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

Shauna Heron1, Michael Emond1, Luc Rousseau1, Kalpdrum Passi1, & Nicholas Schwabe2

1Department of Psychology, Laurentian University

2Compass

# 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 & writing – review & editing. *Luc Rousseau***:** supervision, writing – review & editing, & advisory committee. *Kalpdrum Passi***:** supervision, writing – review & editing, & advisory committee. *Nicholas Schwabe***:** conceptualization, writing – review & editing, & advisory committee

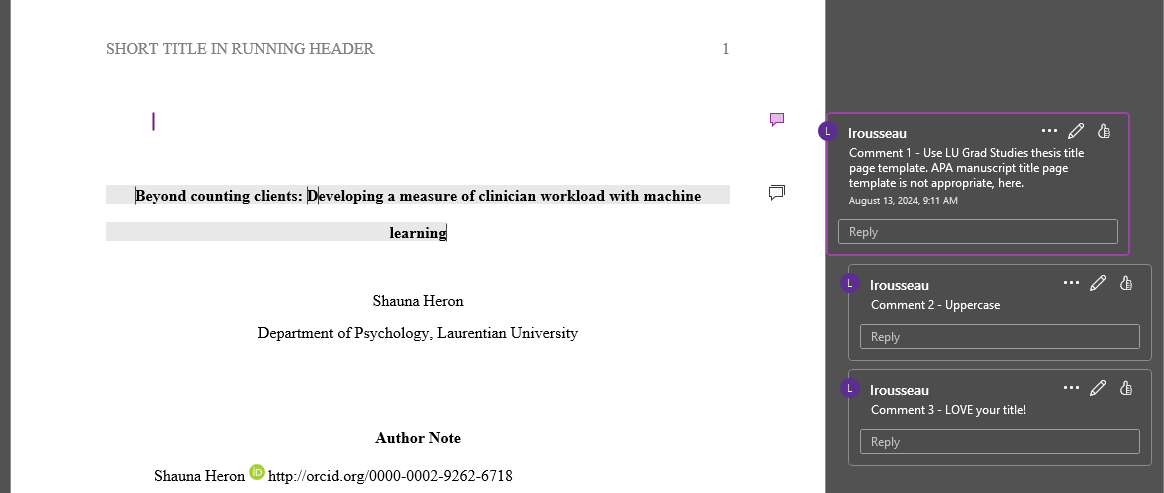
Correspondence concerning this article should be addressed to Shauna Heron, Department of Psychology, Laurentian University, Sudbury, ON, Canada, Email: sheron@laurentian.ca

# Abstract

As community based child and youth mental health service (CYMH) providers face increasing demands for services at the same time as serious workforce shortages, the ability to anticipate and optimize the workload of staff is critical to providing high quality client care that also prioritizes provider well-being ([CMHO, 2019](#ref-cmho2019); [CYMHLAC, 2019](#ref-cymhlac2019)). Yet so far, there are a lack of consistent and reliable strategies for making equitable client assignment, and research that might help us understand the various client factors that contribute to a providers workload–particularly within the Child and Youth Mental Health sector (CYMH), is small. With these gaps in mind, the goal of the current study is to evaluate the utility of machine learning models, trained on information collected at intake, to predict the amount of work an individual client might contribute to a provider’s workload. Specific objectives include: (i) identify significant predictors of client-related workload from intake assessment data; (ii) compare the predictive accuracy of several tree-based machine learning models against traditional linear models; (iii) investigate the relationship between client-related workload indicators (i.e. depression scores or referral source) and workload proxies (i.e. number of direct hours spent with a client); (iv) explore the potential for early prediction of caseweight to inform case assignment and workload management.

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

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



|  |
| --- |
| Note |
| Will be reaching out today for the grad studies thesis title page template!!  Capitalization fixed :) |

# Background

Amidst a growing demand for child and youth mental health services, human resource challenges have been identified as a significant area of concern by Children’s Mental Health Ontario (CMHO) member agencies ([Children’s Mental Health Ontario, 2022](#ref-childrensmentalhealthontario2022)). This is a concern, as without an adequate workforce, children, youth and their families experience longer wait times and gaps in service that ultimately impact access to treatment and mental health outcomes.

Illustratively, a 2020 survey of CMHO agencies found that up to 28,000 children and youth in Ontario were waiting as long as 2.5 years for services ([Children’s Mental Health Ontario, 2020](#ref-childrensmentalhealthontario2020)). At the same time, 83% of agencies reported they were experiencing staffing shortages, with 59% of those vacancies being direct-service clinical staff ([Children’s Mental Health Ontario, 2020](#ref-childrensmentalhealthontario2020)).

Clinical staff are qualified mental health professionals like psychologists, psychometrics, psychotherapists and social workers who work in an interdisciplinary team to provide mental health treatment, services and assessment and to children, youth and their families.

According to the CMHO report, it has become increasingly challenging to attract and retain mental health service providers in the community CYMH sector, citing pressure from the COVID pandemic coupled with an increasing severity of mental health concerns that exacerbate problems that existed before the pandemic like wage disparities between public and private sectors and a lack of providers with the required profesional training or designation to support the complexity of needs. Staff shortages like these lead to existing staff having to manage higher client volumes containing more complex cases, which can lead to increased stress and perpetuates a cycle of provider burnout and absenteeism.

