al.* [170] propose an unsupervised approach for the discovery of patient state. A switching vector autoregressive (SVAR) model was applied to minute-by-minute heart rate and blood pressure measurements, with the goal of patient state estimation and clinical outcome prediction. In the absence of clinical labels for the patient time series, an expectation–maximization algorithm was used to simultaneously segment the patient data into several phenotypic dynamical states and learn parameters of an AR model to best explain each segment. The proportion of time spent within a given dynamical region was then used as an input to a classifier for patient outcome prediction.

This approach has the advantage of automating the process of finding dynamical motifs in patient data in the absence of clinical labels, at the expense of an increase in complexity of the inference and learning algorithm. These methods have a further advantage of maintaining a belief state (that is, a probability distribution over the unobserved state variables) over the true physiological values of a patient when these cannot be directly observed due to artifact. They thus are able to provide the clinician with an estimate of the underlying true physiology, even in the presence of total corruption by noise.

2) Time-Series Search and Clustering: To enable personalized treatments, one may need to query a database for patients who match static and dynamics features of a given patient. Although much work has been performed on relational database searches, the issue of searching though time series is relatively unexplored in critical care data. Time-series search has a broad range of applications from finance to medical informatics, however, robust algorithms for finding predictive patterns in long sequences of nonstationary multivariate time series are sparse [156]. Moreover, robust navigation and mining of physiological time series often requires finding similar temporal patterns of physiological responses. Detection of these complex physiological patterns not only enables demarcation of important clinical events but can also elucidate hidden dynamical structures that may be suggestive of disease processes. Some specific examples where physiological signal search may be useful include real-time detection of cardiac arrhythmias, sleep staging or detection of seizure onset. In all these cases, being able to identify a cohort of patients who exhibit similar physiological dynamics could be useful in prognosis and informing treatment strategies. However, pattern recognition for physiological time series is complicated by changes between operating regimes and measurement artifacts.

A very related topic to time-series similarity is that of time-series clustering. Clustering methods for time-series data is often more challenging than clustering of static data primarily because the distance metric between two time series is less well-defined. Numerous distance metrics have been proposed, including the Euclidean distance, Pearson's correlation factor and dynamic time warping. As categorized by Liao, there are three different approaches for clustering time-series data: using the raw time series as input, using features extracted from the raw data, or by presuming an underlying model of the data [157]. Unsupervised approaches can be used not only as standalone analyses, but also within two-step algorithms to generate features as input for secondary supervised analyses. This is particularly appropriate when it is unclear which aspects of the data may be discriminatory (e.g., within a complex physiologic time series), or when it is suspected that the underlying structure in the data correlates with the desired outcome predictor variable.

Saeed *et al.* transformed patient time series into a symbolic representation using wavelet decomposition and subsequently applied term informativeness techniques [158] to identify similar patterns in blood pressure waveforms. Lehman *et al.* [159] developed a vectorized threshold and gradient-based search engine, which allowed users to identify patients (and episodes) which fit specific criteria. By precomputing maximum values, minimum values, and gradients over multiple scales for all time series for all patients, the authors were able to accurately

identify episodes indicative of acute myocardial infarction, lactic acidosis, acute kidney injury, hemodynamic instability, multiorgan failure, and paroxysmal tachyarrhythmia. Subsequent work by the same authors [160] employed a Gaussian mixture model approach to learn the dynamic patterns in physiology through expectation–maximization. Similarity between segments was computed using the Mahalanobis distance. Sow *et al.* [161] demonstrated that clustering similar patients together using locally supervised metric learning reduced the error in physiology forecasting algorithms.

In [162], Nemati and Ghassemi proposed a framework for distributed identification of dynamical patterns in physiological time series using a switching KF. Moreover, they described a fast and memory-efficient algorithm for learning and retrieval of phenotypic dynamics in large clinical time-series databases. Through simulation they showed that the proposed algorithm is at least an order of magnitude faster that the state of the art, and provided encouraging preliminary results based on real recordings of vital sign time series from the MIMIC-II database. The switching KF framework allows for defining a notion of "similarity" among multivariate physiological time series based on their underlying shared dynamics. Therefore, one may consider two subjects to be similar if their underlying vital sign time series exhibit similar dynamics in response to external (e.g., tilting of body) or internal perturbations (e.g., onset of blood infection). This approach provides an improvement over time-series similarity measures based on trend-detection [163], wavelet-based symbolic representations [164], or Gaussian mixture modeling [160] due to its compact representation and sharing of the model parameters within and across time series.

Hauskrecht *et al.* [165] applied time-series similarity measures for the opposite task: to locate abnormal patients and alert physicians when possible. The authors built a model for many possible clinical treatment actions using archived data collected in a patient's EHR. The model they developed would alert if the probability of an event, either administration of treatment or omission of treatment, strongly differed from the action taken. An example task was heparin delivery, and the model would alert if heparin was given to the current patient when the probability of heparin being given to similar patients in the past was very low. These alerts were generated using a SVM trained for each possible action, and the features were extracted from a 24-h segmentation of patient time-series data.

Saria *et al.* [166] framed neonatal vital signs as having an underlying set of "topics," in an analogous manner to document clustering. This approach allowed the authors to learn the associations between different "words," or features of the signal, and these larger "topics." Such unsupervised analyses provided insight into patient similarities, which can drive the generation of features that are important for discrimination between patient states [147].

Schulam *et al.* [167] took a different approach to a time-series clustering model, in which they defined a set of generative linear prototype functions to describe the behavior of individual clinical features over time for patients with scleroderma (a connective tissue disease). Ross and Dy [168] developed a set of nonparametric models for clustering patient time-series data that use a Dirichlet mixture of GPs, as well as take into account domain knowledge. In their application area of COPD patients, they were able to relate their identified subgroups to the presence of several genetic mutations known to be associated with certain forms of COPD. Though these latter two examples are drawn from applications of chronic disease, similar approaches are relevant for critical care situations.

In some applications, this two-stage procedure—unsupervised feature extraction followed by supervised learning for outcome discrimination—may be suboptimal, since the latent dynamics that are important to the supervised target may only be weakly related to those that are best for explaining the raw statistics of the time series. Additionally, generative approaches to unsupervised feature learning [169], [170] may be hamstrung by the shortcomings of approximate inference, or the underlying models may be underspecified with respect to the nuanced features associated with the outcomes of interest. For instance, in a neurophysiological experiment involving EEG recordings, it may be the case that only a single low amplitude oscillation is the distinguishing feature of successful trials, and therefore a reduced-model specifically trained to capture that oscillation may provide a more parsimonious solution to the problem of predicting outcomes of each trial. It is therefore desirable to learn models of time-series dynamics in which the latent variables are directly tuned towards the supervised task of interest.

In [171], a learning algorithm specifically designed to learn dynamical features of time series that are directly predictive of the associated labels was presented. Rather than depending on label-free unsupervised learning to discover relevant features of the time series, a system that expressly learns the dynamics that are most relevant for classifying time-series labels is built. The goal is to obtain compact representations of nonstationary and multivariate time series, a task frequently referred to as representation learning [172]. To accomplish this, the authors used a connection between DBNs (e.g., the switching VAR model) and ANNs to perform inference and learning in state-space models, in a manner analogous to backpropagation in neural networks [173]. This connection stems from the observation that the directed acyclic graph structure of a state-space model can be unrolled both as a function of time and inference steps to yield a deterministic neural network with efficient parameter

