# 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

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

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

As mental health service providers face increasing demands, the ability to anticipate and optimize the workload of staff is critical to maintaining provider well-being and delivering equitable, effective care. For this reason, It is important to understand providers’ current workload as well as their capacity to take on additional work. Yet so far, there are a lack of consistent and reliable strategies for making equitable client assignment. The primary goal of this study is to evaluate the utility of machine learning models to predict the amount of work individual clients contribute to a clinician’s workload in outpatient child and youth mental health services. Specific objectives include: (i) identifying significant predictors of client-related workload from intake assessment data, (ii) comparing the predictive accuracy of several machine learning models against traditional linear regression, (iii) investigating the relationship between client-related workload indicators and workload proxies, (iv) exploring the potential for early prediction of caseweight to inform workload management.

# Background

In Canada, and similar countries, outpatient Child and Youth Mental Health Services (CYMHS) face a significant volume of service referrals, posing challenges for agencies to deliver high-quality care that is both cost-effective and equitable in its work distribution ([CYMHLAC, 2019](#ref-cymhlac2019)). The typical, publicly funded, CYMH agency is characterized by a high volume of referrals, long wait lists and caseloads containing a mix of clients with varying needs, requiring a range of services ([CMHO, 2019](#ref-cmho2019)). At the same time, client need varies widely, with some requiring more intensive services than others, making it a challenge to distribute the work fairly among providers.

Traditionally, agencies have relied on simple case counts to determine whether a provider has capacity to take on more; however we know that flat counts cannot account for the individual client need nor the amount of support that each case might need. Indeed, the significance of the issue is highlighted by the findings of a 2019 audit of provincially-funded child and youth mental health agencies which out of 11 recommendations, two addressed caseloads specifically ([CMHO, 2019](#ref-cmho2019)). Among other issues, the report report stressed that case counts, a static measure, failed to reflect the varying needs and intervention levels required by different clients.

## 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 staff member at any given time, while *workload* is the amount of time required to serve each case as well as tend to other professional responsibilities such as supervision, professional development, and training ([CMHO, 2019](#ref-cmho2019)).

One line of inquiry looks at the influence of caseload on provider role performance, finding high case loads linked with lower self-reported efficacy ([King et al., 2000](#ref-king2000)) and poorer clinical outcomes ([Lloyd & King, 2004](#ref-lloyd2004)), while smaller caseloads lead to better clinical outcomes and higher retention rates ([Berkel & Knies, 2015](#ref-vanberkel2015)). These findings support earlier research suggesting that caseloads exceeding 20 to 30 clients leads 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, the relationship between caseload and burnout may not be as straightforward ([King et al., 2000](#ref-king2000), [2004](#ref-king2004)). Indeed, King et al. found that case counts were not predictive of clinician burnout, but the *mix* of cases was ([King et al., 2000](#ref-king2000)). He posited that the reason for this may be that providers adapt to an increased number of cases by simply doing less for each case, suggesting a potential ‘dose-response’ relationship between a provider’s time and their effectiveness.

These findings tie into recommendations that came out of a 2019 audit of provincially-funded child and youth mental health agencies in Ontario, which determined that workload was a more comprehensive measure of how providers spend their time than counts alone ([CMHO, 2019](#ref-cmho2019)). In the report, “work” is defined as the hours spent on client-related tasks. Direct hours were defined as face-to-face interactions, phone or video-based communications, and meetings with parents and caregivers. While indirect hours involved client-related tasks like documentation, telephone calls, advocacy, and consultations. The sum of which amounted to the overall “work” attributable to a given client. Today, publicly funded service providers in Ontario require clinicians to report all direct-time time spent in the service of clients ([CMHO, 2019](#ref-cmho2019)).

## A measure of client-related work

Given these findings, if there were factors that could be used to accurately predict the workload associated with each client, agencies might better manage their caseloads. However modeling client characteristics, let alone predicting the work that is driven by them, has proved difficult ([CMHO, 2019](#ref-cmho2019); [King, 2009](#ref-king2009); [Tran et al., 2019](#ref-tran2019)). Despite audits and initiatives (e.g. 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 ([CMHO, 2019](#ref-cmho2019)). Nevertheless, the new guidelines and data have informed the work of several agencies in expanding efforts to understand and track workload in their own organization.

For example, [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 inform new case assignment.