According to several audits, follow ups and recommendations by Ontario’s Auditor General (AG) starting in 2008, a central problem that impedes the ability of agencies to efficiently meet this challenge, is determining reasonable provider to client workload ratios. While agencies have traditionally relied on simple case counts to determine who may or may not have room for another client, the AG reports suggest that counts alone do not convey the complexity of individual need nor the work and time needed to support that need, making it an unreliable strategy for equitable client assignment. However, repeated recommendations for agencies to develop reasonable client to provider workload benchmarks, including subsequent follow-ups and another comprehensive 2016 audit, have produced little progress to date–underscoring the complexity of the problem ([Office of the Auditor General of Ontario, 2016](#X74ef00379ba32932a14081a0b155af9f4f12f08)).

For example, a 2018 follow-up report found that though several agencies had developed and implemented informal caseload benchmarks, in most cases agencies could not demonstrate that they were based on best practices nor were caseloads monitored for reasonableness which might help identify deviations requiring follow-up and correction.

In response to the difficulties that agencies have reported in developing these targets, CMHO, the Ministry of Children and Youth Services (Ministry) and CYMH agencies have partnered to work together towards developing caseload guidelines with a first step to focus on identifying variables that impact client-related work such as case complexity, case acuity, geography and variability in the types of core services delivered ([Office of the Auditor General of Ontario, 2016](#X74ef00379ba32932a14081a0b155af9f4f12f08)). Among other recommendations following the last collaborative 2019 report, it was stressed that case counts as a static measure, failed to reflect the varying needs and intervention levels required by different clients ([CMHO, 2019](#ref-cmho2019)).

## Caseload versus workload

It is important here to distinguish between caseload and workload. In the current paper, we refer to *caseload* as the number of active clients assigned to a provider at any given time, while *workload* is the amount of time a provider spends serving each case as well as attend other professional responsibilities such as supervision, professional development, and training ([CMHO, 2019](#ref-cmho2019)).

In the literature, high case loads have been linked with lower self-reported efficacy ([King et al., 2000](#ref-king2000)) and poorer clinical outcomes ([Lloyd & King, 2004](#ref-lloyd2004)), while smaller caseloads with better clinical outcomes and higher retention rates ([Berkel & Knies, 2015](#ref-vanberkel2015)). These findings supports earlier evidence that suggest caseloads exceeding 20 to 30 clients can lead to “reactive” case management characterized by deficiencies in service planning and family support ([Intagliata, 1982](#ref-intagliata1982); [Lloyd & King, 2004](#ref-lloyd2004)). At the same time, another line of inquiry suggests that the relationship between caseload and burnout may not be as straightforward as it seems ([King et al., 2000](#ref-king2000), [2004](#ref-king2004)). For example, King et al. ([2000](#ref-king2000)) did not find that case counts were predictive of clinician burnout, but the *mix* of cases in regards to individual client-level complexity, *was*. He posited that providers may adapt to high caseloads by simply doing less for each case, suggesting a potential ‘dose-response’ relationship between a provider’s time and their effectiveness ([King, 2009](#ref-king2009)). Importantly, King et al. ([2000](#ref-king2000)) ’s work suggests that the ratio of low to high need cases in any given caseload may be a more important predictor of provider stress than counts alone ([King, 2009](#ref-king2009)). This finding is particularly relevant to the CYMH sector where the mix of cases can vary widely in terms of complexity and acuity ([CMHO, 2019](#ref-cmho2019)).

To illustrate, consider two clinicians each with a caseload of 20 clients. Clinician A, due to their years of experience, has a greater proportion of complex clients requiring intensive weekly therapy sessions, complex case management and frequent crisis interventions, while clinician B’s clients may require simpler interventions and less frequent check-ins. Additionally, Clinician A may have the added responsibility of supervision and attending meetings which further increases their overall workload. Thus, even with the same caseload, the *workload* between them may differ significantly in regards to the intensity and nature of the services required by their clients and other professional duties.

## A measure of client-related work

Given these findings, if we had the ability to accurately predict the workload associated with each client, agencies might better manage provider caseloads. However modeling client characteristics, let alone predicting the work that is driven by them, has proved difficult so far ([CMHO, 2019](#ref-cmho2019); [King, 2009](#ref-king2009); [Tran et al., 2019](#ref-tran2019)). Despite initiatives like the Quadruple Aim Framework meant to improve health outcomes, reduce costs, and improve provider work-life balance, the CYMH sector continues to face challenges in establishing a consistent, standardized measurement system, let alone a dynamic case assignment system that can accurately reflect the demands placed on staff ([Arnetz et al., 2020](#ref-arnetz2020); [CMHO, 2019](#ref-cmho2019)). Nevertheless, the new guidelines *have* informed new policies meant to improve the tracking of client-related work which has enabled several agencies to expand efforts to understand and track workload in their own organization.