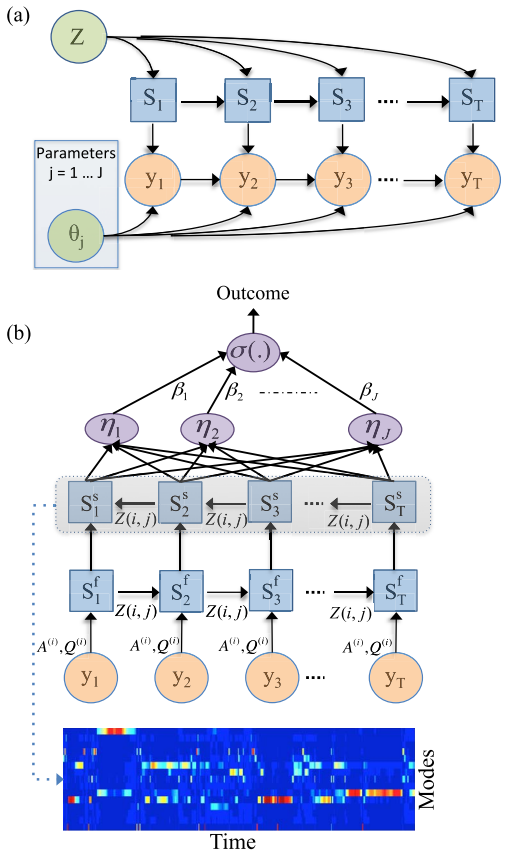

Fig. 5. ***Supervised learning in dynamic Bayesian networks. Graphical model representation of the switching vector autoregressive (switching VAR) is depicted in panel (a). Panels (b) shows the unrolled representation (with respect to time and inference steps) of the two models, with an added logistic regression layer (elliptic nodes) which utilize the marginals over the discrete latent variables as features for time-series classification [an example of inferred marginals is shown at the bottom of the panel (b)]. These unrolled structures, which resemble recurrent neural networks, allow for efficient supervised learning and inference via error backpropagation.***

tying across time (see Fig. 5). In contrast to generative and maximum-likelihood-based approaches to feature learning in time series, the outcome-discriminative learning framework provides the learning algorithm with the outcomes (labels) corresponding to each time-series sample (e.g., supine, slow-tilt, etc.) or the entire time series (responders versus nonresponders), and learns time-series features that are maximally discriminative. The method allowed for combining unsupervised dynamics discovery with supervised fine-tuning to design and initialize a new class of models for dynamic phenotyping, and development of phenotype-informed predictive models.

C. Specific Advances in Modeling

There are some modeling advances that are worth mentioning specifically, as they are particularly useful in the face of the complexity of data found in critical care settings.

1) Non-Parametric Bayesian Approaches: The new field of Bayesian nonparametrics has gained much attention in recent years due to the fact that it offers a tractable means of tackling "big data" problems, where the complexity of models can scale with the increasing size and complexity of the data that are encountered.

As with conventional (parametric) Bayesian methods, nonparametric Bayesian algorithms allow the specification of prior knowledge in a principled manner, but where the distributions involved are typically defined over objects of infinite dimensionality [174]. This yields models that make fewer constraining assumptions about the underlying mechanism assumed to have generated the observed data, and which therefore offer the possibility of scaling to very large data sets that would otherwise not be possible. For example, rather than assuming that a time series of physiological data comprises a number of individual data-points that are independent and identically distributed (i.i.d.) with respect to some underlying probability distribution of constrained parametric form, the Bayesian nonparametric approach is to define a probability distribution over the infinite-dimensional space of functions of which the observed data are an instantiation. That is, we move from the conventional notion of point-by-point analysis, which is the current state-of-the-art in patient monitoring, to one in which entire functions are analyzed (i.e., functional data analysis) [175]. This latter approach closely matches the manner in which human experts perform inference: a clinician will analyze an entire time series by comparing it with the prior knowledge gained from their clinical training and experience, rather than by performing a series of independent decisions on each data point within a time series.

Clifton *et al.* illustrate how patient-specific GP regression can be used to identify patient deterioration much earlier than would be possible using traditional methods [176]. Using wearable ECG and pulse oximetry sensors to acquire data from ambulatory patients recovering from surgery [177], the authors use GPs to model the time series of each vital sign. A functional approach was taken in [178], and related approaches [179]–[181] extend extreme value statistics over highly multivariate spaces, with applications in fusing data from patient monitoring systems. Such methods were shown to perform favorably with respect to nonprobabilistic systems [182].

More recent work in the area of GP-based approaches to critical care [181] demonstrated their use in combining data from wearable sensors with those obtained from manual nursing observations in acute wards. The flexibility of the GP framework was demonstrated by Durichen *et al.* [183], in which multiple time series were fused in a Bayesian nonparametric framework for further improvements in time-series patient monitoring.

The functional approach to data analysis in critical care was used to identify common trajectories of HR and breathing rate following surgery [184]. After fitting a GP to each patient's vital signs, the authors computed a likelihood-based similarity metric between each patient-specific GP (essentially determining the likelihood that one patient's GP accurately models a second patient's time-series data). Hierarchical clustering was then used on the values of the inter-GP similarity metric to group these trajectories. Previously unseen test data were compared to the time-series clusters to determine if the test data were similar to "normal" or "abnormal" clusters. The GP-based approach was able to more accurately discriminate normal from abnormal physiological trajectories than the state-of-the-art dynamic time warping [158]. Such techniques allow for detection of impending physiological deterioration via time-series-based similarity matching of a patient to the existing patients within a database with known outcomes.

2) Global Optimization for Cohort-Specific Parameter Tuning: Many algorithms used for the analysis of physiological signals include hyperparameters that must be selected by the investigator. The ultimate choice of these parameter values can have a dramatic impact on the performance of the approach [185]. Addressing this issue often requires investigators to manually tune parameters for their particular data set. In general, global optimization approaches are best motivated for objective functions which are both costly to evaluate and whose performance is sensitive to parametrization. As concluded in [186], recent advances in global optimization techniques provide an effective and automated framework for tuning parameters of such algorithms, and easily improve upon the default settings selected by experts.

Bayesian optimization (BO) [187] is one such methodology for global optimization that relies on building and querying a relatively inexpensive probabilistic surrogate of a more expensive objective function. In general, the surrogate is a GP, which when combined with observations yields a convenient posterior distribution over functions. Intuitively, the optimization routine proceeds by exploring through seeking regions of high posterior uncertainty in the surrogate and exploiting by evaluating regions with a promising expected value. At each iteration the routine proposes a set of hyperparameters that maximizes the expected improvement over the best result seen. An experiment is run with these hyperparameters and then the surrogate model is updated with the result. This process continues over several iterations until some threshold is reached, or a maximal number of iterations surpassed.

In [186], it was shown that BO can outperform the traditional global optimization techniques such as the standard grid search, multistart scatter search algorithm, and genetic algorithms, given the same computational and time constraints.

3) Growing Volume of Data: Many of the early studies on ICU patient prognosis relied on small samples sizes for model building, but recent trends in hardware and data collection have dramatically increased clinical database sizes. In 1981, the APACHE I system was validated on a data set of 581 admissions, while the APACHE IV system was validated in 2006 on a data set of over 44 000 patients [31], [130].

As the number of examples and feature sets grow larger, fast and efficient algorithms become more important. Fan *et al.* present an efficient method for clustering large amounts of patient data by creating a hierarchical structure [188]. Kale *et al.* present a method they term "kernalized locality-sensitive hashing" for efficiently evaluating various similarity metrics for time-series data [189].

The increasing availability of large volumes of patient data is also making it possible to apply more powerful "data hungry" machine learning techniques to clinical problems. Lasko *et al.* [190] applied a deep learning-based approach to unsupervised learning of phenotypical features in longitudinal sequences of serum uric acid measurements. The resulting unsupervised phenotypic features were passed to a classifier to distinguish the uric acid signatures of gout versus acute leukemia, with a performance level competitive with the gold-standard features engineered by domain experts.

D. Multimodal Data

While the majority of this review has focused upon vital sign data that are commonly available in the ICU, there are many additional sources of data that can be used to improve decision support in critical care. However, care must be taken: there is not always a benefit in incorporating certain types of additional data. For instance, Saria *et al.* found that adding laboratory test values as features did not improve prediction [147], consistent with other studies that have found high amount of correlation among features [138]. The key therefore lies in appropriate combination of additional information available in the patient record.

