# 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 are a lack of efficient and reliable tools to support this decision-making in a way that accounts for variable client need and is cost-effective and fair. Manual review of client records, necessary to fairly and efficiently allocate new clients and monitor existing caseloads, 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 (EHRs) to estimate the number of provider hours that a client may require in the weeks to come (caseweight). Specific objectives include: (i) identifying the features that best predict client-related provider hours from structured demographic, administrative and assessment EHRs at the earliest stages of client contact (i.e., intake screener scores) and at weekly intervals throughout treatment (i.e., aggregated visit counts,  days since last contact); ii) compare tree-based and neural network machine learning algorithms in their ability to predict client-related provider hours; iii) compare the utility of modelling a continuous index of needed provider hours (caseweight) compared to a classification of the same (i.e., low, medium, high); (iv) conduct interpretability analyses to identify and explain the contributions of individual features to model predictions.

*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, human resource challenges have been identified as a significant barrier to providing timely, high-quality care ([CMHO, 2022](#ref-childrensmentalhealthontario2022); [WHO, 2022](#ref-worldme2022)). In 2020, a survey of community child and youth mental health (CYMH) centres in Ontario revealed that 83% of agencies reported staff vacancies, 59% of them direct service, front line positions (i.e., social workers, psychologists, and psychotherapists). This is a concern, as without an adequate and qualified workforce, children, youth and their families experience longer wait times, causing gaps in service that ultimately impact outcomes ([CMHO, 2020](#ref-childrensmentalhealthontario2020); [Comeau et al., 2019](#ref-comeau2019)). 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 were off the wait list ([CMHO, 2020](#ref-childrensmentalhealthontario2020); [CYMHLAC, 2019](#ref-cymhlac2019)).

With over 70% of mental health and addiction problems starting before age seventeen, any delay in access to service is a problem ([CMHO, 2019](#ref-cmho2019), [2020](#ref-childrensmentalhealthontario2020); [WHO, 2022](#ref-worldme2022)). Not only are critical opportunities 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 has increased over the last 30 years by an estimated 90% ([CMHO, 2020](#ref-childrensmentalhealthontario2020); [CYMHLAC, 2019](#ref-cymhlac2019)). At the same time, when demand outpaces staffing, existing providers often end up managing higher client volumes containing more complex cases, which can perpetuate 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 caseloads of providers is critical to improving client outcomes and minimizing provider burnout ([King et al., 2000](#ref-king2000); [King, 2009](#ref-king2009)).

According to recent reports by the Auditor General of Ontario on Child and Youth Mental Health, a vital issue limiting agencies’ ability to meet rising demand is the challenge of monitoring client-to-provider workload ratios in a way that accounts for individual client needs ([Auditor General of Ontario, 2016](#X74ef00379ba32932a14081a0b155af9f4f12f08), [2018](#Xfd2ada954160b70803c93fd53cbcc2893671edf)). Ideally, as case complexity increases, the overall number of cases in a provider’s portfolio (case count) should decrease; however, the administrative resources required to manually evaluate each case across dozens of caseloads are beyond what most public agencies can support ([CMHO, 2019](#ref-cmho2019)). As a result, cases are often assigned under the assumption that each requires a similar level of effort ([CMHO, 2019](#ref-cmho2019)). As a consequence, 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 set a target of 20 cases per provider for counselling services, meaning that providers with fewer than 20 cases have “room” for more, regardless of how many complex cases they have in their overall portfolio.

This “casecount” approach to determining caseloads can result in significant disparities in work, particularly for more experienced clinicians who may be assigned more complex cases due to their expertise ([CMHO, 2019](#ref-cmho2019); [King et al., 2000](#ref-king2000)). Complex cases may include those with severe behavioural challenges, high-risk family situations, or co-occurring mental health and developmental disorders, often requiring additional phone calls to coordinate with schools or other community supports, more frequent consultations with other professionals, longer or more detailed treatment plans, and extended documentation time ([CMHO, 2019](#ref-cmho2019); [King, 2009](#ref-king2009)). Without a systematic way to monitor workload beyond case counts, administrators may unknowingly overburden some staff, assuming they have the capacity for more cases when they may already be overburdened ([King, 2009](#ref-king2009)). A reliance on providers to self-report when they feel overwhelmed creates an uneven system where some clinicians silently manage unsustainable workloads, which can lead to burnout and diminished care quality ([CMHO, 2019](#ref-cmho2019); [King, 2009](#ref-king2009)).

