1-18-2024

Revealing chronic disease progression patterns using Gaussian process for stage inference.

Yanfei Wang
University of Texas Health Science Center at Houston, School of Health Information Sciences, Houston TX, USA

Weiling Zhao

Angela Ross

Lei You

Hongyu Wang

See next page for additional authors

Recommended Citation

Wang, Yanfei; Zhao, Weiling; Ross, Angela; You, Lei; Wang, Hongyu; and Zhou, Xiaobo, "Revealing chronic disease progression patterns using Gaussian process for stage inference." (2024). *Journal Articles*. 90.
https://digitalcommons.library.tmc.edu/uthshis_docs/90

Journal of the American Medical Informatics Association, 2024, **31(2)**, 396–405
https://doi.org/10.1093/jamia/ocad230
Advance access publication 5 December 2023
Research and Applications

Research and Applications

Revealing chronic disease progression patterns using Gaussian process for stage inference

Yanfei Wang, MS[1], Weiling Zhao, PhD[1], Angela Ross, PhD[1], Lei You, PhD[1], Hongyu Wang, PhD[2,3], Xiaobo Zhou, PhD[1,*]

[1]Center for Computational Systems Medicine, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, United States, [2]McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, United States, [3]Cizik School of Nursing, The University of Texas Health Science Center at Houston, Houston, TX 77030, United States

*Corresponding author: Xiaobo Zhou, PhD, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St., Houston, TX 77030 (Xiaobo.Zhou@uth.tmc.edu)

Abstract

Objective: The early stages of chronic disease typically progress slowly, so symptoms are usually only noticed until the disease is advanced. Slow progression and heterogeneous manifestations make it challenging to model the transition from normal to disease status. As patient conditions are only observed at discrete timestamps with varying intervals, an incomplete understanding of disease progression and heterogeneity affects clinical practice and drug development.

Materials and Methods: We developed the Gaussian Process for Stage Inference (GPSI) approach to uncover chronic disease progression patterns and assess the dynamic contribution of clinical features. We tested the ability of the GPSI to reliably stratify synthetic and real-world data for osteoarthritis (OA) in the Osteoarthritis Initiative (OAI), bipolar disorder (BP) in the Adolescent Brain Cognitive Development Study (ABCD), and hepatocellular carcinoma (HCC) in the UTHealth and The Cancer Genome Atlas (TCGA).

Results: First, GPSI identified two subgroups of OA based on image features, where these subgroups corresponded to different genotypes, indicating the bone-remodeling and overweight-related pathways. Second, GPSI differentiated BP into two distinct developmental patterns and defined the contribution of specific brain region atrophy from early to advanced disease stages, demonstrating the ability of the GPSI to identify diagnostic subgroups. Third, HCC progression patterns were well reproduced in the two independent UTHealth and TCGA datasets.

Conclusion: Our study demonstrated that an unsupervised approach can disentangle temporal and phenotypic heterogeneity and identify population subgroups with common patterns of disease progression. Based on the differences in these features across stages, physicians can better tailor treatment plans and medications to individual patients.

Key words: disease progression; Gaussian process; unsupervised learning.

Background and significance

As electronic health records (EHRs) have become more widely adopted, many disease progression models have been developed to understand disease progression. These models serve to stage disease severity, guide management, predict outcomes, and gauge treatment impact, offering early warnings of potential decline. Young et al's "Subtype and Stage Inference"[1] (SuStaIn) algorithm exemplifies this by identifying specific patterns for a disease. However, such disease progression models often rely on linear approaches, which, despite their simplicity, may not fully capture the complex real-world disease progression.

Gaussian processes (GPs) have proven effective in revealing complex, non-linear patterns within medical data over time.[2,3] Numerous studies have employed GPs to predict critical events like readmission and mortality.[4,5] For instance, Cheng et al developed a Bayesian nonparametric model based on multi-output GP regression for hospitalized patient monitoring.[6] Meng et al advanced this work with a multivariate GP model that captures dynamic interrelations among clinical variables.[7] While the above supervised learning methods could learn a correlation structure within and between time series, they were not devised to infer disease stage information from cross-sectional data.

Significance

Chronic diseases progress continuously from health to disease rather than through sudden shifts. However, defining this progression is difficult because staging information is often lacking or not precise. For instance, the Kellgren–Lawrence (KL)[8] grading system, while prevalent in osteoarthritis (OA) research, focuses on structural changes[9] and may miss early-stage symptoms and functional decline in early stages.[10] Similarly, the Liver Imaging Reporting and Data System (LI-RADS)[11] encounters challenges in precise staging, which impacts the assessment of hepatocellular carcinoma (HCC) progression.[12,13] These cases underscore the need for refining disease staging to improve treatment.[14,15] To address these limitations, we present the Gaussian Process for Stage Inference (GPSI), an advanced method capable of inferring the

Received: August 11, 2023. **Revised:** November 6, 2023. **Editorial Decision:** November 17, 2023. **Accepted:** November 20, 2023

ordered sequence of pseudo-stages from unordered, high-dimensional data. Traditional models often fail to capture the complex progression of chronic diseases due to missing or inaccurate staging information. However, GPSI addresses this by constructing a latent space that reflects the inherent continuity of the data, preserving the proximity of related data points, and inferring a sequential order where none was previously apparent.

Gaussian Process for Stage Inference begins by hypothesizing a smooth, latent space where the proximity of points represents the data's inherent continuity. Starting with randomly assigned positions in this latent space, GPSI forms a probabilistic link to the observed data, refining each position to reflect precise probability distributions of the high-dimensional data. Through optimization that maximizes observed data likelihood, GPSI adjusts these positions to align with a logical sequence of disease progression. The result is a series of pseudo-stages, providing a framework for understanding an individual's phase in the disease process without relying on explicit staging data.

The selection of a covariance function is crucial as it articulates our presumptions about the data's sequential nature. A function that ensures high similarity for closely situated data points permits the model to support a sequential interpretation consistent with disease progression.

We apply a multi-faceted validation strategy to validate the robustness and clinical applicability of GPSI. We assess the model's stability by verifying that the inferred pseudo-stages remain consistent across different data perturbations and exhibit a high intraclass correlation when applied to various data subsets. Furthermore, the clinical relevance of GPSI is evaluated by correlating the pseudo-stages with known disease stages, using Kendall's Tau for ordinal association. Meeting these criteria would indicate that the pseudo-stages reflect the disease's latent progression pattern, not merely a product of algorithmic construction.

Methods

Intuition

Due to their inherent complexity, chronic diseases are characterized by diverse subgroups, each with a distinct progression pattern. We developed a two-stage framework for subgroup classification and trajectory mapping to address this heterogeneity. Initially, we utilized k-means clustering to group patients into more uniform subgroups based on inherent similarities. K-means is favored for its simplicity and efficiency,[16–18] making it an excellent preparatory tool for reducing the computational load of the subsequent application of complex models.

After identifying these subgroups, we estimated the stages of disease progression within each subgroup using GPSI. Gaussian Process for Stage Inference, a GP-based unsupervised learning algorithm with an informative prior, enriches our understanding of disease progression by capturing the variation in data and ensuring stability across different observations.

