Gaussian process regression and classification using International Classification of Disease codes as covariates

Sanvesh Srivastava 1, Zongyi Xu 1, Yunyi Li 2, W. Nick Street 3, Stephanie Gilbertson-White 4

1Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa, USA
2McCombs School of Business, University of Texas, Austin, Texas, USA
3Tippie College of Business, The University of Iowa, Iowa City, Iowa, USA
4College of Nursing, The University of Iowa, Iowa City, Iowa, USA

Correspondence
Sanvesh Srivastava, Department of Statistics and Actuarial Science, The University of Iowa, 241 Schaeffer Hall, 20 East Washington Street, Iowa City, IA 52242, USA. Email: sanvesh-srivastava@uiowa.edu

Funding information
Office of Naval Research, Grant/Award Number: ONR-BAA N000141812741; National Science Foundation, Grant/Award Number: DMS-1854667/1854662; National Institute for Nursing Research, Grant/Award Number: P20NR018081

Abstract
In electronic health records (EHRs) data analysis, nonparametric regression and classification using International Classification of Disease (ICD) codes as covariates remain understudied. Automated methods have been developed over the years for predicting biomedical responses using EHRs, but relatively less attention has been paid to developing patient similarity measures that use ICD codes and chronic conditions, where a chronic condition is defined as a set of ICD codes. We address this problem by first developing a string kernel function for measuring the similarity between a pair of primary chronic conditions, represented as subsets of ICD codes. Second, we extend this similarity measure to a family of covariance functions on subsets of chronic conditions. This family is used in developing Gaussian process (GP) priors for Bayesian nonparametric regression and classification using diagnoses and other demographic information as covariates. Markov chain Monte Carlo (MCMC) algorithms are used for posterior inference and predictions. The proposed methods are tuning free, so they are ideal for automated prediction of biomedical responses depending on chronic conditions. We evaluate the practical performance of our method on EHR data collected from 1660 patients at the University of Iowa Hospitals and Clinics (UIHC) with six different primary cancer sites. Our method provides better sensitivity and specificity than its competitors in classifying different primary cancer sites and estimates the marginal associations between chronic conditions and primary cancer sites.

KEYWORDS
Bayesian learning, Gaussian process, Markov chain Monte Carlo, nonparametric regression and classification, string kernels

1 | INTRODUCTION

Focusing on EHR data analysis, we develop GP priors for Bayesian nonparametric regression and classification using a patient’s chronic conditions as covariates. EHRs are useful for organizing information about clinical care and are also used for answering important research questions that cannot be practically addressed through traditional prospective research. The motivating UIHC EHR encodes diagnoses using the 10th edition of
the ICD billing code (hereafter called the ICD code). Following Calderón-Larrañaga et al. (2017), a chronic condition is defined as a set of ICD codes. An ICD code is a three to seven character long string, where the first three characters define the diagnosis category, the next three define the aetiology, anatomic site and severity, and the final character identifies encounters. We develop a family of kernel functions (or similarity measures) on a pair of ICD code subsets using the ICD code and chronic condition definitions. Using any such kernel as a covariance function for a GP prior, we perform Bayesian nonparametric regression and classification using patients' chronic conditions as covariates.

Encoding diagnoses using the ICD coding system has become the most widely adopted medical diagnostic classification system since it was endorsed by the World Health Assembly in 1990. Automated data extraction from EHRs provides a cost-effective and efficient way to gather detailed medical information about chronic conditions compared to the more expensive and time-consuming manual chart review by healthcare providers. In well-curated EHRs, models that include ICD codes as covariates tend to achieve higher predictive accuracy for disease risks and related medical conditions than those without ICD codes (Bastarache, 2021; Horng et al., 2017). Building on these ideas, we achieve an additional but significant improvement in predictive accuracy by exploiting the hierarchical structure of ICD codes and the similarity of ICD codes defining the same chronic condition. Most importantly, the proposed methods require minimal tuning and provide results with uncertainty estimates, so a healthcare provider can easily use them for assisting with patient care.

The literature on Bayesian models for EHR data analysis using ICD codes as covariates is sparsely populated. Henao et al. (2016) have developed a deep hierarchical factor model for identifying latent subgroups of patients. They summarize the EHR data in terms of counts of different medications usage, laboratory tests and diagnoses specified using ICD codes. If the EHR data only contains information about the presence or absence of disease, then Bayesian nonparametric extensions of low-rank matrix factorization perform better in discovering latent disease groups (Ni et al., 2020). Recently, topic models and deep learning have been widely used for automated clustering, regression and classification in EHR data. Topic model-based approaches are unsupervised and focus mainly on automated phenotyping and clustering patients into homogenous subgroups (Ahuja et al., 2022; Li et al., 2020; Song et al., 2022; Zou et al., 2022). The deep learning approaches are concerned with mining clinical notes and not on using the information encoded in ICD codes or on providing uncertainty estimates (Rajkomar et al., 2018). None of these approaches leverage the additional information provided by the hierarchical structure of ICD codes and the chronic condition definition.

Addressing this gap, we develop a kernel function on ICD codes that exploits both sources of information. A widespread practice is to define the similarity of two patients to be proportional to the number of common ICD codes in their diagnoses or Jaccard index. Our similarity measure is superior to this practice in that its value could be positive for a pair of non-identical ICD codes, and its magnitude depends on the degree of overlap between the pair. In fact, our kernel function extends a restricted version of the boundrange kernel, a popular string kernel in text mining (Lodhi et al., 2002; Shawe-Taylor & Cristianini, 2004). The similarity between a pair of ICD codes is defined as the total number of common substrings between them, where a substring always begins at the first character of an ICD code. This similarity measure is extended to a pair of subsets as a weighted average of similarities of all ICD code pairs formed using the two subsets.

We use these similarity measures in defining GP priors. First, we extend the similarity measure to define the equivalents of polynomial, exponential and SE kernels on subsets of ICD codes, which represent chronic conditions. Second, we use these kernels as covariance functions for defining GP priors indexed by subsets of ICD codes. These GP priors are used for fitting Bayesian nonparametric regression and classification models that are tuned for EHR data analysis (Rasmussen & Williams, 2006). We develop MCMC algorithms for automated fitting of these models that provide theoretically guaranteed estimation of posterior uncertainty (Ghosal & van der Vaart, 2017). Minimal tuning requirements make our method extremely attractive for routine biomedical applications and for assisting healthcare providers in predicting risks related to chronic conditions. We verify our claims through empirical studies and show that the proposed method provides a more detailed quantification of the dependence of primary cancer sites on chronic conditions.

## 2 | Motivating EHR Data

Multimorbidity is the simultaneous occurrence of more than one chronic condition. Developing operational measures of multimorbidity is an active research area due to its importance in improving patient care. Calderón-Larrañaga et al. (2017) have operationalized the concept of multimorbidity into a list of 918 ICD codes that map onto a list of 60 chronic conditions. This serves as a motivation for developing a classifier that accounts for the grouping of diagnoses into chronic conditions. They show that this yields better accuracy in predicting primary cancer sites in UIHC EHR data.

The data for this analysis came from the EHR at UIHC relating to the set of patients who received cancer treatment beginning in 2017. The cohort included 18 years or older patients who were diagnosed with malignant solid neoplasm of any type or site. The data included patients' diagnoses encoded as ICD codes and the race and marital status of every patient. We filtered patients who had at least one of the 58 non-cancerous chronic conditions, resulting in a sample size of 1660. The main question here is to identify chronic conditions or diagnoses that predict a primary cancer site. Unfortunately, the Bayesian toolkit provides limited options for answering such biomedical questions. Using the UIHC EHR data as a representative example, we develop GP-based regression and classification models that are tuned for solving such biomedical problems, where the subsets of ICD codes are covariates.
The main goal is to predict marginal associations between chronic conditions and cancer primary sites and a minor goal is to predict the primary cancer site using the ICD codes. Consider an example illustrating the advantages of exploiting the information contained in the ICD codes. The code J301 encodes allergy, and D50, D51 denote different kinds of anaemia. Clearly, using dummy variables for representing these ICD codes is inappropriate because D50, D51 denote the same disease but are assigned different dummy variables. Motivated by the text mining literature, a useful similarity measure for strings is defined by counting the number of matching substrings of different lengths. Accounting for the hierarchical structure of ICD codes, if we modify this definition by enforcing the substrings to always start at the beginning, then the similarities of the three pairs (J301, D50), (J301, D51) and (D50, D51) are 0, 0 and 2, respectively, which are more realistic. In the next section, we develop this idea further to account for the extra structure provided by the 58 chronic conditions while down-weighting the contributions of commonly occurring diagnoses.