In one novel approach, Wiens *et al.* first created a day-by-day patient risk score for becoming infected by *Clostridium difficile* [191]. This risk score was derived from an SVM classifier with $> 10\,000$ features from the patient EHR as input. Features included the reason for admission, demographics, lab results, room location, vital sign measurements, etc. (binary features were created from categorical variables, which accounts for most of the high dimensionality). The authors then modeled this risk score as a time series, using three different approaches (extracted features, similarity metrics, and HMMs) to perform classification. Their methods were able to predict patient risk more successfully than traditional approaches of taking aggregate or daily features,

with AUROCs of up to 0.79 in contrast to the traditional approaches' AUROC of 0.69.

1) Incorporation of Genomic Data: One particular data type that historically has not been used widely in patient decision support is that of genomic data. While our growing understanding of patient genomics and gene expression is likely to greatly improve our ability to treat disease in the future, there are a few medical areas in which machine learning applications of genomics are already being adopted.

Clinical microbiology is one such area, which impacts closely with critical care given the high risk of infection for patients who have extended ICU stays. While *human* genetic information is not yet available in most EHR and clinical decision systems, bacterial and viral DNA analysis is more manageable (due to the smaller size of such genomes when compared with the human genome) and has already started to be incorporated into some hospital systems. Using this available information, machine learning techniques have been employed to predict bacterial and viral phenotypes from the genotype. Prediction of viral drug resistance is a pressing problem for many viruses, such as Human Immunodeficiency Virus (HIV). Both rule-based methods (e.g., ANRS, Rega, and Stanford HIVdb [192]) and machine-learning techniques (e.g., geno2pheno [193]) have been developed to improve genotypic prediction of HIV drug susceptibility. Machine-learning methods have been found to predict more accurately the response of patients to drugs in retrospective analysis than do rule-based methods used for the same task [194].

Machine learning techniques have also been used to predict virulence profiles of clinically relevant microorganisms. In 2014, Laabei *et al.* used whole-genome data to predict the virulence of methicillin resistant *S. aureus* using random forests [195]. Alternative methods for bacterial resistance prediction has been attempted using LR, random forests, and set covering machines [196]–[198].

2) Mining of Free-Text Clinical Notes: Given the explanatory power of physician notes for discounting anomalous measurements (as discussed above) and their ability to capture information not easily obtained elsewhere, there is great potential for clinical notes to improve machine learning-based prediction in the ICU setting.

Lehman *et al.* [199] used a hierarchical Dirichlet process (HDP) to perform patient risk stratification by combining physiologic data and topics learned from unstructured clinical notes. The authors found that the learned topic structures significantly improved the performance of the SAPS-I algorithm for mortality prediction (from 0.72 to 0.82).

Ghassemi *et al.* [200] used a multistep pipeline to predict ICU mortality. They first used latent Dirichlet allocation (LDA) to identify common words and topics recorded in ICU patient notes. They then fit multitask GPs to the proportion of topics observed in each note in each patient's record. Finally, as features for supervised learning to predict mortality, they used the GP hyperparameters, time-averaged topic membership, and a standard ICU-admission clinical scoring system (simplified acute physiology score: SAPS-1), finding that the combination of these features provided improved predictive performance over the clinical scoring system alone.

Ghassemi *et al.* [201] also utilized an unsupervised approach to generate vector space representations of unstructured free-text notes. They investigated the evolution of clinical sentiment and language complexity with respect to several categories including: mortality, time in the hospital, age, race, and gender. Their analysis identified greater positive sentiment for females, unmarried patients, and patients of African ethnicity in the ICU.

Even simple counts of textual terms and completed fields in the EHR can be informative in risk prediction. Nurses have been found to document 0.9–1.5 more optional comments and 6.1 to 10 more vital signs within the 48 h before patient death [202].

V. DISCUSSION

This review has summarized the latest trends in machine learning in critical care. Focus has been given to all components necessary in this field: acquisition of data, assurance of quality, and final analysis. A large amount of effort has been invested in the processing and validation of data acquired within the ICU. Many of these methods are necessary due to the relatively unique format of data collection in the ICU. When developing algorithms in other domains, such as aircraft health monitoring or finance, researchers will specifically collect data for the purpose of analysis. However, most applications of machine learning in the ICU are secondary, that is, the data is collected for a purpose other than for the analysis proposed. Frequently, the data collected is acquired during routine clinical care where there are little to no incentives for acquisition of accurate data. In fact, those who record the data are frequently prevented from auditing and correcting the observations due to extreme time constraints. While advanced data management systems have the opportunity to improve clinical work flow and facilitate higher quality data collection, vendors in the health care field have produced notoriously inefficient systems which lag a great deal behind similar systems in "civilian" areas [205].

The end result is a wealth of data being collected in ICUs across the world daily going to waste [204]. Of the data that has been successfully archived and retrieved, a significant amount of effort must be employed to either transform the data into a usable form or correct a variety of artifacts present. As demonstrated in this review, a number of researchers have developed

excellent techniques which address these data quality issues. These methods have allowed for further processing of the data with confidence, either for outcome prediction, state estimation, or patient alerting.

While machine learning research in critical care has provided the community with a wealth of knowledge on how patient care could be improved by the use of automated algorithms assessing patients, two criticisms arise. First, while many high performance algorithms have been proposed, there has been a paucity of evidence for the efficacy of these algorithms once implemented in ICUs.

Second, an objective analysis would imply that the sophistication of the machine learning methods applied in the critical care domain lag behind those applied in other areas. Many explanations of this could be conceived, including the earlier discussed lack of consistent and reliable data management systems in hospitals. However, we would posit that one of the biggest barriers to research has been the lack of openly available standardized data sets for the purpose of benchmarking machine learning tasks. Recent advances in image classification have been achieved in no small part due to the openly available Imagenet database which contains 456 567 images for classification as of 2014 [206]. No equivalently sized database exists for critical care. Given the complexity and heterogeneity of critical care data, and the variance in clinical practices, millions of patients are needed to identify subcohorts of particular disease processes and the range of applied clinical actions.

Yet, there are notable success stories surrounding open data in the past. The MIT-BIH arrhythmia database [208] galvanized manufacturers into reporting, and consequently improving, performance of their algorithms on ECG signals with arrhythmia. It was clear that, prior to the release of MIT-BIH, the lack of a well-defined database for this purpose not only hindered academic progress on arrhythmia detection, but also hindered the ability of manufacturers to systematically evaluate their methods. Leaps in performance similar to those achieved after the release of MIT-BIH could be attainable in a variety of machine learning tasks after the creation of suitable standardized benchmark datasets. The need for high quality databases in critical care, with information that is complete and accurate, based upon standardized definitions of clinical disorders, interventions, and outcomes has already been recognized [208]. The creation of openly available databases such as MIMIC [22] is a key step toward this goal, and the recent announcement that a subset of the eICU database [30] will be made open to the public demonstrates that this practice is becoming more common. Future directions should strive to define and describe benchmark data sets, much like the PhysioNet/Computing in Cardiology 2012 challenge defined a benchmark data set for mortality prediction [142]. It is worth noting that the benchmark data set for mortality prediction resulted in state-of-the-art algorithms with over 170% higher performance than their severity score predecessors [143].

Many tasks reviewed here would benefit from benchmark data sets and, more generally, further research. A large proportion of work that addressed data corruption was ultimately used for the purpose of false alarm reduction. Drew *et al.* [65] reviewed the issue of alarm fatigue associated with false alarms and suggested alarm algorithms should focus on: using all available ECG leads and extracting at least one lead with high quality data if available, providing contextual alarms based upon multiple features (e.g., only alerting staff to preventricular contractions if the patient has a prolonged QT interval), accommodating and learning from human alarm threshold adjustment, and "smart" defaults which adjust to the patient using some subset of initialization data.

