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

Shauna Heron

Department of Psychology, Laurentian University

# Author Note

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

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 services (CYMH) face significant workforce challenges alongside increased demand for service, the ability to anticipate and optimize the workload of staff is critical to providing high quality care ([CMHO, 2019](#ref-cmho2019); [CYMHLAC, 2019](#ref-cymhlac2019)); yet research into the client-side factors that contribute to provider workload is largely absent in the literature–particularly in the CYMH sector. With this gap in mind, the goal of the current study is to evaluate the utility of machine learning models, trained on mental health information collected at intake, to predict the amount of work an individual client might contribute to a provider’s workload. Specific objectives include: (i) identifying significant predictors of client-related work from intake assessment data; (ii) investigating 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); (iii) comparing the predictive accuracy of several tree-based machine learning models against traditional linear models; (iv) exploring 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

Warning: package 'DiagrammeR' was built under R version 4.3.3

Community-based child and youth mental health services (CYMH) in Ontario face a dual challenge: a growing volume of service referrals alongside significant staffing challenges ([CYMHLAC, 2019](#ref-cymhlac2019)). A 2020 survey of community CYMH centres across Ontario revealed that 83% of agencies reported staffing shortages–59% of them direct-service, clinical roles like psychologists, psychometrists, psychotherapists, and social workers. At the same time, there has been an estimated three-fold increase in the number of youth identifying a need for professional help over the last thirty years ([Comeau et al., 2019](#ref-comeau2019)), resulting in interruptions in service and longer wait times for children and their families ([CMHO, 2020](#ref-cmho2020)).

# Background

Illustratively, in 2020 Children’s Mental Health Ontario (CMHO) reported that over 28,000 children and youth in Ontario were waiting up to 2.5 years for mental health services ([CMHO, 2019](#ref-cmho2019), [2020](#ref-cmho2020); [CYMHLAC, 2019](#ref-cymhlac2019)), some even “aging out” of the system before they were off the wait list. With over 70% of mental health and addiction problems starting before age seventeen, this is a problem. Not only do we miss a critical opportunity for early intervention, but individual and family stress related to mental health challenges are compounded, increasing the burden to a public health care system where hospitalization of youth with mental health and addictions issues has increased by an estimated 90% ([CMHO, 2019](#ref-cmho2019), [2020](#ref-cmho2020); [CYMHLAC, 2019](#ref-cymhlac2019)).

To address these challenges, the Ontario Ministry of Health and Long-Term Care (MOHLTC) has implemented several initiatives meant to increase the flow of young people through the system with the aim of reducing wait times. Today, same-day help can be accessed through youth hubs, walk-in clinics and other rapid access programs and new intake processes have been implemented to identify those at highest risk faster ([CMHO, 2020](#ref-cmho2020)). However, for those that don’t meet the highest-risk threshold but still require intensive services, they will still wait–and much longer than evidence suggests is best practice ([CMHO, 2020](#ref-cmho2020)). According to the most recent Auditor General’s Report on Child and Youth Mental Health a significant barrier to addressing the problem is determining reasonable provider-to-client workload ratios ([Auditor General’s Report, 2016](#X74ef00379ba32932a14081a0b155af9f4f12f08), [2016](#X74ef00379ba32932a14081a0b155af9f4f12f08)).

## Caseload versus workload

It is important to define what we mean by “workload” versus “caseload. In the current thesis, *caseload* refers to the number of active clients assigned to a provider at any given time; while *workload* is the time a provider spends serving each case (e.g. face to face meetings, documentation, communicating with care givers, travel time, etc.) as well as the time they spend on other professional responsibilities such as travel, supervision, professional development, and training ([CMHO, 2019](#ref-cmho2019)).

