# Beyond counting clients: Developing a measure of clinician workload with machine learning

Shauna Heron1,2, Michael Emond1, Luc Rousseau1, Kalpdrum Passi2, & Nicholas Schwabe3

1Department of Psychology, Laurentian University

2Department of Mathematics & Computer Science, Laurentian University

3Compass

# Author Note

Shauna Heron  http://orcid.org/0000-0002-9262-6718

Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: *Michael Emond***:** supervision. *Luc Rousseau***:** advisory committee. *Kalpdrum Passi***:** supervision & advisory committee. *Nicholas Schwabe***:** advisory committee

Correspondence concerning this article should be addressed to Shauna Heron, Department of Psychology, Laurentian University, Sudbury, ON, Canada, Email: sheron@laurentian.ca

# Abstract

As child and youth mental health (CYMH) providers face increasing service demands, anticipating and optimizing staff caseloads is critical to maintaining provider well-being and delivering equitable, high-quality care. However, there is a lack of efficient and reliable tools to support this decision-making in a cost-effective, fair way that accounts for individual client needs. Manual review of client records, necessary to fairly and efficiently allocate new clients and monitor current workload, is untenable in the face of the same workforce shortages. With this gap in mind, we propose examining the utility of leveraging machine learning algorithms trained on electronic mental health records (EHR) to estimate the number of hours per week that individual clients contribute to a provider’s work. Specific objectives include: (i) identify the best features to predict client-related work (case weight) from structured demographic, administrative and assessment EHRs at the earliest stages of client contact (i.e., intake screener scores) and at intervals throughout treatment (i.e., visit counts,  days since last contact, says since last crisis); ii) compare tree-based and neural network machine learning algorithms in their ability to predict client-related work; iii) compare the utility of modelling a continuous index of work (hours per week) compared to a classification of work intensity (i.e., low, medium, high); (iv) explore the potential for early and ongoing prediction of case weight based on individual client need.

*Keywords*: workload, caseload, case management, data science, machine learning, organizational psychology

# Beyond counting clients: Developing a measure of clinician workload with machine learning

Amidst the growing demand for child and youth mental health services in Canada and beyond, human resource challenges have been identified as a significant area of concern ([CMHO, 2022](#ref-childrensmentalhealthontario2022)). In Ontario, a 2020 survey of community child and youth mental health (CYMH) centres revealed that 83% of agencies reported staffing shortages–59% of them direct-service, clinical roles (i.e., psychologists, psychotherapists, and social workers)–which is a concern, as without an adequate workforce, children, youth and their families experience longer wait times and gaps in service that impact access to treatment ([CMHO, 2022](#ref-childrensmentalhealthontario2022)). Illustratively, the same CMHO survey reported that 28,000 children and youth in Ontario were waiting up to 2.5 years for mental health services, some even “aging out” of the system before they are off the wait list ([CMHO, 2019](#ref-cmho2019), [2020](#ref-cmho2020); [CYMHLAC, 2019](#ref-cymhlac2019)). With over 70% of mental health and addiction problems starting before age seventeen, this is a problem([CMHO, 2019](#ref-cmho2019), [2020](#ref-cmho2020)). Not only is a critical opportunity for early intervention missed, but individual and family stress related to mental health challenges are compounded, increasing the burden to a public health care system, where in Ontario, hospitalization of youth with mental health and addiction issues is up by an estimated 90% ([CMHO, 2022](#ref-childrensmentalhealthontario2022); [CYMHLAC, 2019](#ref-cymhlac2019)). At the same time, when demand outpaces staffing, existing providers have to manage higher client volumes containing more complex cases, perpetuating a cycle of provider burnout, absenteeism and high turnover ([Comeau et al., 2019](#ref-comeau2019); [King, 2009](#ref-king2009)). For this reason, the ability to anticipate and monitor the contribution of individual cases to the overall work of providers is critical to improving client outcomes and minimizing provider burnout ([King et al., 2000](#ref-king2000a), [2004a](#ref-king2004), [2004b](#ref-king2004a)).

According to recent reports by the Auditor General on Child and Youth Mental Health, a vital issue limiting agencies’ ability to meet rising demand is the challenge of establishing provider-to-client workload ratios that accurately reflect the varying needs and levels of intervention required by different clients ([Auditor General’s Report, 2016](#X74ef00379ba32932a14081a0b155af9f4f12f08), [2018](#Xfd2ada954160b70803c93fd53cbcc2893671edf)). Ideally, caseloads should decrease as case complexity increases; however, the resources required to evaluate each child’s needs across hundreds of cases manually are beyond what most public agencies can support (CMHO, 2019). As a result, cases are often assigned under the assumption that each requires a similar level of effort, leading to inequitable case distribution. Consequently, some clinicians consistently manage a higher proportion of complex cases than others ([CMHO, 2022](#ref-childrensmentalhealthontario2022); [King, 2009](#ref-king2009)). For example, an agency might target 20 cases per provider for counselling services, meaning that providers with fewer than 20 cases are considered to have “room” for more, regardless of case complexity.

