# Machine Learning Models for Multidimensional Clinical Data

Christina Orphanidou and David Wong

## Abstract

Healthcare monitoring systems in the hospital and at home generate large quantities of rich-phenotype data from a wide array of sources. Typical sources include clinical observations, continuous waveforms, lab results, medical images and text notes. The key clinical challenge is to interpret these in a way that helps to improve the standard of patient care. However, the size and complexity of the data sets, which are often multidimensional and dynamically changing, means that interpretation is extremely difficult, even for expert clinicians.

One important set of approaches to this challenge is Machine Learning Systems. These are systems that analyse and interpret data in a way that automatically recognizes underlying patterns and trends. These patterns are useful for predicting future clinical events such as hospital re-admission, and for determining rules within clinical decision support tools.

In this chapter we will provide a review of machine learning models currently used for event prediction and decision support in healthcare monitoring. In particular, we highlight how these approaches deal with multi-dimensional data. We then discuss some of the practical problems in implementing Machine Learning Systems. These include: missing or corrupted data, incorporation of heterogeneous and multimodal data, and generalization across patient populations and clinical settings. Finally, we discuss promising future research directions, including the most recent developments in Deep Learning.

## 1. Introduction

Advances in the development of smart sensors and intelligent communication systems combined with the proliferation of smart devices and access to cheaper and more effective power and storage mechanisms have led to an explosion in healthcare data in the past few years. In 2012 healthcare data worldwide amounted to approximately 500 petabytes and by 2020 the amount is projected to be 25000 petabytes [1]. The large quantities of data, generated in hospital and at home, present the opportunity to develop data-driven approaches for delivering best practice and improving patient outcomes. The key clinical challenge is to interpret the available data in order to provide better and faster decision-making and thus improve the standard of patient care and, consequently, patient health outcomes. Data-driven systems being developed aim to provide disease diagnosis, offer online patient tracking, identify physiological deterioration, provide risk assessments, as well as predict the occurrence of severe abnormalities such that suitable interventions can be put in place in a timely manner.

One important set of approaches to this challenge is Machine Learning Systems. These are computer algorithms that analyse and interpret data in a way that automatically recognizes underlying patterns and trends. Compared to more traditional statistics-based approaches where prior information about the process to be modelled is required, machine learning favours a black box approach: the relationship between different variables does not need to be fully understood. For instance in disease diagnosis systems, the underlying labelling processes are not particularly important; the system just needs to learn how to replicate them. While this black-box approach does not provide any knowledge into the way the different parameters are associated with outcomes, it is particularly suitable for healthcare monitoring applications where the available information to be processed is very complex. Variables to be combined are often present in a plethora of different formats, such as lab-results, clinical observations, imaging scans, continuous waveforms and more, and the associations between the different variables are not always clearly understood. The human expert, the gold standard of clinical decision-making gains clinical acumen in large part through experience. The basic principle of Machine Learning is not far-off: for a computer to be taught how to perform a task we need to provide it with enough examples of how it should be done. As more information is added to the system, the "experience" grows and the decision making is improved.

The potential of machine learning clinical applications is enormous. In complex medical cases, the inclusion of aggregate data may reveal new information that is not seen by the individual. Machine learning systems additionally offer the possibility of dynamic, online monitoring at home and the hospital and are particularly useful in situations where real-world constraints may restrict the number of clinical staff attending to the patients. Moreover, the ability of machine learning models to analyse massive amounts of constantly refreshed, diverse information in real-time, via Big Data/Deep Learning approaches, offers the potential of quick and effective decision-making at a decreased cost. Additionally, machine learning can provide input which is similar to that of a truly independent expert since it circumvents the confirmation bias of the clinical expert [2]. However, the size and complexity of the data sets, which are often multidimensional and dynamically changing, means that interpretation is extremely difficult, even for expert clinicians. Prediction accuracy depends on the amount of data available to build the system's "experience". Additionally, because the researcher is searching for patterns without knowing what may emerge, findings need to be validated using stringent methods, in order to ensure that they are not occurring by chance.

In this chapter we will introduce the principles of machine learning and review models currently used for event prediction and decision support in healthcare monitoring. In particular, we highlight how these approaches deal with multi-dimensional and heterogeneous data. We then discuss some of the practical problems in implementing Machine Learning Systems. These include: how to process missing or corrupted data and how to process heterogeneous and multimodal data. Finally, we discuss promising future research directions, including the most recent developments in Deep Learning.

## 2. Machine Learning Models

Machine learning models are computer programs that can "learn" important features of a dataset (the *training* set) such that the user can make predictions about other data which were not part of the training set (the *test* set). Applications arising from these models include classifiers which can separate datasets into two or more classes based on attributes measured in each dataset [3] or regression models which can estimate continuous variables. In the context of clinical applications, classifiers have been proposed for disease diagnosis (computer-aided diagnosis-CAD), event prediction, forecasting of patient outcomes, even to predict hospital mortality. Regression models, on the other hand, have been proposed for estimating risk scores and for estimating disease stage and predicting clinical progression.

A machine learning model considers a large set of N D-dimensional *feature vectors* $\{X_1, \ldots, X_N\}$, $X \in R^D$, called the *training* set which is used in order to tune the parameters of an adaptive model. In order to build a machine learning model, for each one of the feature vectors, we need to have a corresponding *target* value $\{Y_1, \ldots, Y_N\}$, $Y \in R^K$. When building a binary classifier, $Y \in \{0,1\}$ (the *label)*, while in the case of building a regression model $Y$ may be multidimensional, takes a continuous value from a usually predefined range. The goal of building a machine learning model is to build a rule which can predict $Y$ given $X$, using only the data at hand. Such a rule is a function $h: X \rightarrow Y$ which is essentially the *machine.* The exact form of the function $h(X)$ is determined during the *training* phase (sometimes also referred to as the *learning* phase), using the training data: this type of learning is called *supervised*. Once the model is trained, it can then be used to determine Y for new values of X, not used in the training set, i.e., the *test* set. The ability to predict Y correctly from new values of X is known as *generalization* and it is a central goal in machine learning and pattern recognition [4].

It is also possible to build a machine learning model based only on the input vectors X, without any corresponding target values. These type of *unsupervised* learning approaches aim to discover groups of similar attributes within the dataset (*clustering*), to determine the distribution of data within the input space (*density estimation*) or to reduce the dimensionality of the input space for the purpose of *visualization* [4].

The steps involved in building a machine learning algorithm are:

- Choosing the analysis model
- Choosing the attributes of the dataset that will comprise the features of the system
- Training the model
- Validating the model on the test data

In the following sections we will review current approaches for addressing every step of the process and discuss considerations related to clinical applications.

## 2.1 Model Selection

The first applications of machine learning in biomedicine were based on Artificial Neural Networks (ANN) and the promise of building systems modelled after the structure and functioning of the brain [5]. The systems showed a lot of promise and led to many applications in biomedical image and signal analysis. However, the complexity and lack of understanding about how the different components of the systems are connected made it difficult to interpret the outputs in a clinical context. Further effort was then made to create linear models which for pattern recognition in biomedicine, which have the advantage of being easier to analyse and interpret in a meaningful way and are also computationally efficient. An example of a linear model, which has been used extensively in biomedical applications, is Support Vector Machines (SVM) which combines the simplicity of a linear process for separating high-dimensional feature data with the sometimes necessary complexity of non-linear modelling of the input data in order to obtain the high-dimensional feature space [6]. It is often the case, however, that clinical data contain high amounts of noise (for example physiological signals obtained via wearable sensors). To address this uncertainty in the data and at the same time to incorporate prior knowledge into the decision-making process, methods based on Bayesian inference have been introduced and have made significant impact in the detection and assessment of disease in biomedicine [5]. In the next sections we will describe the methodologies employed in ANNs, SVMs and Bayesian Networks and discuss the respective advantages and disadvantages of each technique when applied to the analysis of multidimensional clinical data.

### 2.1.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are mathematical models which attempt to simulate the structure and functionality of the brain. The building blocks of such networks are mathematical functions which are interconnected using some basic rules. The parameters of each building block are learnt during the process of training. ANNs have shown great potential in approximating arbitrary functions, however, in some practical applications it was found that the brain-like structure could sometimes impose entirely unnecessary constraints [4]. *Feed-forward neural networks*, e.g., the multilayer perceptron, have shown to be of greatest practical value and have been widely applied to biomedical clinical data analysis. Figure 1 shows a general structure of a feed-forward neural network. In a feed-forward neural network the route from the multidimensional input space to the multidimensional output space involves a series of functional transformations via the so-called *hidden layers*. The first step is to construct M linear combinations of the input variables $x_1, x_{2,} \dots \dots, x_D$ in the form

$$a_j = \sum_{i=1}^{D} w_{ij}^{(1)} x_i + w_{j0}^{(1)} \tag{1}$$

Where $j = 1, \dots \dots, M$ and $M$ is the number of hidden units and the superscript (1) indicates which layer the parameters $w_{ij}^{(1)}$ (the *weights*) and $w_{j0}^{(1)}$ (the *biases*-nodes allowing for any fixed offset on the data) originate from [4]. The outputs $a_j$, known as activations are then transformed using nonlinear activation functions $h(\cdot)$ to give

$$z_j = h(a_j) \tag{2}$$

which are the *hidden units* of the network. The same procedure is then repeated in order to produce the output unit activations:

$$a_k = \sum_{j=1}^{D} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{3}$$

where $k = 1, \dots \dots, K$ and $K$ is the number of outputs. The output unit activations are then transformed using another activation function to give the outputs, $y_k$.

$$y_k = l(a_k) \tag{4}$$

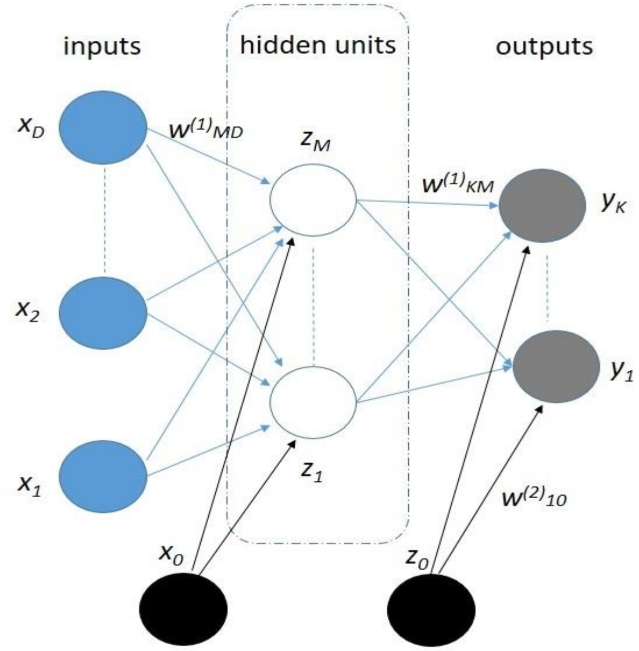


*Figure 1. Feed-forward neural network* (adapted from [4]).

The choice of activation functions in all layers of the network is usually determined by the type of application and data characteristics. Sigmoidal functions are often used (logistic sigmoid or the *tanh* function), especially for binary classification problems [4].

### 2.1.2 Support Vector Machines

In its most common formulation, the Support Vector Machines approach considers N-dimensional patterns $\boldsymbol{x_i}$ and class labels $y_i$ which are trained in order to estimate a function $f: R^N \rightarrow \{\pm 1\}$ such that $f$ will correctly classify new examples $(\boldsymbol{x}, y)$, that is $f(\boldsymbol{x}) = y$ for examples $(\boldsymbol{x}, y)$, which were generated from the same underlying probability distribution $P(\boldsymbol{x}, y)$ as the training data [6]. The SVM classifier is based on the class of hyperplanes

$$(w \cdot \boldsymbol{x}) + b = 0, w \in R^N, b \in R\ , \tag{5}$$

where the decision function is given by

$$f(x) = sign((w \cdot \boldsymbol{x}) + b) \tag{6}$$

The optimal hyper-plane, defined as the one with the maximal margin of separation between the two classes, can be uniquely constructed by solving a constrained optimization problem whose solution **w** has an expansion $w = \sum_i v_i \boldsymbol{x_i}$ in terms of training patterns that lie on the margin, the so-called *support vectors*. Because equations (4) and (5) depend only on dot products between patterns, it is possible to map the training data nonlinearly into a higher-dimensional feature space $F$, via a map Φ, and construct the optimal separating hyper-plane in $F$. This is accomplished by substituting $\Phi(\boldsymbol{x_i})$ for each pattern $\boldsymbol{x_i}$ and simple *kernels k*, such that

$$k(\boldsymbol{x}, \boldsymbol{x_i}) := ((\Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x_i})) \tag{7}$$

The decision boundary then becomes

$$f(x) = sign(\textstyle\sum_{i=1}^{l} v_i \cdot k(\boldsymbol{x}, \boldsymbol{x_i}) + b)\ , \tag{8}$$

where the parameters $v_i$ are computed as the solution of a quadratic programming problem.