According to a 2019 survey of CYMH agencies in Ontario, the most common metric used for estimating provider workload was client counts and the type of service coming in second (i.e., brief services versus counselling and therapy). Case counts are used to determine how many new clients a provider has room for in their overall caseload and signal when a provider has reached their capacity ([CMHO, 2019](#ref-cmho2019)). Today, caseload guidelines for counselling and therapy services range from 12-16 per provider for intensive treatment all the way up to 30 ([Burns et al., 2007](#ref-burns2007)). In practice, case counts should *decrease* as the complexity of individual cases within each providers portfolio increases. However, the resources needed to evaluate the complexity and intensity of each child’s need *before* they are even assigned to a clinician are resources that most public agencies do not have ([CMHO, 2019](#ref-cmho2019)). As a result, cases are most often assigned in a way that assumes that each case represents a similar amount of work. For example, an agency might set a target of 20 cases per provider for counselling and therapy services; meaning that providers *without* 20 cases have room for more.

In Ontario, between 16% and 24% of CYMH agencies reported average caseloads per worker that were at least 50% larger than provincial averages for the same services ([CMHO, 2019](#ref-cmho2019)), which is a problem since high caseloads (more than 20 to 30 clients) are associated with a range of negative outcomes for providers, their clients and the agencies they work for. High caseloads have been linked with self-reported burnout ([Morse et al., 2012](#ref-morse2012)), poor work engagement, low job satisfaction ([Green et al., 2014](#ref-green2014)) and poorer treatment outcomes ([Garman et al., 2002](#ref-garman2002)). Provider burnout, is characterized by emotional exhaustion, depersonalization and a decreased sense of self efficacy which impacts a providers perceived ability to handle job-related stressors ([Kim et al., 2018](#ref-kim2018)). Provider burnout is particularly pronounced in community mental health settings where caseloads are typically larger and rates of clinical complexity and co-morbidity are higher ([Tran et al., 2019](#ref-tran2019)).

At the same time, newer research suggests the more important factor influencing client-related work may be the *mix* of cases rather than flat counts ([King et al., 2000](#ref-king2000)). For instance, King et al. ([2000](#ref-king2000)) found that counts were not as predictive of clinician burnout as the mix of cases. He posited that providers 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 which might explain the poorer outcomes associated with high caseloads ([Kim et al., 2018](#ref-kim2018); [King, 2009](#ref-king2009)). To illustrate: two clinicians could have identical caseloads as far as counts but the amount of work necessary to deliver services could vary wildly between them. For example, one clinician may easily manage a higher number of low-complexity cases but become overwhelmed with just a few high-complexity cases.These findings are supported by the work o

## Predicting client-related work

Given this state of affairs, if we had a way to predict the work associated with a given client in a without requiring dedicated staff to carefully evaluate each case before clinical assignment, agencies might more efficiently evaluate and manage provider caseloads. However modeling client characteristics and 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 such as 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 case assignment system that can accurately reflect the changing demands placed on staff ([Arnetz et al., 2020](#ref-arnetz2020); [CMHO, 2019](#ref-cmho2019)). Nevertheless, the new guidelines *have* informed 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.

## Measuring workload: direct and indirect time

An important workload metric born out of Ontario’s CYMH AG audits and the recommendations that followed, is time (direct and indirect) that is spent with each client–a metric that all publicly funded CYMH service providers in Ontario are now required to report ([Ministry of Health, 2024](#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, travel time, and consultations. The sum of all direct and indirect hours logged by a clinician amounts to the overall “work” attributable to a given client. In our models, we will utilize direct and indirect hours to approximate the work attributed to an individual client. It is important to note that time is only a *proxy* for client-related work as it doesn’t convey differences in the effort necessary to treat different clients, nor the stress associated with each case or even the workplace culture and supports that can influence the perceived workload experienced by the provider.

One Ontario agency that has begun using the new metric to track and manage caseloads, is [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. Compass utilizes an electronic dashboard to monitor caseloads and the associated direct and indirect time logged. While the dashboard has been useful for comparing case counts between clinicians and teams, it offers less insight into the amount of work associated with each case. Moreover, it doesn’t help in assigning *new* cases beyond indicating which providers have higher or lower case counts than others. It does not allow administrators to evaluate the mix of cases (high to low needs) within each caseload, making it difficult to assign new clients in a way that is fair in terms of work. Based on requests from both leadership and clinicians themselves for an assignment tool that accounts for differences in the intensity of services needed, we began to wonder whether psychological screening data collected at intake might be modeled to quantify client complexity. If we could determine a *weight* for each case it might be easier to monitor existing caseloads as well as inform new case assignment.