Quantification of a signal into states is a principled and robust approach which has been shown to work well for both arterial blood pressure artifact detection [77] and ECG beat classification [96]. In terms of artifact detection, many known signal disruptions could be quantified in this way, including calibration artifacts, suctioning artifacts (which occur when a care provider is clearing ventilation equipment for a patient), and motion artifact. The automatic determination of artifact data would facilitate future research on the relationship between physiological dynamics and patient health. In terms of beat detection, previous research has primarily addressed ventricular ectopic beats, but many arrhythmia of interest have yet to be addressed, including atrial ectopics, asystole, atrial fibrillation, atrial flutter, bundle branch block, and so on. In general, there remains a need for openly available high performance algorithms capable of segmenting a physiologic waveform into components (e.g., segmentation of the ECG into "P," "QRS," and "T"). This could be facilitated if equipment manufacturers transmitted their confidence levels in parameter estimates. Such confidence levels could be incorporated into prediction algorithms, which could be used to greatly improve performance.

Mortality prediction models appear to have reached a plateau, with the performance of the latest generation models being fairly close to their predecessors. The primary reason for such is likely the very coarse data used in the model input, usually average values over 24 h. The incorporation of dynamics has been shown to improve these models [170], and future research is warranted in this exciting area. Many of these models could be applied to the technically similar task of predicting readmission, where a high performing model could have many ramifications due to the large economic penalties incurred to hospitals when a patient is readmitted within 30 days.

Looking even further forward, there is an urgent need for integrative and interactive machine learning solutions, with teams of machine learning researchers and

clinicians—who are directly involved in patient care and data acquisition—working in tandem to generate actionable insight and value from the increasingly large and complex critical care data [205]. The data deluge has overwhelmed many clinicians and researchers, and in the future, *smart* hospitals, which utilize machine learning approaches to provide information in a context aware manner, will be necessary [128]. Dimensionality reduction and visualization techniques are exciting areas of research which have the potential of redefining the single sensor single input monitoring approach currently applied in clinical practice. Overall, a growing body of literature [6] is pointing to the clinical utility of big data in critical care to inform prognosis and to provide early predictors of potentially life-threatening conditions in the ICU. As researchers begin to pool resources to generate large open access data sets [22], the "Unreasonable Effectiveness of Data" is beginning to take effect. However, as we note in this article, the nuances of healthcare require extreme care to be taken in the acquisition and processing of critical care data. The meaningful secondary uses of EHRs can only take place if such issues are addressed. Careful consideration of the compartmentalization, corruption, and complexity of clinical data has created a unique climate of research in critical care, which has great potential. ■

REFERENCES

[1] J.-L. Vincent, "Critical care-where have we been and where are we going," *Crit. Care*, vol. 17, p. S2, 2013.

[2] P. Pronovost, D. Angus, T. R. Dorman, K. A. Dremsizov, and T. T. Young, "Physician staffing patterns and clinical outcomes in critically Ill patient: A systematic review," *JAMA*, vol. 288, no. 17, pp. 2151–2162, 2002.

[3] R. Kane, T. Shamliyan, C. Mueller, S. Duval, and T. J. Wilt, "The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis," *Med. Care*, vol. 45, no. 12, pp. 1195–1204, Dec. 2007.

[4] S. M. Pastores and V. Kvetan, "Shortage of intensive care specialists in the united states: Recent insights and proposed solutions," *Revista Brasileira de terapia intensiva*, vol. 27, no. 1, pp. 5–6, 2015.

[5] N. A. Halpern and S. M. Pastores, "Critical care medicine in the United States 2000–2005: An analysis of bed numbers, occupancy rates, payer mix, costs," *Crit. Care Med.*, vol. 38, no. 1, pp. 65–71, 2010.

[6] O. Badawi *et al.*, "Making big data useful for health care: A summary of the inaugural MIT critical data conference," *JMIR Med. Inf.*, vol. 2, no. 2, p. e22, 2014.

[7] G. F. Riley, "Administrative and claims records as sources of health care cost data," *Med. Care*, vol. 47, pp. S51–S55, 2009.

[8] A. E. W. Johnson, A. Kramer, and G. D. Clifford, "Data preprocessing and mortality prediction: The Physionet/CinC 2012 challenge revisited," in *Proc. Comput. Cardiol. Conf.*, 2014, vol. 41, pp. 157–160.

[9] Centers for Medicare & Medicaid Services, "The Health Insurance Portability and Accountability Act of 1996 (HIPAA)," 1996. [Online]. Available: http://www.cms.hhs.gov/hipaa/

[10] F. Caldicott, "Information: To share or not to share. The Information governance review," 2013.

[11] D. C. Ince, L. Hatton, and J. Graham-Cumming, "The case for open computer programs," *Nature*, vol. 482, no. 7386, pp. 485–488, 2012.

[12] R. B. Ness, Joint Policy Committee, "Influence of the HIPAA privacy rule on health research," *JAMA*, vol. 298, no. 18, pp. 2164–2170, 2007.

[13] C. M. O'Keefe, "Privacy and the use of health data-reducing disclosure risk," *Electron. J. Health Inf.*, vol. 3, no. 1, p. 5, 2008.

[14] H. Office for Civil Rights, "Standards for privacy of individually identifiable health information. final rule," *Fed. Register*, vol. 67, no. 157, p. 53181, 2002.

[15] P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam, The Netherlands: Elsevier Science, 2004.

[16] I. Neamatullah *et al.*, "Automated de-identification of free-text medical records," *BMC Med. Inf. Decision Making*, vol. 8, no. 1, p. 32, 2008.

[17] S. N. Murphy *et al.*, "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)," *J. Amer. Med. Inf. Assoc.*, vol. 17, no. 2, pp. 124–130, 2010.

[18] C. Dwork, "Differential privacy," in *Encyclopedia of Cryptography and Security*. New York, NY, USA: Springer-Verlag, 2011, pp. 338–340.

[19] N. Mohammed, X. Jiang, R. Chen, B. C. Fung, and L. Ohno-Machado, "Privacy-preserving heterogeneous health data sharing," *J. Amer. Med. Inf. Assoc.*, vol. 20, no. 3, pp. 462–469, 2013.

[20] A. Goldberger, L. Amaral, and L. Glass, "PhysioBank, PhysioToolkit, PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[21] M. Saeed, C. Lieu, G. Raber, and R. G. Mark, "MIMIC II: A massive temporal ICU patient database to support research in intelligent patient monitoring," *Comput. Cardiol.*, vol. 29, pp. 641–644, 2002.

[22] M. Saeed *et al.*, "Multiparameter intelligent monitoring in intensive care (MIMIC II): A public-access intensive care unit database," *Crit. Care Med.*, vol. 39, no. 5, pp. 952–960, May 2011.

[23] U.S. Food and Drug Administration, "Registration Listing," Jun. 2015. [Online]. Available: http://www.fda.gov/MedicalDevices/DeviceRegulationand Guidance/HowtoMarketYourDevice/RegistrationandListing/ucm134495.htm

[24] K. Lesh, S. Weininger, J. M. Goldman, B. Wilson, and G. Himes, "Medical device interoperability-assessing the environment," in *Proc. Joint Workshop HCMDSS-MDPNP*, 2007, pp. 3–12.

[25] D. Charles, J. King, V. Patel, and M. F. Furukawa, "Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008–2012," *ONC Data Brief*, no. 9, 2013.

[26] N. Black and M. Payne, "Directory of clinical databases: Improving and promoting their use," *Quality Safety Health Care*, vol. 12, no. 5, pp. 348–352, 2003.

[27] C. R. Cooke and T. J. Iwashyna, "Using existing data to address important clinical questions in critical care," *Crit. Care Med.*, vol. 41, no. 3, p. 886, 2013.

[28] T. J. Iwashyna, E. W. Ely, D. M. Smith, and K. M. Langa, "Long-term cognitive impairment and functional disability among survivors of severe sepsis," *JAMA*, vol. 304, no. 16, pp. 1787–1794, 2010.

[29] J. M. Finney, A. S. Walker, T. E. Peto, and D. H. Wyllie, "An efficient record linkage scheme using graphical analysis for identifier error detection," *BMC Med. Inf. Decision Making*, vol. 11, no. 1, p. 7, 2011.

[30] M. McShea, R. Holl, O. Badawi, R. R. Riker, and E. Silfen, "The EICU research institute-a collaboration between industry, health-care providers, academia," *IEEE Eng. Med. Biol. Mag.*, vol. 29, no. 2, pp. 18–25, 2010.

