# Patient Similarity Using Population Statistics and Multiple Kernel Learning

**Bryan Conroy** BRYAN.CONROY@PHILIPS.COM
**Minnan Xu-Wilson** MINNAN.XU@PHILIPS.COM
**Asif Rahman** ASIF.RAHMAN@PHILIPS.COM
*Philips Research North America*
*Cambridge, MA*

## Abstract

We present a multiple kernel learning framework to learn similarity functions that compare physiological state between patients. A powerful ensemble kernel is learned from many base kernels evaluated on individual features. Our proposed framework captures two aspects of patient similarity: that patient similarity should be dependent on clinical context and that similarity should be modulated by the frequency and specificity of individual feature values. We validate our model on ICU data to predict hemodynamic instability and present analyses on using the patient similarity function to construct personalized cohorts. Our experiments show that the statistical properties learned by the kernels functions based on feature population distributions are significantly more predictive than naive stationary kernels (e.g. RBFs). Population-based kernels outperform RBF's in identifying patient cohorts based on abnormality of their vitals and lab measurements and at predicting mortality.

## 1. Introduction

The main question addressed in this paper is how to learn useful similarity functions that compare physiological state between patients. This question, often referred to as "patient similarity", has received increased attention in the age of personalized medicine. Specifically, patient similarity aims to identify cohorts based on a set of patient characteristics (like demographics, vitals, labs, medical history, and treatments) to enable applications like case-based comparisons for clinical decision support and to compare treatments across similar cohorts. However, a notorious challenge of patient similarity is that it is not well-defined – features characterizing patient state may be relevant or irrelevant to patient similarity depending on the clinical context. For example, Creatinine may be hugely important in characterizing kidney failure but largely irrelevant when considering traumatic brain injury. Similarity should therefore be a function of the underlying clinical context of the patient.

A second aspect of patient similarity is that similarity should be value-dependent. Intuitively, for example, two patients with heart rates in the normal range of $70 - 75$ should receive a lower similarity score than two patients with elevated heart rates in the range $120 - 125$. This is an especially crucial concept in healthcare because the interesting and most relevant aspects of patient state typically lie in the abnormal (tails of the distribution). This is analogous to the notion of "term specificity" in information retrieval, in which the

importance of query terms to search is inversely proportional to how frequently the term occurs in a text corpus (inverse document frequency (IDF)).

We propose a novel patient similarity learning framework that attempts to combine both of the above concepts – that patient similarity should be dependent on clinical context and that similarity should be modulated by the frequency of the individual feature values. In addressing the latter, we generalize the notion of "term specificity" to continuous-valued features, which allows us to address similarity on ordinal features – e.g., how similar are two patients with heart rates of 120 and 125? – as opposed to similarity scoring based solely on the presence/absence of binary features. We achieve this by proposing novel kernel functions that are based on the population distribution of the features. The kernels are tuned to amplifying similarities in the tails of the population distribution. Here, population may loosely refer to a certain clinical population – for example, patients in the ICU.

To address the underlying clinical context, we use the multiple kernel learning (MKL) framework (Gonen and Alpaydin (2011)). Specifically, given a training dataset we learn an ensemble kernel over the individual population feature kernels described above that is capable of predicting one or more clinical contextual targets of interest. For example, the kernel may be optimized to predict aspects of cardiovascular health or predict hemodynamic deterioration. As a result, the ensemble kernel is comprised of many base kernels, each of which is tuned to emphasize distribution tails, and ensemble weights assigned to the base kernels are determined by how discriminative each is in predicting a clinical context.

From a clinical perspective, the learned ensemble kernel may be used both for providing a predictive score of the clinical target variable that it was trained on as well as a ranked list of most similar patients from a retrospective database. The list of similar patients may be used to contextualize and explain the predictive score, as well as enable the clinician to perform case-based reasoning by referring back to similar past cases. In addition, personalized cohorts may be constructed to produce statistics and analytics on related clinical questions. For example, to assist with therapy decision support at point-of-care, the clinician may interrogate the interventions given to the patient's personalized cohort.

The paper is organized as follows. In Section 2 we introduce the population-derived kernel functions and derive an explicit feature map based on (Vedaldi and Zisserman (2012)). This allows for an efficient multiple kernel learning algorithm, detailed in Section 2.4, that tunes the ensemble kernel to a particular clinical context of interest. We then present results of our algorithm in Section 3 on predicting hemodynamic instability in the ICU, and present an exploratory analysis of the ensemble kernel in constructing personalized cohorts and visualizing statistics on treatments and outcomes.

## 2. Methods

### 2.1 Notation

We use capital letters to denote random variables ($X$) and lowercase letters to denote instances of the random variable ($x$). We use superscripts when it's necessary to represent multiple samples from a random variable ($x^{(1)}, x^{(2)}, \dots$). When $X$ (resp. $x$) denotes a vector, we use subscripts $X_j$ (resp. $x_j$) to denote the $jth$ element of $X$ (resp. $x$).