Given this state of affairs, if there was a data-driven tool that could quantify workload based on client complexity rather than counts, it might support clinical decision-makers in a fairer distribution of work ([King, 2009](#ref-king2009); [Tran et al., 2019](#ref-tran2019)). However, the development of sophisticated data-driven predictive tools to aid in clinical decision-making has been hampered by several factors: i) a lack of resources across the public health sector generally ([Auditor General of Ontario, 2018](#Xfd2ada954160b70803c93fd53cbcc2893671edf); [CMHO, 2022](#ref-childrensmentalhealthontario2022)), ii) limits imposed by paper-based client record systems ([*Developing Caseload/Workload Guidelines for Ontario’s Child and Youth Mental Health Sector*, 2019](#ref-developi2019)), iii) disagreements on how “workload” should be defined ([*Developing Caseload/Workload Guidelines for Ontario’s Child and Youth Mental Health Sector*, 2019](#ref-developi2019)), and iv) lack of computing power and expertise in modelling complex electronic health record data in ways that remain transparent and interpretable ([Garriga et al., 2022](#ref-garriga2022); [Xiao et al., 2018](#ref-xiao2018)). However, the recent 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 possibility of leveraging EHRs with machine algorithms to improve client outcomes.

With this gap in mind, the current research proposes to explore the feasibility of estimating the time that a given client might need from a provider at intervals across the treatment timeline using information contained in the EHR with the eventual goal of testing whether such predictions provide actual added value to clinical practice. The assumption underlying the research is that historical patterns predict future mental health resource use and that such patterns can be identified in electronic mental health records (EHR) despite their inherent sparseness and systematic bias ([Garriga et al., 2022](#ref-garriga2022)).

## Case-mix History

Across health domains, particularly emergency medicine, various strategies have been employed to manage provider workload by mapping service levels to client characteristics like symptom severity or prior diagnoses ([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 patient (case) records. At the agency level, case records contain various information, including provider-level information like the number of direct and indirect hours associated with individual clients and 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 ([Johnson et al., 1998](#ref-johnson1998); [Tran et al., 2019](#ref-tran2019)). For example, a client accessing long-term counselling and therapy services might be assigned a greater 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) ([*Developing Caseload/Workload Guidelines for Ontario’s Child and Youth Mental Health Sector*, 2019](#ref-developi2019); [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 at all. For instance, a youth reporting thoughts of suicide or other self-harming behaviour will likely index higher than a youth reporting problems remaining focused in school ([CMHO, 2019](#ref-cmho2019)).

Case-mix algorithms are typically conceptual, rules-based frameworks that rely on predefined factors known or hypothesized to affect client care needs ([Tran et al., 2019](#ref-tran2019)). These frameworks are informed by clinical expertise, existing research, or policy guidelines and often use well-defined variables such as demographic characteristics, diagnoses, or treatment types. In contrast, data-driven frameworks employ empirical analysis, leveraging statistical or machine learning techniques to identify patterns and groupings in client populations without relying on prior assumptions ([Garriga et al., 2022](#ref-garriga2022); [Martin et al., 2020](#ref-martin2020); [Tran et al., 2019](#ref-tran2019)). A data-driven approach holds the potential to uncover novel insights that conceptual frameworks may overlook.

While a hybrid approach—combining conceptual expertise for clinical validity with data-driven methods for automation and insight discovery—is ideal, the complexity of modeling EHR data has limited the development of reliable data-driven frameworks, particularly in mental health service delivery ([Tran et al., 2019](#ref-tran2019)). Existing research has primarily focused on acute, inpatient hospital settings, which differ substantially from community-based outpatient care ([Aminizadeh et al., 2023](#ref-aminizadeh2023); [Garriga et al., 2022](#ref-garriga2022); [Tran et al., 2019](#ref-tran2019)). In inpatient settings, conditions often have clear diagnostic criteria and predictable recovery trajectories, such as the relatively fixed timeline and treatment protocol for a broken arm. In contrast, recovery from mental health conditions like anxiety or depression is inherently more subjective, making modelling these data significantly more challenging ([**garriga2023?**](#ref-garriga2023)).

