Predicting Out of Home Placement

A research report on recorded behavioral health placements

ODH

Abstract

Can administrative and clinical records help target limited assessment capacity toward youth likely to enter behavioral health out-of-home care? We study 15,332 New Jersey youth with pre-2015 Child and Adolescent Needs and Strengths (CANS) assessments in a PerformCare cohort selected through service involvement during 2014–2015. We use assessments, demographics, diagnoses, medications, and eligibility records dated before January 1, 2015 to predict recorded admission starts through November 2, 2016, a 672-day window with no lead gap. Lasso models are tuned within an 80% development sample and evaluated on held-out youth. Among the 3,067 test youth, selecting the 100 highest scores identifies 28 youth with a recorded placement, compared with 5.9 expected under random selection, 23 using prior placement history, and 19 using CANS and demographics alone. The top 100 captures 15.4% of test youth with a placement; the gain over the history rule is 5 (95% CI -5 to 15) youth. At the same 100-youth capacity, a calendar-year 2015 analysis also finds higher observed yield than these comparators. The model concentrates recorded placement risk, but its incremental value over a simple history rule at this capacity is uncertain. These are gains in case finding, not estimates of placements prevented or welfare improved. Retrospective cohort selection and uncertain admission completeness limit prospective interpretation.

1 Introduction

What are the factors associated with a youth being placed Out Of Home (OOH)? What explains how long a youth will be placed OOH? And what are the factors associated with a youth being readmitted to OOH? We approach these questions through recorded placements, multiple starts, and occupied days within a fixed window, rather than clinical readmission or individual episode length. We use rich administrative and clinical data on 22,282 New Jersey youth from PerformCare. We have data on youths' socio-demographics, their race and ethnicity, gender, and age, etc., clinical assessments of their needs and strengths (from the Child and Adolescent Needs and Strengths or CANS assessments), diagnostic data codes (from PerformCare's internal system), and the medications that have been prescribed to them, among other such things. This rich data gives us an opportunity to understand the factors that are correlated with OOH utilization.

To shed light on factors that predict a youth being placed OOH in a defined period in the future, we split the data into an observation window and a prediction window. We use information recorded before January 1, 2015 to predict recorded placements from that date through November 2, 2016, a 672-day window. We also examine multiple recorded placements during this window and recorded occupied days. Multiple placements within a fixed window are not the same as readmission following discharge, and occupied days per youth are not the length of an individual stay.

Among youth held out from model development, the placement model has an area under the receiver operating characteristic curve (ROC AUC) of 0.79 (95% CI 0.75 to 0.82). Its precision-recall AUC is 0.211, against an event prevalence of 5.9%. These results describe discrimination within a selected historical cohort. They do not establish that the model can be applied prospectively to all youth served by the system. Cohort selection overlaps the prediction period, historical values of the retained predictors cannot all be verified, and the admission extract does not establish complete ascertainment of ongoing placements.

The rest of the report is organized as follows: we begin by describing the data. We next describe how CANS and other recorded characteristics relate to OOH placement. We then use CANS, demographic, diagnostic, medication, and eligibility measures to fit separate predictive models for the three outcomes. Given the data are high-dimensional, we fit penalized models. We cross-validate the models for tuning the hyper-parameters, with feature selection and imputation repeated within each training fold. We evaluate the fitted models on held-out youth and end by discussing the limits of causal interpretation.

1.1 Related research

Research on out-of-home care has examined how assessments of needs and strengths relate to placement decisions and subsequent outcomes. Chor et al. (2012) compared multidisciplinary team decisions with a CANS-based decision support algorithm for children in residential treatment in Illinois. Agreement between the two approaches was associated with greater clinical improvement, but not differences in length of stay. A subsequent longitudinal study examined agreement between these approaches across placement types and its association with changes in children's well-being (Chor et al. 2015). Both studies concerned children in the child welfare system.

This report examines recorded behavioral health placements among New Jersey youth in the PerformCare extract. It asks whether information recorded before a common forecast date predicts subsequent placement starts and occupied days.

Prediction can inform policy when a decision depends on an unknown future outcome (Kleinberg et al. 2015). Here the decision is whom to assess first when assessment capacity is limited. We compare rankings at the same capacity and distinguish their ability to find youth with future recorded placements from the benefits of offering those youth additional services.

2 Data and forecasting design

2.1 Sample and ascertainment

PerformCare supplied records for 22,282 youth selected through behavioral health OOH use or care management organization involvement during 2014–2015. The available sampling descriptions differ on whether qualifying OOH stays had to begin during those years or could overlap them. The cohort excludes several other settings, including developmental disability and substance use facilities, group homes, treatment homes, detoxification facilities, and SJI facilities. The outcomes here concern the behavioral health facilities represented in this extract, not every kind of out-of-home care.

The sampling criteria partly use events in the prediction period. Consequently, this is a retrospectively selected cohort. Holding out people addresses fitting optimism within that cohort, but does not remove the selection or establish performance in a prospectively enumerated service population.

We limit the analytic universe to youth with a pre-origin CANS assessment. The resulting analytic cohort contains 15,332 youth; assessment availability does not guarantee complete follow-up.

Table 1: Cohort construction and recorded outcomes. Event counts use the primary window.

Stage	Youth
Members in extract	22,282
Members with pre-origin CANS	15,332
Excluded without pre-origin CANS	6,950
Recorded placement	907
Multiple recorded placements	178
Unusable occupied-day outcome	6

All source admission rows have discharge dates. The supplied files do not establish whether ongoing placements were excluded at extraction. We therefore predict recorded completed placements. Absence of a matching admission means no event recorded in this extract; it does not verify that the youth had no placement. This uncertainty cannot be repaired by treating the available rows as ordinary right-censored spells, because an omitted spell has no observed start in this file.

Admission dates run from January 3, 2011 to November 2, 2016; discharge dates extend to November 29, 2016. The primary endpoint is the last recorded admission date. It defines the analysis window but is not proof of complete coverage. The window is January 1, 2015 inclusive to November 3, 2016 exclusive. Sensitivity analyses use 180 days, 365 days, and the 699 calendar days through November 29, 2016. None of these windows supplies a missing ascertainment guarantee. Records with earlier event dates may have been entered later; this is retrospective outcome ascertainment.

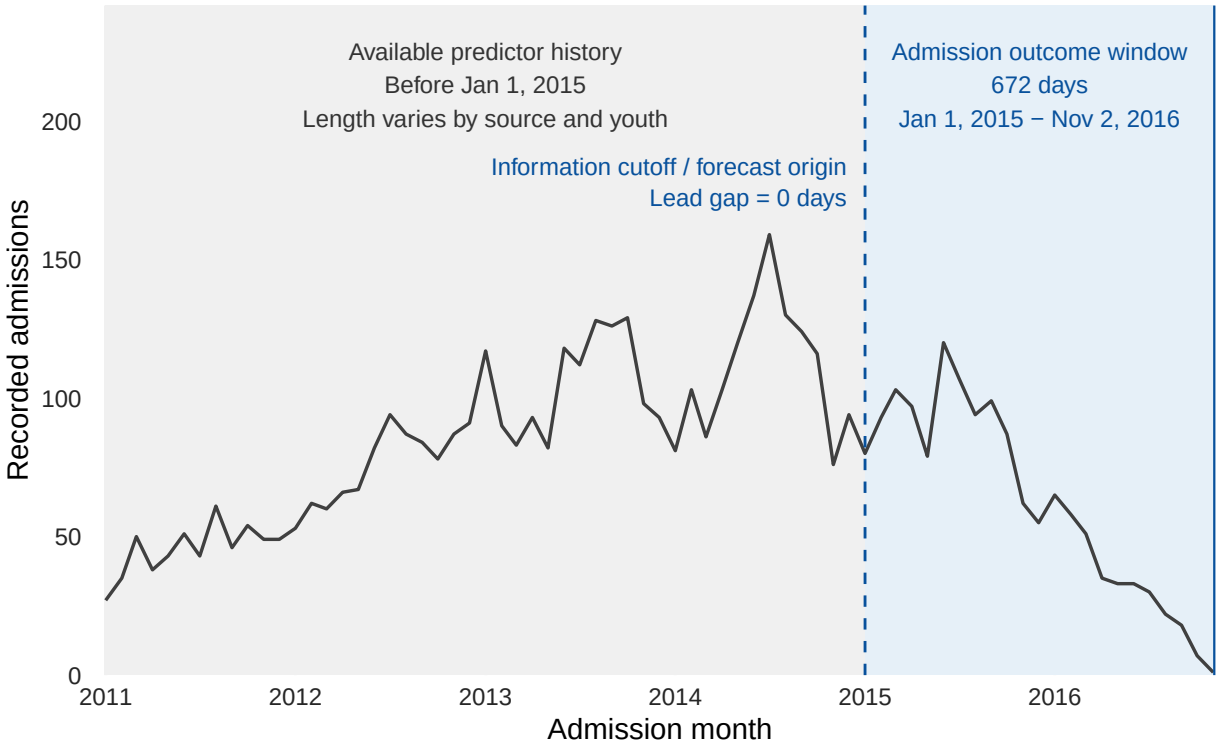


Figure 1: Recorded admission counts by month in the full extract. Gray shading marks the displayed period before the information cutoff; blue shading marks the admission outcome window. The dashed line is the January 1, 2015 cutoff and forecast origin, with no lead gap. The outcome window ends on November 2, 2016. History length varies by source and youth; the development/test split is across youth, not calendar periods. The late decline may reflect extraction or completion selection as well as changes in utilization; these counts do not identify a population trend.