Given this state of affairs, if there was an automated tool that could estimate a weight for each case based on information in the client record, agencies might more efficiently evaluate and manage provider caseloads, which in turn might reduce workload-associated provider burnout and improve client outcomes. However, the development of sophisticated predictive tools to augment clinical decision-making has been hampered by several factors: i) a lack of resources across the public health sector generally, ii) limits imposed by a mostly paper-based client record system, iii) disagreements on how “workload” should be defined and iv) lack of computing power and expertise in modelling complex, electronic health data. However, the transition of CYMH services in Ontario from paper-based health records to electronic records, combined with increased computational power and advances in computer science, has opened the door to leveraging EHR with algorithms to better anticipate and manage healthcare resources and outcomes.

With this gap in mind, the current research proposes to explore the feasibility of predicting the work associated with a given case at different points in the client timeline to examine whether such predictions could provide added value to clinical practice. The assumption underlying the research is that there are historical patterns that predict future mental health resource use and that such patterns can be identified in electronic mental health records (EHR) despite their sparseness, noise, errors, and systematic bias.

## Case-mix Review

Across health domains (e.g., psychiatric and emergency medicine), various strategies have been employed to manage provider workload by determining service levels with client characteristics like symptom severity or prior diagnoses ([Johnson et al., 1998](#ref-johnson1998); [Tran et al., 2019](#ref-tran2019)). These systems assume that although the needs of each individual in a population will be unique, there will be shared characteristics that determine the type of treatment they need (e.g., family counselling versus substance use treatment). These groups represent the mix of cases in a given caseload, which is often used to estimate cost based on the types of care the population needs and the time involved in servicing those needs ([Johnson et al., 1998](#ref-johnson1998); [Tran et al., 2019](#ref-tran2019)).

Case-mix classification systems have been used in the healthcare sector to help payers and agencies monitor costs by categorizing clients based on their expected resource use ([Johnson et al., 1998](#ref-johnson1998); [Tran et al., 2019](#ref-tran2019)). Casemix algorithms assume that though the needs of an individual will be unique, shared characteristics determine the type and intensity of treatment needed (e.g., family counselling versus crisis intervention). Typically, these systems are informed by information contained in case records. At the agency level, case records contain a variety of information, including provider-level information like the number of direct and indirect hours attributable to individual clients, as well as client-level characteristics like diagnoses, treatment history, referral source and presenting symptoms (e.g., crisis intervention versus brief services).

Typically, case-mix systems take one of two approaches to classification ([CMHO, 2019](#ref-cmho2019)). Grouping systems assign people to classes in terms of their expected resource use, with each group having a specific weight (e.g., time-intensive treatment versus brief treatment) relative to the average case in the population. For example, a client accessing long-term counselling and therapy services might be assigned a different weight in terms of expected resource use than a client accessing a one-session brief service. Index systems, on the other hand, combine different case characteristics to provide a value that maps to expected resource use or acuity of needs (e.g. a case weight or case complexity score that ranges from 0, the least complex, to 1, the most complex) ([Tran et al., 2019](#ref-tran2019)). Indexing systems are often used to triage cases by assigning a score to new clients based on answers to an intake assessment. Often, there is a threshold score above which clients are considered acute and may receive services more quickly; at the same time, scores below a specific threshold may not qualify for publicly funded services. For instance, a youth reporting thoughts of suicide or other self-harming behaviour might be scored higher than a youth reporting problems focusing in school.

While the term *algorithm* gives the impression that these are sophisticated mathematical models, they are usually conceptual, rules-based frameworks that rely on manual review of client files rather than an automated, data-driven tool. Although research has begun to explore how machine learning could automate and streamline classification processes, prior work has focused primarily on inpatient acute-care settings, which differ markedly from community-based outpatient settings ([Tran et al., 2019](#ref-tran2019)), where clear diagnostic criteria and predictable recovery paths are more variable. For instance, the healing process for a broken arm has a relatively fixed timeline and treatment protocol. However, recovery from anxiety or depression is more nuanced and individualized, making it inherently more complicated to model.