## Case-mix models of client-related work

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 ([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 (2019) as a starting point, hoping it would lead to discovering more recent work. Unfortunately, the single case that looked at case-mix classification to predict mental health care resource use in 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. Though the study was only a first step, they found the classification algorithms ability to predict accurately on new data was weak and no 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, which was contrary to what they expected ([Martin et al., 2020](#ref-martin2020)).

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 their first appointment. Though they reported an accurate model, their sample consisted 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 ([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

Several measures of workload intensity have been developed to manage caseloads in other specialties, particularly in inpatient 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 lack 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)). In 2020, the SCIPP methodology was used to inform a cluster analysis of 346 interRAI assessments to predict resource “cost” for children and youth with developmental disabilities ([Stewart et al., 2020a](#ref-stewart2020a)). Although the resulting Child and Youth Resource Index (ChYRI) could explain only 30% of the variance in per diem costs for community based services, researchers report it is a successful model in use today ([Stewart et al., 2020b](#ref-stewart2020b)).

Another study that stood out to us was one that examined the feasability of developing a visual representation of the work attributible to individual patients in a hospital setting. The idea was to have a live visualization that could be used to compare workload across clinicians and improve patient assignment. The display was driven by an algorithm that predicted patient-level work based on a combination of clinical assessment scores and the number of orders or “events” (e.g. tests, phone calls, diagnoses) placed in their electronic health record ([Benda et al., 2018](#ref-benda2018)). Though the clinicians evaluated the tool positively, the algorithm underlying the display was found to inadequately account for actual workload ([Benda et al., 2018](#ref-benda2018)).

### A novel approach to modeling case-mix

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 (2018; 2021). 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 are different (hospital stay versus outpatient mental health treatment), Wang et al. ([2021](#ref-wang2021)) 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 from the literature that will inform our research. First, we aim to address the gap in case-mix research by focusing on the specific needs of younger people with mental health concerns in community outpatient settings. Our study targets Child and Youth Mental Health Services (CYMHS), 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, which do not 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 (time spent) per week ([Wang et al., 2021](#ref-wang2021)). This method, supported by the System for Classification of In-Patient Psychiatry (SCIPP) developed in Canada, should better reflect the intensity of work than flat counts ([CMHO, 2019](#ref-cmho2019)).

Additionally, we will account for 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 increase the risk for mental health concerns ([Gaines et al., 2003](#ref-gaines)). To avoid perpetuating marginalization, we will focus on variables that directly drive resource use, such as scale scores or symptom ratings, and exclude variables like race and ethnicity. Following Tran et al. ([2019](#ref-tran2019))’s recommendation to utilize direct measures of client need, we plan to use interRAI screener+ items and scores specifically ([Hirdes et al., 2020](#ref-hirdes2020)). By employing a data-driven approach to predictor selection, we will compare and validate model-selected items against clinical best practices, ensuring fairness and accuracy in our workload predictions.

In addition, we will address the challenges that prior research has had in separating provider and client-side drivers of work. Provider-side influences, such as staff training level and agency policy can confound resource use proxies, while client-side influences are more direct. In addition, @tran2019 suggested that including provider-side variables might reinforce systemic unfairness in case distribution (2000). Therefore, we will focus solely on client-side variables to predict workload, minimizing provider-side influence and enhancing the fairness of workload distribution.

Model evaluation and metrics are another critical area of weakness across the literature. 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 clearly outline our choice in metrics used to evaluate model performance (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 that traditional statistical methods often struggle with ([Joseph et al., 2023](#ref-joseph2023)). EMH data poses unique challenges such as hidden clustering, non-independence of observations, missing data, outliers, and sparse data. Consequently, a significant amount of data wrangling and variable culling is necessary for traditional analysis, resulting in 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 ([Zeleke et al., 2023](#ref-zeleke2023)).