### 2.2 Population-Derived Kernel Similarity Functions

Let $X = (X_1, X_2, \ldots, X_p)$ denote a set of random variables characterizing patient state. These can be any combination of vital signs (Heart Rate, Systolic blood pressure), laboratory values (Sodium, Lactate, Magnesium), comorbidities and demographics, etc. They may also be features extracted from a latent variable model – e.g., features extracted from a hidden layer in a deep neural network. A patient can then be represented by a realization $x$ from $X$. Additionally, patient state may be evaluated over a limited time window so that a realization may represent a patient at a particular point in time, and a patient may therefore exhibit multiple patient state vectors over time.

Our proposed kernel between patient state vectors is based on a very literal interpretation of the meaning of patient similarity: given two patients, their similarity is inversely related to the expected number of patients that lie between them. Since the base kernels are evaluated on individual features $X_j$, this amounts to calculating the expected number of patients that lie in an interval. For example, denote $x_j$ and $z_j$ as the corresponding feature values for two patients with state vectors $x$ and $z$. Then the expected number of patients between them is given by the area under the population distribution, $P(X_j)$, for $X_j$ in that interval. We therefore propose the following kernel on feature $X_j$:

$$k_{j,c}(x, z) = (1 - P(\min(x_j, z_j) \leq X_j \leq \max(x_j, z_j)))^c \tag{1}$$

where $c \geq 1$ is an exponent that controls the speed of decay in similarity. In Section 2.3 we show that the above is a valid kernel and derive an explicit feature map $\Psi_{j,c}(x)$ such that $k_{j,c}(x, z) \approx \langle \Psi_{j,c}(x), \Psi_{j,c}(z) \rangle$.

Figure 1 illustrates examples of the proposed kernel on three features: Hematocrit, Lactic Acid, and Patient age. Figures 1(a)-1(c) show the empirical population cumulative distribution functions (evaluated on ICU patients) for each of the three features, and Figures 1(d)-1(f) show the resulting kernel function from (1) with $c = 5$, plotted as a 2-dimensional heatmap. The population distribution for Hematocrit is fairly symmetric and mono-modal, whereas Lactic Acid and Patient Age both exhibit skew. Patient age is biased towards older patients because the population under consideration is adult ICU patients. In all cases, the kernel is tuned to the tails of the distribution: similarity is augmented if both patients lie in corresponding low-density regions of the distribution.

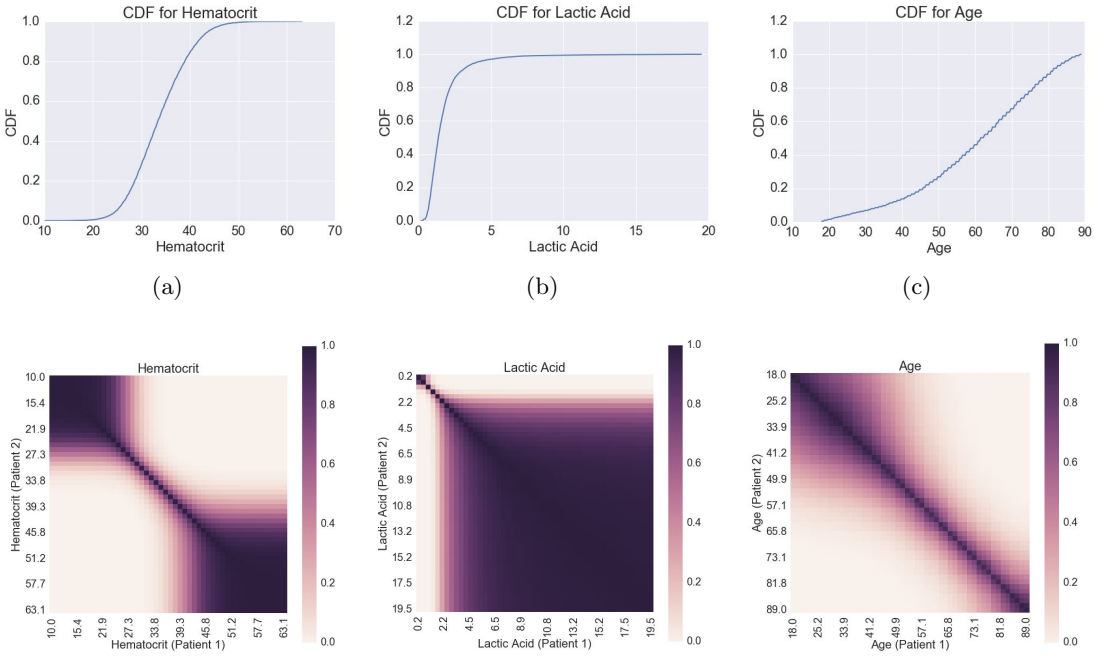
For continuous random variables $X_j$, $P(a \leq X_j \leq b) = P(a < X_j \leq b)$, and so (1) can be simplified to $k_{j,c}(x, z) = (1 - |F_j(z_j) - F_j(x_j)|)^c$, where $F_j(\cdot)$ is the cumulative distribution function of $X_j$. In this case, the kernel reduces to a stationary kernel after converting inputs to quantiles through $F_j$. We found, however, that many measurements in healthcare are either discrete or measured on a heavily quantized scale (e.g., heart rate and blood pressure are usually charted as integers). As a result, even many of the continuously-valued measurements portray a discrete distribution. This distinction is important; otherwise two patients with matching heart rates of 70 (normal range) would receive the same similarity score as two patients with matching heart rates of 120 (abnormally elevated).