The decline in recorded starts during 2016 does not by itself demonstrate bias. These counts follow a cohort selected through service involvement in 2014–2015, rather than all youth entering the system each year. Changes in cohort composition and time spent receiving services could therefore change the counts without a comparable change in the system-wide admission rate. Exclusion of ongoing stays is another possible explanation: later admissions would have less time to end before extraction. The available records cannot distinguish these explanations from a real change in utilization. The supplementary information examines admission counts and stay lengths and repeats the fixed-origin evaluation for January 1–December 31, 2015.

The PerformCare data roughly span 2011–2016. In the analytic cohort, 61.7% of youth are male and 27.4% are recorded as Hispanic or Latino. The source codes 35.8% as other ethnicity; we retain that category separately from recorded non-Hispanic identity. Race is unavailable or declined for 7.9%. Among youth with recorded race, 36.3% are White and 38.9% are Black. Average age at the forecast origin is 14.1 years.

2.2 Outcomes

Recorded placement indicates at least one admission start in the prediction window. It includes both first recorded placements and subsequent placements. We use the admission date directly; a

missing or invalid duration does not turn an observed admission into a non-event.

Multiple recorded placements indicates at least two distinct admission dates within the window. It describes repeated starts during a common forecasting interval. A youth with one admission in the window and an earlier admission is negative on this outcome. In the analytic cohort, 308 youth have exactly this history. The records do not distinguish transfers, continuations, and returns after discharge. A clinical readmission analysis would need a discharge-based origin, a defined time at risk, and an episode-linkage rule; it is not identified by relabeling this outcome.

Recorded occupied days counts the union of valid admission-to-discharge intervals intersecting the prediction window, including stays that began before the forecast origin. Admission days are included and discharge days excluded. Overlaps are counted once, so a youth cannot contribute more days than the window contains. There are 992 valid source stays that begin before the origin and continue into 2015. Including carryover stays and clipping durations to the endpoint measures utilization during the stated window. It is not the completed length of stay conditional on an admission.

The admission file contains 1 exact duplicate and 8 records with discharge before admission. We remove the exact duplicate, retain the starts for binary placement outcomes, and mark occupied days unavailable for any cohort member with an invalid admission record starting before the endpoint, including historical records. This conservative rule does not assume that an impossible discharge date rules out a carryover stay. This excludes 6 youth from occupied-day modeling. We do not convert their unknown duration to zero.

2.3 Predictors and their availability

CANS assessments must have finish, submission, and creation dates before January 1, 2015. This rule retains 99,803 assessments; using finish date alone would retain 99,804. CANS observations available to this analysis begin in 2012, so a uniform three-year observation period is not available for every source or youth. We use each youth's available pre-origin history rather than impose an unsupported common exposure duration.

CANS measures children's needs and strengths on a variety of important dimensions. Life domain functioning is a general assessment of how the youth is doing medically, academically, vocationally, socially, etc. Youth strengths is a general psycho-social assessment, including rating questions such as whether the youth has well-developed interpersonal skills and friends. Caregiver strengths measures aspects of supervision, involvement, knowledge, organization, social resources, and residential stability. Youth risk behavior covers self-harm, delinquency, fire-setting, and related risks. The remaining modules concern acculturation, developmental needs, trauma, substance use, juvenile justice, sexual development, and sex-related problems.

We average the ratings within each module, and again average the ratings over available assessments for each youth. A higher score generally denotes greater needs or fewer strengths on the source's zero-to-three coding. The caregiver and youth needs aggregate combines caregiver needs with youth behavioral and emotional needs. The sexual development and sex-related problem labels follow their item content. These are descriptive domain averages, not validated latent-scale estimates. We also include the number of qualifying pre-origin CANS assessments per youth. This records assessment frequency and may reflect duration of contact with the system as well as clinical needs.

Following the assessment coding guidance, we recode special values -1 and 4 to zero for the main analysis. We preserve ordinary missing values and average observed items; an entirely unobserved

module remains missing. As a sensitivity analysis, we treat -1 and 4 as missing. Zero represents no need on needs items and a strong asset on strengths items, so this recoding does not mean absence of a strength. This alternative changes the universe of scored assessments for optional modules and is not an interchangeable estimate of the same mean. Some questions were introduced later, so item availability varies by assessment date.

Socio-demographics comprise age at the forecast origin, gender, and indicators for White, Black, and Hispanic or Latino identity. Birth dates are known to the month; age uses the first day of that month. Unknown or declined race remains missing rather than being coded as a known absence of a racial identity. The source ethnicity code "Other" has a separate indicator, so it is not pooled with recorded non-Hispanic identity. These characteristics come from a later member snapshot, so their historical values cannot all be verified independently.

CYBER diagnostic codes and medication records must have creation dates before the origin. Eligibility records must have start dates before the origin; their entry dates are not supplied. The pre-origin CYBER data contain 917,452 records. Medication records with missing or later creation dates are excluded. Source updates and the absence of version histories prevent a complete reconstruction of what every record contained on January 1, 2015.

Claims provide a second source of diagnostic information. However, all supplied claims creation and payment dates are in 2016, including claims with earlier service dates. We therefore exclude claims from forecasting predictors.

Table 2: Source coverage. Pre-origin counts apply the stated date rules; they do not establish historical snapshot integrity.

Source	All records	Pre-origin records
CANS	193,329	99,803
CYBER diagnoses	1,447,843	917,452
Medications	354,533	155,075
Eligibility	74,294	55,286

For diagnoses, medications, and eligibility we use indicators for at least one retained record. Zero means no matching pre-origin record in the extract, not confirmed absence of a condition, prescription, or entitlement. Common categories are selected using only the development data: at least 0.5% of training youth, with a minimum of 25 youth. It is recomputed in every internal training fold.

3 Descriptive relationships

We expect youth who are placed OOH to be worse on many of the CANS dimensions than youth who are not placed OOH. For a broad but useful view, we compare the youth-level module means by recorded placement during the prediction window. These are descriptions of the full selected analytic cohort, not estimates of causal effects or a separately validated ranking of predictors.

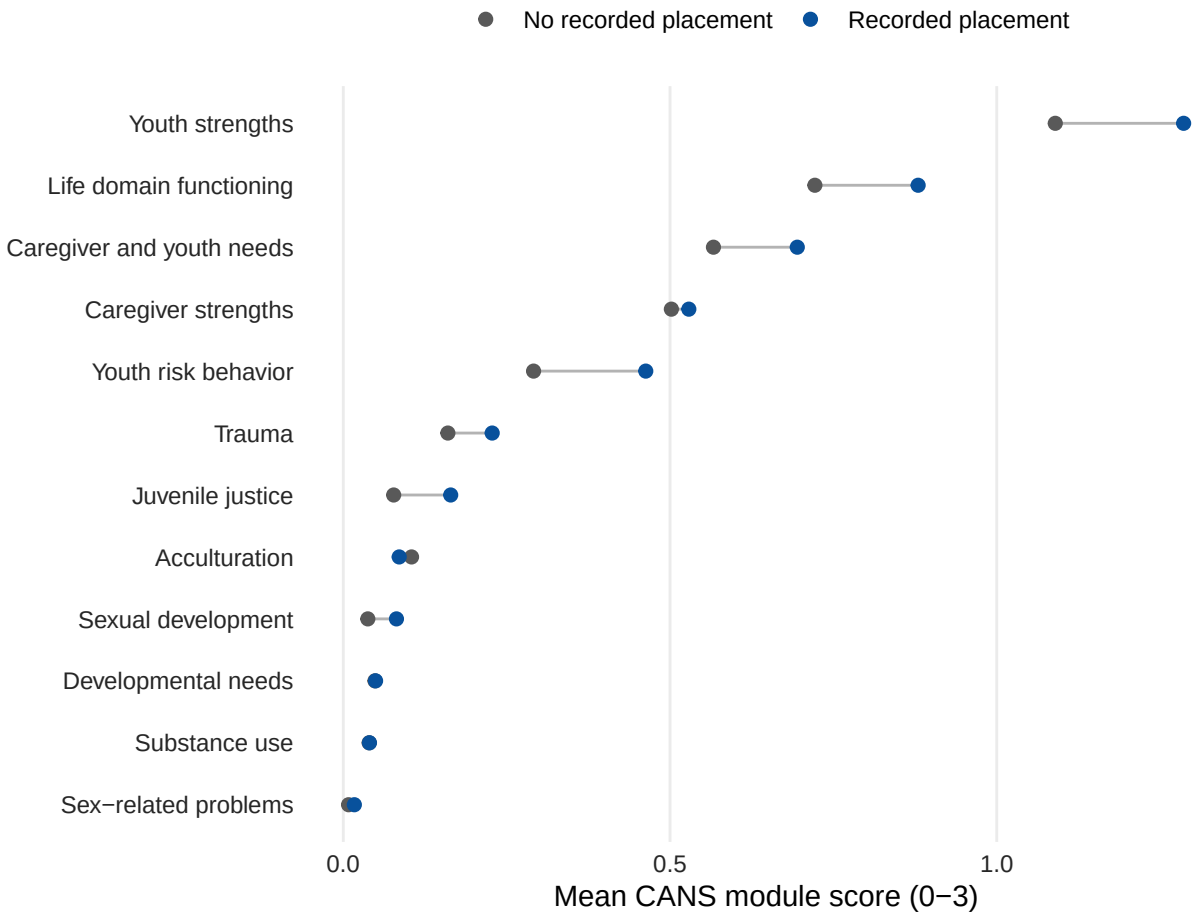


Figure 2: Mean pre-origin CANS scores by recorded placement during the primary window, among youth with available module scores. Each point is an equally weighted mean of youth-level assessment averages. Special values −1 and 4 are recoded to zero. Higher scores generally denote greater needs or fewer strengths.