The challenges of modelling electronic mental health data are underscored by the limited body of research addressing this problem despite its urgency ([Tran et al., 2019](#ref-tran2019)). A 2019 scoping review of case-mix literature in community-based mental health care identified only one study that employed data-driven methods to predict mental health care resource needs in children and youth populations ([Martin et al., 2020](#ref-martin2020); [Tran et al., 2019](#ref-tran2019)). That study analyzed 4,573 client records from 11 UK outpatient CYMH agencies, comparing a conceptual ‘clinical-judgement’ framework to cluster analysis and negative binomial regression to predict the number of appointments a client would attend during treatment ([Martin et al., 2020](#ref-martin2020)). While the data-driven classification did as well as the conceptual classification, the researchers suggest that data quality issues (systematic errors introduced by data entry or subjective ratings) and omission of important individual-level factors that were not contained in the EHR impacted the accuracy of their models ([Martin et al., 2020](#ref-martin2020)).

In a related cohort, researchers attempted to predict the workload associated with client characteristics at a community-based mental health center for the elderly, aiming to develop a more accurate representation of workload than simple case counts could provide ([Baillon et al., 2009](#ref-baillon2009)). Using an eight-item, self-designed Case Weighting Scale (CWS), they identified factors that staff perceived as contributing to time demands. After an initial assessment, clinicians would complete the CWS for each client, assigning scores based on factors such as family support, communication difficulties or risk of harm to self or others. These scores were input into a multiple regression model, which generated an estimate of the total time that the client would need over a four-week period. The model accounted for 58% of the variance in time spent on client-related work, which they considered a success. However, the sample size of only 87 cases raises concerns about the model’s generalizability and accuracy ([Baillon et al., 2009](#ref-baillon2009)). Additionally, inter-rater and re-rater reliability results suggested that the assessments, whether derived from client self-reports or clinicians’ professional opinions, did not consistently align with the time required for client care ([Baillon et al., 2009](#ref-baillon2009)). Nevertheless, the study does provide a basis for understanding how client characteristics might be leveraged to predict workload in mental health care settings–particularly with more sophisticated models.

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

Building on the limitations of traditional approaches like regression-based models in the Case Weighting Scale (CWS) study, machine learning (ML) offers a promising alternative for predicting mental health resource needs. Unlike conventional methods, ML algorithms learn directly from data without prior programming and are equipped to handle the high-dimensional nature of EHRs making them well-suited for mapping complex relationships between client features, such as depression scores or prior no-shows with outcomes like weekly service hours ([An et al., 2023a](#ref-an2023a); [Chen et al., 2023](#ref-chen2023)). Supervised ML models, aim to optimize a function f(x) that predicts an outcome Y (e.g., hours per week) from input features X, minimizing the difference between predictions and actual data. ML’s ability to uncover patterns in messy data presents a clear advantage for addressing the challenges of modelling client characteristics to predict workload ([An et al., 2023b](#ref-an2023); [Chen et al., 2023](#ref-chen2023)).

Within the mental health domain, ML has mainly been used to predict specific events like substance relapse ([Kinreich et al., 2021](#ref-kinreich2021)), self-harm, and suicide risk ([Simon et al., 2018](#ref-simon2018); [Walsh et al., 2017](#ref-walsh2017)). For example, Kinreich et al. ([2021](#ref-kinreich2021)) used ML to predict a change in drinking behaviour in a population diagnosed with alcohol use disorder (AUD). Combining features like brain connectivity, genetic risk scores and demographic information like age, they achieved 86% accuracy in identifying patients whose AUD had gone into remission, enabling clinicians to provide targeted interventions such as additional counselling sessions or closer monitoring ([Kinreich et al., 2021](#ref-kinreich2021)). Another study leveraged ML to monitor patient records and predict crisis relapse in 28-day windows based on EHR data([Garriga et al., 2022](#ref-garriga2022)). The top performing XGBoost model correctly differentiated those at risk from those not at risk for crisis relapse about 80% of the time ([Garriga et al., 2022](#ref-garriga2022)), and in a subsequent post-hoc case study, clinicians rated the predictions as useful for managing patient care in 64% of cases, reporting the estimates helped prioritize patients more effectively, potentially preventing crises ([Garriga et al., 2022](#ref-garriga2022)). Though the authors did not model resource use directly as we hope to do, ‘crisis risk’ served as a proxy for work. By predicting crises, they aimed to anticipate increased resource demand, allowing for better-informed case prioritization and management. Together, these examples demonstrate the utility of ML in identifying high-risk situations, highlighting its potential to enhance resource planning and improve care delivery in mental health settings.