### What is machine learning?

Machine learning (ML) is a subset of artificial intelligence (AI) that focuses on developing algorithms and statistical models enabling computers to perform specific tasks without explicit instructions. Instead of being programmed with a fixed set of rules, the goal of ML systems is to learn patterns in data that will generalize well to new, unobserved data ([Nielsen, 2016](#ref-nielsen2016)). In *supervised* ML, if a model generalizes well, it will make good predictions. Supervised machine learning is concerned with establishing the relationship between an outcome variable Y and a set of predictor variables X. This learning process allows ML models to improve their performance over time as they are exposed to more data

In the context of psychological research, ML can be particularly valuable. While traditional statistical methods rely on predefined equations and assumptions about the data, ML algorithms can detect complex, nonlinear patterns and relationships within large and potentially noisy datasets. This capability makes ML ideal for analyzing electronic health data, where data may be non-linear, have missing values, or exhibit complex interdependencies among variables ([Sheetal et al., 2023](#ref-sheetal2023)). Moreover, while traditional methods are excellent for confirming specific linear relationships and testing hypothesese, ML excels in predicting outcomes and uncovering relationships that can be validated in subsequent studies. Abductive reasoning like this can help drive research forward by generating new hypotheses based on the patterns identified by ML models ([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 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). The clinical flow of clients at Compass involves an intake assessment that determines the level of care required. Clients are then referred to a brief service (one session), counselling and therapy (CT) (intensive treatment without a defined end date), or other services like Day Treatment or group programs. We were interested specifically, in predicting the case weight of individuals who were approved for CT as this is where a better case assignment method is most relevant. Only clients whose initial screening resulted in referral to CT were 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 agencies institutional review boards 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. It will never include 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 and preprocessing; 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) model the relationship between the indicators and proxies with algorithms of varying complexity; and 4) evaluate the model’s 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 .

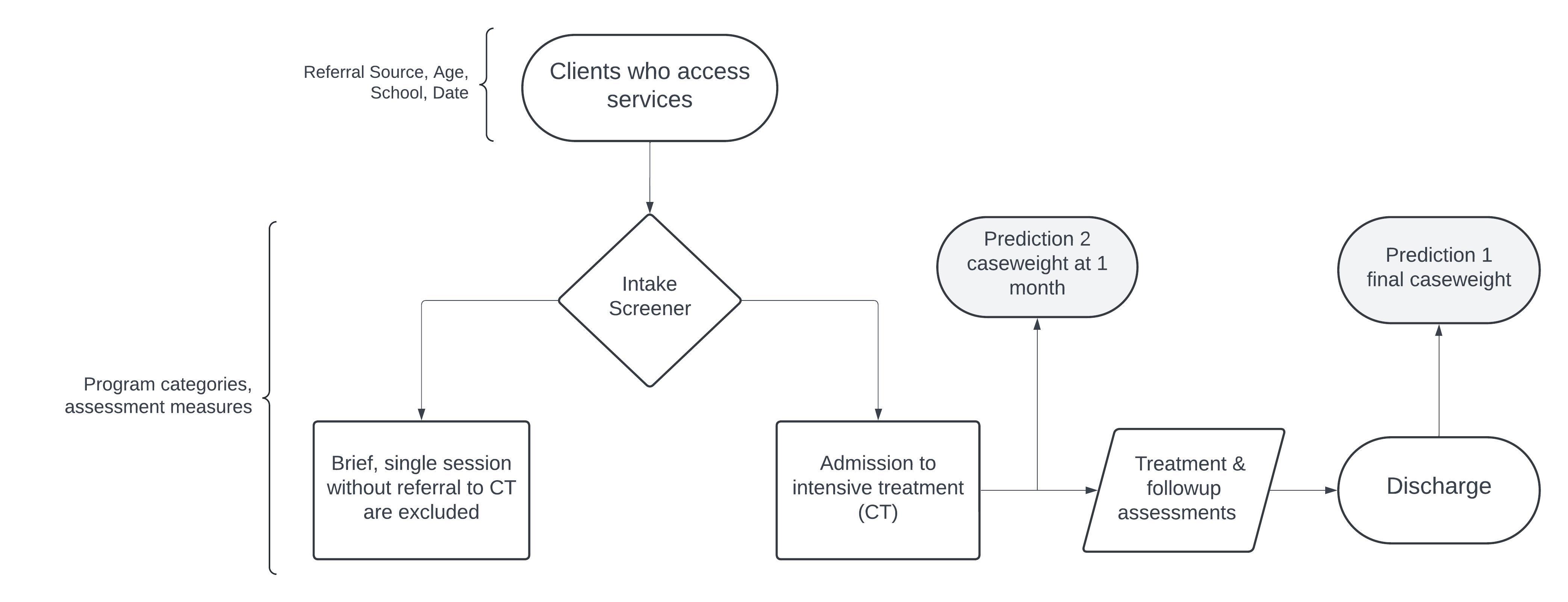
Importantly, feature engineering–the creation of new predictors based on existing variables in the dataset–will occur *after* data splitting on the training data to minimize the risk of data leakage that could inadvertently occur when creating new variables from the full dataset ([Sheetal et al., 2023](#ref-sheetal2023)).

