Improving Forecasts of Suicide Attempts for Patients with Little Data

Genesis Hang
Wellesley College
gh104@wellesley.edu

Hope Neveux
Harvard University
hopeneveux@fas.harvard.edu

Matthew K. Nock
Harvard University
nock@wjh.harvard.edu

Yaniv Yacoby
Wellesley College
yy109@wellesley.edu

Abstract

Ecological Momentary Assessment provides real-time data on suicidal thoughts and behaviors, but predicting suicide attempts remains challenging due to their rarity and patient heterogeneity. We show that single models fit to all patients perform poorly, while individualized models improve performance but still overfit to patients with limited data. To address this, we introduce Latent Similarity Gaussian Processes (LSGPs) to capture patient heterogeneity, enabling those with little data to leverage similar patients' trends. Preliminary results show promise: even without kernel-design, we outperform all but one baseline while offering a new understanding of patient similarity.

1 Introduction and Related Work

Ecological Momentary Assessment (EMA) studies leverage smartphones to capture insights into suicidal thoughts and behaviors (STBs) as they unfold in daily life [1]. In these intensive longitudinal studies, patients are surveyed multiple times daily on their suicidal urges, intent, and affects. This presents opportunities for machine learning (ML) to forecast imminent suicide risk in time for intervention; however, to date, no current approach can do this reliably [2].

Prior work primarily focuses on forecasting suicidal ideation from EMA data (e.g. [3–6]). While forecasting ideation is itself challenging, suicide attempts are even harder to predict due to their low base-rate [7]; even in the largest available datasets (e.g. 600 patients), attempts are rarely captured (e.g. [8]). This severely limits the data available for model training and evaluation. Exacerbating this challenge, recent work shows that patients' paths to suicide ideation are heterogeneous, suggesting that, at the very least, there are many subtypes of at-risk patients, advocating against the use of single models across all patients [9–11], further reducing the number of data points per model.

Here, we show that the same patient heterogeneity found in the prediction of suicidal ideation is found in the prediction of suicidal attempts. We then present a single model to improve forecasts for patients with little data by capturing patient heterogeneity. Our contributions are:

(A) We show that a single model trained on data to predict suicide attempts from all patients performs worse than individualized, per-patient models. Specifically, we show that each patient exhibits a different forecasting trend, that, when combined, conflict with one another, resulting in poor forecasting performance. This underscores the importance of explicitly modeling patient heterogeneity [11, 10]. From these results, we may be tempted to use a different model per patient—
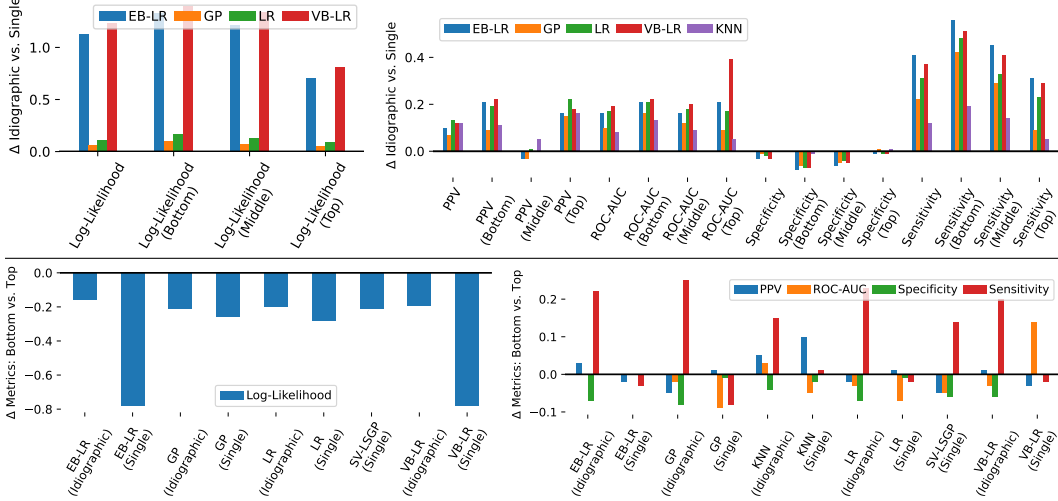
Figure 1: **Top Row: Idiographic models outperform their Single counterparts**—except for specificity, which stayed constant, the magnitude of difference in metrics between the idiographic and single models is almost always positive. **Bottom Row: The 30% patients with fewest data points consistently receive worse forecasts across most metrics and most models than the 30% of patients with most data**—except for sensitivity, the magnitude of difference in metrics computed is usually negative, indicating lower performance.

but per-patient models are prone to severe overfitting for patients with little data. Moreover, they require us to collect enough data per-patients to make informed forecasts.