An important workload metric that was born out of the AG audit and recommendations was a policy change requiring all publicly funded CYMH service providers in Ontario to report all direct-time time spent in the service of clients ([*MOH CYMH Service Description Schedules - Central Forms Repository (CFR)*, n.d.](#ref-mohcymh)). Direct-time is defined as the number of hours spent in face-to-face interactions, phone or video-based communications, and meetings with parents and caregivers while indirect hours involve client-related tasks like documentation, telephone calls, advocacy, and consultations. The sum of direct and indirect hours amounts to the overall “work” attributable to a given client–a metric we will use in our models as an approximation of the work attributed to an individual client.

It is important to stress that the direct and indirect hours logged are only a *proxy* for client-related work, as hours alone may not convey differences in emotional or mental effort required to treat different clients nor the stress associated with a case that may factor in to the *actual* workload experienced by the provider and associated with any given client that is more difficult to measure.

[Compass](https://www.compassne.ca/), the lead child and youth mental health agency in the districts of Sudbury and Manitoulin and the proposed site of the current study, utilizes a dashboard to monitor caseloads and the associated direct and indirect time logged by service providers. While the dashboard has been a useful tool to compare caseloads between clinicians and across teams, it offers less insight into how much work is associated with each case. Moreover, it doesn’t aid in assigning *new* cases beyond indicating which providers have more “space” in their caseload than others. For this reason, the agency wondered whether intake data like age and psychological screener scores could be used to quantify client complexity with a *weight* for each case that might be used to monitor caseloads as well as inform new case assignment.

## How do we model client complexity?

Across domains, various strategies have been employed to manage provider workload, many focusing on increasing client-flow by providing different levels of service determined by presenting characteristics like symptom severity or prior diagnoses ([Johnson et al., 1998](#ref-johnson1998); [Tran et al., 2019](#ref-tran2019)). A popular model is the case-mix classification system, which assigns clients to different categories based on their expected resource use ([CMHO, 2019](#ref-cmho2019)). Case-mix classification systems usually take one of two approaches to modeling client-related work: grouping or index. Grouping systems assign people into classes in terms of their expected resource use, with each group having a specific weight attached to it, representing expected resource use (e.g., high-need versus low-need) relative to the average case in the population; Index systems on the other hand, combine different characteristics of a case to provide a continuous, numerical value which represents expected resource use relative to an average case (e.g., total hours, total length of stay or cost associated with care) ([Tran et al., 2019](#ref-tran2019)).

In reviewing the literature for the current study, we relied on a 2019 scoping review of case-mix classifications for community based mental health care as a starting point, hoping it would lead to discovering more recent work ([Tran et al., 2019](#ref-tran2019)). Unfortunately, the single case that looked at case-mix classification to predict mental health care resource use in CYMH community settings remains the only case of its kind ([Martin et al., 2020](#ref-martin2020); [Tran et al., 2019](#ref-tran2019)). In that study, researchers modeled 4573 client records from eleven UK outpatient community based child and youth mental agencies to predict the number of appointments a client might need ([Martin et al., 2020](#ref-martin2020)). Three classification methods were compared: two data driven (cluster analysis and regression trees) and one conceptual (classification informed by clinical judgement) to predict the number of appointments a client might need. Contrary to what they expected, they found the classification algorithms ability to predict accurately on new data was weak, and not significantly better than clinical judgement ([Martin et al., 2020](#ref-martin2020)). Moreover, they found little statistical evidence to support the idea that client complexity had much to do with differences in resource provision ([Martin et al., 2020](#ref-martin2020)). Researchers cited several reasons for problems explaining the variance in resource use: data quality, omission of important individual-level factors and lack of standardization of practice between providers.

In a different population, another group of researchers tried to predict the workload associated with patients at a community based mental health centre for the elderly ([Baillon et al., 2009](#ref-baillon2009)). Using an 8-item case weighting scale (CWS) that identified factors staff felt contribute to demand for staff time, they built a multiple regression model to assign different weightings to each item based on the strength of its relationship with the outcome (an estimation of time spent on each client logged over a four week period). The model was then used to predict the total time a client would utilize in a four-week period following the first appointment. Though they reported the model as a success–the model accounting for 58% of the variation in time spent on client-related work–the sample was small, consisting of only 87 cases and relied on a statistical method inappropriate for evaluating agreement between model predictions and actual observations, leaving it unclear how accurate the model actually 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)).

### Case-mix in other domains

Though there is less work in the CYMH domain, several measures of workload intensity have been developed to manage caseloads in other specialties, particularly in inpatient medical settings. For example, in general psychiatry, researchers have used factors like sociodemographics, functional ability, and caregiver and social network characteristics to predict service utilization. Though, many of these models similarly lacked adequate evaluations and would not be suitable for all client groups, Tran et al. ([2019](#ref-tran2019)), recommend that it may be useful to experiment with case-mix systems developed in other settings.

One model that the authors of the review suggested may be a good candidate for testing in community settings is Canada’s System for Classification of In-Patient Psychiatry (SCIPP) ([Perlman et al., 2013](#ref-perlman2013); [Tran et al., 2019](#ref-tran2019)). The SCIPP algorithm is a grouping methodology that sorts patients according to clinical characteristics obtained from standardized interRAI assessment data to estimate resource use ([Hirdes et al., 2020](#ref-hirdes2020); [Perlman et al., 2013](#ref-perlman2013)).