Our two-stage framework was designed to adapt the diverse progression patterns—while some subgroups traced a simple, linear path, others navigated a more complex, non-linear course. The GP, a core component of our framework, offers a dynamic non-linear mapping between latent factors and observed data and quantifies the associated uncertainty.

In general, our model aimed to deliver a holistic overview of disease evolution patterns, bypassing the need for predefined staging. Such a model should facilitate detailed analyses of subgroup characteristics and capture a spectrum of disease progressions with uncertainty in our understanding of these patterns.

Detailed summarization

We developed a model that applies GP for non-linear mapping from a latent disease stage to observed EHR data. This mapping is typically defined by the mean function of the GP, and the covariance function describes the covariance between any two points within the latent space. Commonly, a mean function of 0 is adopted for simplicity, particularly with insufficient prior knowledge about the function to be studied. The covariance function then defines the relationship between two latent space points shaped by properties like smoothness and periodicity.

Figure 1A displays our model's design. We first preprocessed the data, incorporating feature selection, handling missing values, and applying z-score normalization based on a healthy control mean and variance. This process began with univariate analysis to identify disease-linked features, employing chi-square tests for categorical and ANOVA for numerical features, only considering P-values below .05 significant. K-nearest neighbors (KNN) followed for imputation, and we normalized features against the control group to highlight deviations of a disease state from the "normal" condition.

Figure 1B outlines the model fitting process. We used k-means clustering for subgroup identification and the elbow method to pinpoint the optimal number of subgroups. Then, we applied GPSI for precise stage estimation within each subgroup, catering to their specific progression profiles. The kernel function in GPSI is intended to capture potential smoothness and correlations between different stages by incorporating existing knowledge or beliefs. Figure 1C illustrates the basic procedure of stage inference with the GPSI, as detailed in the following section. Lastly, we used the Shapley Additive Explanations (SHAP) algorithm[19] to obtain explanations of the features driving disease progression.

Gaussian process for stage inference

Considering a dataset with N patients with P variables, the observed data $\boldsymbol{X} = \{x_i\}_{i=1}^{N}$ are conditionally independent given a mean function $\mu(s)$ and an observation covariance matrix Σ. These are predicted on the latent unobserved disease stage $s = \{s_1, \ldots, s_N\}$, $s_i \in \{1, \ldots, \zeta\}$, where ζ stands for the number of stages. ζ may either be informed by prior knowledge of the disease or deduced from empirical data, for instance, by employing a high Kendall rank correlation coefficient in bootstrap analyses as detailed in Section Internal validation. The corresponding results are discussed in the results section for each specific disease. For each j feature, the mean function μ_j is given an independent GP prior with a

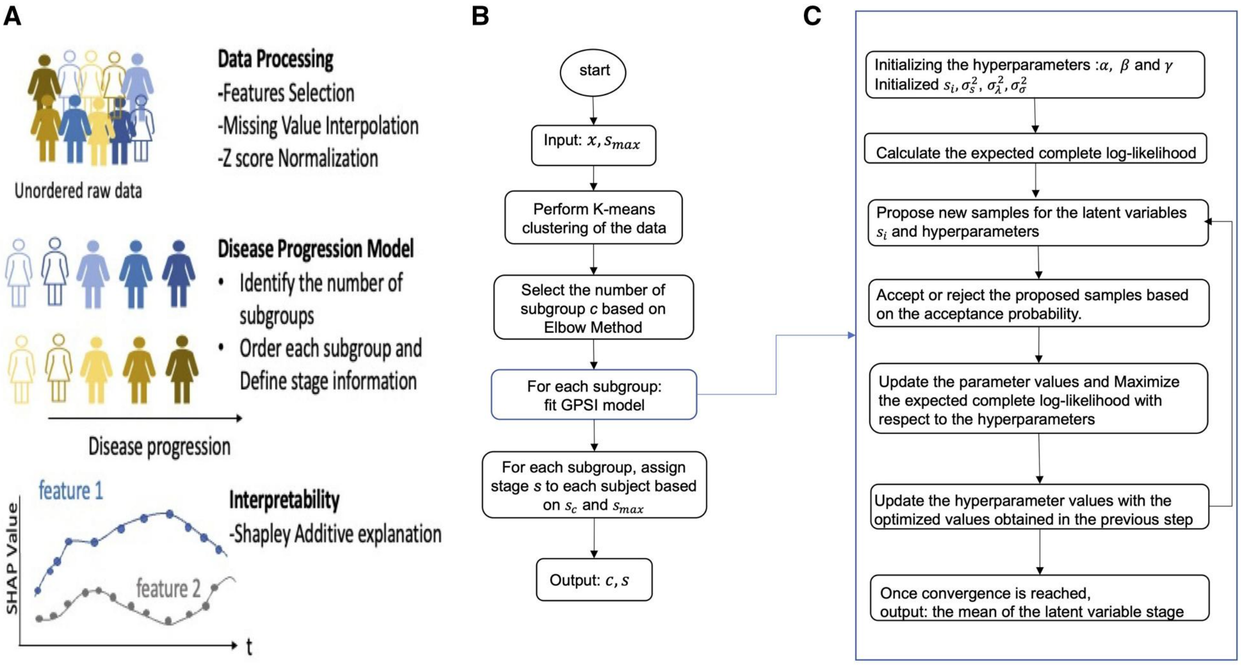

A
Data Processing
-Features Selection
-Missing Value Interpolation
-Z score Normalization
Unordered raw data
Disease Progression Model
• Identify the number of subgroups
• Order each subgroup and Define stage information
Disease progression
Interpretability
-Shapley Additive explanation
feature 1
feature 2
SHAP Value
t
B
start
Input: x, s_{max}
Perform K-means clustering of the data
Select the number of subgroup c based on Elbow Method
For each subgroup: fit GPSI model
For each subgroup, assign stage s to each subject based on s_c and s_{max}
Output: c, s
C
Initializing the hyperparameters :α, β and γ Initialized $s_i, \sigma_s^2, \sigma_\lambda^2, \sigma_\sigma^2$
Calculate the expected complete log-likelihood
Propose new samples for the latent variables s_i and hyperparameters
Accept or reject the proposed samples based on the acceptance probability.
Update the parameter values and Maximize the expected complete log-likelihood with respect to the hyperparameters
Update the hyperparameter values with the optimized values obtained in the previous step
Once convergence is reached, output: the mean of the latent variable stage

Figure 1. Overall schema. (A) The overall schema of the model. The model is divided into data processing, disease progression for each subgroup, and interpretability. (B) The general procedure of model fitting, including subgroup estimation with k-means, determination of the optimal number of subgroups with elbow method, and stage inference with GP in the blue box (C) The detailed procedure for stage estimation using the Gaussian process.