The comparisons use information recorded before the forecast origin. Their interpretation remains specific to the facilities and youth included in the extract. In particular, exclusion of specialized substance use and developmental disability settings can make an association with behavioral health placement differ from an association with all out-of-home care. A negative coefficient on a related measure would not establish that the condition protects against placement.

Next, we track how diagnostic codes, medications, and eligibility differ between youth with and without recorded OOH placement in the prediction period. Appendix tables report member-level prevalence for common categories using the same pre-origin indicators as the models. These describe recorded histories; zero does not confirm the absence of a diagnosis, medication, or eligibility. Claims do not enter these comparisons because their availability at the forecast origin is not established.

3.1 Placement frequency across recorded characteristics

Recorded placement becomes more common as life domain functioning scores indicate greater needs: 1.5% of youth with average scores below 0.5 have a placement in the prediction window, compared

with 11.2% of those scoring at least 1. Placement is also more common among youth with prior recorded admission starts: 4.0% with none, 13.5% with one, and 17.5% with two or more. Prior starts must have occurred and been entered before January 1, 2015. They describe the available history, not a complete account of earlier care.

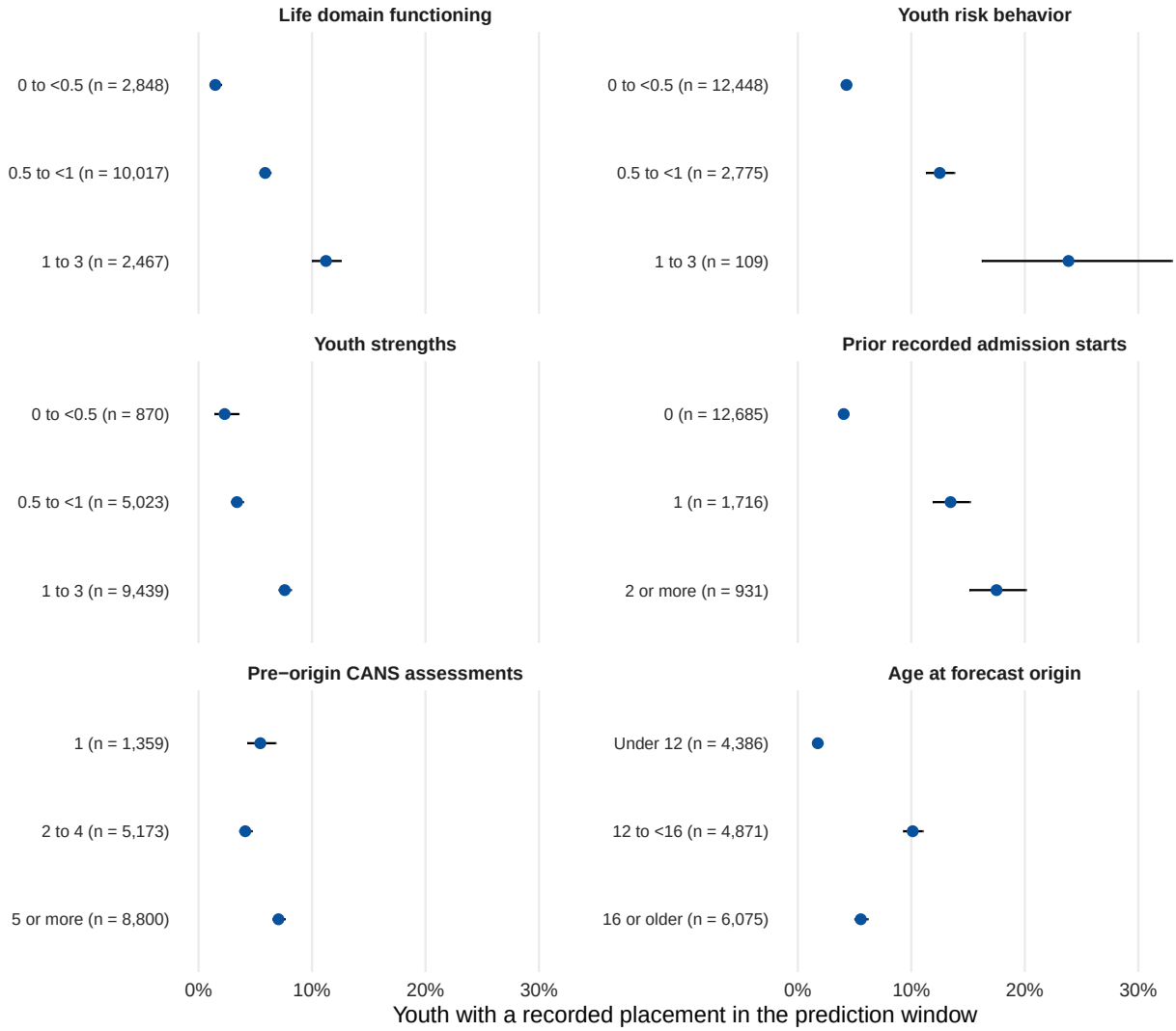


Figure 3: Recorded placement frequency during the primary prediction window by pre-origin characteristics. Labels give the number of youth in each group; bars are conditional 95% exact binomial intervals. CANS scores are youth-level averages; higher scores indicate greater needs or fewer strengths. Groups are descriptive, with no adjustment for other characteristics or cohort selection.

Age and assessment frequency show different patterns. Recorded placement is most common among youth aged 12 to under 16 (10.1%), compared with 1.8% under age 12 and 5.5% at age 16 or older. Assessment frequency does not have a simple increasing relationship with placement. More assessments can reflect clinical needs, time in the system, or documentation practices.

3.2 Recorded medications

Medication histories are part of the analysis. The figure compares eight common entries selected by their prevalence among development youth, without using their relationship with subsequent placement. For the recorded "Risperdal 1 mg" entry, prevalence is 7.3% among youth with a placement during the prediction window and 2.4% among those without one. Such differences can reflect the conditions for which medications were prescribed, their severity, and contact with services. They do not establish medication effects.