## 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 provider 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 like provider-workload and that these patterns are identifiable in electronic mental health records (EHR).

To test these assumptions, we will analyze a retrospective, deidentified dataset from a large child and youth mental health (CYMH) agency in Ontario, Canada, encompassing data from clients served between 2019 and early 2024. Although largely exploratory, the study will be guided by several hypotheses. First, as informed by Garriga et al. ([2022](#ref-garriga2022)), we hypothesize that workload prediction will be weakest early in the client journey when available EHR data is limited to intake screener results and basic demographic information. However, as more data accumulates over the course of treatment—such as session attendance and crisis events—we anticipate prediction accuracy will significantly improve.

Consistent with Garriga et al. ([2022](#ref-garriga2022))’s, we expect that for new clients, factors such as a lack of family support and risk of harm to self or others will most strongly predict provider hours needed. For known clients, we hypothesize that time-based factors, such as the frequency of no-shows and the number of crisis events, will be more predictive of workload demands.

Finally, we expect that the winning machine learning algorithm will outperform a baseline model that will be designed to reflect the way agencies typically estimate resource needs today. This baseline model will rely on the conceptual approach often used in practice, where resource allocation is based on the type of service a client is accessing (e.g., counselling and therapy services being assigned greater weight than brief interventions) ([CMHO, 2019](#ref-cmho2019)). By comparing these approaches, the study aims to evaluate the extent to which data-driven machine-learning models can enhance workload prediction in CYMH settings.

# Methodology

## Overview

This study aims to estimate the weekly provider-hours needed (direct and indirect service) at regular stages in the client journey using machine learning predictive models. The analysis will utilize a retrospective dataset from Compass Child and Youth Family Services, the largest CYMH agency in northern Ontario. Compass serves a culturally and socially diverse population of children, youth, and families, making it a representative setting for this study.

## Data Set

The dataset will include de-identified client records with completed initial intake assessments for clients active between January 1, 2019, and December 31, 2024. Only cases with a completed initial screener will be included to ensure the availability of baseline data for generating meaningful predictions. 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. The de-identified data will include approximately 6000 EHRs containing hundreds of datapoints such as demographic information, referrals, diagnoses, risk and well-being assessments and crisis events for all outpatients. All variables left over after initial variable reduction will be included in the final report. For an overview of the data flow from raw electronic health records (EHRs) to the derived weekly features used in the predictive model structure, see [Figure 1](#fig-datastructure).

Figure 1

Data Flow Pipeline

*Note*. Data flow from raw client records to the derived features used in the predictive model. The top section represents the raw data structure containing rows of client-specific information, including dates, programs, contact types, and contact durations. The middle section visualizes a sample client timeline, mapping key events such as assessment, no-shows, face-to-face contacts, and discharges, which are stored in the EHR. The bottom section shows the weekly aggregate feature set created from these events, with features such as days since last contact and direct hours that were logged for that case in the week prior. The weekly aggregates will be used for model selection and training to predict weekly workload (e.g., weekly caseweight)

## Data Security

Given the sensitivity of mental health data, strict privacy and security measures will be enforced throughout the research process. Necessary ethical approvals will be obtained from relevant ethics boards, including both Compass Child and Youth Family Services and Laurentian University’s institutional review board. An exemption for the use of de-identified data will also be required from both institutions.

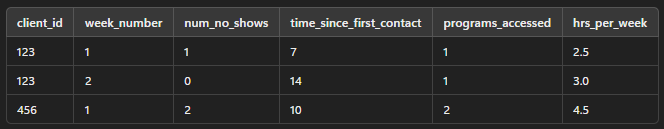
De-identified clinical data will be extracted from Compass’s electronic health information system, which is maintained by the agency. The data will be de-identified at extraction using the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor Method ([OCR, 2012](#ref-rightsocr2012)). This process ensures that all directly identifying information, such as names, addresses, birth dates, and postal codes, is removed. Additionally, unique client identification codes will be encrypted using a hashing system to prevent re-identification.