The difficulties inherent in modelling electronic mental health data are underscored by the fact that only a handful of studies have looked at solving this problem despite an urgent need. Indeed, a 2019 scoping review of casemix literature in community-based mental health care found only a single case that looked at data-driven methods to predict mental health care resource ([Martin et al., 2020](#ref-martin2020); [Tran et al., 2019](#ref-tran2019)). In that study, researchers modelled 4573 client records from eleven UK outpatient CYMH agencies, comparing cluster analysis, regression trees and a conceptual classification based on clinical best practice guidelines to predict the number of appointments a client attended in treatment ([Martin et al., 2020](#ref-martin2020)), finding the data-driven classification no more clinically meaningful than conceptual classification in accounting for number of appointments. There was little evidence to support the idea that either client complexity or context factors (with the exception of school attendance problems) were linked to overall appointment counts ([Martin et al., 2020](#ref-martin2020)). Moreover, the models failed to explain significant variations in resource provision between workers despite clients exhibiting similar characteristics. Data quality problems and omission of critical individual-level factors were cited as potential points of failure, but suggesting their results merited further testing and development.

In a related cohort, researchers tried to predict the work associated with client features at a community-based mental health centre for the elderly ([Baillon et al., 2009](#ref-baillon2009)). Using an 8-item self-designed case weighting scale (CWS), researchers identified factors that staff felt contributed to the demand for time. A multiple regression model was used to assign different weightings to predictors based on the strength of their relationship with the outcome (an estimation of time spent on each client logged over four weeks). The resulting coefficients were then added to a spreadsheet and used to predict the total time a client would utilize in a four-week period following the first appointment based on the eight characteristics. Though the model was reported a success, accounting for 58% of the variance in time spent on client-related work, the sample consisted of only 87 cases leaving it unclear how accurate the model was ([Baillon et al., 2009](#ref-baillon2009); [Mansournia et al., 2021](#ref-mansournia2021)). Moreover, inter-rater and re-rater reliability results indicated that the assessment, whether from a client’s self-report or a professional’s clinical opinion, did not necessarily relate to the amount of time needed by clients ([Baillon et al., 2009](#ref-baillon2009)).

### Machine learning, a novel approach to modeling case-mix

Considering the challenges outlined by prior research in modelling the high-dimensional, sparse data characteristic of EHRs, we next looked to a growing body of research leveraging machine learning algorithms to model electronic health data. Machine learning (ML) is a branch of artificial intelligence that uses statistical techniques that enable computers to learn patterns from data without explicit instructions ([Nielsen, 2016](#ref-nielsen2016)). In mathematical terms, machine learning algorithms use statistical techniques to optimize a model’s parameters. This process involves minimizing a loss function that quantifies the difference between the model’s predictions and the actual data. For example, in supervised learning, the goal is to find a function f(x) that maps input features X to an output Y such that the predicted outcomes are as close as possible to the actual outcome ([Nielsen, 2016](#ref-nielsen2016)). This approach is highly effective for complex tasks like image recognition, natural language processing, and predictive analytics, where traditional rule-based programming is infeasible due to the high dimensionality and variability of the data. ML algorithms are particularly well-suited to model complex, high-dimensional data in EHRs ([An et al., 2023](#ref-an2023); [Chen et al., 2023](#ref-chen2023)).

Within the inpatient mental health domain, machine learning has mainly been used to predict specific events like substance relapse, self-harm and suicide risk. However, a recent study leveraged ML to build a model that *continuously* monitors patient records to predict crisis relapse over 28 days ([Garriga et al., 2022](#ref-garriga2022)). The winning XGBoost regression demonstrated good accuracy in distinguishing between cases that were likely and unlikely to experience a crisis in the next 28 days. Specifically, the model could correctly differentiate those at risk from those not at risk about 80% of the time ([Garriga et al., 2022](#ref-garriga2022)). Moreover, in a subsequent post-hoc case study, healthcare professionals rated the predictions produced by the model as valuable for managing patient care in 64% of cases, helping them to prioritize patients more effectively and potentially prevent crises ([Garriga et al., 2022](#ref-garriga2022)). Though the authors did not model the resource use directly as we hope to do, ‘crisis risk’ served as a proxy for work. By predicting crises, they hoped to anticipate increased resource demand, which might better inform case prioritization and management.

## The current study

Building on insights from Garriga et al. ([2022](#ref-garriga2022)), the current research aims to explore the feasibility of estimating the number of weekly direct hours a case may require, assessed at 28-day intervals. The underlying assumption is that historical patterns can reliably predict future mental health resource use. Such patterns are identifiable in electronic mental health records (EHR) despite challenges such as sparsity, noise, errors, and systematic bias.

To test these assumptions, we will use a retrospective, deidentified dataset from a large child and youth mental health (CYMH) agency in Ontario, encompassing clients served from 2019 to the end of 2023. Although largely exploratory, the study will be guided by several hypotheses. First, based on Garriga et al. ([2022](#ref-garriga2022))’s research, we expect predictions to be weakest early in the client journey when EHR information is limited to an intake screener and basic demographics. However, as more data accumulates across the treatment timeline, we anticipate that prediction accuracy will improve.

Furthermore, we hypothesize that for new clients, mental health acuity features such as externalizing behaviours and a history of self-harm or suicide attempts will be stronger predictors of required resources. For known clients, however, we anticipate that time-based factors, such as the number of no-shows or the number of crisis events, will be more predictive of workload needs.

