Unsupervised Bayesian Inference to Fuse Biosignal Sensory Estimates for Personalizing Care

Tingting Zhu, Marco A. F. Pimentel, Gari D. Clifford, and David A. Clifton

***Abstract*—The role of sensing technologies, such as wearables, in delivering precision care is becoming widely acceptable. Given the very large quantities of sensor data that rapidly accumulate, there is a need to employ automated algorithms to label biosignal sensor data. In many real-life clinical applications, no such expert labels are available, and algorithms for processing sensor data must be relied upon, without access to the "ground truth." It is therefore extremely difficult to choose which algorithms to trust or discard at any point in time, where different algorithms may be optimal for different patients, or even for different points in time for the same patient. We propose two fully Bayesian approaches for fusing labels from independent and potentially correlated annotators (i.e., algorithms or, where available, experts). These are generative models to aggregate labels (i.e., the outputs of the algorithms, such as identified ECG morphology) in an *unsupervised* manner, to estimate jointly the assumed bias and precision of each algorithm *without access to the ground truth*. The latter fused estimate may then be used to infer the underlying ground truth. For the first time in the biomedical context, we show that modeling correlations between annotators, and fusing information concerning task difficulty (such as the estimated quality of the sensor data), improve these estimates with respect to commonly employed strategies in the literature. Also, we adopt a strongly Bayesian approach to inference using Gibbs sampling to improve estimates over the existing state of the art. We present results from applying the proposed pair of models to simulated and two publicly available biomedical datasets, to demonstrate proof-of-principle. We show that our proposed models outperform all existing approaches recreated from the literature. We also show that the proposed methods are robust when dealing with missing values (as often occurs in real-life biomedical applications), and that they are suitably efficient for use in real-time applications, thereby providing the basis for the reliable use of sensors for personalizing the care of the individual.**

***Index Terms*—Bayesian inference, data fusion, personalised modelling, unsupervised learning.**

I. Introduction

With the rapid increase in volume of wearable devices being used in clinical practice, there exists the possibility of personalising the care of individuals, such that care is more closely based on their physiology. This "precision" approach to healthcare is predicated on the fact that biosignal data arising from patient-worn sensors can be used for diagnosis and prognosis in a manner that is sufficiently robust and interpretable for use in a clinical scenario. Replacement of "one-size-fits-all" treatments for personalised equivalents would be greatly supported by the use of wearable healthcare sensors. It is recognised, however, that existing systems suffer from a lack of robustness [1], typically because reliable analysis of the very large resulting datasets of sensor data is often impossible using existing methods. Given the very large quantities of sensor data that rapidly accumulate, there is a need to employ automated algorithms to label biosignal sensor data (e.g., abnormal morphology of the ECG). However, automated algorithms are typically less reliable than gold-standard expert labels; the latter are typically sparse and expensive, and it is usually not feasible to obtain expert labels for real-time data arising from actual patients.

Expert labelling of these datasets that contain sensor data is the gold standard for diagnosing many diseases and providing subsequent care. From a clinical perspective, expert labelling is defined as being the result of experienced physicians annotating an area of interest in some data when the ground truth (e.g., the true time-of-occurrence or duration of an event; a numerical estimate of a physiological measurement, such as a vital sign; or the diameter of an object in a medical image) is not readily available. Expert labels are generally used for training automated algorithms when taking a *supervised* approach to learning. However, experts are relatively scarce, their time is expensive, and the task of labelling is time-consuming. It is thus typically impossible to obtain expert labelling in real-time for most practical healthcare-sensing applications, where time-series rapidly becomes too large in quantity for human interpretation. In such cases, automated algorithms must be

Manuscript received October 2, 2017; revised February 21, 2018; accepted March 20, 2018. Date of publication June 4, 2018; date of current version January 2, 2019. The work of T. Zhu was supported by the RCUK Digital Economy Programme under Grant EP/G036861/1. The work of D. A. Clifton was supported in part by the RAEng, an EPSRC Challenge Award, and in part by the EPSRC-funded "ASPIRE" programme. *(Corresponding author: Tingting Zhu.)*

T. Zhu, M. A. F. Pimentel, and D. A. Clifton are with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX1 2JD, U.K. (e-mail: tingting.zhu@eng.ox.ac.uk; marco.pimentel@eng.ox.ac.uk; david.clifton@eng.ox.ac.uk).

G. D. Clifford is with the Department of Biomedical Informatics, Emory University, Atlanta, GA 30322 USA, and also with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: gari@alum.mit.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2018.2820054

relied upon to label real-time datasets. Expert labelling can be further complicated by the ambiguous definition of what it means to be an "expert" for a given medical application. There is no standardisation for testing measures of clinical competency above a proficient level, both in terms of accuracy and consistency, and no empirical method for measuring levels of clinical expertise, even though the quality of annotation relies heavily on the expert's experience. Therefore, large inter- and intra-annotator variabilities occur when physicians label data, depending on their respective experience and level of training [2], [3].

The "experts" in terms of the work described by this paper could be either humans (such as trained clinicians), or automated algorithms, or some combination of both. Empirical studies have shown that both humans and automated algorithms are likely to have their own bias values regardless of their level of "expertise" [3]–[5]. The bias of each labeller (also termed annotator) is defined as being the average difference between the expert's estimate and the (unseen) true value. In the context of labelling medical data, being reliant on results from automated algorithms to produce a diagnosis (or to make a patient-specific decision for a certain treatment) is typically not sufficient due to large inter- and intra-variabilities between recommendations from such algorithms. Furthermore, when the ground truth is unknown, it is difficult to choose which algorithms to trust or discard, or even how to merge their recommendations to form a consensus output.

In this paper, we propose two generative models for aggregating individual continuous labels in a principled manner and inferring a more reliable *label* than each individual labeller considered independently. We demonstrate our models using two datasets from exemplar biomedical applications where subjective continuous labels of some presumed underlying ground truth are provided by "experts". These "experts" can be independent or potentially-correlated. For the purpose of this paper, we will accept the broad definition of "expert" to include automated algorithms for producing continuous-valued estimates given exemplar biomedical data. In keeping with the literature, we will refer to "annotators" and "experts" interchangeably. Similarly, we will refer to "annotations" and "labels" as referring to the outputs of the experts.

II. Related Work

There exist some key contributions in the literature for modelling continuous-valued labels, and the biases and expertise of each annotator: Raykar *et al.* [6] used an expectation maximisation (EM) method to fuse continuous-valued labels for measuring the diameter of a suspicious lesion from a medical image. Welinder and Perona [4] devised a Bayesian EM framework for fusing binary, multi-valued, and continuous-valued labels, which explicitly modeled the precision of each annotator to account for their varying skill levels, without modelling their bias values. We have extended this Bayesian framework to jointly model the annotators biases and precisions using a Bayesian treatment [7]. In medical imaging, Warfield *et al.* [3] proposed a method for validating segmentation by estimating the bias and variance of each annotator. A similar model was described by Ouyang *et al.* [8] that estimated the quantitative ground truth (such as count and percentage estimation) in a "crowd sensing" setting. Xing *et al.* recently proposed using a Gaussian prior on the bias parameter for the identification of cardiac landmarks in two-dimensional images [9]. However, their model does not cater for missing annotations or the incorporation of physiological features into the model to further improve the estimation of ground truth [6]. Furthermore, existing approaches have no principled means of accounting for uncertainty due to missingness and/or quality of the data. The aforementioned studies [3], [4], [6], [9], [10] serve as the basis of comparison for the proposed models. A comprehensive comparison of such models is presented in the Supplementary Materials Table III. Note that these models are unsupervised methodologies. Other related work concerns the use of supervised approaches. Ensemble methods, such as Bagging, Boosting, AdaBoost, and more recently Rotation Forests, combine the predictions from multiple base learners to form ensembles, which typically achieve more accurate aggregate predictions than the individual base learners. Ensemble learning, in general, is a performance-driven approach which involves a measure of quality of the base learners (e.g., accuracy), either for training or for estimating the weights of each base learner, for which labelled training data (i.e., training data with a ground-truth) are required. Also, their interpretability is critically limited [11]. Our work diverts from this body of literature in that the proposed learning method does not require a ground truth, as commonly found in medical and biomedical datasets.

III. Novelty

In this paper, we built on our previous model [7], and propose a fully-Bayesian approach via Gibbs sampling for fusing continuous-valued labels from independent and/or potentially-correlated annotators to form a consensus in an unsupervised manner. The labels can be either QT intervals of an ECG measured, timestamps of abnormal ECG peaks from a sensor-based signal, or the estimated respiratory rate values from a wearable device. This allows, for the first time, a full distribution over the posterior estimates, due to the use of a strongly-Bayesian approach, and an explicit quantification of our uncertainty in the estimates using a signal quality extension. Additionally, the proposed framework overturns the strong assumptions of existing approaches such that we take into account the precision and bias of the individual annotators, in addition to estimating the correlation that exists between annotators. All existing algorithms from the literature assume that the experts are independent, including our previously proposed model [7]. In this study, we aim to demonstrate that its performance can be improved using the newly proposed models. Allowing such correlations is an important means of representing, for example, the distinction between "experts" and "novices", or between human experts and automated algorithms. To the best of the authors' knowledge, this is the first attempt to perform such task, and we will demonstrate that explicit modelling of these correlations improves model performance with respect to the start-of-the-art. We aim thereby to make a case that the methods proposed in this paper increase robustness in our example sensor-based healthcare applications such that the results can support the use of patient-specific personalised treatments.

IV. Proposed Generative Models

The generative models considered by this paper are stochastic processes that are assumed to have generated the available observations, given some model parameter values. These will be fully-probabilistic, in which we model the joint probability distribution over all parameters.

