Hybrid Feature- and Similarity-Based Models for Prediction and Interpretation using Large-Scale Observational Data

Jacqueline K. Kueper  
Department of Epidemiology and Biostatistics  
Department of Computer Science  
Western University  
London, Ontario, Canada  
JKUEPER@UWO.CA

Jennifer Rayner  
The Alliance for Healthier Communities  
Ontario, Canada  
JENNIFER.RAYNER@ALLIANCEON.ORG

Daniel J. Lizotte  
Department of Epidemiology and Biostatistics  
Department of Computer Science  
Western University  
London, Ontario, Canada  
DLIZOTTE@UWO.CA

Abstract

Introduction: Large-scale electronic health record (EHR) datasets often include simple informative features like patient age together with complex data like care history that are not easily represented as individual features. Such complex data have the potential to both improve the quality of risk assessment and to enable a better understanding of causal factors leading to those risks. For example, increased age may be associated with risk, but that relationship may not persist if we account for care history. We propose a hybrid feature- and similarity-based model for supervised learning that combines feature and kernel learning approaches to take advantage of rich but heterogeneous observational data sources to create interpretable models for prediction and for investigation of causal relationships.

Methods: The hybrid model we propose is fit by convex optimization with a sparsity-inducing penalty on the kernel portion of the model. Depending on the desired interpretation, the model can be fit with feature and kernel coefficients sequentially or simultaneously. We compared our models to solely feature- and similarity-based approaches using synthetic data and using real EHR data from a primary health care organization to predict client risk of future loneliness or social isolation. We also present a new strategy for kernel construction that is suited to high-dimensional indicator-coded EHR data.

Results: The hybrid models had comparable or better predictive performance than the feature- and kernel-based approaches in both the synthetic and clinical case studies. The synthetic experiments showed the performance benefits are highest when feature and kernel data have similar importance for explaining the outcome. The inherent interpretability of the hybrid model was used to learn about model behaviour and the population of interest in the clinical case study by exploring feature and kernel data characteristics of clients with zero, negative, and positive kernel coefficients. We use simple examples to further discuss opportunities and cautions of the two hybrid model forms when causal interpretations are desired.
**Conclusion**: Hybrid feature- and similarity-based models provide an opportunity to capture complex, high-dimensional data within an additive model structure that supports improved prediction and interpretation relative to simple models and opaque complex models. Although this study is focused on health care settings, the methods can be extended for use with other datasets and applications.

1. **Introduction**

Health care settings generate large amounts of data and yet it can be challenging to fully harness these data for machine learning applications. For machine learning tasks with large-scale observational data, there are often known, informative features as well as additional data that may be useful for the task but that are challenging to summarize into meaningful features due to size or complexity. For example, electronic health records (EHRs) capture client characteristics (e.g., year of birth) in structured fields and record information arising from each encounter (e.g., date-stamped diagnosis and procedure codes) in dynamic tables. The former may be well suited for features while the latter high-dimensional, variable length data may be better represented in terms of similarity to other clients in the database. Explainable, reproducible methods that take full advantage of these rich data are needed to support further advancements in the field of machine learning for healthcare (Payrovnaziri et al., 2020; Guo et al., 2020; Ghassemi et al., 2020; Morgenstern et al., 2020).

Feature- and similarity-based models have complimentary characteristics. Feature-based approaches, such as logistic regression (LR), tend to be more familiar to end-users, less susceptible to overfitting, and easier to interpret (e.g., viewing regression coefficients or the structure of a decision tree); however, not all valuable information can be captured with features and model performance may suffer from underfitting, especially for heterogeneous populations. In contrast, similarity-based approaches such as multiple kernel learning have a higher computation cost but can incorporate more complex or time-varying data that may account for additional variability in the outcome (Shawe-Taylor and Cristianini, 2011; Conroy et al., 2017; Gonen et al., 2011). Interpretation of similarity-based approaches is not as straightforward as for feature-based methods, but can include strategies such as summarizing characteristics about the most similar training examples to the one for whom a prediction is being made (Fang et al., 2021; Tonekaboni et al., 2019). Similarity-based approaches are not interpreted on their own for the purpose of causal inference, while feature based approaches may be, either explicitly in estimating a treatment effect or implicitly by interpreting feature coefficients to identify risk factors to intervene on.

We present two variations on an intrinsically interpretable hybrid feature- and similarity-based model (HFSM) and demonstrates their use with synthetic data and with EHR data from a complex primary health care population. The model form is able to support traditional causal interpretations of feature coefficients while reaping additional benefits from similarity based approaches, such as improved absolute risk prediction while maintaining traditional feature interpretations, and adjustment for complex confounders. We demonstrate how the HFSM approach can outperform solely feature- or similarity-based methods while retaining or enhancing interpretability.
Hybrid Feature- and Similarity-Based Models

Generalizable Insights about Machine Learning in the Context of Healthcare

Our primary contributions are through the hybrid model, both in presenting the model structure and in describing the types of supervised learning scenarios where combining feature- and similarity-based approaches within an inherently interpretable model may be beneficial. These include prediction-oriented tasks and exploratory or causal analyses. We describe how our model can be used for epidemiological research and use simple examples to demonstrate situations where feature coefficients may become more or less biased depending on other characteristics of the model—these concepts apply to \textit{HFSM} as well as to any other multivariable/multicomponent model that may be interpreted for decision making.

Additional contributions are made through our clinical case study, wherein we apply \textit{HFSM} using a new strategy for building kernels that assess similarity in terms of both the presence of rare care characteristics and the absence of common care characteristics. Assessing similarity in terms of what expected characteristics are missing may be useful for other settings (e.g., public health, emergency room triage) where two people that deviate from population-level expectations are more similar than if they fit the expected profile. We discuss additional challenges encountered in our applied setting that are relevant to other health care contexts as well, such as the open cohort nature of primary health care and decisions related to features that are informative but rare.

2. Related Work

Our hybrid model approach contributes to two general areas of research: 1) methods designed to incorporate multiple sources or types of data and 2) combining simple and complex models to make a single prediction.

Integrating multiple types of data  

Recommender systems, e.g. for movies or products, are often designed for settings with two distinct types of data: 1) user attributes, such as demographic information and 2) time-varying, high-dimensional information arising from user interactions with a system, such as histories of movie viewings or ratings. Fan et al. (2017) developed \textit{RIT-UA}, which makes predictions based on a weighted linear combination of two similarity scores: one based on a weighted count of common attributes and one based on sigmoid functions applied to historical data about user preferences and ratings. Our \textit{HFSM} approach is designed to handle data with a similar structure; however, the \textit{RIT-UA} generates scalar similarity scores to combine information from the two data sources whereas \textit{HFSM} maintains separate model structure to use information from features directly.

Multiview learning combines multiple data types to improve predictive performance. Lian et al. (2015) proposed a framework that assumes all feature and/or similarity matrices contribute a different “view” of the data. A shared latent factor matrix is learned to serve as a global representation of the data. Multiple kernel learning is a special case within this framework where each view is treated as a kernel matrix (Lian et al., 2015; Shawe-Taylor and Cristianini, 2011; Conroy et al., 2017; Gonen et al., 2011). While there is overlap in the similarity-based part of this approach to \textit{HFSM}, including the possibility to incorporate multiple kernel learning techniques, \textit{HFSM} maintains separation of the feature matrix in a way that also prioritizes interpretation of individual feature coefficients.
Hybrid Feature- and Similarity-Based Models

Combining model types  Boosting approaches may also handle diverse data types by combining complimentary model forms to improve predictions. Hothorn et al. (2010) developed mboost, a component-wise boosting algorithm that combines penalized least square estimates and/or regression tree base learners in an additive model structure. Each component may be applied to all or a subset of data, is weighted in the fitted model, and can be interpreted separately (Hothorn et al., 2021, 2010). Our sequentially-optimized HFSM approach is similar to mboost, but uses different components and does not employ an overall weight for each model component. Building on mboost, Sigrist (2021) developed KTBoost, which learns both a regression tree and a reproducing kernel Hilbert space regression function on all available data in each iteration, and then adds the one that is expected to result in better performance to the ensemble of base learners. This approach does not segregate data and does not allow for feature and kernel coefficients to be jointly optimized as in our simultaneous HFSM approach. A popular gradient boosting technique is XGBoost, which continues to fit new decision tree models to account for residual errors from previous models until performance stops improving (Chen and Guestrin, 2016). XGBoost has demonstrated excellent predictive performance in several settings, but as with the other boosting techniques, the focus is on predictions. Our HFSM is parametric with a fully convex objective function; this supports reproducibility, which is particularly important when interpretation of the model may be used to learn about a population or to support clinical decision making.

3. Methods

We address supervised learning tasks where the outcome of interest $y(o)$ for a particular observation $o$ can be explained partly by constructed features $\phi(o)$ (i.e. scalar or fixed-length vectors that represent a characteristic or property of $o$) and partly by more complex information $\Psi(o)$ (e.g., high-dimensional, time-varying, variable-length data). For example, in a primary health care setting each observation could be a client and the outcome of interest a condition or situation, such as diabetes or food insecurity, that the client is at risk for and early intervention may help to prevent. In this setting, $\phi(o)$-type information may include sociodemographic characteristics and core diagnoses while $\Psi(o)$-type information may include years of encounter data representing the subset of thousands of possible tests, diagnoses, and procedures that the client has received in their lifetime. Additional background on feature- and similarity-based models is provided in Appendix A.1.

The Hybrid Feature- and Similarity-Based Model (HFSM)

The proposed HFSM combines a feature-based component and a similarity-based component with an additive model structure. Prediction $\hat{y}(o)$ for observation $o$ is given by

$$\hat{y}(o) = h \left( \sum_j \phi_j(o) \beta_j + \sum_i \alpha_i \sum_l k_l(\Psi_l(o), \Psi_{l}(a_i)) \right) = h (\phi(o)^T \beta + k(o)^T \alpha)$$

where $j$ indexes the features; $i$ indexes the observations or clients in the training data; $l$ indexes the kernel domains, if there are multiple; and $h$ is a monotonic function, e.g., sigmoid or identity. $k(o)$ is the vector of kernel values between $o$ and each training data point. All analyses in this paper use the sigmoid function to estimate the probability of a
binary outcome occurring. Thus, the estimated probability of an outcome occurring for \( o \) is based on 1) their feature values and the corresponding coefficients (\( \beta ) \) and 2) similarity to clients from the training data and the overall influence (\( \alpha ) \) of each client.

To train the model, we optimize a penalized log likelihood training criterion given by

\[
LL(\beta, \alpha; \lambda) = \left( \sum_i y_i (\phi(o_i)^T \beta + k(o_i)^T \alpha) - \log(1 + e^{\phi(o_i)^T \beta + k(o_i)^T \alpha}) \right) / n - \lambda ||\alpha||_1.
\]

The L1-penalty on \( \alpha \) controls overfitting and produces a sparse model whose kernel component only depends on a subset of the training data; this is different from the original kernel logistic regression formulation which penalizes the norm of the regression function in its Hilbert space but does not induce sparsity (Zhu and Hastie, 2005). Training \( o_i \) that maintain non-zero \( \alpha \) can be thought of as “representatives” for groups of similar clients. We solve this problem using the convex programming language cvxpy in Python (Diamond and Boyd, 2016; Agrawal et al., 2018). An illustrative example and Python code are provided in Appendices A.2 and B, respectively. Required memory for model fitting, assuming \( n \) clients and \( m \) features, is \( O[mn + n^2] \) for HFSM as compared to \( O[mn] \) for LR and \( O[n^2] \) for kernel logistic regression (KLR). Solve time will be compared in the experiments.