To further enhance security, the dataset will remain under the custody of Compass at all times. Data analysis will be conducted solely by the principal researcher on a password-protected machine belonging to Compass. Model results, summary statistics, and visualizations will only include aggregate metrics, focusing on predictor and model performance. No individual scores or identifiers linked to clients or small subgroups will be reported. Approval from Compass will be obtained before any findings are disseminated in external reports or presentations.

## Data Preprocessing

After de-identification, data preprocessing will include cleaning, joining data frames and handling missing values. Decisions regarding missing data will be made on a case-by-case basis, with details on imputation or exclusion documented in the final report. Any data normalization procedures will also be reported. To ensure reproducibility, the Python data scripts used for preprocessing will be publicly available.

All data points will include an associated date and time, reflecting the moment a specific event or assessment occurred. These timestamps will guide the aggregation of each client’s case records into weekly evenly spaced time series for each client, spanning their first interaction with Compass to the last (see [Figure 1](#fig-datastructure)). Features and labels for each week will be computed at the start of the week from data that was aggregated the week before, ensuring temporal consistency and avoiding data leakage. Additionally, static data prone to change over time (e.g., postal code or school board information) will be excluded to mitigate the risk of retrospective leakage. Retrospective data leakage occurs when information from the future (relative to the prediction point in time) inadvertently influences the model during training or evaluation. This typically happens in retrospective studies where datasets contain time-stamped records, and the temporal order of events is not carefully maintained during data preprocessing or feature engineering.



## Data Splitting and cross-validation

To maintain temporal consistency and maximize the generalizability of the models, the plan is to conduct a time-based 80/10/10 split for training, validation, and testing with careful thought to seasonal aspects of our data. Typically, there are fewer clients accessing service in summer months compared to months that school is in. For this reason, utilizing only a half years data for a testing would risk influencing predictions. Data splitting will be based on chronological order, as follows:

Training Data: January 2019 to March 2023 - 79.69% Validation Data: April 2023 to September 2023 - 9.38% Test Data: October 2023 to April 2024. - 10.94%

Data from the first six months of the COVID-19 pandemic may need to be excluded, depending on its irregularity and impact on service delivery. This will be addressed during data cleaning, with details reported in the final documentation.

To tune model parameters and ensure robust evaluation, we will use time-based cross-validation. Cross-validation is a method used to assess how well a model is likely to perform on unseen data. In Cross-validation the training data is divided into sequential, time-based subsets, or “folds,” preserving the chronological order of the data. For each fold, the model parameters will be tuned on earlier time periods and tested on later ones, simulating real-world prediction scenarios where past data is used to forecast future outcomes. “Tuning model parameters” involves adjusting **hyperparameters**, which are internal settings that control how the model learns from the data. Examples include the depth of a decision tree, the number of trees in a random forest, or the learning rate in a neural network. The goal is to find the combination of hyperparameters that minimizes the error between the model’s predictions and the true values. This ensures that the model’s generalizability to new, unseen data has been thoroughly tested, while still accounting for the temporal nature of the dataset.

The test set will act as an unseen control to evaluate the final models’ performance at the very end after training and tuning. It will remain untouched during model development to provide an estimation of how the models will perform in real-world scenarios. 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.

## Feature Generation (Independent Variables)

Features will be extracted from a possible set of approximately 400 variables. A complete list of proposed feature groupings and variables is provided in [Table 1](#tbl-predictors). Following the methodology outlined in Garriga et al. ([2022](#ref-garriga2022)), feature extraction will be categorized into six main types:

**Static or Semi-Static Features.** Demographic data will be represented as fixed values for each case. Age will be treated as a special case, recalculated annually to reflect changes over time.

**Diagnostic Features.** Each client will be assigned their most recent valid diagnosis if any (e.g., developmental disability, psychological disorder, or “undiagnosed”). Diagnoses will be grouped by category, using the latest valid entry up to the end of the training period to prevent data leakage. Any classification codes generated for these features will be documented in the final report.

**EHR Weekly Aggregations.** Weekly records of client-agency interactions will be aggregated for each client. These aggregated features will include counts of interaction types (e.g., appointments, no-shows) and one-hot encoded variables indicating whether a specific event occurred within the week. For one-hot encoding, a value of 1 indicates the event occurred, while 0 indicates it did not.