[31] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, "Acute physiology and chronic health evaluation (apache) iv: Hospital mortality assessment for today's critically ill patients," *Crit. Care Med.*, vol. 34, no. 5, pp. 1297–1310, 2006.

[32] J. E. Zimmerman and A. A. Kramer, "Outcome prediction in critical care: The acute physiology and chronic health evaluation models," *Current Opinion Crit. Care*, vol. 14, pp. 491–497, 2008.

[33] U.S. Dept. Health Human Services, "ICD 9 CM. The International Classification of Diseases. 9. Rev: Clinical Modification.; 1: Diseases: Tabular List.; 2: Diseases: Alphabetic Index.; 3: Procedures: Tabular List and Alphabetic Index," U.S. Government Printing Office, 1980.

[34] K. J. O'malley *et al.*, "Measuring diagnoses: ICD code accuracy," *Health Services Res.*, vol. 40, no. 5p2, pp. 1620–1639, 2005.

[35] J. P. Pestian *et al.*, "A shared task involving multi-label classification of clinical free text," in *Proc. Workshop BioNLP 2007: Biol. Transl. Clin. Lang. Process.*, 2007, pp. 97–104.

[36] L. Bos and K. Donnelly, "Snomed-CT: The advanced terminology and coding system for ehealth," *Stud. Health Technol. Inf.*, vol. 121, pp. 279–290, 2006.

[37] P. L. Elkin *et al.*, "Evaluation of the content coverage of snomed CT: Ability of snomed clinical terms to represent clinical

problem lists," *Mayo Clin. Proc.*, vol. 81, no. 6, pp. 741–748, 2006.

[38] C. J. McDonald *et al.*, "Loinc, a universal standard for identifying laboratory observations: A 5-year update," *Clin. Chem.*, vol. 49, no. 4, pp. 624–633, 2003.

[39] P. Whetzel *et al.*, "Bioportal: Enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications," *Nucleic Acids Res.*, vol. 39, pp. w541–w545, Jul. 2011.

[40] J. D D'Amore *et al.*, "Are meaningful use stage 2 certified EHRS ready for interoperability? Findings from the smart c-CDa collaborative," *J. Amer. Med. Inf. Assoc.*, vol. 21, no. 6, pp. 1060–1068, 2014.

[41] K. Nouira and A. Trabelsi, "Intelligent monitoring system for intensive care units," *J. Med. Syst.*, vol. 36, no. 4, pp. 2309–2318, 2012.

[42] M. Imhoff, M. Bauer, U. Gather, and D. Löhlein, "Statistical pattern detection in univariate time series of intensive care on-line monitoring data," *Intensive Care Med.*, vol. 24, no. 12, pp. 1305–1314, 1998.

[43] M. West, P. J. Harrison, and H. S. Migon, "Dynamic generalized linear models and Bayesian forecasting," *J. Amer. Stat. Assoc.*, vol. 80, no. 389, pp. 73–83, 1985.

[44] C. Becker and U. Gather, "The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules," *Comput. Stat. Data Anal.*, vol. 36, no. 1, pp. 119–127, 2001.

[45] S. Nizami, J. R. Green, and C. McGregor, "Implementation of artifact detection in critical care: A methodological review," *IEEE Rev. Biomed. Eng.*, vol. 6, pp. 127–142, 2013.

[46] C. L. Tsien and J. C. Fackler, "Poor prognosis for existing monitors in the intensive care unit," *Crit. Care Med.*, vol. 25, no. 4, pp. 614–619, 1997.

[47] M. C. Chambrin *et al.*, "Multicentric study of monitoring alarms in the adult intensive care unit (ICU): A descriptive analysis," *Intensive Care Med.*, vol. 25, no. 12, pp. 1360–1366, Dec. 1999.

[48] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter," *Physiol. Meas.*, vol. 29, no. 1, pp. 15–32, Jan. 2008.

[49] G. D. Clifford, J. Behar, Q. Li, and I. Rezek, "Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms," *Physiol. Meas.*, vol. 33, no. 9, p. 1419, 2012.

[50] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. 32, no. 3, pp. 230–236, 1985.

[51] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database," *IEEE Trans. Biomed. Eng.*, vol. BME-33, no. 12, pp. 1157–1165, 1986.

[52] W. Zong, G. Moody, and D. Jiang, "A robust open-source algorithm to detect onset and duration of QRS complexes," in *Proc. Comput. Cardiol.*, 2003, vol. 30, pp. 737–740.

[53] J. Behar, J. Oster, Q. Li, and G. D. Clifford, "ECG signal quality during arrhythmia and its application to false alarm reduction," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 6, pp. 1660–1666, 2013.

[54] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[55] Q. Li and G. D. Clifford, "Signal quality and data fusion for false alarm reduction in the intensive care unit," *J. Electrocardiol.*, vol. 45, no. 6, pp. 596–603, 2012.

[56] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[57] E. Morgado *et al.*, "Quality estimation of the electrocardiogram using cross-correlation among leads," *Biomed. Eng. Online*, vol. 14, no. 1, p. 59, 2015.

[58] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[59] J. R. Quinlan, *C4. 5: Programs for Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2014.

[60] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, 1995, pp. 115–123.

[61] A. Barachant, A. Andreev, and M. Congedo, "The Riemannian potato: An automatic and adaptive artifact detection method for online experiments using Riemannian geometry," in *Proc. TOBI Workshop lV*, 2013, pp. 19–20.

[62] C. L. Tsien, I. S. Kohane, and N. McIntosh, "Building ICU artifact detection models with more data in less time," in *Proc. AMIA Symp.*, 2001, p. 706.

[63] M. Imhoff, S. Kuhls, U. Gather, and R. Fried, "Smart alarms from medical devices in the OR and ICU," *Best Practice Res. Clin. Anaesthesiol.*, vol. 23, no. 1, pp. 39–50, 2009.

[64] M. Cvach, "Monitor alarm fatigue: An integrative review," *Biomed. Instrum. Technol.*, vol. 46, no. 4, pp. 268–277, 2012.

[65] B. J. Drew *et al.*, "Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients," 2014.

[66] W. Zong, G. Moody, and R. Mark, "Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure," *Med. Biol. Eng. Comput.*, vol. 42, no. 5, pp. 698–706, 2004.

[67] G. D. Clifford *et al.*, "Computing in Cardiology," 2015; 42:273–276.

[68] F. Plesinger, P. Klimes, J. Halamek, and P. Jurak, "False alarms in intensive care unit monitors: Detection of life-threatening arrhythmias using elementary algebra, descriptive statistics and fuzzy logic," in *Proc. Comput. Cardiol. Conf.*, 2015, pp. 1–4.

[69] C. H. Antink and S. Leonhardt, "Reducing false arrhythmia alarms using robust interval estimation and machine learning," in *Proc. Comput. Cardiol. Conf.*, 2015, pp. 1–4.

[70] S. Fallet, S. Yazdani, and J.-M. Vesin, "A multimodal approach to reduce false arrhythmia alarms in the intensive care unit," in *Proc. Comput. Cardiol. Conf.*, 2015 pp. 1–4.

[71] G. D. Clifford, W. J. Long, G. B. Moody, and P. Szolovits, "Robust parameter extraction for decision support using multimodal intensive care data," *Philosoph. Trans. A, Math. Phys. Eng. Sci.*, vol. 367, no. 1887, pp. 411–429, Jan. 2009.

[72] P. G. H. Metnitz *et al.*, "SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description," *Intensive Care Med.*, vol. 31, no. 10, pp. 1336–1344, Oct. 2005.

[73] J. W. Tukey, "Exploratory data analysis," 1977.

[74] V. Barnett and T. Lewis, *Outliers in Statistical Data*. New York, NY, USA: Wiley, 1994, vol. 3.

[75] A. Fialho *et al.*, "Disease-based modeling to predict fluid response in intensive care units," *Methods Inf. Med.*, vol. 52, no. 6, pp. 494–502, 2013.