In input space the hyper-plane basically corresponds to a nonlinear decision function whose form is determined by the type of kernel used [6] (see Fig. 2). Depending on the application at hand, different kernels may be used. Commonly used are the radial basis function (RBF) given by

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2\ )/2\sigma^2\ , \tag{9}$$

where σ is a scaling factor [6] and the polynomial given by

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y})^{\mathrm{d}}\ , \tag{10}$$

where d is the order of the polynomial.

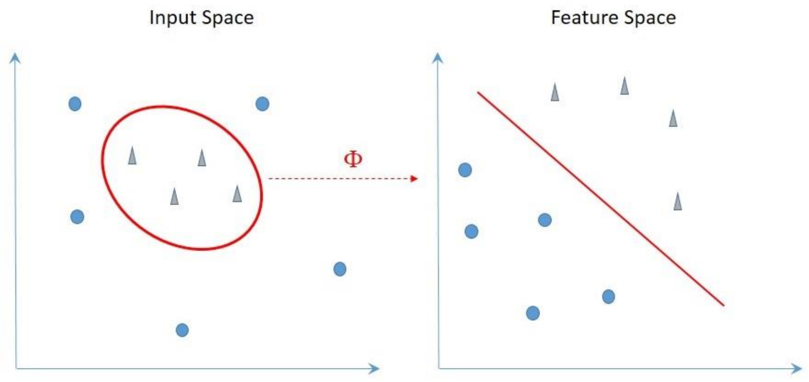


*Figure 2. The kernel trick in Support Vector Machines formulation (adapted from [6]).*

### 2.1.3 Bayesian Networks

Clinical datasets are often noisy and incomplete, making the building of machine learning models challenge. A way of dealing with noisy observations is to consider measured variables as latent states from which noisy observations are made [7]. Because many patient aspects are not directly measurable, state-space approaches have been extensively considered for obtaining reliable estimates of physiological states under uncertain conditions. Kalman Filters (KFs), are a good choice for dealing with noisy data since they treat measurements as noisy observations of an underlying state and update the state only if high confidence in the current state is high, conditioned on the previous observation [7]. Noisy observations are then naturally rejected and not taken into account in the calculation of the state. Bayesian approaches are also able to meet these challenges by incorporating uncertainty into the decision-making process. In the recent years, Bayesian methods have experienced a huge popularity for the development of biomedical applications and have shown promising performance in modelling a range of problems relevant to biological learning [5]. In classification problems, for instance, Bayesian approaches consider class conditional distributions, $P(D/C)$, (where D is the data and C is the class) which can be trained for each different class. The conditional probability of each class given data D can then be calculated using Baye's rule to obtain

$$P(C/D) = \frac{P(D/C)P(C)}{P(D)} \tag{11}$$

Classification of novel examples can then be performed by computing the likelihood over each model.

Bayesian networks are graphical models where each node represents a random variable and each link represents the dependencies between the linked variables. Along with the graphical structure of the network, a joint probability distribution $Pr$ is learnt using the training data; for each random variable $V_i$ represented by a node, a set of conditional probability distributions is determined connecting it to all the nodes it follows (sometimes referred to as the *parent* nodes and symbolized by $\pi(\cdot)$), $\Pr(V_i/\pi(V_i))$. These sets of conditional probability distributions with each other define a unique joint probability distribution that factorises over the graphs structure as [8]:

$$\Pr(V_1, \ldots., V_n) = \prod_{i=1}^{n} \Pr(V_i/\pi(V_i)) \tag{12}$$

A restriction in the graphical models defined is that there can be no directed cycles, i.e., that the structure of the graph does not permit for a path which starts and ends at the same node, for this reason such graphs are also called *acyclic graphs* [4].

An advantage of Bayesian Networks compared to other approaches such as SVMs and ANNs is that they allow for interpretation of the interactions between different variables which makes them easier to combine with findings from clinicians.

## 2.2 Feature extraction and selection

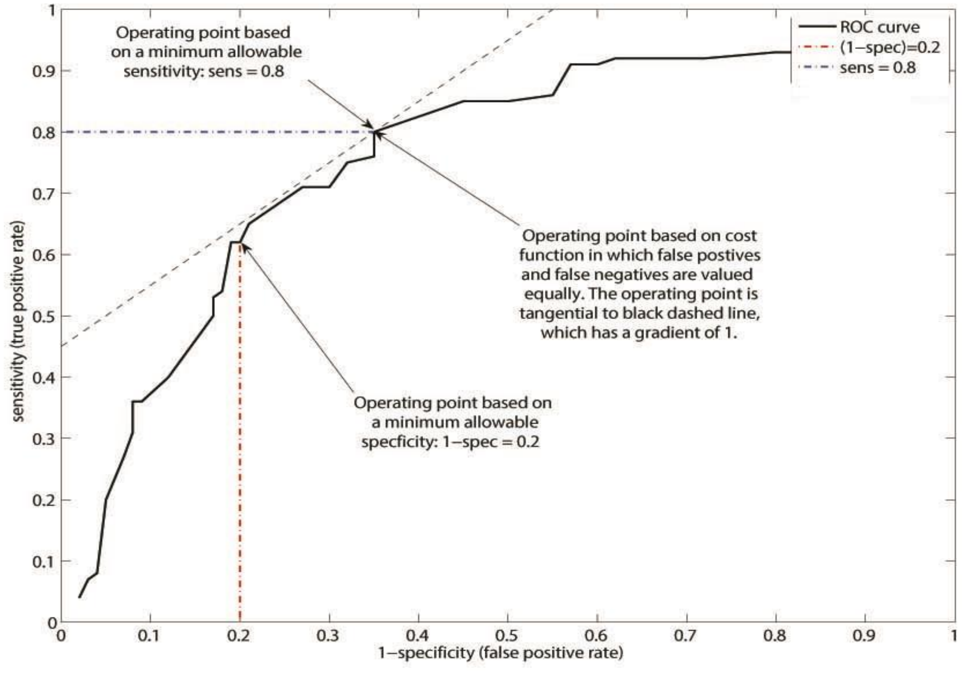
The number and choice of features is critical to the success of a machine learning model. Using too many features relative to the number of "events" may results in *overfitting*, a result of the classifier learning the training data instead of the underlying trends of the data [3]. This is a common problem in healthcare monitoring applications where usually the number of training instances available is small. Using a large number of features also requires a large training dataset in order to reliably estimate the relationships between the multidimensional variables, a phenomenon known as *the curse of dimensionality*. While there is no widely accepted rule for the ratio of features to "events", as a rule of thumb, at least ten events are needed per feature to achieve a reasonable predictive performance [9]. The choice of features is another crucial issue. Depending on the application of the machine learning model, the features selected are usually picked such that they have some bearing on the associated physiological process. These features would be the ones a clinician would review in order to assess the physical state of the patient. For example when building a system which may predict exacerbations in patients with Traumatic Brain Injury (TBI), intracranial pressure (ICP) should be included since it is the most important identifier of an exacerbation. For people with chronic cardiorespiratory problems elevations in heart rate (HR) and respiration rate (RR) and a drop in oxygen saturation (% $SpO_2$) are the most important precursors of an exacerbation. As a result HR, RR and % $SpO_2$ are obvious choices for features to be used in a predictive model for such exacerbations. It is often the case, however, that more abstract characteristics are used as features for a machine learning model. Examples are frequency characteristics of the ECG signal, such as the amount of entropy in different frequency bands, used for the diagnosis of Atrial Fibrillation (AF) [10] or statistical texture features extracted from medical images in order to identify malignant tumours [11]. While these abstract characteristics cannot be directly linked with assessments a clinician would make for making a decision, they often reveal strong links with medically relevant information.

## 2.3 Training, testing and evaluation metrics

The data required for training and testing machine learning models are most often collected via clinical trials. The protocols of these clinical trials need to reflect the requirements of the algorithm to be designed. Collected data naturally need to correspond to the variables defined in the model. Enough data need to be collected, guided by the dimension of the feature space, so that the relationships between the different variables can be reliably derived. Target values need to be carefully defined. For example, when designing systems to identify "events", such as physiological exacerbations, the training data need to be chosen as to contain enough clearly marked occurrences of such events. This is often a challenge when building systems for diagnosing rare events and diseases.

In classification problems, human labelling, the "gold standard" of clinical diagnosis, is usually done by clinicians. Human labelling, however, suffers from inconsistencies, known as intra- and inter-rater variability. To alleviate this, often, multiple raters are used and only data with consistent labels across raters are used for building the classifiers.

Performance evaluation of machine learning algorithms is usually assessed based on predictive accuracy compared to the “gold standard”. Sensitivity, Specificity, and Accuracy are the metrics most often used and in many cases a trade-off between true positive and true negative rate needs to be defined since the “cost” of each different type of error varies depending on the application. Sensitivity, given by TP/(TP+FN) where TP are true positives and FN are false negatives, measures the proportion of poor quality signals that have been correctly identified as such. Specificity, given by TN/(TN+FP) where TN are true negatives and FP are false positives, measures the proportion of good quality signals that have been correctly identified as acceptable. Accuracy corresponds to the proportion of signals that have correctly been classified. Receiver Operating Characteristic (ROC) curves serve as a graphical representation of the trade-offs between the Sensitivities and Specificities of each model specification. While accuracy could be used as a metric for evaluating the performance of the system, the “cost” of a false positive (i.e., a signal identified as acceptable which is actually unacceptable) may be higher in practice than that of a false negative (i.e., an acceptable signal identified as unacceptable). The former would result in a false alert whereas the result of the latter would result in the rejection of signal which could actually have been used. In such situations, decision functions may be defined where the relative cost of each error is weighted or thresholds may be set in the minimum acceptable value of each metric, such that the best model for every application may be chosen.



Operating point based on a minimum allowable sensitivity: sens = 0.8

Operating point based on cost function in which false postives and false negatives are valued equally. The operating point is tangential to black dashed line, which has a gradient of 1.

Operating point based on a minimum allowable specficity: 1−spec = 0.2

*Figure 3. Example ROC curve indicating operating point selection based on different criteria.*

**Cross-Validation**

Cross-validation is a method often used in order to evaluate machine learning models which allows all data points to be used in both the training and testing phases of the model evaluation procedure. In K-fold cross-validation all available data are firstly divided randomly into K different equally sized groups. At each iteration, one group is treated as the test set and the remaining K-1 groups are used as the training set. A model is then trained K times, each time using one of the training sets and the associated test set and the overall accuracy is measured as the average of the accuracy measures of over the K iterations [4].

# 3. Challenges related to clinical applications

For real-life clinical problems, analysis of data is not straightforward, and cannot simply be reduced to the application of standard machine learning algorithms. Whilst data quality and analysis issues are not unique to healthcare, the acquisition of data from human patients brings additional challenges that do not occur in other fields. Two prominent challenges are the analysis of datasets with missing or corrupted data, and the analysis of heterogeneous and high-dimensional data.

## 3.1 Corrupted and Missing Data

Corrupted data occurs when a recorded measurement does not accurately reflect the true state of the object or person being measured. Missing data occurs when there is no recorded measurement at a given point in time. If left unrecognised, analysis that does not take into account missing or corrupted data may to lead to inaccurate decision-making. In the worst cases, this could mean patients being assigned the wrong course of treatment. In the UK healthcare system alone, data of poor quality has previously been linked to unnecessarily cancelled operations, and undetected outbreaks in C. difficile infections [12].

To understand why health data are often of poor quality, it is helpful to first consider the steps typically involved during data collection. In general, the clinical data from patients involves multiple stages (Fig. 4), each allowing possibility of data corruption. Two of these stages are now discussed in greater detail.

**Patient-Medical Device:** In many common scenarios, the first opportunity for data corruption occurs when patient measurements are recorded using medical equipment. Incorrect use of medical equipment has been associated with erroneous measurements in a wide variety of clinical situations. For instance, Boba et al. [13] found that many breast core needle biopsies produced false-negative results due to sampling from an inappropriate site.