A. Modelling the Latent Ground Truth

Suppose that there are N records of N labels (annotations). The underlying ground truth for the ith record, z_i (such as a physiological timestamp of an event, a predicted continuous-valued of an vital sign, and a QT interval value in the ECG), can be assumed to be drawn from a Gaussian distribution with mean a_i and variance $1/b$. The probability density function (pdf) of z_i is defined as follows:

$$p\left(z_i \mid a_i, b\right) = \mathscr{N}\left(z_i \mid a_i, 1/b\right), \tag{1}$$

where a_i can be expressed as a linear regression function $f(\mathbf{w}, \mathbf{x}_i)$ with an intercept w_0: $\mathbf{w}$ are the coefficients of the regression (which includes w_0[1]). $\mathbf{x}_i$ is a column feature vector for the ith record containing d features (i.e., we have an $(N \times d)$-dimensional design matrix, $\mathbf{X} = [\mathbf{x}_1^\intercal; ...; \mathbf{x}_N^\intercal]$). To cater for the modelling of w_0, a scalar value of one was added in the feature matrix (i.e., $\mathbf{x}_i := [1, \mathbf{x}_i]$)). Furthermore, the precision of the ground truth defined as the inverse-variance, b, is assumed to be modeled from a gamma distribution[2] as follows:

$$p\left(b \mid k_b, \vartheta_b\right) = \text{Gamma}\left(b \mid k_b, \vartheta_b\right), \tag{2}$$

where k_b is the shape parameter and ϑ_b is the scale parameter. If one further assumes that the ground truth can be drawn independently from the N records, the conditional pdf of $\mathbf{z}$ as a vector of annotations is given by:

$$p\left(\mathbf{z} \mid \mathbf{x}_i, \mathbf{w}, b_i\right) = \prod_{i=1}^{N} \mathscr{N}\left(z_i \mid \mathbf{x}_i^\intercal \mathbf{w}, 1/b_i\right). \tag{3}$$

z_i may represent a QT interval value of an individual that is drawn from a Gaussian distribution cetered around its mean (i.e., $\mathbf{x}_i^\intercal \mathbf{w}$) (which is determined by the heart rate and age features $\mathbf{x}_i$.) with variance $1/b_i$. For a total of N subjects or N recordings, there are N independent Gaussian.

B. The Independent Annotator Model

Assuming the presence of N recordings, we have a dataset, $\mathbf{D} = [\mathbf{x}_i^\intercal, y_i^{j=1}, \cdots, y_i^{j=R}]_{i=1}^N$, where y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and there are a total of R annotators. In this model, it is assumed that y_i^j is a noisy version of z_i, with a Gaussian distribution $\mathscr{N}(y_i^j \mid z_i, (\sigma^j)^2)$. The motivation for the latter comes from the central limit theorem: given the assumption that the annotations for a given annotator are independent and identically distributed, their residuals (i.e., errors in annotations) derived from the latent ground truth will converge to a Gaussian distribution. In the absence of prior knowledge, this assumption provides a robust and generalisable model for the given data, as will be demonstrated. Here, σ^j is the standard deviation associated with the jth annotator and represents his variance in annotation around ground truth z_i. Furthermore, the bias of each annotator, where it measures the average difference between the estimation and the ground truth, can be modeled as an additional term, denoted as ϕ^j. The pdf of estimating y_i^j can then be written as:

$$p\left(y_i^j \mid z_i, \left(\sigma^j\right)^2\right) = \mathscr{N}\left(y_i^j \mid z_i + \phi^j, 1/\lambda^j\right), \tag{4}$$

where $(\sigma^j)^2$ is replaced with $1/\lambda^j$. Here, λ^j is the precision of the jth annotator, defined as the estimated inverse-variance for annotator j. It is assumed that $y_i^1, \cdots, y_i^R$ are conditionally independent given the ground truth z_i; with the assumption that records are independent, the conditional pdf of $\mathbf{y}$ can be modeled as:

$$p\left(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\lambda}\right) = \prod_{i=1}^{N} \prod_{j=1}^{R} \mathscr{N}\left(y_i^j \mid z_i + \phi^j, 1/\lambda^j\right). \tag{5}$$

The assumption of conditional independence may not be necessarily true in cases where the annotations are generated by algorithms, some of which may be variations in implementation of the same general approach. Nevertheless, this assumption was made to simplify the model and subsequent derivation of the likelihood (see *relaxation of independence assumption* in Section IV-D). The pdf of the bias for annotator j, ϕ^j, is assumed to be drawn from a Gaussian distribution with mean μ_ϕ and variance $1/\alpha_\phi$ [9]:

$$p\left(\phi^j \mid \mu_\phi, \alpha_\phi\right) = \mathscr{N}\left(\phi^j \mid \mu_\phi, 1/\alpha_\phi\right). \tag{6}$$

Although the biases of the annotators might be assumed to have other distributions, such choices are likely to be dataset-dependent. In the absence of any knowledge of the underlying distribution of biases, we adopt the strategy of assuming them to be drawn from a Gaussian distribution. As described earlier, it is assumed that precision values, such as λ^j and α_ϕ, were drawn from a gamma distribution, with parameters k_λ, ϑ_λ, and k_α, ϑ_α, respectively:

$$p\left(\lambda^j \mid k_\lambda, \vartheta_\lambda\right) = \text{Gamma}\left(\lambda^j \mid k_\lambda, \vartheta_\lambda\right). \tag{7}$$

$$p\left(\alpha_\phi \mid k_\alpha, \vartheta_\alpha\right) = \text{Gamma}\left(\alpha_\phi \mid k_\alpha, \vartheta_\alpha\right). \tag{8}$$

In the case when predicting an QT interval for the ith subject: the QT interval, y_i^j, estimated from algorithm j is assumed to

[1] w_0 models the overall offset predicted in the regression, which is different from the annotator specific bias ϕ in the proposed models that will be described in later sections.

[2] A gamma distribution can be defined as $\text{Gamma}(x \mid k, \vartheta) = \frac{1}{\Gamma(k)\vartheta^k} x^{k-1} \exp(-\frac{x}{\vartheta})$, where k is the shape of the distribution and ϑ is the scale of the distribution, $\Gamma(\cdot)$ is the gamma function. The gamma distribution is commonly used to model positive continuous values and it is therefore assumed that precision values are drawn from a gamma distribution.

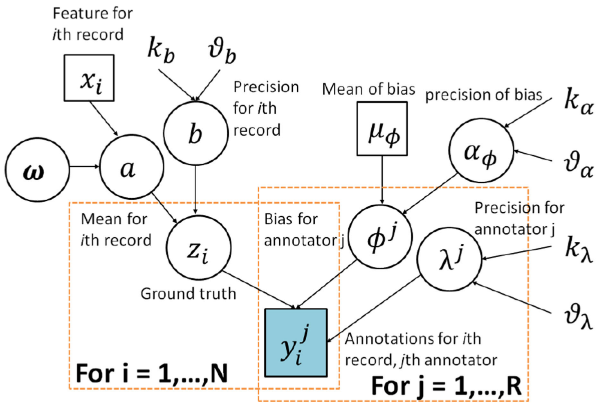


Fig. 1. Graphical representation of the BCLA model [7]: y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and is modeled by the z_i (the unknown underlying ground truth), the ϕ^j (bias), and the λ^j (precision). Furthermore, z_i is drawn from a Gaussian distribution with parameters mean a and variance $1/b$, where a for the ith record is a function of feature vector $\mathbf{x}_i$ as a linear regression function $f(\mathbf{w},\mathbf{x})$, and $\mathbf{w}$ being the coefficients of the regression. ϕ^j is modeled from a Gaussian distribution with mean μ_ϕ and variance $1/\alpha_\phi$. The b, λ^j, and α_ϕ are drawn from a gamma distribution with parameters k_b, ϑ_b, k_λ, ϑ_λ, and k_α, ϑ_α, respectively.

be within $\pm\sqrt{1/\lambda^j}$ away from the truth QT value for subject i (i.e., z_i), with an offset, ϕ^j, that is specific to algorithm j.

C. BCLA - A Joint Model of Ground Truth and Independent Annotators

Bayesian continuous-valued label aggregator (BCLA) [7] is a straightforward means of combining the ground truth and annotator models (see Section IV-A and Section IV-B, respectively). It comprises two key contributions: (i) BCLA provides an unsupervised estimation of the continuous-valued annotations that are valuable for time-series related data, as well as the duration of events for physiological data; (ii) it introduces a unified framework for combining continuous-valued annotations to infer the underlying ground truth, while jointly modelling annotators' bias and precision values. The graphical form of BCLA is presented in Fig. 1. Under the assumption that records are independent, the likelihood of the parameters $\boldsymbol{\theta} = \{\mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\phi}, \alpha_\phi, b, z_i\}$ for a given dataset $\mathbf{D}$ can be formulated as:

$$p(\mathbf{D} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p\left(y_i^1, \cdots, y_i^R \mid \mathbf{x}_i, \boldsymbol{\theta}\right). \tag{9}$$

For the first time, we here propose a fully-Bayesian approach to BCLA modelling using Gibbs sampling, denoted as BCLA-Gibbs. We note that previous methods in the literature are not fully-Bayesian (either using maximum likelihood or maximum-a-posteriori (MAP) approach for estimating model parameters). Our previous work [7], for example, uses a MAP approach (denoted as BCLA-MAP). The posterior probability of the parameters $\boldsymbol{\theta}$ for a given dataset $\mathbf{D}$ can be written using Bayes' theorem as:

$$p(\boldsymbol{\theta} \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}, \tag{10}$$

where

$$\begin{aligned} p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) &= \text{Gamma}(\alpha_\phi \mid k_\alpha, \vartheta_\alpha)\text{Gamma}(b \mid k_b, \vartheta_b) \\ &\times \left[\prod_{j=1}^{R} \mathcal{N}\left(\phi^j \mid \mu_\phi, 1/\alpha_\phi\right) \text{Gamma}(\lambda^j \mid k_\lambda, \vartheta_\lambda)\right] \\ &\times \left[\prod_{i=1}^{N} \mathcal{N}\left(z_i \mid \mathbf{x}_i^\intercal \mathbf{w}, 1/b\right) \prod_{j=1}^{R} \mathcal{N}\left(y_i^j \mid z_i + \phi^j, 1/\lambda^j\right)\right]. \end{aligned}$$

Each parameter in the BCLA-Gibbs likelihood is assumed to be independent, and can therefore be updated in a fully-Bayesian manner by sampling from its conditional posterior distribution with its hyperparameters (denoted by *). The derivations and convergence criterion of BCLA-Gibbs, as well as its implementation are explained in the Supplementary Materials.