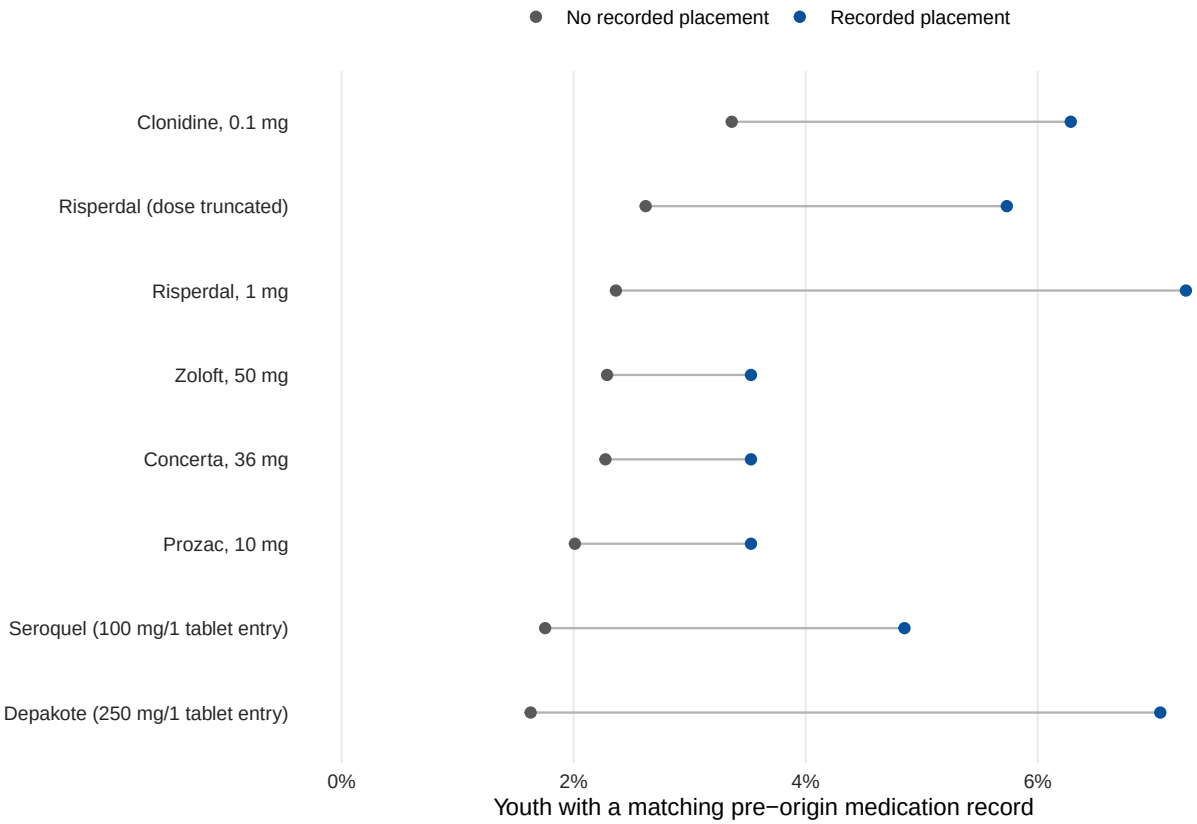


Figure 4: Common pre-origin medication entries by recorded placement during the primary window. Percentages use all youth in each outcome group as denominators; a youth can appear under several entries. Entries are selected and ordered by development-set prevalence. Labels abbreviate the source entries; one Risperdal dose is truncated in the source. A matching record does not establish current use, adherence or a complete medication history.

4 Prediction models and validation

Our main analyses rely on penalized regression models. For the binary placement outcomes we use logistic regression with an L1 penalty, or lasso. For occupied days we use linear regression with an L1 penalty, with predictions bounded below at zero. We rely on lasso regression for reasons to do with producing a sparse feature set. Penalized coefficients are conditional predictive associations; their raw magnitudes across differently scaled predictors are not a common importance scale.

We randomly assign 80% of youth within each primary-window outcome stratum (no placement,

one placement, or multiple placements) to development and the remainder to testing. The development set contains 12,265 youth and the test set contains 3,067. All three outcomes and all sensitivity specifications use the same person-level split. Within development, binary-model folds are stratified by their respective outcomes. Five-fold cross-validation selects the penalty; occupied-day models also allow an unpenalized fit. Diagnosis, medication, and eligibility vocabulary selection, missingness indicators, training medians, and zero-variance filtering are fitted within each internal training fold. Binary models minimize held-out log loss; occupied-day models minimize held-out squared error. The selected specification is refitted on all development youth before predicting the test set. We fit the models in R using glmnet (Friedman et al. 2010).

We report ROC AUC, precision-recall AUC, Brier score, and observed versus predicted event frequencies. Precision-recall measures complement ROC AUC by describing how often positive predictions are correct, which is especially relevant when placements are uncommon (Saito and Rehmsmeier 2015). For one illustrative classification threshold, we maximize Youden's sensitivity-plus-specificity criterion using development cross-validated predictions. This threshold is fixed before testing. It is not a clinical decision rule or a recommendation about relative false-positive and false-negative costs. For probability accuracy, the baseline assigns everyone the development event rate.

To evaluate limited assessment capacity, we rank test youth by predicted placement risk and select the top 50, 100, 200, or 500. We compare these lists with equally sized lists from the CANS-and-demographics model and a simple history rule: more recorded pre-origin admission starts first, then more recent starts. Prior admissions must both begin and be entered before January 1, 2015. Remaining ties are broken randomly without reference to outcomes. Random selection provides a further benchmark, with expected yield equal to capacity times test prevalence. The history rule is a transparent comparator, not an estimate of current clinical practice. Paired bootstrap intervals resample test youth and reconstruct each ranking at the same capacity; they condition on the fitted models and tie rule.

For occupied days, the squared-error comparator predicts the training mean for every test youth. The absolute-error comparator predicts the training median. The median is appropriate for absolute loss, not as the optimal constant under squared loss. We use 1,000 person-level bootstrap resamples of the held-out predictions for percentile intervals. Classification rates and prevalence use exact binomial intervals conditional on their observed denominators. All intervals condition on the fitted model and split; they do not incorporate training-sample, extraction, model-selection, or population-selection uncertainty. They also do not account for dependence among youth who share a household or provider; household membership cannot be established reliably from these files.

5 Results

5.1 Recorded placement and multiple placements

Table 3: Held-out performance with conditional 95% intervals: percentile bootstrap for AUCs and Brier score, exact binomial for classification rates. Classification uses the threshold selected within development.

Metric	Recorded placement	Multiple recorded placements
ROC AUC	0.788 [0.753, 0.822]	0.849 [0.798, 0.894]
Precision-recall AUC	0.211 [0.167, 0.268]	0.050 [0.030, 0.089]
Brier score	0.051 [0.045, 0.058]	0.012 [0.008, 0.015]
Accuracy	0.653 [0.636, 0.670]	0.610 [0.593, 0.627]
Sensitivity	0.786 [0.719, 0.843]	0.917 [0.775, 0.982]
Specificity	0.644 [0.627, 0.662]	0.606 [0.589, 0.624]
Precision	0.122 [0.104, 0.142]	0.027 [0.019, 0.038]

The test set contains 182 youth with at least one recorded placement and 36 with multiple placements. Their respective frequencies are 5.9% and 1.2%. Placement discrimination is 0.79 (95% CI 0.75 to 0.82); discrimination for multiple placements is 0.85 (95% CI 0.80 to 0.89).

5.1.1 What does selecting the top 100 deliver?

Among 100 youth with the highest placement scores, 28 have at least one recorded admission during the 672-day window. Randomly selecting 100 from the same test group would find about 5.9 such youth on average. The model therefore finds about 22 more at the same capacity, or 4.7 times the random-selection yield. The top 100 captures 15.4% of the 182 youth with a recorded placement. It leaves 154 outside the list, while 72 selected youth have no recorded placement.

Table 4: Youth with at least one recorded placement among equally sized selected groups from the same test sample. The outcome window is January 1, 2015 through November 2, 2016. Each method selects its own list of youth. Random selection is an expectation; the other columns are observed counts, not percentages or admission episodes.

Youth selected	Full model	CANS + demographics	Prior placement rule	Random: expected
50	14	12	14	3.0
100	28	19	23	5.9
200	58	42	40	11.9
500	95	86	79	29.7

The more demanding comparison is with a simple rule based on prior placements. Its top 100 includes 23 youth with a later recorded placement, compared with 28 for the full model: a gain of 5 (95% CI -5 to 15) youth. CANS and demographics alone find 19, giving a full-model gain of 9 (95% CI 0 to 17) youth. Thus, the large gain over random selection does not establish a clear improvement over the history rule at this capacity. At capacity 200, the full model finds 58 youth with a placement, compared with 40 using history: a gain of 18 (95% CI 5 to 29) youth. The incremental yield therefore depends on assessment capacity. These comparisons count youth with recorded placements, not the number of admission episodes or placements prevented.

The development-selected classification threshold answers a different capacity question. It flags 1,169 youth (38.1% of the test set), of whom 143 have a placement, and misses 39 youth with a

placement. Its 78.6% sensitivity and 12.2% precision therefore require assessing far more than 100 youth. A service with a fixed assessment budget should judge the ranking at that capacity rather than adopt this threshold by default.

For multiple placements, mean predicted risk is 1.22% and observed test prevalence is 1.17%. Development prevalence is 1.16%. Stratification balances the primary-window outcome categories across development and testing; it does not guarantee calibration or address cohort selection.

Relative to predicting the development event rate for everyone, the reduction in Brier score is 7.8% (95% CI 4.6% to 10.9%) for placement and 0.7% (95% CI -2.2% to 2.9%) for multiple placements. The latter interval does not establish an improvement in probability accuracy despite the model's ROC AUC. These are paired person-bootstrap intervals conditional on the fitted model and development prevalence.