### Modelling client complexity

Across health domains (e.g., psychiatric, emergency medicine, community based mental health) various strategies have been employed to manage provider workload with different levels of service determined by characteristics like symptom severity or prior diagnoses ([Johnson et al., 1998](#ref-johnson1998); [Tran et al., 2019](#ref-tran2019)). These systems assume that though the needs of each individual in a population will be unique, there will be shared characteristics that determine the type of treatment they will need (e.g., family counselling versus substance use treatment). These groups represent the mix of cases or “case-mix” which can be viewed as a proxy for the types of care needs of the population. Case-mix classification systems are most often used in the health care sector to help payers and agencies monitor cost by categorizing clients based on their expected resource use ([CMHO, 2019](#ref-cmho2019)). At the agency level, the data within a case mix may contain the activity of individual providers such as the direct-hours attributable to individual clients within their caseloads and client-level characteristics like diagnoses, treatment history and current presenting symptoms (e.g., crisis intervention versus brief services). It may also contain demographics information like age, school district and the number of services offered. These data points are often stored in an electronic health care database and are captured using some kind of classification system.

## Case-Mix Classification Systems

Case-mix classification systems usually take one of two approaches to evaluating client-related work. Grouping systems assign people into classes in terms of their expected resource use, with each group having a specific weight attached to it (e.g., time-intensive treatment versus brief treatment) 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 maps to expected resource use relative to an average case ([Tran et al., 2019](#ref-tran2019)).

While case-mix systems are widely used in the medical domain, to date, most mental health case-mix classification systems focus on acute care in hospital or other inpatient settings which are distinctly different than community based care in several ways ([Tran et al., 2019](#ref-tran2019)). Typically, community based care involves a team of providers offering a wider range of services. For instance, urgent care, crisis intervention, brief services or longer-term treatment like counselling and therapy as well as group programs and day treatment provided in educational settings. Often these services are provided without clear diagnoses or well-defined treatment protocols, making them more complicated to model. For instance, a medical emergency will often come with a clear diagnoses like “broken arm” which has a recovery window that is easier to estimate. On the other hand, the subjective experience of anxiety or depression as well as the time it will take to recover is far less black and white.

The difficulty of modeling CYMH data is illustrated by the fact that only a handful of studies have looked at solving this problem, despite urgent need. Indeed, a 2019 scoping review of casemix literature in community-based mental health care domains found only a single case that looked at case-mix classification to predict mental health care resource in the CYMH domain ([Martin et al., 2020](#ref-martin2020); [Tran et al., 2019](#ref-tran2019)). In that study, researchers modeled 4573 client records from eleven UK outpatient CYMH agencies, comparing cluster analysis, regression trees and a conceptual classification based on clinical best practice guidelines to accurately predict the number of appointments a client attended in treatment ([Martin et al., 2020](#ref-martin2020)). While the researchers found the conceptual classification was as clinically meaningful as data-driven classification in accounting for number of appointments, they found little evidence to support the idea that either client complexity or context factors (with the exception of school attendance problems) were linked with overall appointment counts ([Martin et al., 2020](#ref-martin2020)). Moveover, the models failed to explain significant variation in resource provision between workers despite clients exhibiting similar characteristics. Data quality problems and omission of important individual-level factors were cited as potential points of failure but suggested their results merited further testing and development.

Another group of researchers tried to predict the workload associated with clients at a community-based mental health centre for the elderly ([Baillon et al., 2009](#ref-baillon2009)). Using an 8-item case weighting scale (CWS) researchers identified factors staff felt contributed to demand for staff time. A multiple regression model was used 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 the model was reported a success, accounting for 58% of the variance in time spent on client-related work, the sample consisted of only 87 cases and relied on a statistical method unsuitable 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)).

Next we looked to other domains where we found several measures of workload intensity had been developed to manage caseloads. For example, in general psychiatry, emergency medicine and nursing, researchers have used factors like sociodemographics, functional ability, and caregiver and social network characteristics to predict service utilization. One approach that Tran et al. ([2019](#ref-tran2019)) suggested may be a good candidate for testing in community settings 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)).