Learning From Incomplete Data Using BCLA-Gibbs

For the N annotations from the R annotators, we should consider the case in which there are missing annotations from different annotators (i.e., not all annotators have labelled all recordings). In such a case, the hyperparameters of the posterior distribution of BCLA-Gibbs can be re-written as follows:

$$z_i \sim \mathcal{N}\left(z_i \middle| a_i^*, \frac{1}{b_i^*}\right), \qquad \phi^j \sim \mathcal{N}\left(\phi^j \middle| \mu_\phi^{j*}, \frac{1}{\alpha_\phi^{j*}}\right),$$

$$\lambda^j \sim \text{Gamma}\left(\lambda^j \middle| k_\lambda^{j*}, \vartheta_\lambda^{j*}\right), \qquad b \sim \text{Gamma}\left(b \,|k_b^*, \vartheta_b^*\right),$$

$$\alpha_\phi \sim \text{Gamma}\left(\alpha_\phi \mid k_\alpha^*, \vartheta_\alpha^*\right).$$

where

$$a_i^* = \frac{(\mathbf{x}_i^\intercal \mathbf{w})\, b + \sum_{j \in V_i} \left[\left(y_i^j - \phi^j\right) \lambda^j\right]}{b + \sum_{j \in V_i} \lambda^j}, \quad b_i^* = b + \sum_{j \in V_i} \lambda^j.$$

$$\mu_\phi^{j*} = \frac{\mu_\phi \alpha_\phi + \lambda^j \sum_{i \in U_j} \left(y_i^j - z_i\right)}{\alpha_\phi + \sum_{i \in U_j} \lambda^j}, \quad \alpha_\phi^{j*} = \alpha_\phi + \sum_{i \in U_j} \lambda^j.$$

$$k_\lambda^{j*} = k_\lambda + \frac{N_j}{2}, \quad \frac{1}{\vartheta_\lambda^{j*}} = \frac{\sum_{i \in U_j} \left(y_i^j - \phi^j - z_i\right)^2}{2} + \frac{1}{\vartheta_\lambda}.$$

$$k_\alpha^* = k_\alpha + \frac{R}{2}, \quad \frac{1}{\vartheta_\alpha^*} = \frac{\sum_{j=1}^{R} (\phi^j - \bar{\phi})^2}{2} + \frac{1}{\vartheta_\alpha}.$$

$$k_b^* = k_b + \frac{N}{2}, \quad \frac{1}{\vartheta_b^*} = \frac{\sum_{i=1}^{N} (z_i - \bar{z})^2}{2} + \frac{1}{\vartheta_b}.$$

Note that U_j is the set of records with annotations provided by the jth annotator, V_i is the set of annotators that provided annotations for the ith record, and N_j is the number of records annotated by the jth annotator. $\mathbf{w}$ can be learnt by finding the zero gradient of the expectation of the complete data log-likelihood as $\mathbf{w} = (\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\intercal)^{-1} \sum_{i=1}^{N} \mathbf{x}_i z_i$. The above allows us to cope robustly with the commonly-encountered difficulties arising from incomplete (or even sparse) labelling, in a principled and probabilistic manner.

D. BCLAc – A Combined Ground Truth and Correlated Annotator Model

The BCLA model assumed that annotators are conditionally independent given the features which defined the ground truth. It does not factor in the possible dependency (or correlation) between individual annotators, which might occur between sets of automated algorithms. For example, as outlined previously, a set of algorithms based on different implementations of the same analytical method might be expected to yield correlated errors in their respective labels. Incorporating a correlation measure into the annotator model could therefore allow for a better aggregation of the inferred ground truth. Annotators who are considered to be anomalous (i.e., those that are highly correlated to other annotators but which have relatively large variances and biases) should be penalised with lower weighting for their labels; conversely, expert annotators (i.e., those that are highly correlated to other annotators but which have relatively small variances and biases) should have their labels weighted more heavily in the model. A generative framework for modelling BCLA with correlated annotators (denoted BCLAc) is now described.

A multivariate normal distribution[3] can be applied to the annotator model, using the covariance matrix (denoted $\mathbf{\Sigma}$) to describe the correlation among annotators, as well as providing a constraint on the biases $\boldsymbol{\phi}$. The Inverse-Wishart (IW) distribution[4] is used as a prior for the covariance matrix $\mathbf{\Sigma}$, and the bias values $\boldsymbol{\phi}$ for all annotators are modeled using a multivariate normal distribution with mean $\boldsymbol{\mu}_{\phi\Sigma}$ and covariance $\mathbf{\Sigma}/k_0$. The graphical representation of BCLA with correlated annotators (denoted BCLAc) is shown in Fig. 2.

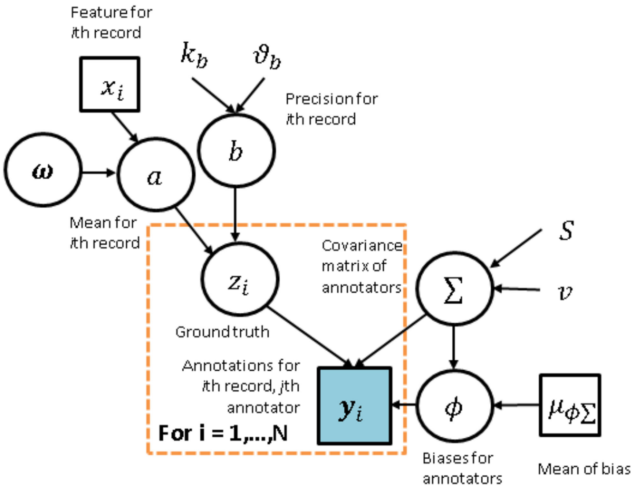


Fig. 2. Graphical representation of the BCLAc model: $\mathbf{y}_i$ corresponds to the annotations provided by all annotators for the ith record, and it is modeled by the z_i (the unknown underlying ground truth), the $\boldsymbol{\phi}$ (biases), and the $\mathbf{\Sigma}$ (covariance matrix). Furthermore, z_i is drawn from a Gaussian distribution with parameters mean a and variance $1/b$, where a for the ith record can be a function of feature vector $\mathbf{x}_i$ and coefficients $\mathbf{w}$. The biases, $\boldsymbol{\phi}$ are modeled from a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_{\phi\Sigma}$ and covariance $\mathbf{\Sigma}$ over a scalar factor k_0. The $\mathbf{\Sigma}$ is drawn from an Inverse-Wishart (denoted IW) distribution with parameters v and $\mathbf{S}$. The b is drawn from a gamma distribution (denoted as Gamma) with parameters k_b and ϑ_b.

As illustrated in Fig. 2, only the annotator model has been modified in BCLAc when compared to the BCLA model, by introducing the covariance measure among annotators. Assuming that each record is independent, the conditional pdf of the modified annotator model with covariance becomes the following:

$$p\left(\mathbf{y} \mid z_i, \boldsymbol{\phi}, \mathbf{\Sigma}\right) = \prod_{i=1}^{N} \mathcal{N}\left(z_i + \boldsymbol{\phi}, \mathbf{\Sigma}\right), \tag{11}$$

where $\mathbf{\Sigma}$ is the covariance matrix of the R annotators and where there are N recordings.

Matrix $\mathbf{\Sigma}$ can be further decomposed into a correlation matrix and the precision values of the annotators. Using the separation strategy proposed by Barnard *et al.* [12], $\mathbf{\Sigma}$ is formulated as:

$$\mathbf{\Sigma} = \mathbf{Q}\boldsymbol{\rho}\mathbf{Q}, \tag{12}$$

where $\mathbf{Q}$ is an R-by-R diagonal matrix with entries being $\frac{1}{\sqrt{\lambda^{j=1}}}, \ldots, \frac{1}{\sqrt{\lambda^{j=R}}}$. Here, λ^j is the precision value for the jth annotator, and $\boldsymbol{\rho}$ is the correlation matrix of the annotation errors among R annotators. The $\boldsymbol{\rho}$ matrix measures the Pearson product-moment correlation coefficients [13], where each element $\rho_{ij} \in [-1, 1]$. The correlation coefficient ρ_{pq} of two sets of N annotations (i.e., $\mathbf{y}^p$ and $\mathbf{y}^q$), each provided by annotator p and q, can be written as:

$$\rho_{pq} = \frac{N\sum_{i=1}^{N} y_i^p y_i^q - \sum_{i=1}^{N} y_i^p \sum_{i=1}^{N} y_i^q}{\left[N\sum_{i=1}^{N} y_i^{p\,2} - \left(\sum_{i=1}^{N} y_i^p\right)^2\right]^{\frac{1}{2}} \left[N\sum_{i=1}^{N} y_i^{q\,2} - \left(\sum_{i=1}^{N} y_i^q\right)^2\right]^{\frac{1}{2}}}. \tag{13}$$

This coefficient $\rho_{pg} = 0$ when both annotators' errors are independent from each other, whereas having $\rho_{pq} \in [-1, 0)$ or $(0, 1]$ indicates that annotators' errors are negatively or positively correlated in a linear manner, respectively (i.e., their errors decrease or increase together throughout recordings).