Finally, we expect the winning machine learning algorithm to outperform the baseline model mimicking the conceptual classification system that is the way that resource use is typically estimated by agencies today (e.g., counselling and therapy will require more provider-hours than a brief service) ([CMHO, 2022](#ref-childrensmentalhealthontario2022)). Finally, we expect the winning machine learning algorithm to outperform the baseline model, which is intended to mimic the conceptual way that resource use is estimated by agencies today (e.g., counselling and therapy will require more provider-hours than a brief service) ([CMHO, 2022](#ref-childrensmentalhealthontario2022)).

# Methodology

Considering we want to predict resource use (hours of direct and indirect service) at regular stages in the client journey, only cases with a completed initial screener will be included in the analysis. Final counts after screening will be reported and added to the flowchart before analysis. The study will be conducted at Compass Child and Youth Family Services, the largest CYMH agency in northern Ontario, serving a culturally and socially diverse population of children, youth and families. The study will utilize a retrospective dataset containing de-identified client records with completed intake assessments who were active between January 1, 2019, and December 31, 2024.

## Data Security

Given the sensitivity of mental health data, data privacy and security will be strictly enforced by obtaining the necessary ethical approvals and maintaining transparency throughout the research process. Relevant ethics boards will obtain the necessary approvals. For the use of deidentified data, an exemption must be granted by both the agency (Compass) and Laurentian’s institutional review board.

De-identified clinical data will be acquired from an electronic health information system belonging to Compass. The institution maintains the EHR database. Data will be deidentified at extraction using the Health Insurance Portability and Accountability Act Safe Harbor Method ([OCR, 2012](#ref-rightsocr2012)). This means that names, addresses, birth dates, postal codes and any other directly identifying information will be stripped from the dataset before any analysis begins. As an added precaution, unique client identification codes will be encrypted with a hashing system, making it nearly impossible to reverse engineer the code to obtain original IDs. Furthermore, the data will not leave the custody of Compass and will only be analyzed by the principal researcher within a password-protected machine belonging to Compass.

The reporting of model results, summary statistics and other visualizations will only include metrics associated with the performance of predictors and the models themselves, never individual scores or any other identifying information that could be linked to clients or smaller subgroups of clients. Furthermore, the researchers will seek approval from Compass before results are shared or utilized in any report or presentation.

## Dataset

The de-identified data will include approximately 6000 records containing demographic information, referrals, diagnoses, risk and well-being assessments and crisis events for all outpatients. Cases younger than five and older than 17 will be excluded, as Compass’ core services are only offered to children and youth under 18. There are no plans to exclude cases based on any other feature, including diagnoses; however, if, for whatever reason, this changes, it will be outlined in the documentation.

## Procedure

The following steps outline the proposed process, which will consist of four phases: 1) data cleaning, preprocessing and exploration; 2) data splitting; 3) aggregating workload proxies (direct and indirect services hours) that will be used as stand-ins for actual workload; 4) identifying and aggregating indicators of workload (i.e., independent variables/features) that will be used to model our proxies; 5) modelling the relationship between the indicators and proxies with machine learning algorithms of varying complexity; and 6) evaluating the accuracy of predictions on the unseen test data (see [Figure 1](#fig-procedure-flow)).

Figure 1

Experimental procedure

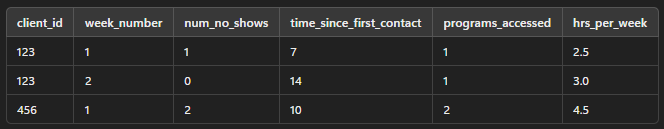
*Note*. Flowchart of the experimental procedure. Data will be split into training and test sets by client ID. The training set will be used to train the models using 10-fold group cross-validation, while the test set will act as a control group to evaluate the models’ performance on unseen data.

## Data Preprocessing

After de-identification, data preprocessing will involve cleaning, joining data frames, handling missing values if necessary, and aggregating features across weeks. All decisions we make regarding missing data, data normalization or any other changes will be decided on a case-by-case basis and reported in our final paper. Pending approval from the host agency, the Python data scripts will also be publicly available.

Importantly, feature engineering–the creation of new predictors based on existing variables in the dataset–will occur *after* data splitting on the training data to minimize the risk of data leakage that could inadvertently occur when creating new variables from the full dataset ([Sheetal et al., 2023](#ref-sheetal2023)). The main criterion for inclusion in the model will be the variable’s availability in the electronic health record (EHR) system at intake. There will be four primary types of predictor: i) time-based, ii) count-based, iii) recent information, and iv) static and semi-static information.

