

RISK STRATIFICATION, SEGMENTATION, AND TIERING TRANSPARENCY DOCUMENT

JULY 2025

TABLE OF CONTENTS

Risk Stratification, Segmentation, and Tiering Transparency Document.....	1
Table of Contents.....	2
I. Executive Summary.....	4
II. Introduction	9
III. Design Principles.....	11
IV. Leadership and Governance.....	13
A. Academic Work Group.....	13
B. Scientific Advisory Council.....	14
C. Stakeholder Engagement.....	14
V. Model Design.....	17
A. Population Inclusion/Exclusion Criteria.....	17
B. Data Sources	19
C. Outcomes.....	19
D. Predictors.....	20
E. Model Training Methods.....	21
F. Bias and Sensitivity Testing Methods.....	24
G. Tiering Analysis Methods.....	25
VI. Model Results.....	31
A. Outcomes Detail	31
B. Predictors Detail.....	38
C. Technical Environments Setup.....	44
D. Model Tuning Results.....	45
E. Model Performance.....	47
F. Tiering Analysis Results.....	56
G. Validation Results.....	65
VII. Maintenance and Monitoring.....	70
A. Monthly RSST Tier File Generation.....	70
B. Dashboard-Based Monitoring (DHCS Access Only).....	71

C. Ongoing Maintenance of the RSST Algorithm.....	71
VIII. Assumptions and Limitations.....	72
IX. Appendix.....	74
A. Glossary of Key Policy Terminology Used	74
B. Glossary of Technical Terms Used.....	77
C. Additional Model Detail.....	79
D. References.....	90

I. EXECUTIVE SUMMARY

1) Overview and Purpose

The Risk Stratification, Segmentation, and Tiering (RSST) Algorithm is a predictive analytics system developed by the California Department of Health Care Services (DHCS) as part of its Population Health Management (PHM) strategy. RSST provides a monthly, data-driven method for identifying Medi-Cal members who may be at increased risk of poor outcomes or underutilization of essential services, and who may benefit from additional outreach or care coordination. Each month, the system assigns each eligible member to a risk tier—low, medium-rising, or high—based on predicted likelihood of future outcomes, allowing for more proactive and equitable deployment of resources across the state.

RSST was designed to create a standardized and transparent method of risk tiering that can be applied uniformly across MCPs statewide. While the algorithm has been developed and validated for implementation, decisions about its formal policy adoption and integration into MCP operations are still forthcoming. The creation of this tool supports DHCS's broader objective to improve consistency, transparency, and equity in how risk is assessed across the Medi-Cal population, and it lays the groundwork for future phases of policy and program rollout.

One innovation of the RSST approach is the inclusion of a domain for underutilization of services. This domain focuses on identifying members who are predicted to not engage with recommended care despite known clinical or behavioral health risks. Unlike many risk models that target only high-utilizing or high-cost members, RSST explicitly seeks to surface members with unmet needs who may be overlooked in traditional frameworks because their lack of utilization makes them appear to be "low risk." In Version 1, underutilization emerged as one of the most significant drivers of high-risk designation—particularly in the adult population.

Importantly, RSST is an evolving algorithm that will be continuously updated and retrained on additional data to improve its accuracy, equity, and usefulness over time. While this initial launch represents a critical step forward, DHCS recognizes that it is an early version in an ongoing journey of refinement and learning.

2) Design Principles, Governance, and Stakeholder Engagement

The development of RSST was anchored in several core design principles: standardization, transparency, equity, and flexibility. The algorithm applies a statewide method to ensure that high-risk members are identified in a standardized method

across all counties and health plans. Transparency was prioritized not only in the algorithm logic and model architecture, but also in the supporting documentation, thresholding approach, and validation methods. Equity was not treated as a one-time metric but as a continuous design consideration, influencing every stage of development—from the fine-tuning of outcomes and predictor variables to the way subgroup performance was evaluated, to the final tiering decisions. Flexibility was embedded by structuring the algorithm around multiple domains and subdomains of risk—physical, behavioral, social, and underutilization—so that future policies could adjust their emphasis across these areas without altering the core technical framework.

The RSST Work Group, led by DHCS, provided strategic direction and policy oversight throughout development. An Academic Work Group (AWG) composed of national experts in health equity, health outcomes, Medicaid analytics, and machine learning provided guidance on technical decisions and subgroup evaluation. A Scientific Advisory Council representing clinical and operational leaders from across the Medi-Cal delivery system helped ensure real-world applicability of the algorithm and supported stakeholder alignment. This multi-level governance approach created space for rigorous analysis, transparent debate, and deliberate policy design.

3) Model Structure and Subdomain Organization

The RSST Algorithm is composed of ten predictive models, five each for adult and pediatric members. Each model focuses on a specific subdomain of risk: adverse physical health events, adverse behavioral health events, underutilization of physical health services, underutilization of behavioral health services, and social risk. Each month, these models generate individual-level probability scores that are then translated into categorical risk tiers.

Subdomain tiers are rolled up into domain-level tiers and a single overall tier using a max-tier logic, ensuring that the highest relevant risk signal is preserved. This structure enables DHCS and MCPs to see not just who is at risk, but what type of risk is being predicted, allowing for differentiated program responses. By modeling each subdomain independently, the algorithm provides transparency in its predictions and enables future policy decisions to selectively emphasize or de-emphasize domains without disrupting the broader infrastructure.

4) Machine Learning Methods

The RSST models were developed using established machine learning methods widely adopted in health services research and public health contexts. These include gradient-boosted decision trees (e.g., LightGBM) and regularized logistic regression, both of

which are known for their performance and interpretability in large, structured datasets. These approaches are supported by academic literature and have been applied in similar Medicaid-focused predictive modeling efforts.

Models were trained and retrospectively evaluated using eight years of Medi-Cal administrative claims and eligibility data from January 2016 through December 2023, covering more than 15 million individuals. Training was performed at the person-month level, with predictor variables constructed using “what was known when” logic. This ensured that when models were evaluated by retrospectively comparing predictions to known outcomes, the predictions from each model relied only on predictors that would have been available at the time of prediction. The evaluation therefore accounted for real claims processing delays and mimicked real-world deployment conditions to predicted risk of a poor outcome in the coming year.

Model performance was optimized using held-out validation data (e.g., to select hyperparameters). Independent test data (not used in model training) were used to evaluate performance and assess generalizability. In order to mitigate impacts of COVID-19, these test data, and all model performance metrics and tiering decisions, were restricted to index dates beginning January 2022 (after which COVID-19-associated disruptions had diminished).

5) Technical Performance

Area under the curve (AUC) scores ranged from 0.75 to 0.94 across the ten models, which compared favorably to relatively similar models, with particularly strong performance in behavioral health and social risk domains. Additional performance evaluation took place during tiering (below).

Performance was also assessed using recall, and number-needed-to-treat (NNT) at the model, domain, and overall tier levels. For example, under the selected tiering configuration, the overall high-risk tier for adults achieved a recall of 0.23 (meaning that, of members who experienced a relevant outcome in the next 12 months, 1 in 4 were identified as high risk), with an NNT of 1.75 (meaning that for every two members flagged as high-risk, approximately one experienced a relevant outcome in the next 12 months). Similar results were observed for pediatrics, with a recall of 0.24 and NNT of 1.97. These metrics showed that the models could reliably support identification of members who went on to experience a relevant outcome in the subsequent year.

6) Tiering Strategy and Scenario Evaluation

After model training and initial performance evaluation were completed, the RSST Work Group turned to the policy challenge of converting subdomain-level risk scores into tier assignments—particularly the high-risk tier, which may drive outreach or services in future phases of implementation. DHCS set a statewide target that no more than approximately 10 percent of the Medi-Cal population should be classified as high-risk, in order to align with existing MCP capacity and ensure policy feasibility. To implement this constraint, the team evaluated three approaches to setting model-specific thresholds.

Option 1 evenly distributed high-risk flags across all five models, regardless of model performance or outcome prevalence. Option 2 applied thresholds to equalize sensitivity (recall) across models, prioritizing balance in outcome capture. Option 3 optimized thresholds to maximize overall recall, favoring models with stronger technical performance. After evaluating the tradeoffs, the Work Group selected Option 2 for both adults and pediatrics. It offered the best alignment with policy goals by maintaining proportional contributions from each domain, supporting transparency in how risk was defined, and achieving strong subgroup performance.

Equity was central to this decision. Throughout the tiering analysis, the Work Group assessed model performance across subgroups defined by race, ethnicity, language, age, sex, and new enrollees. Option 2 was the only option that met or exceeded DHCS's equity performance threshold in all but one subgroup for each population. These deviations were closely examined and attributed to differences in outcome definitions—such as the use of female-only preventive care measures in pediatric underutilization models, which resulted in slightly worse performance for males—rather than model bias. In both adult and pediatric populations, Option 2 also produced a meaningful number of net-new members who would not have otherwise been flagged under existing risk tools, many of whom went on to experience an RSST outcome. Technical performance was assessed using recall, NNT, and AUC.

These results confirmed that Option 2 offered the best combination of predictive value, equity performance, and explainability, and supported DHCS's broader objective to ensure that risk tiering is consistent, transparent, and actionable for this initial release. The overarching tiering methodology is explicitly designed to incorporate flexibility in tiering decisions, including refinement of the options above, as new data become available and the models are updated.

7) Maintenance, Monitoring, and Future Work

RSST risk tiers are refreshed monthly using the most recent data available and are delivered to MCPs via secure file transfer and API, then integrated into the longitudinal member record in the PHM Portal. A separate dashboard is available to DHCS to support internal monitoring and policy evaluation. This dashboard enables detailed analyses of risk distribution over time, including subgroup comparisons aligned with the equity considerations emphasized throughout RSST development.

Validation tests confirmed that the RSST models and infrastructure performed as expected during productionization, with only minor, explainable differences across environments. Underutilization risk prevalence increased modestly when applied to 2024 data, but model logic and performance remained consistent with the original design.

Future updates to RSST may include model retraining on newer data, incorporation of additional predictors, refinement of tiering thresholds, and eventual release of birthing population models. DHCS is developing a formal process for reviewing and implementing such changes, ensuring that future updates preserve the technical integrity and policy alignment of the algorithm.

8) Assumptions and Limitations

RSST Version 1 includes several important limitations. The models rely entirely on Medicaid administrative claims and eligibility data. Clinical EHR data and broader social datasets are not currently integrated. Members enrolled in limited-benefit programs, such as GHPP, CCS, or Family PACT, were excluded due to incomplete data. Dual Medicaid–Medicare eligible members were included in all models except for underutilization, where lack of Medicare claims particularly limited the predictive power of the model. Note, Medicare claims data was not used to develop any of the models, so the reliability of the models for dual members is limited to their Medicaid benefits. Tier thresholds were set using predicted scores from January 2023 and remain fixed for Version 1, meaning the proportion of members flagged as high-risk may vary slightly over time. Internal data transfer lag within the PHM system was not explicitly modeled but is expected to be better understood and addressed in future releases.

9) Conclusion

The RSST Algorithm represents DHCS’s commitment to building a modern, transparent, and policy-aligned approach to risk stratification for the Medi-Cal population. It introduces a replicable framework that can be applied statewide, supports more equitable targeting of care interventions, and lays the foundation for future

enhancements. Version 1 of RSST reflects a careful balance of technical rigor, operational feasibility, and policy vision. This document provides a comprehensive explanation of the algorithm’s design, logic, and intended use, and is intended to serve as a transparent reference point for all stakeholders in the Medi-Cal system.

II. INTRODUCTION

Managed Care Plans (MCPs) in California are currently required to implement their own risk stratification methods as part of their Population Health Management programs. However, most MCPs have historically used disparate tools—many proprietary or commercially developed—resulting in wide variability in how members are identified for assessment or care coordination. To establish a statewide framework for risk tiering and address care gaps within the Medi-Cal program, DHCS developed the Risk Stratification and Segmentation Tiering (RSST) algorithm.