covariance function $K^{(j)}$, where $K^{(j)}(s,\ s')$ denotes the covariance between any two stages $s\ and\ s'$. The model is presented as follows in a hierarchical structure:

$$x_i \sim N(\mu(s_i),\ \Sigma),\ i = 1,\ \ldots,\ N$$

$$\mu_i = GP\left(0,\ K^{(j)}\right),\ j = 1,\ \ldots,\ P$$

$$K^{(j)}\left(s,\ s'\right) = \exp\left(-\lambda_j(s - s')^2\right),\ j = 1,\ \ldots,\ P$$

$$\Sigma = \mathrm{diag}\left(\sigma_1^2,\ \ldots,\ \sigma_P^2\right)$$

$$\lambda_j \ \sim\ \exp(\gamma)$$

$$\sigma_j^2 \ \sim\ invgamma(\alpha, \beta)$$

$$s_i \ \sim\ U(0, 1)$$

where s follows a uniform prior distribution and σ^2 describes the prior variance, which follows an inverse gamma distribution with shape parameter α and scale parameter β. In general, the shape parameter α determines the shape of the distribution, with larger values leading to a more peaked distribution. The scale parameter β controls the spread or scale of the distribution, with larger values resulting in a wider distribution. The parameter λ is the process variance, following an exponential distribution with rate parameter γ and has an intuitive interpretation in the context of curve fitting. The likelihood of x given the latent stage s is conditionally independent across features, so

$$P\left(X|s,\ \sigma^2,\ \lambda\right) = \prod_j^P p(x_j|s, \sigma_j^2, \lambda_j\ )$$

$$p\left(x_j|s, \sigma_j^2, \lambda_j\right) = N(x_j|0,\ K^{(j)}\left(\lambda_j,\ s\right) + \sigma_j^2\ I)$$

Therefore, the objective function for Maximum A Posterior (MAP) estimation is

$$p\left(s, \lambda,\ \sigma^2\right) \ltimes \prod_{j=1}^{P} N(x_j|0,\ K^{(j)}\left(\lambda_j,\ s\right) + \sigma_j^2 I) * \prod_{m=1}^{N} \prod_{n=m+1}^{N} (s_m - s_n)$$

Notably, the kernel function should have a different form for different applications. For example, the kernel function should include a periodic function with sin and cos forms for the features with circadian rhythm.

Inference

We assume the model's unknown parameters to be distributed according to a multivariate Gaussian distribution. To estimate these parameters—which include the stage variables s, rate parameter λ, and variance σ^2—we implemented the Metropolis–Hastings (MH) algorithm using a random-walk strategy, where random perturbations from a normal distribution are used to explore the parameter space. The process begins with the initialization of the model's hyperparameters (α, β, γ) of the model, as well as $s, \lambda,\ \sigma^2$. We then calculate the posterior distribution of these parameters given the observed data and current estimates of hyperparameters.

Following initialization, we compute the complete log-likelihood. Proposed parameter samples are then evaluated: they are accepted if they increase the likelihood, or rejected otherwise, maintaining the current parameter values. If accepted, the parameter set—including latent variables and hyperparameters—is updated to these new proposed values.

The optimization continues with iterations of the MH algorithm to refine the parameter estimation. This iterative process is repeated until convergence is achieved, ensuring the parameter estimates are robust and accurately reflect the underlying data structure.

Internal validation

To ensure the robustness of the inferred stage information, we used 1000 bootstrap resampling across different potential stage numbers. We computed the Kendall rank correlation coefficient (Kendall's Tau)[20] to measure the order consistency between inferred stages from two consecutive samples during each resampling. The goal was to determine the stage count that yielded the highest average Kendall's Tau.

Kendall's Tau was designed to assess the consistency of progression patterns. Kendall's tau values ranged from -1 (100% disagreement in rank) to 1 (100% agreement in rank), while 0 indicated no correlation. We deemed the stage identification robust when the mean Kendall's Tau exceeded 0.5, indicating a strong ordinal association.

Consider two ordering sequences S_A and S_B, for any pair of two observations x_1 and x_2, their positions in the sequences are denoted by x_{1A} and x_{2A} for S_A, and x_{1B} and x_{2B} for S_B, respectively. Moreover, x_1 and x_2 would concordant; if $x_{1A} < x_{1B}$, then $x_{2A} < x_{2B}$. Otherwise, they would be discordant. The Kendall τ is calculated as:

$$\tau = \frac{\text{number of concordant pair} - \text{number of discordant pairs}}{\text{total number of pairs}}$$

Datasets and feature selection

The Osteoarthritis Initiative

The Osteoarthritis Initiative (OAI) is a repository for data on OA, a joint condition known for causing knee pain and stiffness. We leveraged OA imaging and clinical data from the OAI (https://www.niams.nih.gov/grants-funding/funded-research/osteoarthritis-initiative), involving 1356 participants aged between 45 and 79 with recorded imaging scans and demographic information. The distribution of participants is presented in Table S1. The radiographic markers in this study were extracted from MRIs and X-rays, including minimal cartilage thickness, joint space narrowing (JSN), and relative areas of denuded bone.

Adolescent brain cognitive development

Bipolar disorder (BP), also known as manic-depressive illness, manifests through severe mood oscillations, including manic and depressive episodes. Our BP research utilizes data from the Adolescent Brain Cognitive Development Study (ABCD) repository (https://nda.nih.gov/abcd/), selecting a subset of 3211 participants aged 9 to 11, including 1607 females. The participants' distribution is cataloged in Table S1. The study evaluated morphometric data to understand this condition's neurological basis, specifically the cortical volumes across 68 cortical[21] and 40 subcortical regions.[22]

Hepatocellular carcinoma datasets

Hepatocellular carcinoma is the primary form of liver cancer, arising from mutative alterations in liver cells. We sourced HCC-related information from two distinct cohorts: UTHealth and The Cancer Genome Atlas (TCGA). The UTHealth study comprised 144 patients, while TCGA had a sample size of 119. The distribution of participants is presented in Table S1. From MRI scans, we derived 16 quantifiable features for our analysis, including tumor size, margin sharpness, and maximum lesion area, with an extended overview given in Table S2.

Results

Robustness of GPSI in recovering predefined disease progression using synthetic data

In this study, we compared the performance of the GPSI with SuStaIn in recovering a predefined progression using synthetic data. We generated a dataset of 500 samples across five features with 16 predefined ground-truth stages. The general change of features over time is defined in Figure 2A. The details of generating synthetic data are presented in Section S3.

Figure 2B illuminates the comparative performance of GPSI and SuStaIn in recovering predefined feature trajectories. The gray dot represents the results estimated by SuStaIn, the blue dot stands for the ground truth, and the green dot shows the results obtained by GPSI. Between stages 1 and 6, GPSI identified a significant change in feature 2, aligning with the ground truth, contrasted with SuStaIn, which identified feature 4. From stages 6 to 11, both GPSI and SuStaIn concurred in identifying substantive changes in features 1 and 5. Between stages 11 and 16, feature 4 was identified by GPSI, aligning with the ground truth, where SuStaIn identified changes in feature 2. In the end, feature 3 was successfully identified by both methods. Though both methodologies proficiently captured feature shifts during the mid and late stages, GPSI demonstrated additional adeptness at identifying variations during the early stages of progression.