With the exception of static information like presenting concerns or referral sources, all EHR data will include the associated date and time. The date and time refer to the moment when the specific event or assessment occurred—that is, the date and time that there was contact with a client. To prepare the data for the modelling task, each client’s case records will be consolidated weekly according to the date associated with the record. Following this process, evenly spaced time series will be generated for each client, spanning from their first interaction with Compass to the study’s final week. The features and labels generated for each week will be computed using the data from dates prior to that week. Static data susceptible to change over time (for example, postal code or school board information) will be removed to mitigate the risk of retrospective leakage.



### Outcome (target) generation

**Caseweight prediction modeling and evaluation.** The case weight prediction task will be defined as a continuous regression problem to be performed weekly and a continuous classification problem on the same timeline (low, medium, and high intensity). For each week, the model will estimate the weekly hours needed during the upcoming 28 days. A rolling window approach will be applied to allow for a periodic update of the case weight by incorporating newly available data (or the absence of it) at the beginning of each week. This approach is common when predictions are used in real-time and when data are continuously updated ([Garriga et al., 2022](#ref-garriga2022)).

To construct a continuous case-weight prediction target, the sum of client-related direct and indirect hours logged by clinicians and associated with a specific program and client will be aggregated at the same weekly level as the features, based on the time recorded by the worker prior to that week. We also intend to examine which measure of client-related time is most stable and reliable over time: the summed direct and indirect time or either on its own.

### Features (predictors) generation.

We will extract features from a total possible feature set of approximately 250 features. See [Table 1](#tbl-predictors) for a list of proposed feature groupings and variables that we aim to include. Informed by Garriga et al. ([2022](#ref-garriga2022))’s methodology, extraction will be performed according to six data characteristics:

**Static or semi-static features.** Demographics data will be represented as constant values attributed to each case, with age treated as a special case that changes yearly.

**Diagnosis features.** Client will be assigned their latest valid diagnosed developmental disability or psychological disorder (if contained in the record) or an ‘un-diagnosed’ label and then separated into diagnostic groups according to the latest valid diagnosed disorder at the last week of the training set to avoid leakage into the validation and test sets. Any codes created for this feature will be added to the final paper.

**EHR weekly aggregations.** Records related to client-agency interactions will be aggregated on a weekly basis for each client. The resulting features will constitute counts of each type interaction and one-hot encoded according to their categorization. One-hot coding is a type of dummy variable where if a specific event did occur, there will be a 1 for that week, and if not, a value of 0 will be assigned to the feature related to the corresponding event for the corresponding week.

**Time-elapsed features.** For each type of interaction and category and each client week, we will construct a feature that counts the number of weeks since the last occurrence of the corresponding event. If the client has never experienced such an event type up to that point in time, an NA value will be used.

**Last crisis episode descriptors.** For each crisis episode, a set of descriptors will be used to build features for the subsequent weeks until the next crisis occurs. If the client has never had a crisis event up until that point in time, NA values will be used.

**Last assessment descriptors.** For each assessment item, a set of descriptors will be used to build a feature for the subsequent weeks until the following assessment occurs. All clients will have at least one assessment to be included in the study. A penalty will reduce the value of assessment scores the further away from the assessment date they are for each aggregated week.

**Status features.** For specific records, characterized by a start and end date, features for the corresponding weeks will be built by assigning their corresponding value (or category); otherwise, they are set to NA.

In addition to record-based features, we will add the week number (of a year 1-52) to account for seasonality effects.

Table 1

Planned Features (Predictors)

| Time based | Count based | Latest available assessment / contact information | Static/semi-static information |
| --- | --- | --- | --- |
| Weeks since last crisis event | Count of crisis events | Mental health acuity scores (e.g., depression, anxiety, internalizing, externalizing etc.) | Age, gender, school district |
| Weeks since first contact | Count of no-shows | Identified risks such as substance use, self-harm or suicide risk | ADHD/Autism diagnosis |
| Weeks since last no-show | Counts of substances used | Current services accessed | Mental health diagnosis |
| Weeks since last contact | Counts of phone calls | Identified symptoms | Learning disability diagnosis |
| Weeks since substance misuse identified | Count of previous completed services | Previously indicated need (CHAMPS) |  |
| Weeks since self/harm identified | Discharge and referral counts | Recent contact with CAS |  |
| Weeks since suicide risk identified | Number of current services | Recent psychological consult |  |
| Weeks since last discharge |  | Recent diagnosis |  |
| Weeks since last crisis episode |  | Current caregiver and family support |  |
| Weeks since first visit |  |  |  |
| Weeks since referral |  |  |  |

### Data Splitting

To maximize our models’ generalizability, we aim to apply a time-based 80/10/10 training/validation/test split depending on the amount of data left over after data cleaning. Roughly, training data will start in the first week of January 2019 and end in the last week of April 2023. Validation data will start in the first week of May 2023 and end in the last week of October 2023. Test data will start in the first week of November 2024 and end at the end of April 2024. Data from the first 6 months of the COVID-19 pandemic may need to be included due to disruptions in normal service delivery that continued until new policies and procedures could be implemented, depending on the irregularity of these data.