(B) We naturally formalize our observations into a single model to capture patient heterogeneity, grounded in modeling assumptions supported by our analysis and prior work. Our Latent Similarity Gaussian Process (LSGP) posits that patients lie in a latent space in which distance corresponds to similarity in forecasting trends. By inferring patients' locations in this latent space, forecasts for patients with little data intelligently draw on trends from similar patients. While inspired by prior methods (see Section 3), LSGPs have never been previously applied in this context.

(C) Our preliminary results show promise in improving forecasts of suicide attempts from EMA data for patients with little data, and reveal new avenues for understanding patient similarity. Without significant kernel design or hyperparameter search, our approach already matches the best-performing baseline nearly all metrics, showing promise. Furthermore, we introduce a graph-based visualization of patient similarity within the learned latent space, offering insights into individualized risk profiles and potential shared mechanisms.

2 The Geometry of Forecasts for At-Risk Patients

Notation. Let N denote the number of patients in the data. Let $x_i = [n_i \quad r_{i,1} \quad \ldots \quad r_{i,D_x}]^\mathsf{T}$ represent the ith observation in the data, belonging to patient n_i at time t_i, consisting of their responses $r_{i,d} \in \{0, \ldots, 10\}$ to 10-point likert-scale EMA questions. Here, we will use questions about patients' affects, suicidal intent/urge and behaviors—for details on the study and data, see Appendix A. Using patient responses to these questions, our task is to predict $y_i \in \{0, 1\}$—whether patient n_i engaged in any suicide related event (SRE) sometime in the week following t_i. We define an SRE as either a self-injurious behavior with some (non-zero) intention of dying, or a presentation to a hospital with suicidal thoughts to prevent a suicide attempt. Let $\mathcal{D} = X, Y$ represent the entire training data. Let $\mathcal{D}_n = X_n, Y_n$ represent patient n's training data, where $X_n = \{x_i | n_i = n\}$ and $Y_n = \{y_i | n_i = n\}$. Note that every patient has a different amount of data.

Goal. Given $\mathcal{D}$, our goal is to predict whether, given a *new* EMA response, x_n^*, patient n will engage in an SRE sometime in the next week, y_n^*.

Single vs. Idiographic Models. To better understand the geometry of patient classification boundaries, we compare models trained on *all* patient data $(y_n^* | x_n^*, \mathcal{D})$ with a model consisting of a collection of
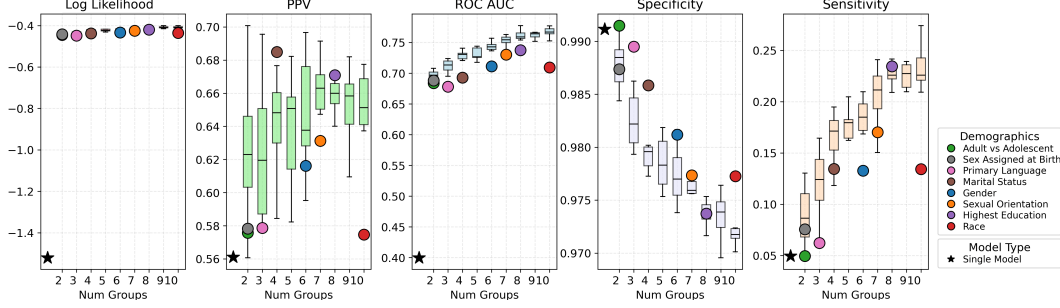
Figure 2: **Patient heterogeneity is so high, even random groupings of patients significantly boosts performance.** We split the data set into G groups, randomly assigning patients, fitting a separate model to each group, and measuring the performance (y-axis) as G increases (x-axis). We repeated this experiment 10 times per G, plotting the distribution of metrics. Finally, we also compare models fit on randomly- vs. demographics-grouped patients (with the same number of G).

models—*one per patient* $(y_n^*|x_n^*, \mathcal{D}_n)$. We refer to the former and latter as a single and an idiographic model, respectively. If idiographic models consistently outperform the single models, this suggests that patients have differing forecasting trends. Even before comparing their performance, we note that idiographic models have one major shortcoming: they cannot be used to make predictions for a new patient n^*; we address this limitation in our method (Section 3).

Baselines and Metrics. We compare our method with several baselines, each used both as a single and idiographic model: Gaussian Process Classification (GP) with a Laplace Approximation, k-Nearest Neighbor Classifier (KNN), Logistic Regression (LR), and Bayesian LR with an empirical Bayes Type II and variational (EB-LR and VB-LR) approximations. For evaluation, we use: Avg. Log-Likelihood, Positive Predictive Value (PPV), Area Under the Receiver Operating Characteristic Curve (ROC-AUC), Specificity, and Sensitivity.