We performed an additional comparison using two simulated datasets with different signal-to-noise ratios (SNRs). Figure 2C shows the inferred stage distribution of each model using data with SNR = 5 and Figure 2D shows results for SNR = 1. In summary, GPSI demonstrated robustness and higher accuracy than SuStaIn in high-noise data.

GPSI application for identifying OA progression patterns

Utilizing the GPSI on data from the OAI, we divided OA progression into 15 distinct stages, revealing two separate subgroups, as depicted in Figure 3A and B. Figure 3C delineates the unique progression patterns of these subgroups, discerned without prior genotype information. Subgroup 1 patients first suffered comorbidity impacts before displaying reduced cartilage thickness and joint space, concluding with substantial subchondral bone area reduction and acute pain. Contrastingly, subgroup 2 began with JSN, followed by a decrease in subchondral bone area and then cartilage thickness, with comorbidities manifesting subsequently.

The reliability of the assigned stages and subgroups was assessed to probe the GPSI's stratification capability. We began our analysis by comparing the distribution of stages inferred through GPSI between the healthy group and the OA group (those with a KL score of ≥2). We observed a distinct separation between the healthy and the disease groups, as illustrated in Figure 3D. In Figure S11, the silhouette plots of OA are presented, revealing the highest silhouette score of 0.55 when $n = 2$. Subsequent analysis of intra-cluster similarity within each stage yielded an average similarity of 69%, as detailed in Figure S7A and B. The heatmap of inter-cluster dissimilarity for each pair of stages is presented in Figure S8A and B, illustrating the distinctions and separations between different stage pairs. The correlation between the inferred

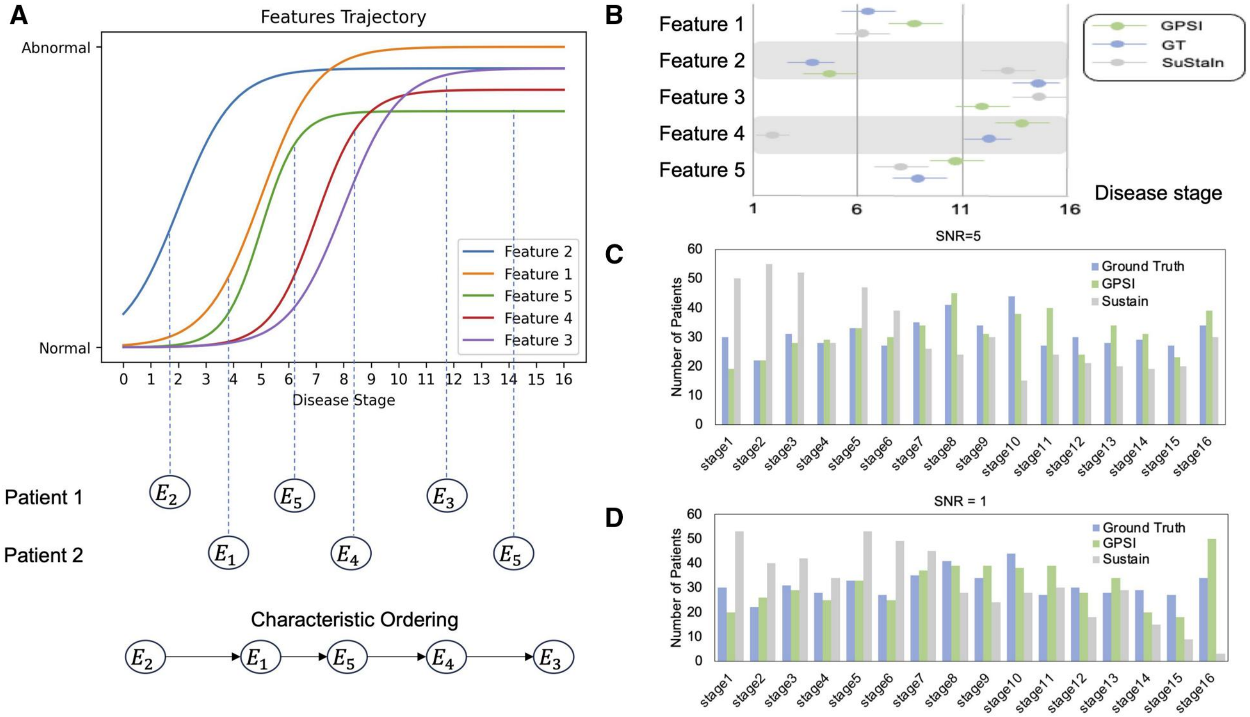


Figure 2. Simulated dataset. (A) The predefined disease progression within five features. (B) The major biomarker changes in each stage. (C) The inferred stage distribution of model (Gaussian Process for Stage Inference [GPSI], SuStaIn and ground truth) using data with signal-to-noise ratio (SNR)=5. (D) The inferred stage distribution of model (GPSI, SuStaIn and ground truth) using data with SNR=1.