Vital sign monitoring is particularly prone to error caused by patient-device interaction. In the case of pulse oximeters, that measure the level of oxygen in the blood ($SpO_2$), corrupted data are often due to poor attachment of the device to the finger [14]. The relative motion between finger and device leads to physiologically implausible measurements known as motion artefact. Motion artefacts are a common problem for other physiological measurements including heart rate (via ECG) and activity detection (via accelerometers) [15,16].

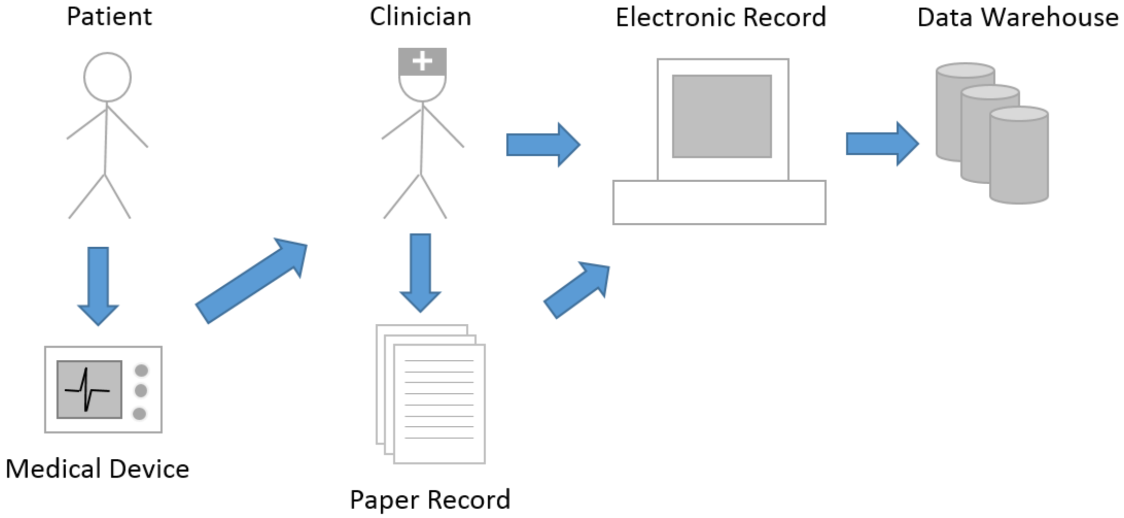


*Figure 4: Data flow during the collection of health data from patients. Multiple stages of data transfer are usually necessary before the final data items are stored securely within a data warehouse.*

The problem of poor device attachment is becoming increasingly important as attempts are made at medium and long-term ambulatory monitoring outside of the hospital environment to reduce pressure on emergency services [17]. In these cases, the state-of-the-art is to apply an adhesive patch to the patient's sternum. Each patch contains a set of integrated sensors that can monitor multiple vital signs concurrently [18,19]. The patients monitored tend to be more physically active than those monitored in-hospital which leads to greater levels of motion artefact. These artefacts are compounded by the practical problem of deterioration of patch adhesion over time.

Even if reliable device attachment can be guaranteed, the accuracy of the medical devices themselves may vary. In many cases, medical instrumentation is subject to regulations that guarantee tolerance (that is, variation from the true value). For instance, the US Food and Drug Administration mandates that all pulse oximeters for medical use have a maximum root mean squared error of <3% over the normal operating range [20]. However, multiple reviews of pulse oximeters have shown much greater variability when tested on healthy individuals [21].

Finally, we note that in some instances, there are multiple clinically-accepted methods for measuring the same data item. This can lead to instances in which variability is due to the measurement method, rather than the patient's true state. Core body temperature may be measure using either oral mercury, oral electronic, or tympanic electronic thermometers [22]. Blood pressure is traditionally measured using a mercury sphygmamometer. During measurement, an inflatable cuff temporarily cuts off blood flow to the arm. As the cuff is deflated, characteristic 'Korotkoff' sounds are used to identify the peak (systolic) and trough (diastolic) of the blood pressure waveform [23]. Modern semi-automatic blood pressure monitors take a different approach. The monitors measure the amplitudes of oscillations in the blood pressure cuff caused by the expansion of the arteries as blood is being forced through. These measurements are then converted into systolic and diastolic blood pressures derived empirically [24]. Pavlik et al. tested semi-automatic against manual methods, showing that the semi-automatic method produced consistently higher blood pressure readings [25]. Similar studies showing discrepancies between the two methods have since been reported for hypertensive patients and home-monitoring devices [26,27]

**Clinical Expert – Paper/Electronic Record:** Having successfully taken a patient's measurements, the next step is for the clinician to interpret and validate the measurement. Error in clinical interpretation can occur through misreading information. For example, the oversight of a decimal point has led to high profile medication errors [28]. In response, national agencies have specific recommendations for numbers, including the avoidance of trailing zeros (to differentiate 1.0 and 10, for example) [29]. Clinical interpretation may also involve combining raw information into aggregate scores. For instance, the overall level of patient severity may be assessed through the use of Early Warning Scores (EWS). These scores assign an integer value to vital sign measurements; the sum of these values forms an EWS that is used to inform level of in-hospital care. These relatively simple calculations have repeatedly been shown to be erroneous approximately 20% of the time [30, 31].

The final clinically-validated data are transcribed to an official clinical record, which may be paper-based or electronic. The process of data transcription is prone to further error. For instance, Callen et al. [32] describe, for an Australian metropolitan hospital, how both handwritten and electronic systems contained clinically significant medication transcription errors in discharge summaries (handwritten: 12.1% and electronic: 13.3% for 13000+ medications). The authors suggest numerous potential factors that may contribute to the level of error, including heavy workload and distractions from the current task. Further causes of transcription error include unintuitive design and lack of training [33,34]

### 3.1.1 Reducing Corrupt Data

The previous section showed the multiple steps required to collect health data, and how each step is prone to data corruption. In instances when data collection is on-going, it is highly desirable to optimise these steps to maximise the reliability of the data for retrospective analysis.

One way to reduce patient-medical device errors is to improve the sensors within the device. Whilst integrated adhesive patches represent the current clinically viable method for recording vital signs, novel techniques are being developed that may reduce the problem of motion artefact. Batchelor et al. tackle motion artefact by using alternative methods of affixation [35]. Their prototype transfer tattoo electrodes provide a strong attachment to the skin and higher durability than traditional electrodes whilst producing similar data reliability (as evidenced through signal-to-noise ratios). Tarassenko et al. [36], as well as multiple others [37,38], take another approach. Rather than ensuring the best possible contact, they attempt to measure vital signs with no patient contact, using video images. Results using the contactless techniques are comparable to traditional monitoring methods for patients at rest, with a mean absolute difference of approximately 3 beats/min for heart rate measurements.

Another way to reduce data corruption is to eliminate the amount of interpretation and transcription of clinical measurements. For vital sign data, incorrect calculations of EWS has been virtually eliminated with the help of electronic data entry at the bedside [39]. These so-called e-Obs systems allow users to type data; the system then automatically calculates the EWS and may provide care recommendations [40,41]. This idea has been extended by medical device manufacturers, who have created integrated vital signs monitors that automatically send the EWS score and vital signs directly to the hospital’s Electronic Patient Record [42].

Automated systems that reduce transcription have become increasingly commonplace in modern healthcare. The rise of Computer Physician Order Entry (CPOE) systems to standardise and automate the ordering of medication has been associated with reductions in nursing transcription errors [43]. Another electronic systems designed to reduce transcription error is the Bloodtrack system for blood transfusions [44]. Bloodtrack electronically allocates compatible red blood cell units by using barcode scanners to identify both the patient and blood packet. The introduction of the system has been associated with improvements in safety checks during blood sample collection, in addition to reduction in time to deliver blood [45,46].

A final way to increase data quality is to identify corrupt data in real-time and encourage human intervention. In the simplest cases, this means preventing a user from entering implausible data. For instance, in the case of CPOE systems, drug-drug interactions data can be used to prevent prescription of potentially dangerous drug combinations [47].

A more sophisticated example is the Phillips Intellivue vital sign monitor. These devices measure multiple types of vital signs continuously and are used to monitor patients at risk of rapid deterioration. These devices generate audible alarms that indicate when the monitor data are unreliable and the vital signs sensors need to be reattached. The technical details for determining periods of unreliable data (typically via Signal Quality Indices) are explained in Section 3.1.2.

Whilst the adoption of processes that reduce corrupt and missing data is helpful for a wide range of healthcare related tasks, such as medical research and hospital management, there are multiple other competing aims within real clinical practice. Most importantly, standards of patient care should not be compromised. Bonnici et al [48] highlight this in the context of wireless sensors, implying that patient choice and comfort is of paramount importance for successful implementation of remote monitoring systems.

Financial cost also needs to be considered. There is a trade-off between using the best (and most costly) equipment, and the level of improvement in staff efficiency and data quality that can be achieved [49]. For some in-hospital electronic solutions, improvements in data quality have been offset by significant increases in time to complete clinical tasks [50].

### 3.1.2 Identifying Corrupt Data

In reality, data corruption cannot be eliminated completely during the measurement and documentation process. Methods are therefore required to process and analyse low quality data. The first step of this process is to correctly identify the corrupt portions of the data set. In particular, it is important to accurately distinguish between an unusual, but true, measurement that may have clinical significance and abnormal data due to artefact. One common way to identify artefactual data is through Signal Quality Indices (SQIs). A SQI is a measure of confidence in the reliability of a data point. Typically, the development of an SQI begins by selecting relevant attributes, or features, that should be present in high quality data. Three popular types of features are:

*Range Checking* compares the data to physiologically plausible ranges. Any measurement outside the range is considered to be low quality. Tat et al. [51] implement range checking to evaluate the quality of an ECG signal. Part of their SQI involves converting the raw signal to a heart rate. Heart rates outside of the range 30-210 are considered bad quality.

*Inference based on simultaneously recorded data* – relies on the fact that measurements are often not independent. For instance, there is strong correlation between continuous blood pressure signal and ECG. Johnson et al. [52] make use of this by developing a SQI for heart rate, in which a high data reliability is estimated when the two signals peak at similar times.

*Comparison to previous values* compares the current data to previously measured values. If the difference between the current and previous values is deemed to be improbable, the measurement is considered to be invalid. Clifton et al. showed one implementation of this approach for tympanic thermometers. By using Bayesian changepoint detection, natural variation in temperature was distinguished from an unexpected step-change in temperature due to calibration error [53].

If the original waveform data are available, more complex features may be used. These may include morphological features (that is, recognition of typical shapes of data), and frequency features generated after the raw signal has been converted into its frequency components using a Fourier transform. For instance, Orphanidou et al. show, for ECG signals, how differences in morphology may be measured using an average cross-correlation between a template and unknown data [54]. A review from [55] indicated that this approach was particularly specific for ECG signal quality in comparison to alternative features.

SQI may be derived from one or more of these features. If multiple features are used, they must be combined in some way to produce a single result. The synthesis of multiple features is often completed using machine learning methods, like the example covered in case study 2 [56].

Setting the threshold between good and bad data quality is itself a challenging problem. In the previous example, the threshold was determined via clinical experts who were asked to label the training data. In many cases however, acceptable signal quality is often task dependent. For instance, in the case of PPG, respiratory rate is often detected using the small amplitude, low-frequency part of the signal – a high threshold on data quality is required. In contrast, heart rate can be computed from portions of the signal that typically have greater amplitude and less prone to random noise. Practically, this means that signal quality indices are highly specific to the clinical setting, a finding supported by Nizami's review of 80 artefact detection methods used in critical care medicine [57].

### 3.1.3 Processing Corrupt and Missing Data

After the identification of corrupt data, there are, broadly, two possible options: to use the corrupted data, or to discard the data.

**Using corrupt data**

Correction of corrupted data is possible when the mechanism by which the data were corrupted is known. For instance, in the case of an ECG signal, the observed signal is often subject to baseline wander caused by respiration, motion, or gradual changes in the ECG electrodes [figure 5]. The baseline wander artefact is known to primarily affect low frequencies, so a high pass filter can be used to eliminate the spurious part of the signal [58].

In the majority of cases, it is not possible to correct artefacts. However, even then, artefact may be non-random and can be used to infer additional useful information. For instance, in the case of EEG signals that measure brain activity, artefactual changes in the signal may be caused by muscle movement as the eyes blink. In many applications, the blink artefact is a nuisance, and there have been many attempts to identify and remove eye-blink artefacts [60,61]. However, for certain applications, knowledge about blinking may be useful. For instance, the clinical standard for determining level of consciousness is partly determined by whether a patient's eyes are open [62]. If one were to attempt to ascertain consciousness level using EEG (as has been attempted by [63,64]), the eye-blink artefact may contain useful information.