The RSST Algorithm was developed to address these gaps and establish a unified, statewide framework for risk tiering within the Medi-Cal program. This approach is distinct from many existing tools in both its technical design and its policy orientation. The RSST Algorithm is:

- » Standardized across all MCPs and regions
- » Transparent and interpretable
- » Specifically tailored to the Medi-Cal population and benefit structure
- » Built using supervised machine learning models that are maintainable and tunable over time
- » Designed with equity in mind—including bias mitigation in model development and subgroup monitoring post-deployment

Unlike most traditional risk models, which often focus on prior cost and utilization, the RSST Algorithm is forward-looking. It estimates the risk of future adverse outcomes, underutilization, and social vulnerability. It also emphasizes transparency—both in how members are flagged and in how results are monitored at the population level by DHCS.

The following table summarizes the core challenges this approach aims to address, and the corresponding solutions implemented through RSST:

Table 1. Challenges of Existing Risk Identification Approaches and RSST Solutions

Challenge of Existing Risk Identification Approaches	RSST Response
Inconsistent stratification across plans	Statewide, standardized method applied to all members
Fragmented or incomplete data	Centralized data from the full Medi-Cal system, including sources not easily available to individual MCPs
Cost/utilization-focused models	Models include adverse events, underutilization, and social risk
Proprietary, opaque algorithms	Transparent and interpretable logic with monthly oversight
Limited attention to equity	Equity-informed model design and ongoing subgroup analysis
Retrospective risk score defined by past events or diagnoses	Forward looking Machine Learning models predict likelihood of future negative outcomes that have not yet occurred
Uniform treatment of risk across patient groups	Risk scores and tiers are differentiated across multiple domains
Static parametric models updated annually or less often	Built using supervised ML for ongoing refinement and updates

These principles also inform the operational model: RSST tier outputs are produced monthly using updated data and delivered for use in the PHM Service. Members in the highest risk tiers (Tier 3) are prioritized for outreach and assessment. The monitoring dashboard—available to DHCS—provides detailed visibility into tier distribution and subgroup patterns, supporting accountability and ongoing refinement.

Taken together, the RSST Algorithm provides a scalable, policy-aligned, and equitable foundation for identifying members most likely to benefit from proactive services and care coordination across the Medi-Cal program.

III. DESIGN PRINCIPLES

The RSST Algorithm was developed to support more equitable, consistent, and actionable risk stratification across the Medi-Cal program. Its design reflects both technical modeling objectives and broader policy values, including transparency, equity, and statewide standardization. This section outlines the key principles that guided the development of the algorithm.

Statewide Standardization with Population-Specific Design

A core goal of RSST is to establish a consistent, statewide approach to identifying Medi-Cal members at elevated risk—ensuring that risk stratification is applied equitably across Managed Care Plans (MCPs), regardless of regional or population differences.

Standardization does not imply a one-size-fits-all model, but rather a shared framework that accommodates population-specific nuances while delivering consistent outputs for programmatic use.

Different populations experience risk differently—and face different consequences from being overlooked. RSST was designed from the outset to recognize the distinct needs of adults, children, and birthing people. Separate models were developed for each stratum to ensure that outcomes and predictors reflect the clinical context, service patterns, and risk factors relevant to each group. For example, children may be flagged for missed well-childcare, while adults may be flagged for gaps in chronic disease or behavioral health care. Separate modeling enables the algorithm to better align with the populations DHCS serves while still promoting a standardized framework that can be applied consistently across all health plans.

Traditional risk models often identify only those who are already high-cost or high-utilizing. RSST was designed to also flag individuals who have clinical indicators suggesting unmet need. These may include, for example, a lack of primary care follow-up after an emergency department visit or a behavioral health diagnosis without a recent prescription refill. This “Underutilization” Domain helps highlight members facing systemic barriers to care who may otherwise go unflagged.

Equity in Model Development

The ideal goal of equity in this context is to ensure that all Medi-Cal members—regardless of race, ethnicity, language, or gender—have a fair opportunity to be accurately identified for services and support. This approach aligns with DHCS’s commitment to fostering an equitable, dignified Medi-Cal system that works for everyone. Equity was a guiding consideration throughout the RSST development

process. This includes selecting outcomes and predictors with input on how risk manifests across populations and analyzing tier outputs by subgroup to detect any potential disparities. DHCS is not assuming this eliminates disparities but is committed to using equity-focused evaluation as a basis for future model refinement. Ongoing subgroup monitoring will continue as part of routine model oversight. Equity was considered at each stage of the model lifecycle—from defining outcomes and predictors, to model training and aggregation logic, to evaluation and ongoing monitoring. This framing helped ensure equity considerations were not siloed but embedded throughout the development and deployment process.

Transparency in Design and Structure

Transparency was a core reason for developing a custom RSST Algorithm. Many commercial risk tools function as black boxes—offering little visibility into how risk scores are calculated or how they relate to member needs. RSST was designed to be fully interpretable, with its logic, structure, and assumptions clearly documented. This document is part of that commitment—offering a public explanation of how the algorithm was designed, what it is intended to do, and how it is being monitored over time.

The RSST structure separates different types of risk into distinct subdomain models and aggregated domains (adverse events, underutilization, and social risk). This design was intentional: it allows for transparency in how each domain contributes to overall risk and creates the flexibility for policymakers to decide how to weight or prioritize each domain. As described in the Tiering Analysis section, this structure supported a policy decision-making process that considered how different types of risk should be elevated or balanced.

Interpretability and Flexibility

The RSST Algorithm uses supervised machine learning techniques that are designed to be flexible (provide as much information about risk as possible from available data), transparent (with all model inputs, outputs, and training procedures clearly defined), reproducible, and have interpretable outputs. This allows the algorithm to be updated over time as new data become available, additional models are introduced, or policy needs evolve. The algorithm's structure supports both transparency and ongoing refinement, ensuring it remains relevant and aligned with Medi-Cal program goals.

IV. LEADERSHIP AND GOVERNANCE

To guide the development of the RSST Algorithm, DHCS established a governance structure designed to ensure evidence-based methods, stakeholder input, and alignment with the needs of Medi-Cal beneficiaries. A dedicated leadership team was formed, with internal leads and sponsors overseeing strategy and execution. The RSST Work Group, under supervision of the Quality and Population Health Management (QPHM) division, led model design. A Scientific Advisory Council (SciAC) provided expert review on major decisions, including model principles, outcomes, predictors, and equity considerations. The PHM Advisory Group—composed of public stakeholders—helped ensure transparency and relevance throughout development. DHCS also received strategic support from external advisors to inform planning and execution.

A. Academic Work Group

The work group is charged with developing the RSST contextual design and approach of the algorithm. The work group (which was brought in 2023 to advise) consists of nationally recognized experts in health equity, health outcomes, Medicaid analytics, and machine learning.

Table 2. Academic Work Group Members

Name	Title	Organization
Maya Petersen – Lead	Professor, Biostatistics and Epidemiology	UC Berkeley
Jonathan Kokstad	Assoc. Professor, Economic Analysis and Policy	UC Berkeley
Michael Barnett	Professor, Health Services Policy and Practice	Brown University
Alejandro Schuler	Asst. Professor, Biostatistics	UC Berkeley
Jacob Wallace	Assoc. Professor, Public Health (Health Policy)	Yale
Anna Zink	Principal Researcher	Chicago Booth (Center for Applied AI)

B. Scientific Advisory Council

Established by DHCS in 2023 to support the RSST Work Group, the Scientific Advisory Council (SciAC) includes representatives from Medi-Cal Managed Care Plans, healthcare delivery systems, care management leaders, and medical academic institutions.

Throughout the design and development of the RSST algorithm the SciAC served as key advisors providing input and guidance.

[Introducing the PHM Initiative RSST Work Group Members and Scientific Advisory Council Members](#)

C. Stakeholder Engagement

The RSST WG relied on the expertise of additional stakeholders through the development of the algorithm (e.g., design, structure, outcomes). This included extensive internal and external stakeholder engagement, consultation and consensus building.

1. Internal Stakeholders

Internal DHCS stakeholders were engaged and consulted through the design of the algorithm, particularly regarding the design and selection of outcomes. This included engagement with the following DHCS divisions and external groups:

Table 3. Internal Stakeholders

DHCS Divisions	External Groups
Director's Office	Medicare Team
Behavioral Health	Birthing Care Pathway
Health Care Benefits and Eligibility	Children and Youth Advisory Council
Enterprise Data and Information Management	
Office of Medicare Innovation and Integration	Social Risk experts
Quality and Population Health Management	
Health Care Financing	Further, MCPs were consulted on the RSST rationale and contextual algorithm design (i.e., framework, predictors, outcomes).
Health Care Delivery Systems	

2. External Stakeholders

External stakeholders were actively engaged throughout the design of the algorithm, particularly through the Medi-Cal Connect Advisor process. These Advisors were MCPs that volunteered time and resources to review and validate Medi-Cal Connect on behalf of the broader Medi-Cal Plan community. The following MCPs participated in these activities:

- » Health Net Community Solutions
- » Inland Empire Health Plan
- » Kaiser Permanente
- » L.A. Care Health Plan

For RSST, Advisors compared 100 high-risk and 100 low-risk members from January 2023 against their internal risk assessments and care management activities conducted during the same year. They reported on concordance and discordance in tiering and shared insights from member assessments completed during that period. This comparison enabled DHCS to validate RSST tiers before launch and informed key policy decisions. Additionally, DHCS gathered feedback on RSST-related policy considerations, including reporting requirements, implementation timelines, and potential barriers from the Advisors as well as from the Medi-Cal Plan community. As the RSST algorithm is prepared for deployment and released, additional engagement will be required. This may include consultations regarding risk tiers, algorithm performance results, equity analyses, and how to improve the performance of the algorithm for marginalized, underrepresented and equity-deserving groups. The DHCS RSST and User and Stakeholder Engagement (USE) teams will work in partnership to engage MCPs post-release. The USE team will lead the engagement coordination, including outreach to additional MCPs beyond the original group of Advisors. The RSST team and project leads will contribute subject matter expertise to inform these efforts.

Table 4. RSST/USE Stakeholder Engagement

Engagement Type	Purpose	MCPs	MCP Rationale
May 2024 - USE Discovery	Engage a small subset of MCPs to validate, understand, and refine initial assumptions on the five identified value propositions	<ul style="list-style-type: none">» Health Net» IEHP» LA Care	Selected for engagement from early adopter list and per the recommendation

Engagement Type	Purpose	MCPs	MCP Rationale
	and core elements of Medi-Cal Connect.		of GWTs strategic advisor.
Sept 2023 - MCP Engagement	Engage a small subset of MCPs to collect input on how RSST may be used at the MCP for purposes of developing the algorithm.	<ul style="list-style-type: none"> » Health Net » HPSM » Kaiser » LA Care » Partnership 	Selected a small group with diverse representation of technology/analytic abilities.
RSST Lead Outreach Sarah Lopez	Introduce Sarah to MCPs and get a sense of the "state of the art" and advanced Plan approaches to understand how the RSST effort's work compares.	<ul style="list-style-type: none"> » Central California Alliance for Health » HPSM » IEHP » Partnership 	Selected for engagement due to maturity of population health and tiering capabilities as well as access to advanced analytics.
Scientific Advisory Council	The Council serves as an advisory group to DHCS and the RSST WG with a goal of guiding the development of the PHM Service RSST algorithm.	<ul style="list-style-type: none"> » Contra Costa Health Services » LA Care » Partnership » IEHP » Health Net » Kaiser 	The SciAC is comprised of individuals representing MCPs, health care delivery systems, care management leaders, and academic medical systems.
PHM Advisory Group	The Advisory Group was comprised of cross-sector stakeholders that provided feedback and made recommendations on the PHM Program and Service, including RSST.	<ul style="list-style-type: none"> » Health Net » HPSM » IEHP » LA Care » Kaiser 	Members include a diverse set of leaders with representation from health plans, providers, counties, state departments,

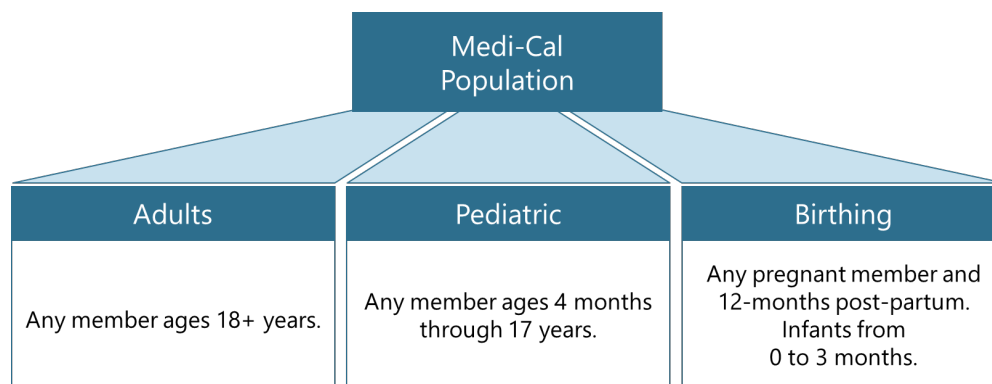
Engagement Type	Purpose	MCPs	MCP Rationale
		» Kern Health Systems » Partnership	consumer organizations, and other groups.