It's instructive to note that the kernel can be readily applied to binary or nominal discrete features. For example, $X_j$ may be a Bernoulli random variable characterizing whether or not the patient exhibits a symptom or presents with a rare condition or comorbidity. In

this case, (1) simplifies to:

$$k_{j,c}(x,z) = \begin{cases} (1 - P(X_j = 1))^c & , x_j = z_j = 1 \\ (1 - P(X_j = 0))^c & , x_j = z_j = 0 \\ 0 & , x_j \neq z_j \end{cases}$$

Thus, the similarity between patients $x$ and $z$ is inversely related to the prevalence (absence) of the clinical condition if both patients have (don't have) the condition, and no similarity if they differ in condition status. The kernel may also be applied to nominal variables on $c$ categories by one-hot encoding, which converts it to $c$ Bernoulli random variables.




(a) (b) (c)

(d) Kernel on Hematocrit (e) Kernel on Lactic Acid (f) Kernel on Patient Age

Figure 1: Examples of the kernel $k_{j,c}(x,z)$ in (1) with $c = 5$ on three features evaluated on adult ICU population: Hematocrit, Lactic Acid, and Patient Age

### 2.3 Explicit Feature Map

In this section we show that the proposed kernel can be expressed as a sum of intersection kernels in a transformed input space. With this representation, we are able to derive an explicit feature map for the kernel following the work in (Vedaldi and Zisserman (2012)). The explicit feature map is *data-independent* and, unlike Nystrom's method, does not require estimating eigenvalues/eigenvectors of an empirical Gram matrix. We will first consider the $c = 1$ case in (1) and then generalize to $c > 1$.

For each kernel $k_{j,c}$, define a $2D$ transformation $x \to (F_j(x), R_j(x))$ defined by:

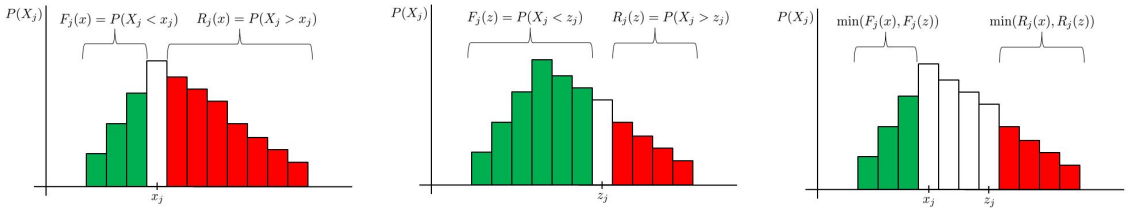
$$F_j(x) = P(X_j < x_j) \quad , \quad R_j(x) = P(X_j > x_j) \tag{2}$$

where, with a slight abuse of notation, $F_j$ is the (strictly less-than) cumulative distribution function of $X_j$, and $R_j$ is the reliability (complementary cumulative) function of $X_j$.

Given this transformation, the kernel in (1) for $c = 1$ can be equivalently expressed as:

$$k_{j,1}(x, z) = \min(F_j(x), F_j(z)) + \min(R_j(x), R_j(z)) \tag{3}$$

Thus, $k_{j,1}(x, z)$ is a sum of two intersection kernels applied in a two-dimensional space $x \to (F_j(x), R_j(x))$. The equivalence is shown visually in Figure 2.




(a) $x \to (F_j(x), R_j(x))$ (b) $z \to (F_j(z), R_j(z))$ (c) $k_{j,1}(x, z)$

Figure 2: Expressing $k_{j,1}(x, z)$ on $X_j$ as a sum of intersection kernels in a transformed space.