Even if artefactual data does not present useful information for the task at hand, a curated and annotated set of known artefacts can be used as exemplars to improve future artefact detection. Lawhern et al. use this approach for EEG signal classification to identify jaw and eye motion artefacts. Sections of EEG signal were first parameterised in an autoregressive model. The model parameters were then were successfully classified using Support Vector Machines, such that real signal was distinguished from multiple types of EEG artefact [65].

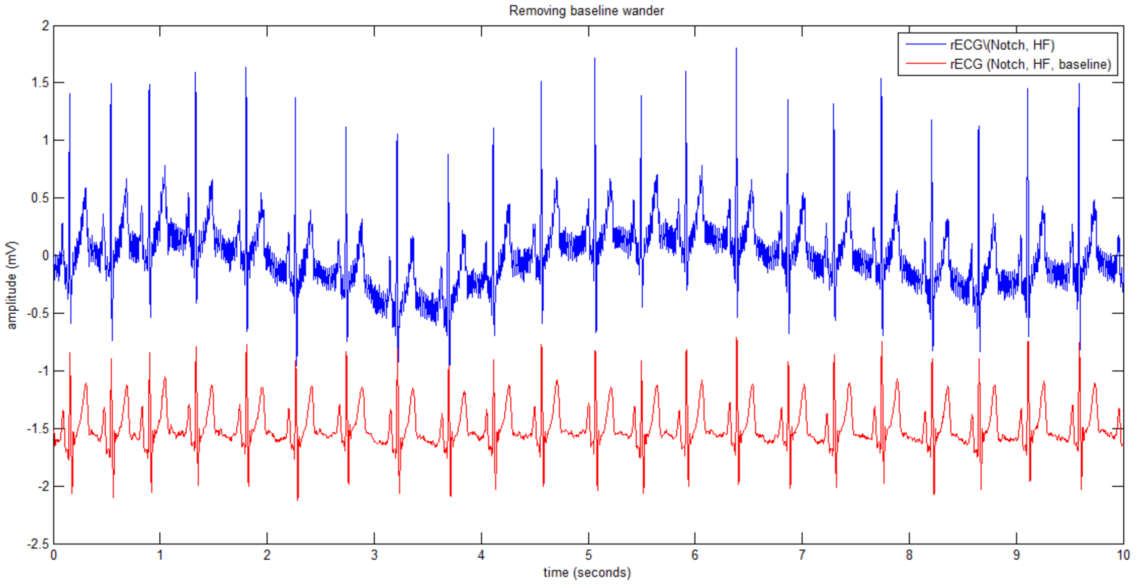


*Fig 5: Baseline wander in an ECG signal. The original signal (in blue) has a low frequency variation associated with respiration rate. By applying a high-pass filter (in red) the baseline wander can be removed to allow easier signal processing [59].*

When data cannot be reasonably corrected, they are often removed from analysis. The way that missing data are handled is of paramount importance; it is easy to inadvertently introduce bias that leads to spurious results. One example of this is an early version of the cardiovascular risk score, QRISK [66]. QRISK (and its successor, QRISK2 [67]) outputs, for a patient, a percentage risk of cardiovascular disease within the next ten years. The output is based on an extensive range of variables including family history, smoking status and age. When data are missing, it attempts to estimate the missing values. However, the effect of this has led to unexpected outcomes, including an implausible null association between cholesterol level and cardiovascular risk [68]. Problems with imputation in QRISK have since been corrected and the algorithm revalidated [69].

Missing data can be categorised in three ways. Data Missing Completely At Random (MCAR) means that data points are missing at random, AND the missing value is independent of any other values in the data set. Data Missing At Random (MAR) means that data points are missing at random, but that the missing data can be partially explained by other variables in the data set. Data Missing Not At Random (MNAR) means that data points are not missing at random, such that the probability of the data being missing is associated with its value. In practice, it is often difficult to distinguish between these three categories, as the missing values are not known. However, knowledge of these missing data mechanisms is useful for recognising potential bias introduced by techniques for analysing incomplete data sets.

### Listwise Deletion

The simplest technique for analysing incomplete data sets is to remove records with missing data, known as listwise deletion. This process ensures that only completed sets of data are used to construct models. Listwise deletion is typically appropriate when only a small percentage of data are missing (e.g. 1%) [70]. Listwise deletion has two significant drawbacks. First, the remaining data set will be biased, with reduced variability in the missing variable, if data are not MCAR. Second, valid information is unnecessarily discarded when there are multiple variables for each data record.

### Data Imputation

Rather than removing missing records, one can attempt to replace the missing data elements with values estimated from the observed data, a process known as data imputation. A simple method for imputation, *mean imputation*, is to replace the missing data with the mean of the all other observations of that variable. This approach has been adopted within clinical software for detection of physiological deterioration from continuous vital signs [71]. In this case, the speed and simplicity of mean imputation was a useful practical method that allowed assessments of deterioration to be conducted in real-time. Mean imputation should be applied with caution, as the variance of the complete data set after imputation will lower than the true value.  To avoid this problem, we may instead randomly sample from the distribution rather than selecting the mean, a process known as *stochastic imputation*. Unbiased sampling from an arbitrary distribution can be achieved using Gibbs sampling, or other Monte Carlo Markov Chain approaches [72].

If data are MAR (not MCAR), the underlying relationship between the observed and missing data can be used to provide a more precise imputation. One such method is *regression mean imputation*. Under this scheme, complete data records are used to regress (typically linear regression) the variable with missing data onto all other variables. The resulting equation is then used to impute the missing data points. Like simple mean regression, this deterministic approach (the unknown value is completely determined by the observed variables) artificially reduces data variability.

The regression approach can be extended to non-continuous variables. For example, logistic regression can be used for categorical variables [73]. In the case of censored data where values outside of a given range are unknown (for instance, due to limited time for study follow-up, or when the dynamic range of a sensor is too small) the Tobit model for truncated regression provides a non-biased estimate [74].

Another simple method for imputation, used in particular for time series data, is ‘last value carried forward’ also known as ‘sample-and-hold’. Under this scheme, missing data are replaced with the last known value. Compared to the mean imputation approaches, sample-and-hold makes the additional assumption that subsequent observations are likely to be more similar than observations taken at random times.

The primary limitation of each of these methods is that the inherent uncertainty of missing data are discarded. More complex approaches, including multiple imputation, maximum-likelihood, and Gaussian process regression, circumvent this by using probability distributions to model the missing data.

Multiple imputation uses stochastic imputation to create multiple possible versions of the missing data record. All versions of the data set are analysed separately, and the outputs are averaged to get an overall result. Standard errors on the output parameters are calculated using Rubin's rules [75]. Rubin's rules take into account the variance of the missing data variable (determined as the variance within the completed records) and the variance of the multiple imputations. The reliability of multiple imputation estimates depends on the number of imputations used. Whilst initial research suggested a small number (3 to 5) imputations, Graham et al. showed that this was insufficient for estimating variance accurately [76]. Instead, they suggest using at least 20 imputations, a number that may increase depending on the overall percentage of missing data [77]. A more detailed tutorial on multiple imputation can be found in [78].

One popular implementation of multiple imputation when more than one variable has missing data is Multiple Implementation by Chained Equations [79]. Under this scheme, all missing data are initialised using mean imputation. Multiple imputation is then used to provide a more accurate estimate for each variable in turn. This process is repeated several times until convergence criteria are met.

An alternative approach is Maximum-Likelihood (ML) estimation. A full description is provided in [80] and is summarised briefly here. ML methods model all of the measured data as a joint probability distribution function, *f*. The distribution function is parameterised by a set of free parameters, $\theta$. For the simple case in which each variable is normally distributed, the joint pdf is a multivariate Gaussian that is fully defined by the mean and covariance.

If each data record, $x_i$ , is independent, then the probability of attaining a given set of n observed data is provided by the likelihood function:

$$L = \prod_{i=1}^{n} f(x_i|\theta) \tag{13}$$

In the case when some of the data records contain missing elements, the probability can be described by marginalising *f* over the missing variables so that the likelihood of single data record, in which the set of M variables are missing, is:

$$\int_M f(x_i|\theta) \tag{14}$$

ML attempts to find the most likely model instance by maximising the likelihood through adjustment of $\theta$. In some instances, the maximum likelihood may be calculated analytically. However, in practice, the likelihood function may be highly non-linear and a closed form solution is not possible. In these cases, the likelihood function is maximised iteratively using methods such as the Expectation-Maximisation algorithm. Because the parameter set, θ, fully describes the variables and the correlations between them, the parameters can then be converted to regression equation parameters if specific instances of imputation are required.

Gaussian process regression extends these principled approaches by taking into consideration temporal relationships. In Gaussian process regression, n data points from a time series are modelled as a single sample from an n-dimensional Gaussian distribution (Figure 6), defined by a covariance matrix. Any missing data are then simply represented by the conditional distribution: *P(missing data | observed data)*, which is also Gaussian.

The elements of the covariance matrix are determined via a covariance function that describes the expected change of the time series through time. In the simplest cases, the covariance functions simply describe how local measurements are highly correlated, and that correlation decreases as data samples become further apart in time. Alternatively, the covariance function can be based on domain-specific knowledge. For instance, Stegle et al. use a covariance function that takes into account periodic circadian rhythm for inferring missing heart rate [81].

Roberts provides a more comprehensive introduction to Gaussian process regression [82]. If multiple time series data are captured simultaneously, correlations between variables can also be modelled. Two similar approaches that model both temporal and inter-variable correlations using Gaussian processes are multi-task GPs and dependent GPs [83,84].

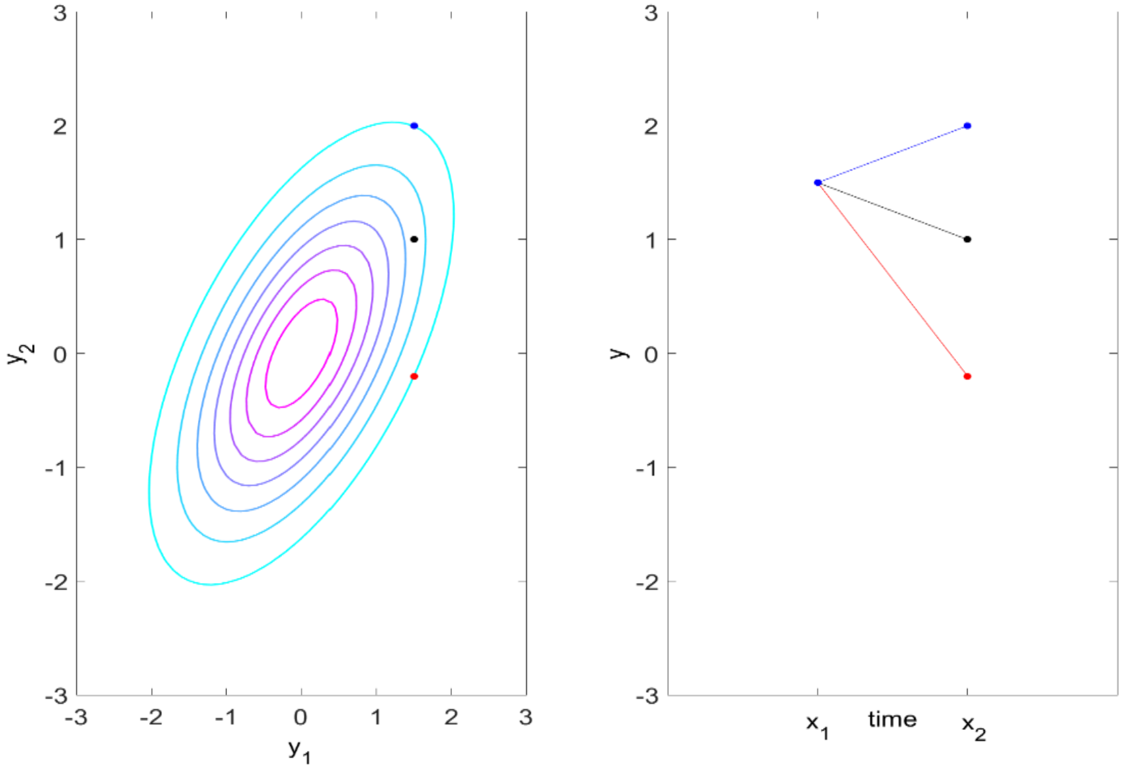


*Figure 6: Simple example of a Gaussian process for two time points. The left figure shows the joint probabilities of all possible pairs of points as a bivariate Gaussian distribution. The right figure shows the time series plots for the three points on the distribution highlighted in blue, black and red.*

Wong et al. show how Gaussian process regression in conjunction with other machine learning methods were used to generate alerts for abnormal vital sign data [85]. Likely distributions of the missing data were first imputed. The distributions were inputs to a model that generated alerts based on the vital sign abnormality, and the level of certainty in the data.