#### 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 collected at intake includes various psychological measures included on the interRAI Screener+ ([Stewart & Babcock, 2020](#ref-stewart2020)) as well as client demographics and administrative data (a full list of potential indicators will be included in the final report).

Figure 1

Clinical pathway

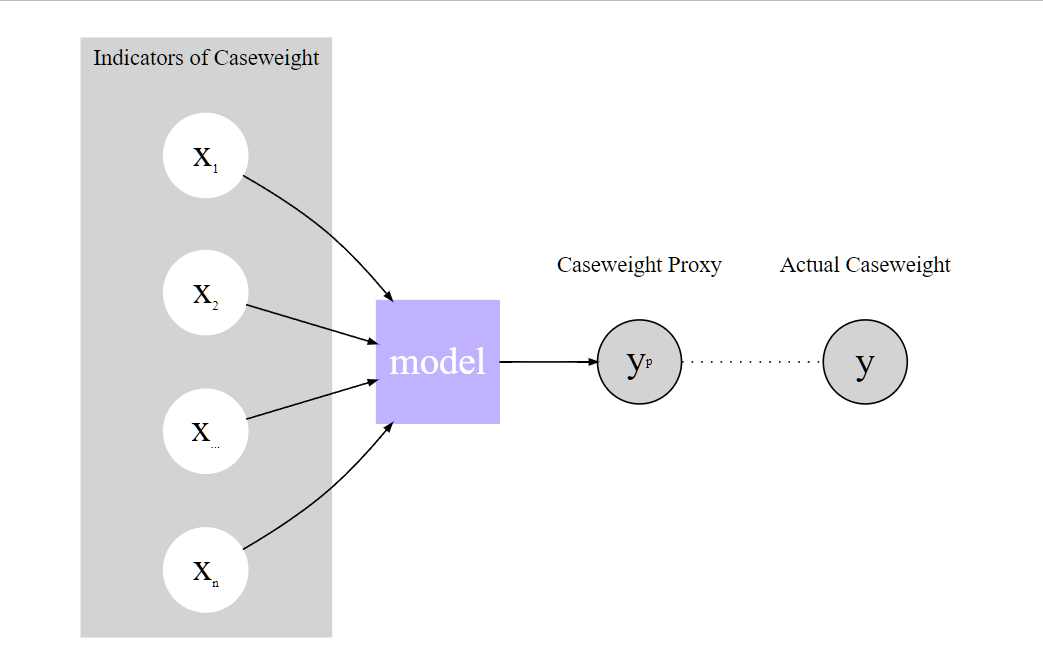


#### Dependent Variables (workload proxies).

Informed by Wang et al., 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 (2000). 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



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, and a three class proxy at the 30th, 60th, and 90th percentile.

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. As a first step, we will 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 one month later or at the next follow-up assessment. This will allow us to test whether providing a workload prediction in the early stages of a visit is feasible.