### Measuring Client-Related Work

interRAI is an international research network “dedicated to developing clinical standards across a variety of health and social service settings” that have developed a large toolkit of instruments that are used by health organizations worldwide to assess people at the point of care across a variety of domains including emergency medicine, emergency psychiatry and children and youth mental health. The resulting data are meant to be used at the agency level for quality improvement activities, benchmarking, program planning and resource planning and at the system level to compare health data across regions and provinces ([Hirdes et al., 2002](#ref-hirdes2002)). In Canada, interRAI is partnered with the Canadian Institute for Health Information (CIHI) who acts as a custodian of interRAI standards for Canada and houses and monitors the collected interRAI data. Many interRAI instruments are used across Canada and internationally, but the ChYMH represents the first assessment specifically designed for children and youth ([Stewart & Hamza, 2017](#ref-stewart2017)).

In Ontario, two instruments are most often used in the CYMH sector: the Child and Youth Mental Health Screener+ (ChYMH-S) and the more comprehensive full ChYMH and its variants. The primary use of the CHYM-S is to support decision making related to triaging, placement and service utilization while the full ChYMH and its associated Collaborative Action Plans (CAPs) are meant to assess, repond to and monitor the strengths, preferences and mental health needs of clients in in-patient and out-patient treatment. It is currently being utilized in Ontario at over 60 mental health agencies.

The full ChYMH includes over 400 items meant to build a comprehensive picture of a child/youth’s strengths, needs, functioning and areas of risk to inform care planning ([Stewart et al., 2022](#ref-stewart2022)), while the ChYMH-S is comprised of 106 items and is intended as a brief screener to identify young people who are in need of more comprehensive assessment. The screener is administered via a semi-structured interview to children and youth between 4-18 years in a variety of settings and is intended to take 15-20 minutes to complete. The current study will focus on screener data collected at intake.

The ChYMH-S includes 34 administrative and tracking items, 26 mental health indicators, 5 substance use indicators, 9 questions related to harm to self and others, 6 behaviour focused items, 1 cognitive item as well as items that look to track social relations, anxiety levels, medications, living arrangements, diagnoses, physical conditions, past interventions and current and past strengths and resilience ([Stewart et al., 2022](#ref-stewart2022)). The Depressive Severity Index, Anxiety Scale, Disruptive/Aggressive Behaviour Scale, Hyperactive/Distraction Scale and Internalizing/Externalizing scales are all included in the ChYMH-S. See table for full list of items. The instrument and the various scales and items it comtains have been tested extensively in research worldwide, demonstrating reliable face, content, construct and predictive validity [lots of detail could go here–please inform as to how much!.

Importantly, several ChYMH scales have demonstrated strong predictive validity. For example, data from over 5000 children and youth placed in psychiatric settings in Ontario found that the Agressive Behaviour Scale was predictive of multiple control interventions, while the Severity of Self Harm Scale (SOS) has proven useful in predicting admission for risk of self-harm in youth between 10-17 years. Moreover, individuals who score higher on scales like the Hyperactive/Distraction scale were more likely to have a provisional diagnosis of ADHD ([Stewart et al., 2022](#ref-stewart2022)).

Though the various scales and items of the ChyMH have demonstrated *some* predictive utility, it remains unclear how well these items might predict the actual work required to serve any given client. Within the CYMH domain we found one example of interRAI data being used to develop an algorithm to predict resource cost for children and youth with developmental disabilities with cluster analysis ([Stewart et al., 2020](#ref-stewart2020)). Though the resulting Child and Youth Resource Index (ChYRI) could only explain 30% of the variance in per diem costs for community based services ([Stewart et al., 2020](#ref-stewart2020)), the algorithm was deemed a success. Nevertheless, a lack of explanation of how the analysis was conducted, makes it unclear where the algorithm could be improved.

It is important to note, we are not interested in predicting client-related work for budgetary reasons, but instead our efforts are in response to requests from clinicians themselves to develop a resource allocation system that accounts not only for current capacity in regards to counts, but the differences in need (and thereby work) between clients so that caseloads between clinicians are more equitable and mindful of the work that is already on their plate.

### A machine learning approach to modeling case-mix

Considering the limitations and lack of transparency in regards to the SCIPP and ChYRI development, we looked to academia and specifically the medical domain which has experimented with several ways of modelling electronic health data (EHD) (e.g. interRAI scores, test results, diagnoses) to predict patient-related work for managing professional caseloads and other hospital resources. ….<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8967950/>.

One study that stood out to us, utilized…

Building on Benda et al. ([2018](#ref-benda2018))’s study, Wang et al. ([2021](#ref-wang2021)) focused on improving the underlying algorithm with various machine learning algorithms known for their robustness in modeling sparse, highly heterogeneous data features. To this end, both regression and classification algorithms were applied to model several proxies for workload (length of stay, number of events, and density of events) and high versus low demand patients. The accuracy of prediction for low versus high length of stay was 70% with information from the first hour, 73% from the first two hours and 83% with data from the entire visit. Though the domain and temporal aspects were different (hospital stay versus outpatient mental health treatment), their methodology demonstrates the potential of leveraging machine learning techniques to predict client-related workload from information collected at intake.