Finding 1: Patients exhibit conflicting classification boundaries. Fig. 1 (top-row) shows that, across all baselines and metrics (except specificity), idiographic models outperform their single counterpart. As such, we may be tempted to just use separate models; however, as Fig. 1 (bottom-row) shows, both single and idiographic models make worse forecasts for the patients with less data. Single model predictions are most influenced by patients with more data, generalizing poorly to patients with less data; idiographic model overfit to patients with little data.

Finding 2: Patient heterogeneity may not well accounted for by discrete groups. Given the heterogeneity of forecasting trends, we may want to find *discrete groupings* of patients with similar trends. Methods like Subgroup Group Iterative Multiple Model Estimation (S-GIMME) [12] do exactly this and have been applied to similar data (e.g. [13, 14, 11]). However, as we show here, our patient trends are so heterogeneous, *even randomly grouping patients significantly boosts performance*, increasing the risk of mis-grouping patients and misleading scientific conclusions. To show this, we conduct a simple experiment: we split the data set into G groups by *randomly* assigning each patient's data to a group, fitting a separate model to each group, and measuring the performance of the models as G increases. As Fig. 2 shows, with increasing G, all metrics improve (except for specificity, which is accounted for by increased sensitivity). Even more interestingly, models trained on patients grouped by demographics tend to perform worse than those on randomly-grouped patients on nearly all metrics. This leads us to question whether our demographic data explain patient heterogeneity. We next propose a model that captures a *continuous* notion of patient similarity.

3   Method

There are many pathways to suicide; even among mental health disorders, conditions such as major depression, generalized anxiety, post-traumatic stress, and borderline personality disorder each may present distinct mechanisms leading to elevated suicide risk [15, 16]. Moreover, within a single diagnosis, each patient's unique life circumstances—e.g. shaped by social determinants and individual differences—may further contribute to patient heterogeneity [10]. To capture this probabilistically, we must account for each patient's individual forecasting trajectory while enabling those with limited

data to leverage data from others without imposing a one-size-fits-all solution. We address this by embedding patients in a latent space, where proximity reflects similarity in risk trajectories; forecasts for patients with little data can thus intelligently "borrow trends" from their neighbors.

Latent Similarity Gaussian Processes (LSGPs). We naturally arrive at the model,

$$\text{Ⓐ } z_n \sim p(z) = \mathcal{N}(0, \mathbb{I}_{D_z}), \qquad \text{Ⓑ } \hat{x}_i | x_i, Z = [r_{i,1} \ \dots \ r_{i,D_x} \ z_{n_i,1} \ \dots \ z_{n_i,D_z}]^\intercal, \quad (1)$$

$$\text{Ⓒ } F | \hat{X}; \theta \sim \mathcal{N}(0, K_\theta(\hat{X}, \hat{X})), \qquad \text{Ⓓ } y_i | f_i \sim \text{Bernoulli}(\text{sigmoid}(f_i)), \qquad (2)$$

where $\hat{x}_i$ represents the concatenation of the inputs x_i with the latent variable z_{n_i} corresponding to patient n_i, $\hat{X}$ is a matrix consisting of all x_i's as rows, $K_\theta(\cdot, \cdot)$ is the kernel matrix computed on rows of its arguments with hyperparameters θ, F is a concatenation of all function values f_i corresponding to each x_i, and $\mathbb{I}_{D_z}$ is an identity matrix of width D_z.

Related Models. Our model bears similarity to several existing models, including (i) GP with Latent Covariate [17] or Covariate GP Latent Variable Models [18], but adapted to have *multiple* observations per latent variable, (ii) Multi-Group GPs [19], but in which the "group" is both *continuous and latent*, or (iii) Meta-Learning GPs [20], but without the *control signal*. None of these models have been previously applied in this context.

Sparse Variational LSGPs. Analytical inference is impossible due to the non-Gaussianity of the likelihood and the large number of observations (14763 from $N = 77$ patients), so we apply the sparse variational formulation of GPs [21] to our model, replacing Ⓒ above with:

$$\text{Ⓒ.1 } U; W, \theta \sim \mathcal{N}(0, K_\theta(W, W)), \qquad (3)$$

$$\text{Ⓒ.2 } F | U, X, Z; W, \theta \sim \mathcal{N}(\Psi \cdot U, K_\theta(\hat{X}, \hat{X}) - \Psi \cdot K_\theta(\hat{X}, W)^\intercal), \qquad (4)$$

where $\Psi = K_\theta(\hat{X}, W) \cdot K_\theta(W, W)^{-1}$. In this formulation, $W \in \mathbb{R}^{M \times (D_x + D_z)}$ is a matrix of M inducing point locations used to "summarize" the training data, enabling more efficient inference.