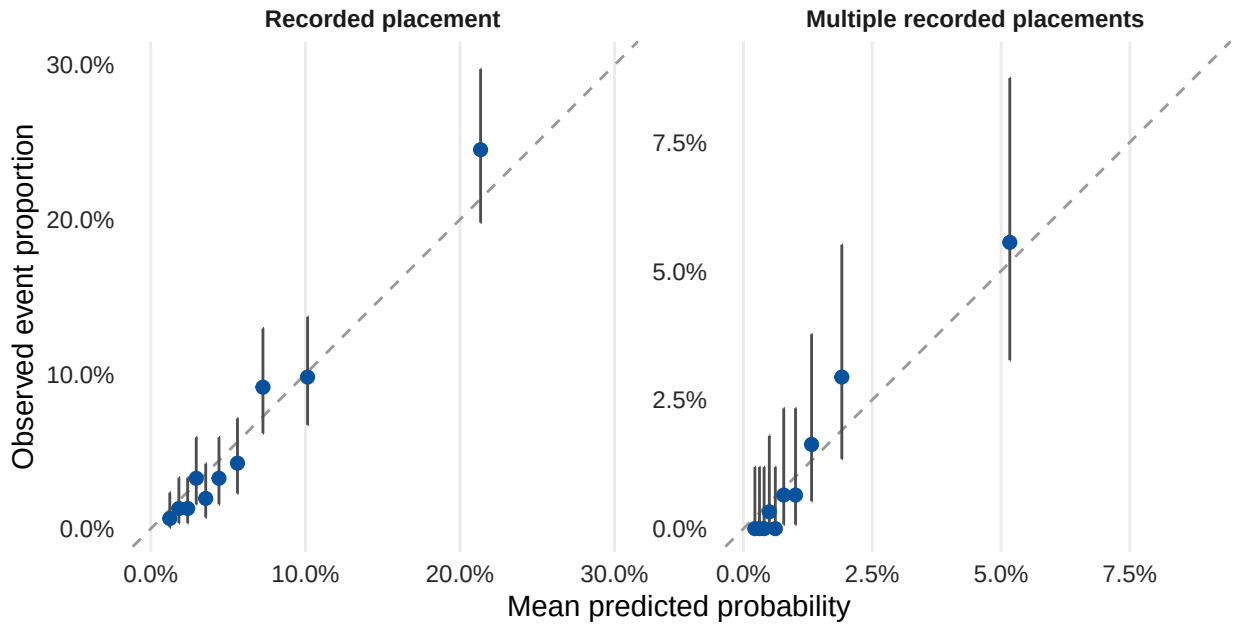


Figure 5: Calibration on held-out youth, separately for the two binary outcomes. Points compare mean predicted probability with the observed event proportion within ten equal-count prediction bins. Vertical intervals are exact 95% binomial intervals conditional on those bins; the dashed line marks equality. Axis ranges differ across outcomes to make rare-event calibration readable; the horizontal and vertical ranges match within each panel. These estimates concern the selected cohort and recorded outcomes.

5.2 Recorded occupied days

Table 5: Occupied-day prediction among held-out youth with usable duration outcomes. RMSE is root mean squared error; MAE is mean absolute error.

Metric	Estimate	95% lower	95% upper
Model RMSE (days)	87.66	81.64	93.83
Model MAE (days)	43.80	41.19	46.47
Training-mean baseline RMSE (days)	95.43	88.27	102.38
Training-median baseline MAE (days)	28.26	24.92	31.62
Fractional MSE reduction versus training mean	0.16	0.12	0.19

The model's root mean squared error is 87.7 days, compared with 95.4 days for the training-mean predictor. The reduction in mean squared error is 15.6%. Mean absolute error is 43.8 days for the model, versus 28.3 days for the training-median baseline. The model performs worse on absolute error despite improving squared error; the latter gives more weight to large errors. This comparison concerns total recorded occupied days during the window, including zeros for youth without recorded occupancy. It does not measure prediction of an admitted youth's eventual discharge date.

5.3 Horizon and coding sensitivity

Table 6: Occupied-day sensitivity analyses with conditional 95% bootstrap intervals. MSE reduction is a fraction relative to the training-mean baseline. Horizon changes alter the outcome.

Setting	RMSE (days)	MAE (days)	MSE reduction
Main: 672 days	87.7 [81.6, 93.8]	43.8 [41.2, 46.5]	0.156 [0.121, 0.190]
CANS + demographics	89.5 [83.2, 95.8]	45.5 [42.9, 48.2]	0.120 [0.089, 0.148]
Special codes missing	87.6 [81.5, 93.8]	43.7 [41.1, 46.4]	0.158 [0.124, 0.192]
180 days	38.2 [36.0, 40.2]	19.0 [17.8, 20.1]	0.196 [0.160, 0.227]
365 days	64.9 [61.2, 68.5]	33.0 [31.1, 35.0]	0.180 [0.146, 0.211]
699 days	87.9 [81.8, 94.0]	43.9 [41.3, 46.6]	0.156 [0.120, 0.189]

Table 7: Recorded placement sensitivity analyses: test events and conditional 95% bootstrap intervals. Intervals describe individual estimates, not paired differences between models.

Setting	Events	ROC AUC	PR AUC	Brier
Main: 672 days	182	0.788 [0.753, 0.822]	0.211 [0.167, 0.268]	0.051 [0.045, 0.058]
CANS + demographics	182	0.767 [0.731, 0.802]	0.165 [0.133, 0.213]	0.053 [0.046, 0.060]
Special codes missing	182	0.792 [0.758, 0.824]	0.210 [0.166, 0.266]	0.051 [0.045, 0.058]
180 days	94	0.800 [0.758, 0.840]	0.114 [0.079, 0.173]	0.029 [0.023, 0.034]
365 days	151	0.792 [0.754, 0.826]	0.167 [0.129, 0.220]	0.044 [0.037, 0.050]
699 days	182	0.788 [0.753, 0.822]	0.211 [0.167, 0.268]	0.051 [0.045, 0.058]

Table 8: Multiple recorded placements sensitivity analyses: test events and conditional 95% bootstrap intervals. Intervals describe individual estimates, not paired differences between models.

Setting	Events	ROC AUC	PR AUC	Brier
Main: 672 days	36	0.849 [0.798, 0.894]	0.050 [0.030, 0.089]	0.012 [0.008, 0.015]
CANS + demographics	36	0.797 [0.727, 0.867]	0.057 [0.026, 0.143]	0.011 [0.008, 0.015]
Special codes missing	36	0.839 [0.790, 0.885]	0.046 [0.027, 0.084]	0.012 [0.008, 0.015]
180 days	7	0.805 [0.665, 0.889]	0.006 [0.002, 0.013]	0.002 [0.001, 0.004]
365 days	16	0.901 [0.861, 0.940]	0.035 [0.016, 0.077]	0.005 [0.003, 0.008]
699 days	36	0.849 [0.798, 0.894]	0.050 [0.030, 0.089]	0.012 [0.008, 0.015]

For recorded placement, the full model's ROC AUC is 0.788, compared with 0.767 for CANS and demographics alone. Precision-recall AUCs are 0.211 and 0.165, respectively; Brier scores are 0.0515 and 0.0528. These metrics assess different aspects of prediction. The paired comparisons and repeated-split analysis below assess the full-versus-reduced differences directly. The 180-day multiple-placement result rests on only 7 test events, and the 365-day result rests on 16. No admissions are recorded after the primary endpoint, so extending the window to 699 days leaves both binary outcomes unchanged; it can change occupied days.

The shorter horizons answer different forecasting questions, with fewer opportunities for an event. Treating special CANS values as missing changes which assessments contribute to module means. Removing diagnosis, medication, and eligibility indicators changes the predictor set. These comparisons do not test performance in a new calendar cohort or service system.

5.4 Model comparison across splits

We compare the full and reduced binary models on matched test youth, then repeat the complete development, tuning, and evaluation procedure over 20 stratified splits, including the primary split. Each model uses development-only preprocessing and tuning within each split. The two models share split membership and tuning folds.

Table 9: Paired comparisons on the primary test split. Positive gains favor the full model: full minus reduced for AUCs, reduced minus full for Brier score. Intervals condition on the fitted models and split.

Outcome	Measure	Full-model gain [95% CI]
Placement	ROC AUC	0.0217 [-0.0017, 0.0458]
Placement	PR AUC	0.0459 [0.0164, 0.0809]
Placement	Brier score	0.00131 [0.00009, 0.00252]
Multiple placements	ROC AUC	0.0512 [0.0079, 0.0947]
Multiple placements	PR AUC	-0.0071 [-0.0723, 0.0151]
Multiple placements	Brier score	-0.00018 [-0.00043, 0.00003]

Table 10: Full-model gains across 20 stratified splits. Positive gains favor the full model. The range describes observed resampling variation, not a confidence interval.

Outcome	Measure	Median gain	Range	Full better
Placement	ROC AUC	0.0343	[0.0142, 0.0473]	20/20
Placement	PR AUC	0.0271	[-0.0114, 0.0541]	19/20
Placement	Brier score	0.00099	[0.00009, 0.00178]	20/20
Multiple placements	ROC AUC	0.0365	[-0.0199, 0.0711]	17/20
Multiple placements	PR AUC	0.0029	[-0.0071, 0.0208]	12/20
Multiple placements	Brier score	0.00002	[-0.00019, 0.00019]	12/20

The full placement model improves ROC AUC in 20 of 20 splits, PR-AUC in 19, and Brier score in 20. For multiple placements, the corresponding counts are 17, 12, and 12. Added record indicators provide more consistent gains for placement; improvements for multiple placements depend more on the metric and split.

These repeated splits measure stability of the model comparison within the supplied cohort. Test sets overlap across repetitions, so the repetitions are not independent external validations. They do not correct cohort selection, incomplete ascertainment, or unavailable historical snapshots.

The supplementary information examines rolling forecasts: 365 days of history, a 30-day lead gap, and a 180-day admission window, with alternative histories, gaps and outcome windows. It trains models on earlier periods and evaluates them on later snapshots of held-out youth.