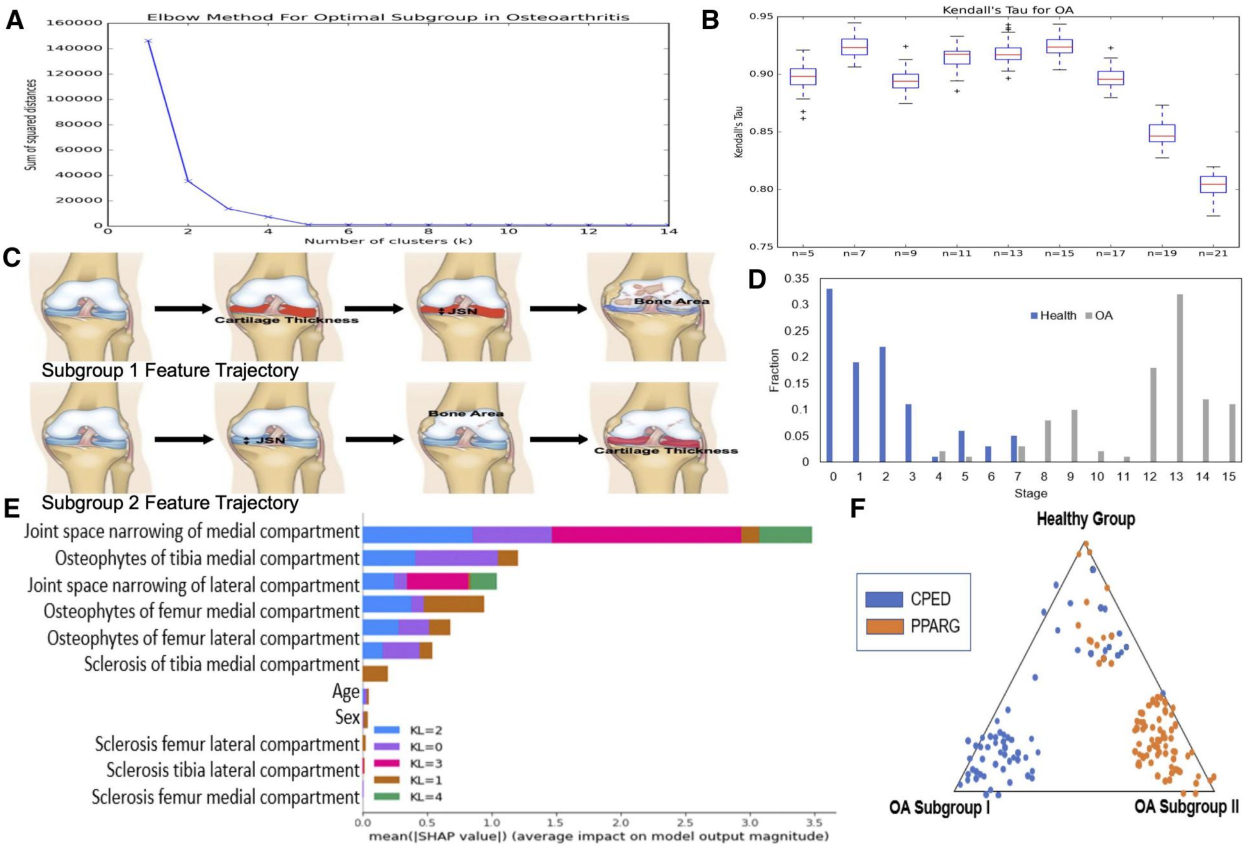
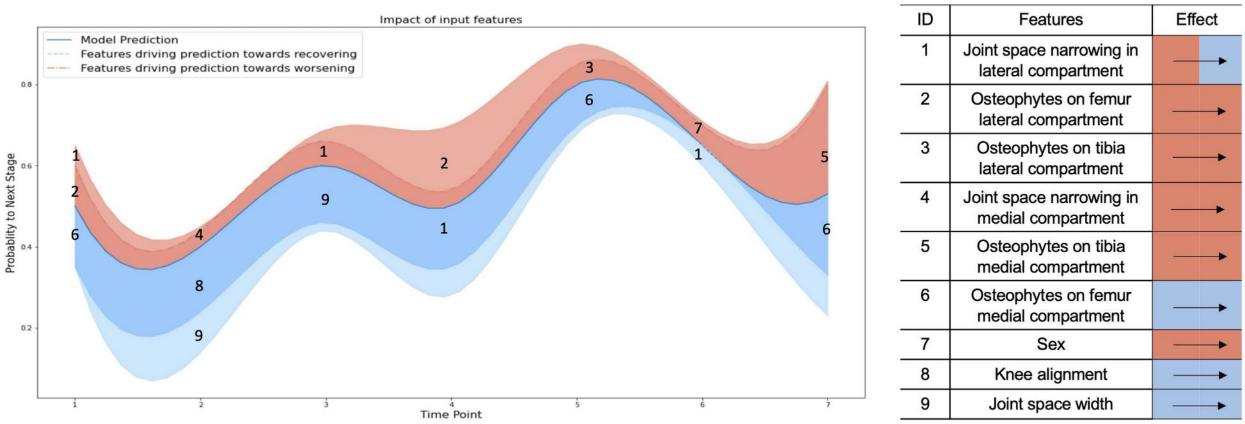


Figure 3. Osteoarthritis (OA) dataset. (A) Elbow methods for subgroup estimation. (B) Boxplot of the Kendall rank correlation of the increasing number of maximized stages. (C) Two major OA development patterns identified by Gaussian Process for Stage Inference (GPSI). (D) The distribution of healthy patients and OA patients. (E) The importance (SHAP value) of each imaging biomarker in predicting the risk of progression to the severer stage. (F) Scatter plot of three major diagnostic groups in two genotypes.



ID	Features	Effect
1	Joint space narrowing in lateral compartment	🟥🟦
2	Osteophytes on femur lateral compartment	🟥 →
3	Osteophytes on tibia lateral compartment	🟥 →
4	Joint space narrowing in medial compartment	🟥 →
5	Osteophytes on tibia medial compartment	🟥 →
6	Osteophytes on femur medial compartment	🟦 →
7	Sex	🟥 →
8	Knee alignment	🟦 →
9	Joint space width	🟦 →

Figure 4. Dynamic feature importance. Dynamic importance of each imaging biomarker. Shaded areas show the three most important input features towards either non-survival (pink) or survival (blue) during the seven-time points. High opacity reflects high relative feature importance. Numbers are used to identify the features; labels are added whenever a feature is outranked by another.

stage and the KL was quantified using Kendall's Tau, yielding an average value of 0.72 (Figure S9A). Following this, the 15 stages were mapped into four-class KL grades, with the heatmap of the OA confusion matrix being presented and demonstrating an accuracy of 0.78 (Figure S9B).

Since the KL system is insensitive to changes in the disease progression,[10] we also compared the inferred stages with the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain score and disability score[23] in Figure S9C and D. The WOMAC pain score is the subscale for evaluating the level of pain the individual experiences during various activities like walking, using stairs, lying in bed, sitting or lying, and standing upright. The WOMAC disability score is the physical function subscale for assessing an individual's difficulty when performing daily activities, including ascending or descending stairs, rising from sitting, standing, bending, and walking. Although the pain score and disability score are obtained from questionnaires, which inherently have a subjective component, we still can see that patients might experience an initial increase in pain or disability with stage progression.

Figure 3E illustrates the contribution of each image feature to different KL grades calculated by SHAP. For example, JSN is crucial in the middle stage, while osteophytes in the tibia medial compartment are essential in the early stage. Figure 4 shows the dynamic contribution of features over time based on the available seven years of follow-up data. Decreasing variable values in the salmon ribbon represent the patient's knee gradually deteriorating, while the variables in the blue ribbon indicate a recovering status of the knee. In this case, we found that narrowing joint space in the medial compartment led to worsening knee status, with knee alignment as the most crucial feature to indicate recovery. Some features, like the lateral compartment's JSN, predominated the prediction of deterioration early and recovery later.

We further checked the genotypes of identified subgroups (Figure 3F) using genome-wide association study (GWAS) with PLINK.[24] Peroxisome proliferator-activated receptor gamma (PPARG) genotypes were the main contributors to subgroup 2, whereas cadherin-like and PC-esterase domain containing 1 (CPED1) genotypes were the main contributors to subgroup 1. Early studies have shown that mutations in the PPARG gene can increase the risk of developing obesity.[25] In addition, overexpression of PPARG in mice leads to increased adiposity and insulin resistance.[26] Several studies have suggested that CPED1 can regulate bone density through different mechanisms depending on age and sex.[27–29]

GPSI application for identifying BP progression patterns

The GPSI segmented BP progression into 15 developmental stages, categorized into two distinct subgroups (Figure 5A and B). Figure 5C shows the disease development sequence of type I (BP-I) and type II (BP-II). In the case of BP-I, significant atrophy first occurs in the central sulcus. Other parietal and frontal areas become involved subsequently, while the cuneus is only affected in the late stages. Contrastingly, in BP-II, significant atrophy first occurs in the cuneus, followed by the subsequent involvement of the cingulate and central sulcus.