The biases of individual annotators are now assumed to be drawn from a multivariate normal distribution constrained by $\mathbf{\Sigma}$, with conditional probability density:

$$p\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}, \mathbf{\Sigma}\right) = \mathcal{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}, \mathbf{\Sigma}/k_0\right), \tag{14}$$

where $\boldsymbol{\mu}_{\phi\Sigma}$ is the prior mean for $\boldsymbol{\phi}$, and k_0 is a positive scalar that expresses our belief on $\boldsymbol{\mu}_{\phi\Sigma}$.

The posterior of the parameter $\boldsymbol{\theta}_c = \{\boldsymbol{\phi}, \mathbf{\Sigma}, b, z_i\}$ for a given dataset $\mathbf{D}$ can be written using Bayes' theorem as:

$$p\left(\boldsymbol{\theta}_c \mid \mathbf{D}\right) = \frac{p\left(\mathbf{D} \mid \boldsymbol{\theta}_c\right) p\left(\boldsymbol{\theta}_c\right)}{\int_{\boldsymbol{\theta}_c} p\left(\mathbf{D} \mid \boldsymbol{\theta}_c\right) p\left(\theta\right) d\boldsymbol{\theta}_c}, \tag{15}$$

[3] The probability density function of the d-dimensional multivariate normal distribution can be defined as $\mathcal{N}(z \mid \boldsymbol{\mu}, \mathbf{\Sigma}) = (2\pi)^d |\mathbf{\Sigma}|)^{(-1/2)} \exp(-\frac{1}{2}(z - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(z - \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is 1-by-d and $\mathbf{\Sigma}$ is a d-by-d symmetric positive-definite matrix.

[4] The probability density function of a Inverse-Wishart distribution for d-by-d symmetric positive-definite matrices X and T, and where v as a scalar greater than or equal to d, is $\frac{|\mathbf{S}|^{v/2}|X|^{-(v+d+1)/2}\exp(-\frac{1}{2}\mathrm{trace}(\mathbf{S}X^{-1}))}{2^{vd/2}\Gamma_d(v/2)}$, where $\Gamma_d(v/2) = \pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma(\frac{v+1-i}{2})$ is a multivariate gamma function.

where

$$p\left(\mathbf{D} \mid \boldsymbol{\theta}_c\right) p\left(\boldsymbol{\theta}_c\right) = \mathcal{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}, \boldsymbol{\Sigma}/k_0\right) \mathrm{IW}\left(\boldsymbol{\Sigma} \mid v, S\right)$$
$$\times \mathrm{Gamma}\left(b \mid k_b, \vartheta_b\right)$$
$$\times \left[\prod_{i=1}^{N} \mathcal{N}\left(z_i \mid a_i, 1/b\right) \mathcal{N}\left(\mathbf{y_i} \mid z_i + \boldsymbol{\phi}, \boldsymbol{\Sigma}\right)\right].$$

The Gibbs sampler can be used to estimate the parameter of the covariance matrix $\boldsymbol{\Sigma}$ directly, without modelling the precision and correlation individually. The ground truth z_i can be estimated using the precision values derived from the estimated $\boldsymbol{\Sigma}$. The posterior of biases $\boldsymbol{\phi}$ and covariance $\boldsymbol{\Sigma}$ can be modeled jointly using a conjugate prior defined by the multivariate normal inverse-Wishart distribution. Each parameter in the likelihood described in equation (15) can therefore be updated by sampling from its conjugate prior distribution with posterior hyperparameters (denoted by *) as follows:

$$z_i \sim \mathcal{N}\left(z_i \mid a_i^*, \frac{1}{b_i^*}\right), \qquad \boldsymbol{\phi} \sim \mathcal{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}^*, \boldsymbol{\Sigma}_\phi^*\right),$$
$$b \sim \mathrm{Gamma}\left(b \mid k_b^*, \vartheta_b^*\right), \qquad \boldsymbol{\Sigma} \sim \mathrm{IW}\left(\boldsymbol{\Sigma} \mid v^*, \mathbf{S}^*\right).$$

where

$$a_i^* = \frac{\left(\mathbf{x}_i^{\mathsf{T}}\mathbf{w}\right) b + \sum_{j\in V_i}\left[\left(y_i^j - \phi^j\right)\lambda^j\right]}{b + \sum_{j\in V_i}\lambda^j}, \quad b_i^* = b + \sum_{j\in V_i}\lambda^j.$$

$$\boldsymbol{\mu}_{\phi\Sigma}^* = \frac{k_0\boldsymbol{\mu}_{\phi\Sigma}}{k_0 + N} + \frac{\mathbf{U}\bar{\mathbf{y}}_b}{k_0 + \mathbf{U}}, \quad \boldsymbol{\Sigma}_\phi^* = \frac{\boldsymbol{\Sigma}}{k_0 + N}.$$

$$k_b^* = k_b + \frac{N}{2}, \quad \frac{1}{\vartheta_b^*} = \frac{\sum_{i=1}^{N}\left(z_i - \bar{z}\right)^2}{2} + \frac{1}{\vartheta_b}, \quad v^* = v + N,$$

$$\mathbf{S}^* = \mathbf{S} + \sum_{i=1}^{N}\left(\mathbf{y}_i - z_i - \bar{\mathbf{y}}_b\right)^T\left(\mathbf{y}_i - z_i - \bar{\mathbf{y}}_b\right)$$
$$+ \frac{k_0 N}{k_0 + N}\left(\bar{\mathbf{y}}_b - \boldsymbol{\mu}_{\phi\Sigma}\right)^T\left(\bar{\mathbf{y}}_b - \boldsymbol{\mu}_{\phi\Sigma}\right).$$

We recall that V_i is the set of annotators that provided annotations for the ith record. $\mathbf{U}$ is a 1-by-R vector, and each of its elements indicates the total number of annotations provided by a respective annotator, excluding missing annotations. $\bar{\mathbf{y}}_b = [\bar{y}_b^{j=1}, \cdots, \bar{y}_b^{j=R}]$, where $\bar{y}_b^j = \frac{1}{N_j}\sum_{i=1}^{N}(y_i^j - z_i)$, is the sample mean difference between the inferred ground truth and annotations across N_j recordings provided by the jth annotator (excluding records with missing annotations). $\boldsymbol{\mu}_{\phi\Sigma}$ is the prior mean for $\boldsymbol{\phi}$, and k_0 defines the belief on this prior mean. $\mathbf{S}$ is proportional to the prior mean for $\boldsymbol{\Sigma}$, and v defines our belief concerning $\mathbf{S}$. v also must satisfy the condition that $v > R - 1$. The precision values, $\boldsymbol{\lambda}$ for R annotators, can be estimated as being $[\mathrm{diag}\left(\boldsymbol{\Sigma}\right)]^{-1}$. $\mathbf{w}$ can be learnt from the complete data log-likelihood as described before. After convergence of the Gibbs sampler, the value of each parameter can be approximated by calculating the mean. See Supplementary Materials for details concerning the implementation of BCLAc.

V. Data Description and Methodology Evaluation

A. Data Description

This section introduces datasets from exemplar clinical applications involving the potential personalisation of patient care using health sensor data, and which would directly benefit from the improvements of robustness offered by the work described in this paper.

1) Simulated QT Datasets With Correlated Annotators: A total of 1,000 simulated patient records were generated, each having five annotators with correlated QT interval (i.e., the distance between the beginning of a Q wave and the end of a T wave on an electrocardiogram) annotations. The number of annotators and annotations are selected here for the purpose of demonstration. Application to other datasets using different numbers of annotators and annotations will be described later. The simulated dataset was generated using the following parameter values: the *true* annotation for each record, $z_i \sim \mathcal{N}(400, 40)$, which has a mean, $a = 400$ ms and a standard deviation $b^{-1/2} = 40$ ms. The gold standard of the simulated QT dataset was defined following the ICH E14 clinical guidelines [14]. No additional features were considered in this simulation (i.e., $x_i = 1$) as it was solely used to investigate the performance of the model in estimating correlation between annotators, but an intercept was assumed for $f(\mathbf{w}, \mathbf{x})$. Furthermore, it was assumed that $\alpha_\phi \sim \mathrm{Gamma}(3, 0.0005)$, and that $b \sim \mathrm{Gamma}(3, 0.0002)$. The simulated annotations of the five annotators for a particular record were $\mathbf{y}_i \sim \mathcal{N}\left(z + \boldsymbol{\phi}, \boldsymbol{\Sigma}\right)$. The $\boldsymbol{\Sigma} \sim \mathrm{IW}(\tau, 5)$ for five annotators, where τ is assumed to be a diagonal matrix with diagonal elements being the expected mean of the $\mathrm{Gamma}(4, 0.0003)$; their biases, $\boldsymbol{\phi} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\phi\Sigma}, \boldsymbol{\Sigma}/k_0\right)$, has a 0 ms mean vector and a covariance $\boldsymbol{\Sigma}/k_0$. The k_0 describes the confidence of the mean $\boldsymbol{\mu}_{\phi\Sigma}$ of the distribution of biases $\boldsymbol{\phi}$ and was set to $k_0 = 1$ for the purposes of illustration. The *true* precision values of individual annotators, $\boldsymbol{\lambda}$, can be estimated by finding the inverse of the values in the diagonal elements of $\boldsymbol{\Sigma}$ where each of their correlations is $\rho = 1$. The correlation matrix of the annotation errors is then obtained through decomposing the $\boldsymbol{\Sigma}$ matrix, and the correlation of a pair of annotators (e.g., a_1 and a_2) can be estimated as follows:

$$\rho_{a_1,a_2} = \frac{\Sigma_{a_1,a_2}}{1/\sqrt{\left(\lambda_{a_1}\lambda_{a_2}\right)}}. \tag{16}$$