For $c \geq 1$, we can use the fact that $k_{j,c}(x, z) = k_{j,1}(x, z)^c$ and apply the binomial expansion on (3) to obtain:

$$k_{j,c}(x, z) = \sum_{i=0}^{c} \begin{pmatrix} c \\ i \end{pmatrix} \min(F_j(x), F_j(z))^i \min(R_j(x), R_j(z))^{(c-i)} \tag{4}$$

An explicit feature map for (4), denoted $\Psi_{j,c}(x)$ can be derived from explicit feature maps $\tilde{\Psi}_i(x)$ for kernels $\min(x, z)^i$, $i = 0, 1, \ldots, c$, which are given in the Appendix. Then by the additive/multiplicative combination of kernels (Vedaldi and Zisserman (2012)), the explicit feature map for $k_{j,c}(x, z)$ is given by:

$$\Psi_{j,c}(x) = \bigoplus_{i=0}^{c} \begin{pmatrix} c \\ i \end{pmatrix} \left[ \tilde{\Psi}_i(F_j(x)) \otimes \tilde{\Psi}_{c-i}(R_j(x)) \right] \tag{5}$$

where $\oplus$ is the direct sum of feature spaces and $\otimes$ is the Kronecker product.

There are two important points to note about the above explicit feature maps. The first is that the transformation $x \to (F_j(x), R_j(x))$ only depends on the distribution of $X_j$ through cumulative and (complementary cumulative) distribution functions and not on the probability mass function. As a result, we can calculate the transformation via empirical estimates of the cumulative distributions directly, which bypasses binning problems inherent to pmf estimation. Specifically, let $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ denote a sample of patient state vectors. Then $F_j(x)$ (resp. $R_j(x)$) can be estimated via $\sum_i \mathbb{I}(x_j^{(i)} < x_j)$ (resp. $\sum_i \mathbb{I}(x_j^{(i)} > x_j)$).

The second point to note is that the dimensionality of the explicit feature map may far exceed the number of distinct values, denoted $m$, that a feature can take on. For example, $X_j$ is Bernoulli ($m = 2$) or an ordinal discrete random variable taking on one of $m$ values. In these cases, we can project the feature map into an $m$-dimensional space spanned by the feature maps for the $m$ distinct values. This results in an $m$-dimensional feature space regardless of the original dimensionality of $\Psi_{j,c}$.

### 2.4 Training Algorithm - Multiple Kernel Learning

Up to this point, we have specified the family of base kernels using the population distributions for each input feature. In this section, we seek to learn a powerful ensemble kernel

that is guided by a supervised learning problem to predict a clinical variable of interest from the patient state vector. For example, the clinical variable of interest may be one or more variables characterizing a diagnosis or some measure of organ health. To achieve this we use the multiple kernel learning framework (Gonen and Alpaydin (2011)).

Let $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ denote a training sample with $x^{(i)} \in \mathbb{R}^p$ the input features (patient state vector) and $y^{(i)}$ a clinical variable of interest to be predicted. The techniques from Section 2.3 may first be applied in an unsupervised manner on each of the input features $j = 1, \ldots, p$ to obtain base kernels $k_{j,c}(\cdot, \cdot)$ and their explicit feature map approximations $\Psi_{j,c}(\cdot)$ so that $k_{j,c}(x, z) \approx \langle \Psi_{j,c}(x), \Psi_{j,c}(z) \rangle$. Via the kernel trick, kernelized predictive models in the input space are equivalent to linear predictive models in the explicit feature space. Therefore, we seek to train a generalized additive model of the form:

$$E(y|x) = g^{-1}\left(\sum_{j=1}^{p} f_j(x)\right) \tag{6}$$

where $E(y|x)$ is the conditional mean of the target $y$ given the data $x$, $g$ is a link function and each $f_j(x)$ is a linear transformation in the feature space for the $jth$ kernel – i.e., $f_j(x) = \langle w_j, \Psi_{j,c}(x) \rangle$, which is a nonlinear function of $x_j$.

As outlined above, we allow the patient state vectors to be incomplete due to missing feature values. To account for this, we assign a neutral similarity value $k_{j,c}(x, z) = 0$ if $x_j$ or $z_j$ is missing so that each base kernel "abstains" from ascribing a similarity score if its dependent feature is missing. In terms of the explicit feature map, this amounts to mapping missing feature values to the origin ($\Psi_{j,c}(x) = 0$ if $x_j$ is missing). To capture information in the measurement patterns, we augment the patient state vector with missingness features for each of the original patient state features. When deriving the base kernels on the missingness features, they are each treated as Bernoulli random variables.

The dimensionality of the predictive model is the sum of the dimensionalities of the individual base kernel's feature maps, which can be quite high. We therefore regularize the weights using a ridge penalty: $\sum_j ||w_j||^2$, which results in a convex problem, and we optimize using stochastic average gradient (Schmidt et al. (2013)). We note that techniques from (Lu et al. (2014)) can be used to further speed up training.

After training the predictive model, we seek to infer the ensemble kernel from the learned feature space weight vectors $w_1, \ldots, w_p$. There are a number of ways to do this. One possibility is to construct an ensemble kernel $k(x, z) = \sum_j \alpha_j k_{j,c}(x, z)$ with a static weighting where $\alpha_j = ||w_j||^2 / \sum_j ||w_j||^2$. In this way, a base kernel's contribution to the ensemble is based on the relative energy of the predictive model weights in its feature space.