## Themes in the literature

In planning our approach to modeling client-related workload, several themes emerged in the literature that we hope to address in our research.

First the most commonly cited problem in …. was the difficulties in modeling EMHD.

First, we address a gap in case-mix research by focusing on the specific needs of younger people served by mental health providers in a community mental health outpatient setting. Our study targets child and youth mental health services specifically, addressing the unique requirements of this demographic.

We also intend to tackle the issue of inconsistent outcome measures. Previous studies often used flat proxy measures for workload, such as length of stay or number of appointments which fails to capture the variance in work intensity throughout the treatment period ([CMHO, 2019](#ref-cmho2019); [Martin et al., 2020](#ref-martin2020)). To address this, we will calculate a resource use measure on a per-diem basis, predicting resource use (direct and indirect time logged) per week ([Wang et al., 2021](#ref-wang2021)). This method, supported by the System for Classification of In-Patient Psychiatry (SCIPP) developed in Canada, might better reflect the intensity of work than flat counts ([CMHO, 2019](#ref-cmho2019)).

Additionally, we will address several limitations related to client-related workload indicators. In many studies, variables like gender or race are used to categorize client need, which can raise concerns about fairness and predictor bias. While race and gender may correlate with resource use, these correlations can be confounded by systemic marginalization factors that may increase the risk for mental health concerns ([Gaines et al., 2003](#ref-gaines2003); [Tran et al., 2019](#ref-tran2019)). To avoid perpetuating this marginalization, we will focus on variables that drive resource use directly such as scale scores or symptom ratings, and exclude variables thay may indicate race, culture or ethnicity. Following Tran et al. ([2019](#ref-tran2019))’s recommendation to favour direct measures of client need, we plan to use interRAI screener+ items and scores specifically ([Hirdes et al., 2020](#ref-hirdes2020)).

Furthermore, to avoid some of the challenges faced in prior research when including both client and provider-side drivers of work, as indicated in their casemix recommendations that suggest including provider-side variables (i.e. years of experience or preferred modality) in models risks reinforcing systemic unfairness in case distribution, we will focus solely on client-side variables to predict workload.

Model evaluation and metrics were another critical area of weakness across the literature ([Tran et al., 2019](#ref-tran2019)). Few studies employed robust cross-validation methods and even fewer tested their models on unseen data ([Tran et al., 2019](#ref-tran2019)). We will utilize cross-validation folds within our training set for model building and hold back a test set of unseen data for final model evaluation, ensuring a control for determining model accuracy. We will also split our data group-wise ensuring that clients in the training set are not including in the test set. This will ensure that we have a more robust measure of how well our model would perform on “unseen” clients. We will also clearly outline our choice in metrics used to evaluate model performance and document all R code for reproducibility (see Methods).

Finally, issues relating to the complexities involved in modeling electronic health data was evident across all studies we looked at. We attempt to address this issue by utilizing machine learning (ML) algorithms which are better-suited to handle the high-dimensionality and heterogeneity of electronic mental health (EMH) data ([Joseph et al., 2023](#ref-joseph2023)). EMH data poses unique challenges such as hidden clustering, non-independence of observations, missing data and outliers. Consequently, a significant amount of data wrangling and variable culling is necessary for traditional analysis, often resulting in significant data loss and models with limited generalizability. In contrast, ML methods, such as random forests, gradient boosting, and neural networks, can better manage non-independent observations, non-normal distributions, and multicollinearity among predictor variables–albeit with varying levels of interpretability ([Zeleke et al., 2023](#ref-zeleke2023)).

### What is machine learning?

Machine learning is a branch of artificial intelligence where computers are trained to perform tasks by identifying patterns within data instead of relying on explicitly programmed instructions. 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 that maps input features to an output such that the predicted outcomes are as close as possible to the true 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.

This capability makes ML particularly suitable for analyzing electronic health data, where the data may be sparse, exhibit non-linearity, missing values, and complex interdependencies among variables ([An et al., 2023](#ref-an2023); [Chen et al., 2023](#ref-chen2023)). Unlike traditional methods that might only confirm specific linear relationships, machine learning algorithms can identify high-dimensional interactions and non-obvious patterns. Techniques such as regularization and cross-validation further enhance the robustness of these models, maximizing their ability to generalize well to unseen data ([Chen et al., 2023](#ref-chen2023)).

Moreover, ML excels in both predictive accuracy and discovering new relationships that might be validated in subsequent studies. By leveraging algorithms that optimize for predictive performance, such as gradient boosting or ensemble methods, researchers can uncover patterns with the potential to drive new hypotheses. This *abductive* reasoning, informed by the patterns identified in ML models, may advance research by providing new directions for future investigation ([Sheetal et al., 2023](#ref-sheetal2023)).

## Purpose

In summary, the current research aims to extend previous work in developing a measure of client-related work while addressing the recommendations and limitations of the same studies. We propose a quasi-experimental data science framework to test our hypotheses and evaluate the models’ predictive performance . Models will be trained on one set of data, while another set of “unseen” future data will act as a control to evaluate model performance.