Stochastic Variational Inference (SVI). We learn the W and θ that minimize the divergence between an approximate and true posterior [21]:

$$W^*, \theta^*, \phi^* = \text{argmin}_{W, \theta, \phi} D_{\text{KL}} \left[q(F, U, Z; W, \theta, \phi) || p(F, U, Z | U, X, Z, Y; W, \theta) \right], \qquad (5)$$

using the variational family, $q(F, U, Z; W, \theta, \phi) = p(F | U, X, Z; W, \theta) \cdot q(U; \phi) \cdot \prod_{n=1}^N q(z_n; \phi)$, where ϕ are the parameters of full-covariance Gaussian $q(U; \phi) = \mathcal{N}(\mu_\phi, \Sigma_\phi)$ and mean-field Gaussians $q(z_n; \phi)$. This is equivalent to maximizing the evidence lower bound (ELBO) [22]:

$$\mathcal{L} = \sum_i \mathbb{E}_{q(f_i | X; W, \theta)} \left[\log p(y_i | f_i) \right] - D_{\text{KL}}[q(U; \phi) || p(U; W, \theta)] - \sum_{n=1}^N D_{\text{KL}}[q(z_n; \phi) || p(z)] \quad (6)$$

wherein the expectation is approximated via Monte Carlo by sampling $q(f_i | X; W, \theta) = \mathbb{E}_{q(U; \phi)} \left[p(f_i | U, X, Z; W, \theta) \right] \cdot \prod_{n=1}^N q(z_n; \phi)$, in which the expectation is computed analytically [23]: $\mathcal{N} \left(\psi_i \cdot \mu_\phi, \text{diag}(\psi_i \cdot (\Sigma_\phi - K_\theta(W, W)) \cdot \psi_i^\intercal) \right)$, with $\psi_i = K_\theta(\hat{x}_i^\intercal, W) \cdot K_\theta(W, W)^{-1}$. Since the first term of $\mathcal{L}$ can be estimated via mini-matching with just $O(M^3)$ per gradient step [23–25].

Computing Latent Similarity. We can better understand patient similarity, even in high dimensional latent spaces, by treating the LSGP's covariance as a graph, provided that the kernel over $\hat{x}$ can be decomposed into a product of kernels applied to x and z: $K_\theta(\hat{X}, \hat{X}') = K_\theta^x(X, X') \odot K_\theta^z(Z, Z')$. Specifically, we compute the covariance matrix between *patients* (not observations) by applying the latent-space kernel, K_θ^z, to the means of $q(z_n; \theta)$. We then treat this covariance as an adjacency matrix, in which patients are nodes and edge weights equal the covariance between the patients.

Exploring Similarity Explained by Demographics. To deepen our understanding of patient similarity, we can now explore how it aligns with demographics or social determinants of health in two ways. (1) Qualitatively: we visualize the similarity graph, pruning edges with covariances below a chosen threshold to reduce clutter. We then color vertices by group (whether demographic or other), as well as edges connecting patients within the same group, keeping remaining edges black (see Fig. 3). The relative number of colorful to black edges visually communicates how similar patients are to members of the same group. (2) Quantitatively: we can compare similarity within vs. between demographic groups using *modularity* [26]. Modularity, Q, lies in the range $[-1, 1]$,

Table 1: **Comparison of Methods on Test Metrics.** We report log-likelihood by stratifying patients into bottom, middle, and top thirds based on number of data points. Idiographic models consistently outperform their Single counterparts. The SV-LSGP outperforms all baselines except for VB-LR.