Figure 5D and F illustrates that the distribution of stages inferred by GPSI exhibits variations between different diagnostic groups, implying distinct patterns among the healthy, BP-I, and BP-II groups. In Figure S12, the silhouette plots of BP are presented, revealing the highest silhouette score of 0.43 when $n = 2$. Subsequent analysis of intra-cluster similarity within each stage yielded an average similarity of 90%, as detailed in Figure S7C and D. The heatmap of inter-cluster dissimilarity for each pair of stages is presented in Figure S8C and D, illustrating the distinctions and separations between different stage pairs.

Figure 5E presents how each image feature contributes to a different BP type calculated by SHAP. Bipolar II disorder has two significant features: mean diffusivity (MD) within the parcellation of sub-adjacent white matter and MD within the parcellation of cortical gray matter. In contrast, age and average longitudinal diffusivity (LD) within the parcellation of sub-adjacent white matter are critical factors in BP-I. Mean diffusivity reflects the average magnitude of water diffusion within a given voxel or region of interest, while LD demonstrates the degree of water diffusion in the axial direction.

GPSI-predicted HCC progression pattern reproduction in two independent clinical datasets

Hepatocellular carcinoma development had nine stages without distinct subgroups using UTHealth data (Figure 6A and

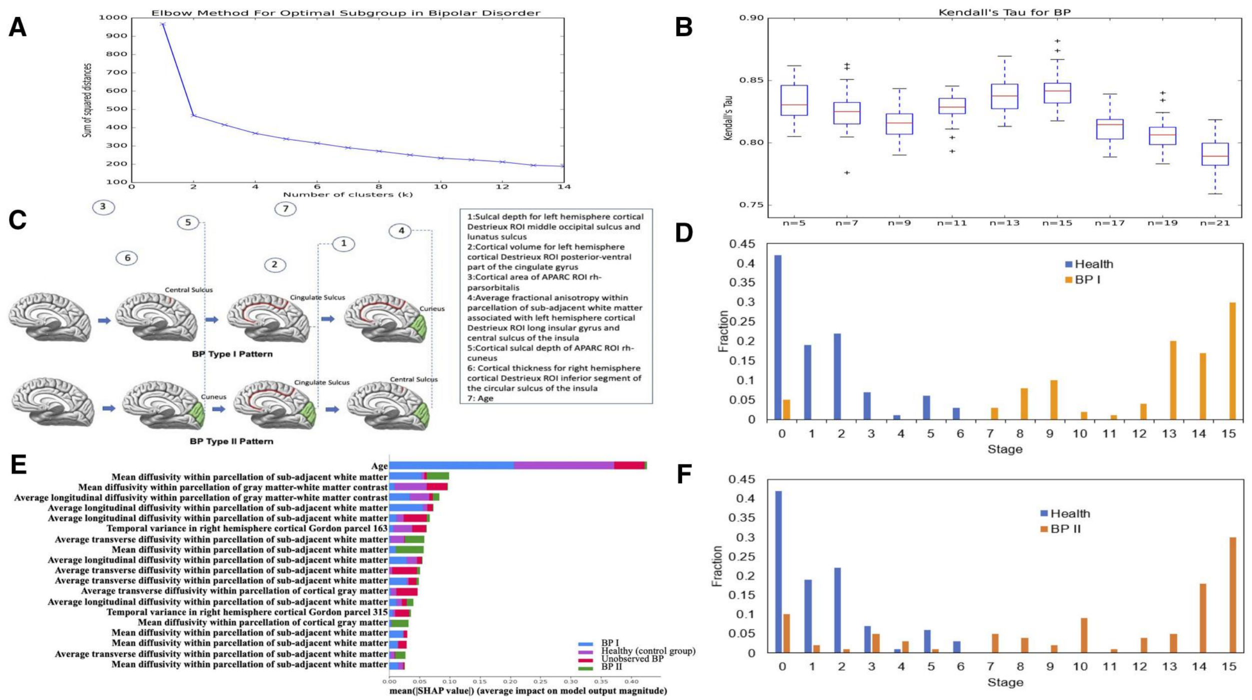


Figure 5. Bipolar disorder (BP) dataset. (A) Elbow methods for subgroup estimation. (B) Boxplot of the Kendall rank correlation of the increasing number of maximized stages. (C) Two major BP development patterns identified by Gaussian Process for Stage Inference (GPSI). (E) The relative importance (Shapley Additive Explanations [SHAP] value) of each imaging biomarker in predicting the risk of progression to the severer stage. (D) The distribution of control (healthy) group and BP type I patients. (F) The distribution of control (healthy) group and BP type II patients.

C). The results were similar in the TCGA dataset (Figure 6B and D). Figure 6E displays the HCC progression patterns identified by the GPSI. The trajectory was identical in the TCGA dataset, so it is not shown.

We initiated our analysis by comparing the distribution of GPSI inferred stages across each LR grade, observing a consistent increase in the inferred stage as the LR grade increases (Figure 6F and G). Following this, Kendall's Tau was utilized to delve deeper into the UTHealth and TCGA datasets. The boxplot depicted in Figure S10A illustrates the Kendall rank correlation between inferred stages and LR grade, revealing averages of 0.95 and 0.94 for the UTHealth and TCGA datasets, respectively.

We then mapped the nine stages into five-class LR grades, and the heatmap of the HCC confusion matrix for UTHealth and TCGA achieved average accuracies of 0.67 and 0.61, respectively (Figure S10B and C). The mapping was as follows: stage 1 to LR1, stages 2 and 3 to LR2, stages 4–6 to LR3, stages 7 and 8 to LR4, and stage 9 to LR5. The challenge in distinguishing LR3 and LR4 is the primary reason causing the low accuracy. The LR3 and LR4 patients are in the middle of developing HCC, where their lesions are nonspecific. Even experienced radiologists sometimes struggle when classifying these two, and the actual probabilities of HCC associated with LR3 and LR4 remain obscure.[12]

A more granular analysis of the staging includes an exploration of intra-cluster similarity within each stage, revealing an average similarity above 84% (Figure S7E and F) and a heatmap of inter-cluster dissimilarity among all stage pairs, showcased in Figure S8E and F.

Figure 6H shows how each image feature contributed to different LR grades in the UTHealth dataset. For example, the tumor's hypodense halo and area ratio over the liver were crucial in the LR4 stage, while the number of lesions was essential in the early stage. Figure 6I presents how each image feature contributed to different LR grades in the TCGA. The hypodense halo still played a crucial role in the LR4 stage, while the number of lesions was a critical factor in the early stage.

Discussion

We developed the GPSI approach to uncover chronic disease progression patterns and assess features' dynamic contributions to the diseases. Unlike methods dependent on clinical staging, GPSI extracts progression timelines directly from cross-sectional data, negating the need for longitudinal datasets. Gaussian Process for Stage Inference's innovation lies in its allowance for non-linear feature trajectories and its ability to integrate additional covariates, unveiling pertinent trends in feature progression.