6 Discussion

With capacity to assess 100 youth, the full model identifies more youth with subsequent recorded placements than random selection. The incremental gain over prior placement history is smaller and uncertain. That comparison matters for deciding whether the added data collection, linkage, and model maintenance are worthwhile. A prospective assessment should compare equal-capacity allocation rules, including current practice, and measure the services delivered and their consequences.

A higher placement probability does not imply a larger benefit from intervention. Allocating services to improve welfare requires evidence about how treatment changes outcomes and what it costs (Kitagawa and Tetenov 2018). These records do not identify those effects. Some placements may be appropriate and beneficial, so reducing placements is not itself a sufficient welfare objective. The present estimates establish case-finding yield within the recorded cohort; they cannot establish cost savings, placements prevented, or which youth would benefit most from additional support.

The analyses that we have presented here are predictive but not causal. It is one thing to say that someone who scores poorly on "Life Domain Functioning" is likelier to be placed Out of Home, and another to say that poor "Life Domain Functioning" caused the child to be placed out of home. The difference is subtle but important. A child who scores poorly on "Life Domain Functioning" likely has a host of other problems, some diagnosed, some undiagnosed, and some not in the data. So, a positive coefficient on "Life Domain Functioning" does not mean that it is poor score on that, that is causing the child to be placed OOH. It could well be poverty or some other factor that we don't account for. The other problem that complicates inferences is that for variables that we do account for, we cannot be sure of the functional form. For all these reasons, it is important to interpret the results with caution.

The analysis supports discrimination among youth in this historical cohort, subject to the recorded-outcome definition. It does not validate a discharge-time readmission score, a model for eventual episode length, or a claim that applying the scores will prevent placements. The later snapshot, unavailable extraction query, outcome-dependent sample construction, and excluded facility types constrain the interpretation independently of model performance. Bootstrap intervals quantify only conditional test-sample uncertainty; they do not resolve these design limitations.

Next steps for analyses would include thinking about ways of including yet more data on the children, including how their families are doing, economically, medically, and psychologically, and what happens to children in the real world. Before prospective use, the immediate data requirements are a cohort enumerated at the forecast origin, historically available predictor snapshots, and a complete admission-and-discharge registry that retains ongoing stays. A discharge-based readmission question also requires a declared return window and a rule distinguishing transfers from new episodes.

References

Chor, Ka Ho Brian, Gary M. McClelland, Dana A. Weiner, Neil Jordan, and John S. Lyons. 2012. "Predicting Outcomes of Children in Residential Treatment: A Comparison of a Decision Support Algorithm and a Multidisciplinary Team Decision Model." *Children and Youth Services Review* 34 (12): 2345–52. https://doi.org/10.1016/j.childyouth.2012.08.016.

Chor, Ka Ho Brian, Gary M. McClelland, Dana A. Weiner, Neil Jordan, and John S. Lyons. 2015. "Out-of-Home Placement Decision-Making and Outcomes in Child Welfare: A Longitudinal Study." *Administration and Policy in Mental Health and Mental Health Services Research* 42 (1): 70–86. https://doi.org/10.1007/s10488-014-0545-5.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Kitagawa, Toru, and Aleksey Tetenov. 2018. "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice." *Econometrica* 86 (2): 591–616. https://doi.org/10.3982/ECTA13288.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95. https://doi.org/10.1257/aer.p20151023.

Saito, Takaya, and Marc Rehmsmeier. 2015. "The Precision-Recall Plot Is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLOS ONE* 10 (3): e0118432. https://doi.org/10.1371/journal.pone.0118432.

Appendix

Recorded diagnoses, medications, and eligibility

The following tables show up to eight of the most prevalent categories from each source among categories retained by the final development-set frequency rule. Ordering uses prevalence in the full analytic cohort for descriptive display; it does not feed model selection. Percentages have youth, not source records, as their denominators. Differences are percentage points, calculated as placement minus no placement. The eligibility code DYFS refers to the Division of Child Protection and Permanency. Entries such as "None Identified" and unspecified diagnostic codes are retained record categories, not confirmed conditions; they can reflect documentation practices. Medication labels are the recorded strings, which can be truncated and do not establish a harmonized drug or dose.

Table 11: Recorded diagnosis history by placement outcome in the analytic cohort.

Characteristic	No placement (%)	Placement (%)	Difference (pp)
SUPPGR: PROBLEMS WITH PRIMARY SUPPORT GROUP	71.2	80.5	9.3
NONE: None Identified	66.4	71.8	5.3
799.9: Other unknown and unspecified cause of morbidity and mortality	64.1	75.1	11.0
EDUCAT: EDUCATIONAL PROBLEMS	52.3	63.9	11.6
SOCENV: PROBLEMS RELATED TO THE SOCIAL ENVIRONMENT	44.0	59.5	15.5
313.81: Oppositional defiant disorder	41.7	63.1	21.4
OTHER: OTHER PSYCHOSOCIAL AND ENVIRONMENTAL PROBLEMS	36.0	42.8	6.8
V71.09: Observation for other suspected mental condition	35.0	41.2	6.2

Table 12: Recorded medication history by placement outcome in the analytic cohort.

Characteristic	No placement (%)	Placement (%)	Difference (pp)
clonidine 0.1 mg oral tablet (clonidine hydrochlor	3.4	6.3	2.9
risperdal 0.5 m	2.6	5.7	3.1
risperdal 1 mg	2.4	7.3	4.9
zoloft 50 mg oral tablet (sertraline hydrochloride	2.3	3.5	1.2
concerta 36 mg	2.3	3.5	1.3
prozac 10 mg oral capsule (fluoxetine hydrochlorid	2.0	3.5	1.5
depakote divalproex sodium 250 mg/1 oral tablet, d	1.6	7.1	5.4
seroquel quetiapine fumarate 100 mg/1 oral tablet,	1.8	4.9	3.1

Table 13: Recorded eligibility history by placement outcome in the analytic cohort.

Characteristic	No placement (%)	Placement (%)	Difference (pp)
NJFCF	43.8	29.9	-13.9
AFDC	27.8	20.1	-7.7
DYFS	21.8	35.6	13.8
SSI	21.7	27.5	5.8
PE	3.9	1.8	-2.2

Source date ranges and missingness

Table 14: Observed date ranges in source files. Extremes describe the supplied values, including errors; they do not establish coverage.

Field	First	Last
Admission	2011-01-03	2016-11-02
Discharge	2007-07-21	2016-11-29
CANS submitted	2012-03-01	2016-11-29
CANS finish	2012-03-01	2023-01-05
Claims service start	2008-11-25	2016-12-16
Claims service end	2008-11-25	2061-04-01
Claims creation	2016-02-01	2016-07-18
Claims payment	2016-02-03	2016-07-20
Medication creation	2012-11-14	2017-10-16

Table 15: Item missingness by CANS submission year in the full extract. Missingness varies across calendar years.

Year	Assessments	Sleep missing (%)	Bullied missing (%)	Gambling missing (%)
2012	16,248	100.0	100.0	100.0
2013	33,535	100.0	100.0	100.0
2014	50,020	7.5	7.5	7.5
2015	54,470	0.0	0.0	0.0
2016	39,056	0.0	0.0	0.0

Predicting Out of Home Placement

Supplementary information: forecast windows and rolling evaluation

ODH

Calendar-year 2015 endpoint

The decline in recorded admissions during 2016 raises a question about the longer outcome window. We therefore repeat the fixed-origin analysis using admissions from January 1 through December 31, 2015. The information cutoff remains January 1, 2015, with no lead gap. The same youth remain in development and testing, and the predictor histories are unchanged. Models are refitted and tuned within development for the one-year outcome. This comparison changes the outcome window, not the cohort.

Table 1: Placement prediction with a common January 1, 2015 origin and different endpoints. Both models are evaluated on the same test youth. Brackets give conditional 95% bootstrap intervals for each estimate, not for differences between windows.

Measure	Through Nov. 2, 2016	Calendar year 2015
ROC AUC	0.788 [0.753, 0.822]	0.792 [0.754, 0.826]
PR AUC	0.211 [0.167, 0.268]	0.167 [0.129, 0.220]
Brier score	0.051 [0.045, 0.058]	0.044 [0.037, 0.050]
Placement frequency	5.9%	4.9%

Among the 3,067 test youth, 151 have a recorded admission during 2015 (4.9%). ROC AUC is 0.792 and precision-recall AUC is 0.167. The top 100 one-year model scores identify 26 youth with a 2015 admission, compared with 4.9 expected under random selection, 18 using the prior-placement rule, and 14 using CANS and demographics. The paired gain over the history rule is 8 (95% CI -2 to 16) youth.

Table 2: Youth with at least one recorded admission during calendar year 2015 among equally sized selected groups. Rankings use pre-2015 information; model tuning uses only development youth.

Youth selected	Full model	CANS + demographics	Prior placement rule	Random: expected
50	10	10	12	2.5
100	26	14	18	4.9
200	43	36	31	9.8
500	76	71	63	24.6

At a smaller capacity of 50 youth, the history rule finds 12 youth with a 2015 placement and the full model finds 10. The model's relative yield therefore depends on assessment capacity.