Alternatively, a dynamic ensemble can be constructed in which the ensemble weights depend on one (or both) of the kernel input arguments. For example, given a query patient state vector $x$, $k(x, \cdot)$ may be used as a ranking function to retrieve similar patients from a large retrospective database. In this case, it may be more useful to construct an ensemble that is tuned to the query patient. For example, $k(x, z) = \sum_j \alpha_j(x) k_{j,c}(x, z)$, where $\alpha_j(x)$ depends on the query patient $x$. This results in an asymmetric kernel function ($k(x, z) \neq k(z, x)$), in which the first argument is treated as the query patient and the second argument is a candidate similar patient from a database of patients.

For similar patient retrieval, we propose to use the following ensemble weighting:

$$\alpha_j(x) = |f_j(x)| \tag{7}$$

in which case the ensemble weights characterize the strength of the prediction made by each base kernel's feature space on the query patient $x$.

A block diagram of the entire learning framework is presented in Figure 3.

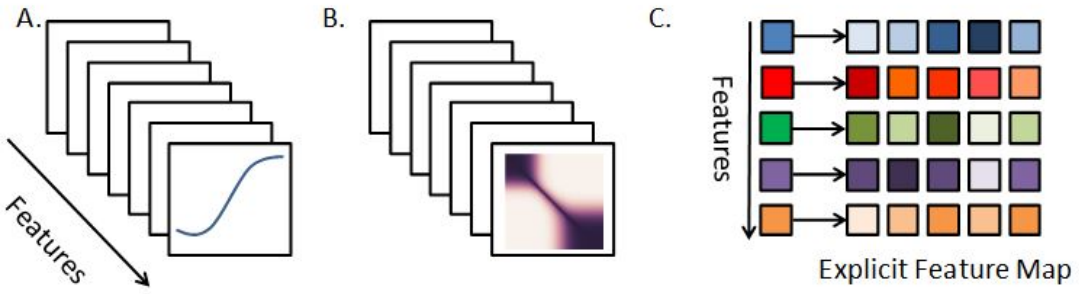



Figure 3: Learning framework block diagram. (A) For each feature, a cumulative distribution function (CDF) is estimated via training data; (B) The CDF for each feature induces a CDF kernel (Section 2.2); (C) Each feature is then transformed into a higher-dimensional space via its kernels explicit feature map (Section 2.3). These explicit maps are concatenated to form the high-dimensional feature space used by the multiple kernel learning algorithm (Section 2.4).

## 3. Results

### 3.1 Data Cohort

Patient data were obtained from the eResearch Institute (McShea et al. (2010)); the data included records from 40,883 patients from across 25 hospitals. Episodes of hemodynamic instability were identified based on a set of charted clinical interventions that was developed from a strong consensus among of group of experienced intensive care physicians. This included the administration of inotropic or vasopressor medications, administration of at least 2.4L of fluid (colloid or crystalloid) over 8 hours, and administration of packed red blood cells (PRBC's). For further details, see the work of (Conroy et al. (2016)).

For purposes of training and validation, the patients ICU stays were divided into 6 h segments, and these segments were labeled as either stable or unstable. Unstable segments were the 6 h period prior to any intervention described above. Stable segments were chosen from patients who had none of the above interventions or who ended their ICU stay with at least 18 h without an intervention. A 6 h segment was chosen at random from these stable periods for the stable segments. This criteria resulted in 49,256 labeled segments (44,019 stable; 5,237 unstable). For each segment, we extracted a total of 61 features that comprised vital signs, lab values, and demographic information about the patient. For each segment, the data extracted 1 h prior to intervention was used for training.

### 3.2 Performance of Kernelized Predictive Model

We first evaluated the quality of the explicit feature map in approximating the kernel functions. Supplementary Figure 1 plots the mean absolute error, $\text{mean}(|k_{j,c}(x,z) - \langle \Psi_{j,c}(x), \Psi_{j,c}(z) \rangle|)$, as a function of the dimensionality of the feature map for the kernels defined on heart rate

with $c \in \{1, 2, 5, 10\}$. Larger values of the kernel parameter $c$, which acts to speed the decay in similarity, require higher dimensionality to achieve the same approximation error.

We then trained the generalized additive model (6) to predict hemodynamic status (0 = stable, 1 = unstable). Model performance was tested using 10-fold cross-validation, and the folds were generated to avoid splitting segments of a patient across training and test sets. For each feature, the kernel parameter $c = 5$ was chosen. Based on Supplementary Figure 1, we selected the dimension of the feature map to be 500, which achieves an error of less than 0.02. The model achieved a cross-validated AUC $= 0.881 \pm 0.004$, which slightly outperforms the best model in (Conroy et al. (2016)) (AUC = 0.8772), which uses boosting.