V. MODEL DESIGN

A. Population Inclusion/Exclusion Criteria

The RSST algorithms were originally designed to produce monthly risk tiers on three subpopulations within California’s Medi-Cal population – one risk tier for each subpopulation. These subpopulations included:

- » **Adult population:** Any eligible Medi-Cal member, aged 18+
- » **Pediatric population:** Any eligible Medi-Cal member, aged 4 months through 17 years of age
- » **Birthing population:** Any eligible Medi-Cal member, who is pregnant or up to 12 months post-partum. Any infant, Medi-Cal member aged 0-3 months.



Eligible members were restricted to Medi-Cal members, not those who solely enrolled via FPACT, CSS, or GHPP. Members with Medicare dual eligibility were included, but their sub-domains of risk modeling were restricted to Adverse Events and Social Risk. Ages were calculated at the time of the most recent monthly data processing.

For the RSST version 1.0 launch, modeling was restricted to the Adult and Pediatric populations only, with the intention of adding the birthing population to the next version release.

Table 5: Characteristics and demographics in Jan 2023

	Adult	Pediatric
Sex – no. (%)		
Male	3,945,095 (44.5%)	2,329,503 (51.2%)
Female	4,930,234 (55.5%)	2,219,539 (48.8%)
Age		
Mean (sd)	43 (18.40)	9 (5)
Race/Ethnicity – no. (%)		
Am. Indian or Alaska Native	35,939 (0.4%)	13,830 (0.3%)
Asian	912,370 (10.3%)	252,153 (5.5%)
Black or African American	651,099 (7.3%)	285,213 (6.3%)
Native Hawaiian or Pacific Islander	107,863 (1.2%)	39,400 (0.9%)
Other	3,783,527 (42.6%)	2,508,718 (55.1%)
Two or More Races	515,482 (5.8%)	299,658 (6.6%)
Race Unknown	1,057,495 (11.9%)	543,973 (12.0%)
White	1,811,554 (20.4%)	606,097 (13.3%)
Language – no. (%)		
Primary Language English	6,085,288 (68.6%)	3,073,996 (67.6%)
Primary Language Spanish	2,028,567 (22.9%)	1,323,350 (29.1%)
Primary Language Other	761,474 (8.6%)	151,696 (3.3%)
Enrollee Status – no. (%)		
Dual Eligible	1,524,937 (17.2%)	104 (0.0%)
Not Dual Eligible	7,350,392 (82.8%)	4,548,938 (100.0%)

B. Data Sources

Restricted to data available at the time of model training, the RSST models (v1.0) used data from two core sources: Medicaid Administrative Claims Data and Eligibility Data. Both data sources were formatted by DHCS to conform to the All-Payer Claims Data Common Data Layout (APCD-CDL).

Historic data files were delivered in bulk for training, and monthly incremental files are to be delivered on a monthly cadence post launch:

- » An extract of all Medi-Cal claims, including professional, facility, dental, and pharmacy claims
- » An extract of Medi-Cal eligibility/enrollment data, including demographic and product enrollment information
- » A supplemental member eligibility (SUME) extract, which was used to extract death date (used for mortality prediction) and create a dual-enrolled status indicator (Note: future model iterations will draw more information from SUME)

Representing eight years of data from 2016-2023, these data files were passed through an MPI engine that grouped multiple CIN values into a single PersonId value, allowing the data to be modeled at the person-level. We partitioned data into non overlapping time-based windows with index dates spanning January 2019 to January 2023. See 'Data Splitting' section for details on specific splits across datasets.

C. Outcomes

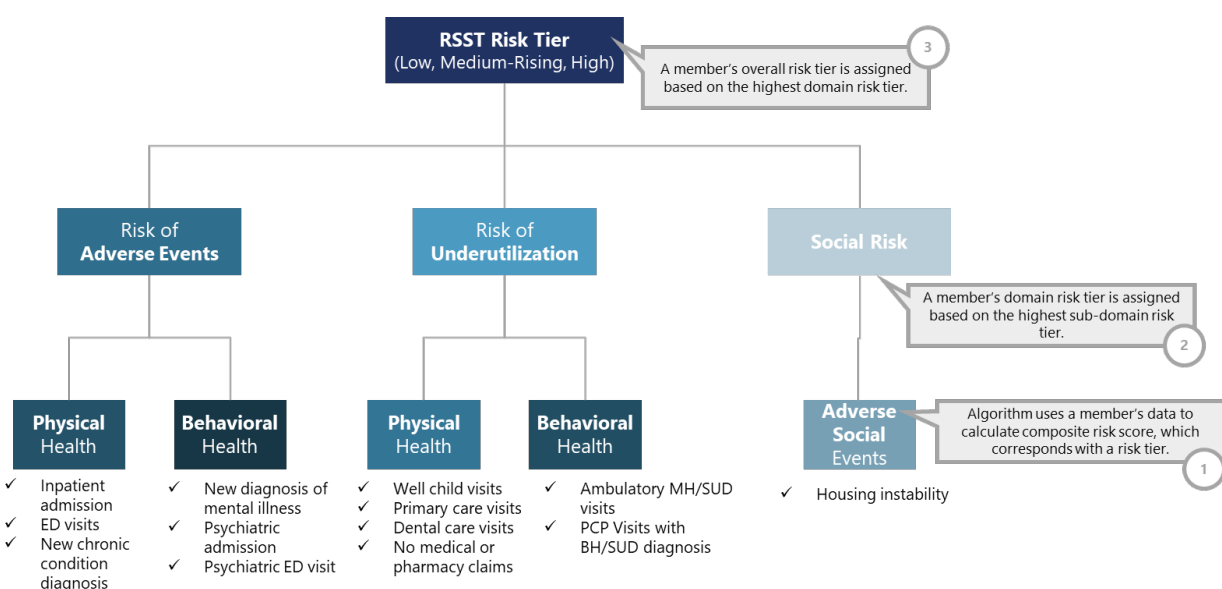
For RSST V1, binary outcomes were developed for the adult and pediatric Medi-Cal populations. Each subdomain-level model predicts whether a member will experience one or more of these outcomes in the 12 months following a defined index date. If a member experienced at least one outcome, they were assigned a composite outcome value of 1; otherwise, they were assigned a value of 0.

Outcomes were organized within three overarching domains: Risk of Adverse Events, Risk of Underutilization, and Social Risk. Within each domain, outcomes were grouped into clinically and programmatically relevant subdomains, including physical health, behavioral health, and adverse social events such as housing instability.

The RSST Work Group led the design of these outcomes, supported by analytic staff and with consultation from subject matter experts across DHCS, including the Enterprise Data and Information Management (EDIM) Division and the Data Services Branch (DSB). The Work Group sought to align outcome definitions with existing DHCS business logic

and ensure outcomes could be implemented consistently across the Medi-Cal population. Wherever possible, definitions were based on data elements and rules already in use at the state level.

Each proposed outcome was tested in historical Medi-Cal data to confirm feasibility, appropriate prevalence, and clarity of construction. Prevalence rates were reviewed by DHCS and academic SMEs to confirm that results were reasonable and interpretable, and subgroup breakdowns (e.g., by race, ethnicity, sex, language, enrollment type) were used to identify potential data quality issues. These equity-focused reviews helped flag patterns to be aware of but did not lead to major structural changes at the outcome level.



Note: A complete listing of RSST V1 outcomes with definitions can be found below in the [Outcomes Detail](#) section.

D. Predictors

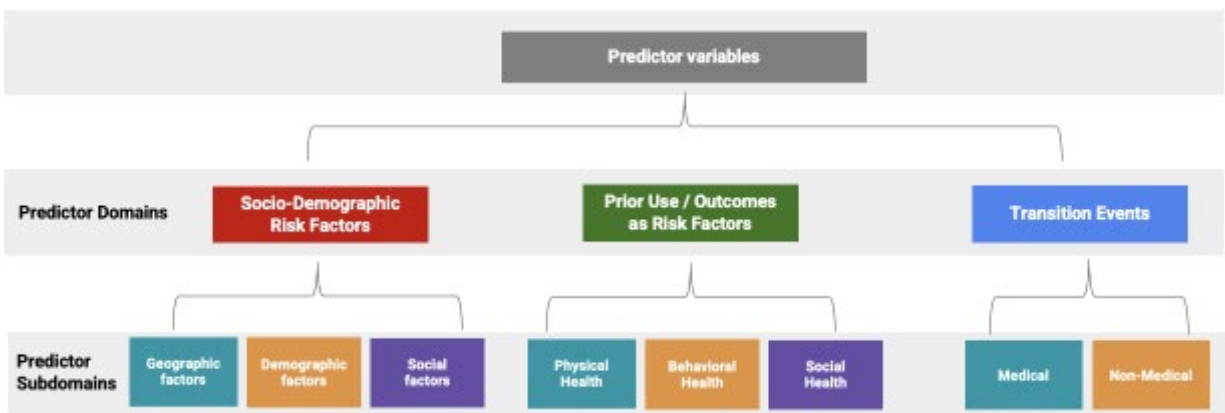
Predictors were developed using lookback windows prior to the defined index date. Depending on the predictor, the lookback window was defined in the past 3 months, 12 months, or "ever", where "ever" incorporates the full available history, which extends back no earlier than January 2016. Lookback window information for each predictor can be found in the "Predictor Details" section. Predictors were grouped into one of four domains:

- » Socio-Demographic Risk Factors
- » Prior Use

- » Transition Events
- » Outcomes as Risk Factors

Socio-Demographic Risk Factors were grouped in geographic, demographic, and social subdomains and included information such as age, race, gender, employment status, etc. **Prior Use** risk factors were grouped into physical, behavioral, and social health subdomains and included indicators and counts on attributes such as chronic condition diagnoses and healthcare service utilization. **Transition Events** were grouped into medical and non-medical subdomains and included indicators such as switching Medicaid plans, moving counties, aging into the adult population, etc.

Outcomes as Risk Factors is a special case where the outcomes we are predicting were adapted for use as prior period predictors. In many cases this involved transformation from a binary outcome (e.g. Indicator of any ED visits in the next 12 months) to a continuous predictor (e.g. Count of ED visits in the past 3mo, 12mo, or “ever” – full available history).



Note: A complete list of the RSST V1 Predictors can be found below in the [Predictors Detail](#) section.

E. Model Training Methods

Machine learning models were used to provide flexibility in learning complex patterns from the data. We evaluated three supervised learning algorithms, selected for their balance between performance, scalability (up to 100s of millions of rows), and interpretability.

1. Models

a. Regularized Logistic Regression (LR)

Logistic regression with elastic net penalty was implemented with stochastic gradient descent optimization, allowing tuning over a wide range of regularization strengths (C) and mixing parameters (l1_ratio).

b. XGBoost

A high-performance implementation of gradient boosting, optimized for parallel training. Hyperparameters included number of estimators, learning rate, and tree depth.

c. LightGBM

A leaf-wise histogram-based gradient boosting algorithm, selected for its speed and memory efficiency on large datasets. Similar hyperparameters were explored as in XGBoost for comparison.