3 | KERNEL FUNCTIONS FOR DIAGNOSES

3.1 | Kernel function on ICD codes

Consider a general description of the structure of ICD codes. Let $t$ be an ICD code, $t^*$ be the longest valid ICD code in the database, $|t|$ denote the number of characters in $t$ and $t_{1,i}$ be the contiguous substring of length $i$ that starts at the first character of $t$. The first character of $t$ is $t_{1,1}$ and $t_{1,i} = t$ if $i > |t|$. Let $T$ be the set of all ICD codes, $A$ and $D$ be the sets of letters and numbers that are used to represent an ICD code and $B = A(\{A \cup D\})^{|t| - 1}$ be set of all alphanumeric substrings that have a letter as their first character and have lengths 1, 2, ..., $|t^*|$. Then, $B$ contains all contiguous substrings of ICD codes in $T$ that start at the first character in an ICD code and $|T| < |B| = |A(\{A \cup D\})^{|t| - 1}|$.

The kernel function on two ICD codes is defined as a Euclidean dot product between their feature maps. Let $\Psi(t)$ be the feature map of an ICD code $t \in T$ and $\mathcal{F} = \{\Psi(t) : t \in T\}$ be the feature space. Then, for every $t \in T$, $\Psi(t) \in \{0, 1\}^{|t^*| |B|}$ and is defined as

$$\Psi : T \rightarrow \mathcal{F}, \Psi(t) = \{\Psi_b(t)\}_{b \in B}, \Psi_b(t) = \{1(b = t_{1,1}), ..., 1(b = t_{|B|})\}^T,$$

where $1(\cdot)$ is $1$ if $\cdot$ is true and $0$ otherwise. The kernel function on $T \times T$ is defined using $\Psi(\cdot)$ and known $|t^*| \times |t^*|$ symmetric positive semi-definite matrices $(A_b)_{b \in B}$ as

$$k_0(t, t') = \sum_{b \in B} \Psi^*_b(t) A_b \Psi_b(t') = \Psi^T(t) A \Psi(t'), t, t' \in T,$$

where $A = \text{diag}(A_b : b \in B)$ is a block diagonal matrix. Two useful special cases of (2) are obtained by setting $A_b$ to be a diagonal matrix in (2) such that (i) $A_b = I_b |A|$ and (ii) $(A_b)_{a=0} = \lambda^a I$ for every $b \in B$ and $a = 1, 2, ..., |t^*|$, where $I$ is an identity matrix, $\lambda_b > 0$ for every $b \in B$, and $\lambda \in (0, 1)$. We show in the Supporting Information that $k_0(t, t' | A)$ in (2) is a valid kernel function on $T \times T$.

The kernel in (2) is a type of string kernel, which is widely used in text mining (Shawe-Taylor & Cristianini, 2004). Two popular string kernels are spectrum and boundrange kernels (Lodhi et al., 2002). The $k$-spectrum kernel defines the similarity as the number of matching substrings between two strings of length $k$ exactly. The $k$-boundrange kernel instead defines the similarity to be the number of matching substrings between two strings of length less than or equal to $k$. The substrings can be non-contiguous in both kernels, and the substring matches can be weighted by a factor that decays exponentially with the substring length. The kernel in (2) is a restricted version of $|t^*|$-boundrange kernel where the substrings always start at the beginning of an ICD code.

3.2 | Kernel function on subsets of ICD codes

We now define the kernel function on subsets of ICD codes that represent diagnoses. Let $2^T$ be the power set of $T$, $t = \{t_1, ..., t_r\}$ be an element of $2^T$, $\Psi(t)$ be the feature map of $t$ and $\mathcal{F}_p = \{\Psi(t) : t \in 2^T\}$ be the feature space of subsets of ICD codes. The feature vector $\Psi(t)$ and the kernel function $k_1$ on $2^T \times 2^T$ are defined using $\Psi(\cdot)$ in (1) and $k_0(\cdot, \cdot | A)$ in (2), respectively, as

$$\Psi(t) = \sum_{t \subseteq t} \Psi(t), k_1(t, t' | A) = \sum_{t \subseteq t, t' \subseteq t} k_0(t, t' | A) = \Psi^T(t) A \Psi(t'),$$

for $t, t' \in 2^T$, where $\Psi(t)$ is a $(|t^*| |B|)$-dimensional vector with entries in $\{0\} \cup \mathbb{N}$. The kernel $k_1$ is a derived subsets kernel with the base kernel $k_0$; see Definition 9.41 in Shawe-Taylor and Cristianini (2004, p. 317).
In the motivating EHR data, covariates include subsets of ICD codes with an additional structure specified by their relation to a chronic condition; see Calderón-Larrañaaga et al. (2017) for the detailed definitions. Let \( C \) be the total number of chronic conditions and \( T_c \in \mathbb{Z}^2 \) be the subset of ICD codes that defines the chronic condition \( c \ (c = 1, \ldots, C) \). A patient has chronic condition \( c \) if the patient’s diagnoses consists of an ICD code that belongs to \( T_c \). Accounting for the \( C \) chronic conditions, represent a patient’s diagnoses as the set \( T = \{t_1, \ldots, t_c, \ldots, t_C\} \in \mathbb{Z}^2 \), where \( t_c \subseteq T_c \). A kernel function on \( (2^T)^C \times (2^T)^C \) that accounts for the structure imposed by \( C \) chronic conditions is

\[
k_2(T, T' | \mathbf{w}, \Lambda) = \sum_{c=1}^{C} \omega_c \kappa_2(t_c, t'_c | \Lambda), T, T' \in (2^T)^C, \omega_c > 0, c = 1, \ldots, C,
\]

where \( \Lambda \) is defined in (2), \( \mathbf{w} = (w_1, \ldots, w_C) \) and \( \omega_c \) indicates the importance of chronic condition \( c \) in defining the similarity between two patients with diagnoses \( T \) and \( T' \), respectively. The kernel \( \kappa_2 \) assumes the effect of \( C \) chronic conditions on the response is additive, but other valid methods for combining kernels \( \kappa_1(t_1, t'_1 | \Lambda), \ldots, \kappa_2(t_c, t'_c | \Lambda) \) can be also used in (4).

The kernel \( \kappa_2 \) in (4) is affected by the number of elements in \( T \) and \( T' \). The value of \( \kappa_1 \) in (3) is impacted by \( t \) in that \( \|\mathbf{Y}(t)\| \) increases with the cardinality of \( t \), where \( \| \cdot \| \) is the Euclidean norm. This implies that \( \kappa_2 \) in (4) is large if the cardinality of \( T \) is large; therefore, the length of feature maps impacts the similarity measure instead of the chronic conditions. This undesirable effect is removed by normalizing the feature maps of \( T, T' \) before computing the dot product, resulting in the kernel function

\[
k(T, T' | \mathbf{w}, \Lambda) = \frac{\kappa_2(T, T' | \mathbf{w}, \Lambda)}{\left\{k_2(T, T' | \mathbf{w}, \Lambda)\right\}^{1/2} \left\{k_2(T, T' | \mathbf{w}, \Lambda)\right\}^{1/2}}. \tag{5}
\]

All kernels in this paper are derived from \( \kappa \) in (5), which depends on the kernels \( \kappa_1, \kappa_2 \) and \( \kappa_3 \) defined earlier.

The following proposition states that \( \kappa \) is a valid kernel function on \((2^T)^C \times (2^T)^C\) and yields a metric on the same space using the duality between distance and kernel functions. We use the metric for defining the equivalents of polynomial, exponential and squared exponential (SE) kernels on \((2^T)^C \times (2^T)^C\), respectively.

**Proposition 1.** Let \( \{A_b\}_{b \in B} \) and \( \mathbf{w} \) be defined as in (2) and (4). \( T, T' \in (2^T)^C, \sigma^2 > 0, \phi > 0 \) and \( d(T, T' | \mathbf{w}, \Lambda) = \{\kappa(T, T' | \mathbf{w}, \Lambda) + \kappa(T', T | \mathbf{w}, \Lambda) - 2\kappa(T, T' | \mathbf{w}, \Lambda)\}^{1/2} \).

Then, given \( \Lambda \) and \( \mathbf{w} \),

1. \( \kappa(\cdot, \cdot | \mathbf{w}, \Lambda) \) and \( d(\cdot, \cdot | \mathbf{w}, \Lambda) \) are valid kernel and distance functions on \((2^T)^C \times (2^T)^C\);
2. if \( \Lambda = R^{-1} R \), where \( R \) is an upper triangular matrix, then the feature map of \( \kappa \) is

\[
\Phi_w(T) = \frac{\Psi_{wR}(T)}{\|\Psi_{wR}(T)\|} = \{w_1^{1/2} \mathbf{R} \Psi(t_1), \ldots, w_C^{1/2} \mathbf{R} \Psi(t_c)\}^\top, \tag{6}
\]

where \( \Psi_{wR}(T) \) is a \((C|t^*|B)\)-dimensional vector; and
3. the equivalents of polynomial kernel of order \( s \) and \( \gamma \)-exponential kernel on \((2^T)^C \times (2^T)^C, \) respectively, are defined as

\[
k_s(T, T' | \sigma^2, \mathbf{w}, \Lambda) = \sigma^2 \left(1 + \kappa(T, T' | \mathbf{w}, \Lambda)\right)^s, s \in \mathbb{N};
k^\gamma_s(T, T' | \sigma^2, \phi, \mathbf{w}, \Lambda) = \sigma^2 e^{-\phi d(T, T' | \mathbf{w}, \Lambda)^\gamma}, \gamma \in [1, 2], \tag{7}
\]

where \( \sigma^2 \) and \( \phi \) play the role of variance and inverse length-scale parameters, respectively, and \( \kappa^\gamma_s \) reduces to the exponential and SE kernels when \( \gamma \) equals 1 and 2, respectively.