### interRAI Child and Youth Mental Health Assessment Tool

interRAI is an international research network that develops clinical standards across a variety of health and social service settings that have developed a large toolkit of instruments 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 collected 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 act as a custodian of interRAI standards and houses and monitors the collected interRAI data. Many interRAI instruments are used across Canada and internationally, but the Children and Youth Mental Health assessment (ChYMH) represents the first assessment designed specifically 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, respond to and monitor the strengths, preferences and mental health needs of clients in in-patient and out-patient treatment. The ChYMH products are currently being utilized in over 60 mental health agencies across Ontario including Compass.

The full ChYMH includes over 400 items that together are meant to build a comprehensive picture of a client’s strengths, needs, functioning and areas of risk ([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 utilize screener data collected at intake to model client-related complexity and resulting work.

Importantly, the ChYMH scales have demonstrated good 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) was useful in predicting admission for risk of self-harm in youth between 10-17 years. In addition, 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 demonstrate some predictive utility, it remains unclear how well these items might predict the actual work required to serve a 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 a 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 nonetheless deemed a success and is still in use today. However, a lack of explanation of how the analysis was conducted, as well as public availability of resulting fit statistics, makes it unclear how and where the algorithm could be improved.

Moreover, unlike the goal of the ChYRI, we are not interested in predicting client-related service cost. Instead our efforts are driven by the need to reduce wait times and make client assignment fairer and mindful of the work that is already on each clinicians plate.

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

Considering several limitations outlined by prior research in modeling the high-dimensional, sparse data characteristic of digital client information systems, we next looked to a growing body of research that utilizes machine learning algorithms to model electronic health data.

Machine learning (ML) is a branch of artificial intelligence that uses statistical techniques to enable computers to learn from data and make predictions without being explicitly programmed ([Nielsen, 2016](#ref-nielsen2016)). ML algorithms are particularly well-suited to modeling the complex, high-dimensional data found in electronic health records (EHR) and have been used to predict a variety of outcomes in health care, from patient readmission to disease diagnosis ([An et al., 2023](#ref-an2023); [Chen et al., 2023](#ref-chen2023)).This capability makes ML particularly suitable for our purposes where the data is sparse, missing values, and has complex interdependencies among variables ([An et al., 2023](#ref-an2023); [Chen et al., 2023](#ref-chen2023)). Moreoever, ML, unlike traditional methods that might only confirm specific linear relationships, can identify high-dimensional interactions and non-obvious patterns. Techniques such as regularization and cross-validation further enhance the robustness of these models and maximizes their ability to generalize well to unseen data ([Chen et al., 2023](#ref-chen2023)).Beyond prediction, ML also excels in discovering new relationships that might be validated in subsequent studies. By leveraging algorithms that optimize for predictive performance we may uncover patterns that could drive new hypotheses ([Sheetal et al., 2023](#ref-sheetal2023)).

A study that stood out to us examined the feasibility of utilizing ML algorithms to drive a visual representation of the work attributable to individual patients in a hospital setting ([Benda et al., 2018](#ref-benda2018)). The idea was to build a live, dynamic visualization that could be used to compare cases and workloads across clinicians to improve patient assignment. The display was driven by an algorithm that predicted patient-level work based on a combination of diagnoses and the number of orders or “events” (e.g., tests, phone calls, diagnoses) found in a patient’s electronic health record ([Benda et al., 2018](#ref-benda2018)). Although clinicians evaluated the tool positively, the algorithm underlying the display was found to inadequately account for actual workload, suggesting more refinements were needed ([Benda et al., 2018](#ref-benda2018)).

Building on Benda et al. ([2018](#ref-benda2018))’s study, Wang et al. ([2021](#ref-wang2021)) focused on improving the dashboards underlying algorithm using several machine learning algorithms known for their robustness in modeling the sparse, heterogeneous data found in electronic health records. Both regression and classification algorithms were used to model several proxies for workload: i) overall length of stay, ii) number of events (e.g., tests ordered, medication administered), iii) density of events (count of events divided by length of stay) and iv) a binary outcome indicating high versus low demand patients. The accuracy of the model in predicting low versus high length of stay (LOS) was 70% with information from the first hour, 73% from the first two hours and 83% with data from the entire visit. Importantly, Wang et al. ([2021](#ref-wang2021)) ’s work demonstrated the potential for machine learning techniques to predict client-related work from information collected at intake–which is what we are interested in doing and so resulted in the methodology we chose.