We can also examine what was learned by the model by visualizing the nonlinear features, $f_j(x)$, of the generalized additive model in (6). These nonlinear features have two interpretations: first, they model the log-odds of hemodynamic instability and therefore serve as instability risk functions (positive values indicates higher risk); and second, they serve as the kernel weights in the asymmetric ensemble kernel in (7). Supplementary Figure 2 plots the nonlinear features learned for Noninvasive Systolic Blood Pressure, Hematocrit, and Shock Index, which were identified as the three most important features (as measured by $||w_j||$). Both systolic blood pressure and hematocrit resemble smoothed step functions, in which risk for hemodynamic instability increases below a critical threshold (approximately 100 for systolic blood pressure and 30% for Hematocrit). Both of these thresholds make sense clinically and agree with previous analyses (Conroy et al. (2016)). Shock index exhibits a more complicated sinusoidal risk pattern, with lower risk transitioning to higher risk above 0.7 and peaking at 0.9.

### 3.3 Patient Similarity for Cohort Building and Exploration

We next explored the use of the patient similarity function learned by the asymmetric ensemble kernel in Section 2.4 to addressing secondary, but related questions to the task of hemodynamic instability prediction. To do this, we construct a personalized cohort for a given query patient, which we define as the 200 most similar patients, as ranked by the ensemble kernel. Future work will examine fine-tuning the cohort size based on the ordered similarity scores themselves. We can then evaluate relevance by comparing interventions and outcomes of the personalized cohort to the true intervention/outcome of the query patient. In the following, we refer to the population-based kernel as "CDF kernel".

We contrast the results against developing an ensemble kernel based on a standard stationary kernel – the RBF kernel on individual features $k_j(x, z) = \exp(-\gamma(x_j - z_j)^2)$. The RBF kernel was also trained to predict hemodynamic status in the exact same manner as was described in Section 3.2. The kernel hyperparameter $\gamma$ was tuned locally around the default value $\gamma = 1$ to achieve similar predictive performance (cross-validated AUC $= 0.874 \pm 0.007$). The explicit feature map used for the RBF kernel was based on the random Fourier features given in (Rahimi and Recht (2007, 2008)).

We first explore the "term specificity" property of the population-based kernel functions – retrieval of similar patients based on abnormality of feature values (Figure 4). To achieve this, we first grouped hemodynamically unstable patients by the intervention they eventually received (PRBC, fluid, inotrope, or pressor). For each group, we then compare the population distribution (black) of feature values to the distribution returned by the per-

sonalized cohort for each patient in the group (red: CDF kernel, green: RBF kernel). The top row in Figure 4 visualizes the distribution of Shock Index, Systolic Blood Pressure, and Hematocrit for patients that received pressors or PRBC. Patients that need pressors are more likely to have abnormal shock index and systolic blood pressure (blue), similarly patients needing PRBC transfusions have abnormal hematocrit levels (blue) compared to the whole population (black). The personalized cohort retrieved by the CDF kernel (red) better reflects the clinical evidence by retrieving patients with abnormal feature values that better reflects the intervention group distribution (blue), compared to the RBF kernel (green). We quantified the difference in kernel distributions from the population distribution using the KL divergence, and the bottom row of Figure 4 plots the results for the PRBC and Pressor groups. The proposed CDF kernel achieves higher divergence than the RBF kernel for both groups, indicating the cohort returned by the RBF kernel is more similar to the population distribution than the CDF kernel – which better approximates abnormal feature values seen in the intervention group distribution.