The proposition’s proof is in the Supporting Information with other proofs. The polynomial kernel has finite dimensional feature. If \( s > 1 \), then the polynomial kernel \( \kappa_s \) is obtained by a non-linear embedding of the kernel \( \kappa \). The dimension of the feature space for \( \kappa_s \) is \((C|t^*|B)^{s+1}\) (Proposition 9.2, Shawe-Taylor & Cristianini, 2004). The feature space of \( \kappa^\gamma_s \) lies in a Hilbert space of functions, where the ‘smoothness’ of a
function depends on the chosen kernel. We use the SE kernel in our experiments due to its popularity; see the Supporting Information for the precise definition of smoothness and related details.

Figure 1 illustrates the utility of $\kappa$, $\kappa_1^e$, $\kappa_2^e$ in capturing patients' similarity in the UIHC EHR data. The patients with common primary sites have similar sets of chronic conditions. Let $T_i$ be the ICD codes for the patient $i$, $n$ be the sample size and $K(\cdot, \cdot)$ be a kernel function on $(2^T)^C \times (2^T)^C$. Then, the kernel matrix defined using $K_n$ is an $n \times n$ matrix with $(i, j)$th entry $K(T_i, T_j)$. To highlight the similarity of patients, we have grouped them into six blocks depending on their primary cancer sites. The kernel matrix with $K = \kappa_2$ in (4) fails to capture the similarities of patients with common primary sites because instead of the diagnoses, their number determines the patient similarities (Figure 1(a)). On the other hand, the kernel

![Kernel matrices](image)

**Figure 1** Heatmaps of kernel matrices for patients in the UIHC EHR data. The colour palette represents 0 and 1 using dark blue and dark red colours, respectively. The first four kernel matrices in (a)-(d) set $\Lambda = I$, $\sigma^2 = 1$, $\phi = 1$ and $\omega_2 \times \chi^2(T_c, T_c)$ for kernels $\kappa_2, \kappa, \kappa_1^e$ and $\kappa_2^e$, where $T_c$ is the set of ICD codes defining the $c$th chronic condition. The matrices in (e) and (f) are obtained using spectrum and boundrange kernels with substring length 3. The block correlation structures are better captured in (b), (c) and (d) than in (a), (e) and (f), indicating the superiority of $\kappa$, $\kappa_1^e$ and $\kappa_2^e$ kernel functions.
matrix with $K = \kappa$ in (5) shows high similarities for diagnostically similar patients, where the diagonal blocks denote patients with common primary sites and the off-diagonal blocks capture similarities of patients with cancers at two different sites (Figure 1b). The same observation is true for the $\gamma$-exponential kernel in (7) with $\gamma$ equals 1 and 2, respectively (Figure 1c,d). The diagonal and off-diagonal blocks are clearest for the radial basis function kernel. The spectrum and boundrange kernel matrices fail to capture similarities of patients except those with cancer in the brain and nervous system, where the diagnoses are very similar compared to cancers at other sites (Figure 1e,f).

4 | MODELLING WITH SUBSETS OF ICD CODES AS COVARIATES

4.1 | Set-up

Consider the set-up for regression and classification problems with ICD codes as covariates. Let $n$ be the sample size and $(y_i, x_i, T_i)$ be the data for subject $i$ ($i = 1, ..., n$), where $y_i$ is the response, $x_i^\top = (x_{i1}, ..., x_{ip})$ is the vector of $p$-dimensional covariates excluding ICD codes with $x_{i1} = 1$ and $T_i = \{t_{i1}, ..., t_{ic}\} \in (2^C)^C$. For every $i$, $t_{ic}$ denotes the diagnoses of subject $i$ that belong to chronic condition $c$, where $t_{ic} = 0$ denotes the absence of condition $c$ ($c = 1, ..., C$). In regression and classification problems, $y_i \in \mathbb{R}$ and $y_i \in \{0, 1\}$, respectively, and $(x_{i2}, ..., x_{ik})^\top \in \mathbb{R}^{p-1}$ for every $i$. We also predict the responses and estimate the covariate effects for a given collection of covariates including ICD codes $(x^*, T^*)$ ($i = 1, ..., m$), where $T^*_i \in (2^C)^C$.

We model the effects of $x_i$ and $T_i$ independently. Let $\beta \in \mathbb{R}^p$ and $f : (2^C)^C \rightarrow \mathbb{R}$. Then, the covariate effects of $x_i$ and $T_i$ are $x_i^\top \beta$ and $f(T_i)$, respectively, and the overall covariate effect for subject $i$ as $x_i^\top \beta + f(T_i)$. In practice, $\beta$ and $f$ are unknown. Choosing a Bayesian approach, we estimate the posterior distributions of $\beta$ and $f$ given $\{(y_i, x_i, T_i)\}^n_{i=1}$. We use a parametric model for estimating the effects of $x_i$ because our focus is on nonparametric regression and classification using $T \in (2^C)^C$. The Supporting Information details the model extension that replaces $x_i^\top \beta + f(T_i)$ by $g(x_i, T_i)$, where $g$ is a function estimated using $\{(y_i, x_i, T_i)\}^n_{i=1}$.

4.2 | Regression using subsets of ICD codes as covariates

Consider the GP-based nonparametric regression model. Let $\epsilon_i$ denote the idiosyncratic error for subject $i$, $\epsilon_i$s are independent and identically distributed (iid) with mean 0 and $\tau^2 > 0$ denotes the error variance. Then, the model assumes that

$$y_i = x_i^\top \beta + f(T_i) + \epsilon_i, \ \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \tau^2), \ \text{i} = 1, ..., n, \quad (8)$$

where $N(a,b)$ denotes the Gaussian density with mean $a$ and variance $b$. We put a GP prior on $f(\cdot)$ with 0 mean function and covariance function $\kappa_f(\cdot, \cdot | \theta)$ on $(2^C)^C \times (2^C)^C$, where $\theta$ are the parameters specifying $\kappa_f$. The model is completed by putting a prior distribution on $(\beta, \tau^2, \theta)$ and its form depends on the choice of $\kappa_f$.

We develop posterior inference algorithms for $\kappa_f$ equalling $\kappa^j_f$ defined in Proposition 1. We exclude the polynomial kernel because it yields a finite dimensional regression model, which is less flexible than the one based on $\kappa^j_f$. Setting $w = (1, ..., 1)$ and $A_b$ to be an identity matrix ($b \in B$) in (6) and (7), we define

$$\kappa^j_{\text{exp}}(T, T^*) = \kappa^j_1(T, T^*), \quad \kappa^j_{\text{SE}}(T, T^*) = \kappa^j_2(T, T^*), \quad T, T^* \in (2^C)^C. \quad (9)$$

The parameter for $\kappa^j_{\text{exp}}$ and $\kappa^j_{\text{SE}}$ is $\theta = (\alpha^2, \phi)$, where $\sigma^2, \phi$ are positive scalars. Denote the kernel matrix as $K_\theta$ with $K_\theta = \kappa_{\text{exp}}(T, T^*)$ or $K_\theta = \kappa_{\text{SE}}(T, T^*)$, where $i, j \in \{1, ..., n\}$. Let $y = (y_1, ..., y_n)^\top$, $X$ be the $n \times p$ matrix with $x_i^\top$ as its ith row, $f = (f(T_1), ..., f(T_n))^\top$ and $\epsilon = (\epsilon_1, ..., \epsilon_n)^\top$. Then, the hierarchical Bayesian model for $y$ based on (8) and parameter expansion is defined as

$$y = X\theta + f + \epsilon, \ \epsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I), \quad (\beta, \sigma^2) \sim \mathcal{N}(0, \sigma^2 I), \quad f \mid \theta \sim \mathcal{N}(0, K_\theta), \quad (10)$$

where $\beta$ and $f$ are assumed to be independent a priori, $\alpha = \tau^2/\sigma^2$ is the inverse of signal-to-noise ratio and $\phi \in (a_\phi, b_\phi)$, where $0 < a_\phi < b_\phi$. We assign a prior on $(\phi, \alpha)$ through $(u_1, u_2) \in \mathbb{R}^2$ as

$$\phi = a_\phi + (b_\phi - a_\phi)/(1 + e^{-u_1}), \quad \alpha = e^{u_1}, \quad (u_1, u_2)^\top \sim \mathcal{N}(0, \text{diag}(b_1, b_2)), \quad (11)$$
where $b_1, b_2 > 0$. The parameters $\sigma^2, \phi$ are non-identified in $\kappa_{\exp, \kappa_{SE}}$, but this does not affect the inference on $f$ or prediction of the response (Rasmussen & Williams, 2006).