## Themes and gaps in the literature

In planning our approach to modeling client-related work, we hope to address several problems and limitations that emerged in the literature. The first, which dictated our chosen machine learning (ML) methodology, relates to the complexities that arise from modeling electronic health data (EMH) ([Joseph et al., 2023](#ref-joseph2023)). Hidden clustering, non-independence of observations and multi-collinearity are all challenges that ML methods like random forests and gradient boosted forests can often better manage; albeit with varying levels of explainability ([An et al., 2023](#ref-an2023); [Chen et al., 2023](#ref-chen2023); [Zeleke et al., 2023](#ref-zeleke2023)). Though there is often a trade-off of interpretability with increased predictive accuracy, modern post-hoc methods like SHapley Additive exPlanations (SHAP) offer an alternative way of understanding the contribution of inidividual features to a specific prediction ([Lundberg & Lee, 2017](#ref-lundberg)).

We also intend to tackle the lack of inconsistent outcome measures found in the literature. Previous studies often relied on flat measures such as length of stay or number of appointments as a proxy for workload which doesn’t 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 work measure on a per-diem basis, based on the total time logged in service of the client per week, which we hope will better reflect the intensity of work than a flat count ([CMHO, 2019](#ref-cmho2019); [Wang et al., 2021](#ref-wang2021)).

Additionally, we intend to address concerns highlighted by Tran et al. ([2019](#ref-tran2019)) where client-level indicators like gender and race are included in models and which raise concerns about fairness and predictor bias. Tran et al. ([2019](#ref-tran2019)) importantly pointed out that while race and gender may *correlate* with resource use, the relationships are confounded by marginalization that may be drive increased risk for mental health concerns ([Gaines et al., 2003](#ref-gaines2003); [Tran et al., 2019](#ref-tran2019)). To avoid these pitfalls, we will focus solely on client-side drivers of work, specifically the mental health acuity features collected at intake, which have been shown to be more reliable drivers of resource use ([Perlman et al., 2013](#ref-perlman2013); [Tran et al., 2019](#ref-tran2019)). We plan to use interRAI screener+ items and scores specifically ([Hirdes et al., 2020](#ref-hirdes2020)).

Furthermore, we intend to focus solely on client-side drivers of work as recommended by Tran et al. ([2019](#ref-tran2019)) who suggest including provider-side variables like years of experience or preferred therapeutic modality risks reinforcing systemic unfairness in case distribution where the more experienced clinicians may have the greater bulk of complex clients. Moreover, provider related information is not available in the client record at intake.

Finally, to avoid considerable validation problems found across the literature, we will utilize cross-validation for model training and tuning and will hold back a test set of unseen data for final model evaluation. The test set we hold back will serve as a control for determining model generalizability ([Costa et al., 2015](#ref-costa2015); [Reid et al., 2021](#ref-reid2021); [Tran et al., 2019](#ref-tran2019)). We will also split our data in a group-wise fashion, ensuring that clients with multiple rows will only be in either the training set or the test set, never both. This will ensure that we have a more robust measure of how well the final model will generalize to “unseen” clients. We will also clearly outline our choice of models, metrics and all R code will be made available for reproducibility (see Methods).

## Hypotheses

– ask for advice here –

Though the current study will be largely exploratory in nature, we 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 than simple linear regression, we predict that tree-based machine learning models will be more accurate predictors of client-related work than linear regression. Moreover, we anticipate that mental health acuity features such as depression and anxiety scores will contribute more substantially to explaining the variance in client-related work than age. This hypothesis stems from existing literature indicating that these variables are critical drivers of resource use ([Perlman et al., 2013](#ref-perlman2013); [Tran et al., 2019](#ref-tran2019)). Finally, we anticipate that models will more accurately predict caseweight (hours spent directly with the client, divided length of stay), than flat proxies (e.g., overall length of stay) as a ratio 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)).