All models were trained to generate probabilistic predictions for binary outcomes in a multi-label setting.

2. Data Splitting

The data were partitioned into training, validation, and test sets, with a buffer window to ensure that no individuals appeared with an outcome label that overlapped with the test set. The data were partitioned into the following distinct time-based windows:

- » **Tuning and Hyperparameter Search Set** (Jan 2019 – Jan 2020): Used to explore hyperparameter configurations across all models and train initial models for validation and model selection.
- » **Validation Set** (Feb 2020 – Jan 2021): Used to perform model selection between models trained on the Search Set and for sensitivity analyses to investigate model stability & generalizability.
- » **Buffer Window** (Feb 2021 – Dec 2021): Excluded from all training and evaluation to prevent label leakage. This window ensures that no individual appears with an outcome label that overlaps into the test set.
- » **Test Set** (Jan 2022 – Jan 2023): Held out for final evaluation only, simulating real-world performance in a temporally shifted future.

These dates reflect the index dates; historical predictor data before (3 months, 12 months, and "ever") and outcome data after (12 months) the index dates were included in the datasets. This structure was designed to ensure that all model comparisons and selections were based on realistic future-facing performance, and to avoid

contamination of future labels during training. In order to mitigate impact of the COVID-19 pandemic on the generalizability of results to current settings, the independent Test Set that was used for all tiering decisions and final performance evaluations used data from a period when COVID-related disruptions to the health care system had substantially decreased. In additional sensitivity analyses, sub-domain-specific model AUCs were stable when evaluated on the Test Set compared to the Validation Set (which included the period of maximal COVID-19 impact), supporting stability of model performance.

Logistic Regression was implemented with elastic net penalty and stochastic gradient descent optimization, and was tuned over a wide range of regularization and mixing parameters. Models were trained using a GPU-accelerated implementation of logistic regression from the cuML library, with the 'qn' solver (quasi-Newton optimization) and a maximum of 100 iterations per trial. Optimization was performed using Optuna, with single-threaded tuning and early stopping and pruning conditions, consistent with the tree-based models.

The XGBoost model used gradient boosting, optimized for parallel training, while the LightGBM model used a leaf-wise gradient boosting algorithm. The number of estimators, learning rate, and tree depth were tuned as hyperparameters. Training and evaluation for the XGBoost Model were performed using DaskQuantileDMatrix to support GPU-native batching. The gradient based sampling method was enabled to stabilize learning under class imbalance. For the LightGBM model, hyperparameter trials were parallelized using Dask, and tuning was conducted using Optuna with the same early stopping and pruning strategies.

The models were evaluated for overall model performance using AUC, and prespecified groups were evaluated for algorithmic bias, including race/ethnicity and sex, among others. Calibration curves were developed to compare predicted probabilities with observed outcome rates within each subgroup. NNT, Precision, and Recall were calculated across the range of possible tiering thresholds to evaluate model performance, and SHAP values were generated for the top predictors to evaluate feature importance.

Model training and evaluation were performed using a distributed GPU compute environment, optimized for large-scale data processing, including 96 vCPUs, 768 GiB system memory, 8 NVIDIA V100 GPUs (32GB each), and 100 Gbps bandwidth.

Parallel execution and resource orchestration were handled by Dask (version 2024.7.1), which enabled distributed scheduling, spill control, and GPU-to-GPU communication at scale. GPU-backed computation was managed using the RAPIDS ecosystem (RAPIDS AI,

2024), specifically cuDF for dataframe operations and CuPy for array computations. RAPIDS provides a suite of GPU-accelerated libraries for data science workflows, enabling efficient manipulation of large datasets and seamless integration with Python-based machine learning pipelines.

F. Bias and Sensitivity Testing Methods

Equity was identified by DHCS as a key principle of the model development process, and academic leaders in the space were consulted for their opinions and recommendations at several stages such as outcome development, model training, tiering, and monitoring.

Before model training commenced, the WG evaluated prevalence rates of individual outcomes by subgroup, and consulted with DHCS subject matter experts, available literature, and clinical experts on these prevalence rates.

The Work Group approached model training and tiering evaluation from two angles: (1) whether models performed consistently across demographic subgroups (e.g., similar recall, AUC), and (2) whether high-risk flags were equitably distributed across those subgroups to reach members with relevant outcomes across models, to support equitable prioritization of members for services. During model training, model performance results were reviewed for each subgroup and evaluated by the WG.

During Tiering Analysis, the Work Group assessed model performance across equity subgroups using a relative benchmark. The general guidance adopted by DHCS was that subgroup recall values should fall within $\pm 20\%$ of the statewide average recall. This threshold was used as a screening tool to highlight meaningful differences and inform decision-making around tiering options. While not a definitive standard for fairness, this approach supported a consistent, interpretable evaluation of subgroup-level variation during the selection process.

Equity Subgroups were considered across Race, Ethnicity, Sex, Language, and Access:

- | | |
|-------------------------------------|---------------------------------------|
| » New Medicaid Enrollee | » Asian |
| » Primary Language English | » Black or African American |
| » Primary Language Spanish | » Native Hawaiian or Pacific Islander |
| » Female | » Other Race |
| » Male | » Unknown Race |
| » Hispanic or Latino | » White |
| » American Indian or Alaskan Native | |

Note: We initially included Medicare Duals and Housing Insecure members in the above list but ultimately decided not to include them in the formal equity evaluation during the Tiering Analysis for the following reasons. Medicare Duals were excluded from the Underutilization domain due to data availability limitations (unavailable Medicare claims) which could have resulted in false positives – flagging members as underutilizing services even when those services were delivered but covered by Medicare and thus not visible in the data. Housing Insecurity is the only outcome in the social domain in V1, and prior housing insecurity was a strong predictor of future housing insecurity, thus, housing insecure members by design have a high likelihood of becoming high risk in the model.

When subgroup-level disparities were identified during development, the Work Group considered several remediation options, including refining outcome definitions. Work group evaluation of disparities and selection of remediation approach, if any, was based on case-by-case evaluation incorporating contextual understanding, recognizing that not all disparities in performance reflect algorithmic bias. For example, the inclusion of a pediatric Chlamydia screening measure improved predictive performance for females but was not applicable to males. After reviewing the data and model outputs, the Work Group concluded this pattern reflected real-world clinical relevance and chose to retain the outcome.

G. Tiering Analysis Methods

1) Overarching Goals of Tiering

The RSST Algorithm was developed to help DHCS identify Medi-Cal members who are most likely to benefit from proactive services and outreach. To support implementation, DHCS established an initial high-risk tier ceiling: no more than 10% of the total Medi-Cal population should be classified as high risk in Version 1. This target was informed by policy guidance and stakeholder input and reflects what is operationally feasible for managed care plans to assess and engage. Over time, this ceiling may be adjusted as additional tiering strategies and capacity are introduced.

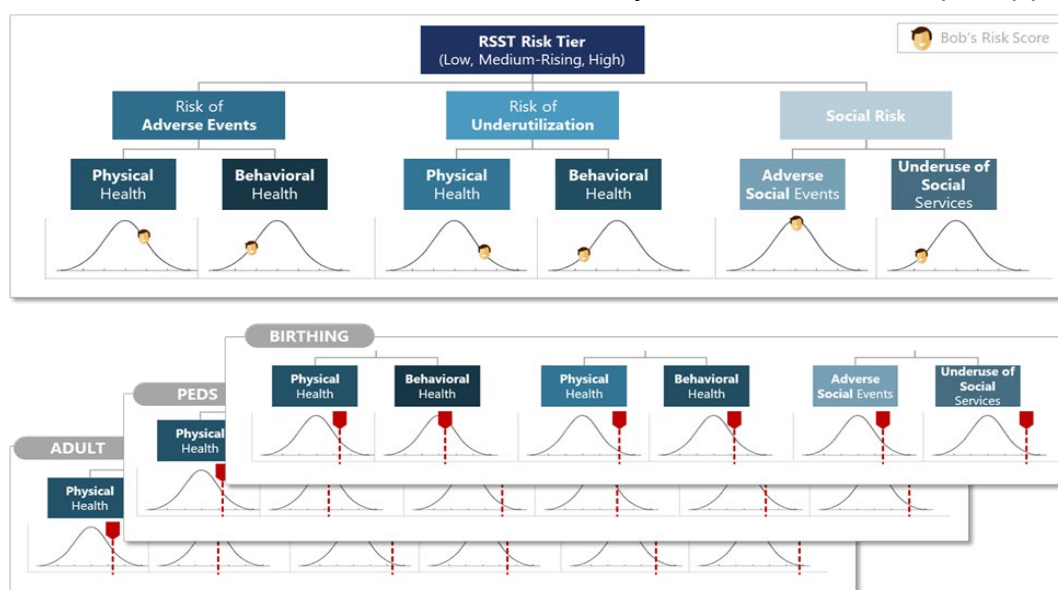
To ensure a consistent definition of high risk across plans and regions, the algorithm outputs a numeric risk score for each of five subdomain models, separately for adults and pediatrics. These scores are translated into tier assignments—low, medium, or high—using fixed threshold values. In total, 20 thresholds are applied: one for each subdomain across the two population strata. This structure enables a common definition of “high risk” that is applied uniformly, supporting alignment in how resources and services are allocated statewide.

To define these thresholds, the RSST Work Group used predicted risk scores from January 2023—the most recent month in the training environment for which subsequent 12-month outcomes were knowable. High-risk thresholds were set at the subdomain level such that, in combination, they would flag approximately 10% of the Medi-Cal population as high risk. Given that this was done on historic data, but launching on the current population, an extra step was taken to translate 2023 subdomain-level high risk rates to 2025 corresponding risk score thresholds.

After that adjustment, these values are held fixed over time rather than recalculated each month, allowing the definition of high risk to remain stable even as population characteristics shift. This means that the flagged proportion may vary slightly in future months and monitoring and adjustments will be needed on a regular basis for operational continuity.

In contrast, the medium-risk threshold was set using a simpler method—at the point corresponding to 50% recall in each subdomain model. The selection of high-risk thresholds was treated as a key policy decision—just as impactful as the model design itself—because it is eventually intended to determine which members are prioritized for outreach and assessment.

Note: The placement of “medium” risk threshold should be considered a pre-decisional placeholder, while the team conducts further analyses to refine this simple approach.



Note: The above visual illustrates the tiering thresholds across the underlying subdomain models. The visual includes Birthing population, as well as Social Underuse subdomain model, neither of which are included in V1 launch, but are planned for future versions.

2) Tiering Options Considered

Because the RSST Algorithm produces separate model outputs for each of the five subdomains, the RSST Work Group considered multiple options for how to define “high risk” at the subdomain (i.e. model) level, and the implications of each of these options for member-level high risk tiers. Each option prioritizes a different principle for how high risk tiers were determined across the subdomains and members flagged for assessment.



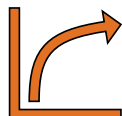
Option 1 flagged the same percentage of members as high risk from each subdomain. This was the most straightforward and interpretable approach but did not adjust for model performance and tended to over-represent members flagged by models predicting less common outcomes. As a result, more members flagged as high risk under this option may not go on to experience an adverse outcome.

Option 2 aimed to balance risk tier performance across subdomains by ensuring each model flagged high-risk members at the same level of recall (sensitivity)—that is, flagging a similar share of members who actually go on to experience a poor outcome in that subdomain. This resulted in different proportions of members flagged across models, providing a more complete representation of how effectively each model identifies true cases.

Option 3 prioritized overall recall (for at least one outcome in any subdomain) by concentrating high-risk flags in the models with the highest model performance. This option pulled more high-risk members from models that were better able to identify future adverse outcomes, potentially maximizing the number of members flagged who would go on to need services, but allowing the share of members flagged who actually go on to experience a poor outcome in that subdomain to vary across subdomains.

Each of these options was evaluated for performance (recall and NNT) under fixed resource constraints, such as targeting approximately 10% of the population as high risk. While the total number of members flagged was consistent across options, the distribution of who was flagged and why varied significantly, with tradeoffs in equity, interpretability, and predictive value.