**Fitting and interpretation** We consider two variations on fitting HFSM that have different interpretations: 1) HFSM-Sequential (HFSM-Seq), which learns the feature coefficients fixing \( \alpha = 0 \) and then fixes them before learning the kernel coefficients, and 2) HFSM-Simultaneous (HFSM-Sim), which optimizes the feature and kernel coefficients jointly. The simultaneous model fit is expected to result in better predictive performance since there is more flexibility to maximize the objective function, but the resulting model has a more complex causal interpretation. In HFSM-Seq, the feature coefficients represent their impact on the outcome adjusted for all of the other features in the model but averaged over the information in the kernel, whereas HFSM-Sim feature coefficients are additionally adjusted for the information in the kernel. We discuss the implications for interpretation below. If the feature and kernel matrices are orthogonal, the models produced by the two procedures will be identical.

A series of four illustrative examples contrast the performance and interpretation of HFSM-Seq and HFSM-Sim in terms of causal inference. For each example there is a binary outcome \( y \), one continuous feature \( X_1 \sim N(0, 1) \) that maintains a direct relationship with \( y \), and a binary feature \( X_2 \) whose relationship with \( y \) and \( K \) is manipulated. For simplicity, \( K \) is unpenalized and constructed from a linear kernel function applied to a single binary variable. There are four examples, represented in Figure 1:

1. **Independent contributions.** \( P(Y) = \varsigma(0.25 - 1X_1 + 2X_2 + 3K) \) where \( X_2 \sim B(p = 0.5) \), and \( K \sim B(p = 0.5) \).
2. The kernel acts as a **confounder** between the second feature and the outcome. \( P(Y) = \varsigma(0.25 - 1X_1 + 3K) \) where \( K \sim B(p = 0.5) \) and \( P(X_2) = \varsigma(2K) \).
3. The kernel is a **collider** between the outcome and on the second feature. \( P(Y) = \varsigma(0.25 - 1X_1) \) where \( P(K) = \varsigma(3y + 2X_2) \) and \( X_2 \sim B(p = 0.5) \).
4. The kernel is a **mediator** between the second feature and the outcome. \( P(Y) = \varsigma(0.25 - 1X_1 + 3K) \) where \( P(K) = \varsigma(2X_2) \) and \( X_2 \sim B(p = 0.5) \).
For each example, feature coefficients are compared for \textit{HFSM-Seq}, \textit{HFSM-Sim}, and \textit{LR} fit on 3,000 training examples and predictive performance is compared based on area under the receiver operating characteristic curve (AUROC) for 1,000 new test examples.

![Diagram showing data generating mechanisms for four examples: Example 1, Example 2, Example 3, and Example 4.]

Figure 1: Data generating mechanisms used to contrast sequential and simultaneous hybrid model optimization

As seen in Table 1(a) subtable, \textit{HFSM-Sim} had the best predictive performance for all examples; however, as seen in Table 1(b) subtable the corresponding \textit{HFSM-Sim} feature coefficient estimates could be closer to the truth, further from the truth, or similar to the feature coefficients learned in \textit{HFSM-Seq}. While the impacts of adjusting the feature coefficients by the kernel are predictable for these simple experiments, in practice the direction of bias, if any, may be hard to determine. This uncertainty is analogous to situations with solely feature-based approaches where the relationships between the features and the outcome are unknown, or when automatic feature selection methods are used (Brookhart et al., 2010; Greenland and Morgenstern, 2001; Shrier and Platt, 2008; Austin and Tu, 2004).
An interpretation advantage of the *HFSM-Sim* approach over solely feature-based approaches is the opportunity to adjust for more complex types of confounding information than can be adequately captured through features. When the information captured by the kernel is uncertain, *HFSM-Seq* can be used to maintain straightforward feature coefficient interpretation while still improving the absolute risk prediction through the addition of the kernel. Thus, hybrid models may be used much in the same way that logistic regression can be used for prediction or inference depending on whether one focuses on the predicted outcome or model coefficients, respectively, with the advantage of using a kernel to account for additional variability.

The kernel coefficients $\alpha$ may be also be informative in and of themselves. Whereas applying an L1-penalty to features is a form of feature selection, applying an L1-penalty to $\alpha$ selects “representative observations” to include while adjusting for the features. The higher the penalty, the fewer observations are allowed. The most influential clients in the training data (highest magnitude $\alpha$) can be investigated to explore kernel behaviour. A prediction for an individual client is based on their feature values and corresponding $\beta$, and then will be further increased or decreased depending on similarity in terms of the kernel to clients in the training data that have non-zero $\alpha$. Similarity to clients with positive $\alpha$ will increase the predicted probability while similarity to clients with negative $\alpha$ will decrease the predicted probability of the outcome.
4. Evaluation

The performance of the HFSM approach is compared to solely feature- or similarity-based approaches with 1) a simulation study of three synthetic data scenarios (S1-S3) that vary the relative importance of the feature- and kernel-based data and 2) a clinical case study with EHRs from a primary health care organization in Ontario.

4.1. Simulation Study

This study compares the HFSM approaches to the two most direct sub-component models as in an ablation study. The ADEMP framework was followed for planning and reporting on this study (Morris et al., 2019). The data generating mechanism is based on a parametric model that most closely corresponds to Example 1 (independent contributions) in the illustrative examples above, with four binary features and additional complex information that cannot be well represented by features in a linear model. This latter information is based on the classic Monk-1 data problem, which includes 6 categorical variables (two-level variables: \(a_3, a_6\); three-level variables: \(a_1, a_2, a_4\); four-level variable: \(a_5\)) and an outcome \(M\) that is the result of the Boolean statement \((a_1 = a_2) \lor (a_5 = 1)\) (Thrun et al., 1991). The outcome is used in the data generating mechanism. For models with a similarity-based component, the RBF kernel function will be applied to the six categorical variables. The RBF kernel was selected due to its popularity in machine learning applications and its use in previous work with Monk’s data problems (Belanche and Villegas, 2013; Márquez, 2014).

The data generating mechanism is
\[
P(Y) = \varsigma(\beta_0 + 0.3X_1 + 0.4X_2 + 0.6X_3 + 0.7X_4 + \delta M)
\]
Coefficients were decided such that if \(\beta_0 = 0, \delta = 0, \text{ and } \sum_{m=1}^{4} \beta_m = 2\) then \(P(Y)\) ranges from 0 to 0.88 and \(\delta\) can be used to further increase the maximum probability of the outcome. Across the three experiments the \(\beta\) are fixed and \(\delta\) is changed to vary the relative importance of the feature- and kernel-based data. The intercept \(\beta_0\) is used to bring the prevalence of the outcome below 50% to be more similar to most clinical outcomes. Three scenarios are set up with 10,000 observations generated from each, which is similar to the number of clients expected across a few small primary health care clinics.

1. Kernel has a similar effect to a single feature: \(\delta = \text{mean}(\beta)\) and \(\beta_0 = -1.5\)
2. Kernel has a similar effect to the set of features: \(\delta = \text{sum}(\beta)\) and \(\beta_0 = -2.1\)
3. Kernel has a larger effect than the set of features: \(\delta = 2 \cdot \text{sum}(\beta)\) and \(\beta_0 = -3.2\)

For each scenario, a nested cross validation (CV) procedure is implemented whereby for each of five outer folds, the outer fold training data are split 75/25 into inner fold training and validation data. To reduce random variation between the models, the same outer and inner CV folds will be used for each model; seeds are re-set between scenarios. Hyperparameters, if any, are selected through a grid search for the best AUROC on the inner validation data. Models are then re-trained with the selected hyperparameter(s) on all outer fold training data, and predictions of the target binary outcome are made on the outer fold test data. Folds were trained in parallel using the python package multiprocessing (McKerns and Aivazis, 2010; McKerns et al., 2011).

Four models are compared: 1. Feature only model: \(LR\), 2. Similarity only model: \(KLR\), 3. \(HFSM\)-Seq, and 4. \(HFSM\)-Sim. For kernel-containing models, three candidate hyperparameters \(\sigma\) for the RBF kernel are considered (0.01, 0.1, 1). These values provide a range of similarity patterns on the Monk’s data, based on a measure of matrix diagonal
Hybrid Feature- and Similarity-Based Models

dominance and visual exploration of the RBF kernel calculated for a random sample of 1000 observations (Appendix C). For each $\sigma$, five candidate values for the L1-regularization strength $\lambda$ on the kernel coefficients, ranging from 0.001 to 1, are considered.

Model performance is compared using measures averaged across the five outer test folds. The primary metric of interest is discrimination, assessed through AUROC. Secondary metrics of interest include the area under the precision recall curve (AUPRC), calibration plot slopes and intercepts, and time to re-train the model with selected hyperparameters. Predictive performance metrics are also calculated for a “best possible model” that makes predictions based on applying the known coefficients to all data. Hyperparameters and parameters are viewed and compared between models.

The expected trends emerge across the three scenarios: the hybrid models always perform similar to or better than the single component models, with a notable advantage for the second scenario (Table 2). Selected hyperparameters and learned parameters are in Appendix B. In the two extreme scenarios, the hybrid models perform similarly to whichever single-component model captures the more important portion of the data. In the “feature heavy” scenario 1, $HFSM$ performance comes with a large increase in computation time (seconds vs hours) as compared to $LR$. In the “kernel heavy” scenario 3, the discrimination performance of $LR$ approaches a dummy classifier while $KLR$ shows similar predictive performance to the hybrid models and increased fitting time as compared to $HFSM$-Sim. In the “middle ground” scenario 2, $HFSM$ demonstrates the best discrimination and precision-recall performance, but neither $HFSM$ version outperformed $LR$ in terms of calibration. For all models with a feature component, as the relative importance of the kernel-based data increased, the feature coefficient estimates got further from the truth (Appendix C).

4.2. Clinical Case Study

We present a case study with EHR data from the Alliance for Healthier Communities, which provides inter-professional, team-based primary health care through Community Health Centres (CHCs) across Ontario, Canada (Albrecht, 1998; Alliance for Healthier Communities, 2020). All CHCs record standardized sociodemographical information (e.g., birth date, education, household income) and appointment details (e.g., care provider type, diagnosis codes) in a centralized, structured EHR database. We use de-identified data from January 1, 2009 to December 31, 2019 to predict two-year risk of first incidence loneliness or social isolation for middle-aged clients being served by the “urban-at-risk” (UAR) peer group of CHCs. This subgroup of CHCs provides care to clients with pre-existing substance use, homelessness, or mental health challenges. Although this case study is primarily intended to test the proposed methods, we selected this outcome as there are a range of services and programs offered through these CHCs that may help mitigate the risk of this outcome, such as their social prescribing initiatives (Hsiung, 2022; Nowak and Mulligan, 2021; Mulligan et al., 2020). This study was approved by XXX ethics board (project ID XXX).