In summary, analysis of missing data remains a complex problem. The optimal choice of data imputation technique depends on the specific problem, and should consider practical problems such as speed of implementation in addition to accuracy of the imputed data. Simple imputation may be used for some cases, but care should be taken to ensure that biases are identified. The fundamental limitation of such approaches is that a single value is used to represent the missing data point – thereby losing information about the uncertainty or 'missingness'. By contrast, stochastic approaches such as multiple imputation, maximum likelihood and Gaussian process regression attempt to model missing data as distributions. Whilst these are more principled methods, they require more complex and time-consuming calculations.

## 3.2 Integrating Heterogeneous Data Sources

Typical care for a patient in a modern health service involves multiple types of data collected from a disparate range of sources. Data analytics that combines multiple sources of data, a process known as data fusion, is useful for two main reasons. First, the multiple data sources may provide *complementary* information that provides a more complete description of the problem. Secondly, the multiple sources may measure the same event, providing redundant information. Whilst the data itself may not provide any new information, the independent sources can be used to corroborate a given measurement and may be useful for assessing data corruption via Signal Quality Indices.

In practice, the process of integrating data sources is fraught with difficulty. We now address three issues in data analysis with multiple sources and address how they are commonly dealt with in practice.

### 3.2.1 Dimensionality and Feature Selection

Combining multiple data sources into a super-set used for analysis increases both the volume and variety of data items. As the number of variables increases, the amount of data required to derive meaningful results increases exponentially – a phenomenon commonly referred to as the Curse of Dimensionality. Figure 7 demonstrates this phenomenon for categorical data. In the example, three data points are represented by red squares. For two variables (a 2D data space), this represents coverage of 3/9 possible states. The addition of a third variable (a 3D data space) means that the same number of data points represents a much smaller proportion of the possible states, 3/27.

One solution to the Curse of Dimensionality is to determine an *optimal* subset of variables, a process known as *feature selection*. If done correctly, only unimportant variables are discarded. Multiple methods for feature selection have been proposed in the literature, and

Saeys et al. provide a detailed discussion of a wide range of feature selection techniques [86].

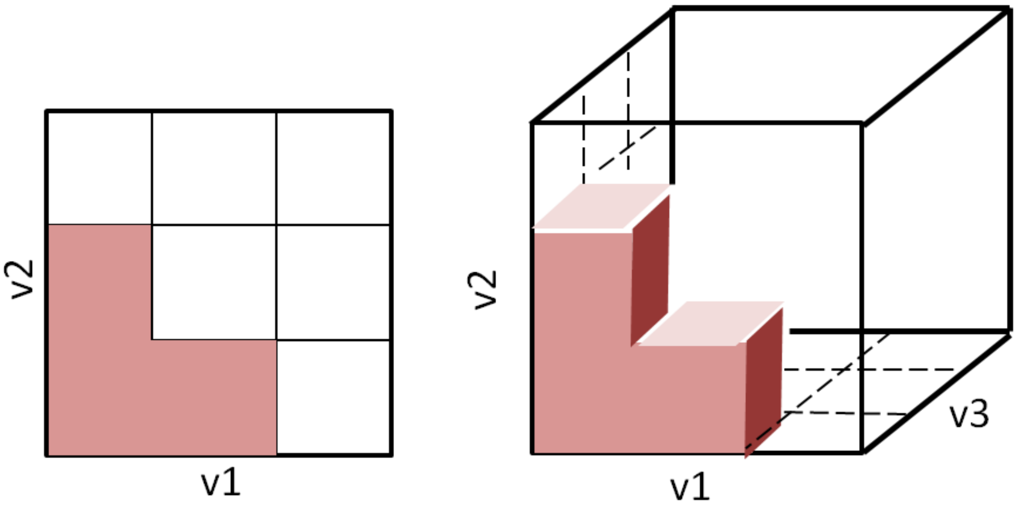


*Figure 7: An example of the Curse of Dimensionality. (a) Variables v1 and v2 can take three possible values each. The three data points (in red) provide examples in 3/9 (33%) of possible combinations. (b) The addition of a new variable, v3, reduces the coverage to 3/27(11%) of possible combinations.*

The most conceptually simple of feature selection techniques is Univariate Feature Selection and is appropriate when the target variable is known (i.e. supervised learning). In this method, each variable is taken in turn to see how it correlates with the target variable. Variables that have poor correlation are discarded. One drawback of this approach is that data redundancy is not considered.

Another intuitive way of selecting features is Backward feature elimination/Forward feature construction. In backward feature elimination, all variables are initially included in a model that tries to explain the target variable. After this, one input is removed from the initial set of n, and the model is re-run. There are n possible variables to remove, leading to n different model results. The model that best describes the target variable is kept, and the associated input variable is discarded. This process is repeated until a pre-determined criterion is met. Forward feature construction uses the same iterative approach, but instead begins with only one input variable, and adds the most useful variable at each iteration.

One model that lends itself well to feature elimination and construction methods is random forests, an extension of the decision tree algorithm [87]. In the case of feature elimination, a random forest model is first applied to all input variables. The variable that is least informative is the first to be removed. The level of information is determined through a scoring function. The function may include information about how many times a feature appears in the individual decision trees, and the classification accuracy of each tree. Specific examples of scoring function are described in [88].

Unlike the univariate feature selection, both feature elimination and construction naturally deal with redundant information. If once an input variable is included, a similar input will only be incorporated if it provides significant additional information. Both approaches are also ‘greedy’ algorithms. This means that they select the best available choice on each iteration. Such approaches only guarantee an optimal solution under specific conditions [89].

The methods described so far find a subset of the original input variables. However, in some cases, input variables may themselves be based on some smaller set of unmeasured, latent variables. The process of recasting the initial input variables into the smaller set of latent variables is known as dimensionality reduction. Each new variable will be a function of the initial inputs, and may have no inherent meaning itself. Mathematically, dimensionality reduction can be considered a transformation of the initial data space into a feature space that can be used to describe most of the variance within the data set. Because these techniques solely rely on properties of the input data, they can be applied without reference to an output target variable.

One common dimensionality reduction techniques is Principal Component Analysis [90]. PCA transforms data sets described by N-input variables to a data set described by M features. The M features are linear combinations of the input variables. They are derived by projecting the dataset onto the eigenvectors corresponding to the M largest eigenvalues. Figure 8 demonstrates PCA for a simple case in which N=M=2 (b), and N=2, M=1 (c). For (b), the PCA output results in a linear rotation of the data so that the data align with the principal axes. In (c), the axis with smaller range (and smaller eigenvalue) is discarded. The resulting dataset is dependent on y_1 only.

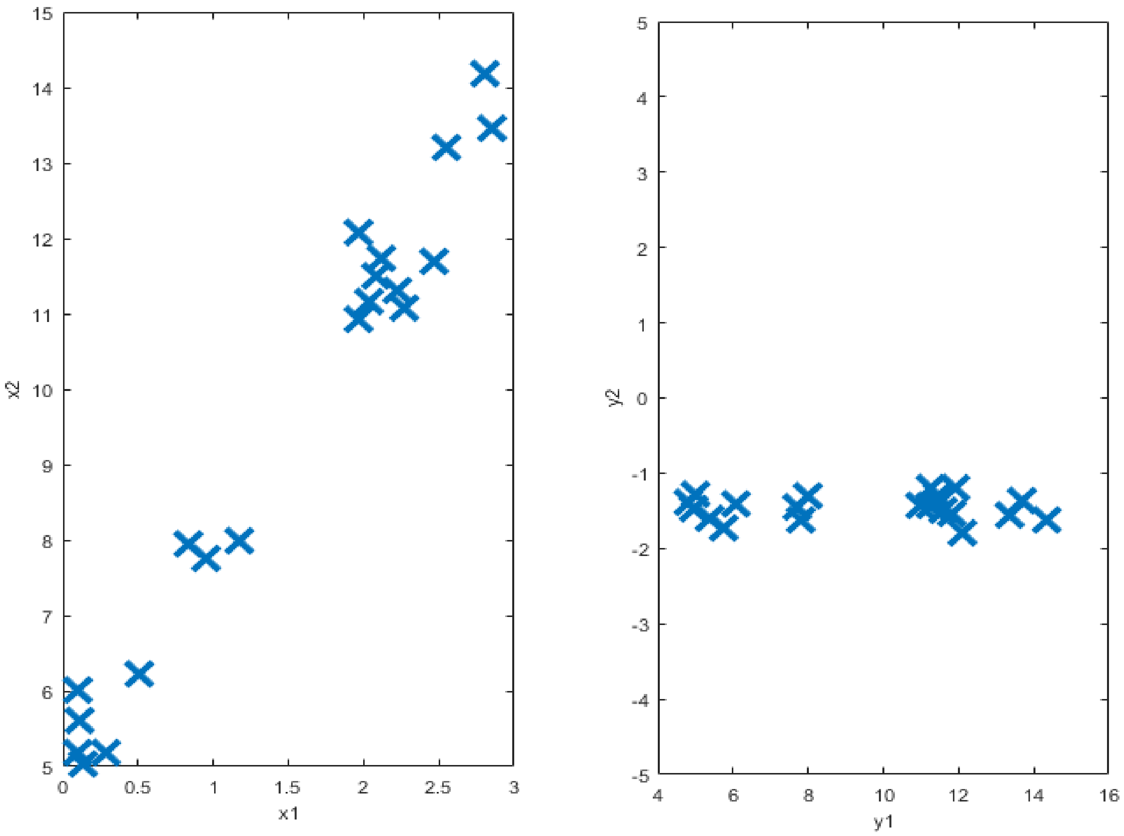


*Figure 8: Demonstration of principle component analysis with two features .Initially, the data are fully described in terms of x1 and x2. PCA linearly transforms the data along the directions with greatest variance. In this case [y1 y2] = [a b; c d][x1 x2]. After the transformation, data are primarily a function of y1, and y2 can be ignored with minimal loss of information.*

Sammon maps [91] are another dimensionality reduction technique that attempts to maintain the Euclidean distance between points in the initial feature space, and a reduced-feature output space. In this case, the axes of the output space represent non-linear combinations of the original features. The method considers all distances between data points so that a data set with *n* records requires *n!* computations – intractable for large values of *n*. For such data sets, approximations may be derived using a sparse set of comparisons [92] or by explicitly learning the transform function [93]. Due to its reliance on Euclidean distance, Sammon maps are not easily applicable to data sets that contain categorical or binary features.

### 3.2.2 Data Fusion of Heterogeneous Data

The combination of multiple sources and types of data to provide a single, more informative, variable is known as data fusion. Many of the key data fusion concepts were developed within robotics and have since been adapted to healthcare data. A detailed discussion of data fusion can be found in [94]. Here, we describe two specific data fusion approaches, their engineering application, and their subsequent use for healthcare problems.

Kalman filters have found wide-spread application in many different data fusion problems using clinical data [94]. The Kalman filter (KF) is a recursive linear estimator which calculates estimates for a continuous valued state that evolves over time on the basis of existing observations of the state [94]. The underlying assumptions are Bayesian (explained in 1.2.1.3) where estimations of parameters are made based on conditional probabilities. In the case of the KF the condition parameters are the probabilistic observations of the values of the variable in time. The evolution of the parameter of interest $\boldsymbol{x}(\boldsymbol{t})$ is, thus, described using an explicit statistical model. Another statistical model is also used to describe the way that the observations, $\boldsymbol{z}(\boldsymbol{t})$, are related to $\boldsymbol{x}(\boldsymbol{t})$. The gains of the KF are then chosen such that the resulting estimate of the parameter of interest $\widehat{\boldsymbol{x}}(\boldsymbol{t})$ minimises mean-squared error and is thus the conditional mean, $\widehat{\boldsymbol{x}}(\boldsymbol{t}) = \boldsymbol{E}[\boldsymbol{x}(\boldsymbol{t})/\boldsymbol{Z}^{\boldsymbol{t}}]$**.** This means that the estimated value is calculated as an average and not as a most likely value as in other probabilistic approaches. Because of the explicit description of process and observations, and the consistent use of statistical measures of uncertainty, the Kalman Filter framework makes it possible to incorporate different sensor models into the basic form of the algorithm. Additionally, at each point in time, it is possible to evaluate the role each sensor plays in the performance of the system, making it an ideal approach for data fusion.