	Method	Avg. Log-Likelihood				ROC-AUC	PPV	Sensitivity	Specificity
		Bottom	*Middle*	*Top*	*All*	*All*	*All*	*All*	*All*
Single	**KNN**	N/A	N/A	N/A	N/A	0.70 ± 0.01	0.61 ± 0.02	0.22 ± 0.02	0.97 ± 0.00
	RBF-GP	-0.62 ± 0.01	-0.49 ± 0.01	-0.36 ± 0.00	-0.43 ± 0.00	0.74 ± 0.00	0.66 ± 0.04	0.15 ± 0.01	0.98 ± 0.00
	LR	-0.66 ± 0.00	-0.53 ± 0.01	-0.38 ± 0.00	-0.45 ± 0.00	0.68 ± 0.01	0.60 ± 0.06	0.05 ± 0.01	0.99 ± 0.00
	EB-LR	-1.87 ± 0.00	-1.72 ± 0.00	-1.09 ± 0.00	-1.56 ± 0.00	0.68 ± 0.01	0.61 ± 0.03	0.05 ± 0.01	0.99 ± 0.00
	VB-LR	-1.87 ± 0.00	-1.72 ± 0.00	-1.09 ± 0.00	-1.56 ± 0.00	0.68 ± 0.01	0.61 ± 0.05	0.05 ± 0.01	0.99 ± 0.00
Idiographic	**KNN**	N/A	N/A	N/A	N/A	0.78 ± 0.01	0.73 ± 0.02	0.34 ± 0.03	0.97 ± 0.00
	RBF-GP	-0.52 ± 0.01	-0.42 ± 0.01	-0.31 ± 0.00	-0.37 ± 0.00	0.84 ± 0.00	0.73 ± 0.01	0.37 ± 0.01	0.97 ± 0.00
	LR	-0.49 ± 0.01	-0.40 ± 0.01	-0.29 ± 0.01	-0.34 ± 0.01	0.85 ± 0.01	0.73 ± 0.01	0.36 ± 0.02	0.97 ± 0.00
	EB-LR	-0.54 ± 0.01	-0.51 ± 0.00	-0.38 ± 0.01	-0.43 ± 0.00	0.84 ± 0.01	0.71 ± 0.02	0.46 ± 0.03	0.96 ± 0.00
	VB-LR	-0.47 ± 0.01	-0.38 ± 0.01	-0.28 ± 0.01	-0.33 ± 0.01	0.87 ± 0.00	0.73 ± 0.01	0.42 ± 0.02	0.96 ± 0.00
	SV-LSGP	-0.50 ± 0.01	-0.40 ± 0.01	-0.29 ± 0.01	-0.35 ± 0.01	0.85 ± 0.01	0.73 ± 0.01	0.37 ± 0.02	0.97 ± 0.00

Gender: Modularity $Q = 0.12$ Sexual Orientation: Modularity $Q = 0.13$

Legend
- woman
- Unknown
- non-binary/gender non-conforming
- transgender man
- other
- man
- transgender woman

Legend
- prefer not to answer
- Unknown
- bisexual
- straight
- gay/lesbian
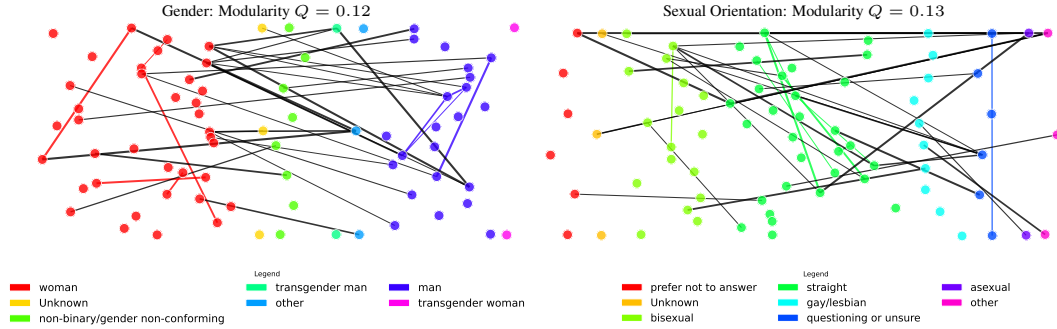- questioning or unsure
- asexual
- other

Figure 3: **Graphs of Patient Similarity.** Color of nodes represent group membership. Edges are black if connecting nodes of different groups; thickness indicates magnitude of covariance. The modularity for all graphs is close to 0, indicating balanced connections within/between groups.

with $Q = 1$ indicating only within-group edges, $Q = -1$ indicating only between-group edges, and $Q = 0$ when balanced. Originally proposed to identify community structures in networks, modularity has recently been used on covariance matrices, for example, to identify connectivity of brain regions (e.g. [27, 28]); here, we compute it on our similarity graph. Given a covariance matrix, $K \in \mathbb{R}^{N \times N}$, group membership g_n for every patient n, Q is computed via:

$$Q = \frac{1}{2S} \sum_{n,n'} \left[K_{n,n'} - \frac{\gamma}{2S} \cdot \left(\sum_i K_{i,n'} \right) \cdot \left(\sum_j K_{n,j} \right) \right] \cdot \mathbb{I}(g_n = g_{n'}), \tag{7}$$

where $S = \sum_{n,n'} K_{n,n'}$ and γ is the "resolution parameter" that controls the influence of between-group vs. within-group edges (commonly set to 1).

4 Experiments, Results, and Future Work

Without significant kernel design, our approach already outperforms all but one baseline, showing promise. We compare our method to baselines in ability to better forecast SREs one week in advance (details in Appendix B). Table 1 shows that, only having naively experimented with *a single kernel*, our method already nearly matches the best performing baseline on nearly all metrics. We anticipate that a future investigation into the inductive biases of different kernels will allow our method to outperform all baselines, since the LSGP generalizes both the GP and LR-based baselines.