Cohort The cohort of interest includes ongoing primary care clients at UAR CHCs without the outcome at baseline. To restrict the sample to new or newly returning, mid- to long-term clients, only those whose first event is recorded in 2010 or later and who had at least one event three years from the first recorded event were eligible for inclusion. Primary health care is provided at all stages of life and health and social isolation or loneliness may
Hybrid Feature- and Similarity-Based Models

Table 2: Simulation Study Results

| Scenario 1: Kernel data have similar effect to a single feature | LR | KLR | HFSM-Seq | HFSM-Sim | Best \(^1\) |
|---------------------------------------------------------------|----|-----|----------|----------|---------|
| AUROC                                                        | 0.647 | 0.504 | 0.648 | 0.647 | 0.655 |
| AUPRC                                                        | 0.571 | 0.436 | 0.572 | 0.573 | 0.581 |
| Calibration Slope                                            | -0.025 | -0.362 | -0.032 | -0.013 | -0.008 |
| Calibration Intercept                                        | 1.035 | -0.517 | 1.038 | 1.031 | 0.989 |
| Time (hours)                                                 | < 1 | 6.553 | 7.815 | 7.177 |

| Scenario 2: Kernel data have similar effect to the set of features | LR | KLR | HFSM-Seq | HFSM-Sim | Best \(^1\) |
|--------------------------------------------------------------------|----|-----|----------|----------|---------|
| AUROC                                                              | 0.614 | 0.712 | 0.725 | 0.726 | 0.781 |
| AUPRC                                                              | 0.587 | 0.672 | 0.708 | 0.710 | 0.759 |
| Calibration Slope                                                  | -0.001 | 0.043 | 0.044 | 0.008 | 0.027 |
| Calibration Intercept                                              | 0.993 | 1.415 | 1.305 | 1.257 | 1.017 |
| Time (hours)                                                       | 0.001 | 9.901 | 10.574 | 9.316 |

| Scenario 3: Kernel data have a larger effect than the set of features | LR | KLR | HFSM-Seq | HFSM-Sim | Best \(^1\) |
|---------------------------------------------------------------------|----|-----|----------|----------|---------|
| AUROC                                                               | 0.558 | 0.872 | 0.877 | 0.877 | 0.903 |
| AUPRC                                                               | 0.538 | 0.825 | 0.840 | 0.846 | 0.879 |
| Calibration Slope                                                   | -0.001 | 0.010 | 0.012 | -0.058 | 0.018 |
| Calibration Intercept                                               | 0.980 | 1.557 | 1.575 | 1.534 | 1.018 |
| Time (hours)                                                        | < 1 | 7.148 | 7.419 | 5.111 |

\(^1\)Best = Hardcoded true coefficients applied to all data

occur at any point, so we randomly selected two-year periods from each client’s observation history to serve as the prediction interval. Feature and kernel input data used to make predictions are from the first recorded event to the beginning of the randomly selected prediction interval. The start of the prediction interval had to be at least one year from the first recorded event as the first year of care provision in this population is associated with a distinct risk profile, likely due to “catch-up” on unresolved care and diagnoses (Kueper et al., 2022). We restrict our cohort to those 45-64 years old at the end of their baseline period as age is associated with the outcome and may influence the risk factors and potential interventions to help someone at high-risk.

Feature choices We identified 19 candidate features based on evidence in the literature (Altschul et al., 2021; Doryab et al., 2019; Koning et al., 2017; Nicholson, 2012; National Academies of Sciences, Engineering, and Medicine, 2020; Holt-Lunstad, 2017; World Health Organization, 2021a), perceived importance with input from Alliance stakeholders, and feasibility to construct with available data. Features from the client characteristic table were handled with complete case analysis if under 1% missingness and with a missingness indicator approach otherwise for 1) client was asked the question and preferred not to respond and 2) client was never asked. Features constructed using International Classification of Disease (ICD-10) (World Health Organization, 2020) and Electronic Nomenclature and Classification Of Disorders and Encounters for Family Medicine (ENCODE-FM) (ENCODE-FM,
vocabularys were assumed absent if no appropriate codes were present during baseline. Three of these features had under 1% prevalence in baseline data and were excluded from the model (people with the features present were not excluded from the cohort); we performed indirect standardization to assess sex-adjusted risk in associated subpopulations.

We also constructed a feature to represent general clinical complexity as the count of the number of chronic conditions identified as important for multimorbidity research in primary care present during baseline (Fortin et al., 2017), scaled to 0,1 range. This type of non-specific, complex information is what we design the kernels to capture; this composite feature represents what we may try to include instead for a solely feature-based model.

Kernel choices In addition to the specific conditions identified for features, there is a sense that general health complexity may be positively associated with the outcome. We use three types of kernel input data based on appointment-associated care characteristics to capture this additional complex information: 1) the provider type(s) involved in care (e.g., nurse practitioner, social worker), 2) the service type(s) provided during an appointment, which represents the general type of care functions provided (e.g., assessment, treatment management) without specifying conditions, and 3) both 1 and 2. There are many ways data could be pre-processed and combined for kernel inputs; we work with sets and add together distinct codes experienced at least once during baseline care.

A valid kernel function must be symmetric and result in a positive semi-definite kernel matrix. Additional properties that we want our kernel function to have include:

1. Holding all else constant, two clients who both have or do not have a specific code should be more similar than when only one of them has the code present.
2. Two clients who do not have a code that is common in the population of interest should be more similar than two people who both have the common code present.
3. Two clients who have a rare code present should be more similar than if they did not, but sharing in the absence of rare codes should not have a large impact on similarity.

We developed kernel functions based on Gower (1971)’s work on the coefficient of similarity. The similarity between two individuals $i$ and $j$ based on character $c$ can be assigned the similarity score $S_{i,j,c}$ ranging from 0 (no similarity) to 1 (the same). An indicator $\delta_{i,j,c}$ is used to represent whether or not a comparison can be made. In the case of a binary variable, if one or both people have the variable present the indicator will be 1; if neither person has the variable present the indicator is 0 and $S_{i,j,c}$ will be set to 0. Gower (1971) further demonstrates that a weight can be introduced for each code $w_c$; if there are no missing values and all $w_c \geq 0$ the following allows for a positive semidefinite similarity matrix with entries $S_{i,j} = \sum_{c=1}^{v} S_{i,j,c,w_c}/\sum_{c=1}^{v} \delta_{i,j,c,w_c}$.

The commonly-used Jaccard (J) similarity is equivalent to setting $w_c = 1$ for all $c$. In the case where both people’s sets are empty, we set $S_{i,j}$ to 1. It meets the first of our desired properties and is the first candidate kernel function we use. The second property can be addressed by reverse coding common data elements and assigning weights such that only common codes are considered; the third property is addressed by maintaining traditional coding based on presence and setting weights such that only rare codes are considered. We use a cut-off based on prevalence in the training data to define common (prevalence $\geq 0.70$) and rare (prevalence $< 0.30$) codes such that codes above/below the threshold are assigned $w_c = 1$ and remaining codes are assigned $w_c = 0$. Our second candidate kernel function (SCR) adds together the “common absence” and “rare presence” similarity scores.
Hybrid Feature- and Similarity-Based Models

**Application 1: Prediction** We assess predictive performance using a similar nested CV procedure as was used for the simulation studies, with 80/20 splits to define inner training/validation data for each of five outer folds. The same four models of interest (LR, KLR, HFSM-Seq, HFSM-Sim) are compared alongside two additional models: LR-E, which is LR with the extra count of chronic conditions feature; and a more complex model (XGBoost) that includes all features and all kernel data input represented as dummy variables. Hyperparameters for kernel-containing models selected based on a grid search for the highest AUROC on the inner fold validation data include L1 penalty strength (0.0001, 0.001, or 0.01), kernel data inputs (providers involved, service types, or both), and kernel function (J or SCR).

**Application 2: Interpretation** To demonstrate model interpretability, we re-train HFSM-Seq and HFSM-Sim on all data using the Jaccard kernel function on both types of data, using the mode of the selected L1 penalty in Application 1 divided by five to scale for the increase in amount of data. We examine how feature coefficients change between the two models similar to in the illustrative examples. We then move our focus to HFSM-Seq to examine the type of information captured by the kernel after accounting for the features. We split the cohort into clients with positive, negative, and zero-valued kernel coefficients. Feature-based characteristics are compared with descriptive table-based summaries across the three strata. Kernel-based characteristics are explored by applying non-negative matrix factorization with five topics, using Python package sklearn.decomposition.NMF and the KL divergence distance metric to each of the three strata (Pedregosa et al., 2011).

**Results for Clinical Case Study**

There were 5,070 eligible clients with a 5.4% cumulative incidence (n=276) of the outcome across all client-specific two-year prediction intervals. See Appendix D for a cohort flow diagram and table of select baseline characteristics.

**Application 1: Prediction** Performance metrics are in Table 3. The SCR kernel function was selected three times for HFSM-Seq and KLR, and two times for HFSM-Sim; for all models the combined provider and service type data were selected the majority of the time (Appendix D).

|                | LR  | LR-E | KLR | HFSM-Seq | HFSM-Sim | XGBoost |
|----------------|-----|------|-----|----------|----------|---------|
| AUROC          | 0.753 | 0.754 | 0.734 | 0.774     | 0.778     | 0.727   |
| AUPRC          | 0.146 | 0.148 | 0.139 | 0.185     | 0.184     | 0.137   |
| Calibration Slope | 0.852 | 0.848 | 0.698 | 0.788     | 0.875     | 0.868   |
| Calibration Intercept | -0.367 | -0.378 | -0.788 | -0.521   | -0.294   | -0.621  |
| Time (minutes) | < 1 | < 1 | 42 | 115 | 89 | < 1 |

Results are averaged across the five outer folds.

General trends across predictive performance metrics from worst to best are KLR and XGBoost, LR, HFSM-Seq, and then HFSM-Sim. For discrimination and precision-recall...
performance, \textit{HFSM-Seq} and \textit{HFSM-Sim} were best. Calibration was best for \textit{HFSM-Sim}; all models tended to under-estimate risk. While there are some instances of a model having notably worse performance in terms of calibration, there are no instances of one model that had very large performance gains over all other models; discrimination performance is comparable across all models. LR and LR-E performed similarly on all metrics. Kernel containing models were the least efficient, even with pre-computed kernel matrices, but still ran within a feasible amount of time.

\textbf{Application 2: Interpretation} In general, \textit{HFSM-Seq} feature coefficients (Appendix D) are larger in magnitude and consistent in direction to those of \textit{HFSM-Sim}, suggesting that the kernel information adjusts for some of the feature relationships. We suspect that for some of the coefficients with the largest change in magnitude the kernel (capturing health care use information) is acting as a mediator (e.g., stable housing), and for others it is acting as a confounder (e.g., depression or anxiety). This or colliding bias could also explain features where associated coefficients tended to increase in magnitude when adjusted for the kernel (e.g., primary language) and features where there was a qualitative change (e.g., food insecurity; coefficient switched from positive to negative after adjusting for the kernel). Importantly, we did not set up this model to be causal and do not know which feature coefficients are closer to the truth; we would not want to deploy it in its current form for clinical decision making.