The MCMC algorithm for posterior inference and predictions in (10) is a Gibbs-type sampling algorithm, with a step that draws $(\phi, \alpha)$ using elliptical slice sampling (ESS) (Nishihara et al., 2014). Let $f^* = ((\mathbf{T}_1^T)^\top, ..., (\mathbf{T}_n^T)^\top)^\top$ be the value of $f$ and $y^* = (y_1^*, ..., y_m^*)^\top$ be the response at the test predictors and diagnoses $(x_i^*, T_i^T)$ $(j = 1, ..., m)$. For notational convenience, define $\mathbf{K}(\phi)$, $\mathbf{K}_*$ and $\mathbf{K}_{**}$ to be the matrices satisfying $(\mathbf{K}(\phi))^T = (\kappa(T_i, T_j)/(\sigma^2), (\mathbf{K}_*)^T = (\kappa(T_i, T_j^T))/\sigma^2)$, where $i, j \in \{1, ..., n\}$ and $i, j \in \{1, ..., m\}$, and $\mathbf{C}_{yy} = \mathbf{K}(\phi) + \alpha I$. If $\mathbf{L}$ is a lower triangular matrix such that $\mathbf{C}_{yy} = \mathbf{LL}^T$, then define

$$\hat{y} = \mathbf{L}^{-1} y, \hat{X} = \mathbf{L}^{-1} X, \hat{\eta} = \mathbf{L}^{-1} \eta, \hat{\beta} = (\hat{X}^\top \hat{X})^{-1} \hat{X}^\top \hat{y}, \hat{\phi} = \hat{X} \hat{\beta}.$$ \hspace{1cm} (12)

The MCMC algorithm for drawing $\beta, f^*, K_*, y^*$ from their posterior distributions cycles through the following four steps until convergence to its stationary distribution.

1. Draw $(\sigma^2, \beta)$ given $y, \phi, \alpha$ as

$$\sigma^2 | y, \phi, \alpha \sim \frac{\| y - \hat{\phi} \|^2}{\chi^2_{n-p}}, \beta | \sigma^2, y, \phi, \alpha \sim N(\hat{\beta}, \sigma^2(\hat{X}^\top \hat{X})^{-1}) \hspace{1cm} (13)$$

where $\chi^2_{n-p}$ is the $\chi^2$ random variable with $n - p$ degrees of freedom.

2. Draw $(u_1, u_2)$ given $y, \sigma^2, \beta$ are drawn using ESS with the prior $(u_1, u_2) \sim N(0, \text{diag}(b_1, b_2))$ and the likelihood

$$l(u_1, u_2) = \frac{1}{2\sigma^2 | \text{det}(\mathbf{K}(\phi) + \alpha I) |^{1/2}} e^{-\frac{1}{2}(y - \hat{X} \phi)^\top (\mathbf{K}(\phi) + \alpha I)^{-1} (y - \hat{X} \phi)}.$$ \hspace{1cm} (14)

Define $\phi$ and $\alpha$ given $(u_1, u_2)$ as in (11) and set $\tau^2 = \sigma^2 \alpha$.

3. Draw $f^*$ given $y, \sigma^2, \phi, \alpha$ from $N(m_*, V_*)$, where

$$m_* = \mathbf{K}_*^T \mathbf{C}_{yy}^{-1} (y - \mathbf{X} \phi), V_* = \alpha^2 \left( \mathbf{K}_* - \mathbf{K}_*^T \mathbf{C}_{yy}^{-1} \mathbf{K}_* \right). \hspace{1cm} (15)$$

4. Draw $y^*$ given $X^*, \beta, f^*, \tau^2$ from $N(X^* \beta + f^* + \tau^2 I)$.

We collect MCMC draws of $\beta, f^*, y^*, \sigma^2, \phi, \tau^2$ post convergence. The derivation of this algorithm is in the Supporting Information along with the others. The algorithm in Steps 1–4 is a variant of Gibbs sampling algorithm for posterior inference in univariate spatial linear models in that we replace the Metropolis–Hastings step by an ESS step in (2) and T replaces a spatial location (Banerjee et al., 2014). Unlike the Metropolis–Hastings step, the ESS step is free of any proposal tuning, which is preferred in automated applications.

### 4.3 Classification using subsets of ICD codes as covariates

The nonparametric classification model is based on logistic regression. It assumes that

$$\log \frac{\Pr(y_i = 1)}{\Pr(y_i = 0)} = \mathbf{x}_i^\top \beta + f(T_i), \hspace{0.5cm} y_i \in \{0, 1\}, i = 1, ..., n, \hspace{1cm} (16)$$

where $y_i$ is the $i$th response. The choice of priors on $\beta, f(\cdot)$ and the kernels remain the same as in (10), but the prior on $\theta = (\phi, \sigma^2)^\top$ is assigned through $(u_1, u_2) \in \mathbb{R}^2$ as

$$\phi = a_\phi + (b_\phi - a_\phi)/(1 + e^{-u_1}), \sigma^2 = e^{u_2}, (u_1, u_2)^\top \sim N(0, \text{diag}(b_1, b_2)), \hspace{1cm} (17)$$

where $b_1, b_2 > 0$. We use Polya-Gamma data augmentation (PG-DA) for posterior inference on $\beta, f^*, K_*$ and prediction of $y^*$ (Polson et al., 2013). This set-up ensures that the MCMC algorithms for inference and predictions in classification and regression models are very similar.
Specifically, we introduce Polya-Gamma random variables $a_1, \ldots, a_n$ specific to every observation such that $a_i$ is marginally distributed as PG(1, 0), where PG is the Polya-Gamma distribution with parameters $b = 1$ and $c = 0$, respectively. Define $\bar{y} = (y_1 - 1/2, \ldots, y_n - 1/2)^T$, an $n \times 1$ vector $b = (\beta, f)^T$, an $n \times 1$ vector $\omega = (a_1, \ldots, a_n)^T$, an $n \times n$ diagonal matrix $\Omega = \text{diag}(\omega)$, pseudo responses $z = \Omega^{-1}\bar{y}$. Then, the conditional likelihood of $b$ given $y, \omega$ and the associated model are defined as

$$L(y|\omega) \propto e^{-\frac{1}{2}([Ab - z]^T \Omega^{-1}[Ab - z]),}$$

respectively. Furthermore, Theorem 1 in Polson et al. (2013) implies that $a_i$ given $b$ and the $i$th row of $\Omega$, denoted as $a_i \sim \text{PG}(1, |a_i^T b|)$.

The arguments for drawing the parameters and latent variables in (16) follow from arguments similar to those in Section 4.2. Marginalizing over $f$ in (18) implies that $z$ given $y, \omega, \phi, a^2$ is distributed as $N(\bar{X} \beta, C_{zz})$, where $C_{zz} = K_\theta + \Omega^{-1}$, and it plays the same role as $C_{zy}$ in (10). Following (12), if $L$ is a lower triangular matrix such that $C_{zz} = LL^T$, then define

$$z = L^{-T}z, \hat{X} = L^{-1}\bar{X}, \hat{\beta} = (\hat{X}^T \hat{X})^{-1}\hat{X}^T z.$$

The MCMC algorithm for drawing $\beta, f^*, K, \ldots, y^*$ from their posterior distributions cycles through the following six steps until convergence to its stationary distribution.

1. Draw $\beta$ given $y, \omega, \phi, a^2$ as $N(\bar{X}^T \bar{X})^{-1}$. 
2. Draw $(u_1, u_2)$ given $y, \omega, \beta$ using ESS following (14), where the likelihood and prior for $(u_1, u_2)$ are defined in (18) and (17). Define $\phi$ and $\sigma^2$ given $(u_1, u_2)$ using (17).
3. Draw $f$ given $y, \omega, \beta, \phi, \sigma^2$ as $N(m, \Sigma)$, where

$$V = K_\phi - K_\phi C_{zz}^{-1} K_\phi, \ m = V \Sigma (z - X \beta).$$

4. Draw $f^*$ given $y, \beta, \sigma^2, \phi$ from $N(m, \Sigma)$, where

$$m^* = K_\phi^* - K_\phi^* C_{zz}^{-1} (z - X \beta), \ V^* = K_{\theta^*} - K_{\theta^*} C_{zz}^{-1} K_{\theta^*}.$$

where $\{K_{\theta^*}\}_{i} = \kappa_i(T_i, T_i^*)$, and $\{K_{\theta^*}\}_{ij} = \kappa_i(T_j^*, T_j^*)$, where $i \in \{1, \ldots, n\}$ and $j, j' \in \{1, \ldots, m\}$.
5. Draw $y^*_i$ given $X^*, \beta^*, f^*$ from Bernoulli($p_i^*$) for $i = 1, \ldots, m$, where $p_i^* = e^{\beta^T x_i^*} / (1 + e^{\beta^T x_i^*})$ and $v_i = x_i^T \beta + (f^*)_{i}$.
6. Draw $a_i$ given $X, \beta, f$ from PG($1, |x_i^T \beta + (f)|$) for $i = 1, \ldots, n$.