The estimated correlations of errors of BCLAc for five annotators are shown in Fig. 3(a). It differs from the correlation matrix derived directly from the annotations provided by these annotators as shown in Fig. 3(b). As shown in Fig. 3(b), annotations from all five annotators are positively correlated with $\boldsymbol{\rho} > 0$. This phenomenon can be explained by the fact that the correlation measures a change in trend (i.e., an increase or decrease of two variables together, or an increase and decrease inversely) for two sets of annotations, taken from a pair of annotators. As long as the annotations both increase similarly, the correlation coefficients would therefore be positive. However, the true correlation of errors among annotators is not based on the *absolute* variation of values about their respective means; instead, it is an

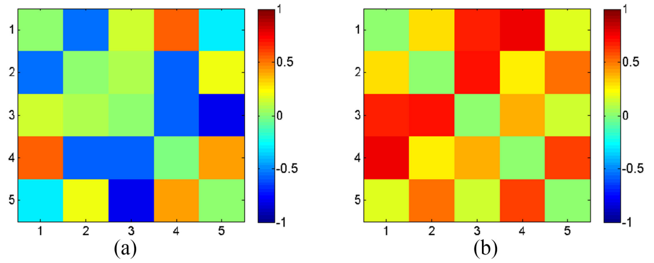


Fig. 3. Correlation of errors of a simulated dataset: (a) The estimated correlation matrix of errors from five annotators using equation (16). (b) The correlation matrix derived directly from the annotations provided by these annotators. Note that each correlated matrix is subtracted from an identity matrix to remove the correlation of an annotator with itself.

estimation of correlation obtained from the *difference* between the annotations provided and the underlying ground truth. That is, Fig. 3(a) captures this variation of labels around the latent ground truth, which is different to the variation of the absolute values of labels about their means, where the latter is shown in Fig. 3(b).

To test the robustness of the BCLAc-Gibbs approach, it was applied to six simulated QT datasets generated from the BCLAc model. Simulated datasets of 100, 500, and 1000 records were generated with parameter values described above, each with (i) five and (ii) 10 correlated QT interval annotations from corresponding annotators.

2) Publicly-Available QT Dataset: We also used the 2006 PhysioNet "Computing in Cardiology Challenge" QT dataset [15]: 548 ECG records were provided with 38,621 annotations of the QT interval sourced from: 20 human annotators in Division 1 of the dataset; 48 automated algorithms in Division 2; and 21 automated algorithms in Division 3. An additional division (Division 4) was defined as being the union of the labels from all automated algorithms from Divisions 2 and 3. Division 1 was used in the competition to generate the reference annotations, and so we therefore focused on the analysis of the sets of automated labels (i.e., Divisions 2, 3, and 4). The records were obtained from 290 subjects (209 men with mean age of 55.5 and 81 women with mean age of 61.6), each represented by between one and five recordings. About 20% of the subjects were healthy controls, and the rest of subjects had a variety of ECG morphologies with QT intervals ranging from 256 ms to 529ms. Diagnostic classifications are detailed elsewhere [16].

Not all annotators provided annotations for all records, and where a minimum of 33% of the annotators labelled any single recording. For the work described in this paper, only those annotators that had provided annotations for more than 50% of records were used, which thereby corresponds to 40 annotators in Division 2, 15 annotators in Division 3, and 55 annotators in Division 4.

3) Capnobase Respiratory Rate (RR) Dataset: The Capnobase dataset [17] was collected from subjects undergoing elective surgery and routine anaesthesia. It consists of photoplethysmogram (PPG) recordings from a pulse oximeter and capnometry data ($F_s = 300$ Hz), from 59 children (median age: 9, range: 1–17 years) and 35 adults (median age: 52, range: 26–76 years). We used the set as described in [18], which has 42 recordings of 8-minutes duration (336 minutes in total) containing reliable recordings of spontaneous breathing or controlled ventilation. The capnometric waveform was used as the reference for RR estimates derived from the PPG (we note that estimating RR from a pulse oximeter is an important application of physiological monitoring using wearable devices) [18].

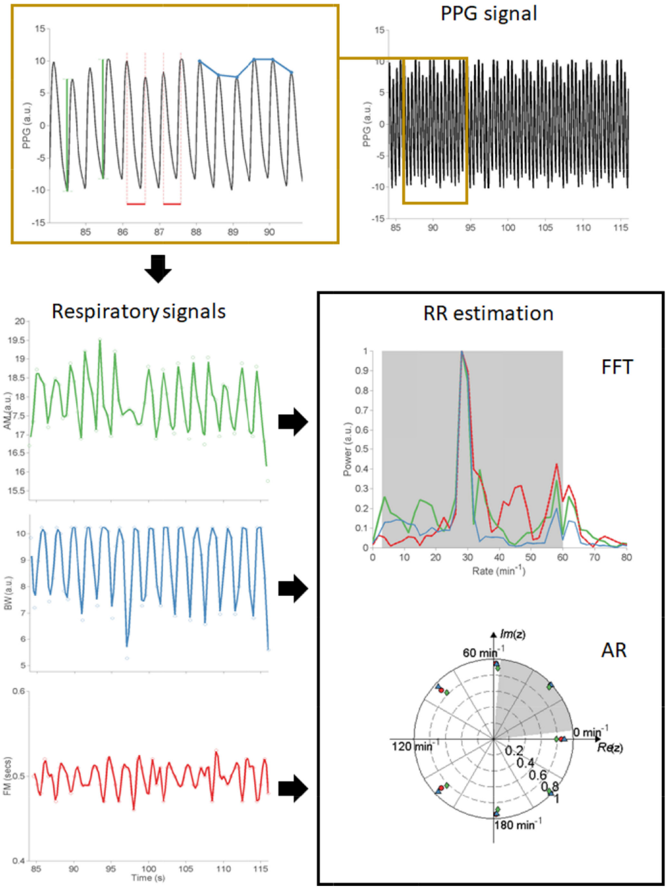


Fig. 4. Example of a 32-second PPG window used for RR estimation. AM (green), BW (blue), and FM (red) respiratory modulations are extracted; for the FFT-based method, the power spectrum is calculated for each modulation using FFT, and the maximum power is selected within the physiologically-plausible RR range (grey area); for the AR-based method, the poles for each modulation are determined using an AR model, and the dominant pole within the plausible range of RR (grey area) is selected. Our proposed models are then used to combine estimates, and provide a final "fused" value for that window.

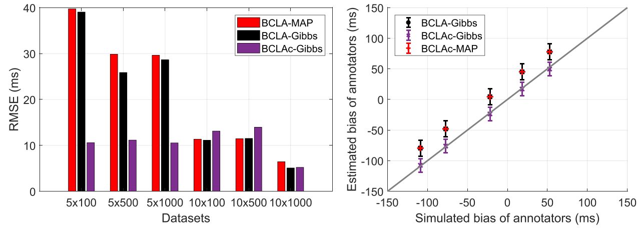
RR was computed for 32-second windows, with successive windows having 29 s overlap. To extract the three respiratory-induced modulations (AM, BW, FM), beat detection was performed on the PPG using a segmentation algorithm [19]. The latter produces a series of maximum and minimum intensities for each pulse. As shown in Figure 4, the series of maximum intensities of the PPG pulses was used for extracting the BW timeseries. The (max-min) amplitude was used to derive the AM timeseries. The intervals between successive beats were used to extract the FM timeseries.

RR was estimated using two different spectral approaches that have been used in the literature: Fourier analysis (FFT) and autoregressive (AR) modelling. Spectral analysis requires evenly-sampled data, and so each timeseries (corresponding to BW, AM and FM) was first re-sampled at 4 Hz using linear interpolation. The frequency spectra of the resulting respiratory signals were calculated. The frequency at which the maximum intensity of each spectrum is obtained within the frequency range of interest (corresponding to 3 to 60 beats-per-minute, or bpm), was taken as corresponding to the respiratory frequency (Fig. 4). For the AR method, an AR model of order 7 was fitted to each timeseries. The respiratory frequency was identified as that corresponding to the pole with the greatest magnitude within the plausible range of frequencies for respiration. We note that for each window, for each approach (FFT and AR), three RR estimates were determined (i.e., on each from the BW, AM, and FM signals).

B. Methodology Evaluation

1) Simulated and Real QT Datasets: The precision values $\boldsymbol{\lambda}$ and biases $\boldsymbol{\phi}$ inferred by BCLA-Gibbs and BCLAc-Gibbs were compared with those estimated using BCLA-MAP. For the estimation of the ground truth, the root-mean-squared-error (RMSE) was calculated using the "gold standard" reference provided. Results were additionally compared with the following methodologies, which represent the existing state-of-the-art: (i) an EM-based formulation proposed by Raykar *et al.* [6] (denoted as EM-R); (ii) *Scalar Simultaneous Truth and Performance Level Estimation* (denoted "sSTAPLE") proposed by Warfield *et al.* [3]; (iii) Mean voting between the annotations provided by the "experts"; (iv) Median voting between the annotations as above. To enable us to assess the performance of the proposed models as a function of the number of annotators, a random number of annotators was selected 500 times. This was repeated with the number of the annotators being $R = \{3, 5, 7, 9\}$ in Division 4. The minimum number of annotators was chosen to be $R = 3$ to allow for obtaining results via the median voting approach. A Friedman test ($p < 0.01$) was applied to the bootstrapped RMSEs, to provide a comparison between the various methods.

2) RR Dataset: In the context of personalised care, RRs were inferred for each subject individually over all windows. The mean-absolute-error (MAE) of the inferred RR estimates from the PPG across all subjects using the proposed fusion frameworks, BCLA and BCLAc, were compared to the "gold standard" reference RR values from capnography. The BCLA and BCLAc models were applied to the RR estimates extracted using the conventional FFT-and AR-based algorithms outlined earlier [20] for each individual subject. Additionally, the performance of BCLA-Gibbs and BCLAc-Gibbs were compared to that of "Smart Fusion" (i.e., the benchmarking algorithm in this field proposed by Karlen *et al.* [18]), and with the best-performing (lowest-MAE) single algorithm (denoted "Best"), EM-R, sSTAPLE, BCLA-MAP, and also the traditional naïve mean and median voting approaches. To further analyse the performance of different voting algorithms with varied signal qualities, a comparison was made in which we compute the MAE results of (i) RR estimates for all windows; (ii) only RR estimates for those windows following the criteria considered by "Smart Fusion" (denoted $*$)[5]; (iii) only RR estimates for noisy windows (i.e., those which are rejected by "Smart Fusion") were considered. We note that the sets of windows (i) = (ii) $\bigcup$ (iii).