Our goal is to better understand the relationship between client-related workload indicators, workload proxies, and client complexity in the CYMH sector. Additionally, we will explore the potential of early predictions to improve client assignment, balance workload, and potentially identify overloaded providers. Given the pressing need in the CYMH sector for a case-management system that accurately and fairly assesses caseload we feel our proposed study can positively contribute to the findings.

## Hypotheses

The study will be guided by several hypotheses.

First, given past research indicating machine learning techniques are better able to capture complexities and patterns within EMH data, we predict that tree-based machine learning models will be better predictors of client-related work across all workload proxies than linear regression or cluster analysis.

Second, we anticipate that mental health acuity features such as depression scores at intake, will be more significant predictors of client-related work than age alone. This hypothesis stems from existing literature indicating that these variables are critical determinants of resource use ([Perlman et al., 2013](#ref-perlman2013); [Tran et al., 2019](#ref-tran2019)).

Third, we predict that workload indicators (independent variables) will explain a higher proportion of the variance in flat proxies (total hours in service or total days in service) than dynamic proxies (caseweight: total hours divided by the number of weeks). Though we believe a ratio better represents the intensity and frequency of demands placed on providers than flat counts alone, these scores may compound worker-level effects that the models will have difficulty accounting for ([Wang et al., 2021](#ref-wang2021)).

# Methodology

All clients who completed an intake assessment at Compass Child and Youth Family Services between January 1, 2019 and December 31, 2023 will be screened for inclusion. Compass is a large, publicly funded mental health agency for youth and families located in northern Ontario, Canada. The general flow of new clients into Compass is illustrated in [Figure 1](#fig-client-selection).

Considering we want to predict the amount of resource use a client will require based on information collected at intake (demographics information and interRAI screener scores), only clients whose initial screening resulted in referral to Counselling and Therapy (CT) will be included in the analysis. Predicting who may or may not need Brief versus CT services is a prediction task for another study. Final counts after screening will be reported and added to the flowchart after analysis.

## Data Security

Given the sensitivity of mental health data, ensuring data privacy and security by obtaining necessary ethical approvals and maintaining transparency throughout the research process will be strictly enforced. Necessary approvals from relevant ethics boards will be obtained. An exemption must be granted by both the agency and Laurentian’s institutional review board for the use of de-identified data.

Deidentified clinical data will be acquired from an electronic health information system belonging to Compass. The EHR database is maintained by the institution. 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, birthdates, full postal codes, clinical notes and any other directly identifying information will be stripped from the dataset before any analyses begins. As an added precaution, unique client identification codes will be encrypted with a hashing system that makes it near 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 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.

## Procedure

The following steps outline the proposed experimental process which will consist of four, broad phases: 1) data collection, preprocessing and exploration; 2) identifying a list of workload proxies (output/dependent variables) that could be used as stand-ins for actual workload; 3) identifying and extracting indicators of workload (i.e. independent variables/features) that could be used to model our proxies; 3) modeling the relationship between the indicators and proxies with algorithms of varying complexity; and 4) evaluating and comparing the models’ performance on a set of unseen data (see [Figure 3](#fig-procedure-flow)).

### Data Collection & Preprocessing

After deidentification, data preprocessing will involve cleaning, joining dataframes, handling missing values, and narrowing items to only information available at intake. All decisions we make in regard to missing data, data normalization or any other changes will be reported in our final paper. Moreover, the final report will include the R code necessary to replicate these steps .

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)).

#### Independent Variables (features/indicators of caseweight).

The main criterion for inclusion will be the variable’s availability in the electronic health record (EHR) system at intake. Information collectedincludes items from the interRAI Screener+ and full ChYMH screener. The interRAI ChYMH is a clinician-rated tool that is completed based on a semi-structured clinical interview [ADD BETTER DESCRIPTION HERE] and includes several scale scores, items as well as demographics information ([Stewart & Hamza, 2017](#ref-stewart2017)). A full list of all variables will be included in the final report.

Figure 1

Clinical pathway

*Note*. Flow chart of client selection process. Clients who complete an intake assessment will be screened for inclusion. Only clients referred to Counselling and Therapy (CT) services will be included in the analysis. Predictions will include: client-related work at follow-up assessment and client-related work at end of treatment.

ADD TABLE OF POTENTIAL INDICATOR VARIABLES

#### Dependent Variables (workload proxies).

Informed by Wang et al. ([2021](#ref-wang2021)), we intend to measure the influence of our workload indicators on Length of Service (weeks), Hours of Direct Time, Indirect Time and combined Direct and Indirect Time as well as a case density score (caseweight) calculated by dividing the number of hours (direct and indirect) spent with a client by the number of weeks in service (see [Equation 1](#eq-caseweight)).

Density scores should reflect the frequency/intensity of demands better than a flat count of weeks in service or length of stay ([Wang et al., 2021](#ref-wang2021)). As such, we predict it will be a better indicator of work intensity, tempo and complexity. Length of Service (LoS) will be calculated as the number of days from assessment to the point at which they were discharged. LoS will not include the time a client waited for service on a waitlist as this reflects a provider-side driver of work that would bias our model. See [Figure 2](#fig-caseweightmodel) for a visualization of the relationship between variables.