The comparison changes both the outcome and the fitted score. To separate these changes, we also keep the longer-window model's ranking fixed and count only 2015 admissions. Its top 100 contains 21 youth with a 2015 admission, compared with 26 for the one-year model. The paired difference is 5 youth (95% CI -2 to 9). The longer-window model thus retains ranking information for the 2015 outcome. Refitting does not improve yield at every capacity, as the following comparison shows.

Table 3: Youth with a 2015 admission identified by the annual and longer-window models. Both use the same test youth and annual outcomes. Brackets give paired conditional 95% bootstrap intervals for annual minus longer-window yield.

Youth selected	One-year score	Longer-window score	Gain from refitting [95% CI]
50	10	10	0 [-4, 4]
100	26	21	5 [-2, 9]
200	43	47	-4 [-9, 4]
500	76	77	-1 [-6, 3]

These results do not establish an optimal forecast horizon. The one-year outcome window falls within the period used to select the cohort, so moving the endpoint earlier does not resolve outcome-dependent selection. An earlier endpoint also does not guarantee complete admission coverage.

Rolling forecasts: question and timing

Can a model fitted to an earlier period predict recorded admissions for other youth at a later date? We examine this question by moving the observation and outcome windows forward through the extract. This supplements the main report's fixed-origin analysis. It remains an analysis of recorded admissions in the selected PerformCare cohort.

Three time intervals define each forecast:

- **History:** records from the 365 days before the information cutoff.
- **Lead gap:** 30 days between the cutoff and the start of the outcome window.
- **Outcome window:** any recorded admission start during the next 180 days.

The information cutoff is the date when the prediction could be made from the retained records. It is not necessarily the date of the youth's last assessment. For example, the first outcome window begins on January 01, 2015; its information cutoff is December 02, 2014. An admission during the lead gap is not part of the target and does not enter the predictors. We do not remove youth because they had such an admission. The question concerns any recorded admission start during the target window, whether first or subsequent; it is not readmission after discharge.

A six-month outcome window is a useful planning question to examine, not an empirically established optimum. Taquet et al. (2025) used a 180-day measurement period to predict psychiatric hospitalization over the following 180 days. In a pediatric inpatient setting, Poweleit et al. (2026) evaluated psychiatric readmission within 30, 90, and 180 days after discharge. These studies concern different populations and outcomes. They motivate examining shorter windows but do not validate a particular horizon or a 30-day gap for the PerformCare cohort. Here, a year of history allows more time for intermittent assessments; the lead gap tests whether useful signal remains with advance notice. We also examine a 180-day history, no gap, and 90- and 365-day outcome windows, changing one quantity at a time.

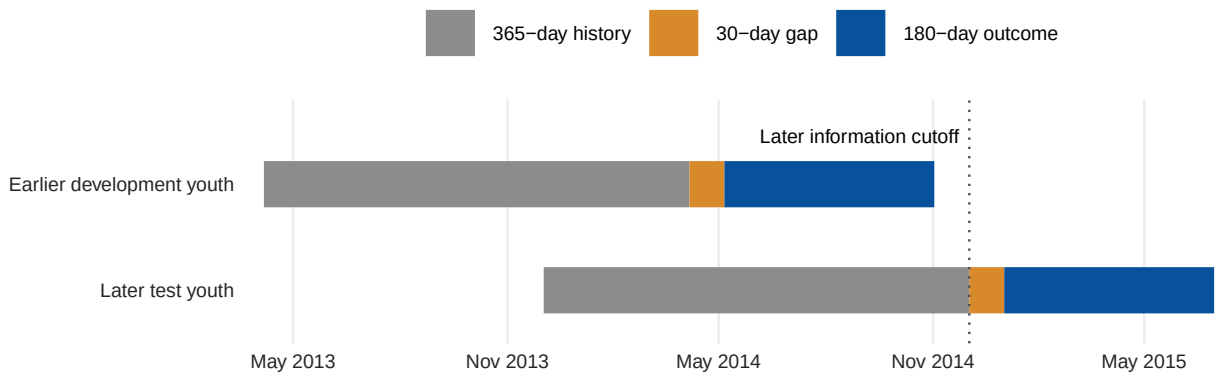


Figure 1: First default rolling forecast. Each row uses 365 days of history, a 30-day lead gap and a 180-day outcome window. The earlier training outcome window ends 30 days before the dotted line, the information cutoff for the later forecast. Development and test rows represent disjoint groups of youth.

Earlier training and later evaluation

We divide youth listed in the member roster into development and test groups once, without using their outcomes. CANS records for youth absent from that roster do not enter this rolling analysis. We randomly assign 80% to development. A youth remains on the same side at every date. This exploratory analysis uses a separate person-level split. At each quarterly target start, the model is trained on an earlier snapshot of development youth and evaluated on a later snapshot of test youth. Eligibility at each snapshot requires a CANS assessment finished in its history window and created and submitted before its cutoff. Eligibility and sample sizes can therefore change across dates and specifications.

The earlier training outcome window ends 30 days before the later information cutoff. This additional interval allows some time for admission records to be entered. Only training admissions with creation dates before that later cutoff count as positive training labels. The buffer does not guarantee that all admissions have been recorded. Evaluation uses recorded admission starts within the later target window.

Table 4: Rolling schedule for the 365-day history, 30-day gap, and 180-day outcome specification. Asterisks mark exclusive boundaries: dates on or after a cutoff or endpoint are excluded.

Training outcomes begin	Training outcomes end*	Information cutoff*	Target begins	Target ends*
2014-05-06	2014-11-02	2014-12-02	2015-01-01	2015-06-30
2014-08-04	2015-01-31	2015-03-02	2015-04-01	2015-09-28
2014-11-03	2015-05-02	2015-06-01	2015-07-01	2015-12-28
2015-02-03	2015-08-02	2015-09-01	2015-10-01	2016-03-29
2015-05-06	2015-11-02	2015-12-02	2016-01-01	2016-06-29
2015-08-05	2016-02-01	2016-03-02	2016-04-01	2016-09-28

Predictors include CANS module averages and assessment counts, demographics with age at each cutoff, recent diagnosis, medication and eligibility records, and the number and recency of prior recorded admission starts. CANS assessment finish dates must fall within the history window, with finish, creation and submission all before the cutoff. Diagnosis and medication entries use creation

dates; eligibility uses start dates because entry dates are unavailable. Prior admissions must start within the history window and have been entered before the cutoff. These features describe recorded histories, not confirmed active diagnoses, medication use, eligibility or complete placement histories. Later demographic snapshots and unavailable record versions remain limitations.

We fit logistic lasso models (Friedman et al. 2010). Each training snapshot has one observation per youth. Five-fold outcome-stratified cross-validation selects the penalty by log loss; category selection, imputation, missingness indicators and constant-column removal are repeated within each training fold. Models are then refitted on all development youth at that earlier date. Test youth never enter training, preprocessing or tuning. Each forecast uses one earlier training snapshot; the windows slide rather than accumulating all earlier observations.

We retain target windows ending no later than November 2, 2016, the last recorded admission date. We require the earlier training history to begin no earlier than November 14, 2012, the first medication creation date. This excludes the January 2015 target for the 365-day outcome specification; its final feasible target is October 2015. The date restriction avoids mechanically using a history preceding that source, but does not establish complete observation for each youth.

Results by forecast date

Across the default forecasts, ROC AUC ranges from 0.769 to 0.803. The fraction of test youth with a recorded admission declines from 3.2% in the first target window to 0.9% in the last. Mean predicted risk exceeds observed admission frequency at 6 of the 6 dates. Ranking performance therefore does not imply that the predicted probabilities are well calibrated to these later recorded outcomes.

Table 5: Later-period discrimination for the default rolling specification. Brackets give conditional 95% percentile bootstrap intervals from 1,000 resamples of test youth within each forecast date.

Target begins	Training youth	Test youth	Events	ROC AUC	PR-AUC	Prevalence
2015-01-01	8,392	2,792	89	0.774 [0.730, 0.820]	0.091 [0.065, 0.129]	3.2%
2015-04-01	9,658	2,857	81	0.794 [0.753, 0.838]	0.085 [0.060, 0.116]	2.8%
2015-07-01	10,620	2,953	71	0.799 [0.748, 0.847]	0.094 [0.061, 0.157]	2.4%
2015-10-01	11,230	2,969	57	0.769 [0.705, 0.826]	0.082 [0.044, 0.163]	1.9%
2016-01-01	11,479	2,952	38	0.785 [0.702, 0.854]	0.054 [0.029, 0.098]	1.3%
2016-04-01	11,753	2,870	27	0.803 [0.702, 0.890]	0.065 [0.029, 0.122]	0.9%

Table 6: Probability accuracy and average risk. The constant comparator predicts the earlier training prevalence for every test youth; smaller Brier scores indicate better probability accuracy.