Timed-based cross-validation will be used to tune model parameters. The cross-validation folds will be created with a portion of the training data, subdividing it into ten subsets, or “folds,” that will preserve the same time structure. Evaluations will be conducted on weekly predictions, and each week’s results will be used to build confidence intervals for the evaluated metrics. All reported results will be computed using the test set if not otherwise indicated. The test set will act as a control group to evaluate the models’ performance on “unseen data” at the end of the training and tuning processes. Keeping the test set separate and untouched during training ensures that our final evaluation provides a better estimation of how the models will perform in practice. This final step is crucial for assessing the models’ generalizability and for identifying any over-fitting that may have occurred during training.

A final and complete list of all variables will be included in the final report.

### Model Selection

We plan to utilize the following supervised machine-learning algorithms for both regression and classification problems: i) Random Forest (RF) for its ability to handle large datasets with high dimensionality; ii) XGBoost, known for its predictive accuracy on tabular datasets; iii) Random Forest; iv) Feed Forward Neural Network, and finally iv) linear and logistic regression will be used to model our clinical baseline. All algorithms will be trained on the same aggregated training set and cross-validation folds and evaluated on the same test set. These algorithms were chosen based on their success in modelling similarly complex, tabular data types and may grow to include other models in the final paper ([Salditt et al., 2023](#ref-salditt2023); [Sheetal et al., 2023](#ref-sheetal2023)). See [Table 2](#tbl-models) for a list of proposed models.

**Machine Learning Models and Classifiers.** We plan to utilize versions of the following supervised machine-learning algorithms depending on the specific target outcome. Since we will be modelling and comparing a continuous and a classification outcome, this approach will allow us to assess performance across different types of predictions. XGBoost, an implementation of gradient boosting machines (GBMs), will serve as our primary algorithm due to its ability to handle missing data and robustness to scaling factors, eliminating the need for imputation or scaling. GBMs build a sequence of decision trees, where each tree improves on the performance of prior iterations, making them well-suited for our case weight prediction task. To benchmark performance, we will compare XGBoost to a selection of state-of-the-art machine learning classifiers, including logistic regression, naive Bayes, random forest, and neural networks (specifically, multi-layer perceptrons), all of which have been successfully applied to similar prediction tasks with electronic health records (EHRs). We will apply standard scaling and imputation as needed for these classifiers to ensure comparable conditions. We will conduct 100 hyperparameter optimization trials for each classifier to identify optimal parameters, with detailed search spaces provided in the supplementary materials.

**Hyperparameter Tuning and Feature Selection.** To optimize the models’ hyper-parameters, we will use a Bayesian optimization approach to maximize the area under the receiver operating characteristic curve (AUROC) for classification outcomes and the mean squared error (MSE and MAE) for continuous outcomes. Specifically, we will use Hyperopt, a sequential model-based optimization algorithm that applies Bayesian optimization via the Tree-structured Parzen Estimator, which accommodates a variety of distributions across search spaces. This flexibility will make Hyperopt particularly effective for tuning hyper-parameters across all classifiers. We will use the same approach for feature selection, grouping features by information gain and adding a binary indicator to determine whether each feature should be selected for the model. See [Table 1](#tbl-predictors) for details on feature groupings and the feature selection.

Table 2

Planned Machine Learning Models

| Regression & Classification Models Proposed | Justification |
| --- | --- |
| Clinical baseline |  |
| XGBoost |  |
| Regression (linear & multivariable logistic) |  |
| Random Forest |  |
| Naïve Bayes |  |
| Feed Forward Neural Network |  |

## Validation and Testing

Final models will be statistically compared and evaluated on the test set using appropriate performance metrics depending on whether it is a regression task (mean absolute error or root mean squared error) or classification task (accuracy, precision, recall and area under the curve). The evaluations will help determine each model’s accuracy, generalizability and robustness ([Salditt et al., 2023](#ref-salditt2023); [Wang et al., 2021](#ref-wang2021)). Final models will also be analyzed to identify which predictors were the most important in terms of estimating client-related work.

Furthermore, to enhance the interpretability of our model, we plan to implement SHapley Additive exPlanations (SHAP) for feature analysis ([Lundberg & Lee, 2017](#ref-lundberg)). SHAP is a method that helps quantify the contribution of each feature to the model’s predictions, providing insights into how specific client characteristics and historical data points influence predicted weekly clinician hours. Interpretability is essential in a mental health care setting, as decisions directly impact client care and resource allocation ([Feretzakis et al., 2024](#ref-feretzakis2024)). Clinicians and administrators need to understand not only the predicted workload but also the driving factors behind each prediction to ensure fair, personalized, and transparent decision-making. For instance, if certain factors like recent diagnoses or patterns of no-shows are highly influential, this can guide intervention strategies and inform staffing decisions tailored to client needs. SHAP’s ability to provide such detailed, interpretable explanations makes it a critical tool for ensuring that the model’s predictions are aligned with clinical understanding and ethical care practices ([Feretzakis et al., 2024](#ref-feretzakis2024)).