### Data Splitting

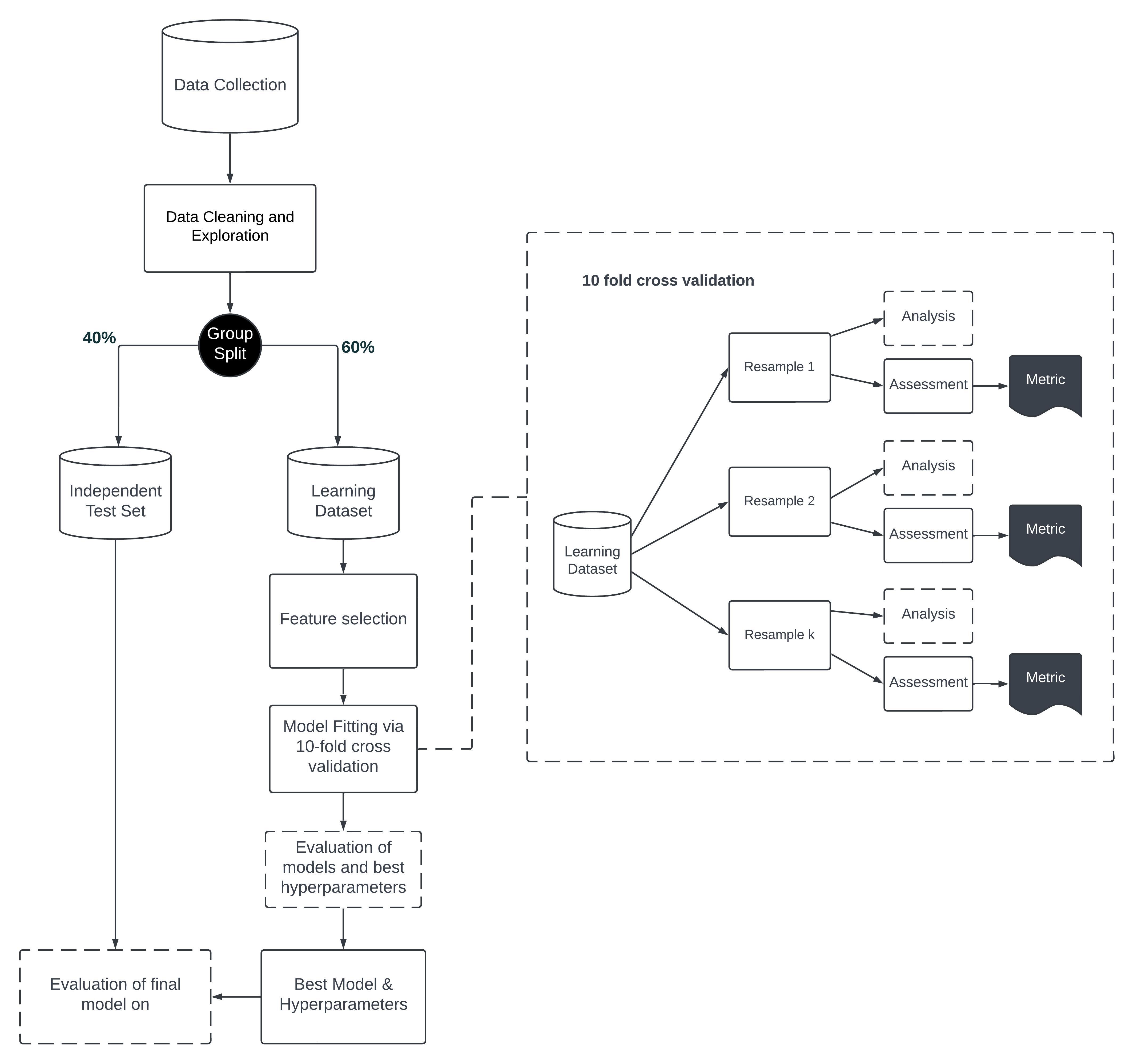
To ensure the robustness and generalizability of our models, data will be randomly split by groups (client ID). This method ensures that clients in the training set, which is used to train the models, are not included in the test set, thus preventing data leakage and providing a more unbiased evaluation of model performance ([Figure 1](#fig-client-selection)).

The training set will be used to train the models using 10-fold group cross-validation. In group cross-validation, the 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 grouped 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 realistic measure of how the models will perform in real-world scenarios. This step is crucial for assessing the models’ generalizability and for identifying any overfitting that may have occurred during training.

Figure 3

Experimental procedure



### 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. 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 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, our study’s cross-sectional nature, focusing on data collected at intake, may not capture the dynamic changes in patient 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

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>

CMHO. (2019). *Developing caseload/workload guidelines for ontario’s child and youth mental health sector*. Children’s Mental Health Ontario. <https://www.cmho.org/wp-content/uploads/2019/05/CYMH-LAC-Report-2019.pdf>

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.cymhlac.ca/wp-content/uploads/2019/05/CYMH-LAC-Report-2019.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.health.govt.nz/publication/new-zealand-mental-health-classification-and-outcomes-study-final-report>

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>

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>

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>

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., & Babcock, S. E. (2020). InterRAI child and youth mental health-screener (ChYMH-s): A psychometric evaluation and validation study. *Child Psychiatry and Human Development*, *51*(5), 769–780. <https://doi.org/10.1007/s10578-020-01003-7>

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

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>