Target begins	Lasso Brier	Constant Brier	Observed admissions	Mean predicted risk
2015-01-01	0.0305	0.0313	3.2%	4.7%
2015-04-01	0.0272	0.0276	2.8%	3.4%
2015-07-01	0.0227	0.0235	2.4%	2.9%
2015-10-01	0.0187	0.0190	1.9%	3.1%
2016-01-01	0.0133	0.0131	1.3%	2.9%
2016-04-01	0.0094	0.0095	0.9%	2.2%

Table 7: Assessment timing among test youth, measured from assessment finish to target start. The last column counts youth whose training outcome would become positive only using later- entered admissions.

Target begins	Days since last CANS: median [IQR]	Training positives unavailable at cutoff
2015-01-01	88 [58, 182]	1
2015-04-01	85 [56, 188]	3
2015-07-01	86 [56, 177]	2
2015-10-01	84 [56, 156]	1
2016-01-01	88 [59, 177]	0
2016-04-01	87 [59, 205]	2

Table 8: Paired probability-accuracy comparison. Improvement is constant-risk Brier score minus lasso Brier score on the same test youth; positive values favor the lasso.

Target begins	Brier improvement [95% CI]
2015-01-01	0.00086 [-0.00005, 0.00187]
2015-04-01	0.00042 [-0.00031, 0.00119]
2015-07-01	0.00084 [0.00017, 0.00165]
2015-10-01	0.00034 [-0.00044, 0.00124]
2016-01-01	-0.00021 [-0.00089, 0.00053]
2016-04-01	0.00017 [-0.00030, 0.00062]

A 30-day lead gap is a minimum separation from the information cutoff, not the age of the last available clinical assessment. Across the default forecasts, median time from the last CANS assessment to target start ranges from 84 to 88 days. The recency table shows how much earlier assessments actually occurred. The probability table also distinguishes ranking from calibration: good discrimination can coexist with predicted risks that exceed the frequency of recorded admissions.

Outcome coverage

Every admission in the supplied file has a discharge date. This does not establish whether the extraction query excluded ongoing stays, but it makes complete late-period ascertainment uncertain. The table shows both the decline in recorded starts and the distribution of recorded stay lengths by admission year. These are counts within a cohort selected through service involvement during 2014–2015. The files do not enumerate all youth served in later years or establish how long each remained under observation. Consequently, the counts do not estimate a system-wide admission rate.

Table 9: Recorded admission starts and episode durations by admission year in the full extract. Invalid negative durations are excluded only from duration summaries. The last year ends on November 2, 2016.

Year	Recorded starts	Valid durations	Median days	90th percentile days
2011	546	544	309	649
2012	911	910	301	606
2013	1,269	1,267	285	553
2014	1,329	1,328	246	506
2015	1,076	1,075	240	415
2016	353	352	84	221

The sampling descriptions consistently anchor selection to 2014–2015, but differ on whether qualifying placements began in those years or overlapped them. Such a cohort need not have stable service exposure in 2016: youth may leave services, age out, or have fewer subsequent admissions. A second possible mechanism is exclusion of unfinished stays. Under that mechanism, later starts would be observed preferentially when stays are short. Recorded stays beginning in 2016 are shorter than those beginning in 2014 or 2015, which is consistent with this mechanism but does not establish it. A real change in service use could also contribute. The available extraction documentation and records cannot identify how much of the decline comes from each source.

Table 10: Observed coverage of default target windows. The final column measures time from the exclusive target endpoint to the day after the last recorded discharge; it is not verified follow-up.

Target begins	Target ends (exclusive)	Recorded starts in full extract	Days remaining in discharge record
2015-01-01	2015-06-30	567	519
2015-04-01	2015-09-28	574	429
2015-07-01	2015-12-28	493	338
2015-10-01	2016-03-29	373	246
2016-01-01	2016-06-29	270	154
2016-04-01	2016-09-28	169	63

The first three default target windows end during 2015. They provide an earlier-period view of the results, although they are still drawn from the selected cohort. An earlier endpoint does not guarantee complete admission coverage.

Alternative histories, gaps and outcome windows

180-day history

Table 11: 180-day history; other timing choices retain their default values. Brackets give conditional 95% bootstrap intervals.

Target begins	Test youth	Events	ROC AUC	PR-AUC	Brier	Prevalence
2015-01-01	2,228	89	0.766 [0.719, 0.809]	0.121 [0.080, 0.175]	0.0373	4.0%
2015-04-01	2,223	79	0.801 [0.760, 0.840]	0.097 [0.072, 0.132]	0.0337	3.6%
2015-07-01	2,319	69	0.776 [0.715, 0.829]	0.109 [0.069, 0.171]	0.0277	3.0%
2015-10-01	2,415	54	0.743 [0.673, 0.809]	0.109 [0.046, 0.208]	0.0216	2.2%
2016-01-01	2,356	36	0.788 [0.701, 0.866]	0.053 [0.030, 0.093]	0.0156	1.5%
2016-04-01	2,171	25	0.836 [0.736, 0.916]	0.123 [0.036, 0.261]	0.0109	1.2%

No lead gap

Table 12: No lead gap; other timing choices retain their default values. Brackets give conditional 95% bootstrap intervals.

Target begins	Test youth	Events	ROC AUC	PR-AUC	Brier	Prevalence
2015-01-01	2,825	91	0.784 [0.741, 0.822]	0.089 [0.065, 0.123]	0.0313	3.2%
2015-04-01	2,870	83	0.812 [0.764, 0.852]	0.101 [0.068, 0.140]	0.0273	2.9%
2015-07-01	2,954	75	0.799 [0.751, 0.842]	0.102 [0.067, 0.162]	0.0238	2.5%
2015-10-01	2,976	63	0.757 [0.695, 0.815]	0.073 [0.042, 0.131]	0.0207	2.1%
2016-01-01	2,943	41	0.794 [0.717, 0.859]	0.063 [0.036, 0.113]	0.0140	1.4%
2016-04-01	2,795	27	0.819 [0.710, 0.907]	0.071 [0.033, 0.131]	0.0095	1.0%

90-day outcome

Table 13: 90-day outcome; other timing choices retain their default values. Brackets give conditional 95% bootstrap intervals.

Target begins	Test youth	Events	ROC AUC	PR-AUC	Brier	Prevalence
2015-01-01	2,792	46	0.785 [0.736, 0.833]	0.043 [0.029, 0.066]	0.0166	1.6%
2015-04-01	2,857	47	0.750 [0.683, 0.814]	0.045 [0.028, 0.072]	0.0160	1.6%
2015-07-01	2,953	43	0.813 [0.754, 0.871]	0.061 [0.035, 0.125]	0.0141	1.5%
2015-10-01	2,969	38	0.773 [0.708, 0.829]	0.039 [0.021, 0.101]	0.0127	1.3%
2016-01-01	2,952	23	0.772 [0.659, 0.865]	0.029 [0.013, 0.066]	0.0079	0.8%
2016-04-01	2,870	15	0.839 [0.738, 0.922]	0.101 [0.011, 0.267]	0.0051	0.5%

365-day outcome

Table 14: 365-day outcome; other timing choices retain their default values. Brackets give conditional 95% bootstrap intervals.

Target begins	Test youth	Events	ROC AUC	PR-AUC	Brier	Prevalence
2015-04-01	2,857	125	0.793 [0.760, 0.825]	0.126 [0.095, 0.166]	0.0416	4.4%
2015-07-01	2,953	101	0.795 [0.753, 0.832]	0.117 [0.083, 0.168]	0.0334	3.4%
2015-10-01	2,969	80	0.774 [0.723, 0.822]	0.111 [0.065, 0.177]	0.0269	2.7%

Interpretation

These comparisons examine temporal transport within the supplied cohort. They do not establish prospective performance in the service population. The cohort was selected through service involvement during 2014–2015, and the extract may omit ongoing placements. The decline in recorded admissions during 2016 can therefore reflect how the files were assembled as well as changes in utilization. Moving the windows does not repair those limitations.

The smallest test event count across specifications is 15; short-window estimates based on few admissions should be read with their uncertainty and denominators.

The same test youth can contribute at several forecast dates, and outcome windows overlap. Results are reported separately by date; they are not pooled as independent observations. Intervals

condition on each fitted model and split and omit uncertainty from training, record availability, cohort selection and family or provider dependence. Differences across specifications can also reflect changing eligible youth. All five specifications share the April, July and October 2015 target dates; the 365-day outcome has no results for the other dates. Comparisons across horizons must retain this calendar distinction and the different outcome definitions.

References

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Poweleit, Ethan A., Jeffrey R. Strawn, Rhonda D. Szczesniak, et al. 2026. "Predicting Psychiatric Readmission of Children and Adolescents Using Machine Learning." *Journal of Child and Adolescent Psychopharmacology* 36 (8): 483–92. https://doi.org/10.1177/10445463261466534.

Taquet, Maxime, Seena Fazel, and A John Rush. 2025. "Transdiagnostic Early Warning Score for Psychiatric Hospitalisation: Development and Evaluation of a Prediction Model." *BMJ Mental Health* 28 (1): e301622. https://doi.org/10.1136/bmjment-2025-301622.