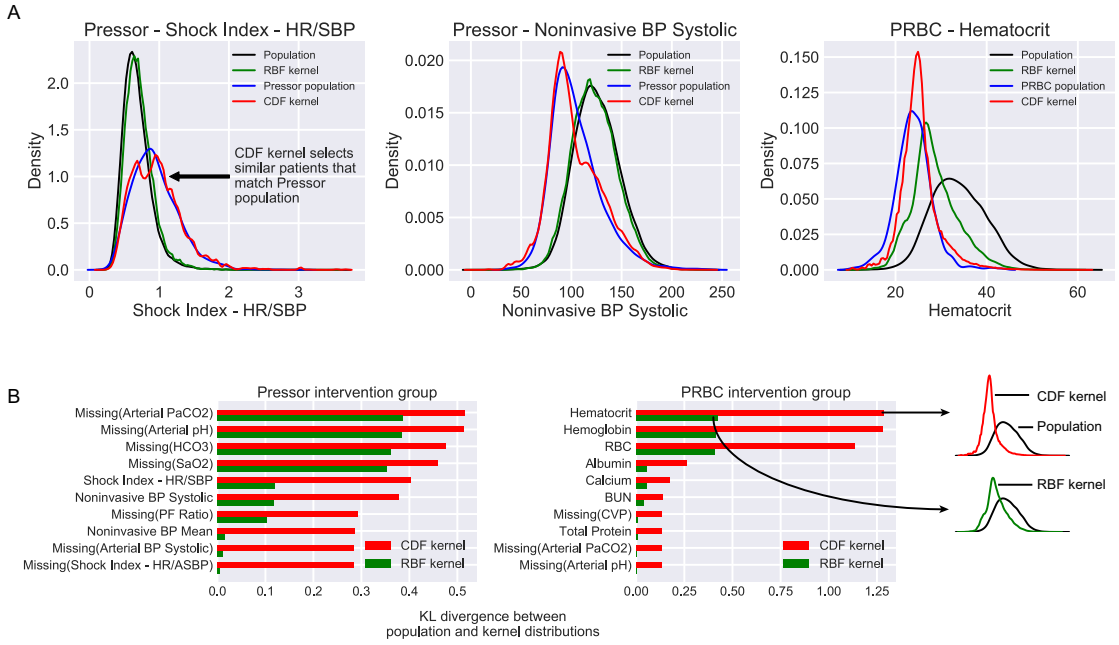
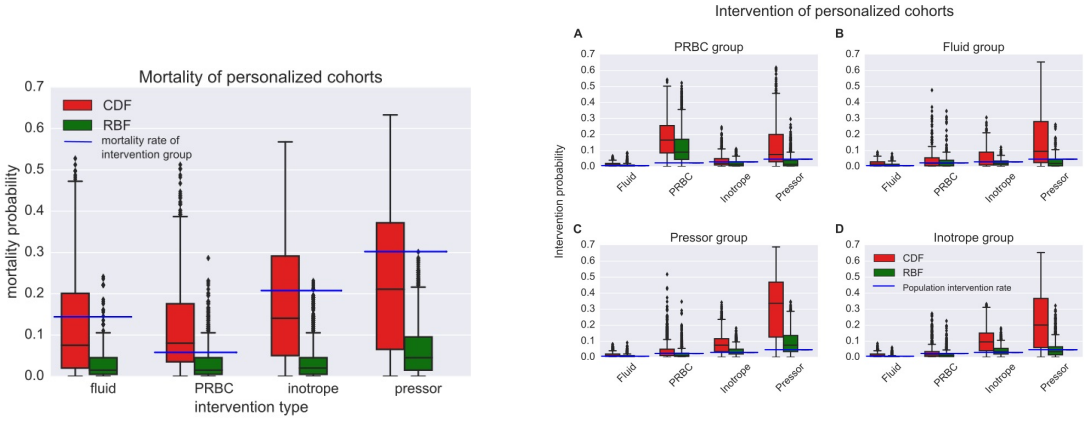



Figure 4: Term-specificity property of the population-based kernel functions. A) Comparing distributions of Shock Index, Systolic Blood Pressure, and Hematocrit for Pressor and PRBC intervention groups. Population distribution (black) versus the intervention group distribution (blue) and the distributions retrieved by the personalized cohorts for the CDF (red) and RBF (green) kernels. The CDF kernel retrieves patients with feature values that are most similar to the intervention group (red and blue). B) KL divergence between population distribution of feature values and the distribution returned by the the personalized cohorts. The CDF kernel has a higher divergence from the population indicating a larger difference in the feature values of the retrieved patients.

We next explored the relevance of the personalized cohorts in retrieving similar interventions and outcomes to the query patient. For each intervention group, we compare the

population distribution of outcome/intervention type to the distribution returned by the personalized cohorts of each patient in the intervention group. Figure 5a plots the mortality rate of personalized cohorts for each intervention group compared to the true mortality rate of that intervention group. The RBF-based similar patient cohort does not resemble the entire intervention group in terms of mortality rate. The CDF-based personalized cohorts have mortality rates that are closer to the intervention group's true mortality rate.

We also compared each query patient's intervention given to the distribution of interventions given to its personalized cohort. These are aggregated by query patient intervention and shown in Figure 5b. For patients in the PRBC group (Figure 5b,A), the dominant intervention given to similar patients was PRBC, followed by Pressor. For Pressor patients(Figure 5b,C), the personalized cohort was dominated by Pressor patients, followed by Inotrope patients. In both cases, the effect was more strongly reflected using the CDF kernel rather than the RBF kernel. The CDF and RBF kernels both have difficulty in discerning the fluids and inotrope patient groups (Figure 5b,B,D), which may be due to a limitation of the available feature data – an area for future work.




(a) (b)

Figure 5: (a) Mortality rate of personalized cohort for each intervention group. Compared to RBF (green), CDF-based cohorts (red) have mortality rates that are closer to the true mortality rate observed in a given intervention group. (b) Interventions given to personalized cohort, grouped by intervention.

## 4. Conclusion

We presented a multiple kernel learning framework to learn similarity functions that compare physiological state between patients. The learning procedure is based on two desirable properties: (1) that the similarity can be tuned to a particular clinical context; and (2) that the similarity reflects a generalized notion of "term specificity" – that shared abnormal feature values should receive an amplified similarity score. We validated our model on real ICU data to predict hemodynamic instability and explored the relevance of the patient similarity function for crafting personalized cohorts and presenting statistics. A natural extension of our approach is to allow the features characterizing patient state to be learned simultaneously; e.g., the hidden layers of a deep learning network which are updated by back-propagation during training. This can be used to learn interactions between dependent features that strongly influence the population distributions.