Examining \textit{HFSM-Seq}, $\alpha$ coefficients ranged from -1.70 to 1.30 and when rounded to five significant digits there were 5038 zero, 13 positive, and 19 negative. Feature and outcome values stratified across these three groups are in Appendix D. Distinct trends for clients with negative $\alpha$, and thus decrease the predicted probability for similar clients, are that none had the outcome, all lived in an urban geography, and they tended to have higher levels of obesity than the other strata. Clients with positive $\alpha$, and thus increase the predicted probability for similar clients, all had English as their primary language and tended to have lower household income and higher levels of stable housing, substance use, smoking or tobacco use, and food insecurity. The top ten weighted terms from NMF for the three subgroups are in Appendix D. The group with negative $\alpha$’s had a unique topic characterized by diagnosis and treatment with physician and nurse providers; a topic related to counselling and foot care with counsellor and chiropodists; and one related to counselling with nurse practitioners. The group with positive $\alpha$ had a unique topic strongly characterized by external referral and consult; and a topic strongly characterized by social worker, nurse practitioner, and individual counselling. Codes related to diagnosis, treatment, and management are not as prominent in topics as for the other groups. The zero $\alpha$ topics include one strongly characterized by community resources and community health workers, which only weakly enter topics for the other subgroups.

\textbf{5. Discussion}

The \textit{HFSM} approach captures relationships within large-scale observational data in an interpretable form when some but not all data and desired information to capture are suitable for simple feature representation. Simulation studies confirmed that \textit{HFSM} is best suited for situations where the feature- and kernel-based data are both important for the outcome, and our clinical case study demonstrated how it can be used to build a predictive
model and develop understanding of risk drivers within a complex primary health care population.

### 5.1. Hybrid Model Methodology

The predictive performance of \textit{HFSM-Sim} is always expected to be as good or better than \textit{HFSM-Seq}, assuming appropriate set up and tuning, while \textit{HFSM-Seq} provides more certainty in feature coefficient interpretation when the role of the kernel in terms of causal structures is uncertain. If the goal is to prioritize absolute risk predictions, \textit{HFSM-Sim} is recommended; however, if the model is intended to support decision making with interpretation of feature coefficients, then greater care is needed. Feature coefficients learned under \textit{HFSM-Seq} as shown in this paper are adjusted for each other and averaged over the kernel, so can be interpreted analogously as for \textit{LR}. An additional option is to fix some or all of the feature coefficients based on previous research studies or epidemiological analyses, and learn the rest from the training data. The feature coefficients for \textit{HFSM-Sim} are adjusted for each other and adjusted for the kernel matrix, which may be favourable for interpretation in situations such as when the kernel is constructed to adjust for complex confounders. The degree to which the kernel and feature matrices are independent will determine the difference between \textit{HFSM-Seq} and \textit{HFSM-Seq}.

The Illustrative Examples demonstrating feature coefficient changes depending on the “causal role” of the kernel reinforce the importance of careful modelling not only for explicit causal inference but also when a risk prediction model might be interpreted as an “upstream” treatment effect model. For example, if feature coefficients are interpreted as identifying modifiable risk factors (e.g., hypertension, smoking) important for a client’s estimated outcome, and then inform risk prevention strategy selection. Formal techniques for multiple causal inference, such as the deconfounder approach informed by a directed acyclic graph based on clinical and epidemiological input, may be useful here (Wang et al., 2018; Wang and Blei, 2019; Greenland et al., 1999). Future work is needed to determine what “pragmatic” level of causality is sufficient to support decisions in these settings. Future work could explore interactions between features and the kernel, other outcomes types (e.g., time to event), an L2 penalty, and multilevel modelling.

### 5.2. Clinical Case Study

Predicting the rare outcome of social isolation or loneliness in middle-aged clients served by UAR CHCs is a challenging supervised learning problem. The \textit{HFSM} models performed as well or better than solely feature- and similarity-based models, including \textit{XGboost}, while providing superior interpretability. Our proposed kernel function that calculates similarity based on the absence of common codes and the presence of rare codes was selected more often than Jaccard, which considers presence of all codes. We used a basic cut-off to define rare and common codes, but future work could expand this to obtain more sophisticated weightings. For example, Belanche and Villegas (2013) define the similarity score for a present code based on the probability of that code, which may also be useful when applied to the inverse of common codes.

We explored model behaviour and learned about the cohort by viewing feature values and NMF-derived topics on kernel data stratified by positive, negative, or zero $\alpha$. Of note,
the kernel data included provider types; some, e.g., social work, may have a higher index of suspicion for the outcome and care for correlated conditions, but these data are only from baseline and all providers can code the outcome. We could extend this work into the causal setting and intentionally set up the model to use the features and kernel to capture relationships between some or all of the features and the outcome. Future work could compare coefficients from HFSM that e.g., uses to kernel to adjust for complex potential confounders, against other approaches or known effect estimates.

In addition to insights about HFSM and the kernel, our clinical case study demonstrates insights relevant to future primary health care machine learning applications. In contrast to settings where care is initiated due to a problem (e.g., cancer diagnosis, emergency room visit), primary health care is sought out during all stages of health, there is variability in visit patterns, and risk patterns change across across the care history due to cumulative and acute factors (World Health Organization, 2021b; Kueper et al., 2022). Outcomes such as ours are relevant across the entire care trajectory, which induces challenges for determining a prediction interval as “lifetime risk” is unhelpful. Future research is needed on the best way to define prediction intervals for these types of outcomes in primary health care; we selected two-year prediction interval periods to support generalizability of the resulting model across the care continuum within a time frame that allows preventative intervention, but the most appropriate choice will depend on context.

A significant challenge we encountered, which is relevant to other health sectors as well, is rare features. Three features representing characteristics that literature suggests are risk factors for social isolation and loneliness had less than 1% prevalence in the baseline cohort data. The standardized morbidity ratios (SMR), representing the ratio of observed to expected number of outcome cases based on sex-specific rates in the remaining eligible population, showed higher than expected risk in each of the rare feature sub-populations (Sensory Disability SMR = 2.44; Social Phobia SMR = 3.27; Dementia or Alzheimer’s Disease SMR = 2.11). The number of clients with these characteristics is too small to meaningfully do statistics with, and we could not find an explainable artificial intelligence framework that addresses this type of scenario. If the model were implemented, it would be important to communicate to care providers that these risk factors are not considered by the model, such as with a flag when making predictions for a client with one of the characteristics present. Qualitative research could investigate risks and needs of these subpopulations specific to the health care setting of implementation.

Limitations We have not provided confidence intervals or hypothesis tests particularly in the case of HFSM-Sim because although the objective function is convex, the non-smoothness of the L1 penalty is expected to require use of techniques like the m-out-of-n bootstrap (Bickel and Sakov, 2008) or potentially a selective inference framework (Chen et al., 2021) to account for non-regularity in the estimators. Developing these is beyond the scope of this work.

In our clinical case study, we restrict based on age and CHCs within the UAR peer group; however, remaining variability within these strata were not taken into account. Feature construction included use of the client characteristic table, which includes rich sociodemographic information but is not time stamped in our data extraction. Outcome recording was not blinded and can only be considered a proxy for “true” social isolation and
loneliness. The majority of clients excluded for having less than three years of observation had their first event in 2017 or later so there was not enough calendar time for sufficient observation. There were 1,430 clients with a first event early enough and who met other eligibility criteria. If we were to proceed with this model we would perform sensitivity analyses to assess whether there is bias due to this “true” loss to follow up.

6. Conclusion

We presented a hybrid feature- and similarity-based model that combines well-established approaches (logistic regression and kernel regression) into a single machine learning model. The hybrid approach provides a way to take advantage of large-scale datasets information about features where the relationship with the outcome can be specified in a linear model (e.g., known informative risk factors or structured one-time question fields), as well as more complex data that may be better captured in terms of similarity to other training examples (e.g., historical data on care and diagnoses received). Maintaining separation of feature and similarity based components supports interpretability of the final model, and the option to fix or learn feature coefficients in advance of the kernel coefficients provides additional control over feature coefficient interpretation. This inherent model interpretability and the reproducibility due to a fully convex objective functions supports the extension of model use from prediction to causal inference tasks both within health care and in other domains with complex data and causal structures.

Acknowledgements

This study was supported by a Canadian Institutes of Health Research Canadain Graduate Student - Doctoral Award. CIHR was not involved in the conduct of the research.

References

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1): 42–60, 2018. ISSN 2330-7706. doi: 10.1080/23307706.2017.1397554. URL https://doi.org/10.1080/23307706.2017.1397554. Publisher: Taylor & Francis eprint: https://doi.org/10.1080/23307706.2017.1397554.

Dennise Albrecht. Community health centres in Canada. *Leadership in Health Services*, 11(1):5–10, January 1998. ISSN 1366-0756. doi: 10.1108/1366075810202596. URL https://doi.org/10.1108/1366075810202596. Publisher: MCB UP Ltd.

Alliance for Healthier Communities. Alliance for Healthier Communities, 2020. URL https://www.allianceon.org/.

Drew Altschul, Matthew Iveson, and Ian J. Deary. Generational differences in loneliness and its psychological and sociodemographic predictors: an exploratory and confirmatory machine learning study. *Psychological Medicine*, 51(6):991–1000, 2021. ISSN 0033-2917. doi: 10.1017/S0033291719003933. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8161432/.
Peter C. Austin and Jack V. Tu. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology*, 57(11):1138–1146, November 2004. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2004.04.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0895435604001118.

Lluís A Belanche and Marco A Villegas. Kernel functions for categorical variables with application to problems in the life sciences. In *Proceedings of the 16th International Conference of the Catalan Association for Artificial Intelligence*, volume 256, pages 171–180, Catalonia, Spain, 2013. IOS Press. doi: https://doi.org/10.3233/978-1-61499-320-9-171.

Peter J. Bickel and Anat Sakov. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, 18(3):967–985, 2008. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24308525.

Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.

M. Alan Brookhart, Til Stürmer, Robert J. Glynn, Jeremy Rassen, and Sebastian Schneeweiss. Confounding control in healthcare database research: challenges and potential approaches. *Medical Care*, 48(60):S114–S120, 2010. ISSN 0025-7079. doi: 10.1097/MLR.0b013e3181dbebe3. URL http://journals.lww.com/00005650-201006001-00018.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL https://dl.acm.org/doi/10.1145/2939672.2939785.

Yiqun T. Chen, Sean W. Jewell, and Daniela M. Witten. More powerful selective inference for the graph fused lasso, 2021. URL https://arxiv.org/abs/2109.10451.

Bryan Conroy, Minnan Xu-Wilson, and Asif Rahman. Patient similarity using population statistics and multiple kernel learning. In *Proceedings of Machine Learning for Healthcare 2017*, volume 68. JMLR, 2017.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of machine learning research*, 17:1–5, 2016. ISSN 1532-4435. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4927437/.