Figure 2

Modeling Caseweight–client-related work

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

ADD TABLE OF WORKLOAD PROXIES

If time and the data structure permits, we will attempt to construct a third “event count” proxy that tallies the total number of appointments, phone calls, and other case events associated with each case then divides by the number of weeks in service. Time and approval permitting, we would also like to consult with clinical managers to obtain a list of indicators that, in their expert opinion, would signal a client who could demand more work during the screening interview—these will be used to externally validate the variable choices of the models.

We will model our proxies as continuous “caseweights” and as classes split into two and three evenly distributed classes separately for classification. For example, a two class outcome (low versus high resource demand) will be split at the median.

In practice, regression will be used to model the number of hours per week a client will utilize across their episode of care, while binary classification will predict whether the stay is shorter or longer than the median for that program. Informed by Wang et al. ([2021](#ref-wang2021)), we will first include data collected over the entire program length to determine whether the indicators have any utility in modelling the workload proxies. Then, we will use each client’s first assessment to predict case weight at follow up assessment (typically 3 months later). This will allow us to test whether a workload prediction in the earliest stages of a visit is even feasible.

### Data Splitting

To ensure the robustness and generalizability of our models, data will be randomly split by groups (client ID). This method ensures that clients in the training set, which is used to train the models, are not included in the test set, thus preventing data leakage and providing a more unbiased evaluation of model performance ([Figure 3](#fig-procedure-flow)).

The training set will be used to train the models using 10-fold group cross-validation. In group cross-validation, the training data is divided into 10 subsets, or “folds,” while preserving the grouping structure. This technique helps tune the models by iteratively training on nine folds and validating on the remaining one, ensuring each group is used for validation exactly once. Group cross-validation is particularly beneficial when dealing with repeated measurement data, as it maintains the integrity of the group structure and prevents information from the test folds from leaking into the training process.

The test set will act as a control group to evaluate the models’ performance on unseen data at the very end of the training process. By keeping the test set separate and untouched during training, we ensure that our final evaluation provides a better estimation of how the models will perform in practice. This step is crucial for assessing the models’ generalizability and for identifying any overfitting that may have occurred during training.

Figure 3

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.

### Model Selection

We propose testing both regression and classification algorithms to model our data. We plan to use the following supervised machine-learning algorithms: i) Random Forest (RF) for its ability to handle large datasets with high dimensionality; ii) XGBoost, known for its high performance and predictive accuracy on tabular datasets; iii) lasso regression to manage and select relevant predictors while handling high multicollinearity (this list may grow). The supervised machine-learning classification algorithms will be trained using the TidyModels suite of packages in R Studio. These algorithms were chosen based on their success modeling similarly complex, tabular data types ([Salditt et al., 2023](#ref-salditt2023); [Sheetal et al., 2023](#ref-sheetal2023)).

### Validation and Testing

All trained models will be statistically evaluated on the test set using the following performance metrics: i) Mean Absolute Error (MAE), and ii) Root Mean Squared Error (RMSE) for continues outcomes. For categorical outcomes, we will rely on accuracy and area under the curve (AUC). We will also look at precision and specificity, and for categorical outcomes with more than 2 categories, we will examine the F-1 Score. These evaluations will help determine the accuracy, generalizability and robustness of each model ([Salditt et al., 2023](#ref-salditt2023); [Wang et al., 2021](#ref-wang2021))

### Final Feature Importance Analysis

The final models will be analyzed to identify the most significant predictors of client-related workload. This will involve examining the feature importance scores from the best-performing models.

### Software and Tools

We will use R Statistical Software and the Tidyverse and TidyModels suite of packages for data manipulation and model building (R Core Team, 2024; Khun & Wickham 2020). This choice aligns with our familiarity with R and the study’s specific requirements. R Quarto Markdown will be used for documentation and reproducibility .

# Limitations and Challenges

While our study aims to advance the modeling of client-related workload, 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 present in the data itself. Systematic biases in the initial data collection process, such as underreporting 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 on the accuracy and thoroughness of data entry by providers. Missing data and inconsistencies are inherent challenges that could affect the robustness of our models. 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 minimizing systemic unfairness, 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.

Lastly, because our study focuses on data collected at intake, we may fail to capture dynamic changes in client needs and resource use over time. Longitudinal studies would be an excellent next step to understand how workload evolves throughout the treatment period, providing a more dynamic view of resource allocation and client needs.

# Conclusion

Given the pressing need in the CYMH sector for a case-management system that accurately and fairly assesses caseload we feel our proposed study can positively contribute to the findings. By advancing our understanding of the relationship between client characteristics and related resource need, we hope to contribute to the broader goal of optimizing mental health services to ensure that young people and their families receive timely care tailored to their individual needs.