Example successful applications can be found in [95] and in [96]. In the former, a KF framework was employed in order to fuse heart rate (HR) estimates extracted from different signals based on individual Signal Quality metrics for each signal. In the latter, an extension of the basic KF framework was used, the Factorial Switching Kalman Filter (FSKF) which applies a third set of variables, in addition to the *observations* and *states* of the classic KF framework, called the *factors*. The FSKF in this case, was used in order to estimate the true values of vital signs in the Neonatal Intensive Care (NICU) at times where the measurements were obscured by artefact. The *factors* incorporated into the system in this case were related to possible system failures causing artefact, such as probe dropouts, incubator open, etc. These factors had a range of possible settings and at each given point, the existing setting was taken into account in the estimation model.

One other approach to data fusion is novelty detection. Novelty detection methods are used when we wish to classify normal and abnormal data records, but only have very few abnormal training examples. In this case, the challenge is often to accurately differentiate between extreme, but normal data, and truly abnormal data. When a single source of information is used, differentiation is sometimes impossible. For instance, for heart rate data, a low value can either indicate good health, or underlying problems with the heart's electrical activity. In these cases, integrating data from multiple different sensors allows for more accurate classification. Multi-sensor data fusion is an approach that has been applied to traditional engineering applications such as jet engine monitoring [97], and adapted for use with healthcare data. One specific data fusion algorithm for vital sign monitoring is described in greater detail in Section 4.1.

Whilst these multi-sensor data fusion approaches are useful in specific healthcare settings, they do not address one of the unique aspects for healthcare data analysis: the rich variety of data acquired. Health data sets often contain variables of multiple data types - a property known as heterogeneity. A data type defines the set of values that a data item may take; common data types include text, binary, categorical, ordinal and continuous (or floating point). Whilst some data fusion and machine learning methods may be adapted (e.g. see [98]), most are unable to deal with multiple data types simultaneously.

One promising approach for fusing heterogeneous data is Multiple Kernel Learning (MKL). In traditional kernel learning methods, such as SVMs (see Section 2.1.2), a kernel function outputs a measure of similarity, given a pair of data inputs. The kernel function is typically chosen a priori, based on known properties of the data. MKL methods differ by learning and using the optimum linear sum of a family of kernel functions:

$$K = \sum_i \ b_i k_i \tag{15}$$

Due to the property of kernel functions, K is also a kernel function, so standard techniques can then be applied (for example, [99]). If each individual kernel is tailored for use with particular data types, the kernel, K, provides a blend that allows us to process fuse multiple data types optimally.

MKL has been successfully applied to heterogeneous data in the healthcare setting. One example, from Ye et al. [100] showed how MRI image data could be fused with patient demographics and genomic results to help diagnose Alzheimer’s disease more effectively than by using any one data type.

### 3.2.3 Technical and Sociological Issues

The use of multiple data sources brings many practical obstacles that are exacerbated in the healthcare space. Currently, clinical data are collected by multiple devices, belonging to multiple organisations. The data are stored in separate databases, a phenomenon referred to as data compartmentalisation. The barriers to successful data sharing are manifold.

First, the linkage of databases poses ethical and legal concerns. Many countries have legislation that governs the use of personal health information [101]. In the UK, this means that personal data can only be used under specific circumstances. Individual databases may be released more generally if the data set is anonymised so that individuals cannot be identified. One prominent example of anonymised health care data is MIMIC-II, which provides access to over 30,000 de-identified hospital patient records. If data are not de-identified properly, there remains a risk that individuals can be identified by piecing together information from complementary data sources. For instance, Gymrek et al. [102]

showed how de-identified genomic data could be traced to individuals by linkage with genealogy databases.

Secondly, there are non-trivial technical challenges in linking databases. Many of these, including minimising levels of data redundancy, are addressed through the academic discipline of data integration theory (See [103] for further information). Practically, successful data integration requires well-defined standards to ensure that database fields can be interpreted unambiguously, and that the field contents are harmonious. One such standard, SNOMED-CT, provides a comprehensive collection of codes for medical terms that could to help structure database items [104]. Unfortunately, the use of multiple competing standards has hindered data integration. Most notably, until recently, the majority of UK healthcare IT systems used an alternative dictionary, Read codes [105].

Finally, database linkage may require cooperation between competitive system manufacturers. In some cases, the data providers may simply disagree with the intended use of the data [106]. More typically, there may be willingness to share data sets, but details such as ownership of the data or the rights of any generated intellectual property [107] means that data sharing agreements are often complex.

# 4 Case Studies

## 4.1 Application of Machine Learning for the prediction of patient deterioration in an emergency department

Studies have shown that patients experiencing adverse events in hospitals (such as cardiac arrest or admission to the ICU) present with abnormal vital signs before the event, with many of those, presenting abnormalities up to 24h in advance [108]. Because the current standard of recording vital sign observations is paper-based and observations are taken intermittently outside of the ICU, these abnormalities are often missed, especially in busy clinical environments. Additionally, current alerting strategies rely on single parameters or in the calculation of Early Warning Scores (EWS) based on a rule-based pre-set thresholds. The first case study we review concerns the automation of the process of calculating the health status of the patient via a data-driven, rather than rule-based, machine learning model. Firstly, we will present the approach and then discuss its application in an emergency department.

### 4.1.1 System Overview

The system, initially presented in [71], tracks patient status in real time by fusing the patient's vital sign data from monitors in a general ward. The parameters used are heart rate, breathing rate, blood pressure, arterial oxygen saturation ($SaO_2$) and skin temperature. With the exception of blood pressure, vital signs were measured every 5s. Blood pressure was measured every 30 mins using an inflatable cuff placed over the medial artery. The proposed system does not extract any rules connecting these five parameters to patient deterioration, it simply *learns* a model of normality directly from the available data. The system fuses the five vital signs in order to produce a single-parameter representation of patient status, the Patient Status Index (PSI). This PSI is calculated via a *probabilistic model of normality* in five dimensions, previously *learnt* from the vital sign data taken from a representative sample of high-risk adult patients. The PSI is calculated continuously and whenever the vital signs fall outside the learnt envelope of normality, an alert is generated [71]. The aim behind the development of such a system is for use as a real-time early warning system for triggering the intervention of a Medical Emergency Team (MET). Such systems can be integrated with existing patient monitors or a central station on the relevant ward to facilitate the simultaneous monitoring of a large number of patients without increasing the burden of the clinical staff.

### 4.1.2 Training data and pre-processing

The training data set included 3500 h of vital sign data collected from 150 general-ward patients at the John Radcliffe Hospital, Oxford (average length of stay of 24h per patient), who were classified as "high-risk" based on a set of assessments proposed by the attending clinicians. The feature vector was defined as $\boldsymbol{x} = \{x_{1,},\ldots,x_5\}$, the vector of the five vital signs. Because the units and dynamic ranges of each parameter are different (i.e. an increase of $0.5^{o}$ C in temperature is more significant than an increase of 0.5 mmHg in blood pressure or 0.5 beats per minute (bpm) in heart rate), the vital signs were normalized before forming the feature vector **x**. Observation of the data revealed that all except for arterial oxygen saturation ($SaO_2$) followed a near-Gaussian distribution ($SaO_2$ was one-sided as it cannot exceed 100%) so pre-processing included a standard zero-mean, unit variance normalization. To deal with the noise in the data caused by patient movement, data were short-term median filtered. Median filtering was also used in order to deal with missing parameter streams that occurred. With the exception of blood pressure, which was only measured every 30 minutes, if no valid measurement of a parameter was acquired for 1 minute, the value from a historic median filtered was used, derived from the most recent 5 minutes of valid data. If the gap in a measurement persisted for 30 minutes (possibly because of a disconnected probe), then the mean of the training set was used instead. In effect, any missing parameter was replaced by the 'most normal' or 'expected' value in the parameter vector, $\boldsymbol{x}$.

### 4.1.3 Model overview

The model of normality was defined as the unconditional probability density function, $\widehat{p}\,(\boldsymbol{x})$, and was estimated using the training data using a combination of k-means clustering and Parzen windows. Initially, the k-means clustering algorithm is used in order to select 500 cluster centers from the tens of thousands of normalized feature vectors in the training set. Each center $\boldsymbol{x_j}$ (also called a prototype pattern) is then a kernel in the Parzen windows estimator of the pdf given by:

$$\hat{p}(\boldsymbol{x}) = \frac{1}{N(2\pi)^{d/2}\sigma^{d}} \sum_{j=1}^{N} exp\left(\frac{-\|x-x_j\|^2}{2\sigma^2}\right) \qquad (16)$$

Where each spherical kernel has the same global width $\sigma$ and $d$ is equal to 5. The PSI then quantifies departures from normality so that alerts can be generated when the index increases above a threshold value. The PSI is then calculated by transforming the probability so that abnormality increases on a vertical scale:

$$PSI = log_e \left[\frac{1}{\hat{p}\ (x)}\right] \tag{17}$$

A PSI of 3.0, corresponding to a probability value of 0.05 was chosen for the alerting threshold and an alert was generated when the PSI was above this threshold of 3.0 for 4 out of 5 minutes.

### 4.1.4 Testing of the PSI in the Emergency Department

The system described was validated on several different clinical trials. Here, we present its validation on the emergency department (ED) of a medium-sized teaching hospital [109]. In this particular study, the aim was to investigate whether employment of the PSI, as calculated using the learnt model, would be able to detect patient deterioration and generate an alert earlier than the standard practice of manually recording vital sign and Track and Trigger (T&T) data (also known as Early Warning Scores). Data were collected from adults entering the resuscitation room, 'majors' and observation ward of the ED. Heart rate, blood pressure, respiratory rate (RR), oxygen saturation ($SpO_2$), temperature and Glasgow Coma Scale (GCS) as well as paper T&T scores were collected retrospectively from observation charts. For calculation of the PSI, continuous vital sign data (RR, heart rate, blood pressure, $SpO_2$) were acquired using bedside monitors and saved to a server. The "gold standard" of patient deterioration was captured by recording escalations of care. This was done by two clinicians who retrospectively and independently reviewed the clinical notes to identify escalations. In the case of disagreement, a third clinician reconciled the discrepancies.

### 4.1.5 Results

Out of the 400 patients for whom continuous vital signs were collected and PSI scores were calculated, 35 had an escalation after arriving at the ED. 15 of them had no PSI score at the time around the escalation either because of equipment failure, unavailability for monitoring at the time of escalation, or because their escalation was deemed to have been due to ongoing conditions rather than a new deterioration occurring at the ED. Of the remaining 20 patients who experienced deterioration while in the ED, 15 were detected by the PSI. PSI greatly outperformed T&T and there were many cases where the PSI would predict deterioration before the traditional paper-based T&T.

### 4.1.6 Conclusions

This study highlighted the potential of machine-learning data fusion approaches for predicting events in patients. Because such scores can be calculated continuously deterioration can be detected earlier compared to systems relying on documented intermittent observations. Additionally, small deviations in one or more parameters can be recognised promptly whereas in current practice an alert would be generated when a parameter shows a big deviation from normality. This study also showcased some of the problems in using automatic machine-learning based systems using clinical data, in that data are very often incomplete or absent altogether due to practical problems (power failure, server failure, movement of patients or monitors or removal of leads from the patient). Machine learning approaches need to have the flexibility to allow for this kind of failures if they are to be incorporated systematically into clinical practise.

## 4.2 Application of Machine Learning for Assessing the Clinical Acceptability of Electrocardiograms

The explosion of m-health applications both in the developing and developed world has the potential to deliver information and decision support to people that would not otherwise have had access to medical treatment, it is important that stringent quality controls are put into place such that the measurements that reach the untrained recipient are reliable and that noisy measurements are not used.  The second case study we will review concerns the creation of a system which is intended to provide real-time feedback on the diagnostic quality of the Electrocardiogram (ECG) and prompt an inexperienced or lay user to make adjustments in the recording of the data until the quality is sufficient that a reliable medical diagnosis can be made, primarily or arrhythmias [56]. The study was developed as part of the PhysioNet/Computing in Cardiology challenge 2011, which further specified that the algorithm should be efficient enough to be able to run in near real-time on a mobile phone.