Moreover, GPSI distinctively accounts for the uncertainties in event sequencing attributable to measurement noise and population diversity. In contrast to machine learning approaches that provide only point estimates, GPs offer a predictive mean and variance, quantifying uncertainty in data point positioning. This uncertainty quantification is vital for informed clinical decision-making, particularly in understanding the confidence level in predicted outcomes.

When applying GPSI to the OAI dataset, we identified two distinct subgroups where further GWAS analysis indicated that these two subgroups corresponded to two gene groups: PPAR and CPED1. Subgroup 1 predominantly shared a CPED1 genotype, suggesting a connection to bone-

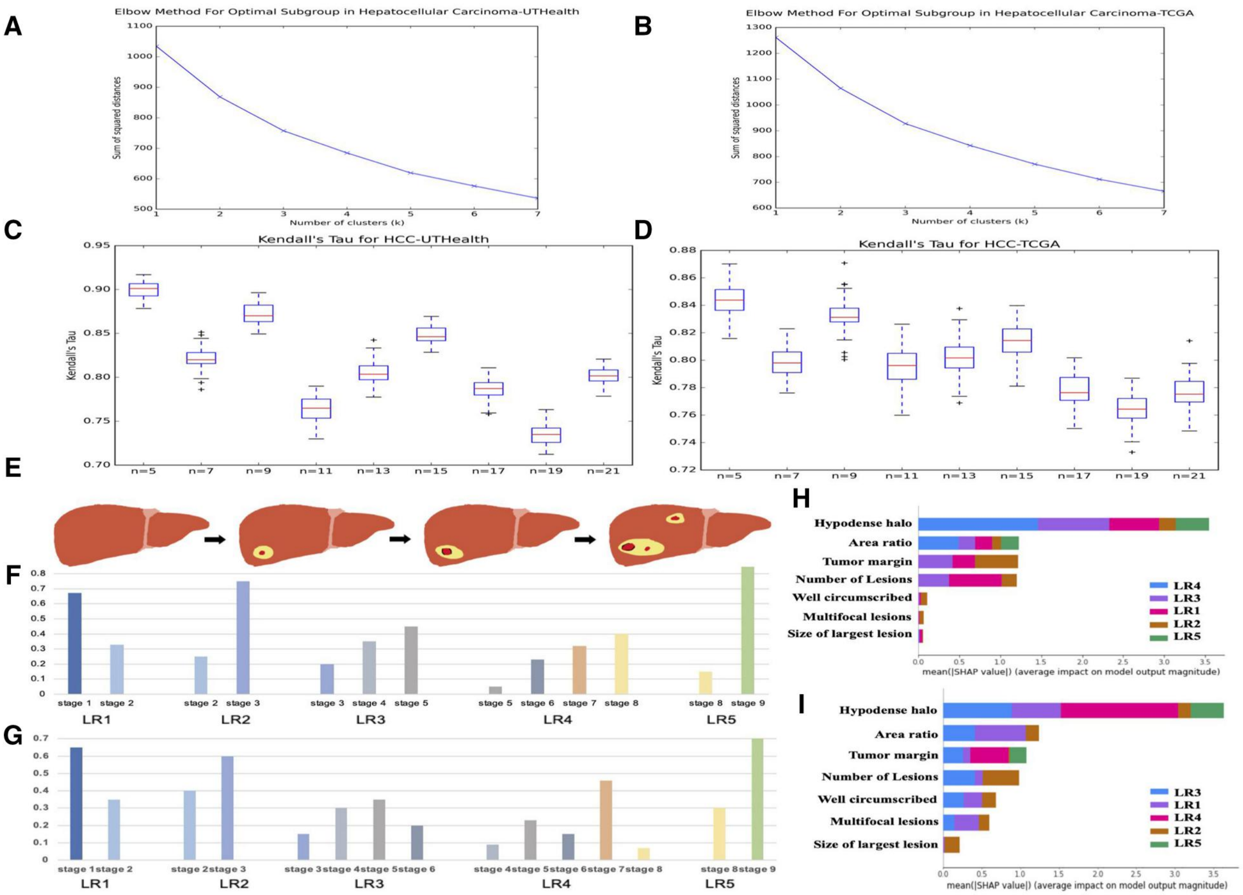


Figure 6. Hepatocellular carcinoma (HCC) dataset. (A) Elbow methods for subgroup estimation using UTHealth. (B) Elbow methods for subgroup estimation using TCGA. (C) Boxplot of the Kendall rank correlation of the increasing number of maximized stages using UTHealth. (D) Boxplot of the Kendall rank correlation of the increasing number of maximized stages using TCGA. (E) HCC development patterns identified by Gaussian Process for Stage Inference (GPSI) using UTHealth and TCGA dataset. (F) The comparison between GPSI staging vs. Liver Imaging Reporting and Data System (LI-RADS) grading system in UTHealth. (G) The comparison between GPSI staging vs LI-RADS grading system in TCGA. (H) The relative importance (Shapley Additive Explanations [SHAP] value) of each imaging biomarker in predicting the LI-RADS in UTHealth dataset. (I) The relative importance (SHAP value) of each imaging biomarker in predicting the LI-RADS in TCGA dataset.

remodeling pathways. In contrast, subgroup 2 shared a PPAR genotype, indicating an association with obesity-related pathways. The regulatory roles of PPARγ in fat cell differentiation and inflammation suggest its potential impact on metabolic and inflammatory aspects of OA.[30,31] Conversely, CPED1's involvement in matrix remodeling may affect cartilage integrity in the disease.[28,27,32,33]

Our analysis with GPSI also demonstrated a clear distinction between healthy individuals and those with bipolar I and II disorders, correlating specific white and gray matter changes with each condition. This underscores GPSI's ability to refine stage information, providing a sharper diagnostic classification across disease stages.

The refined staging information of the GPSI provided a more accurate diagnosis classification, especially in the early and late stages. In the OA case, the KL grade failed to capture the nuanced progression of the disease.[10] For example, we found that comorbid conditions differentiated patients with higher pain scores but the same KL grade. This indicates the role comorbidities play in pain exacerbation and the disease's interaction with other health issues.

Similarly, in HCC analysis, GPSI helped to discern patterns in lesions that could indicate disease stages, aiding in the classification of ambiguous cases like LR3 and LR4 lesions. We observed that tumor margin irregularities and size provided significant clues for staging.

However, GPSI's current form does not support out-of-sample predictions, presenting challenges when integrating new patient data with an existing model. Future work will explore back constraints to enable new data projection into a pre-established latent stage space without retraining the model. Such enhancements will aid in the practical application of GPSI, allowing for continued use as new data emerges. Another concern is that although GPSI offers a robust way to capture intricate patterns in data, it may not always be necessary for linear progressions. It is imperative to conduct exploratory data analysis (EDA) to discern the underlying data relationship prior to application. Gaussian Process for Stage Inference's sensitivity to initial settings also calls for multiple initializations to ensure result stability, which we plan to address in further developments.