[76] N. Aleks *et al.*, "Probabilistic detection of short events, with application to critical care monitoring," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 49–56.

[77] J. A. Quinn, C. K. Williams, and N. McIntosh, "Factorial switching linear dynamical systems applied to physiological condition monitoring," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1537–1551, 2009.

[78] K. Georgatzis and C. K. Williams, "Discriminative switching linear dynamical systems applied to physiological condition monitoring," 2015. [Online]. Available: http://arxiv.org/abs/1504.06494

[79] C. W. Hug, G. D. Clifford, and A. T. Reisner, "Clinician blood pressure documentation of stable intensive care patients: An intelligent archiving agent has a higher association with future hypotension," *Crit. Care Med.*, vol. 39, no. 5, pp. 1006–1014, May 2011.

[80] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast-but is it good?: Evaluating non-expert annotations for natural language tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2008, pp. 254–263.

[81] T. Zhu, A. E. W. Johnson, J. Behar, and G. D. Clifford, "Crowd-sourced annotation of ECG signals using contextual information," *Ann. Biomed. Eng.*, vol. 42, no. 4, pp. 871–884, 2014.

[82] B.-U. Kohler, C. Hennig, and R. Orglmeister, "The principles of software QRS detection," *IEEE Eng. Med. Biol. Mag.*, vol. 21, no. 1, pp. 42–57, 2002.

[83] S. Jakob *et al.*, "Detection of artifacts in monitored trends in intensive care," *Comput. Methods Programs Biomed.*, vol. 63, no. 3, pp. 203–209, 2000.

[84] P. Yang, G. A. Dumont, and J. M. Ansermino, "Sensor fusion using a hybrid median filter for artifact removal in intraoperative heart rate monitoring," *J. Clin. Monitor. Comput.*, vol. 23, no. 2, pp. 75–83, 2009.

[85] J. Allen and A. Murray, "Assessing ECG signal quality on a coronary care unit," *Physiol. Meas.*, vol. 17, no. 4, p. 249, 1996.

[86] W. Kaiser and M. Findeis, "Novel signal processing methods for exercise ECG," *Proc. IJBEM*, vol. 2, Special Issue on Electrocardiography, in Ischemic Heart Disease, 2000.

[87] L. Chen, T. McKenna, A. Reisner, and J. Reifman, "Algorithms to qualify respiratory data collected during the transport of trauma patients," *Physiol. Meas.*, vol. 27, no. 9, p. 797, 2006.

[88] A. E. W. Johnson, J. Behar, F. Andreotti, G. D. Clifford, and J. Oster, "Multimodal heart beat detection using signal quality indices," *Physiol. Meas.*, vol. 36, no. 8, p. 1665, 2015.

[89] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[90] D. F. Sittig and M. Factor, "Physiologic trend detection and artifact rejection: A parallel implementation of a multi-state Kalman filtering algorithm," *Comput. Methods Programs Biomed.*, vol. 31, no. 1, p. 1–10, 1990.

[91] J. M. Feldman, M. H. Ebrahim, and I. Bar-Kana, "Robust sensor fusion improves heart rate estimation: Clinical evaluation," *J. Clin. Monitor.*, vol. 13, no. 6, pp. 379–384, 1997.

[92] M. H. Ebrahim, J. M. Feldman, and I. Bar-Kana, "A robust sensor fusion method for heart rate estimation," *J. Clin. Monitor.*, vol. 13, no. 6, pp. 385–393, 1997.

[93] L. Tarassenko *et al.*, "Medical signal processing using the software monitor," in *Proc. DERA/IEE Workshop Intell. Sensor Process.*, 2001, pp. 3/1–3/4.

[94] S. Challa and D. Koks, "Bayesian and Dempster-Shafer fusion," *Sadhana*, vol. 29, no. 2, pp. 145–174, 2004.

[95] T. Wartzek, C. Brueser, M. Walter, and S. Leonhardt, "Robust sensor fusion of unobtrusively measured heart rate," *IEEE J. Biomed. Health Inf.*, vol. 18, pp. 654–660, 2013.

[96] J. Oster *et al.*, "Semi-supervised ECG beat classification and novelty detection based on switching Kalman filters," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 9, pp. 2125–2134, 2015.

[97] D. M. Eddy, "Variations in physician practice: The role of uncertainty," *Health Affairs*, vol. 3, no. 2, pp. 74–89, 1984.

[98] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. New York, NY, USA: Wiley, 2014.

[99] A. Vesin *et al.*, "Reporting and handling missing values in clinical studies in intensive care units," *Intensive Care Med.*, vol. 39, no. 8, pp. 1396–1404, 2013.

[100] J. A. Sterne *et al.*, "Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls," *BMJ*, vol. 338, 2009. [Online]. Available: http://dx.doi.org/10.1136/bmj.b2393

[101] S. J. Shah *et al.*, "Phenomapping for novel classification of heart failure with preserved ejection fraction," *Circulation*, vol. 131, no. 3, pp. 269–279, Jan. 2015.

[102] A. K. Waljee *et al.*, "Comparison of imputation methods for missing laboratory data in medicine," *BMJ Open*, vol. 3, no. 8, 2013, e002847.

[103] S.-H. Kim, H.-J. Yang, S.-H. Kim, and G.-S. Lee, "Physiocover: Recovering the missing values in physiological data of intensive care units," *Int. J. Contents*, vol. 10, no. 2, pp. 47–58, 2014.

[104] Q. Long and B. A. Johnson, "Variable selection in the presence of missing data: Resampling and imputation," *Biostatistics*, vol. 16, no. 3, pp. 596–610, Jan. 2015.

[105] R. P. Moreno *et al.*, "SAPS 3-From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission," *Intensive Care Med.*, vol. 31, no. 10, pp. 1345–1355, Oct. 2005.

[106] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychol. Methods*, vol. 7, no. 2, p. 147, 2002.

[107] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Stat. Soc. B, Methodol.*, vol. 39, pp. 1–38, 1977.

[108] D. B. Rubin, "Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse," in *Proc. Surv. Res. Methods Sec. Amer. Stat. Assoc.*, 1978, vol. 1, pp. 20–34.

[109] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York, NY, USA: Wiley, 2004.

[110] S. Chevret, S. Seaman, and M. Resche-Rigon, "Multiple imputation: A mature approach to dealing with missing data," *Intensive Care Med.*, vol. 41, no. 2, pp. 348–350, 2015.

[111] L. Clifton *et al.*, "Gaussian process regression in vital-sign early warning systems," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 6161–6164.

[112] T. A. Lasko, "Nonstationary Gaussian process regression for evaluating clinical laboratory test sampling strategies," in *Proc. AAAI Conf. Artif. Intell.*, Jan. 2015, vol. 2015, pp. 1777–1783.

[113] R. P. Dellinger *et al.*, "Surviving sepsis campaign: International guidelines for management of severe sepsis and septic shock, 2012," *Intensive Care Med.*, vol. 39, no. 2, pp. 165–228, 2013.

[114] P. K. Lindenauer, T. Lagu, M.-S. Shieh, P. S. Pekow, and M. B. Rothberg, "Association of diagnostic coding with trends in hospitalizations and mortality of patients with pneumonia, 2003–2009," *JAMA*, vol. 307, no. 13, pp. 1405–1413, 2012.

[115] Y. Halpern, Y. Choi, S. Horng, and D. Sontag, "Using anchors to estimate clinical state without labeled data," in *Proc. AMIA Annu. Symp.*, 2014, vol. 2014, p. 606.

[116] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2008, pp. 213–220.

[117] T. Gjørup, H. S. Kelbaek, D. L. Nielsen, S. Kreiner, and J. Godtfredsen, "Reproducibility of electrocardiographic interpretation in patients with suspected myocardial infarction. A controlled study of the effect of a training trial," Tech. Rep. 1, 1994.

[118] R. Bond *et al.*, "Eye tracking technology and the 12-lead electrocardiogram: Where the experts look?" in *Proc. 39th Annu. Conf. Int. Soc. Computerized Electrocardiogr.*, 2014.