### 4.2.1 Training Data and Annotating

Data to support development and evaluation of challenge entries were collected by the Sana Project and provided freely via PhysioNet. The dataset includes 1500 10s recordings of standard 12-lead ECGs, which were sampled at 500 Hz for a minimum of 10s by nurses, technicians and volunteers with varying degrees of experience. 1000 recordings were available as training data and 500 recordings as test data. Each ECG recording was annotated by a minimum of 3 and a maximum of 18 annotators who assigned a rating to the sample related to its quality. The final label of each sample was determined by the average rating and some pre-set thresholds, such that recordings were divided into three classes: acceptable (70%), unacceptable (30%) and indeterminate (<1%). Because of the discrepancy in the number of records from the acceptable and unacceptable classes, bootstrapping was employed in order to increase the samples in the “unacceptable” class by using additive real noise to clean data taken from other ECG databases. In the resulting database, 20,000 10s ECG samples were used for training and 10,000 for testing, both sets of which were balanced for “acceptable” and “unacceptable” recordings [56].

### 4.2.2 Model Overview

Each ECG channel was down-sampled to 125 Hz using an anti-aliasing filter and QRS detection was performed using two different open source QRS detectors. Next, seven quality indices were extracted for each one of the 12 leads, resulting in 84 features per recording. Those 7 indices were the percentage of beats detected on each lead which were detected on all leads, the percentage of beats detected by the one QRS detector that were also detected by the other, the relative power in the QRS complex, the third moment (skewness) of the distribution, the fourth moment (kurtosis) of the distribution, the percentage of the signal which appeared to be a flat line and the relative power in the baseline. All features which were not given by percentages were normalized by subtracting the median such that all features were in the range of [0 1]. The features were then used to train a classifier using two different models: a Support Vector Machine (SVM) and a standard feed-forward Multi-Layer Perceptron Neural Network (MLPNN). Classifiers were tested using all 12 leads simultaneously, i.e. using 84 features and using a single lead only, i.e., using only 7 features. Additionally, different combinations of the seven features were tested in order to find the best.

### 4.2.3 Results

For the single-lead case, the best overall results were obtained for the SVM with a classification accuracy of 96.5% on the test data and corresponding Sensitivity and Specificity of 97.2% and 95.8%, respectively, using only four out of the seven features. For the 12-lead case, the best results were given using five out of the seven features, using the SVM with an accuracy of 95.9% and Sensitivity and Specificity of 96.0% and 95.8%, respectively.

### 4.2.4 Conclusions

The proposed system achieved training accuracies of 98% and test set accuracies up to 97% which indicate that extremely accurate classification of noisy ECGs is possible. Important improvements were noted when the training sets were balanced using artificial data. Lastly, on inspection of the incorrectly classified data, it was found that the labels were ‘borderline’ and could be relabelled either way, and the test accuracy was considered to approach 100%.

# 5. Concluding Remarks and Future Directions

In this chapter, we provided an overview to the current processes and techniques used to analyse heterogeneous and high dimensional healthcare data. We addressed some of the technical challenges and highlighted the real-world issues that are unique to healthcare.

As health data analysis continues to develop as a research field, one may expect to see new analysis methods tailored towards big health data. Currently, we have seen a trend towards the redevelopment of adaptation of traditional machine learning approaches for use with large data sets. IBM Watson’s success in natural language processing follows on from a wealth of previous research. Similarly, Google’s Deepmind extends artificial neural network methods via the field of deep learning. These new approaches have already shown great promise in other fields. Most notably, in early 2016, the AlphaGo program used deep learning (in combination with other methods) to defeat the world-class Lee Sedol at the game of Go – a scenario thought improbable 10 years ago. Both IBM and Google have since expressed interest in healthcare data. Whilst output from both parties has been limited at the time of writing, there is precedent for using deep learning methods on medical images [110,111].

As machine learning methods are applied to increasingly large datasets, we expect the associated challenges to also increase. In particular, the current trend is towards using data collected within routine clinical care – potentially providing datasets many orders of magnitude larger than from research studies. As these routinely-collected datasets are not carefully curated, resulting data are very likely to be of lower quality, such that corrupt and missing entries are more common. The combination of larger datasets and poorer data quality means that automated methods of reliably and accurately processing missing data will be increasingly necessary.

The use of routine data also offer new opportunities to link multiple sources of data. Whilst we have touched on the benefits of data fusion, future research is likely to bring together data from surprisingly disparate sources. For instance, recent research is starting to link consumer research data from supermarket loyalty cards with health data. The increasing number of data features means that robust methods must be found to ensure that the underlying features are not lost amongst the plethora of variables. The complexities and disparities need to be carefully considered by the research community, so that the potential of machine learning applications in clinical data may be reached. Once many of these issues are resolved, machine learning has the potential to deliver a step-change in the manner in which the monitoring of patients and diagnosis of disease is performed for a sustainable future of healthcare management.

# References

**[1]** Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Affairs. 2014 Jul 1;33(7):1115-22.

**[2]** Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Affairs. 2014 Jul 1;33(7):1163-70.

**[3]** Foster KR, Koprowski R, Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research-commentary. Biomedical engineering online. 2014 Jul 5;13(1):1.

**[4]** Bishop CM. Pattern recognition and Machine Learning. Springer-New York 2006.

**[5]** Sajda P. Machine learning for detection and diagnosis of disease. Annu. Rev. Biomed. Eng. 2006 Aug 15;8:537-65.

**[6]** Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their Applications. 1998 Jul;13(4):18-28.

**[7]** Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. Proceedings of the IEEE. 2016 Feb;104(2):444-66..

**[8]** Lucas PJ, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artificial intelligence in medicine. 2004 Mar 1;30(3):201-14.

**[9]** Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology. 1996 Dec 31;49(12):1373-9.

**[10]** Asgari S, Mehrnia A, Moussavi M. Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine. Computers in biology and medicine. 2015 May 1;60:132-42.

**[11]** Oliver A, Freixenet J, Marti R, Pont J, Pérez E, Denton ER, Zwiggelaar R. A novel breast tissue density classification methodology. IEEE Transactions on Information Technology in Biomedicine. 2008 Jan;12(1):55-65.

**[12]** Healthcare Commission. Report of the healthcare Commision's visit to Maidstone and Tunbridge Wells NHS Trust on 12 and 13 December 2007. Retrieved 20-Jul-2016 url: http://webarchive.nationalarchives.gov.uk/20060502043818/http://healthcarecommission.org.uk/_db/_documents/Maidstone_and_Tunbridge_Wells_follow_up_visit_report_-_Dec_07.pdf

**[13]** Boba M, Kołtun U, Bobek-Billewicz B, Chmielik E, Eksner B, Olejnik T. False-negative results of breast core needle biopsies–retrospective analysis of 988 biopsies. Polish Journal of Radiology. 2011 Jan;76(1):25.

**[14]** Clarke GW, Chan AD, Adler A. Effects of motion artifact on the blood oxygen saturation estimate in pulse oximetry. InMedical Measurements and Applications (MeMeA), 2014 IEEE International Symposium on 2014 Jun 11 (pp. 1-4). IEEE.

**[15]** Hamilton PS, Curley MG, Aimi RM, Sae-Hau C. Comparison of methods for adaptive removal of motion artifact. InComputers in Cardiology 2000 2000 (pp. 383-386). IEEE.

**[16]** Yang CC, Hsu YL. A review of accelerometry-based wearable motion detectors for physical activity monitoring. Sensors. 2010 Aug 20;10(8):7772-88.

**[17]** Celler BG, Sparks RS. Home Telemonitoring of Vital Signs—Technical Challenges and Future Directions. IEEE journal of biomedical and health informatics. 2015 Jan;19(1):82-91.

**[18]** Hernandez-Silveira M, Ahmed K, Ang SS, Zandari F, Mehta T, Weir R, Burdett A, Toumazou C, Brett SJ. Assessment of the feasibility of an ultra-low power, wireless digital patch for the continuous ambulatory monitoring of vital signs. BMJ open. 2015 May 1;5(5):e006606.

**[19]** Steinhubl SR, Feye D, Levine AC, Conkright C, Wegerich SW, Conkright G. Validation of a portable, deployable system for continuous vital sign monitoring using a multiparametric wearable sensor and personalised analytics in an Ebola treatment centre. BMJ Global Health. 2016 Jul 1;1(1):e000070.

[**20**] SO80601-2-61:2011: Medical electronical equipment — Particular requirements for basic safety and essential performance of pulse oximeter equipment. International Organization for Standardization, Geneva, Switzerland

**[21]** Milner QJ, Mathews GR. An assessment of the accuracy of pulse oximeters. Anaesthesia. 2012 Apr 1;67(4):396-401.

**[22]** Modell JG, Katholi CR, Kumaramangalam SM, Hudson EC, Graham D. Unreliability of the infrared tympanic thermometer in clinical practice: a comparative study with oral mercury and oral electronic thermometers. Southern medical journal. 1998 Jul;91(7):649-54.

**[23]** Beevers G, Lip GY, O'Brien E. Blood pressure measurement: Part II--conventional sphygmomanometry: Technique of auscultatory blood pressure measurement. British Medical Journal. 2001 Apr 28;322(7293):1043.

**[24]** Baker PD, Westenskow DR, Kück K. Theoretical analysis of non-invasive oscillometric maximum amplitude algorithm for estimating mean blood pressure. Medical and biological engineering and computing. 1997 May 1;35(3):271-8.

**[25]** Pavlik VN, Hyman DJ, Toronjo C. Comparison of Automated and Mercury Column Blood Pressure Measurements in Health Care Settings. Journal of clinical hypertension (Greenwich, Conn.). 2000 Mar;2(2):81-6.

**[26]** Wong WC, Shiu IK, Hwong TM, Dickinson JA. Reliability of automated blood pressure devices used by hypertensive patients. Journal of the Royal Society of Medicine. 2005 Mar 1;98(3):111-3.

**[27]** Akpolat T, Dilek M, Aydogdu T, Adibelli Z, Erdem DG, Erdem E. Home sphygmomanometers: validation versus accuracy. Blood pressure monitoring. 2009 Feb 1;14(1):26-31.

**[28]** Thimbleby H. Improving safety in medical devices and systems. InHealthcare Informatics (ICHI), 2013 IEEE International Conference on 2013 Sep 9 (pp. 1-13). IEEE.

**[29]** Thimbleby H. Ignorance of interaction programming is killing people. interactions. 2008 Sep 1;15(5):52-7.

**[30]** Wilson SJ, Wong D, Clifton D, Fleming S, Way R, Pullinger R, Tarassenko L. Track and trigger in an emergency department: an observational evaluation study. Emergency Medicine Journal. 2012 Mar 22:emermed-2011.

**[31]** Prytherch DR, Smith GB, Schmidt P, Featherstone PI, Stewart K, Knight D, Higgins B. Calculating early warning scores—a classroom comparison of pen and paper and hand-held computer methods. Resuscitation. 2006 Aug 31;70(2):173-8.

**[32]** Callen J, McIntosh J, Li J. Accuracy of medication documentation in hospital discharge summaries: A retrospective analysis of medication transcription errors in manual and electronic discharge summaries. International journal of medical informatics. 2010 Jan 31;79(1):58-64.

**[33]** Wallace DR, Kuhn DR. Failure modes in medical device software: an analysis of 15 years of recall data. International Journal of Reliability, Quality and Safety Engineering. 2001 Dec;8(04):351-71.

**[34]** Obradovich JH, Woods DD. Special section: Users as designers: How people cope with poor HCI design in computer-based medical devices. Human Factors: The Journal of the Human Factors and Ergonomics Society. 1996 Dec 1;38(4):574-92.

**[35]** Batchelor JC, Casson AJ. Inkjet printed ECG electrodes for long term biosignal monitoring in personalized and ubiquitous healthcare. In2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2015 Aug 25 (pp. 4013-4016). IEEE.

**[36]** Tarassenko L, Villarroel M, Guazzi A, Jorge J, Clifton DA, Pugh C. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. Physiological measurement. 2014 Mar 28;35(5):807.

**[37]** Takano C, Ohta Y. Heart rate measurement based on a time-lapse image. Medical engineering & physics. 2007 Oct 31;29(8):853-7.

**[38]** Verkruysse W, Svaasand LO, Nelson JS. Remote plethysmographic imaging using ambient light. Optics express. 2008 Dec 22;16(26):21434-45.

**[39]** Pullinger R, Wilson S, Way R, Santos M, Wong D, Clifton D, Birks J, Tarassenko L. Implementing an electronic observation and early warning score chart in the emergency department: a feasibility study. European journal of emergency medicine: official journal of the European Society for Emergency Medicine. 2016 Feb 17.