Table 6. Summary of Tiering Options Considered

	Option 1 	Option 2 	Option 3 
Highest Priority	Having the same percentage of members flagged as high risk from each subdomain	Balancing the level of recall (sensitivity) across all subdomains	Maximizing total level of recall (sensitivity) across all subdomains
<i>Because outcomes across subdomains are not equally common and each algorithm has different performance...</i>			
Implications	<ul style="list-style-type: none"> » Comparatively over-weights less-common outcomes among those identified as high risk » May result in more people flagged as high risk who do not go on to have a poor outcome 	<ul style="list-style-type: none"> » Results in different percentages of members flagged as high risk in each domain 	<ul style="list-style-type: none"> » May identify more overall people flagged as high risk who would go on to have a poor outcome

a. Tiering Technical Methods

1. Option 1

To implement Option 1, we operationalized a thresholding strategy that applied a fixed top-K approach within each subdomain model. This strategy aimed to flag a consistent percentage of individuals—such as the top 1%, 5%, or 10%—as high-risk each month in the test set.

We began by calculating monthly risk score rankings for each subdomain. For every person-month, individuals were ranked in descending order of predicted risk. A fixed proportion of individuals (e.g., top 1%) were designated as high-risk based on this rank. The lowest risk score within the flagged group was extracted to serve as the month-specific cutoff. These cutoffs were then averaged across months to produce a stable threshold per subdomain model. This step ensured temporal smoothing and reduced volatility in the high-risk designations over time. Once the subdomain thresholds were computed, a member was flagged as high-risk within a subdomain if their subdomain predicted risk score met or exceeded the corresponding threshold.

To support implementation under multiple policy configurations, we ran this process using three global targeting constraints: 5%, 8%, and 10% of the total Medi-Cal population. For each setting, the appropriate top-K percentages were applied per

subdomain to ensure that cumulative high-risk designation across the state remained within the specified constraint.

2. Option 2

To operationalize Option 2, we implemented an iterative threshold selection process designed to achieve consistent recall across subdomains while adhering to a global constraint on the total size of the high-risk tier. This approach directly supported the goal of equitable identification across risk domains, ensuring that no single model disproportionately influenced high-risk designation due to differences in outcome prevalence or model performance.

The implementation proceeded in two stages. First, we estimated the risk threshold required to achieve a target recall value using a binary search algorithm for each subdomain model. Candidate thresholds were sampled from the range of predicted risk values within each model, and recall was computed based on true positives and false negatives. The search continued until the threshold produced recall within a pre-specified tolerance of $\pm 2\%$ of the target recall.

In the second stage, we calculated the percent of members flagged as high-risk using the selected subdomain thresholds. If the total exceeded or fell short of the global target (e.g., 5% of all members), the recall target was adjusted, and the process was repeated. This loop continued iteratively, adjusting the recall target in small increments and re-running the binary search for each subdomain model until the final selected set of thresholds collectively satisfied both the balanced recall constraint and the population-level targeting constraint ($\pm 0.5\%$ of the specified 5%, 8%, or 10% total).

This approach yielded a threshold set that provided uniform recall across domains, supported transparent documentation of tradeoffs, and served as a strong candidate for policy options focused on equity and broad-based targeting.

3. Option 3

To implement Option 3, we operationalized a strategy designed to identify the combination of subdomain thresholds that would maximize overall recall while ensuring that the total number of flagged individuals remained within a global targeting constraint (e.g., 10% of the Medi-Cal population).

Threshold selection was performed using Optuna, which systematically searched over combinations of pre-defined candidate thresholds. In each optimization trial, thresholds were set for each subdomain. These thresholds were applied to generate binary flags indicating whether each individual was predicted to be high-risk in each subdomain. These subdomain-level flags were then coalesced to create a composite high-risk

designation across subdomains, and this designation was compared to observed outcomes to calculate recall. The primary objective of the optimization was to maximize this recall score.

Every trial was also subject to the initial hard constraint: the proportion of members flagged as high-risk must fall within a fixed margin ($\pm 0.5\%$) of the specified global target (e.g., 5%, 8%, or 10%). Trials that produced target rates outside this range were pruned early to avoid unnecessary computation. This combination of constraint-based pruning and performance-driven scoring allowed the optimization algorithm to efficiently converge toward high-recall configurations that remained operationally feasible with the global targets. The final result of this process was a set of optimized subdomain thresholds that jointly produced the highest recall achievable under the imposed constraint.

4. Computing Final Metrics

After identifying thresholds and assigning high risk flags among all members enrolled in Medi-Cal in January 2023, we created a derived variable to indicate whether the member had experienced an outcome in the following 12 months. This enabled us to compute the empirical measures used for tiering evaluation such as recall, NNT, and distribution of risk by across risk domains.

3) Tiering Analysis Evaluation Criteria

Options were evaluated and scored based on five key principles and associated empirical measures laid out ahead of time (see below) to drive selection of tiers most aligned with the DHCS vision for the RSST algorithm. These principles were used to assess the advantages and disadvantages of each option to drive policymakers' decision of which option to select for the v1 release. While the measures in the list are objective, the relative balancing of these principles was a subjective policy decision that required iterative, data-informed discussions.

Principle	Description	Empirical Measure
Technical Performance	The predictive performance of the RSST algorithm overall, compared with industry standards and subject matter expertise.	<ul style="list-style-type: none"> AUC (Area Under the ROC Curve): Measures the trade-off between the model's recall and its specificity (true negative rate) across different thresholds. Recall: Percent of members with an adverse event who are flagged as high-risk. NNT: Number needed to treat before identifying a member who had an adverse event.
Balance of Risk Types	The balance across adverse events, underutilization, and social domains.	<ul style="list-style-type: none"> Distribution of domains among high-risk members: Among members who are flagged as high-risk, what percent of members are flagged because of each domain.
Equity	The predictive performance of the RSST algorithm by subgroups of interest.	<ul style="list-style-type: none"> Distribution of high-risk by subgroup: Among subgroups, what percentage of members would be flagged as high-risk. Comparative recall by subgroup: Compared to the overall State Weighted Average, who recall performs across each subgroup.
Potential Benefits	The possible benefits of driving improved health outcomes, timely interventions, and better resource allocation.	<ul style="list-style-type: none"> Net new assessments: Number of members who would receive a net new assessment who would not have received one otherwise. Count of net new needs assessments on members who went on to have an event.
Potential Costs and Impacts	The estimated costs of RSST needs assessments and possible direct impacts to stakeholders.	<ul style="list-style-type: none"> Estimated Costs Overall and by Plan: Estimate the net-new needs assessments and high-risk TCS events that would have been triggered by the RSST algorithm, in the context of 2023 experience.

VI. MODEL RESULTS

A. Outcomes Detail

1) Outcome Development Process

Each outcome proposed for RSST V1 was evaluated for feasibility of construction, consistency across populations, and overall utility in supporting meaningful risk stratification. Several outcomes were ultimately excluded due to data limitations or incomplete logic at the time of model development. For example, colorectal cancer screenings and influenza vaccinations require additional clinical data and could not be reliably identified solely through the available claims data.

DHCS subject matter experts were consulted throughout the process to ensure that outcome definitions were consistent with internal measurement logic and program standards. For example:

- » The primary care visit outcome used a DHCS-specified taxonomy of provider types to reflect established definitions.
- » The dental outcome logic was refined in collaboration with a DHCS dental SME to ensure appropriate coding of fluoride and preventive care.
- » Housing insecurity was defined using a DHCS-supplied list of address strings associated with shelters and transitional housing programs.

Several outcomes also underwent significant revision to improve specificity. The underutilization domain in particular required substantial refinement. Initial definitions were too broad, capturing large portions of the population and limiting the ability of the models to distinguish higher-risk members. The Work Group introduced more targeted eligibility criteria for specific underutilization outcomes to better facilitate identification of clinically relevant subgroups. For example:

- » The adult primary care visit underutilization outcome was narrowed to include only members with recent hospital or ED use who did not have a follow-up visit.
- » The pediatric dental care underutilization outcome was limited to children with elevated clinical risk, including those with cardiac conditions, developmental delay, autism spectrum disorder, or recent ED or inpatient visits for dental concerns, who did not have a dental visit.

- » The adult care disengagement outcome was defined as adults with no claims history in the two years prior to the index date who also had no claims in the following year.

These adjustments helped ensure that the outcomes used in RSST V1 were both actionable and appropriate for use in predictive modeling across the Medi-Cal population. Definitions for each outcome are provided below, grouped by strata, domain, and subdomain.

2) Adult Outcome Detail

a. Adverse Events: Physical

Outcome	Description	Data Source(s)
Mortality	The occurrence of death for an individual.	Supplemental Member Eligibility
Morbidity	Y/N Indicator of a net new increase of ≥ 1 in Charlson Comorbidity Index (CCI) count indicating an individual's departure from a state of physiological or psychological well-being. The CCI includes 17 conditions including diabetes, chronic pulmonary disease, congestive heart failure, cancer, and liver disease.	Medical Claims
All Cause Inpatient Admission	Y/N Indicator of any unplanned hospital admission, regardless of diagnosis or reason for admission.	Medical Claims
3 or More All Cause ED Visits	Y/N indicator that member had 3 or more visits to an ED or ED setting in a 12 month period, regardless of primary reason for the visit.	Medical Claims

b. Adverse Events: Behavioral

Outcome	Description	Data Source(s)
Care for Unintentional Drug Overdose	Healthcare encounter for care related to unintentional drug overdose.	Medical Claims
Injection Drug Related Adverse Event	Healthcare encounter for complications typically related to injection drug use within 90 days of an event with a confirmed opioid diagnosis.	Medical Claims
Care for Intentional Self Harm	Healthcare encounter for care related to intentional self-harm.	Medical Claims
Psychiatric Admission	Hospitalization where primary diagnosis is a psychiatric disorder (i.e., inpatient admission with primary diagnosis of mental health conditions).	Medical Claims

Outcome	Description	Data Source(s)
Two or More Psychiatric ED Visits	Two or more visits in ED or ED setting where primary reason for the visit is a psychiatric condition.	Medical Claims
Co-Occurring High ED Utilization and MH/SUD Care	Members with one or more Short Doyle visits and three or more all cause ED visits for any condition.	Medical Claims, Supplemental Claims

c. Underutilization: Physical

Outcome	Description	Data Source(s)
Underuse of Primary Care Visit (among members with hospital usage)	Among adult members who had at least 1 ED Visit or Inpatient Admission in the 12 months prior to index date, Y/N Indicator that the member did not have at least 1 outpatient visit to a primary care provider, based on DHCS definition of a PCP visit in the next 12 months.	Medical Claims
Underuse of Dental Care (among a subset of members with prior dental related usage)	Members who had a visit with a dental diagnosis in a non-dental setting or have diagnoses indicating higher risk (i.e. cancer, cardiac) in prior year who did NOT have 1+ dental care visit.	Medical Claims
Underuse of STI Screening (Chlamydia) Among Females 16-24 Years	Among female members aged 16–24 years as of the index month, Y/N Indicator that the member did NOT receive a chlamydia screening in the 12 month period post-index month. Note for adult population this is 18-24 yr olds.	Medical Claims
Underuse of Appropriate Pharmacotherapy for Common Indications (among members taking medications for those conditions)	Among adult members who had at least 1 relevant diagnosis code and 1 disease-specific medication fill in the year pre-index date, separately for Hypertension and Diabetes, this is a Y/N indicator of sub-optimal medication days covered (<80% PDC) in the year post-index date for those medications.	Medical Claims, Pharmacy Claims
Underuse of Asthma Controller Medications (among members with Asthma)	Y/N Indicator that members who were diagnosed with Asthma in the 12 months prior to index date, did NOT have a ratio of controller medications to total asthma medications of 0.50 or greater during the 12 months post index date.	Medical Claims, Pharmacy Claims
No Medical or Pharmacy Claims	Y/N Indicator that an adult member with no medical or pharmacy claims in the 2 years prior to index date will also have no medical or pharmacy claims in the 12 months post index date.	Medical Claims, Pharmacy Claims