Afsaneh Doryab, Daniella K Villalba, Prerna Chikersal, Janine M Dutcher, Michael Tuminia, Xinwen Liu, Sheldon Cohen, Kasey Creswell, Jennifer Mankoff, John D Creswell, and Anind K Dey. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: Statistical analysis, data mining and machine learning of smartphone and fitbit data. *JMIR mHealth and uHealth*, 7(7):e13209, July 2019. ISSN 2291-5222. doi: 10.2196/13209. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6685126/.
ENCODE-FM. Electronic Nomenclature and Classification Of Disorders and Encounters for Family Medicine, 2020. URL http://aix1.uottawa.ca/~fammed/fmcencod.htm.

Xiaoping Fan, Zhijie Chen, Liangkun Zhu, Zhifang Liao, and Bencai Fu. A novel hybrid similarity calculation model. Scientific Programming, 2017:e4379141, December 2017. ISSN 1058-9244. doi: 10.1155/2017/4379141. URL https://www.hindawi.com/journals/sp/2017/4379141/. Publisher: Hindawi.

Hao Sen Andrew Fang, Ngiap Chuan Tan, Wei Ying Tan, Ronald Wihal Oei, Mong Li Lee, and Wynne Hsu. Patient similarity analytics for explainable clinical risk prediction. BMC Medical Informatics and Decision Making, 21(1):207, July 2021. ISSN 1472-6947. doi: 10.1186/s12911-021-01566-y. URL https://doi.org/10.1186/s12911-021-01566-y.

Martin Fortin, José Almirall, and Kathryn Nicholson. Development of a research tool to document self-reported chronic conditions in primary care. Journal of Comorbidity, 7(1):117–123, January 2017. ISSN 2235-042X. doi: 10.15256/joc.2017.7.122. URL https://doi.org/10.15256/joc.2017.7.122.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, Irene Y. Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. AMIA Summits on Translational Science Proceedings, 2020:191–200, May 2020. ISSN 2153-4063. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233077/.

Mehmet Gonen, Ethem Alpaydın, Boun Edu Tr, and Boun Edu Tr. Multiple kernel learning algorithms. Journal of Machine Learning Research, 12:2211–2268, 2011.

J. C. Gower. A general coefficient of similarity and some of its properties. Biometrics, 27(4):857–871, 1971. ISSN 0006-341X. doi: 10.2307/2528823. URL https://www.jstor.org/stable/2528823. Publisher: [Wiley, International Biometric Society].

Sander Greenland and Hal Morgenstern. Confounding in health research. Annual Review of Public Health, 22(1):189–212, May 2001. ISSN 0163-7525, 1545-2093. doi: 10.1146/annurev.publhealth.22.1.189. URL http://www.annualreviews.org/doi/10.1146/annurev.publhealth.22.1.189.

Sander Greenland, Judea Pearl, and James M. Robins. Causal diagrams for epidemiologic research. Epidemiology, 10(1):37–48, January 1999. ISSN 1044-3983. doi: 10.1097/00001648-199901000-00008. URL https://insights.ovid.com/crossref?an=00001648-199901000-00008.

Yuqi Guo, Zhichao Hao, Shichong Zhao, Jiaqi Gong, and Fan Yang. Artificial intelligence in health care: Bibliometric analysis. Journal of Medical Internet Research, 22(7):e18228, July 2020. ISSN 1438-8871. doi: 10.2196/18228. URL http://www.jmir.org/2020/7/e18228/.

Julianne Holt-Lunstad. The potential public health relevance of social isolation and loneliness: Prevalence, epidemiology, and risk factors. Public Policy & Aging Report,
Carlos García Márquez. *Multivariate Kernel Functions for Categorical Variables*. PhD thesis, Universitat Politècnica de Catalunya, Barcelona, 2014.

National Academies of Sciences, Engineering, and Medicine. Social Isolation and loneliness in older adults: Opportunities for the health care system. In *Risk and Protective Factors for Social Isolation and Loneliness*. The National Academies Press, Washington, DC, USA, February 2020. URL https://doi.org/10.17226/25663. Publication Title: Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System.

Nicholas R. Nicholson. A review of social isolation: An important but underassessed condition in older adults. *The Journal of Primary Prevention*, 33(2):137–152, June 2012. ISSN 1573-6547. doi: 10.1007/s10935-012-0271-2. URL https://doi.org/10.1007/s10935-012-0271-2.

Dominik Alex Nowak and Kate Mulligan. Social prescribing. *Canadian Family Physician*, 67(2):88–91, February 2021. ISSN 0008-350X. doi: 10.46747/cfp.670288. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8324130/.

Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiwen Liu, and Zhe He. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27(7):1173–1185, July 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa053. URL https://doi.org/10.1093/jamia/ocaa053.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Stuart Russell and Peter Norvig. *Artificial Intelligence A Modern Approach*. Pearson Education Inc., 3 edition, 2010. ISBN 978-0-13-604259-4.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2011. ISBN 978-0-511-80968-2.

Ian Shrier and Robert W Platt. Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, 8(1):70, December 2008. ISSN 1471-2288. doi: 10.1186/1471-2288-8-70. URL https://bmcmethodol.biomedcentral.com/articles/10.1186/1471-2288-8-70.

Fabio Sigrist. KTBoost: Combined kernel and tree boosting. *Neural Processing Letters*, 53(2):1147–1160, April 2021. ISSN 1370-4621, 1573-773X. doi: 10.1007/s11063-021-10434-9. URL https://link.springer.com/10.1007/s11063-021-10434-9.
S. B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. De Jong, S. Dzeroski, S. E. Fahlman, D. Fisher, R. Hamann, K. Kaufman, S. Keller, I. Kononenko, J. Kreuziger, R. S. Michalski, T. Mitchell, P. Pachowicz, Y. Reich, H. Vafaie, W. Van De Welde, W. Wenzel, J. Wnek, and J. Zhang. The MONK’s Problems a performance comparison of different learning algorithms. Technical report, Carnegie Mellon University, 1991.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In Proceedings of Machine Learning Research, page 21, 2019.

Yixin Wang and David M. Blei. The blessings of multiple causes. Journal of the American Statistical Association, 114(528):1574–1596, 2019. URL http://arxiv.org/abs/1805.06826. arXiv: 1805.06826.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M. Blei. The Deconfounded Recommender: A causal inference approach to recommendation. arXiv:1808.06581 [cs, stat], August 2018. URL http://arxiv.org/abs/1808.06581. arXiv: 1808.06581.

World Health Organization. ICD-10 Version:2019, 2020. URL https://icd.who.int/browse10/2019/en.

World Health Organization. Social isolation and loneliness among older people: advocacy brief. World Health Organization, Geneva, 2021a. ISBN 978-92-4-003074-9. URL https://apps.who.int/iris/handle/10665/343206.

World Health Organization. Primary health care, 2021b. URL https://www.who.int/westernpacific/health-topics/primary-health-care.

Ji Zhu and Trevor Hastie. Kernel logistic regression and the import vector machine. Journal of Computational and Graphical Statistics, 14(1):185–205, 2005. ISSN 10618600. URL http://www.jstor.org/stable/27594104.
Appendix A. Additional Background and Example

This appendix provides further background on feature- and similarity-based models, and provides a simple illustrative example of how HFSM combines the two.

A.1. Background

Feature-based supervised models, such as trees, learn a mathematical function that takes features as input and provides an estimate about the outcome as output (Bishop, 2006; Russell and Norvig, 2010). To use a feature-based approach for the above scenario, the \( \Psi(o) \)-type information must be converted into \( \phi(o) \)-type information. This can be done in a data-driven and/or theoretical\(^1\) way. For example, a data-driven dimensionality reduction approach such as topic modeling could be applied to client histories of diagnostic codes and the resulting topic weights for each client used as features. A theoretical approach may include identifying known risk-factors for the outcome based on research literature or clinical expertise, and then collapsing specific subsets of codes to generate binary indicators for whether or not the client has ever experienced each risk factor. Oftentimes feature construction loses or misses information and in general the extent to which this is a disadvantage will depend on the complexity of the data and predictive task.

Similarity-based approaches, such as nearest-neighbour and kernel methods, can handle both \( \phi(o) \)-type and \( \Psi(o) \)-type information as inputs (Bishop, 2006; Russell and Norvig, 2010). Instead of learning explicit relationships between individual inputs and the outcome, these methods use similarities or distances between observations by assuming that similar \( o \) are likely to experience the same outcome, and if the same \( o \) is entered into the model twice the same prediction will result. There are potentially two challenging aspects of \( \Psi(o) \) that make similarity-based approaches attractive: the dimensionality and the proper form. In some situations, the best form for \( \Psi(o) \) is known but its dimension is too large or challenging to construct with traditional feature-based approaches; other times, even if \( \Psi(o) \) is a manageable size, the proper way to incorporate it into a model is unknown.

Kernels can handle both of these challenges and are the similarity-based approach used in the remainder of the paper. A kernel function \( k : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R} \) expresses the inner product between two inputs that have been implicitly mapped to some high-dimensional feature space (Bishop (2006); Russell and Norvig (2010)). The feature mapping defines the notion of similarity captured by the scalar output, and can be non-linear with infinite dimensions; the mapping does not need to be made explicit to use a kernel function. The notion(s) of similarity to capture, and whether any data pre-processing is warranted, will depend on the specific scenario. The simplest kernel function is the linear kernel, \( k(o_i, o_j) = o_i^T o_j \), which can be used to create a dual formulation of linear regression that uses \( o \) directly as features. An example of a more complex kernel function is the Gaussian or Radial Basis Function (RBF), \( k(o_i, o_j) = e^{-\frac{|o_i - o_j|^2}{2\sigma^2}} \), which has a feature space with infinite dimensions and one hyperparameter \( \sigma \) to be tuned (Bishop, 2006). The RBF kernel was used in our simulation study experiments; additional kernel functions were introduced and proposed in the clinical case study section.

\(^1\) Here, we are referring to medical or social theories of health rather than statistical or computer science theory.
A.2. Illustrative Example: Hybrid Feature- and Similarity-Based Model

The feature- and similarity-based parts of the model can represent primal and dual forms, respectively (Bishop, 2006). Thus, a model where some features are combined using an unpenalized linear kernel will be equivalent to a model where all features are entered in logistic regression. To demonstrate this, 10,000 observations were generated according to the data generating mechanism $P(Y) = \sigma(0.25 - 1X_1 + 2X_2)$ where $X_{1,2} \sim N(0,1)$. Logistic regression was fit with an intercept, $X_1$, and $X_2$; HFSM was fit with the intercept and $X_1$ maintained as features and $X_2$ included with a linear kernel.