Insights from Patient Similarity Graphs. We visualize the similarity of patients in Figs. 3 and 4 across different demographic groups and find that modularity for all graphs is close to 0, indicating balanced similarity within/between groups. This implies that these demographic factors do not explain patient similarity. This result aligns with Fig. 2, showing random groupings outperform demographic groupings. Interestingly, however, in nearly all graphs, the largest similarities are between patients of different groups (mostly black edges remained after pruning). In future work, we hope to explore patient similarity based on other factors to understand their clinical significance.

Acknowledgments and Disclosure of Funding

References

[1] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[2] Evan M Kleiman, Catherine R Glenn, and Richard T Liu. The use of advanced technology and statistical methods to predict and prevent suicide. *Nature reviews psychology*, 2(6):347–359, 2023.

[3] Evan M Kleiman, Brianna J Turner, Szymon Fedor, Eleanor E Beale, Jeff C Huffman, and Matthew K Nock. Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of abnormal psychology*, 126(6):726, 2017.

[4] Ewa K Czyz, Cheryl A King, Nadia Al-Dajani, Lauren Zimmermann, Victor Hong, and Inbal Nahum-Shani. Ecological momentary assessments and passive sensing in the prediction of short-term suicidal ideation in young adults. *JAMA Network Open*, 6(8):e2328005–e2328005, 2023.

[5] Chang Lei, Diyang Qu, Kunxu Liu, and Runsen Chen. Ecological momentary assessment and machine learning for predicting suicidal ideation among sexual and gender minority individuals. *JAMA network open*, 6(9):e2333164–e2333164, 2023.

[6] Shirley B Wang, Ruben DI Van Genugten, Yaniv Yacoby, Weiwei Pan, Kate H Bentley, Suzanne A Bird, Ralph J Buonopane, Alexis Christie, Merryn Daniel, Dylan DeMarco, et al. Building personalized machine learning models using real-time monitoring data to predict idiographic suicidal thoughts. *Nature Mental Health*, pages 1–10, 2024.

[7] Kathryn R Fox, Xieyining Huang, Eleonora M Guzmán, Kensie M Funsch, Christine B Cha, Jessica D Ribeiro, and Joseph C Franklin. Interventions for suicide and self-injury: A meta-analysis of randomized controlled trials across nearly 50 years of research. *Psychological bulletin*, 146(12):1117, 2020.

[8] Ewa K Czyz, Cheryl A King, and Inbal Nahum-Shani. Ecological assessment of daily suicidal thoughts and attempts among suicidal teens after psychiatric hospitalization: Lessons about feasibility and acceptability. *Psychiatry research*, 267:566–574, 2018.

[9] Evan M Kleiman, Brianna J Turner, Szymon Fedor, Eleanor E Beale, Rosalind W Picard, Jeff C Huffman, and Matthew K Nock. Digital phenotyping of suicidal thoughts. *Depression and anxiety*, 35(7):601–608, 2018.

[10] Aleksandra Kaurin, Alexandre Y Dombrovski, Michael N Hallquist, and Aidan GC Wright. Integrating a functional view on suicide risk into idiographic statistical models. *Behaviour research and therapy*, 150: 104012, 2022.

[11] Daniel DL Coppersmith, Evan M Kleiman, Alexander J Millner, Shirley B Wang, Cara Arizmendi, Kate H Bentley, Dylan DeMarco, Rebecca G Fortgang, Kelly L Zuromski, Joseph S Maimone, et al. Heterogeneity in suicide risk: Evidence from personalized dynamic models. *Behaviour research and therapy*, 180:104574, 2024.

[12] Kathleen M Gates, Stephanie T Lane, E Varangis, K Giovanello, and K Guiskewicz. Unsupervised classification during time-series model building. *Multivariate behavioral research*, 52(2):129–148, 2017.

[13] Stephanie T Lane, Kathleen M Gates, Hallie K Pike, Adriene M Beltz, and Aidan GC Wright. Uncovering general, shared, and unique temporal patterns in ambulatory assessment data. *Psychological methods*, 24 (1):54, 2019.

[14] Christian A Webb, Laura Murray, Anna O Tierney, and Kathleen M Gates. Dynamic processes in behavioral activation therapy for anhedonic adolescents: Modeling common and patient-specific relations. *Journal of Consulting and Clinical Psychology*, 2023.

[15] Keith Hawton and Kees Van Heeringen. *The international handbook of suicide and attempted suicide*. John Wiley & Sons, 2000.

[16] Hilario Blasco-Fontecilla, Maria Rodrigo-Yanguas, Lucas Giner, Maria Jose Lobato-Rodriguez, and Jose De Leon. Patterns of comorbidity of suicide attempters: an update. *Current psychiatry reports*, 18(10):93, 2016.