(a) (b)

Fig. 5. (a) Plots of RMSE using different strategies for simulated datasets. In each dataset, the number of annotators and number of annotations are indicated on the left and right of "×" respectively (i.e., we show $R \times N$). (b) A comparison of the simulated and inferred bias values of $R = 5$ annotators in the dataset of 5×1000. The diagonal line indicates a perfect match between the simulated and estimated results. The results are plotted with one standard deviation from the mean.

VI. Results and Discussion

A. Simulated QT Datasets

BCLA-Gibbs took approximately 350 s to generate 10^4 draws of the 2,500 annotations created from $R = 5$ annotators for $N = 500$ recordings using MATLAB R2011a on a 3.3 GHz Intel Xeon processor. In comparison, BCLAc-Gibbs took approximately 404 s to run the same number of draws. Half of the samples were discarded as burn-in.

When $R = 5$ annotators were considered, RMSE results as shown in Fig. 5(a) varied depending on how the correlation of errors among annotators was generated. In the simulated datasets as described earlier, we assumed that the uncertainty of the bias values $\boldsymbol{\phi}$ was controlled by the covariance matrix $\boldsymbol{\Sigma}$. In comparison, the BCLA-Gibbs approach assumed the bias values $\boldsymbol{\phi}$ were modeled independently from the covariance matrix $\boldsymbol{\Sigma}$. It was therefore expected that BCLA-Gibbs would provide less reliable estimation of the true bias values $\boldsymbol{\phi}$. The BCLA-Gibbs approach performed consistently worse than BCLAc-Gibbs as it over-estimated the bias values $\boldsymbol{\phi}$ of the annotators constantly, as shown in Fig. 5(b). However when there were $R = 10$ annotators, the BCLA-Gibbs approach was sufficient to provide reliable estimation without introducing the correlation of errors among annotators. In comparison, results obtained with BCLA-MAP were slightly worse than those obtained for BCLA-Gibbs when estimating RMSEs; BCLA-MAP also provides less reliable results when estimating the bias values than BCLA-Gibbs. In contrast to BCLA-MAP, the BCLA-Gibbs model not only provides estimates but also produces confidence in its estimation – this is a key advantage of fully-Bayesian inference.

[5] i.e., those windows where RR estimates with one standard deviation exceeding four bpm were discarded.

TABLE I
THE PARAMETERS OF BCLA AND BCLAc FOR MODELLING THE CAPNOBASE RR AND QT DATASETS

Symbol	Definition	Value	
		RR	**QT**
k_b	shape of Gamma distribution for b	3	3 *
ϑ_b	scale of Gamma distribution for b	0.006	0.0002 *
S	the scale matrix of Inverse-Wishart distribution for $\mathbf{\Sigma}$	$\frac{1}{\lambda_M}I$	
v	the degrees of freedom of Inverse-Wishart distribution for $\mathbf{\Sigma}$	R	
k_0	a scale factor of $\mathbf{\Sigma}$	10	
$\boldsymbol{\mu}_{\phi\Sigma}$	mean vector of the bias distribution	0	10 $^\diamond$
μ_ϕ	mean of the bias distribution	variable †	10 $^\diamond$
k_λ	shape of Gamma distribution for $\boldsymbol{\lambda}$	3 ‡	4 *
ϑ_λ	scale of Gamma distribution for $\boldsymbol{\lambda}$	0.02 ‡	0.003 *
k_α	shape of Gamma distribution for α_ϕ	5	3 $^\diamond$
ϑ_α	scale of Gamma distribution for α_ϕ	0.1	0.0005 $^\diamond$

TABLE II
RMSEs (ms) OF THE INFERRED LABELS USING DIFFERENT STRATEGIES IN THE QT DATASET

Division (number)	**Mean**	**Median**	**EM-R**	**sSTAPLE**	**BCLA-**		**BCLAc-**
					MAP	**Gibbs**	**Gibbs**
2 (40)	16.38	15.52	15.40	15.16	12.71	12.06	11.95
3 (15)	32.14	20.55	24.64	20.54	17.41	16.20	16.16
4 (55)	18.28	15.10	16.85	15.13	12.65	11.66	11.66

B. QT Dataset

BCLA-Gibbs took approximately 152 s to generate 5,000 draws of the 548 annotations, each provided from $R = 5$ algorithms using the same system as before. With the same dataset, BCLAc-Gibbs took approximately 70 s to run 2,000 draws until convergence. When fusing a large number of algorithms (e.g., $R = 55$ with $N = 26,922$ annotations), BCLA-Gibbs required 5,000 draws and took about 500 s, whereas BCLAc-Gibbs ran for 290 s to compute 3,000 draws. Both methods discarded the first half of the samples as burn-in, as before. The resulting values of the parameters and hyperparameters of the BCLA and BCLAc models for the QT dataset are shown in Table I.

In the table above: b is the precision for the estimate of the ground truth. $\mathbf{\Sigma}$ is the covariance matrix. $\frac{1}{\lambda_M}I$ is an R × R diagonal matrix with entries obtained from a scalar, λ_M, as the inverse of the mean of the Gamma$(k_\lambda, \vartheta_\lambda)$. $\boldsymbol{\lambda}$ refers to annotator/signal-specific precisions. For the BCLA-Gibbs approach only: α_ϕ is the precision for the estimate of the bias from ground truth. RR dataset: the values with ‡ are determined with the assumption that the RR estimates provided by the best modulation signal are ±2 bpm away from the reference [21], [22]. The values with † are estimated from the median RR estimates provided by the algorithms. QT dataset: the values with * are determined with the assumption that the annotations provided by the best performing algorithm is ±5 ms away from the reference [14]. The values with $^\diamond$ are derived from [23]–[25]. The values with * are derived from [26]–[28].

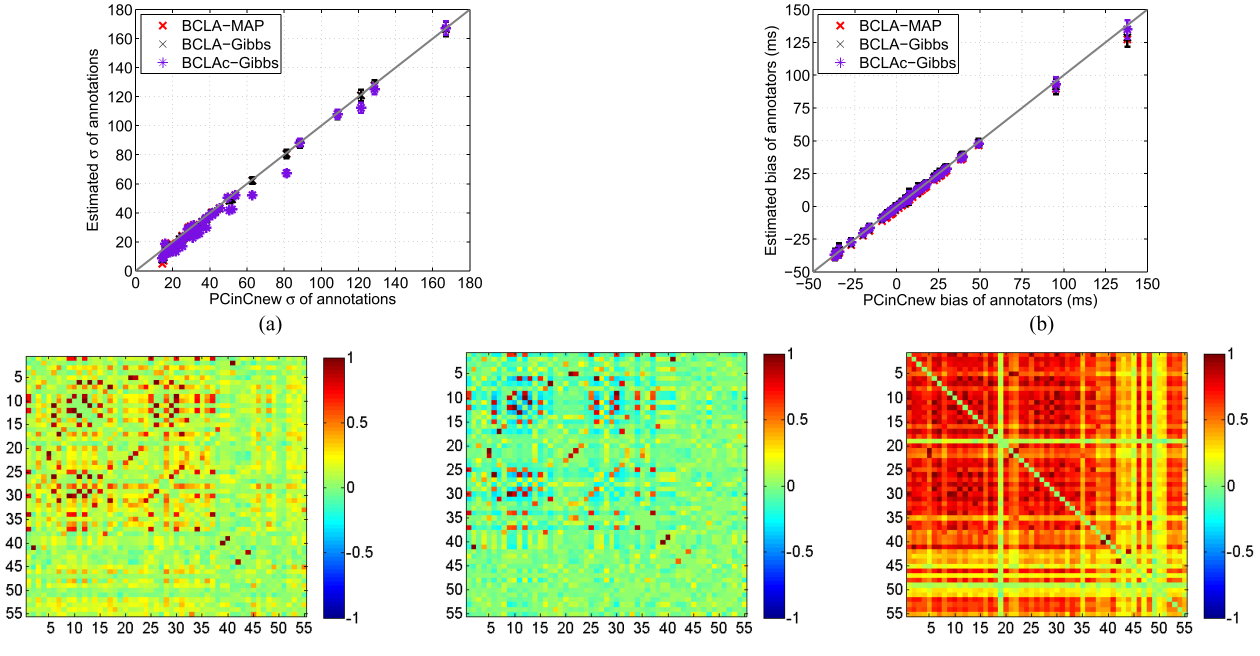
Both the BCLA and BCLAc methods produced accurate estimation of the bias values for Division 2 and 4, an example of Division 4 is shown in Fig. 6(b). More results are demonstrated in the Supplementary Materials. In the case of Division 3 (see Fig. 6(b) in Supplementary Materials), BCLAc-Gibbs produced more accurate estimation of the bias values in comparison to those computed using BCLA-Gibbs and BCLA-MAP. However for σ prediction, the BCLA-Gibbs and BCLA-MAP methods were more reliable than BCLAc-Gibbs. This might be due to the fact that both the correlation of errors and the precision values were jointly modeled as a whole using one distribution (i.e., the IW distribution) in the BCLAc-Gibbs model, limiting its accuracy, where one small imperfect estimation of the correlation of errors would directly affect the estimation of the precision values. In comparison to BCLA-MAP, the BCLA-Gibbs and BCLAc-Gibbs models provide accurate estimates along with estimations of confidence in its result – this is a key advantage of the fully-Bayesian inference methods proposed in this paper.