Table 4: Illustrative Example Feature Coefficients

|       | LR  | HFSM |
|-------|-----|------|
| $\beta_0$ | 0.24 | 0.24 |
| $\beta_1$ | -1.04 | -1.04 |
| $\beta_2$ | 2.04 | NA   |

Table 4 shows the learned coefficients whereby the intercept and $X_1$ coefficients are the same for the two models. The unpenalized $\alpha$’s from HFSM range from -0.001 to 0.001. As expected, predictions based on the two model forms are also equivalent (not shown).
Appendix B. Model Code

The following Python code can be used to fit the four main models we used: logistic regression (M1), kernel logistic regression (M2), hybrid model sequential fit (M3), and hybrid model simultaneous fit (M4).

```python
import numpy as np
import cvxpy as cp
from sklearn.metrics import roc_auc_score
from scipy.special import expit

# Function to fit feature only model
# @param Xtrain the training feature data
# @param yTrain training binary outcome
# @return betas and auc on training data
# Note that cp.logistic(x) is log(1 + exp(x)), not sigmoid

def FIT_M1(Xtrain, yTrain, save=False, fnHead=None):
    beta = cp.Variable((Xtrain.shape[1], 1))
    problemM1 = cp.Problem(cp.Maximize(cp.
        sum(
            cp.multiply(yTrain, (Xtrain @ beta))
        - cp.logistic((Xtrain @ beta)))/Xtrain.shape[0]))
    problemM1.solve(verbos=False, solver=cp.ECOS)
    # Get training AUC value
    aucTrain = roc_auc_score(yTrain, expit(Xtrain @ beta.value))
    print(f"\n***DONE M1 FIT***
Status of M1 problem: {problemM1.status}
and Optimal value: {problemM1.value}
and solve time: {problemM1._solve_time}
**Training AUC: {aucTrain}
"
    if save:
        np.save(fnHead + ".Betas.npy", beta.value)
        np.save(fnHead + ".SolveTime.npy", problemM1._solve_time)
        np.save(fnHead + ".OptValue.npy", problemM1.value)
        np.save(fnHead + ".Status.npy", problemM1.status)
        np.save(fnHead + ".aucTrain.npy", aucTrain)
    return beta.value, aucTrain

# Function to fit kernel only model with L1 penalty
# @param Ktrain precomputed training kernel
# @param yTrain training outcome
# @param l1 strength of L1 penalty for alphas
# @param fnHead start path to save object
# including directory and foldO
# @return alphas, auc on training data

def FIT_M2(Ktrain, yTrain, l1, save=False, fnHead=None):
    alpha = cp.Variable((Ktrain.shape[1], 1))
    lam = cp.Parameter(nonneg=True, value=l1)
```

---

24
problemM2 = cp.Problem(cp.Maximize(cp.sum(
    cp.multiply(yTrain, (Ktrain @ alpha))
    - cp.logistic(Ktrain @ alpha))/Ktrain.shape[0]
    - lam * cp.norm(alpha, 1))))

problemM2.solve(verbose=False, solver=cp.ECOS)

# Get training AUC value
aucTrain = roc_auc_score(yTrain, expit(Ktrain @ alpha.value))

print(f"***DONE_M2_FIT_WITH_LAM_{l1}***")
print(f"Status of M2 problem: {problemM2.status}"
      f"and optimal value: {problemM2.value}"
      f"and solve time: {problemM2._solve_time}"
      f"Training AUC: {aucTrain}"
)

if (save):
    np.save(fnHead + "_Alphas.npy", alpha.value)
    np.save(fnHead + "_SolveTime.npy", problemM2._solve_time)
    np.save(fnHead + "_OptimalValue.npy", problemM2.value)
    np.save(fnHead + "_Status.npy", problemM2.status)
    np.save(fnHead + "_aucTrain.npy", aucTrain)

return alpha.value, aucTrain

# Function to fit HFSM-Seq with L1 penalty
# Betas are fit first and fixed while learning alphas
# @param Xtrain precomputed training kernel
# @param Ktrain precomputed training kernel
# @param yTrain training outcome
# @param l1 strength of L1 penalty for alphas
# @param fnHead start path to save object
# including directory and foldO
# @return betas, alphas, auc on training data

def FIT_M3(Xtrain, Ktrain, yTrain, l1, fixedBeta=None, save=False, fnHead=None):
    if (fixedBeta==None):
        # learn the betas ignoring alphas
        fixedBeta, aucNotUsed = FIT_M1(Xtrain, yTrain, save=True, fnHead=fnHead + "_m1Part")

        # betas are set up as fixed parameter for learning alphas
        betaM1 = cp.Parameter(fixedBeta.shape, value=fixedBeta)

        # Alphas are learned
        alpha = cp.Variable((Ktrain.shape[0], 1))

        # L1 penalty strength is fixed parameter
        lam = cp.Parameter(nonneg=True, value=l1)

        # problem to solve
        problemM3 = cp.Problem(cp.Maximize(cp.sum(
            cp.multiply(yTrain, (Ktrain @ alpha + Xtrain @ betaM1))
            - cp.logistic(Ktrain @ alpha + Xtrain @ betaM1))
            / Ktrain.shape[0]
            - lam * cp.norm(alpha, 1))))

        # call the solver; default max iters is 10,000
        problemM3.solve(verbose=False, warm_start=True, solver=cp.ECOS)

        aucTrain = roc_auc_score(yTrain, expit(Xtrain @ betaM1.value + Ktrain @ alpha.value))

        print(f"***DONE_M3_FIT_WITH_LAM_{l1}***")

        if (save):
            np.save(fnHead + "_Alphas.npy", alpha.value)
            np.save(fnHead + "_SolveTime.npy", problemM3._solve_time)
            np.save(fnHead + "_OptimalValue.npy", problemM3.value)
            np.save(fnHead + "_Status.npy", problemM3.status)
            np.save(fnHead + "_aucTrain.npy", aucTrain)

        return alpha.value, aucTrain
Hybrid Feature- and Similarity-Based Models

f"\nStatus of problem: {problemM3.status}"
f"and optimal value: {problemM3.value}"
f"and solve time: {problemM3._solve_time}"
f"\n**Training AUC: {aucTrain}"}

if (save):
    np.save(fnHead + "_BetasM1.npy", betaM1.value)
    np.save(fnHead + "_Alphas.npy", alpha.value)
    np.save(fnHead + "_SolveTime.npy", problemM3._solve_time)
    np.save(fnHead + "_OptimalValue.npy", problemM3.value)
    np.save(fnHead + "_Status.npy", problemM3.status)
    np.save(fnHead + "_aucTrain.npy", aucTrain)

return betaM1.value, alpha.value, aucTrain

# Function to fit HFSM-Sim with L1 penalty
# @param Xtrain precomputed training kernel
# @param Ktrain precomputed training kernel
# @param yTrain training outcome
# @param l1 strength of L1 penalty for alphas
# @param fnHead start path to save object
# including directory and foldO
# @return betas, alphas, auc on training data data

def FIT_M4(Xtrain, Ktrain, yTrain, l1, save=False, fnHead=None):
    # Variables can be scalars, vectors, or matrices
    beta = cp.Variable((Xtrain.shape[1], 1))
    # vector of values (n,1) to fit
    alpha = cp.Variable((Ktrain.shape[0], 1))
    # Parameter - this one is positive scalar for lam
    lam = cp.Parameter(nonneg=True, value=l1)
    # problem to solve
    problemM4 = cp.Problem(cp.Maximize(cp.sum(
        cp.multiply(yTrain, (Ktrain @ alpha + Xtrain @ beta))
        - cp.logistic(Ktrain @ alpha + Xtrain @ beta))
        / Ktrain.shape[0]
        - lam * cp.norm(alpha, 1)))
    # call the solver; default max iters is 10,000
    problemM4.solve(verbosity=False, warm_start=True, solver=cp.ECOS)
    aucTrain = roc_auc_score(yTrain,
        expit(Xtrain @ beta.value + Ktrain @ alpha.value))
    print(f"\n*** DONE M4 FIT WITH LAM {l1} ***"
        f"\nStatus of problem: {problemM4.status}"
        f"and optimal value: {problemM4.value}"
        f"and solve time: {problemM4._solve_time}"
        f"\n**Training AUC: {aucTrain}"}

if (save):
    np.save(fnHead + "_Betas.npy", beta.value)
    np.save(fnHead + "_Alphas.npy", alpha.value)
    np.save(fnHead + "_SolveTime.npy", problemM4._solve_time)
    np.save(fnHead + "_OptimalValue.npy", problemM4.value)
    np.save(fnHead + "_Status.npy", problemM4.status)
    np.save(fnHead + "_aucTrain.npy", aucTrain)

return beta.value, alpha.value, aucTrain
Appendix C. Simulation Study Details

C.1. RBF Kernel Sigma Selection

Three candidate hyperparameter values for the RBF kernel were selected to provide a range of diagonal dominance as assessed by the following equation:

\[ DD = \sum \frac{|\text{diagonals}|}{|\text{off} - \text{diagonals}|} \]

For a matrix with \( i, j = n \) observations:

\[ DD = \sum_{i,j} \frac{|a_{i,i}|}{|a_{i,j}| - |a_{i,i}|} \]

For a kernel matrix \( a_{i,i} = 1 \) and the range will be \( \left[ \frac{n}{(n-1)}, \infty \right) \).

In addition to looking at the above scalar measure, heat plots for RBF kernels generated with a range of \( \sigma \) values on a random sample of 1000 observations of the six variables in MONK’s data problems. Example plots in Figure 2 provide another view at how varying the \( \sigma \) values alters the similarity captured by the RBF kernel.

![Figure 2: Heatmap demonstrating similarity of RBF kernel with various \( \sigma \) on a random sample.](image-url)

\( \sigma = 0.01 \) \hspace{1cm} \( \sigma = 0.1 \) \hspace{1cm} \( \sigma = 1 \)
### Table 5: Synthetic Data Scenario 1: Selected Hyperparameters

|                     | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|---------------------|--------|--------|--------|--------|--------|
| **Sigma for RBF Kernel** |        |        |        |        |        |
| LR                  | NA     | NA     | NA     | NA     | NA     |
| KLR                 | 1.000  | 1.000  | 1.000  | 1.000  | 1.000  |
| HFSM-Seq            | 0.100  | 1.000  | 0.100  | 0.100  | 1.000  |
| HFSM-Sim            | 0.100  | 1.000  | 0.100  | 0.100  | 0.100  |
| **L1 Penalty Strength** |        |        |        |        |        |
| LR                  | NA     | NA     | NA     | NA     | NA     |
| KLR                 | 0.032  | 0.001  | 0.001  | 1.000  | 1.000  |
| HFSM-Seq            | 0.001  | 0.001  | 0.001  | 0.001  | 0.001  |
| HFSM-Sim            | 0.001  | 0.001  | 0.001  | 0.001  | 0.001  |

### Table 6: Synthetic Data Scenario 1: Model Interpretation

**a** Average Feature Coefficients

|        | LR     | HFSM-Sim | True |
|--------|--------|----------|------|
| $\beta_0$ | -1.263 | -1.021   | -1.500 |
| $\beta_1$ | 0.320  | 0.319    | 0.300 |
| $\beta_2$ | 0.440  | 0.444    | 0.400 |
| $\beta_3$ | 0.568  | 0.567    | 0.600 |
| $\beta_4$ | 0.695  | 0.695    | 0.700 |

**b** Average Kernel Coefficients

|        | KLR     | HFSM-Seq | HFSM-Sim |
|--------|---------|----------|----------|
| Non-0  | 262.600 | 197.400  | 180.400  |
| Max    | 0.000   | 0.014    | 0.005    |
| Min    | -0.019  | -0.014   | -0.019   |
| Mean   | -0.004  | 0.000    | -0.005   |
| Median | -0.003  | -0.002   | -0.006   |