# 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 that we want to predict the amount of resource use a client will require based on information collected at intake (demographics information and ChYMH-S 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 require Brief versus CT services is a prediction task for another study. Final counts after screening will be reported and added to the flowchart before analysis.

## Data Security

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

De-identified 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 de-identified 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 analysis 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 identifying information that could be linked to clients or smaller subgroups of clients. Furthermore, the researchers will seek approval from Compass before results are shared or utilized in any report or presentation.

## Procedure

The following steps outline the proposed process which will consist of four 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 if necessary, 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 decided on a case by case basis and reported in our final paper. Moreover, the final report will include the R code necessary to replicate these steps. Pending approval, the data scripts will also be made publicly available.

### Data Splitting

To maximize the generalizability of our models, data will be split by group (unique client ID) and stratified on the outcome. Stratification will ensure that the outcome distribution is preserved in the training and test sets. Grouping the split on unique IDs will ensure that clients with multiple assessments are included in either the training or test sets (never both) allowing for a more accurate evaluation of model performance. Group cross-validation is particularly beneficial when dealing with repeated measurement data like ours, as it maintains the integrity of the group structure and prevents information from the test folds leaking into the training process ([Figure 3](#fig-procedure-flow)). The split ratio will be 80/20: 80% of data in the training set and 20% in the test set.

10-fold cross-validation will be used to tune the models. The cross-validation folds are created with a portion of the training data, subdividing it into 10 subsets, or “folds,” that will preserve the same stratified-grouping structure. This will allow the models to be iteratively trained on nine folds and then validated on the remaining one, ensuring each group is used for validation exactly once.

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 final step is crucial for assessing the models’ generalizability and for identifying any overfitting that may have occurred during training.

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

Importantly, feature engineering–the creation of new predictors based on existing variables in the dataset–will occur *after* data splitting on the training data to minimize the risk of data leakage that could inadvertently occur when creating new variables from the full dataset ([Sheetal et al., 2023](#ref-sheetal2023)).

The main criterion for inclusion in the model will be the variable’s availability in the electronic health record (EHR) system at intake. Features (variables) will include items and scales from the interRAI ChYMH Screener+. The interRAI ChYMH is a clinician-rated tool that is completed based on a semi-structured clinical interview and includes several scale scores, administrative items as well as other demographics information ([Stewart & Hamza, 2017](#ref-stewart2017)).

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. The instrument and the various scales and items it contains have been tested extensively in research worldwide, demonstrating reliable face, content, construct and predictive validity ([Stewart et al., 2022](#ref-stewart2022); [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.

#### Targets (dependent variables or workload proxies).

Informed by Wang et al. ([2021](#ref-wang2021)), we propose to measure the influence of our workload indicators on i) length of service (weeks), ii) hours of time spent directly with the client or caregivers either in person on the phone or video call (direct time), iii) hours spent indirectly on client-related work like writing a treatment plan or filling out case notes (indirect time), iii) combined direct and iv) a case density score (caseweight) calculated by dividing the sum of direct and indirect time divided by the number of weeks the client spent in service (see [Equation 1](#eq-caseweight)). 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, etc.) in the 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

Outcome proxies will be modeled as a i) continuous index scores or a ii) classification grouping (low versus high). In practice, regression could be used to predict the number of hours a client might need per week across the episode of care, while the classification model will provide a quick flag for cases that are more high intensity than others.

As a first step, we will train models on data collected over the entire program length to determine whether the indicators have any utility in modelling the workload proxies (some clients will have multiple assessments across the treatment period). Next, we will train on data *only* from the first assessment to predict workload captured 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 feasible. We believe this will be a more difficult task to accurately model, as such, a simple binary prediction of low versus high intensity may be easier for the models to accurately predict than a continuous index score.