**Time-Elapsed Features.** For each event type and week, a feature will record the number of weeks since the last occurrence of the event. If the event has never occurred up to that point, the feature will be set to NA.

**Last Crisis Episode Descriptors.** Details from the most recent crisis episode (e.g., type, severity, resolution) will be used to create features for subsequent weeks until the next crisis occurs. If no crisis has occurred, the feature will be set to NA.

**Last Assessment Descriptors.** For each assessment item, features will be created based on the most recent assessment data, with values decaying over time to reflect diminishing relevance. This decay will apply until the next assessment occurs. All clients will have at least one assessment to ensure inclusion in the study.

**Status Features.** For records with a start and end date (e.g., program intake and discharge), features will assign values (or categories) corresponding to the active weeks. For weeks where the record is not applicable, the feature will be set to NA.

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

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

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 |  |  |  |

## Target Generation (Dependent Variable)

The caseweight prediction task will involve two modeling approaches: a continuous regression problem to estimate weekly provider hours and a classification problem to categorize workload intensity into low, medium, and high levels. Examining both approaches allows for flexibility in how predictions are used in practice ([Wang et al., 2021](#ref-wang2021)). The continuous regression model provides precise estimates of weekly hours, which are valuable for detailed planning and resource allocation. In contrast, the classification model simplifies workload prediction into actionable categories, which may be more practical for agencies to integrate into decision-making workflows, especially in contexts where exact estimates are less critical or harder to act on ([Wang et al., 2021](#ref-wang2021)).

Predictions will be generated weekly, with the model estimating the average weekly provider hours required for the upcoming 28 days using information from weeks prior. To support periodic updates, a rolling window approach will be applied, incorporating newly available data (or the absence of data) at the beginning of each week. This approach, commonly used in real-time predictive systems, allows for continuous refinement of predictions as additional information becomes available ([Garriga et al., 2022](#ref-garriga2022)).

The target variable for the regression task will be constructed by aggregating client-related direct and indirect hours logged by clinicians every Friday. These hours will be summed at the weekly level, corresponding to the feature engineering timeline, and aligned with the time recorded prior to each prediction week to prevent data leakage. We will also examine the stability and reliability of the target measure in two forms: the combined total of direct and indirect hours and the number of direct hours on its own, which may be a more stable measure of client-related work than non-direct hours which clinicians may not log consistently.

## Model Selection

A range of supervised machine learning algorithms were selected to address both regression (continuous provider hours) and classification (categories of provider hours) tasks. Models were selected based on how well-suited they are to handling high-dimensional, tabular datasets like electronic health records (EHRs).

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees during training and outputs either the most common classifications or the average predictions from individual trees. RF was chosen for its ability to handle large datasets with numerous features, manage missing data effectively, and capture complex, non-linear relationships. Its built-in feature importance metrics also enhance interpretability, making it a strong candidate for understanding which variables drive predictions.

XGBoost, a highly efficient implementation of gradient boosting machines (GBMs), was selected due to its superior predictive accuracy, scalability, and ability to handle sparse datasets with missing values. Gradient boosting combines weak learners (typically decision trees) iteratively, optimizing for residual errors at each step to minimize a specified loss function. XGBoost’s regularization techniques, such as shrinkage and column sampling, help prevent overfitting, while its computational efficiency makes it well-suited for large datasets .

Feed-forward neural networks (FNNs), a class of deep learning models, were included for their flexibility in modeling complex non-linear interactions among variables. FNNs consist of interconnected layers of nodes where each node applies an activation function to transform input data. These networks are particularly useful when relationships between variables are intricate and not easily captured by tree-based methods.

Recurrent neural networks (RNNs) were added to leverage the sequential nature of the dataset. Unlike FNNs, RNNs include recurrent connections that allow the model to retain information about previous inputs, enabling it to capture temporal dependencies in time-series data. This makes RNNs particularly well-suited for tasks where past events influence future outcomes, such as predicting changes in weekly provider workload based on prior patterns.

Furthermore, a baseline model will be implemented to replicate how new clients are typically assigned in practices without sophisticated casemix algorithms for comparison. The baseline will rely on a simplified feature set, containing the programming they are accessing and their age. By evaluating all of the models against this baseline, we can better estimate whether machine learning approaches offer any improvement over traditional methods of estimating provider workload.