### Table 7: Synthetic Data Scenario 2: Selected Hyperparameters

|                     | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|---------------------|--------|--------|--------|--------|--------|
| **Sigma for RBF Kernel** |        |        |        |        |        |
| LR                  | NA     | NA     | NA     | NA     | NA     |
| KLR                 | 1.000  | 1.000  | 1.000  | 1.000  | 1.000  |
| HFSM-Seq            | 1.000  | 1.000  | 1.000  | 1.000  | 1.000  |
| HFSM-Sim            | 1.000  | 1.000  | 1.000  | 1.000  | 1.000  |
| **L1 Penalty Strength** |        |        |        |        |        |
| LR                  | NA     | NA     | NA     | NA     | NA     |
| KLR                 | 0.001  | 0.001  | 0.001  | 0.001  | 0.001  |
| HFSM-Seq            | 0.001  | 0.001  | 0.001  | 0.001  | 0.001  |
| HFSM-Sim            | 0.001  | 0.001  | 0.001  | 0.001  | 0.001  |
Table 8: Synthetic Data Scenario 2: Model Interpretation

(a) Average Feature Coefficients

|       | LR   | HFSM-Sim | True |
|-------|------|----------|------|
| $\beta_0$ | -0.844 | -0.557 | -2.100 |
| $\beta_1$ | 0.185  | 0.198   | 0.300  |
| $\beta_2$ | 0.377  | 0.408   | 0.400  |
| $\beta_3$ | 0.443  | 0.457   | 0.600  |
| $\beta_4$ | 0.566  | 0.605   | 0.700  |

(b) Average Kernel Coefficients

|       | KLR  | HFSM-Seq | HFSM-Sim |
|-------|------|----------|----------|
| Non-0 | 1696.200 | 1732.800 | 1631.000 |
| Max   | 0.082 | 0.085   | 0.082    |
| Min   | -0.098 | -0.096  | -0.109   |
| Mean  | -0.002 | -0.001  | -0.010   |
| Median| 0.001  | 0.001   | -0.006   |

Table 9: Synthetic Data Scenario 3: Selected Hyperparameters

| Fold   | Sigma for RBF Kernel | L1 Penalty Strength |
|--------|-----------------------|---------------------|
|        | LR                   | KLR                |                |
|        | NA                   | NA                 |                |
| Fold 1 |                      |                    |                |
| Fold 2 | LR                   | KLR                |                |
|        | NA                   | 1.000              | 1.000          |
| Fold 3 |                      | HFSM-Seq           |                |
|        |                     | 1.000              | 1.000          |
| Fold 4 |                      | HFSM-Sim           |                |
|        |                     | 1.000              | 1.000          |
| Fold 5 |                      |                    |                |

Table 10: Synthetic Data Scenario 3: Model Interpretation

(a) Average Feature Coefficients

|       | LR   | HFSM-Sim | True |
|-------|------|----------|------|
| $\beta_0$ | -0.459 | 0.414   | -3.200 |
| $\beta_1$ | 0.150  | 0.160   | 0.300  |
| $\beta_2$ | 0.163  | 0.203   | 0.400  |
| $\beta_3$ | 0.235  | 0.325   | 0.600  |
| $\beta_4$ | 0.280  | 0.385   | 0.700  |

(b) Average Kernel Coefficients

|       | KLR  | HFSM-Seq | HFSM-Sim |
|-------|------|----------|----------|
| Non-0 | 2575.800 | 2599.600 | 2546.400 |
| Max   | 0.094 | 0.102   | 0.092    |
| Min   | -0.264 | -0.269  | -0.297   |
| Mean  | -0.004 | -0.003  | -0.021   |
| Median| 0.006  | 0.006   | -0.013   |
Appendix D. Clinical Case Study Details

Cohort details

Figure 3: Clinical case study cohort flow diagram
Loss to Follow Up Of the 10,687 people excluded for having less than three years between their first and last care records, 6,276 (58.7%) had their first event in 2017 or later so there was not enough calendar time for sufficient observation; bias due to their exclusions is expected to be minimal. The remaining 4,411 (41.3%) are “true” loss to follow-up under a more traditional research study paradigm; we do not know if they stopped receiving care altogether or if they switched to another health care organization. After applying additional eligibility criteria there are 1,430 clients and among them there are 108 cases of the outcome of which 22 (16.9%) occur at least one year from the first recorded event. If our study was more application than methods testing focused we would perform sensitivity analyses to assess whether there is bias due to these lost to follow up as in a real world setting the future length of care when applying a predictive model is unknown.

Table 11: Clinical case study baseline features

| Feature                | Values                          | n (%)       |
|------------------------|---------------------------------|-------------|
| Sex                    | Female                          | 2379 (46.92)|
|                        | Male                            | 2691 (53.08)|
| Rural Residence        | Rural                           | 1011 (19.94)|
|                        | Urban                           | 3942 (77.75)|
|                        | Missing                         | 117 (2.31)  |
| Household Income       | $0 to $14,999                   | 1254 (24.73)|
|                        | $15,000 to $24,999              | 454 (8.95)  |
|                        | $25,000 to $34,999              | 274 (5.40)  |
|                        | $35,000 to $59,000              | 535 (10.55) |
|                        | $60,000 or more                 | 600 (11.83) |
|                        | Do not know                     | 274 (5.40)  |
|                        | Prefer not to answer            | 587 (11.58) |
|                        | Missing                         | 1092 (21.54)|
| Household Composition  | Couple                          | 1897 (37.42)|
|                        | Other Family                    | 519 (10.24) |
|                        | Unrelated housemates            | 217 (4.28)  |
|                        | Sole Member                     | 1205 (23.77)|
|                        | DoNotKnowOrOther                | 255 (5.03)  |
|                        | Prefer not to answer            | 57 (1.12)   |
|                        | Missing                         | 920 (18.15) |
| Education Completed    | Post-secondary or equivalent    | 1717 (33.87)|
|                        | Secondary or equivalent         | 1849 (36.47)|
|                        | Less than high school           | 395 (7.79)  |
|                        | DoNotKnowOrOther                | 269 (5.31)  |
|                        | Prefer not to answer            | 54 (1.07)   |
|                        | Missing                         | 786 (15.50) |
| Language               | English                         | 4691 (92.52)|
|                        | French                          | 82 (1.62)   |
|                        | Other                           | 297 (5.86)  |
| LGBTQ                  | Lgbtq                           | 67 (1.32)   |
|                        | Non-Lgbtq                       | 1084 (21.38)|
|                        | Missing                         | 3919 (77.30)|
Hybrid Feature- and Similarity-Based Models

| Condition                        | True | Count (Percentage) |
|----------------------------------|------|--------------------|
| Immigrated                       | True | 627 (12.37)        |
| Physical Disability              | True | 240 (4.73)         |
| Depression or Anxiety            | True | 410 (8.09)         |
| Chronic Urinary Problem          | True | 852 (16.80)        |
| Obesity                          | True | 737 (14.54)        |
| Personality Disorder             | True | 145 (2.86)         |
| Stable Housing                   | True | 556 (10.97)        |
| Substance Use                    | True | 753 (14.85)        |
| Smoking or Tobacco Use           | True | 1454 (28.68)       |
| Food Insecurity                  | True | 200 (3.94)         |

Application 1: Prediction

Table 12: Clinical case study selected hyperparameters

|                         | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|-------------------------|--------|--------|--------|--------|--------|
| **L1 Penalty Strength** |        |        |        |        |        |
| KLR                     | 1e-04  | 1e-03  | 1e-04  | 1e-04  | 1e-04  |
| HFSM-Seq                | 1e-04  | 1e-03  | 1e-03  | 1e-03  | 1e-04  |
| HFSM-Sim                | 1e-03  | 1e-03  | 1e-03  | 1e-03  | 1e-04  |
| **Kernel Function & Data** |    |        |        |        |        |
| KLR                     | SCR,PIST2 | J_ST   | J_PIST2 | SCR,PIST2 | SCR,PIST2 |
| HFSM-Seq                | SCR,ST  | SCR,PIST2 | J_ST   | SCR,PIST2 | J_PIST2 |
| HFSM-Sim                | J_ST    | SCR,PIST2 | J_PIST2 | SCR,PIST2 | J_PIST2 |

*J = Jaccard similarity score
SCR = Common and rare code similarity score
ST = Service type data; PI = Provider type data
PIST2 = both data types*
## Application 2: Inference/Interpretation

Table 13: Feature Coefficients for Models Re-Trained on All Data

| Variable                        | Values                          | HFSM-Seq | HFSM-Sim |
|---------------------------------|---------------------------------|----------|----------|
| Intercept                       |                                 | -4.275   | -5.268   |
| Sex                             | Male                            | -0.261   | -0.202   |
| Rural Residence                 | Urban                           | 0.462    | 0.227    |
|                                 | Missing                         | 0.774    | 0.636    |
| Household Income                | $15,000 to $24,999              | 0.137    | 0.036    |
|                                 | $25,000 to $34,999              | -0.492   | -0.444   |
|                                 | $35,000 to $59,000              | -0.948   | -0.781   |
|                                 | $60,000 or more                 | -1.640   | -1.508   |
|                                 | Do not know                     | 0.517    | 0.446    |
|                                 | Prefer not to answer            | -0.627   | -0.521   |
|                                 | Missing                         | -0.310   | 0.061    |
| Household Composition           | Other Family                    | 0.659    | 0.608    |
|                                 | Unrelated housemates            | 0.743    | 0.677    |
|                                 | Sole Member                     | 0.909    | 0.833    |
|                                 | Do not know or other            | 0.069    | -0.056   |
|                                 | Prefer not to answer            | 0.443    | 0.455    |
|                                 | Missing                         | 0.742    | 0.462    |
| Education Level                 | Secondary or equivalent         | 0.011    | -0.075   |
|                                 | Less than high school           | 0.127    | 0.056    |
|                                 | Do not know or other            | 0.292    | 0.294    |
|                                 | Prefer not to answer            | 0.081    | 0.334    |
|                                 | Missing                         | -0.268   | -0.156   |
| Primary Language                | French                          | 0.114    | 0.671    |
|                                 | Other                           | 0.448    | 0.509    |
| LGBTQ                           | Non-Lgbtq                       | -0.031   | 0.009    |
|                                 | Missing                         | 0.593    | 0.534    |
| Immigrated                      | True                            | 0.277    | 0.102    |
| Physical Disability             | True                            | -0.249   | -0.188   |
| Depression or Anxiety           | True                            | 0.629    | 0.379    |
| Chronic Urinary Problem         | True                            | 0.098    | 0.079    |
| Obesity                         | True                            | 0.145    | 0.019    |
| Personality Disorder            | True                            | 0.179    | 0.114    |
| Stable Housing                  | True                            | 0.934    | 0.626    |
| Substance Use                   | True                            | 0.225    | 0.114    |
| Smoking or Tobacco Use          | True                            | 0.178    | -0.063   |
| Food Insecurity                 | True                            | 0.116    | -0.106   |
Table 14: Client characteristics stratified by kernel coefficient