We collect MCMC draws of $\beta, f^*, y^*, \sigma^2, \phi$ post convergence.

Our MCMC algorithm is similar to other algorithms based on the PG-DA strategy. For example, Polson et al. (2013) Section 5, tab. 3) present an application of the PG-DA strategy for nonparametric regression with a negative binomial response and a GP prior, but a linear predictor similar to $x_i^T \beta$ in (8) is absent in their model. Second, Wang and Roy (2018) develop a sampling algorithm based on the PG-DA strategy for posterior inference in a Bayesian logistic linear mixed model with independent and Gaussian random effects and prove its geometric ergodicity. Our MCMC algorithm is similar to theirs in that $f(\cdot)$ plays the role of nonparametric random effect, and the GP prior on $f(\cdot)$ induces dependence among patients with similar diagnoses. Overall, the MCMC algorithm above broadens the range of applications of the PG-DA strategy.

5 | EXPERIMENTS

5.1 | Set-up

We evaluate the performance of GP-based classification in (16) using three simulation studies and an analysis of the UIHC EHR data. In this section, the covariance kernel of the GP is SE, and the focus is on classification because our main goal is to assess questions relevant to the UIHC EHR data; the Supporting Information contains the results for GP-based regression using (8). In the simulated and real-data analyses, the cutoff for predicting the response is estimated using receiver operating characteristics (ROC) curve. We compare the predictive performance of all the methods using accuracy, area under the ROC (AUC), sensitivity and specificity on the test data.

We compare GP-based classification with parametric and nonparametric models. The parametric competitors are logistic regression and its regularized versions that penalize the regression coefficients using the lasso and ridge penalties. The covariates in these models are dummy
variables that indicate the presence of ICD codes or phecodes in a patient’s diagnosis. Phecodes are a phenotypic restructuring and harmonization of ICD codes and are widely used in EHR data analysis (Bastarache, 2021). We use the mapping from ICD10 codes to phecodes in Wu et al. (2018) for UIHC EHR data analysis. Because the SE kernel in (5) is a restriction of the boundrange kernel, we use support vector machine (SVM) and kernel ridge regression (KRR) with the spectrum and boundrange kernels as the nonparametric competitors. We also use the kernel function in (5) for defining feature vectors. Set \( \phi = 1, 2^1 = 1 \) in the kernel matrix \( K_\text{S} \). If \( \text{UD}^T \) is the singular value decomposition of \( K \), then the first \( r \) columns of \( \text{UD}^{1:2} \) are the feature vectors that are used as covariates in random forest, logistic regression and its penalized extensions, where \( r \) is selected using the scree plot.

All the experiments are performed in R. We used glmnet (Friedman et al., 2010) for regularized logistic regression, ranger (Wright & Ziegler, 2017) for random forest and kernlab (Karatzoglou et al., 2004) for SVM and KRR. The performance metrics are evaluated using the pROC package (Robin et al., 2011). The tuning parameters in all the methods are selected using the recommended settings. We use the spectrum and boundrange kernels in kernlab with 3 and 4 as the length of the matching substring, respectively. This means that the matches in the two kernels include the first three and four contiguous characters in a pair of ICD codes, which are also the most informative (Calderón-Larrañaga et al., 2017); therefore, we expect the results for kernlab and our method to be very similar. We use ESS by setting \( a_0, b_0, b_1 \) and \( b_2 \) in (17) to be 0.5, 1 and 2, respectively. The MCMC algorithms run for 10,000 iterations. The draws from the first 5000 iterations are discarded as burn-ins, and every fifth draw in the remaining chain is chosen for posterior inference and prediction.

5.2 Simulated data analysis

We present three simulation studies. The first two are based on simple parametric models, whereas the third is based on a nonparametric model based on the UIHC EHR data. We expect all methods to perform comparably in the first two simulations and differ significantly in the third one. All the simulations are replicated 10 times.

5.2.1 First two simulations

Both simulations have four chronic conditions. The first chronic condition is denoted using the codes in \( \{A, B, AA, BB, BA, AB\} \), where every code denotes a ‘diagnosis’. Replacing \( \{A, B\} \) with \( \{C, D\} \), \( \{E, F\} \) and \( \{G, H\} \) in this set gives the set of codes for the second, third and fourth chronic conditions, respectively. The chronic conditions are further structured into two groups: the first group includes the first and third chronic conditions, and the second group includes the remaining two. We assign nine codes to every patient, where six out of the nine codes are from the 12 codes defining the first or second group of chronic conditions and the remaining three come from the other chronic condition group. The probability of \( y = 1 \) is higher than that of \( y = 0 \) if codes from the second chronic condition group are in the majority and vice versa.

The first simulation study is based on logistic regression. Let \( z_{ij}^{(c)} \) be a dummy variable indicating the presence of the \( j \)th code in the \( c \)th chronic condition in the \( i \)th patient \( c = 1, \ldots, 4; j = 1, \ldots, 6; i = 1, \ldots, n + m \), where codes follow the dictionary order, \( n \) is the training data size and \( m \) is the testing data size. The first and second group dummy variables for the \( i \)th patient are \( \{z_{ij}^{(c)} : j = 1, \ldots, 6; c = 1, 3\} \) and \( \{z_{ij}^{(c)} : j = 1, \ldots, 6; c = 2, 4\} \). The response \( y_i \) is simulated independently from Bernoulli \( (p_i) \), where

\[
\log \frac{p_i}{1-p_i} = 0.1 + 0.2x_i + \sum_{c=1}^{4} \sum_{j=1}^{6} (-1)^{c-1} z_{ij}^{(c)} x_i \overset{\text{ind}}{\sim} N(0,1),
\]

for \( i = 1, \ldots, n + m \). The second simulation is a slight modification of the first. Dropping the \( x_i \beta \) term in (22), we modify it to only include the interaction of \( \{AB, BA\}, \{CD, DC\}, \{EF, FE\} \) and \( \{GH, HG\} \), respectively, as

\[
\log \frac{p_i}{1-p_i} = 2 \left( z_{ij}^{(2)} z_{ij}^{(2)} + z_{ij}^{(4)} z_{ij}^{(4)} \right) - \left( z_{ij}^{(1)} z_{ij}^{(1)} + z_{ij}^{(3)} z_{ij}^{(3)} \right),
\]

and \( y_i = 1(p_i > 0.5) \) for \( i = 1, \ldots, n + m \).

In the first two simulations, GP with SE covariance kernel is among the best performers (Table 1). The first simulation has a linear decision boundary in terms of the 24 codes. On the other hand, the second simulation has a quadratic decision boundary depending on the interactions between \( \{AB, BA\}, \{CD, DC\}, \{EF, FE\} \) and \( \{GH, HG\} \). Logistic regression and its regularized versions perform the best in the first simulation, but their performance deteriorates slightly in the second simulation due to the presence of interactions between dummy variables. Random forest using the codes as covariates performs worst in both simulations. After including SE kernel-based covariates, random forest achieves the same
### Table 1 Performance comparisons for the three simulation studies.

#### First simulation

| Method                                      | AUC  | Accuracy | Sensitivity | Specificity |
|---------------------------------------------|------|----------|-------------|-------------|
| SE-GP                                       | .96  | .96      | .96         | .96         |
| Logistic regression with SE kernel-based covariates |      |          |             |             |
| No penalty                                  | .95  | .95      | .96         | .95         |
| Ridge penalty                               | .95  | .96      | .96         | .95         |
| Lasso penalty                               | .95  | .95      | .96         | .95         |
| Logistic regression                         |      |          |             |             |
| No penalty                                  | .96  | .96      | .96         | .96         |
| Ridge penalty                               | .96  | .96      | .96         | .96         |
| Lasso penalty                               | .96  | .96      | .96         | .96         |
| Kernel ridge regression                     |      |          |             |             |
| Spectrum                                    | .95  | .96      | .96         | .95         |
| Boundrange                                  | .95  | .95      | .96         | .95         |
| Support vector machine                      |      |          |             |             |
| Spectrum                                    | .95  | .96      | .96         | .96         |
| Boundrange                                  | .95  | .95      | .96         | .95         |
| Random forest                               |      |          |             |             |
| Default                                     | .50  | .50      | .99         | .01         |
| SE kernel-based covariates                  | .95  | .95      | .96         | .95         |