Outcome	Description	Data Source(s)
(among members with no claims history)		

d. Underutilization: Behavioral

Outcome	Description	Data Source(s)
Underuse of MH/SUD Office Visits (among members with BH/SUD)	Members with an established MH/SUD diagnosis in prior year who did NOT have an ambulatory or preventive office visit (with any provider) with a diagnosis of MH/SUD.	Medical Claims
Underuse of MH/SUD-Related Primary Care (among members with MH/SUD conditions)	Members with an established MH/SUD diagnosis in prior year who did NOT have a PCP visit with a diagnosis of MH/SUD.	Medical Claims
Underuse of Antidepressant Medications (among members with the condition and past medication fills)	Among adult members with a Major Depression diagnosis code and an Antidepressant Medication fill in the year prior to the index date, this is an indicator that the member had <180 days supply of those medications in the year post-index date.	Medical Claims, Pharmacy Claims
Underuse of Antipsychotics (among members with schizophrenia)	Indicator of whether adult members who had a schizophrenia diagnosis code AND who were dispensed any antipsychotic medications in the prior 12 months, remained on those medications, defined as having <80% PDC (<292/365d) within the prediction period by any antipsychotic medication (oral or long lasting injection) in the 12 months post-index date.	Medical Claims, Pharmacy Claims
Underuse of Opioid Agonist Therapy (among members with OUD)	Y/N indicator of those who do NOT fill any prescription for an appropriate opioid agonist in the 12 months post-index date, among adult members with an OUD diagnosis and a relevant medication fill in the 12 months pre-index date.	Medical Claims, Pharmacy Claims

e. Social: Adverse Events

Outcome	Description	Data Source(s)
Housing Instability	DHCS business logic measuring a proxy of housing instability using a combination of ICD10 z-codes, as well as how an individual's address is classified upon Medi-Cal enrollment, and the type of address used (e.g. homeless shelter, group housing).	Medical Claims (z-codes), Eligibility (street address)

3) Pediatric Outcome Detail

a. Adverse Events: Physical

Outcome	Description	Data Source(s)
All Cause Inpatient Admission	Y/N Indicator of any unplanned hospital admission, regardless of diagnosis or reason for admission.	Medical Claims
3 or more all cause ED visits	Y/N indicator that member had 3 or more visits to an ED or ED setting in a 12 month period, regardless of primary reason for the visit.	Medical Claims
New Diagnosis of common chronic illness	Y/N Indicator of a pediatric patient increasing their rolling count of chronic illnesses anytime during the post-index 12 months	Medical Claims
Morbidity	Y/N Indicator of a pediatric patient's movement to a level of higher medical complexity sometime during the rolling 12-month post-index period, as defined by the Pediatric Medical Complexity Algorithm (PMCA), such as moving from no chronic disease to a noncomplex chronic disease (NC-CD) state, or from a noncomplex chronic disease to complex chronic disease (C-CD) state.	Medical Claims

b. Adverse Events: Behavioral

Outcome	Description	Data Source(s)
New Diagnosis of mental illness	Y/N indicator that a member with no MH diagnoses (based on Claims based) in the 12 months prior to index date, had a claim with a MH diagnosis code in the 12 months post index date.	Medical Claims
New diagnosis of developmental delay	Y/N indicator that a member with no DD diagnoses (based on CDPS mapping table) in the 12 months prior to index date, had a claim with a DD diagnosis code in the 12 months post index date.	Medical Claims
New diagnosis of SUD	Y/N indicator that a member with no SUD diagnoses (based on Claims based and F10-F19 ICD10Dx codes) in the 12 months prior to index date, had a claim with a SUD diagnosis code in the 12 months post index date.	Medical Claims
Psychiatric Admission	Hospitalization where primary diagnosis is a psychiatric disorder (i.e., inpatient admission with primary diagnosis of mental health conditions).	Medical Claims
Psychiatric ED Visit	One or more visits in ED or ED setting where any diagnosis code on the claim for the visit indicates a psychiatric condition.	Medical Claims
Care for Drug Overdose	Healthcare encounter for care related to unintentional drug overdose.	Medical Claims

Outcome	Description	Data Source(s)
Care for Intentional Self Harm	Healthcare encounter for care related to intentional self-harm.	Medical Claims

c. Underutilization: Physical

Outcome	Description	Data Source(s)
Underuse of STI Screening (Chlamydia) among females 16- 24 years	Among female members aged 16–24 years as of the index month, Y/N Indicator that the member did NOT receive a chlamydia screening in the 12-month period post-index month. Note that for pediatric population this is 16-17yr olds.	Medical Claims
Underuse of Well child visits (for children in the first 30 months of life)	Members aged 16 to 30 months at the end of the outcome period who did NOT have at least 2 well-care visits. Note: Members less than 16 months of age at the end of the outcome period were not eligible to be included in this measure as they were not included in the pediatric population (due to being 0 to 3 months of age) at the beginning of the outcome period.	Medical Claims
Children and Adolescents with no claim's utilization	Among members aged 30 months to 17 years old with no medical or pharmacy claims in the year prior to index date, this is a Y/N Indicator that the member also had no medical or pharmacy claims in the year post index date.	Medical Claims, Pharmacy Claims
Underuse of Topical Fluoride and/or Dental Care	Among a subset of pediatric members with any of the following conditions in the year prior to the index date, 1) a cardiac diagnosis (based on PMCA body systems), 2) an ED visit or Inpatient admission with a dental diagnosis code (primary or secondary position), 3) an autism spectrum disorder diagnosis, 4) a developmental delay diagnosis, This is a Y/N indicator that those members will NOT receive at least one topical fluoride application (CPT/CDT codes) or at least one dental claim (CDT codes) in the 12 months post-index date,	Medical Claims, Dental Claims
Underuse of Immunizations	Members aged 0 to 12 months at the start of the outcome period who did NOT have at least one dose of 6 different immunizations recommended for this age bracket.	Medical Claims, Pharmacy Claims
Underuse of Asthma Controller Medications	Among members with an Asthma diagnosis code in the 12 months prior to index date, this is a Y/N Indicator that the member did NOT have a	Medical Claims, Pharmacy Claims

Outcome	Description	Data Source(s)
(among members with Asthma)	ratio of controller medications to total asthma medications of 0.50 or greater during the 12 months post index date.	

d. Underutilization: Behavioral

Outcome	Description	Data Source(s)
Underuse of PCP Visits (among members with BH/SUD)	Patients with an established MH/SUD diagnosis in prior year who did NOT have a PCP visit in the prior year or the current year.	Medical Claims
Underuse of MH/SUD related office visits (among members with MH/SUD conditions)	Members with an established MH/SUD diagnosis in prior year who did NOT have an ambulatory or preventive office visit (with any provider) with a diagnosis of MH/SUD in the prior year or the current year.	Medical Claims
Underuse of Metabolic Screenings (among members with Antipsychotics)	Among members with 2 or more antipsychotic prescriptions in the 12 months prior to the index date, this is a Y/N indicator that the member DID NOT receive the appropriate metabolic monitoring (both cholesterol and glucose tests) services in the 12 months post index date.	Medical Claims, Pharmacy Claims
Underuse of Follow-up visits (among members with an admission or ED visit for MH/SUD)	Among members with a psychiatric ED visit or psychiatric admit, this is a Y/N indicator that the member DID NOT have any appropriate follow-up visits within 30 days of the hospital visit, in the 12 months post index date.	Medical Claims
Underuse of ADHD follow-up care (among members taking ADHD medications)	Among members who had at least 210 days supply of ADHD medications in the 12 months prior to index date, this is a Y/N indicator that the member DID NOT receive an adequate follow up visit in the 12 months post index date.	Medical Claims, Pharmacy Claims

e. Social: Adverse Events

Outcome	Description	Data Source(s)
Housing Instability	DHCS business logic measuring a proxy of housing instability using a combination of ICD10 z-codes, as well as how an individual's address is classified upon Medi-Cal enrollment, and the type of address used (e.g. homeless shelter, group housing).	Medical Claims (z-codes), Eligibility (street address)

B. Predictors Detail

1) Socio-Demographic Risk Factors

Predictor	Description	Data Source	Strata	Lookback
months_enrolled	Total number of distinct Medi-Cal enrolled months per unique member, in the full history ("ever"), and within a rolling 3- and 12-month lookback from index date.	Eligibility	Both	3mo, 12mo, ever
disability_status	Composite Y/N indicator of a member having either ABD (Age Blind Disabled) or SPD (Senior and Persons with Disability) flags in their data in the 12 months prior to the index date.	Eligibility	Both	ever
gaps_in_enrollment	Number of distinct, continuous gaps in eligibility of at least 1 month in duration within a rolling 12-month lookback prior to the index date.	Eligibility	Both	12mo, ever
age	Whole number age as of the last day of the index month, based on date of birth value.	Eligibility	Both	
primary_language_indicators (n= 30 languages + missing)	Latest recorded primary member language as included in the APCD-CDL Member Eligibility layout, as of the index month.	Eligibility	Both	
race_category_indicators (n=8)	Latest recorded member race as included in the APCD-CDL Member Eligibility layout, as of the index month, with DHCS code mapping logic applied.	Eligibility	Both	
ethnicity_category (n=2)	Latest recorded member ethnicity as included in the APCD-CDL Member Eligibility layout, as of the index month, with DHCS code mapping logic applied.	Eligibility	Both	
sex (M, F, U)	Latest recorded sex as of the index month, as included in the APCD-CDL Member Eligibility layout CDLME018 provided by DHCS	Eligibility	Both	
county of residence indicator (n=58)	Latest recorded member county of residence as of the index month	Eligibility	Both	
aid_code_category_indicators (n=8)	Each member's aid code category was used as a categorical predictor, reflecting eligibility classification, and was one-hot encoded into binary indicators representing categories such as ACA expansion, adoption/foster care, CHIP, long-term care, parents/caretaker relatives and children, seniors and	Eligibility	Both	

Predictor	Description	Data Source	Strata	Lookback
	persons with disabilities, undocumented, and unknown.			
is_medicare_dual	Indicator of Medicaid/Medicare Dual enrollment via appearance of Medicare ID	Supplemental Member Eligibility	Both	
line_of_business_n on_medicaid	Y/N indicator if member was enrolled with DHCS via a non-MEDICAID line of business or "product id" in the 12 months prior to the index date (e.g. GHPP, CCS, FPACT).	Eligibility	Both	
avg_persons_per_ housing_unit	Average number of people per housing unit, based on member's ZCTA of residence (ACS 2017).	American Community Survey (ACS), 2017	Both	
median_income	Median annual income, based on member's ZCTA of residence (ACS 2017).	American Community Survey (ACS), 2017	Both	
pct_25_hs_degree	Percentage of residents age 25+ with at least a high school diploma, based on member's ZCTA of residence (ACS 2017).	American Community Survey (ACS), 2017	Both	
pct_adults_employ ed	Percentage of adults (age 16+) who are employed, based on member's ZCTA of residence (ACS 2017).	American Community Survey (ACS), 2017	Both	
pct_bach_degree	Percentage of residents with a bachelor's degree or higher, based on member's ZCTA of residence (ACS 2017).	American Community Survey (ACS), 2017	Both	
pct_hh_w_2_paren ts	Percentage of households with children that have two parents, based on member's ZCTA of residence (ACS 2017).	American Community Survey (ACS), 2017	Both	
pct_hh_w_automob ile	Percentage of households with access to at least one vehicle, based on member's ZCTA of residence (ACS 2017).	American Community Survey (ACS), 2017	Both	

Predictor	Description	Data Source	Strata	Lookback
pct_income_over_2x_poverty	Percentage of families with income at or above 200% of the federal poverty level, based on member's ZCTA of residence (ACS 2017).	American Community Survey (ACS), 2017	Both	