A further advantage of the BCLAc model is its ability to measure the correlation of errors among algorithms (experts). An example of the inferred correlation of errors using BCLAc-Gibbs is shown in Fig. 6(c) for $R = 55$ algorithms in Division 4. Although the exact coefficient values were not fully recovered, BCLAc-Gibbs was able to identify the key relationship of correlation in annotation between algorithms, while a direct estimation of the correlation of annotations failed to do so. When comparing the estimation of the ground truth, the resulting RMSEs are given in Table II, where it may be seen that BCLAc-Gibbs produced the smallest errors, effectively outperforming all other voting strategies.

Fig. 7 shows a further evaluation of the accuracies (in terms of RMSE) of different voting strategies as a function of the number of annotators. The results were generated by sub-sampling (i.e., bootstrap with replacement) annotators 500 times with $R = \{3, \cdots, 9\}$ annotators selected from Division 4. A non-parametric Friedman test was conducted to compare all methods and rendered a chi-squared value of 253.27, 348.84, 427.79, and 431.68 for selecting $R = \{3, 5, 7, 9\}$ annotators, respectively, which were significant ($p < 0.01$). We further performed a post-hoc test using Dunn & Sidàk's Approach [29] to determine if the mean RMSEs of BCLA-Gibbs were significantly different from other methods; no significant difference ($p > 0.01$) was observed. Nevertheless, BLCA-Gibbs outperformed all other voting strategies with the least mean or median RMSEs when number of annotators varied from three to nine. RMSEs of 26.64 ± 10.32 ms, 21.10 ± 5.64 ms, 19.26 ± 4.83 ms, 18.34 ± 4.03 ms were obtained for selecting $R = \{3, 5, 7, 9\}$ annotators, respectively).

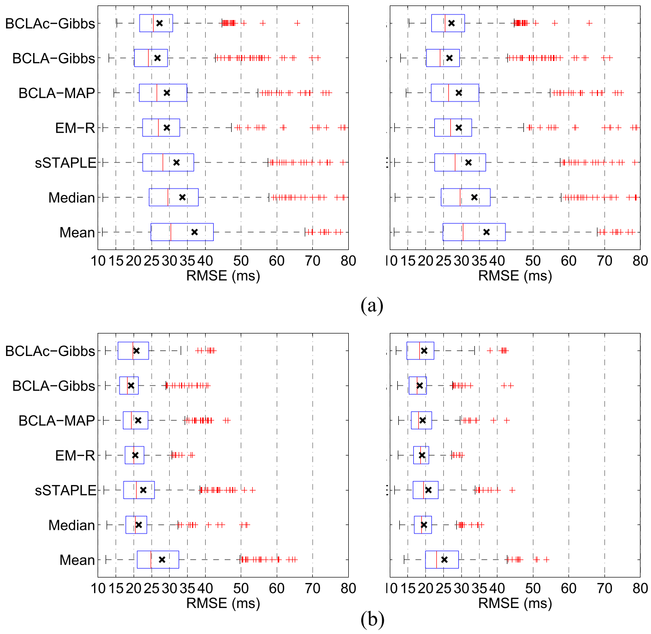
Inspecting the performance of BCLAc-Gibbs, it produced smaller RMSEs when compared with the BCLA-MAP approach when $R < 9$: 27.22 ± 8.34 ms versus 29.26 ± 11.41 ms, 22.49 ± 6.28 ms versus 23.73 ± 7.67 ms, 20.77 ± 6.15 ms versus 21.24 ± 6.26 ms, and 19.58 ± 5.87 ms versus 19.15 ± 4.40 ms for $R = \{3, 5, 7, 9\}$ annotators, respectively. Although the performance of BCLAc-Gibbs was sub-optimal when compared to BCLA-Gibbs, it only required approximately half of the sampling time, and it provided an additional modelling of the correlation of errors among annotators.

When working with incomplete data, our proposed models assume that annotations are missing at random given the observed data, as there was no direct correlated relationship found between MAEs in QT interval estimation and the length of the QT intervals. Future work will focus on simulating



(a) (b) (c)

Fig. 6. A comparison of the QT reference and inferred (a) σ and (b) bias of each annotator for Division 4. The precision can be estimated by taking $1/(\sigma)^2$. The diagonal (grey) line indicates a perfect match between the reference and estimated results. The results are plotted with one standard deviation from the mean for BCLA-Gibbs and BCLAc-Gibbs. (c) A comparison of the reference and inferred correlation of errors among 55 algorithms for Division 4 in the QT dataset: in each row, the reference $\boldsymbol{\rho}$ (left), is compared with its inferred values using BCLAc-Gibbs (middle), and the correlation estimated directly from the annotations provided (right). Note that each correlated matrix is subtracted from an identity matrix to remove the correlation of an algorithm with itself.



(a)

(b)

Fig. 7. Plots of the RMSE results of using different voting approaches, shown as a function of the number of automated annotators. Each plot was generated by randomly sampling the annotators 500 times. The median RMSE is indicated in red '-' within each box, while the black '×' indicates the mean RMSE. The span of each box represents the interquartile range. The outliers are shown as red '+'. (a) Selection of three (left) and five (right) annotators. (b) Selection of seven (left) and nine (right) annotators.

various degrees of incompleteness and explore the performance of our models with percentage of missing annotations.

C. RR Dataset

BCLA-Gibbs took approximately 38 s to generate 5,000 draws of the $N = 900$ annotations provided from $R = 6$ algorithms. Similarly, BCLAc-Gibbs took approximately 68 s, and the first 3,000 samples was discarded as burn-in for both methods. The average time for 5,000 iterations using BCLA-MAP was under 2 s, similar to EM-R and sSTAPLE. The resulting values of the parameters and hyperparameters of the BCLA and BCLAc models for the Capnobase dataset were described in Table I.

Fig. 8 shows MAE results across 42 subjects. In the case where all windows or only those with standard deviation greater than four bpm were considered, the BCLA and BCLAc methods outperformed the single best-performing algorithm with least MAE across subjects (i.e., BA in the table). In the case when discarding windows with standard deviation greater than four bpm (with 55.4% of the windows thereby remaining), the proposed BCLA and BCLAc methods outperformed "Smart Fusion" but were slightly worse than the BA*.

Furthermore, the results show that BCLAc-Gibbs has similar performance when compared to the BCLA approaches, with their mean MAEs being closest to the "theoretical best" (TB) algorithm (i.e., selecting the best algorithm with least MAE per subject which is of course impossible in practice, because it

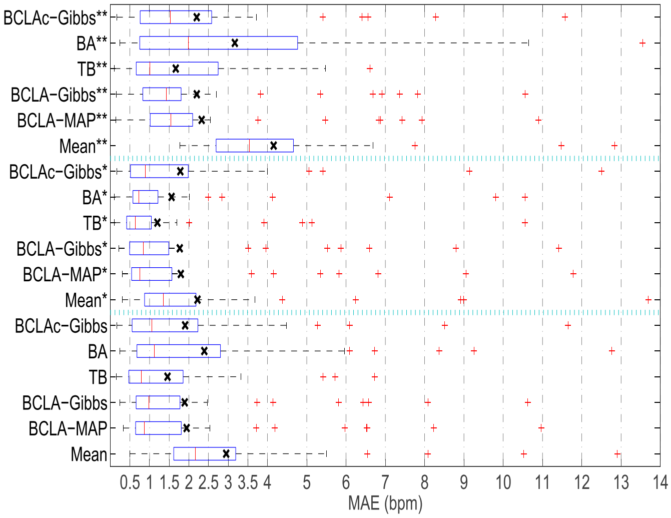
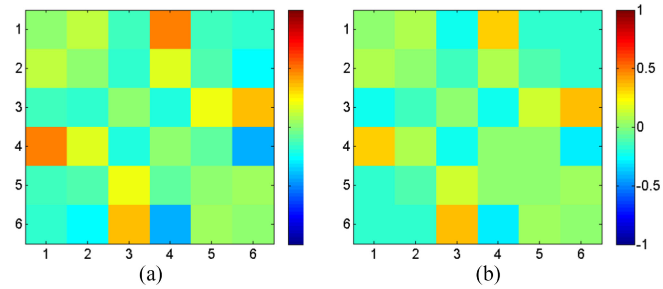


Fig. 8. Box plot of MAEs across 42 subjects for different fusion approaches. The median MAEs are indicated in red '-' in each box, while the black '×' indicates the mean MAE. The span of each box indicate the interquartile MAE range over 42 subjects for each fusion method when combining six algorithms. Notation: BA – the single best-performing algorithm with least MAE across all subjects; TB – theoretical best algorithm with least MAE selected per subject; Mean* – the "Smart Fusion". Note that results associated with * indicate the MAEs were derived from windows excluding those have a standard deviation greater than four bpm, while those associated with ** indicate the results derived only from windows that have a standard deviation greater than four bpm.

requires knowledge of the ground truth) for the three different groups. For using all windows, the MAE improved with BCLA-MAP (i.e., smaller error) over "Smart Fusion" for 69.1% of subjects. When windows were excluded if the standard deviation of RR estimates was greater than four, BCLA-MAP has a MAE improved for 81.0% of subjects. Furthermore, the MAE improved with BCLA-Gibbs over "Smart Fusion" for 73.8% of subjects when using all available windows, and 83.3% of subjects for excluding those with a large standard deviation as aforementioned. In addition, BCLA-Gibbs had 61.9% of subjects with smaller MAEs than those of BCLA-MAP using all windows when fusing $R = 6$ algorithms. When considering BCLAc-Gibbs using all windows, it had 17 subjects (40.5%) with smaller MAEs than BCLA-Gibbs, and had 22 and 30 subjects (corresponding to 52.4% and 71.4% of subjects respectively) with smaller MAEs than BCLA-MAP and "Smart Fusion". In comparison to BCLA, our BCLAc-Gibbs model provides correlation of errors among algorithms (see Fig. 9): as BCLAc-Gibbs is a more complex model where it has to learn an additional parameter (i.e., the correlation of errors), a larger set of annotators should be expected to be required to better infer the ground truth.