## References


Bryan Conroy, Larry Eshelman, Cristhian Potes, and Minnan Xu-Wilson. A dynamic ensemble approach to robust classification in the presence of missing data. *Machine Learning*, pages 443–463, 2016.

Mehmet Gonen and Ethem Alpaydin. Muliple kernel learning algorithms. *Journal of Machine Learning Research*, pages 2211–2268, 2011.

Zhiyun Lu, Avner May, Kuan Liu, Alireza Bagheri Garakani, Dong Guo, Aurelien Bellet, Linxi Fan, Michael Collins, Brian Kingsbury, Michael Picheny, and Fei Sha. How to scale up kernel methods to be as good as deep neural nets. *arXiv*, 2014.

M McShea, R Holl, O Badawi, RR Riker, and E Silfen. The eicu research institute - a collaboration between industry, health-care providers, and academia. *IEEE Eng Med Biol Mag*, pages 18–25, 2010.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 2007.

A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: replacing minimizations with randomization in learning. *Advances in Neural Information Processing Systems*, 2008.

Mark Schmidt, Nicholas LeRoux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *arXiv*, 2013.

Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE transactions on pattern analysis and machine intelligence*, pages 480–492, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J.R. Statist. Soc. B*, pages 49–67, 2006.

# Supplementary Material for "Patient Similarity Using Population Statistics and Multiple Kernel Learning"

## Appendix A.

### Explicit Feature Map Kernels

The explicit feature map (6) depends on explicit feature maps for the kernels $\min(x,z)^i$, $i = 0, 1, \ldots, c$, which we derive here using results from (Vedaldi and Zisserman (2012)). Since $\min(x,z)^i = \min(x^i, y^i)$, we can express the feature map $\tilde{\Psi}_i(x)$ in terms of the feature map $\tilde{\Psi}(x)$ of the intersection kernel $\min(x,z)$ as:

$$\tilde{\Psi}_i(x) = \tilde{\Psi}(x^i)$$

The explicit feature map for the intersection kernel, $\tilde{\Psi}(x)$ is given in (Vedaldi and Zisserman (2012)).

### Supplementary Figures

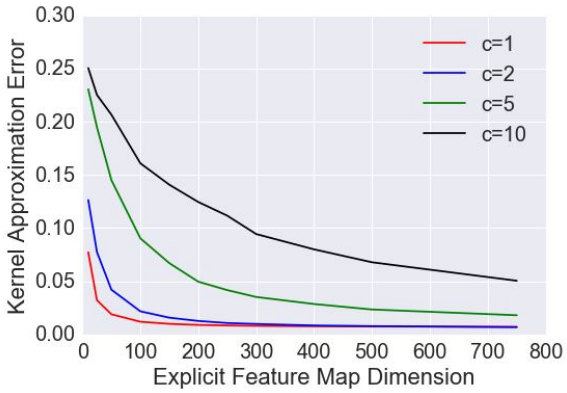
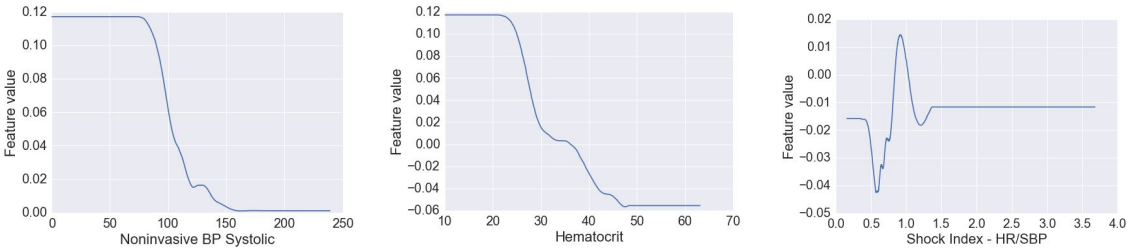



Figure 1: Kernel Approximation Error: Comparison of true kernel function $k_{j,c}(x,z)$ for heart rate and its explicit feature map approximation $\langle \Psi_{j,c}(x), \Psi_{j,c}(z) \rangle$. Mean absolute error, mean $(|k_{j,c}(x,z) - \langle \Psi_{j,c}(x), \Psi_{j,c}(z) \rangle|)$, is plotted against the dimensionality of the feature map $\Psi_{j,c}$ for kernel parameter $c \in [1, 2, 5, 10]$. Larger values of $c$ require higher dimensionality.

## References

Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE transactions on pattern analysis and machine intelligence*, pages 480–492, 2012.




(a) (b) (c)

Figure 2: Nonlinear features learned in generalized additive model.