## Software and Tools

A Python kernel will be used for model building and evaluation (Khun & Wickham 2020). Quarto Markdown will be used for documentation and reproducibility inside Positron IDE ([Van Rossum & Drake, 1995](#ref-vanrossum1995)).

# Limitations and Challenges

While our study aims to enhance understanding of client-related workload over time based on historical and real-time changes in client needs, several limitations should be acknowledged. First, our data is derived from a specific subset of the population—young people with mental health concerns in community outpatient settings—which may limit the generalizability of our findings to other demographics or healthcare settings. Additionally, although we are employing machine learning techniques to handle the complexity of electronic health data, these methods are not immune to biases inherent to the data itself. Systematic biases in the initial data collection process, such as underreporting, data entry errors, or misclassification, could influence the model’s predictions.

Moreover, our reliance on electronic health records means that the quality and completeness of the data are contingent upon the accuracy and thoroughness of data entry made by providers. Missing data and inconsistencies are inherent challenges that could affect the robustness of our models. Additionally, many of the scale scores may be influenced by the subjective interpretation of the provider who administered the assessment. While we will attempt to reduce these issues, there is no guarantee that all biases can be fully mitigated.

Another limitation is the exclusion of provider-side variables from our models. While this decision is aimed at maximizing fairness in case allocation, it also means that potentially valuable information about resource utilization influenced by provider characteristics is not considered. This could impact the comprehensiveness and accuracy of our workload predictions. In future iterations, it might be interesting to introduce a feedback loop where staff perception of workload is accounted for with a weekly or monthly “caseload satisfaction” measure.

Finally, while predictive accuracy and interpretability are crucial, a prospective cohort study would be necessary as a next step to evaluate how effectively the model supports clinical decision-making in practice. Such a study would allow us to track how predictions influence clinician workload distribution and client outcomes over time. It would provide a deeper understanding of its practical benefits and potential drawbacks in a live clinical setting. Garriga et al. ([2022](#ref-garriga2022)) demonstrated this approach effectively, showing that prospective cohort studies can offer insights into the model’s impact on workflow, clinician satisfaction, and client care quality. In future research, implementing a cohort study could help validate the model’s usefulness and refine it for improved applicability in mental health care settings.

# Conclusion

In conclusion, this research represents a crucial step toward addressing the complex and growing demands within mental health services with a data-driven approach. By developing a machine learning model to predict clinician workload, we aim to offer actionable insights that support fair resource distribution and responsive service delivery. Our approach will not only contribute to the field of mental health care by enhancing our understanding of workload drivers but also align with the pressing need for scalable, automated and efficient care solutions. Ultimately, this research has the potential to improve outcomes for clinicians and clients alike, ensuring that mental health care services are equipped to meet the needs of vulnerable populations with greater precision and equity.

# References

An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. *Sensors*, *23*(9), 4178. <https://doi.org/10.3390/s23094178>

Auditor General’s Report. (2016). *Office of the Auditor General Annual Report* (pp. 110–147). <https://www.auditor.on.ca/en/content/annualreports/arreports/en16/v1_301en16.pdf>

Auditor General’s Report. (2018). *1.01 Child and Youth Mental Health*. <https://www.auditor.on.ca/en/content/annualreports/arreports/en18/v2_101en18.pdf>

Baillon, S. F., Simpson, R. G., Poole, N. J., Colledge, R. J., Taub, N. A., & Prettyman, R. J. (2009). The development of a scale to aid caseload weighting in a community mental health team for older people. *Journal of Mental Health*, *18*(3), 253–261. <https://doi.org/10.1080/09638230802522981>

Chen, X., Chen, H., Nan, S., Kong, X., Duan, H., & Zhu, H. (2023). Dealing with missing, imbalanced, and sparse features during the development of a prediction model for sudden death using emergency medicine data: Machine learning approach. *JMIR Medical Informatics*, *11*, e38590. <https://doi.org/10.2196/38590>

CMHO. (2019). *Developing caseload/workload guidelines for ontario’s child and youth mental health sector*. Children’s Mental Health Ontario. <https://iknow-oce.esolutionsgroup.ca/api/ServiceItem/GetDocument?clientId=A1B5AA8F-88A1-4688-83F8-FF0A5B083EF3&documentId=6aeb8d26-16ac-4353-b49c-c21b131b735c>

CMHO. (2020). *Kids can’t wait* (p. 12).

CMHO. (2022). *Addressing urgent workforce challenges in child and youth mental health* (p. 4).