2) Transition Events

Predictor	Description	Data Source	Strata	Lookback
moving_different_address	Indicator for the month that a member moves to a different address than where they lived in the prior month (for model purposes, this will be a Y/N if member moved in a 12 month pre-index period)	Eligibility	Both	12mo
moving_different_county	Indicator for the month that a member moves to a different geographic county than where they lived in the prior month (for model purposes, this will be a Y/N if member moved in a 12 month pre-index period)	Eligibility	Both	12mo
new_medicaid_enrollee	Indicator of the earliest Medi-Cal enrollment three month period for the member, looking back to the full available history up to 1/1/2016	Eligibility	Both	12mo
returning_medicaid_enrollee	Indicator for the first three months of a member regaining Medi-Cal enrollment after an Enrollment Gap of any duration 1 month or more	Eligibility	Both	12mo
snf_indicator	Y/N indicator of any SNF claims utilization in the 12mo prior to the index date	Medical Claims	Both	12mo
switch_aid_category	Y/N indicator if member switches their Medicaid aid category in the last 12 months	Eligibility	Both	12mo
switch_job	Y/N indicator if member switches their line of business (product id) in the last 12 months	Eligibility	Both	12mo
switch_medical_plans	Y/N indicator that a member has changed which Medi-Cal Managed Care Plan (MCP) "Parent" they are managed by during the 12 month period prior to the index date; for Fee for Service (FFS) members, we flag if a member moves between FFS and managed care.	Eligibility	Both	12mo
transition_to_ihss	Y/N indicator that a member has any in home supportive services (IHSS) provided to them in the 12	Medical Claims	Both	12mo

Predictor	Description	Data Source	Strata	Lookback
	months prior to index date, using aid codes available on the APCD-CDL medical claims layout from DHCS.			
pediatric_to_adult	Indicator for the month that a member aged into the adult population strata. For the purposes of modeling this will result in a Y/N indicator applied to the entire pre-index 12 month lookback period if member turned 18 within that pre-index period	Eligibility	Adult Only	12mo

3) Prior Use

Predictor	Description	Data Source	Strata	Lookback
ip_claims	Count of Inpatient visits on distinct service dates, within three overlapping lookback periods: 3 months, 12 months prior to index date, as well as a full "ever" lookback to earliest data available pre index date. Based on place of service (POS) codes in claims data.	Medical Claims	Both	3mo, 12mo, ever
office_claims	Count of Office visits on distinct service dates, within three overlapping lookback periods: 3 months, 12 months prior to index date, as well as a full "ever" lookback to earliest data available pre index date. Based on place of service (POS) codes in claims data.	Medical Claims	Both	3mo, 12mo, ever
op_claims	Count of Outpatient visits on distinct service dates, within three overlapping lookback periods: 3 months, 12 months prior to index date, as well as a full "ever" lookback to earliest data available pre index date. Based on place of service (POS) codes in claims data.	Medical Claims	Both	3mo, 12mo, ever
bh_claims	Y/N indicator if member has at least 1 Short Doyle / Behavioral Health claim in a rolling 12 month lookback pre-index date.	Medical Claims	Both	12mo
ccs diagnosis & procedure code groupings binary flags (n=365)	Diagnosis: CCS groupings of a member's diagnosis codes, primary and secondary, in 12 month lookback period prior to index month (without regard to clinical setting). This flags for each member a Y/N indicator for each of the ccs minor categories. Procedure: CCS groupings of a member's procedure codes, in 12 month lookback period prior to index	Medical Claims	Both	12mo

Predictor	Description	Data Source	Strata	Lookback
	month (without regard to clinical setting). This flags for each member a Y/N indicator for each of the CCS minor categories specific to procedure codes.			
count_of_rx_fill	Count of prescription fills in the prior 12 months	Pharmacy Claims, Medispan	Both	12mo
dme_use	Y/N indicator of "any" DME/Devices claims in the 12 months prior to index date	Medical Claims	Both	12mo
rx Medispan drug class indicators (n=144)	Y/N Indicator of each distinct Medispan drug class that member has a filled prescription claim for in the 12 months prior to the index date.	Pharmacy Claims, Medispan	Both	12mo
rx_fill_hit_2000; rx_fill_hit_500	Two binary Y/N indicators of Total insurance paid amounts on prescription fills in the 12 months prior to the index date hits a) \$500 and b) \$2000 respectively.		Both	12mo

4) Outcomes as Risk Factors

Predictor	Description	Data Source	Strata	Lookback
all_cause_ed_visits	Count of distinct Emergency Department person-claim-service-dates in the 12 months prior to index date, using the same underlying logic as the all cause ED outcome.	Medical Claims	Both	3mo, 12mo, ever
all_cause_inpatient_admission	Count of distinct inpatient admission dates in the past 12 months, regardless of diagnosis.	Medical Claims	Both	3mo, 12mo, ever
dental_care	Binary indicator for whether the member had any dental care encounter in the past 12 months.	Medical Claims, Dental Claims	Both	3mo, 12mo, ever
drug_overdose	Count of distinct drug overdose event dates in the past 12 months.	Medical Claims	Both	3mo, 12mo, ever
intentional_self_harm	Binary indicator for whether there was any intentional self-harm event recorded in the past 12 months.	Medical Claims	Both	3mo, 12mo, ever
mh_sud_visits	Count of distinct outpatient or professional visits for mental health or substance use diagnoses in the past 12 months.	Medical Claims	Both	3mo, 12mo, ever
pcp_visits_bh_sud_diagnosis	Count of primary care visits with a behavioral health or substance use diagnosis in the past 12 months.	Medical Claims	Both	3mo, 12mo, ever

Predictor	Description	Data Source	Strata	Lookback
psychiatric_admission	Count of distinct psychiatric inpatient admissions in the past 12 months.	Medical Claims	Both	3mo, 12mo, ever
psychiatric_ed_visit	Count of distinct emergency department visits with psychiatric diagnosis in the past 12 months.	Medical Claims	Both	3mo, 12mo, ever
housing_instability	Binary indicator for any evidence of housing instability in the past 12 months, using DHCS business logic for address string values and ICD10 z-codes.	Medical Claims, Eligibility	Both	12mo
std_screening_chlamydia	Binary indicator for whether the member received chlamydia screening in the past 12 months.	Medical Claims	Both	12mo
adult_morbidity	Count of distinct service dates for selected adult morbidity diagnoses in the past 12 months. Specifically diseases within the Charlson Comorbidity Index logic.	Medical Claims	Adult Only	3mo, 12mo, ever
injection_drug_related_adverse_event	Count of distinct events indicating injection drug use complications in the past 12 months.	Medical Claims	Adult Only	3mo, 12mo, ever
primary_care_visit	Count of distinct primary care visits in the past 12 months.	Medical Claims	Adult Only	3mo, 12mo, ever
antidepressant_medication_use	Count of distinct antidepressant medication fill dates in the past 12 months.	Pharmacy Claims	Adult Only	12mo
antipsychotics_schizophrenia_use	Count of distinct antipsychotic medication fills for schizophrenia-related treatment in the past 12 months.	Pharmacy Claims	Adult Only	12mo
opioid_agonist_therapy_oud	Count of distinct fill dates involving opioid agonist therapy for opioid use disorder in the past 12 months.	Pharmacy Claims	Adult Only	12mo
pharmacotherapy_diabetes	Count of distinct fill dates for diabetes treatment in the past 12 months.	Pharmacy Claims, Medical Claims	Adult Only	12mo
pharmacotherapy_hypertension	Count of distinct fill dates for hypertension treatment in the past 12 months.	Pharmacy Claims, Medical Claims	Adult Only	12mo
glucose_cholesterol_screenings	Count of distinct screening events for glucose or cholesterol in the past 12 months.	Pharmacy Claims, Medical Claims	Pediatric Only	3mo, 12mo, ever
new_diagnosis_developmental_delay	Binary indicator for any new developmental delay diagnosis first appearing in the past 12 months.	Medical Claims	Pediatric Only	3mo, 12mo, ever

Predictor	Description	Data Source	Strata	Lookback
new_diagnosis_mental_illness	Binary indicator for any new mental illness diagnosis first appearing in the past 12 months.	Medical Claims	Pediatric Only	3mo, 12mo, ever
new_indicator_diagnosis_sud	Binary indicator for any new substance use disorder diagnosis first appearing in the past 12 months.	Medical Claims	Pediatric Only	3mo, 12mo, ever
pmca_body_systems_progressive	Count of distinct progressive body systems flagged by PMCA logic in the past 12 months.	Medical Claims	Pediatric Only	3mo, 12mo, ever
well_child_visits	Count of distinct well-child visit dates in the past 12 months.	Medical Claims	Pediatric Only	3mo, 12mo, ever
asthma_medication	Binary indicator for any asthma medication dispensed in the past 12 months.	Pharmacy Claims	Pediatric Only	12mo, ever
immunization_count	Count of distinct immunization events recorded in claims history.	Pharmacy Claims, Medical Claims	Pediatric Only	ever
net_new_autism_diagnosis	Binary indicator for any new autism diagnosis first appearing in the past 12 months.	Medical Claims	Pediatric Only	12mo, ever
net_new_neurological_diagnosis_no_autism	Binary indicator for any new neurological diagnosis (excluding autism) first appearing in the past 12 months.	Medical Claims	Pediatric Only	12mo, ever
topical_fluoride	Binary indicator for whether topical fluoride treatment was provided in the past 12 months.	Medical Claims, Dental Claims	Pediatric Only	12mo, ever

C. Technical Environments Setup

Model training and evaluation were performed using a distributed GPU compute environment, optimized for large-scale data processing, including 96 vCPUs, 768 GiB system memory, 8 NVIDIA V100 GPUs (32GB each), and 100 Gbps bandwidth.

Parallel execution and resource orchestration were handled by Dask (version 2024.7.1), which enabled distributed scheduling, spill control, and GPU-to-GPU communication at scale. GPU-backed computation was managed using the RAPIDS ecosystem (RAPIDS AI, 2024), specifically cuDF for dataframe operations and CuPy for array computations. RAPIDS provides a suite of GPU-accelerated libraries for data science workflows, enabling efficient manipulation of large datasets and seamless integration with Python-based machine learning pipelines.

D. Model Tuning Results

To identify high-performing configurations across multiple model architectures, we implemented an adaptive and parallelized hyperparameter optimization framework using Optuna, in combination with Dask and GPU-accelerated training backends. The optimization objective was to maximize the area under the ROC curve (ROC-AUC) on a held-out validation set. This tuning strategy was applied independently to each model and for each outcome, enabling architecture-specific configurations tailored to individual risk signals.

All tuning experiments were conducted using consistent validation data drawn from a fixed temporal window (Feb 2020 – Jan 2021), ensuring fair comparisons under conditions of temporal drift. Training data (the “search set”) for the putative models with different hyperparameters were drawn from a separate subset of the training period (Jan 2019 – Jan 2020).

To support efficient search across high-dimensional hyperparameter spaces while maintaining tractability, we randomly sampled 10–30% of the training data (depending on overall size) to create a search set. The sampled data were split into a search dataset, which included data from Jan 2019–Jan 2020, and a validation dataset drawn from a 30% random sample from Feb 2020–Jan 2021. This structure ensured that models were tuned on a reduced but representative subset of earlier data and evaluated on temporally newer data—effectively testing their ability to generalize under real-world deployment conditions.

Trials were distributed across multiple GPUs using Dask, and all computation was orchestrated through a central Dask distributed client to manage task execution, memory allocation, and GPU scheduling.

1) Optimization Framework

Optuna’s Tree-structured Parzen Estimator (TPE) sampler with multivariate optimization was used to guide the sampling process. Early stopping and resource-aware pruning were implemented using the HyperbandPruner, enabling efficient allocation of computational effort toward the most promising regions of the search space.

All trials were managed through Dask’s distributed client, allowing for parallel trial execution across 8 GPUs. Models were evaluated using ROC-AUC computed on the validation set, with each trial’s results logged and persisted.

Key trial management techniques included:

- » Timeout-based stopping: Each tuning session was limited to one hour (3600 seconds).
- » Performance-based early stopping: Tuning was terminated if no improvement was observed over 10 consecutive trials.
- » Pruning-based termination: Studies were halted if 5 consecutive trials were pruned, suggesting convergence.