#### Second simulation

| Method                                      | AUC  | Accuracy | Sensitivity | Specificity |
|---------------------------------------------|------|----------|-------------|-------------|
| SE-GP                                       | 1.00 | .99      | 1.00        | .98         |
| Logistic regression with SE kernel-based covariates |      |          |             |             |
| No penalty                                  | .90  | .82      | .77         | .86         |
| Ridge penalty                               | .90  | .82      | .81         | .84         |
| Lasso penalty                               | .90  | .82      | .75         | .89         |
| Logistic regression                         |      |          |             |             |
| No penalty                                  | .90  | .82      | .78         | .86         |
| Ridge penalty                               | .90  | .82      | .80         | .84         |
| Lasso penalty                               | .90  | .83      | .77         | .87         |
| Kernel ridge regression                     |      |          |             |             |
| Spectrum                                    | .92  | .86      | .86         | .86         |
| Boundrange                                  | .91  | .84      | .83         | .85         |
| Support vector machine                      |      |          |             |             |
| Spectrum                                    | .92  | .86      | .86         | .86         |
| Boundrange                                  | .91  | .84      | .83         | .85         |
| Random forest                               |      |          |             |             |
| Default                                     | .50  | .45      | 1.00        | .00         |
| SE kernel-based covariates                  | .88  | .88      | .85         | .92         |

#### Third simulation

| Method                                      | AUC  | Accuracy | Sensitivity | Specificity |
|---------------------------------------------|------|----------|-------------|-------------|
| SE-GP                                       | .94  | .91      | .90         | .92         |
| Logistic regression with SE kernel-based covariates |      |          |             |             |
| No penalty                                  | -    | -        | -           | -           |
| Ridge penalty                               | .76  | .73      | .66         | .82         |
| Lasso penalty                               | .75  | .73      | .67         | .81         |
5.2.2 Third simulation

For this simulation, we used ICD codes of patients in the UIHC EHR data with the brain and other nervous system or breast as the cancer primary sites. For the \(i^{th}\) (pseudo) patient in this subset, let \(z_i = \|\mathbf{y}(T_i)\|^2\) and \(\delta_i = 1\) if the patient with \(T_i\) code has brain and other nervous system cancer and \(\delta_i = -1\) otherwise, where \(\mathbf{y}(T_i)\) is defined in (3). Because \(T_i\)'s are included in the simulation, we use logistic regression models with phecodes and \(\log(1 + \text{phecode count})\) as predictors. The response \(y_i\) is simulated independently from Bernoulli (\(p_i\)), where

\[
\log \frac{p_i}{1 - p_i} = 0.1 + 0.2x_i + 3\tan(z_i) + 3\delta_i, \quad x_i \sim N(0, 1),
\]

for \(i = 1, \ldots, n + m\). We simulate the data after setting \(n = 1000\) and \(m = 100\).

GP with SE covariance kernel is still among the top performers (Table 1). This simulation has a non-linear periodic decision boundary; therefore, the performance of random forest and parametric models, including logistic regression and its regularized, deteriorate significantly. Including phecodes as predictors fails to yield any noticeable difference in the performance. While all the kernel-based methods are suited for modelling the non-linear periodic decision boundary in (24), the SE kernel is better tuned than spectrum and boundrange kernels for modelling the hierarchical structure of ICD codes. This implies that all kernel-based methods perform well, but the GP with SE kernel performs better than SVM and KRR with spectrum and boundrange kernels.

### TABLE 1 (Continued)

| Third simulation | AUC | Accuracy | Sensitivity | Specificity |
|------------------|-----|----------|-------------|-------------|
| Logistic regression |     |          |             |             |
| No penalty       | .89 | .84      | .85         | .84         |
| Ridge penalty    | .90 | .86      | .85         | .88         |
| Lasso penalty    |     |          |             |             |
| Logistic regression with phecodes |     |          |             |             |
| Lasso penalty    | .90 | .86      | .85         | .85         |
| Ridge penalty    |     |          |             |             |
| Logistic regression with \(\log(1 + \text{phecode count})\) |     |          |             |             |
| No penalty       | .56 | .52      | .61         | .57         |
| Kernel ridge regression |     |          |             |             |
| Spectrum         | .84 | .80      | .82         | .78         |
| Boundrange       | .81 | .78      | .81         | .74         |
| Support vector machine |     |          |             |             |
| Spectrum         | .85 | .81      | .82         | .81         |
| Boundrange       | .83 | .79      | .77         | .81         |
| Random forest    |     |          |             |             |
| Default          | .85 | .86      | .95         | .76         |
| SE kernel-based covariates | .89 | .89      | .91         | .88         |

Note: Every entry in the table is the average of its values across 10 replications. SE-GP is the proposed method, and SE kernel-based features are obtained from the kernel matrix estimated using the proposed method. If a method fails to produce results, then we indicate this using ‘\(-\)’.
### TABLE 2  Performance in the classification model with $y = 1$ denoting the presence of cancer in the brain and other nervous system.

|                         | AUC  | Accuracy | Sensitivity | Specificity |
|-------------------------|------|----------|-------------|-------------|
| SE-GP                   | .93  | .92      | .86         | .93         |
| Logistic regression     |      |          |             |             |
| Logistic regression with SE kernel-based covariates |      |          |             |             |
| No penalty              | .91  | .88      | .82         | .89         |
| Lasso penalty           | .91  | .89      | .81         | .91         |
| Ridge penalty           | .90  | .90      | .81         | .91         |
| Logistic regression     |      |          |             |             |
| Logistic regression with phecode |      |          |             |             |
| Lasso penalty           | .93  | .91      | .86         | .92         |
| Ridge penalty           | .94  | .92      | .89         | .92         |
| Logistic regression     |      |          |             |             |
| Logistic regression with log(1 + phecode count) |      |          |             |             |
| No penalty              | .53  | .42      | .74         | .39         |
| Support vector machine  |      |          |             |             |
| Boundrange              | .94  | .91      | .88         | .92         |
| Spectrum                | .94  | .90      | .90         | .90         |
| Random forest           |      |          |             |             |
| Default                 | .50  | .11      | 1.00        | .00         |
| SE kernel-based covariates | .72  | .92      | .45         | .98         |

Note: Every entry in the table is the average of its values across 10 replications.

### TABLE 3  Performance in the classification model with $y = 1$ denoting the presence of cancer in the breast.

|                         | AUC  | Accuracy | Sensitivity | Specificity |
|-------------------------|------|----------|-------------|-------------|
| SE-GP                   | .77  | .72      | .75         | .71         |
| Logistic regression     |      |          |             |             |
| Logistic regression with SE kernel-based covariates |      |          |             |             |
| No penalty              | .69  | .54      | .80         | .51         |
| Lasso penalty           | .68  | .59      | .74         | .56         |
| Ridge penalty           | .68  | .58      | .74         | .56         |
| Logistic regression     |      |          |             |             |
| Logistic regression with phecode |      |          |             |             |
| Lasso penalty           | .74  | .66      | .74         | .66         |
| Ridge penalty           | .73  | .72      | .69         | .73         |
| Logistic regression     |      |          |             |             |
| Logistic regression with log(1 + phecode count) |      |          |             |             |
| No penalty              | .52  | .53      | .43         | .69         |
| Support vector machine  |      |          |             |             |
| Boundrange              | .68  | .67      | .63         | .68         |
| Spectrum                | .74  | .67      | .75         | .66         |
| Random forest           |      |          |             |             |
| Default                 | .50  | .11      | 1.00        | .00         |
| SE kernel-based covariates | .53  | .80      | .18         | .88         |

Note: Every entry in the table is the average of its values across 10 replications.
### TABLE 4: Performance in the classification model with $y = 1$ denoting the presence of cancer in the urinary system.

| Method                                                                 | AUC | Accuracy | Sensitivity | Specificity |
|-----------------------------------------------------------------------|-----|----------|-------------|-------------|
| Logistic regression with SE kernel-based covariates                   |     |          |             |             |
| No penalty                                                            | .75 | .68      | .71         | .68         |
| Lasso penalty                                                         | .75 | .70      | .70         | .70         |
| Ridge penalty                                                         | .75 | .75      | .63         | .77         |
| Logistic regression                                                    |     |          |             |             |
| Lasso penalty                                                         | .69 | .89      | .36         | .97         |
| Ridge penalty                                                         | .74 | .76      | .68         | .77         |
| Logistic regression with phecode                                       |     |          |             |             |
| Lasso penalty                                                         | .66 | .83      | .41         | .88         |
| Ridge penalty                                                         | .69 | .72      | .65         | .73         |
| Logistic regression with log(1 + phecode count)                        |     |          |             |             |
| No penalty                                                            | .54 | .58      | .55         | .56         |
| Support vector machine                                                |     |          |             |             |
| Boundrange                                                            | .60 | .71      | .43         | .75         |
| Spectrum                                                              | .68 | .76      | .53         | .79         |
| Random forest                                                         |     |          |             |             |
| Default                                                               | .50 | .12      | 1.00        | .00         |
| SE kernel-based covariates                                            | .58 | .89      | .17         | .98         |

Note: Every entry in the table is the average of its values across 10 replications.