Comeau, J., Georgiades, K., Duncan, L., Wang, L., & Boyle, M. H. (2019). Changes in the prevalence of child and youth mental disorders and perceived need for professional help between 1983 and 2014: Evidence from the ontario child health study. *The Canadian Journal of Psychiatry*, *64*(4), 256–264. <https://doi.org/10.1177/0706743719830035>

CYMHLAC. (2019). *Realizing the potential: Strengthening the ontario mental health system for children, youth and their families*. Child & Youth Mental Health Lead Agency Consortium. <https://www.neofacs.org/files/Realizing-the-Potential-2018-2019-Provincial-Priorities-Report-EN-web-1.pdf>

Feretzakis, G., Sakagianni, A., Anastasiou, A., Kapogianni, I., Bazakidou, E., Koufopoulos, P., Koumpouros, Y., Koufopoulou, C., Kaldis, V., & Verykios, V. S. (2024). Integrating Shapley Values into Machine Learning Techniques for Enhanced Predictions of Hospital Admissions. *Applied Sciences*, *14*(13), 5925. <https://doi.org/10.3390/app14135925>

Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*, *28*(6), 1240–1248. <https://doi.org/10.1038/s41591-022-01811-5>

Johnson, L. M., Richards, J., Pink, G. H., & Campbell, L. (1998). *Case-Mix Tools for Decision Making in Healthcare*.

King, R. (2009). Caseload management, work-related stress and case manager self-efficacy among victorian mental health case managers. *The Australian and New Zealand Journal of Psychiatry*, *43*(5), 453–459. <https://doi.org/10.1080/00048670902817661>

King, R., Le Bas, J., & Spooner, D. (2000). The impact of caseload on the personal efficacy of mental health case managers. *Psychiatric Services*, *51*(3), 364–368. <https://doi.org/10.1176/appi.ps.51.3.364>

King, R., Meadows, G., & Le Bas, J. (2004a). Compiling a caseload index for mental health case management. *The Australian and New Zealand Journal of Psychiatry*, *38*, 455–462. <https://doi.org/10.1111/j.1440-1614.2004.01388.x>

King, R., Meadows, G., & Le Bas, J. (2004b). Compiling a caseload index for mental health case management. *The Australian and New Zealand Journal of Psychiatry*, *38*, 455–462. <https://doi.org/10.1111/j.1440-1614.2004.01388.x>

Lundberg, S., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. <https://doi.org/10.48550/arXiv.1705.07874>

Mansournia, M. A., Waters, R., Nazemipour, M., Bland, M., & Altman, D. G. (2021). Bland-altman methods for comparing methods of measurement and response to criticisms. *Global Epidemiology*, *3*, 100045. <https://doi.org/10.1016/j.gloepi.2020.100045>

Martin, P., Davies, R., Macdougall, A., Ritchie, B., Vostanis, P., Whale, A., & Wolpert, M. (2020). Developing a case mix classification for child and adolescent mental health services: The influence of presenting problems, complexity factors and service providers on number of appointments. *Journal of Mental Health (Abingdon, England)*, *29*(4), 431–438. <https://doi.org/10.1080/09638237.2017.1370631>

Nielsen, D. (2016). *Tree boosting with XGBoost* [PhD thesis]. <https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2433761/16128_FULLTEXT.pdf>

OCR. (2012). *Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule*. <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>

Salditt, M., Humberg, S., & Nestler, S. (2023). Gradient tree boosting for hierarchical data. *Multivariate Behavioral Research*, *58*(5), 911–937. <https://doi.org/10.1080/00273171.2022.2146638>

Sheetal, A., Jiang, Z., & Di Milia, L. (2023). Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers. *Applied Psychology*, *72*(3), 1339–1364. <https://doi.org/10.1111/apps.12435>

Tran, N., Poss, J. W., Perlman, C., & Hirdes, J. P. (2019). Case-mix classification for mental health care in community settings: A scoping review. *Health Services Insights*, *12*, 1178632919862248. <https://doi.org/10.1177/1178632919862248>

Van Rossum, G., & Drake, F. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Wang, X., Blumenthal, H. J., Hoffman, D., Benda, N., Kim, T., Perry, S., Franklin, E. S., Roth, E. M., Hettinger, A. Z., & Bisantz, A. M. (2021). Modeling patient-related workload in the emergency department using electronic health record data. *International Journal of Medical Informatics*, *150*, 104451. <https://doi.org/10.1016/j.ijmedinf.2021.104451>

# Appendix

TABLES GO HERE

Figure A1

Modeling Caseweight–client-related work

*Note*. Using indicators of client-related work (e.g. depression scores, anxiety scores, etc.) in the electronic health record (EHR)to predict workload proxies. Adapted from *Predictors of Workload*, by Wang et al. ([2021](#ref-wang2021)).