**[40]** Wong D, Bonnici T, Knight J, Morgan L, Coombes P, Watkinson P. SEND: a system for electronic notification and documentation of vital sign observations. BMC medical informatics and decision making. 2015 Aug 13;15(1):1.

**[41]** Smith GB, Prytherch DR, Schmidt P, Featherstone PI, Knight D, Clements G, Mohammed MA. Hospital-wide physiological surveillance–a new approach to the early identification and management of the sick patient. Resuscitation. 2006 Oct 31;71(1):19-28.

**[42]** Meccariello M, Perkins P, Quigley LG, Rock A, Qiu J. Vital Time Savings: Evaluating the Use of an Automated Vital Signs Documentation System on a Medical/Surgical Unit. J Healthc Inf Manag 2010 24(4):46-51

**[43]** Mekhjian HS, Kumar RR, Kuehn L, Bentley TD, Teater P, Thomas A, Payne B, Ahmad A. Immediate benefits realized following implementation of physician order entry at an academic medical center. Journal of the American Medical Informatics Association. 2002 Sep 1;9(5):529-39

**[44]** Murphy MF, Fraser E, Miles D, Noel S, Staves J, Cripps B, Kay J. How do we monitor hospital transfusion practice using an end- to- end electronic transfusion management system?. Transfusion. 2012 Dec 1;52(12):2502-12

**[45]** Davies A, Staves J, Kay J, Casbard A, Murphy MF. End- to- end electronic control of the hospital transfusion process to increase the safety of blood transfusion: strengths and weaknesses. Transfusion. 2006 Mar 1;46(3):352-64.

**[46]** Staves J, Davies A, Kay J, Pearson O, Johnson T, Murphy MF. Electronic remote blood issue: a combination of remote blood issue with a system for end- to- end electronic control of transfusion to provide a "total solution" for a safe and timely hospital blood transfusion service. Transfusion. 2008 Mar 1;48(3):415-24.

[**47**] Resetar E, Reichley RM, Noirot LA, Dunagan WC, Bailey TC. Customizing a commercial rule base for detecting drug-drug interactions. InAMIA 2005.

[**48**] Bonnici T, Orphanidou C, Vallance D, Darrell A, Tarassenko L. Testing of wearable monitors in a real-world hospital environment: What lessons can be learnt?. In2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks 2012 May 9 (pp. 79-84). IEEE.

[**49**] Wahlster P, Goetghebeur M, Kriza C, Niederländer C, Kolominsky-Rabas P. Balancing costs and benefits at different stages of medical innovation: a systematic review of Multi-criteria decision analysis (MCDA). BMC health services research. 2015 Jul 9;15(1):1.

[**50**] Yeung MS, Lapinsky SE, Granton JT, Doran DM, Cafazzo JA. Examining nursing vital signs documentation workflow: barriers and opportunities in general internal medicine units. Journal of clinical nursing. 2012 Apr 1;21(7- 8):975-82.

[**51**] Tat TH, Xiang C, Thiam LE. Physionet challenge 2011: improving the quality of electrocardiography data collected using real time QRS-complex and T-wave detection. In2011 Computing in Cardiology 2011 Sep 18 (pp. 441-444). IEEE.

[**52**] Johnson AE, Behar J, Andreotti F, Clifford GD, Oster J. Multimodal heart beat detection using signal quality indices. Physiological measurement. 2015 Jul 28;36(8):1665.

[**53**] Clifton DA, Wong D, Clifton L, Wilson S, Way R, Pullinger R, Tarassenko L. A large-scale clinical validation of an integrated monitoring system in the emergency department. IEEE journal of biomedical and health informatics. 2013 Jul;17(4):835-42.

[**54**] Orphanidou C, Fleming S, Shah SA, Tarassenko L. Data fusion for estimating respiratory rate from a single-lead ECG. Biomedical Signal Processing and Control. 2013 Jan 31;8(1):98-105.

[**55**] Daluwatte C, Johannesen L, Galeotti L, Vicente J, Strauss DG, Scully CG. Assessing ECG signal quality indices to discriminate ECGs with artefacts from pathologically different arrhythmic ECGs. Physiological Measurement. 2016 Jul 25;37(8):1370.

[**56**] Clifford GD, Behar J, Li Q, Rezek I. Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. Physiological measurement. 2012 Aug 17;33(9):1419.

[**57**] Nizami S, Green JR, McGregor C. Implementation of artifact detection in critical care: A methodological review. IEEE reviews in biomedical engineering. 2013;6:127-42.

[**58**] Kaur M, Singh B. Comparison of different approaches for removal of baseline wander from ecg signal. InProceedings of the International Conference & Workshop on Emerging Trends in Technology 2011 Feb 25 (pp. 1290-1294). ACM.

[**59**] ECG baseline wander. Reproduced with permission from *http://joachimbehar.comuv.com/ECG_tuto_1.php*

[**60**] Hoffmann S, Falkenstein M. The correction of eye blink artefacts in the EEG: a comparison of two prominent methods. PLoS One. 2008 Aug 20;3(8):e3004.

[**61**] Li Y, Ma Z, Lu W, Li Y. Automatic removal of the eye blink artifact from EEG using an ICA-based template matching approach. Physiological measurement. 2006 Mar 14;27(4):425.

[**62**] Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. The Lancet. 1974 Jul 13;304(7872):81-4.

[**63**] Paul DB, Rao GU. Correlation of Bispectral Index with Glasgow Coma Score in mild and moderate head injuries. Journal of clinical monitoring and computing. 2006 Dec 1;20(6):399-404.

[**64**] Beridze M, Khaburzania M, Shakarishvili R, Kazaishvili D. Dominated EEG patterns and their prognostic value in coma caused by traumatic brain injury. Georgian Med News. 2010 Sep;186:28-33

[**65**] Lawhern V, Hairston WD, McDowell K, Westerfield M, Robbins K. Detection and classification of subject-generated artifacts in EEG signals using autoregressive models. Journal of neuroscience methods. 2012 Jul 15;208(2):181-9.

[**66**] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. Bmj. 2007 Jul 19;335(7611):136.

[**67**] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, Brindle P. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. Bmj. 2008 Jun 26;336(7659):1475-82.

[**68**] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009 Jun 29;338:b2393.

[**69**] Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. Bmj. 2010 May 13;340:c2442.

[**70**] Allison PD. Missing data: Quantitative applications in the social sciences. British Journal of Mathematical and Statistical Psychology. 2002;55(1):193-6.

[**71**] Tarassenko L, Hann A, Young D. Integrated monitoring and analysis for early warning of patient deterioration. British journal of anaesthesia. 2006 Jul 1;97(1):64-8.

[**72**] Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. Applied Statistics. 1992 Jan 1:337-48

[**73**] Kirkwood BR. Essentials of medical statistics. Blackwell Scientific Publications; 1988.

[**74**] Tobin J. Estimation of relationships for limited dependent variables. Econometrica: journal of the Econometric Society. 1958 Jan 1:24-36.

[**75**] Rubin DB. Multiple imputation for nonresponse in surveys. John Wiley & Sons; 2004 Jun 9.

[**76**] Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science. 2007 Sep 1;8(3):206-13.

[**77**] Bodner TE. What improves with increased missing data imputations?. Structural Equation Modeling. 2008 Oct 22;15(4):651-75.

[**78**] Schafer JL. Multiple imputation: a primer. Statistical methods in medical research. 1999 Feb 1;8(1):3-15.

[**79**] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?. International journal of methods in psychiatric research. 2011 Mar 1;20(1):40-9.

[**80**] Allison PD. Handling missing data by maximum likelihood. InSAS global forum 2012 Apr 22 (Vol. 23).

[**81**] Stegle O, Fallert SV, MacKay DJ, Brage S. Gaussian process robust regression for noisy heart rate data. IEEE Transactions on Biomedical Engineering. 2008 Sep;55(9):2143-51.

[**82**] Roberts S, Osborne M, Ebden M, Reece S, Gibson N, Aigrain S. Gaussian processes for time-series modelling. Phil. Trans. R. Soc. A. 2013 Feb 13;371(1984):20110550.

[**83**] Dürichen R, Pimentel MA, Clifton L, Schweikard A, Clifton DA. Multitask Gaussian processes for multivariate physiological time-series analysis. IEEE Transactions on Biomedical Engineering. 2015 Jan;62(1):314-22.

[**84**] Boyle, P. and Frean, M., 2004. Dependent gaussian processes. In *Advances in neural information processing systems* (pp. 217-224).

[**85**] Wong D, Clifton DA, Tarassenko L. Probabilistic detection of vital sign abnormality with Gaussian process regression. InBioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on 2012 Nov 11 (pp. 187-192). IEEE.

[**86**] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. bioinformatics. 2007 Oct 1;23(19):2507-17.

[**87**] Breiman L. Random forests. Machine learning. 2001 Oct 1;45(1):5-32.

[**88**] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.

[**89**] Edmonds J. Matroids and the greedy algorithm. Mathematical programming. 1971 Dec 1;1(1):127-36.

[**90**] Joliffe IT, Morgan BJ. Principal component analysis and exploratory factor analysis. Statistical methods in medical research. 1992 Mar 1;1(1):69-95.

[**91**] Sammon JW. A nonlinear mapping for data structure analysis. IEEE Transactions on computers. 1969 May 1;18(5):401-9.

[**92**] Wong D, Strachan I, Tarassenko L. Visualisation of high-dimensional data for very large data sets. InWorkshop Mach. Learn. Healthcare Appl., Helsinki, Finland 2008.

[**93**] Lowe D, Tipping ME. Neuroscale: novel topographic feature extraction using RBF networks. Advances in Neural Information Processing Systems. 1997:543-9.

[**94**] Durrant-Whyte H, Henderson TC. Multisensor data fusion. InSpringer Handbook of Robotics 2008 (pp. 585-610). Springer Berlin Heidelberg.

[**95**] Li Q. and Clifford G. D., Signal quality and data fusion for false alarm reduction in the intensive care unit, Journal of Electrocardiology, 45(6):596-603, Nov 2012.

[**96**] Williams C, Quinn J, McIntosh N. Factorial switching Kalman filters for condition monitoring in neonatal intensive care.

[**97**] Clifton DA, Bannister PR, Tarassenko L. A framework for novelty detection in jet engine vibration data. InKey engineering materials 2007 (Vol. 347, pp. 305-310). Trans Tech Publications.

[**98**] Ma M, Gonet R, Yu R, Anagnostopoulos GC. Metric representations of data via the Kernel-based Sammon Mapping. InThe 2010 International Joint Conference on Neural Networks (IJCNN) 2010 Jul 18 (pp. 1-7). IEEE.

[**99**] Hu M, Chen Y, Kwok JT. Building sparse multiple-kernel SVM classifiers. IEEE Transactions on Neural Networks. 2009 May;20(5):827-39.

[**100**] Ye J, Chen K, Wu T, Li J, Zhao Z, Patel R, Bae M, Janardan R, Liu H, Alexander G, Reiman E. Heterogeneous data fusion for alzheimer's disease study. InProceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining 2008 Aug 24 (pp. 1025-1033). ACM.

[**101**] Blumenthal D. Launching hitech. New England Journal of Medicine. 2010 Feb 4;362(5):382-5.

[**102**] Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013 Jan 18;339(6117):321-4.

[**103**] Lenzerini M. Data integration: A theoretical perspective. InProceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems 2002 Jun 3 (pp. 233-246). ACM.

[**104**] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics. 2006 Jan;121:279.

[**105**] O'Neil M, Payne C, Read J. Read Codes Version 3: a user led terminology. Methods of information in medicine. 1995 Mar;34(1-2):187-92.

[**106**] Lopez AD. Sharing data for public health: where is the vision?. Bulletin of the World Health Organization. 2010 Jun;88(6):467

[**107**] van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, Heymann D, Burke DS. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014 Nov 5;14(1):1.

[**108**] Goldhill DR, White SA, Sumner A. Physiological values and procedures in the 24 h before ICU admission from the ward. Anaesthesia. 1999 Jun 1;54(6):529-34.

[**109**] Wilson SJ, Wong D, Pullinger RM, Way R, Clifton DA, Tarassenko L. Analysis of a data-fusion system for continuous vital sign monitoring in an emergency department. European Journal of Emergency Medicine. 2016 Feb 1;23(1):28-32.

[**110**] Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. InInternational Conference on Medical Image Computing and Computer-assisted Intervention 2013 Sep 22 (pp. 411-418). Springer Berlin Heidelberg.

[**111**] Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, Pal C, Jodoin PM, Larochelle H. Brain tumor segmentation with deep neural networks. Medical Image Analysis. 2016 May 19.