[17] Chunyi Wang and Radford M Neal. Gaussian process regression with heteroscedastic or non-gaussian residuals. *arXiv preprint arXiv:1212.6246*, 2012.

[18] Kaspar Märtens, Kieran Campbell, and Christopher Yau. Decomposing feature-level variation with covariate Gaussian process latent variable models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4372–4381. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/martens19a.html`.

[19] Didong Li, Andrew Jones, Sudipto Banerjee, and Barbara E. Engelhardt. Bayesian multi-group gaussian process models for heterogeneous group-structured data. *Journal of Machine Learning Research*, 26(30): 1–34, 2025. URL `http://jmlr.org/papers/v26/23-0291.html`.

[20] Steindór Sæmundsson, Katja Hofmann, and Marc Peter Deisenroth. Meta reinforcement learning with latent variable gaussian processes. *arXiv preprint arXiv:1803.07551*, 2018.

[21] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009.

[22] Michalis Titsias and Neil D. Lawrence. Bayesian gaussian process latent variable model. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 844–851, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL `https://proceedings.mlr.press/v9/titsias10a.html`.

[23] James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable Variational Gaussian Process Classification. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360, San Diego, California, USA, 09–12 May 2015. PMLR. URL `https://proceedings.mlr.press/v38/hensman15.html`.

[24] James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 282–290, Arlington, Virginia, USA, 2013. AUAI Press.

[25] Vidhi Lalchand, Aditya Ravuri, and Neil D. Lawrence. Generalised gplvm with stochastic variational inference. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7841–7864. PMLR, 28–30 Mar 2022. URL `https://proceedings.mlr.press/v151/lalchand22a.html`.

[26] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[27] Je-Yeon Yun, Premika SW Boedhoe, Chris Vriend, Neda Jahanshad, Yoshinari Abe, Stephanie H Ameis, Alan Anticevic, Paul D Arnold, Marcelo C Batistuzzo, Francesco Benedetti, et al. Brain structural covariance networks in obsessive-compulsive disorder: a graph analysis from the enigma consortium. *Brain*, 143(2):684–700, 2020.

[28] Gopalkumar Rakesh, Mark W Logue, Emily Clarke-Rubright, Courtney C Haswell, Paul M Thompson, Michael D De Bellis, Rajendra A Morey, and Delin Sun. Network centrality and modularity of structural covariance networks in posttraumatic stress disorder: a multisite enigma-pgc study. *Brain Connectivity*, 13 (4):211–225, 2023.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[30] Amazasp Shaumyan. GitHub: AmazaspShumik/sklearn-bayes: Python package for Bayesian Machine Learning with scikit-learn API. `https://github.com/AmazaspShumik/sklearn-bayes?tab=readme-ov-file#contributions`. [Accessed 01-09-2025].

[31] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.

[32] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

A Overview of the EMA Data

Participants. A total of 623 unique participants presenting with suicidal thoughts and/or recent suicidal behavior were recruited from two hospitals in the Boston area—315 adults (ages 18+) from a psychiatric emergency service, and 308 adolescents (ages 12-19) from a psychiatric inpatient unit. Participants were excluded if they did not own an iOS/Android smartphone, they presented any factor that impaired their ability to provide informed consent/assent, an inability to speak or write English fluently, a gross cognitive impairment due to florid psychosis, intellectual disability, dementia, acute intoxication, or extremely agitated or violent behavior.

Consent, Compensation, and IRB. After agreeing to participate, individuals signed consent/assent forms, answered an initial questionnaire, and installed the LifeData application on their mobile devices, which prompted them with brief self-report questionnaires. Participants received $10 for completing the initial questionnaire and earned $1 for each EMA survey they submitted. The study was approved by our institutions' IRB.

Surveys. Smartphone surveys assessed participants' current experience of suicidal thinking—urge, intent, and ability to resist suicidal urges—as well as 17 affective states—negative, hopeless, trapped, isolated, burdensome, angry, self-hate, agitated, worried, numb, fatigued, humiliated, desire to escape, desire to avoid, energetic, and positive—on a 0-10 likert scale. These surveys were sent to participants 6-times per day for three months, with the first and last sent at fixed times decided in collaboration with each participant, and the remaining surveys sent at randomized times between the first and last surveys. In addition to these surveys, participants could always opt to fill in additional surveys, for example, to report a suicide attempt, non-suicidal self-injury, or another event they deemed important. They study was monitored by a risk-monitoring team in real-time to intervene when participants indicate high suicidal intent (details available upon request).

Recording SREs. An SRE was recorded in the data if it was reported by the patient in the survey, if it was reported by the risk-monitoring team, or if it was reported in the patient's electronic health record (consensus coded by two trained BA-level reviewers with supervision by a doctoral-level clinician with expertise in assessing/treating STBs).