### TABLE 5: Performance in the classification model with $y = 1$ denoting the presence of cancer in the respiratory system.

| Method                                                                 | AUC | Accuracy | Sensitivity | Specificity |
|-----------------------------------------------------------------------|-----|----------|-------------|-------------|
| Logistic regression with SE kernel-based covariates                   |     |          |             |             |
| No penalty                                                            | .67 | .56      | .78         | .52         |
| Lasso penalty                                                         | .67 | .57      | .76         | .53         |
| Ridge penalty                                                         | .67 | .56      | .78         | .52         |
| Logistic regression                                                    |     |          |             |             |
| Lasso penalty                                                         | .81 | .74      | .73         | .74         |
| Ridge penalty                                                         | .83 | .75      | .83         | .74         |
| Logistic regression with phecode                                       |     |          |             |             |
| Lasso penalty                                                         | .80 | .73      | .75         | .72         |
| Ridge penalty                                                         | .79 | .74      | .76         | .74         |
| Logistic regression with log(1 + phecode count)                        |     |          |             |             |
| No penalty                                                            | .58 | .64      | .50         | .65         |
| Support vector machine                                                |     |          |             |             |
| Boundrange                                                            | .77 | .70      | .73         | .70         |
| Spectrum                                                              | .82 | .75      | .78         | .74         |
| Random forest                                                         |     |          |             |             |
| Default                                                               | .50 | .16      | 1.00        | .00         |
| SE kernel-based covariates                                            | .55 | .84      | .13         | .97         |

Note: Every entry in the table is the average of its values across 10 replications.
TABLE 6  Performance in the classification model with $y = 1$ denoting the presence of cancer in female genital system.

| Method                                                | AUC  | Accuracy | Sensitivity | Specificity |
|-------------------------------------------------------|------|----------|-------------|-------------|
| SE-GP                                                 | .86  | .83      | .72         | .86         |
| Logistic regression with SE kernel-based covariates   |      |          |             |             |
| No penalty                                            | .78  | .73      | .66         | .75         |
| Lasso penalty                                         | .78  | .75      | .65         | .78         |
| Ridge penalty                                         | .78  | .75      | .66         | .78         |
| Logistic regression                                   |      |          |             |             |
| Lasso penalty                                         | .84  | .77      | .75         | .77         |
| Ridge penalty                                         | .85  | .81      | .77         | .82         |
| Logistic regression with phecode                      |      |          |             |             |
| Lasso penalty                                         | .81  | .72      | .79         | .71         |
| Ridge penalty                                         | .81  | .77      | .76         | .77         |
| Logistic regression with $\log(1 + \text{phecode count})$ |      |          |             |             |
| No penalty                                            | .60  | .54      | .68         | .50         |
| Support vector machine                                |      |          |             |             |
| Boundrange                                            | .77  | .75      | .68         | .77         |
| Spectrum                                              | .83  | .80      | .70         | .83         |
| Random forest                                          |      |          |             |             |
| Default                                               | .50  | .21      | 1.00        | .00         |
| SE kernel-based covariates                            | .70  | .84      | .45         | .95         |

Note: Every entry in the table is the average of its values across 10 replications.

TABLE 7  Summary of the marginal associations between 58 chronic conditions and the six primary cancer sites.

| Kernel                        | Chronic condition               | Estimate         |
|-------------------------------|---------------------------------|------------------|
| Brain and other nervous system| Other neurological diseases      | $0.8075, (0.5199, 0.9649)$ |
| SE-GP                         | Other neurological diseases      | $0.8057$         |
| Boundrange                    | Other neurological diseases      | $0.6570$         |
| Spectrum                      | Other neurological diseases      |                  |
| Breast                        | Depression and mood diseases    | $0.6909, (0.2960, 0.9451)$ |
| SE-GP                         | Other cardiovascular diseases    | $0.5799$         |
| Boundrange                    | Osteoporosis                     | $0.5884$         |
| Spectrum                      | Allergy                          | $0.6005$         |
| Spectrum                      | Depression and mood diseases     | $0.6454$         |
| Spectrum                      | Venous and lymphatic diseases    | $0.7290$         |
| Urinary system                |                                 |                  |
| SE-GP                         | Other respiratory diseases       | $0.7973, (0.5421, 0.9483)$ |
| Boundrange                    | COPD, emphysema and chronic bronchitis | $0.8815, (0.6181, 0.9880)$ |
| Spectrum                      | Other respiratory diseases       | $0.6901$         |
| Respiratory system            | COPD, emphysema and chronic bronchitis | $0.7575$         |
| SE-GP                         | Other respiratory diseases       | $0.5024$         |
5.2.3 | Summary

Our simulation studies suggest that kernel-based methods are better suited for applications in practice where we expect interactions among the diagnoses. Furthermore, GP with the SE kernel is easily extended to account for any biomedical information, is tuned for modelling the structure of ICD codes and produces similar results as SVM or KRR with spectrum and boundrange kernels in the absence of any additional structure. The distinguishing feature of the proposed method is that the posterior MCMC draws can be used for quantifying uncertainty in predictions and parameter estimates, which is key in biomedical applications; therefore, we conclude that the GP with SE covariance kernel is better than other kernel-based methods in quantifying uncertainty and in modelling the effect of chronic conditions on the response.

5.3 | Real-data analysis

The UIHC EHR data contains information about 1660 patients, including their diagnoses, marital status and the primary cancer site. There are six types of cancer sites: brain and other nervous system (brain), breast, urinary system, respiratory system, female genital system and digestive system. Calderón-Larrañaga et al. (2017) define 58 chronic conditions using ICD codes that do not include any diagnosis related to neoplasms or primary site. For the feature map in (6) that defines the SE kernel, we set \( C = 58, R = I \) and \( w_c \propto \kappa_1(T_c, T_c) \), where \( T_c \) is the set of ICD codes defining...
c-th chronic condition. We include marital status as the only demographic predictor, so \( p = 2 \) in (16). The major and minor goals of the analysis are to estimate the marginal associations of the 58 chronic conditions with the cancer sites and to predict the cancer primary site using the patient diagnoses and marital status, respectively.

We use the methods from the previous section to achieve both goals. Five classification models are used for predicting the six primary sites using \( y = 1 \) to denote the presence of cancer in brain, breast, urinary system, respiratory system and female genital system, respectively, in the five models. As in the third simulation, we also include logistic regression models with phecodes and \( \log(1 + \text{phecode count}) \) as predictors. For every model, all the methods are trained on 80% of the full data, their performance is evaluated on the excluded data and the set-up is replicated 10 times. There is no principled way of using chronic conditions as covariates in logistic regression and its penalized extension, so we only use KRR, SVM and SE-GP for estimating the marginal associations between the 58 chronic conditions and primary cancer site. To this end, we include \( T_1, \ldots, T_{58} \) as the diagnoses of 58 pseudo patients in the test data and summarize the marginal associations using the estimate of \( \Pr(y = 1 | T_i) \), where \( T_i \) is the diagnosis for the c-th pseudo patient \( (c = 1, \ldots, 58) \). Unlike SVM and KRR, SE-GP also provides 95% credible intervals for the marginal associations using the MCMC draws of \( \Pr(y = 1) \).

SE-GP is among the top performers in all five classification models (Tables 2–6). We do not provide results for KRR and logistic regression with no penalty because the former fails due to a line search error, and the latter cannot be used because the number of dummy variables is larger than the sample size. Calderón-Larrañaga et al. (2017) have developed a comprehensive list of chronic conditions after a careful study, so we expect a subset of them to be related to primary cancer sites. This is confirmed by relatively high accuracy of logistic regression with the ridge and lasso penalties for predicting cancer in the brain and urinary system; however, these methods do not use the information encoded in ICD codes, so their sensitivity and specificity vary a lot; for example, logistic regression with the lasso penalty has a high accuracy of .90 for predicting cancer in urinary system, but its specificity in the same model is .36.

On the other hand, SE-GP is among the top performers in terms of accuracy, sensitivity and specificity in all the five classification models. This also implies that the AUC of SE-GP is the highest among all methods in all the five classification models. The features obtained from the SE kernel matrix are also promising in that if we use them as covariates in logistic regression and its penalized extensions, then the AUCs are relatively large in all the five classification models. Logistic regression models with phecode predictors also perform similarly, whereas those with \( \log(1 + \text{phecode count}) \) perform worse. The performance of the default random forest is the worst in all five classification models. AUC and accuracy increase after including SE kernel-based covariates, but the results are still poorer than all other methods.

SE-GP also performs better than SVM with spectrum and boundrange kernels because SE kernel accounts for the additional structure provided by the 58 chronic conditions. The SE-GP is a restriction of the boundrange kernel that is tuned for EHR data analysis, so its sensitivity and specificity are much higher than those of the SVM in all the five classification models. Most importantly, the estimates of marginal associations between chronic conditions and primary cancer sites agree closely with those obtained using the SVM with spectrum and boundrange kernels (Table 7). A key feature of the SE-GP approach is that it provides 95% credible intervals in addition to the point estimates, which are important for characterizing uncertainty in biomedical applications. Additionally, these credible intervals cover the corresponding estimates obtained using SVM. Based on our simulation and real-data analysis results, we conclude that SE-GP outperforms its competitors in estimating the marginal associations among chronic conditions and primary cancer sites and in predicting the cancer primary site using diagnoses and demographic information as predictors.