Conclusion

In sum, we developed an unsupervised approach for studying and characterizing progression patterns for three chronic diseases in unique detail. Based on the differences in these features across stages, physicians can better tailor treatment plans and medications to individual patients. In short, refining disease staging can help advance medical research by allowing for more precise and targeted studies, leading to a better understanding of disease mechanisms.

Author contributions

YW contributed to the study design, developed the model, and wrote the manuscript. WZ and RA provided feedback on the manuscript. YL provided image processing guidance, and HW provided clinical support. XZ provided supervision and contributed to the editing of the manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

Research reported in this publication was supported by a CTSA pilot grant (Wang), National Institutes of Health (NIH) under Grants [R01DE027027, R01GM123037, U01AR069395, and R01CA241930]; National Science Foundation (NSF) under grants [2217515 and 2326879], and Dr Carl V. Vartian Chair Professorship Funds to Dr Zhou from the University of Texas Health Science Center at Houston.

Conflicts of interest

None declared.

Data availability

The OAI, ABCD, and TCGA datasets are publicly available. The UTHealth dataset is available upon reasonable request from the authors.

Code availability

All models were created with Python and common packages such as GPy. The codes for the GPSI are available upon request from the authors.

References

1. Young AL, Marinescu RV, Oxtoby NP, et al.; Alzheimer's Disease Neuroimaging Initiative (ADNI). Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat Commun*. 2018;9(1):4273.
2. Michalis T, Neil DL. *Bayesian Gaussian Process Latent Variable Model*. PMLR. 844-851.
3. Ahuja Y, Wen J, Hong C, Xia Z, Huang S, Cai T. A semi-supervised adaptive Markov Gaussian embedding process (SAMGEP) for prediction of phenotype event times using the electronic health record. *Sci Rep*. 2022;12(1):17737.
4. Colopy GW, Pimentel MAF, Roberts SJ, et al. Bayesian Gaussian processes for identifying the deteriorating patient. In: *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. August 2016:5311–5314.
5. Fradi A, Feunteun Y, Samir C, et al. Bayesian regression and classification using Gaussian process priors indexed by probability density functions. *Inform Sci*. 2021;548:56-68.
6. Cheng L-F, Dumitrascu B, Darnell G, et al. Sparse multi-output Gaussian processes for online medical time series prediction. *BMC Med Inform Decis Mak*. 2020;20(1):152.
7. Meng R, Soper B, Lee HKH, et al. Nonstationary multivariate Gaussian processes for electronic health records. *J Biomed Inform*. 2021;117:103698.
8. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. *Ann Rheum Dis*. 1957;16(4):494-502.
9. Spector TD, Cooper C. Radiographic assessment of osteoarthritis in population studies: whither Kellgren and Lawrence? *Osteoarthritis Cartilage*. 1993;1(4):203-206.
10. Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren–Lawrence classification of osteoarthritis. *Clin Orthop Relat Res*. 2016;474(8):1886-1893.
11. Kielar AZ, Chernyak V, Bashir MR, et al. LI-RADS 2017: an update. *J Magn Reson Imaging*. 2018;47(6):1459-1474.
12. Kim Y-Y, Choi J-Y, Sirlin CB, et al. Pitfalls and problems to be solved in the diagnostic CT/MRI Liver Imaging Reporting and Data System (LI-RADS). *Eur Radiol*. 2019;29(3):1124-1132.
13. Arif-Tiwari H, et al. MRI of hepatocellular carcinoma: an update of current practices. *Diagn Interv Radiol*. 2014;20(3):209-221.
14. Conklin JE, Lieberman JV, Barnes CA, et al. Disease staging: implications for hospital reimbursement and management. *Health Care Financ Rev*. 1984;1984(Suppl):13-22.
15. Vidula N, Peppercorn J. Clicking away to capture cancer staging—the benefits and challenges of completing standardized staging modules. *JCO Oncol Pract*. 2023;19(10):835-838.
16. Andreev VP, Helmuth ME, Liu G, et al. Subtyping of common complex diseases and disorders by integrating heterogeneous data. Identifying clusters among women with lower urinary tract symptoms in the LURN study. *PloS one*, 2022;17(6):e0268547.
17. Gregory A, Xu Z, Pratte K, Lee S, et al. Clustering-based COPD subtypes have distinct longitudinal outcomes and multi-omics biomarkers. *BMJ Open Respir Res*. 2022;9(1):e001182.
18. Zubair M, Iqbal MA, Shil A, et al. An improved K-means clustering algorithm towards an efficient data-driven modeling. *Ann Data Sci*. 2022:1-20. https://doi.org/10.1007/s40745-022-00428-2
19. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint*, 2018. arXiv:1802.03888
20. Kendall MG. A new measure of rank correlation. *Biometrika*. 1938;30(1-2):81-93.
21. Desikan RS, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*. 2006;31(3):968-980.
22. Fischl B, Salat DH, Busa E, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*. 2002;33(3):341-355.
23. Frederick W, Sheldon XK. Rasch analysis of the Western Ontario MacMaster Questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. *Ann Rheum Dis*. 1999;58(9):563.
24. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575.
25. Celi FS, Shuldiner AR. The role of peroxisome proliferator-activated receptor gamma in diabetes and obesity. *Curr Diab Rep*. 2002;2(2):179-185.
26. Norris AW, Chen L, Fisher SJ, et al. Muscle-specific PPARgamma-deficient mice develop increased adiposity and insulin resistance but respond to thiazolidinediones. *J Clin Invest*. 2003;112(4):608-618.
27. Gómez AE, Addish S, Alvarado K, et al. Multiple mechanisms explain genetic effects at the CPED1-WNT16 bone mineral density locus. *Curr Osteoporos Rep*. 2023;21(2):173-183.

28. Kemp JP, Medina-Gomez C, Estrada K, et al. Phenotypic dissection of bone mineral density reveals skeletal site specificity and facilitates the identification of novel loci in the genetic regulation of bone mass attainment. *PLoS Genet*. 2014;10(6):e1004423.
29. Chesi A, Mitchell JA, Kalkwarf HJ, et al. A trans-ethnic genomewide association study identifies gender-specific loci influencing pediatric aBMD and BMC at the distal radius. *Hum Mol Genet*. 2015;24(17):5053-5059.
30. Stienstra R, Duval C, Müller M, et al. PPARs, obesity, and inflammation. *PPAR Res*. 2007;2007:95974.
31. Wang S, Lin Y, Gao L, et al. PPAR-γ integrates obesity and adipocyte clock through epigenetic regulation of Bmal1. *Theranostics*. 2022;12(4):1589-1606.
32. Chesi A, Mitchell JA, Kalkwarf HJ, et al. A genomewide association study identifies two sex-specific loci, at SPTB and IZUMO3, influencing pediatric bone mineral density at multiple skeletal sites. *J Bone Miner Res*. 2017;32(6):1274-1281.
33. Mitchell JA, Chesi A, Cousminer DL, et al. Multidimensional bone density phenotyping reveals new insights into genetic regulation of the pediatric skeleton. *J Bone Miner Res*. 2018;33(5):812-821.