[119] S. M. Salerno, P. C. Alguire, and H. S. Waxman, "Competency in interpretation of 12-lead electrocardiograms: A summary and appraisal of published evidence," *Ann. Internal Med.*, vol. 138, no. 9, pp. 751–760, 2003.

[120] G. Clifford, F. Azuaje, and P. McSharry, *Advanced Methods and Tools for ECG Data Analysis*. Boston, MA, USA: Artech House, 2006.

[121] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *J. Roy. Stat. Soc. C, Appl. Stat.*, vol. 28, no. 1, pp. 20–28, 1979.

[122] V. Raykar *et al.*, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 889–896.

[123] V. C. Raykar *et al.*, "Learning from crowds," *J. Mach. Learn. Res.*, pp. 1297–1322, 2010.

[124] T. Zhu, J. Behar, T. Papastylianou, and G. D. Clifford, "CrowdLabel: A crowdsourcing platform for electrophysiology," in *Proc. Comput. Cardiol. Conf.*, 2014, vol. 41, pp. 789–792.

[125] P. Welinder and P. Perona, "Online crowdsourcing: Rating annotators and obtaining cost-effective labels," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 25–32.

[126] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "The multidimensional wisdom of crowds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2424–2432.

[127] T. Zhu *et al.*, "Fusing continuous-valued medical labels using a Bayesian model," 2015. [Online]. Available: http://arxiv.org/abs/1503.06619

[128] A. Holzinger, C. Röcker, and M. Ziefle, "From smart health to smart hospitals," in *Smart Health*, Lecture Notes in Computer Science. Berlin, Germany: Springer-Verlag, 2015, vol. 8700, pp. 1–20.

[129] G. S. Power and D. A. Harrison, "Why try to predict ICU outcomes?" *Current Opinion Crit. Care*, vol. 20, no. 5, pp. 544–549, 2014.

[130] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence, "APACHE-acute physiology and chronic health evaluation: A physiologically based classification system," *Crit. Care Med.*, vol. 9, pp. 591–597, 1981.

[131] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, and E. A. Draper, "APACHE II: A severity of disease classification system," *Crit. Care Med.*, vol. 13, pp. 818–829, 1985.

[132] W. A. Knaus *et al.*, "The APACHE III prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults," *Chest*, vol. 100, no. 6, pp. 1619–1636, 1991.

[133] J. R. LeGall *et al.*, "A simplified acute physiology score for ICU patients," *Crit. Care Med.*, vol. 12, no. 11, pp. 975–977, 1984.

[134] J. R. LeGall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (SAPS-II) based on a European North-American multicenter study," *JAMA*, vol. 270, no. 24, pp. 2957–2963, Dec. 22, 1993.

[135] S. Lemeshow, D. Teres, and H. Pastides, "A method for predicting survival and mortality of ICU patients using objectively derived weights," *Crit. Care Med.*, vol. 13, pp. 519–525, 1985.

[136] S. Lemeshow, D. Teres, and J. Klar, "Mortality probability model (MPM II) based on an international cohort of intensive care unit patients," *JAMA*, vol. 270, pp. 2478–2486, 1993.

[137] T. L. Higgins *et al.*, "Assessing contemporary intensive care unit outcome: An updated mortality probability admission model (MPM0-III)," *Crit. Care Med.*, vol. 35, no. 3, pp. 827–835, Mar. 2007.

[138] A. E. W. Johnson, A. A. Kramer, and G. D. Clifford, "A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy," *Crit. Care Med.*, vol. 41, no. 7, pp. 1711–1718, 2013.

[139] R. Dybowski, P. Weller, R. Chang, and V. Gant, "Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm," *Lancet*, vol. 347, no. 9009, pp. 1146–1150, Apr. 1996.

[140] G. Clermont, D. Angus, S. DiRusso, M. Griffin, and W. Linde-Zwirble, "Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models," *Crit. Care Med.*, vol. 29, no. 2, pp. 291–296, 2001.

[141] L. S. Wong and J. D. Young, "A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks," *Anaesthesia*, vol. 54, no. 11, pp. 1048–1054, Nov. 1999.

[142] I. Silva, G. B. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, "Predicting in-hospital mortality of ICU patients: The PhysioNet/ Computing in Cardiology Challenge 2012," *Comput. Cardiol.*, vol. 39, pp. 245–248, 2012.

[143] A. E. W. Johnson *et al.*, "Patient specific predictions in the intensive care unit using a Bayesian ensemble," *Comput. Cardiol.*, vol. 39, pp. 249–252, 2012.

[144] L. Citi and R. Barbieri, "Physionet 2012 challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm," *Comput. Cardiol.*, vol. 39, pp. 257–260, 2012.

[145] R. Pirracchio *et al.*, "Mortality prediction in intensive care units with the super ICU learner algorithm (sicula): A population-based study," *Lancet Respirat. Med.*, vol. 3, no. 1, pp. 42–52, 2015.

[146] C. C. Mayer *et al.*, "Selection of entropy-measure parameters for knowledge discovery in heart rate variability data," *BMC Bioinf.*, vol. 15, p. S2, 2014.

[147] S. Saria *et al.*, "Integration of early physiological responses predicts later illness severity in preterm infants," *Sci. Transl. Med.*, vol. 2, no. 48, pp. 48–65, 2010.

[148] M. M. Ghassemi *et al.*, "A data-driven approach to optimized medication dosing: A focus on heparin," *Intensive Care Med.*, vol. 40, no. 9, pp. 1332–1339, 2014.

[149] M. M. Ghassemi, M. B. Westover, R. G. Badawi, O. Mark, and S. Nemati, "Personalized medication dosing via sequential regression: A focus on heparin," *Amer. J. Respirat. Crit. Care*, 2015.

[150] S. Nemati and R. Adams, "Identifying outcome-discriminative dynamics in multivariate physiological cohort time series," in *Advanced State Space Methods for Neural and Clinical Data*. Cambridge, U.K.: Cambridge Univ. Press, 2015, p. 283.

[151] S. K. Nachimuthu, A. Wong, and P. J. Haug, "Modeling glucose homeostasis and insulin dosing in an intensive care unit using dynamic Bayesian networks," in *Proc. AMIA Annu. Symp.*, 2010, vol. 2010, p. 532.

[152] A. Smith, M. West, K. Gordon, M. Knapp, and I. Trimble, "Monitoring kidney transplant patients," *The Statistician*, vol. 32, pp. 46–54, 1983.

[153] U. Gather, M. Imhoff, and R. Fried, "Graphical models for multivariate time series from intensive care monitoring," *Stat. Med.*, vol. 21, no. 18, pp. 2685–2701, 2002.

[154] C. Williams, J. Quinn, and N. Mcintosh, "Factorial switching Kalman filters for condition monitoring in neonatal intensive care," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1513–1520.

[155] M. van der Heijden, M. Velikova, and P. J. Lucas, "Learning Bayesian networks for clinical time series analysis," *J. Biomed. Inf.*, vol. 48, pp. 94–105, 2014.

[156] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM SIGKDD*, vol. 12, no. 1, pp. 40–48, 2010.

[157] T. W. Liao, "Clustering of time series data—A survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, 2005.

[158] J. D. Rennie and T. Jaakkola, "Using term informativeness for named entity detection," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2005, pp. 353–360.

[159] L. Lehman, T. Kyaw, G. Clifford, and R. Mark, "A temporal search engine for a massive multi-parameter clinical information database," in *Proc. Comput. Cardiol.*, 2007, pp. 637–640.

[160] L. Lehman, M. Saeed, G. Moody, and R. Mark, "Similarity-based searching in multi-parameter time series databases," in *Proc. Comput. Cardiol.*, 2008, pp. 653–656.

[161] D. M. Sow *et al.*, "Real-time analysis for short-term prognosis in intensive care," *IBM J. Res. Develop.*, vol. 56, no. 5, pp. 3:1–3:10, 2012.