## DISCUSSION

We have introduced a novel GP-based approach for nonparametric regression and classification using ICD codes as covariates. Unlike existing methods that mainly rely on summary statistics, including phecode-based predictors, the proposed method leverages the hierarchical structure of ICD codes. This leads to better predictive performance compared to the traditional methods. A distinguishing aspect of our approach is using kernel functions to exploit the similarity between ICD codes based on their relation to a chronic condition. This idea is critical in estimating and utilizing the marginal associations between chronic conditions and outcomes, such as phenotypes, survival status, symptoms and treatment effectiveness. We have developed efficient Monte Carlo algorithms for automated applications of the proposed method in biomedical settings.

Our approach can be extended in several ways. First, the sample size in the motivating application is small, so repeated computations and evaluation of the kernel function is relatively inexpensive. This, however, becomes problematic as the sample size becomes moderately large. The kernels developed in this work can be immediately extended for nonparametric regression and classification in large sample settings using low-rank kernels based on inducing points (Quiñonero-Candela & Rasmussen, 2005). Second, every element of the kernel matrix is a similarity measure between a pair of patients. This can be used as input in an algorithm for unsupervised learning using ICD codes, such as clustering of subsets, data visualization and dimension reduction (Shawe-Taylor & Cristianini, 2004, Chapters 6 and 8). The PG-DA strategy outlined in Section 4.3 is easily extended to model zero-inflated negative binomial responses following Neelon (2019). Finally, we are exploring the application of the product partition model with regression on diagnoses, where the similarity measure on diagnoses is defined using the kernel matrix.
AUTHOR CONTRIBUTIONS
Sanvesh Srivastava conceived the Gaussian process regression and classification models, developed the Gibbs sampling algorithms, implemented them in R and wrote the first draft of the manuscript. Zongyi Xu conducted the simulation and real-data analyses and wrote parts of the experiments section. Yunyi Li co-developed the similarity measures on ICD code subsets and implemented them in R. W. Nick Street co-authored the manuscript, especially the presentation of the practical impacts of the methods. Stephanie Gilbertson-White posed the question of developing a regression model where ICD codes are predictors and co-authored the manuscript.

ACKNOWLEDGEMENTS
We thank UIHC for sharing the EHR data. Sanvesh Srivastava’s research is supported by grants from the Office of Naval Research (ONR-BAA N000141812741) and the National Science Foundation (DMS-1854667/1854662). Stephanie Gilbertson-White’s research is supported by the University of Iowa, College of Nursing Center on Advancing Multimorbidity Science (CAMS) funded by the National Institute for Nursing Research (P20NR018081). The R code for simulations is available at https://github.com/blayes/ehr-data-analysis.

CONFLICT OF INTEREST STATEMENT
The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT
Data sharing is not applicable to this article as no new data were created or analysed in this study.

ORCID
Sanvesh Srivastava https://orcid.org/0000-0002-5483-9579

REFERENCES
Ahuja, Y., Zou, Y., Verma, A., Buckeridge, D., & Li, Y. (2022). MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record. Journal of Biomedical Informatics, 134, 104190.
Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). Hierarchical modeling and analysis for spatial data: CRC Press.
Bastarache, L. (2021). Using phecodes for research with the electronic health record: From PheWAS to PheRS. Annual Review of Biomedical Data Science, 4, 1–19.
Calderón-Larrañaga, A., Vetrano, D. L., Onder, G., Gimeno-Feliu, L. A., Coscollar-Santaliestra, C., Carfi, A., Pisciotta, M. S., Angleman, S., Melis, R. J. F., & Santoni, G. (2017). Assessing and measuring chronic multimorbidity in the older population: A proposal for its operationalization. Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences, 72(10), 1417–1423.
Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22.
Ghosal, S., & van der Vaart, A. (2017). Fundamentals of nonparametric Bayesian inference, Vol. 44: Cambridge University Press.
Henaou, R., Lu, J. T., Lucas, J. E., Ferranti, J., & Carin, L. (2016). Electronic health record analysis via deep Poisson factor models. The Journal of Machine Learning Research, 17(1), 6422–6453.
Horng, S., Sonntag, D. A., Halpern, Y., Jernite, Y., Shapiro, N. I., & Nathanson, L. A. (2017). Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PloS One, 12(4), e0174708.
Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab—an S4 package for kernel methods in R. Journal of Statistical Software, 11(9), 1–20.
Li, Y., Nair, P., Lu, X. H., Wen, Z., Wang, Y., Dehaghi, A. A. K., Miao, Y., Liu, W., Ordog, T., & Biernacka, J. M. (2020). Inferring multimodal latent topics from electronic health records. Nature Communications, 11(1), 2536.
Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. Journal of Machine Learning Research, 2(Feb), 419–444.
Neelon, B., & van der Vaart, A. (2017). Fundamentals of nonparametric Bayesian inference, Vol. 44: Cambridge University Press.
Ni, Y., Muller, P., & Ji, Y. (2020). Bayesian double feature allocation for phenotyping with electronic health records. Journal of the American Statistical Association, 115(532), 1620–1634.
Nishihara, R., Murray, I., & Adams, R. P. (2014). Parallel MCMC with generalized elliptical slice sampling. The Journal of Machine Learning Research, 15(1), 2087–2112.
Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using polya-gamma latent variables. Journal of the American Statistical Association, 108(504), 1339–1349.
Quiñonero-Candela, J., & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. Journal of Machine Learning Research, 6(Dec), 1939–1959.
Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., & Sundberg, P. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1(1), 18.
Rasmussen, C. E., & Williams, CKI (2006). Gaussian processes for machine learning: MIT Press.
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). proc: An open-source package to analyze and compare ROC curves. BMC Bioinformatics, 12, 77.
Shawe-Taylor, J., & Cristianini, N. (2004). Kernel methods for pattern analysis: Cambridge University Press.
Song, Z., Hu, Y., Verma, A., Buckeridge, D. L., & Li, Y. (2022). Automatic phenotyping by a seed-guided topic model. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, pp. 4713–4723.

Wang, X., & Roy, V. (2018). Analysis of the Polya-Gamma block Gibbs sampler for Bayesian logistic linear mixed models. Statistics & Probability Letters, 137, 251–256.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77(1), 1–17.

Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Bastarache, L., Denny, J. C., Theodoratou, E., & Wei, W.-Q. (2018). Developing and evaluating mappings of ICD-10 and ICD-10-CM codes to phecodes. bioRxiv.

Zou, Y., Pesaran, A., Song, Z., Verma, A., Buckeridge, D. L., & Li, Y. (2022). Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model. Scientific Reports, 12(1), 17868.

AUTHOR BIOGRAPHIES

Sanvesh Srivastava is a tenured associate professor in Statistics and Actuarial Science at the University of Iowa. Specializing in Bayesian methods and computational algorithms for big data, his research is particularly tailored for genomics, medical imaging and electronic health records data. His research focuses on developing flexible and efficient Bayesian models and Monte Carlo algorithms for complex and large data sets, leveraging distributed computations, latent dependencies and nonparametric modelling. His research has been supported by grants from several US federal funding agencies.

Zongyi Xu is a Ph.D. student in the Department of Statistics and Actuarial Science at the University of Iowa. His research interest is in targeted Bayesian applications for electronic health records data analysis.

Yunyi Li is a Ph.D. student at UT Austin's McCombs School of Business and specializes in fairness and interpretability in socio-technical systems. Her research addresses bias in supervised learning data labels and its downstream impact. She also explores methods for mitigating label bias and enhancing bias-aware data collection. Her work aims to bring accountability to machine learning systems through algorithmic and behavioural approaches.

W. Nick Street is the Henry B. Tippie Research Professor in Business Analytics and associate dean for research and Ph.D. programmes in the Tippie College of Business, with joint appointments in the Computer Science Department and the College of Nursing. His research interests are in algorithmic approaches to machine learning and data mining, applied to business and healthcare analytics problems. His recent work has focused on ensemble construction methods, federated learning, counterfactual reasoning and personalized healthcare decision making. He has published over 130 journal, conference and workshop papers and is the prior recipient of the INFORMS Data Mining Prize and an NSF CAREER award.

Stephanie Gilbertson-White is a tenured associate professor in Nursing at the University of Iowa, specializing in cancer symptom science. With over 23 years of clinical and 19 years of research experience, her primary goals are to understand cancer symptom trajectories and mechanisms and to offer tailored symptom management for underserved advanced cancer patients. Her research encompasses patient-reported biomarkers of immunoreactivity and currently focuses on psychological and biological factors leading to distressing symptoms in patients with cancer and multiple chronic conditions. She also tests interventions for rural cancer patients through her project, OASIS. Concurrently, she practices as an advanced practice registered nurse in palliative care at the Holden Comprehensive Cancer Center, University of Iowa Hospital and Clinics.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Srivastava, S., Xu, Z., Li, Y., Street, W. N., & Gilbertson-White, S. (2023). Gaussian process regression and classification using International Classification of Disease codes as covariates. Stat, 12(1), e618. https://doi.org/10.1002/sta4.618