# 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>

Arnetz, B. B., Goetz, C. M., Arnetz, J. E., Sudan, S., vanSchagen, J., Piersma, K., & Reyelts, F. (2020). Enhancing healthcare efficiency to achieve the quadruple aim: An exploratory study. *BMC Research Notes*, *13*, 362. <https://doi.org/10.1186/s13104-020-05199-8>

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>

Benda, N. C., Blumenthal, H. J., Hettinger, A. Z., Hoffman, D. J., LaVergne, D. T., Franklin, E. S., Roth, E. M., Perry, S. J., & Bisantz, A. M. (2018). Human factors design in the clinical environment: Development and assessment of an interface for visualizing emergency medicine clinician workload. *IISE Transactions on Occupational Ergonomics and Human Factors*, *6*(3-4), 225237. <https://doi.org/10.1080/24725838.2018.1522392>

Berkel, R. van, & Knies, E. (2015). Performance Management, Caseloads and the Frontline Provision of Social Services. *Social Policy & Administration*, *50*(1), 59–78. <https://doi.org/10.1111/spol.12150>

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>

Children’s Mental Health Ontario. (2020). *Kids can’t wait: Report on waitlists and wait times for child and youth mental health care in ontario*. <https://cmho.org/wp-content/uploads/CMHO-Report-WaitTimes-2020.pdf>

Children’s Mental Health Ontario. (2022). *Addressing urgent workforce challenges in child and youth mental health* (p. 4).

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>

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>

Gaines, P., Bower, A., Buckingham, B., Eagar, K., Burgess, P., & Green, J. (2003). *New Zealand Mental Health Classification and Outcomes Study: Final Report*. <https://www.lattice.co.nz/documents/final-report-june03-caos.pdf>

Hirdes, J. P., Everdingen, C. van, Ferris, J., Franco-Martin, M., Fries, B. E., Heikkilä, J., Hirdes, A., Hoffman, R., James, M. L., Martin, L., Perlman, C. M., Rabinowitz, T., Stewart, S. L., & Van Audenhove, C. (2020). The interRAI suite of mental health assessment instruments: An integrated system for the continuum of care. *Frontiers in Psychiatry*, *10*. <https://doi.org/10.3389/fpsyt.2019.00926>

Hirdes, J. P., Smith, T. F., Rabinowitz, T., Yamauchi, K., Pérez, E., Telegdi, N. C., Prendergast, P., Morris, J. N., Ikegami, N., Phillips, C. D., Fries, B. E., & Resident Assessment Instrument-Mental Health Group. (2002). The Resident Assessment Instrument-Mental Health (RAI-MH): inter-rater reliability and convergent validity. *The Journal of Behavioral Health Services & Research*, *29*(4), 419–432. <https://doi.org/10.1007/BF02287348>

Intagliata, J. (1982). Improving the quality of community care for the chronically mentally disabled: The role of case management. *Schizophrenia Bulletin*, *8*(4), 655–674. <https://doi.org/10.1093/schbul/8.4.655>

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

Joseph, J. W., Leventhal, E. L., Grossestreuer, A. V., Chen, P. C., White, B. A., Nathanson, L. A., Elhadad, N., & Sanchez, L. D. (2023). Machine learning methods for predicting patient-level emergency department workload. *The Journal of Emergency Medicine*, *64*(1), 83–92. <https://doi.org/10.1016/j.jemermed.2022.10.002>

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. (2004). 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>

Lloyd, C., & King, R. (2004). A survey of burnout among australian mental health occupational therapists and social workers. *Social Psychiatry and Psychiatric Epidemiology*, *39*(9), 752–757. <https://doi.org/10.1007/s00127-004-0808-7>

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>

*MOH CYMH Service Description Schedules - Central Forms Repository (CFR)*. (n.d.). <https://forms.mgcs.gov.on.ca/en/dataset/on00161>

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>

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

Perlman, C. M., Hirdes, J. P., Barbaree, H., Fries, B. E., McKillop, I., Morris, J. N., & Rabinowitz, T. (2013). Development of mental health quality indicators (MHQIs) for inpatient psychiatry based on the interRAI mental health assessment. *BMC Health Services Research*, *13*, 15. <https://doi.org/10.1186/1472-6963-13-15>

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>

Stewart, S. L., Celebre, A., Head, M. J., James, M. L., Martin, L., & Fries, B. E. (2020). A Case-Mix System for Children and Youth With Developmental Disabilities. *Health Services Insights*, *13*, 1178632920977899. <https://doi.org/10.1177/1178632920977899>

Stewart, S. L., Celebre, A., Semovski, V., Hirdes, J. P., Vadeboncoeur, C., & Poss, J. W. (2022). The interRAI child and youth suite of mental health assessment instruments: An integrated approach to mental health service delivery. *Frontiers in Psychiatry*, *13*. <https://doi.org/10.3389/fpsyt.2022.710569>

Stewart, S. L., & Hamza, C. A. (2017). The child and youth mental health assessment (ChYMH): An examination of the psychometric properties of an integrated assessment developed for clinically referred children and youth. *BMC Health Services Research*, *17*(1), 82. <https://doi.org/10.1186/s12913-016-1970-9>

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>

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>

Zeleke, A. J., Palumbo, P., Tubertini, P., Miglio, R., & Chiari, L. (2023). Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: A gradient boosting algorithm analysis. *Frontiers in Artificial Intelligence*, *6*, 1179226. <https://doi.org/10.3389/frai.2023.1179226>

# Appendix

ADD TABLES OF VARIABLES HERE