| Variable                        | Values                      | Zero Alpha | Positive Alpha | Negative Alpha |
|---------------------------------|-----------------------------|------------|---------------|---------------|
| n                               | 5038 (100%)                 | 13 (100%)  | 19 (100%)     |               |
| Loneliness/Social Isolation     | Present                     | 270 (5.36%) | 6 (46.15%)    | 0 (0%)        |
| Sex                             | Female                      | 2362 (46.88%) | 7 (53.85%)  | 10 (52.63%)   |
|                                  | Male                        | 2676 (53.12%) | 6 (46.15%)  | 9 (47.37%)    |
| Rural Residence                 | Rural                       | 1010 (20.05%) | 1 (7.69%)  | 0 (0%)        |
|                                  | Urban                       | 3912 (77.65%) | 11 (84.62%) | 19 (100%)     |
|                                  | Missing                     | 116 (2.3%)   | 1 (7.69%)     | 0 (0%)        |
| Household Income                | $0 to $14,999               | 1240 (24.61%) | 7 (53.85%)  | 7 (36.84%)    |
|                                  | $15,000 to $24,999          | 451 (8.95%)  | 3 (23.08%)    | 0 (0%)        |
|                                  | $25,000 to $34,999          | 272 (5.4%)   | 1 (7.69%)     | 1 (5.26%)     |
|                                  | $35,000 to $59,000          | 533 (10.58%) | 1 (7.69%)     | 1 (5.26%)     |
|                                  | $60,000 or more             | 596 (11.83%) | 0 (0%)        | 4 (21.05%)    |
|                                  | Do not know                 | 273 (5.42%)  | 0 (0%)        | 1 (5.26%)     |
|                                  | Prefer not to answer        | 584 (11.59%) | 1 (7.69%)     | 2 (10.53%)    |
|                                  | Missing                     | 1089 (21.62%) | 0 (0%)     | 3 (15.79%)    |
| Household Composition           | Couple                      | 1886 (37.44%) | 5 (38.46%)  | 6 (31.58%)    |
|                                  | Other Family                | 516 (10.24%) | 0 (0%)       | 3 (15.79%)    |
|                                  | Unrelated housemates        | 214 (4.25%)  | 2 (15.38%)    | 1 (5.26%)     |
|                                  | Sole Member                 | 1197 (23.76%) | 5 (38.46%)  | 3 (15.79%)    |
|                                  | Do not know/Other           | 254 (5.04%)  | 0 (0%)        | 1 (5.26%)     |
|                                  | Prefer not to answer        | 55 (1.09%)   | 1 (7.69%)     | 1 (5.26%)     |
|                                  | Missing                     | 916 (18.18%) | 0 (0%)       | 4 (21.05%)    |
| Education Level                 | Post-secondary or equiv     | 1705 (33.84%) | 4 (30.77%)  | 8 (42.11%)    |
|                                  | Secondary or equivalent     | 1837 (36.46%) | 5 (38.46%)  | 7 (36.84%)    |
|                                  | Less than high school       | 392 (7.78%)  | 1 (7.69%)     | 2 (10.53%)    |
|                                  | Do not know/Other           | 266 (5.28%)  | 3 (23.08%)    | 0 (0%)        |
|                                  | Prefer not to answer        | 54 (1.07%)   | 0 (0%)        | 0 (0%)        |
|                                  | Missing                     | 784 (15.56%) | 0 (0%)       | 2 (10.53%)    |
| Primary Language                | English                     | 4660 (92.5%) | 13 (100%)    | 18 (94.74%)   |
|                                  | French                      | 81 (1.61%)   | 0 (0%)       | 1 (5.26%)     |
|                                  | Other                       | 297 (5.9%)   | 0 (0%)       | 0 (0%)        |
| LGBTQ                           | Lgbtq                       | 66 (1.31%)   | 1 (7.69%)     | 0 (0%)        |
|                                  | Non-Lgbtq                   | 1077 (21.38%) | 2 (15.38%)  | 5 (26.32%)    |
|                                  | Missing                     | 3895 (77.31%) | 10 (76.92%)  | 14 (73.68%)   |
| Immigrated                      | True                        | 624 (12.39%) | 2 (15.38%)    | 1 (5.26%)     |
| Physical Disability             | True                        | 239 (4.74%)  | 1 (7.69%)     | 0 (0%)        |
| Depression or Anxiety           | True                        | 408 (8.1%)   | 2 (15.38%)    | 0 (0%)        |
| Chronic Urinary Problem         | True                        | 848 (16.83%) | 2 (15.38%)    | 2 (10.53%)    |
| Obesity                         | True                        | 732 (14.53%) | 1 (7.69%)     | 4 (21.05%)    |
| Personality Disorder            | True                        | 144 (2.86%)  | 1 (7.69%)     | 0 (0%)        |
| Stable Housing                  | True                        | 549 (10.9%)  | 4 (30.77%)    | 3 (15.79%)    |
| Substance Use                   | True                        | 745 (14.79%) | 4 (30.77%)    | 4 (21.05%)    |
| Smoking or Tobacco Use          | True                        | 1443 (28.64%) | 8 (61.54%)  | 3 (15.79%)    |
| Food Insecurity                 | True                        | 195 (3.87%)  | 4 (30.77%)    | 1 (5.26%)     |
Following are the top 10 codes for each topic from non-negative matrix factorization on provider type and service type (PIST2) data for sub-cohorts of clients with positive, negative, and zero $\alpha$ coefficients.

**NEGATIVE $\alpha$**

**Topic 1 with top 10 weights**  
[('Diagnostic test request', 1.46), ('Intermediate assessment', 1.46), ('Physician', 1.42), ('Nurse', 1.3), ('Discussion regarding the treatment plan', 1.3), ('Health advice/instructions', 1.23), ('Case management/coordination', 1.17), ('Minor assessment', 1.03), ('Discussion regarding the diagnostic findings', 0.96), ('General assessment', 0.91)]

**Topic 2 with top 10 weights**  
[('discussion', 1.41), ('Recommendation/assistance', 1.29), ('Basic support', 0.88), ('Forms completion', 0.87), ('Internal referral', 0.78), ('Information provision about community resources', 0.75), ('counselling', 0.72), ('Internal consultation', 0.54), ('Case management/coordination', 0.44), ('Counselor', 0.42)]

**Topic 3 with top 10 weights**  
[('Counselor', 0.75), ('Individual counselling', 0.74), ('Forms completion', 0.73), ('Foot care', 0.66), ('Chiropodist', 0.66), ('Client intake/interview', 0.59), ('Service access coordinator', 0.49), ('Blank Services (grandfathered)', 0.48), ('Preventive care', 0.47), ('Medication prescription', 0.47)]

**Topic 4 with top 10 weights**  
[('Periodic health examination', 1.1), ('Client intake/interview', 0.88), ('medication prescription', 0.75), ('Nurse Practitioner (RN-EC)', 0.65), ('discussion', 0.64), ('Discussion regarding the diagnostic findings', 0.46), ('Discussion regarding the treatment plan', 0.43), ('Diagnostic test request', 0.37), ('Intermediate assessment', 0.37), ('Preventive care', 0.37)]

**Topic 5 with top 10 weights**  
[('Individual counselling', 1.47), ('Nurse Practitioner (RN-EC)', 1.05), ('internal referral', 0.6), ('Minor assessment', 0.58), ('External referral', 0.57), ('Dietitian/Nutritionist', 0.55), ('assessment', 0.55), ('Health advice/instructions', 0.52), ('Discussion regarding the treatment plan', 0.41), ('Medication renewal', 0.41)]

**POSITIVE $\alpha$**

**Topic 1 with top 10 weights**  
[('Consultation (grandfathered)', 1.25), ('Health advice/instructions', 1.12), ('referral', 1.05), ('discussion', 1.05), ('Advocacy', 1.05), ('Internal consultation', 1.05), ('Physician', 1.02), ('Nurse', 1.02), ('assessment', 0.97), ('Basic support', 0.97)]

**Topic 2 with top 10 weights**  
[('External consultation', 0.74), ('External referral', 0.61), ('Minor assessment', 0.6), ('Social worker', 0.57), ('Transportation assistance', 0.49), ('Individual counselling', 0.49), ('Intermediate assessment', 0.45), ('Information provision about community resources', 0.41), ('medication prescription', 0.39), ('Community Health Worker', 0.35)]

**Topic 3 with top 10 weights**  
[('Preventive care', 0.93), ('Client intake/interview', 0.68), ('Discussion regarding the treatment plan', 0.61), ('Chronic illness monitoring', 0.6), ('Discussion regarding the diagnostic findings', 0.6), ('assessment', 0.58), ('Basic support', 0.58), ('Community Health Worker', 0.57), ('care', 0.55), ('Health advice/instructions', 0.54)]

**Topic 4 with top 10 weights**  
[('Social worker', 0.8), ('Registered Practical Nurse (RPN)', 0.8), ('Individual counselling', 0.79), ('Case management/coordination', 0.66), ('health examination', 0.59), ('Minor assessment', 0.55), ('External referral', 0.5), ('Physician', 0.46), ('Nurse', 0.46), ('Outreach Worker', 0.45)]

**Topic 5 with top 10 weights**  
[('Minor assessment', 0.85), ('Outreach Worker', 0.62), ('Case management/coordination', 0.6), ('Social worker', 0.59), ('General assessment', 0.56), ('Dietitian/Nutritionist', 0.46), ('Foot care', 0.46), ('Diagnostic test request', 0.46), ('Discussion regarding the diagnostic findings', 0.46), ('Registered Practical Nurse (RPN)', 0.37)]
Zero $\alpha$

**Topic 1 with top 10 weights**  
[('Health advice/instructions', 5.89), ('Nurse Practitioner (RN-EC)', 5.34), ('Discussion regarding the treatment plan', 4.96), ('Intermediate assessment', 4.85), ('Minor assessment', 4.46), ('Nurse', 4.29), ('Physician', 3.85), ('Diagnostic test request', 3.79), ('medication prescription', 3.75), ('Discussion regarding the diagnostic findings', 3.61)]

**Topic 2 with top 10 weights**  
[('Basic support', 2.98), ('Advocacy', 2.9), ('Recommendation/assistance', 2.74), ('discussion', 2.58), ('counselling', 2.32), ('Consultation (grandfathered)', 2.26), ('assessment', 2.04), ('Triage', 1.85), ('referral', 1.82), ('Internal consultation', 1.63)]

**Topic 3 with top 10 weights**  
[('General assessment', 2.99), ('internal referral', 2.77), ('Individual counselling', 2.56), ('Physician', 2.29), ('Internal consultation', 2.21), ('Dietitian/Nutritionist', 2.16), ('Nurse', 2.13), ('External referral', 1.93), ('Diagnostic test request', 1.81), ('Consultation (grandfathered)', 1.73)]

**Topic 4 with top 10 weights**  
[('care', 2.42), ('Mental health care', 2.08), ('Preventive care', 1.93), ('Chronic illness monitoring', 1.88), ('Individual counselling', 1.85), ('Dietitian/Nutritionist', 1.4), ('assessment', 1.36), ('Blank Services (grandfathered)', 1.36), ('Dispensing medication', 1.33), ('counselling', 1.33)]

**Topic 5 with top 10 weights**  
[('Information provision about community resources', 3.65), ('Community Health Worker', 2.09), ('Client intake/interview', 2.01), ('Case management/coordination', 1.99), ('Forms completion', 1.96), ('Recommendation/assistance', 1.86), ('Social worker', 1.5), ('Individual counselling', 1.24), ('internal referral', 1.23), ('Health advice/instructions', 1.21)]