Each model will be trained on the same training set and evaluated using identical cross-validation splits to ensure consistency in comparisons. Hyperparameter optimization will be conducted for all algorithms, with 100 trials per model, focusing on minimizing Mean Absolute Error (MAE) for regression tasks and maximizing the Area Under the Receiver Operating Characteristic Curve (AUROC) for classification tasks. This process will ensure that the models are fine-tuned to achieve optimal performance.

All models will be compared against the baseline and one another to assess relative performance across both regression and classification tasks. Detailed hyperparameter search spaces and tuning procedures will be documented in the supplementary materials. ([Salditt et al., 2023](#ref-salditt2023); [Sheetal et al., 2023](#ref-sheetal2023)).

## 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

Python will be used as the primary programming language for model development and evaluation with support from R Statistical Software ([Van Rossum & Drake, 1995](#ref-vanrossum1995)). Quarto Markdown will facilitate documentation and ensure reproducibility, with all workflows executed within the Positron IDE environment ([**positron?**](#ref-positron)). Positron is a next-generation data science integrated development environment (IDE) developed by Posit PBC. It is built on Code OSS and designed to support multiple programming languages, including R and Python, providing an extensible and familiar environment for reproducible authoring and publishing.

# 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

Aminizadeh, S., Heidari, A., Toumaj, S., Darbandi, M., Navimipour, N. J., Rezaei, M., Talebi, S., Azad, P., & Unal, M. (2023). The applications of machine learning techniques in medical data processing based on distributed computing and the internet of things. *Computer Methods and Programs in Biomedicine*, *241*, 107745. <https://doi.org/10.1016/j.cmpb.2023.107745>

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

An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023a). 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 of Ontario. (2016). *Office of the Auditor General Annual Report* (pp. 110–147). Ministry of Children; Youth Services. <https://www.auditor.on.ca/en/content/annualreports/arreports/en16/v1_301en16.pdf>

Auditor General of Ontario. (2018). *Child and Youth Mental Health Follow-Up Report* (pp. 14–28). Ministry of Children, Community; Social Services. <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: Report on waitlists and wait times for child and youth mental health care in ontario*. Children’s Mental Health Ontario. <https://cmho.org/wp-content/uploads/CMHO-Report-WaitTimes-2020.pdf>

CMHO. (2022). *Addressing urgent workforce challenges in child and youth mental health*. Children’s Mental Health Ontario. <https://cmho.org/wp-content/uploads/CMHO-Workforce-FINAL.pdf>

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>

*Developing caseload/workload guidelines for Ontario’s child and youth mental health sector*. (2019).

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*. Canadian Institute for Health Information. <https://secure.cihi.ca/free_products/Case_Mix_Tools_e.pdf>

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>

Kinreich, S., McCutcheon, V. V., Aliev, F., Meyers, J. L., Kamarajan, C., Pandey, A. K., Chorlian, D. B., Zhang, J., Kuang, W., Pandey, G., Viteri, S. S.-S. de, Francis, M. W., Chan, G., Bourdon, J. L., Dick, D. M., Anokhin, A. P., Bauer, L., Hesselbrock, V., Schuckit, M. A., … Porjesz, B. (2021). Predicting alcohol use disorder remission: a longitudinal multimodal multi-featured machine learning approach. *Translational Psychiatry*, *11*(1), 1–10. <https://doi.org/10.1038/s41398-021-01281-2>

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

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>

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*. U.S. Department of Health; Human Services. <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>

Simon, G. E., Johnson, E., Lawrence, J. M., Rossom, R. C., Ahmedani, B., Lynch, F. L., Beck, A., Waitzfelder, B., Ziebell, R., Penfold, R. B., & Shortreed, S. M. (2018). Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records. *The American Journal of Psychiatry*, *175*(10), 951–960. <https://doi.org/10.1176/appi.ajp.2018.17101167>

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.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clinical Psychological Science*, *5*(3), 457–469. <https://doi.org/10.1177/2167702617691560>

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>

WHO. (2022). *World Mental Health Report: Transforming Mental Health for All* (p. 58). World Health Organization.

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, *25*(10), 1419–1428. <https://doi.org/10.1093/jamia/ocy068>

# 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)).