Data Inclusion in Analysis. We kept all SREs for which there was at least one EMA survey in the week prior. We kept data from all patients that had at least 3 SREs and 3 non-SREs to ensure we can include one of each in the train/validation/test split (see Appendix B). Due to the low base-rate of SREs, this left us with $N = 77$ patients who contributed a total of 14763 complete EMA surveys.

B Experimental Setup

Data Splits. We divided the data into 50%, 25%, and 25% sized-sets for training, validation, and test, respectively. We ensured that there was at least one SRE and one non-SRE in each set. As such, we assume that for our method to be used in practice, patients must have at least one recorded SRE in their data. We created these cuts of the data 5 times, conducting all experiments on each cut of the data, and reporting the mean $\pm$ standard deviation of all metrics.

Random Restarts. For each of cut of the data, we ran each method 5 times, each with a random seed. We selected the best performing random restart on the validation Log-Likelihood.

Hyperparameter Selection. We performed grid search over the following parameters, selecting them based on ROC-AUC on the validation set:

- **KNN:** Neighbors $k \in \{1, 2\}$, which performed best in our preliminary experiments, and distance $\in \{\text{Minkowski}, \text{Manhattan}\}$. We used the default parameters from `scikit-learn` [29] for the remaining parameters.

- **LR:** Default parameters from `scikit-learn` [29] but with a maximum of 5000 iterations until convergence.

- **VB-LR:** We trained for a maximum of 5000 iterations until convergence, with the rate and scale α, β on the Gamma prior on precision of the coefficients both $\in \{1.0, 2.0\}$, and with the remaining parameters set to the defaults from Shaumyan [30].

- **EB-LR:** We trained for a maximum of 5000 iterations, with the initial precision of prior distribution $\alpha \in \{3.0, 2.0, 1.0, 1e-3, 1e-6, 1e-9, 1e-12\}$, and with the remaining parameters set to the defaults from Shaumyan [30].

- **GP:** We used the default GP hyperparameters from `scikit-learn` [29], which uses an automatic relevance determination (ARD) kernel. We additionally set `max_iter_predict` $= 5000$, as well as `n_restarts_optimizer` $= 1$, which selects across two kernel hyperparameter initializations—default and random.

- **SV-LSGP:** We use $M = 2000$ inducing points, $D_z = 3$, mini-batch size $B = 150$. We fit the model with 15000 gradient steps and a learning rate of 0.005. Finally, we used a kernel that factorizes as $K_\theta(\widehat{X}, \widehat{X}') = K_\theta^x(X, X') \cdot K_\theta^z(Z, Z')$, with K_θ^z as an ARD kernel and K_θ^x as a "state-dependent" linear kernel, defined by:

$$k_\theta(\hat{x}, \hat{x}') = b_\theta(z) \cdot b_\theta(z') + v_\theta(z) \cdot v_\theta(z') \cdot (x - c_\theta(z))^\mathsf{T} \cdot (x' - c_\theta(z')), \tag{8}$$

where $b_\theta(\cdot)$, $v_\theta(\cdot)$, and $c_\theta(\cdot)$ are each a neural network with parameters θ. This kernel allows patients to have differing priors depending on their location in the latent space.

Software. We implemented the SV-LSGP in NumPyro [31] and Jax [32].

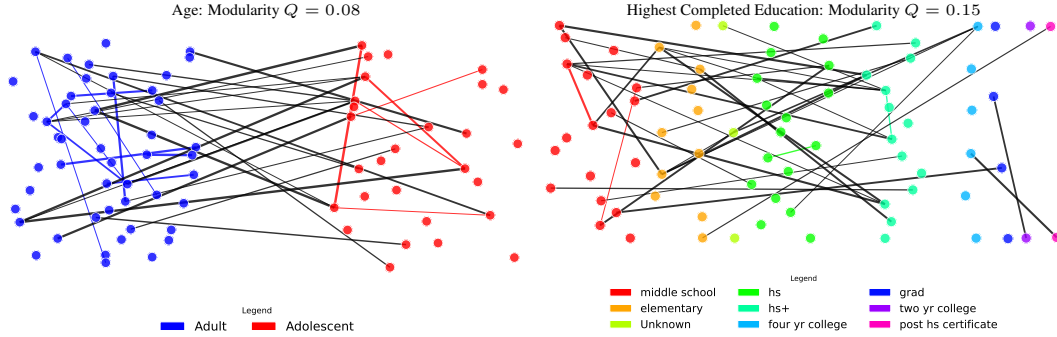## C   Additional Visualizations of Patient Similarity

Figure 4: **Graphs of Patient Similarity.** Color of nodes represent group membership. Edges are black if connecting nodes of different groups; thickness indicates magnitude of covariance. The modularity for all graphs is close to 0, indicating balanced connections within/between groups.