[162] S. Nemati and M. M. Ghassemi, "A fast and memory-efficient algorithm for learning and retrieval of phenotypic dynamics in multivariate cohort time series," in *Proc. IEEE Int. Conf. Big Data*, 2014, pp. 41–44.

[163] R. K. Avent and J. D. Charlton, "A critical review of trend-detection methodologies for biomedical monitoring systems," *Crit. Rev. Biomed. Eng.*, vol. 17, no. 6, pp. 621–659, 1990.

[164] M. Saeed and R. Mark, "A novel method for the efficient retrieval of similar multiparameter physiologic time series using wavelet-based symbolic representations," in *Proc. AMIA Annu. Symp.*, 2006, pp. 679–683.

[165] M. Hauskrecht *et al.*, "Outlier detection for patient monitoring and alerting," *J. Biomed. Inf.*, vol. 46, no. 1, pp. 47–55, 2013.

[166] S. Saria *et al.*, "Learning individual and population level traits from clinical temporal data," in *Proc. NIPS, Predictive Models in Personalized Medicine Workshop*, 2010, DOI: 10.1.1.232.390.

[167] P. Schulam, F. Wigley, and S. Saria, "Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery," 2015.

[168] J. Ross and J. Dy, "Nonparametric mixture of Gaussian processes with constraints," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1346–1354.

[169] L. H. Lehman *et al.*, "A physiological time series dynamics-based approach to patient monitoring and outcome prediction," *IEEE J. Biomed. Health Inf.*, vol. 19, no. 3, pp. 1068–1076, 2015.

[170] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *Proc. 2nd ACM SIGHIT Int. Health Inf. Symp.*, 2012, pp. 389–398.

[171] S. Nemati and R. Adams, "Supervised learning in dynamic Bayesian networks," Tech. Rep., 2014.

[172] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[173] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cogn. Model.*, vol. 323, pp. 533–536, 1988, DOI: 10.1038/323533a0.

[174] E. Phadia, *Prior Processes and Their Applications: Nonparametric Bayesian Estimation*. New York, NY, USA: Springer-Verlag, 2013.

[175] J. Shi and T. Choi, *Gaussian Process Regression Analysis for Functional Data*. London, U.K.: Chapman & Hall, 2011.

[176] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 193–197, 2013.

[177] C. Orphanidou *et al.*, "Signal quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring," *IEEE J. Biomed. Health Inf.*, vol. 19, no. 3, pp. 832–838, 2015.

[178] D. A. Clifton, L. Clifton, S. Hugueny, D. Wong, and L. Tarassenko, "An extreme function theory for novelty detection," *IEEE J. Sel. Top. Signal Process.*, vol. 7, no. 1, pp. 28–37, 2013.

[179] L. Clifton, D. Clifton, and M. Pimentel, "Gaussian processes for personalised e-health monitoring with wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 193–197, 2013.

[180] D. Clifton *et al.*, "Pinning the tail on the distribution: A multivariate extension to the generalised Pareto distribution," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2011, pp. 1–6.

[181] L. Clifton *et al.*, "Probabilistic novelty detection with support vector machines," *IEEE Trans. Reliab.*, vol. 63, no. 2, pp. 455–467, 2014.

[182] D. Clifton, D. Wong, L. Clifton, R. Pullinger, and L. Tarassenko, "A large-scale clinical validation of an integrated monitoring system in the emergency department," *IEEE Trans. Inf. Technol. Biomed.*, vol. 17, no. 4, pp. 835–877, 2013.

[183] R. Duerichen *et al.*, "Multitask Gaussian processes for multivariate physiological time-series analysis," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 314–322, 2015.

[184] M. A. Pimentel, D. A. Clifton, and L. Tarassenko, "Gaussian process clustering for the functional characterisation of vital-sign trajectories," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2013, pp. 1–6.

[185] J. Behar, A. E. Johnson, J. Oster, and G. Clifford, "An echo state neural network for foetal ECG extraction optimised by random search," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013.

[186] M. Ghassemi, L. H. Lehman, J. Snoek, and S. Nemati, "Global optimization approaches for parameter tuning in biomedical signal processing: A focus of multi-scale entropy," Tech. Rep., 2014.

[187] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2951–2959.

[188] J. Fan, K. Mei, J. Peng, N. Zheng, and L. Gao, "Hierarchical classification of

large-scale patient records for automatic treatment stratification," 2015.

[189] D. C. Kale *et al.*, "An examination of multivariate time series hashing with applications to health care," in *Proc. IEEE Int. Conf. Data Mining*, 2014, pp. 260–269.

[190] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, irregular clinical data," *PloS One*, vol. 8, no. 6, 2013, Art. ID e66341.

[191] J. Wiens, J. Guttag, and E. Horvitz, "Patient risk stratification for hospital-associated c. diff as a time-series classification task," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012.

[192] T. F. Liu and R. W. Shafer, "Web resources for HIV type 1 genotypic-resistance test interpretation," *Clin. Infectious Diseases*, vol. 42, no. 11, pp. 1608–1618, 2006.

[193] N. Beerenwinkel *et al.*, "Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3850–3855, 2003.

[194] M. C. Prosperi *et al.*, "Investigation of expert rule bases, logistic regression, non-linear machine learning techniques for predicting response to antiretroviral treatment," *Antivir. Ther.*, vol. 14, no. 3, pp. 433–442, 2009.

[195] M. Laabei *et al.*, "Predicting the virulence of MRSA from its genome sequence," *Genome Res.*, vol. 24, no. 5, pp. 839–849, 2014, DOI: 10.1101/gr.165415.113.

[196] L. Rishishwar, R. A. Petit, C. S. Kraft, and I. K. Jordan, "Genome sequence-based discriminator for vancomycin-intermediate staphylococcus aureus," *J. Bacteriol.*, vol. 196, no. 5, pp. 940–948, 2014.

[197] A. Drouin *et al.*, "Learning interpretable models of phenotypes from whole genome sequences with the set covering machine," in *Proc. Neural Inf. Process. Syst. Comput. Biol. Workshop*, 2014.

[198] K. E. Niehaus *et al.*, "Machine learning for the prediction of antibacterial susceptibility in mycobacterium tuberculosis," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Inf.*, 2014, pp. 618–621.

[199] L.-W. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark, "Risk stratification of ICU patients using topic models inferred from unstructured progress notes," in *Proc. AMIA Annu. Symp. Proc.*, 2012, vol. 2012, p. 505.

[200] M. Ghassemi *et al.*, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data," 2015.

[201] M. Ghassemi, R. Mark, and S. Nemati, "A visualization of evolving clinical sentiment using vector representations of clinical notes," Tech. Rep., 2015.

[202] S. A. Collins *et al.*, "Relationship between nursing documentation and patients' mortality," *Amer. J. Crit. Care*, vol. 22, no. 4, pp. 306–313, 2013.

[203] K. D. Mandl and I. S. Kohane, "Escaping the EHR trap—The future of health IT," *New England J. Med.*, vol. 366, no. 24, pp. 2240–2242, 2012.

[204] L. A. Celi, R. G. Mark, D. J. Stone, and R. A. Montgomery, "'Big data' in the intensive care unit. Closing the data loop," *Amer. J. Respirat. Crit. Care Med.*, vol. 187, no. 11, pp. 1157–1160, 2013.

[205] A. Holzinger and I. Jurisica, "Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions," in *Proc. Interactive Knowl. Disc. Data Mining Biomed. Inf.*, 2014, pp. 1–18.

[206] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[207] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, 2001.

[208] N. Black, "High-quality clinical databases: Breaking down barriers," *Lancet*, vol. 353, no. 9160, pp. 1205–1206, 1999.

ABOUT THE AUTHORS

Alistair E. W. Johnson, photograph and biography not available at the time of publication.

Mohammad M. Ghassemi, photograph and biography not available at the time of publication.

Shamim Nemati, photograph and biography not available at the time of publication.

Katherine E. Niehaus, photograph and biography not available at the time of publication.

David A. Clifton, photograph and biography not available at the time of publication.

Gari D. Clifford (Senior Member, IEEE), photograph and biography not available at the time of publication.