Our proposed models assume that each window in the latent ground truth model is independent. This potentially limits their applications in area where a label produced by an algorithm has a Markov dependence on the previous labels, such as time dependency between windows. Future work could extend the current ground-truth model to Gaussian process regression, where the window-dependent features can be modeled using a covariance function (i.e., a kernel) to describe either linear or non-linear dependency among windows.



(a)
(b)

Fig. 9. Example of correlation matrix of errors among six algorithms estimated for a subject: (a) computed with reference provided and (b) computed using BCLAc-Gibbs without knowing the reference. Note that each correlated matrix is subtracted from an identity matrix to remove the correlation of an algorithm with itself.

D. Signal Quality Extension

A signal quality index (SQI) can be seen as a measure of task difficulty: noisy records/windows are harder to label due to noise contamination of the signals, while clean records can be seen as easier tasks for labelling. Experts are able to "filter" noise to some degree and provide consistently reliable annotations across data of differing noise levels, whereas non-experts might mistake noise for intrinsic features of the signals. Hence, an independent variable can be introduced into our proposed models that acts as a probabilistic score to better infer the underlying ground truth of a physiological signal.

To demonstrate proof-of-concept, we have included a scaling factor, t_i^j, in the range of (0, 1] as an indication of record-specific and annotator-specific SQI in BCLA-Gibbs (denoted as BCLA-SQI), where y_i^j can be drawn from a normal distribution defined as $\mathcal{N}(y_i^j \mid z_i + \phi^j, (t_i^j \lambda^j)^{-1})$. Different from BCLA-Gibbs, the estimation of z_i and ϕ^j are now dependent on the signal quality t_i^j:

$$z_i \sim \mathcal{N}\left(z_i \mid a_{t_i}^*, \frac{1}{b_{t_i}^*}\right), \quad \phi^j \sim \mathcal{N}\left(\phi^j \mid \mu_{t_\phi}^{j*}, \frac{1}{\alpha_{t_\phi}^{j*}}\right).$$

where

$$a_{t_i}^* = \frac{(\mathbf{x}_i^\mathsf{T}\mathbf{w})\, b + \sum_{j \in V_i} \left[\left(y_i^j - \phi^j\right) \lambda^j t_i^j\right]}{b + \sum_{j \in V_i} \lambda^j t_i^j},$$

$$b_{t_i}^* = b + \sum_{j \in V_i} \lambda^j t_i^j.$$

$$\mu_{t_\phi}^{j*} = \frac{\mu_\phi \alpha_\phi + \sum_{i \in U_j} \left(y_i^j - z_i\right) \lambda^j t_i^j}{\alpha_\phi + \sum_{i \in U_j} \lambda^j t_i^j},$$

$$\alpha_{t_\phi}^{j*} = \alpha_\phi + \sum_{i \in U_j} \lambda^j t_i^j.$$

Our preliminary results on the signal quality extension are shown in the Supplementary Materials.

VII. Conclusion

This paper has proposed two Bayesian generative models for aggregating automated labels to form a consensus where subjective continuous annotations of some presumed underlying ground truth are provided, but where the desired ground truth is not readily available in practice. This is motivated by the need to improve the robustness of methods using biosignals acquired from sensor data, such that the data can be used to support precision medicine.

Simulated and two clinical datasets were considered as exemplars to validate the proposed methods for aggregating the outputs of a group of mixed imperfect automated algorithms in an unsupervised manner. The results of the proposed models had optimal performance over the other comparison voting strategies in all datasets. When incorporating the modelling of potentially-correlated annotators, it was shown that annotators can be grouped based on their correlated decision-making process. For example, we might identify sets of annotators that perform well ("trained experts") from those that perform not well ("novices"). Both proposed models were robust in dealing with missing values, and there is no need for additional pre-processing to discard noisy data. No training data need to be "held out" for optimising the parameters of the models, and no prior knowledge of the performance of each algorithm was given. It is important to note that the increased performance of our models can be explained by their ability to fit the data in a different manner for different individuals (records or subjects). This means the proposed models do not contain fixed parameters (i.e., precision and bias of an algorithm); rather they adapt to the given data from an individual by, for example, assigning a higher weight to a set of labels (which can be the results of one algorithm) that are relevant to that individual; for a different individual, the set of annotations from the same labeller can have a lower weight. This leads to a better understanding of an individuals physiology, which in turn, may lead to better informed decisions for personalised care.

Acknowledgment

This paper is compliant with RCUK "open data" policy because the underlying data are all publicly available datasets.

References

[1] G. D. Clifford and D. Clifton, "Wireless technology in disease management and medicine," *Annu. Rev. Med.*, vol. 63, pp. 479–492, 2012.

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

[3] S. K. Warfield, K. H. Zou, and W. M. Wells, "Validation of image segmentation by estimating rater bias and variance," *Philosoph. Trans. Roy. Soc. A, Math., Physical Eng. Sci.*, vol. 366, pp. 2361–2375, 2008.

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

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

[6] V. C. Raykar *et al.*, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, 2010.

[7] T. Zhu, N. Dunkley, J. Behar, D. A. Clifton, and G. D. Clifford, "Fusing continuous-valued medical labels using a Bayesian model," *Ann. Biomed. Eng.*, vol. 43, no. 12, pp. 2892–2902, 2015.

[8] R. W. Ouyang, L. Kaplan, P. Martin, A. Toniolo, M. Srivastava, and T. J. Norman, "Debiasing crowdsourced quantitative characteristics in local businesses and services," in *Proc. 14th Int. Conf. Inf. Process. Sens. Netw.*, 2015, pp. 190–201.

[9] F. Xing, S. Soleimanifard, J. L. Prince, and B. A. Landman, "Statistical fusion of continuous labels: identification of cardiac landmarks," in *Proc. Int. Soc. Opt. Photon. Med. Imag.*, 2011, pp. 7962–796 206.

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

[11] S. Hara and K. Hayashi, "Making tree ensembles interpretable," arXiv:1606.05390, 2016.

[12] J. Barnard, R. McCulloch, and X. L. Meng, "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage," *Statistica Sinica*, vol. 10, no. 4, pp. 1281–1312, 2000.

[13] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proc. Roy. Soc. London*, vol. 58, pp. 240–242, 1895.

[14] International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, "Guidance for Industry E14: Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non- Antiarrhythmic Drugs," 2014.

[15] G. B. Moody, H. Koch, and U. Steinhoff, "The physionet/Computers in Cardiology Challenge 2006: QT interval measurement," in *Proc. Comput. Cardiol. Conf.*, 2006, pp. 313–316.

[16] A. Schnabel, R. Bousseljot, and D. Kreiseler, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB uber das Internet," *Biomedizinische Technik*, vol. 40, no. 1, pp. 317–318, 1995.

[17] W. Karlen, M. Turner, E. Cooke, G. Dumont, and J. M. Ansermino, "Capnobase: Signal database and tools to collect, share and annotate respiratory signals," in *Proc. Annu. Meeting Soc. Technol. Anesthesia*, 2010, p. 25.

[18] W. Karlen, S. Raman, J. Ansermino, and G. Dumont, "Multiparameter respiratory rate estimation from the photoplethysmogram," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 7, pp. 1946–1953, Jul. 2013.

[19] B. N. Li, M. C. Dong, and M. I. Vai, "On an automatic delineator for arterial blood pressure waveforms," *Biomed. Signal Process. Control*, vol. 5, no. 1, pp. 76–81, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809409000470

[20] P. Charlton, T. Bonnici, L. Tarassenko, D. Clifton, R. Beale, and P. Watkinson, "An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram," *Physiological Meas.*, vol. 37, no. 4, pp. 610–626, Mar. 2016.

[21] A. Breakell and C. Townsend-Rose, "The clinical evaluation of the respicheck mask: A new oxygen mask incorporating a breathing indicator," *Emergency Med. J.*, vol. 18, no. 5, pp. 366–369, 2001.

[22] G. Drummond, A. Bates, J. Mann, and D. Arvind, "Validation of a new non-invasive automatic monitor of respiratory rate for postoperative subjects," *Brit. J. Anaesthesia*, vol. 107, pp. 462–469, 2011.

[23] N. P. Hughes, "Probabalistic models for automated ECG interval analysis," Ph.D. dissertation, Inst. Biomed. Eng., Univ. Oxford, Oxford, U.K., 2006.

[24] J. P. Couderc *et al.*, "Highly automated QT measurement techniques in 7 thorough QT studies implemented under ICH E14 guidelines," *Ann. Noninvasive Electrocardiol.*, vol. 16, no. 1, pp. 13–24, 2011.

[25] W. Andrew *et al.*, "Variability of QT interval measurements in OpioidDependent patients on methadone," *Can. J. Addiction Med.*, vol. 2, pp. 10–16, 2014.

[26] M. Malik, P. Fãrbom, V. Batchvarov, K. Hnatkova, and A. J. Camm, "Relation between QT and RR intervals is highly individual among healthy subjects: Implications for heart rate correction of the QT interval," *Heart*, vol. 87, no. 3, pp. 220–228, 2002.

[27] I. Goldenberg *et al.*, "QT interval: How to measure it and what is "normal"," *J. Cardiovascular Electrophysiol.*, vol. 17, no. 3, pp. 333–336, 2006.

[28] G. D. Clifford, F. Azuaje, and P. E. McSharry, *Advanced Methods and Tools for ECG Analysis* (Series Engineering in Medicine and Biology). Norwood, MA, USA: Artech House, Oct. 2006.

[29] Y. Hochberg and A. C. Tamhane, *Multiple Comparison Procedures*. Hoboken, NJ, USA: Wiley, 2008.