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 plan to utilize the following supervised machine-learning algorithms for both regression and classification problems: i) Random Forest (RF) for its ability to handle large datasets with high dimensionality; ii) XGBoost, known for its high performance and predictive accuracy on tabular datasets, iii) LASSO and Ridge regression for their ability to manage high multicollinearity, and finally iv) linear regression to serve as a baseline model for continuous outcomes and generalized logistic regression for binary outcomes. All algorithms will be trained on the same training set and cross validation folds and evaluated on the same test set using the TidyModels suite of packages in R Studio. These algorithms were chosen based on their success modeling similarly complex, tabular data types and may grow to include other models in the final paper ([Salditt et al., 2023](#ref-salditt2023); [Sheetal et al., 2023](#ref-sheetal2023)).

### Validation and Testing

Final models will be statistically compared and 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). 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 models will also be analyzed to identify which predictors were the most significant predictors of client-related workload using SHAP scores.

### 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. During the model building process, there is a chance we may use Python as well in the RStudio environment and will report and document this choice thoroughly if we do ([Van Rossum & Drake, 1995](#ref-vanrossum1995)).

# Limitations and Challenges

While our study aims to advance our understanding 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 from biases present in the data itself. Systematic biases in the initial data collection process, such as under reporting, 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. Moreover, many of the scale scores may be influenced by 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 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 modeling static, historical data, we won’t capture dynamic changes in client needs that may impact resource use over time. More sophisticated real-time modeling techniques might be an interesting next step to explore how workload changes throughout the treatment period based on real-time changes in client need which might potentially providing a more dynamic picture of resource allocation and client needs.

Finally, given the pressing need for a case-management tool that can more accurately and fairly assesses client-related work, we think our proposed study is a timely addition. Not only does our research have the potential to advance our understanding of the relationship between client characteristics and resource use, it contributes to the broader project of optimizing mental health services in a way that maximizes the chance that young people and their families receive high quality, timely care, while minimizing the risk of provider burnout.

# 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>

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

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>

Burns, T., Yiend, J., Doll, H., Fahy, T., Fiander, M., & Tyrer, P. (2007). Using activity data to explore the influence of case-load size on care patterns. *The British Journal of Psychiatry: The Journal of Mental Science*, *190*, 217–222. <https://doi.org/10.1192/bjp.bp.106.025940>

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

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

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

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>

Costa, A. P., Poss, J. W., & McKillop, I. (2015). Contemplating case mix: A primer on case mix classification and management. *Healthcare Management Forum*, *28*(1), 12–15. <https://doi.org/10.1177/0840470414551866>

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>

Garman, A. N., Corrigan, P. W., & Morris, S. (2002). Staff burnout and patient satisfaction: evidence of relationships at the care unit level. *Journal of Occupational Health Psychology*, *7*(3), 235–241.

Green, H., Barkham, M., Kellett, S., & Saxon, D. (2014). Therapist effects and IAPT Psychological Wellbeing Practitioners (PWPs): a multilevel modelling and mixed methods analysis. *Behaviour Research and Therapy*, *63*, 43–54. <https://doi.org/10.1016/j.brat.2014.08.009>

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>

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>

Kim, J. J., Brookman-Frazee, L., Gellatly, R., Stadnick, N., Barnett, M. L., & Lau, A. S. (2018). Predictors of burnout among community therapists in the sustainment phase of a system-driven implementation of multiple evidence-based practices in children’s mental health. *Professional Psychology, Research and Practice*, *49*(2), 131–142. <https://doi.org/10.1037/pro0000182>

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>

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

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

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

Ministry of Health. (2024). *MOH child and youth mental health service description schedules*. <https://forms.mgcs.gov.on.ca/en/dataset/on00161>

Morse, G., Salyers, M. P., Rollins, A. L., Monroe-DeVita, M., & Pfahler, C. (2012). Burnout in mental health services: A review of the problem and its remediation. *Administration and Policy in Mental Health*, *39*(5), 341–352. <https://doi.org/10.1007/s10488-011-0352-1>

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>

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>

Reid, G. J., Stewart, S. L., Barwick, M., Cunningham, C., Carter, J., Evans, B., Leschied, A., Neufeld, R. W. J., St. Pierre, J., Tobon, J., Vingilis, E., & Zaric, G. (2021). Exploring Patterns of Service Utilization Within Children’s Mental Health Agencies. *Journal of Child and Family Studies*, *30*(2), 556–574. <https://doi.org/10.1007/s10826-020-01859-2>

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>

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

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

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