2) Model-Specific Search Spaces

a. XGBoost (Gradient Boosted Trees)

XGBoost was trained using the GPU-accelerated histogram method (`tree_method='hist'`, `device='cuda'`). The hyperparameter space included:

- » `n_estimators`: 1 to 3000
- » `learning_rate`: 10 log-spaced values from 0.001 to 0.1
- » `max_depth`: {2, 4, 6, 8, 10, 12, 14, 16}

Training and evaluation were performed using `DaskQuantileDMatrix` to support GPU-native batching. To stabilize learning under class imbalance, sampling methods and max delta steps were determined (i.e. `sampling_method='gradient_based'` and `max_delta_step=1`).

Validation AUC was computed using `cuML`'s `roc_auc_score` to ensure consistency with the RAPIDS pipeline.

b. LightGBM

LightGBM was optimized using its GPU-enabled training backend, with similar hyperparameter settings to XGBoost for consistency:

- » `n_estimators`: 1 to 3000
- » `learning_rate`: 10 log-spaced values between 0.001 and 0.1
- » `max_depth`: {2, 4, 6, 8, 10, 12, 14, 16}

The model was configured with `boosting_type='gbdt'` and `tree_learner='data'`, enabling distributed gradient boosting suitable for large datasets. Although LightGBM supports GPU training, explicit GPU device settings (e.g., `device='gpu'`) were not enforced.

Hyperparameter trials were parallelized using Dask with `n_jobs=2`, and tuning was conducted using Optuna with the same early stopping and pruning strategies as in XGBoost. Validation followed the same protocol, using probabilistic predictions on a time-split validation set.

c. Logistic Regression (Quasi-Newton with Elastic Net)

We tuned logistic regression using an elastic net penalty, enabling regularization along the full spectrum between L1 (lasso) and L2 (ridge). The search space consisted of:

- » C: log-uniform from $1e-5$ to $1e5$
- » l1_ratio: categorical values in $\{0, 0.01, 0.5, 0.99, 1.0\}$

Models were trained using a GPU-accelerated implementation of logistic regression from the cuML library, with the 'qn' solver (quasi-Newton optimization) and a maximum of 100 iterations per trial. Optimization was performed using Optuna with single-threaded tuning ($n_jobs=1$) and the same early stopping and pruning conditions as for the tree-based models.

Due to its relative computational efficiency, logistic regression was able to support a larger number of trials within the same time budget, making it a robust and interpretable baseline across outcomes. Performance was evaluated using ROC-AUC on a held-out validation set.

3) Final Model Selection

For each model and outcome, the trial with the highest ROC-AUC on the validation set was selected. The selected hyperparameters were then used to retrain the model on the combined training and validation periods (Jan 2019 – Jan 2021). This retrained model was evaluated on the temporally held-out test set (Jan 2022 – Jan 2023) for final performance reporting. Lastly, XGBoost consistently demonstrated the highest performance across most of the subdomain models. In a few cases where it did not perform best, LightGBM yielded slightly superior results. However, given the marginal differences in performance and the superior scalability of XGBoost for multi-GPU inference, XGBoost was selected as the final model for all subdomains.

E. Model Performance

Each subdomain model was reviewed using a standardized evaluation framework developed by the academic working group. This framework assessed models across key performance domains:

- » **Calibration:** Visual inspection of calibration curves to evaluate alignment between predicted risk and observed outcomes.

- » **Discrimination:** Metrics such as AUC, as well as precision, recall, and specificity—particularly at the top 5% risk threshold—were used to assess model separation and baseline operational utility.
- » **Feature Interpretability:** Review of feature importances to confirm that models prioritized clinically and contextually relevant inputs.
- » **Risk Distribution:** Assessment of risk score distributions and outcome co-occurrence to understand model behavior across the population.
- » **Hyperparameter Tuning:** Review of tuning results and parameter sensitivity to ensure performance gains were both robust and reproducible.

Models were approved for use only after meeting minimum thresholds for performance, calibration, and interpretability as defined by the working group.

Table 7. Summary of Model Performance Metrics Across Subdomains¹

Model	AUC	Accuracy	Precision	Recall	Specificity
Adults Behavioral Health Adverse Events	0.938	0.955	0.173	0.697	0.958
Adults Physical Health Adverse Events	0.879	0.928	0.595	0.367	0.978
Adults Social Risk Adverse Events †	0.936	0.959	0.343	0.689	0.966
Adults Behavioral Underutilization	0.805	0.316	0.961	0.066	0.993
Adults Physical Health Underutilization	0.769	0.421	0.941	0.075	0.992
Peds Behavioral Health Adverse Events	0.754	0.915	0.274	0.218	0.961
Peds Physical Health Adverse Events	0.775	0.919	0.341	0.261	0.965
Peds Social Risk Adverse Events †	0.923	0.952	0.080	0.700	0.954

¹ For the underutilization subdomains, these metrics are calculated only among persons eligible for at least one underutilization outcome.

Model	AUC	Accuracy	Precision	Recall	Specificity
Peds Behavioral Underutilization	0.782	0.254	0.969	0.062	0.993
Peds Physical Health Underutilization	0.763	0.466	0.933	0.081	0.992 ²

1) Adult Adverse Events (Physical Health)

The composite outcome for adult physical health adverse events had a prevalence of 8% in the adult population (N = 8,875,329). This composite included key drivers such as all-cause inpatient admissions (8%) and all-cause emergency department visits (6%). Additional outcomes included mortality (1%) and new morbidity, as measured by a rise in the Charlson Comorbidity Index (1%).

The final model selected was XGBoost, with `n_estimators = 1210`, `learning_rate = 0.0359`, and `max_depth = 12`. The model displayed excellent calibration. Predicted probabilities closely tracked observed event rates across all bins, with minimal deviation from the ideal diagonal line. The consistent fit supports the reliability of the risk scores for use in tiering and prioritization (see Appendix for calibration curves).

The model achieved an AUC of 0.879 and, at the model-specific top 5% risk threshold, a precision of 0.595, recall of 0.367, specificity of 0.978, and accuracy of 0.928, indicating high model performance and sufficient discrimination for a relatively common composite outcome (See Table 7).

Subgroup performance was stable across demographic segments. Recall was similar between males (38%) and females (36%) and remained consistent for the English (40%) language group. Higher recall was observed among American Indian or Alaska Native (42%) and Black or African American (43%) subgroups, with strong performance across race and ethnicity categories more broadly. Reductions in recall were observed among Spanish-speaking (31%) and new Medicaid enrollees (30%), potentially reflecting patterns of historical utilization and data availability. Full subgroup results are available in the Appendix.

² All metrics except AUC were calculated using each model's specific top 5% risk threshold, representing the highest-risk individuals in the population.

†The selected models were not the top performers in terms of AUC but were retained due to their close performance to the best (LightGBM) and the scalability benefits of XGBoost, which supports multi-GPU training and inference at production scale.

2) Adult Adverse Events (Behavioral Health)

The composite outcome for adult behavioral health adverse events had a prevalence of 1% in the adult population. It encompassed multiple rare but clinically significant outcomes. Psychiatric admissions and having three or more all-cause ED visits and at least one Short-Doyle encounter were the most common, each with a prevalence of approximately 1%. Other components—intentional self-harm, drug overdose, and injection-related adverse events—were less frequent (all <1%).

The final model selected for this subdomain was XGBoost with `n_estimators = 2833`, `learning_rate = 0.0599`, and `max_depth = 6`. The model demonstrated strong calibration across the full range of predicted probabilities. Predicted risks tracked closely with observed event rates, with most points aligning with the diagonal line of perfect calibration. Mild overprediction was observed in mid-to-high probability bins, but the overall fit supports the model's use for risk stratification and downstream tiering decisions (see Appendix for calibration curve).

The model achieved an AUC of 0.938 and, at the model-specific top 5% risk threshold, a precision of 0.173, recall of 0.697, specificity of 0.958, and accuracy of 0.955, demonstrating high model performance and effective discrimination of high-risk individuals within a low-prevalence outcome space.

Subgroup analysis at the top 5% risk threshold showed consistent performance across demographic strata. Recall was comparable between males (71%) and females (68%). Performance remained strong among Hispanic (65%) and non-Hispanic (72%) populations. Recall exceeded 73% among American Indian or Alaska Native (73%) and Black or African American (73%) individuals. We observed reduced recall among Spanish-speaking (49%) and newly enrolled Medicaid members (53%), most likely consistent with patterns of data availability and historical utilization.

3) Adults Underutilization (Behavioral Health)

The composite outcome for adult behavioral health underutilization had a prevalence of 7% among the total adult strata (N = 8,875,329) and a prevalence of 72% among adults (N = 861,901) eligible for at least one underutilization outcome. The remainder of metrics reported are among the eligible population. This composite was designed to flag gaps in behavioral health care and included: visits to primary care providers with behavioral health or substance use diagnoses (24%), ambulatory mental health or substance use treatment visits (52%), and pharmacologic management such as

antidepressants (16%), antipsychotics for schizophrenia (5%), and opioids for opioid use disorder (2%). These individual outcomes were assessed against a shared denominator defined by the composite inclusion criteria.

The final model was XGBoost, with `n_estimators` = 552, `learning_rate` = 0.0359, and `max_depth` = 16. The model demonstrated strong overall calibration, with predicted probabilities closely matching observed event rates across the risk spectrum. While slightly underestimating risk at the lower end, the model's predictions aligned well with observed frequencies across most bins, indicating good reliability for identifying underuse patterns.

The model achieved an AUC of 0.805 and, at the model-specific top 5% risk threshold, a precision of 0.961, recall of 0.066, specificity of 0.993, and accuracy of 0.316.

Subgroup performance for the behavioral health underutilization model remained stable across demographic groups, with precision consistently above 95% and recall ranging from 4.7% to 8.9%. Recall was highest among Spanish-speaking individuals (8.9%), American Indian or Alaska Native (8.7%), and Black or African American (7.9%) subgroups. Female individuals exhibited lower recall (5.5%) compared to males (7.9%), while English and Spanish speakers performed comparably overall. The model maintained specificity above 99% across all groups.

4) Adults Underutilization (Physical Health)

The composite outcome for adult physical health underutilization had a prevalence of 31% among the total adult strata (N = 8,875,329) and a prevalence of 63% among eligible individuals (N = 4,290,776). The remainder of metrics reported are among the eligible population. This composite identified gaps in routine and preventive care delivery, including missed primary care visits (18%), lack of any claims (17%), and insufficient dental care (8%). Additional measures included gaps in chlamydia screening (14%), pharmacy fills for hypertension (8%) and diabetes (5%), and suboptimal asthma medication use (3%). These outcome-specific rates reflect a shared denominator, as inclusion criteria were applied at the composite level.

The final model selected was XGBoost, with `n_estimators` = 1671, `learning_rate` = 0.0359, and `max_depth` = 10. The calibration curve revealed slight underprediction across the probability range, with observed event rates consistently exceeding predicted probabilities. The model achieved an AUC of 0.769 and, at the model-specific top 5% risk threshold, a precision of 0.941, recall of 0.075, specificity of 0.992, and accuracy of 0.421.

Subgroup analysis demonstrated consistent performance across demographic strata, with precision above 93% for all groups and recall ranging from 4% to 11%. Recall was highest among American Indian or Alaska Native (11.6%), Black or African American (10.0%), and White (8.6%) populations. Lower recall was observed among new Medicaid enrollees (4.4%) and Hispanic or Latino individuals (6.6%). Performance across sex and language groups remained similar, with recall for females at 7.6% and English speakers at 7.7%. The model preserved high specificity ($\geq 99\%$) across all subgroups.

5) Adults Social Risk (Adverse Events)

The composite outcome for adult social risk adverse events had a prevalence of 3% in the adult population. This outcome is defined entirely by a single event: documented housing insecurity, which, accordingly, accounts for the full composite.

The final model selected was XGBoost, with `n_estimators = 1868`, `learning_rate = 0.0359`, and `max_depth = 10`. The model demonstrated excellent calibration, with predicted probabilities closely aligned with observed event rates across the probability spectrum. The calibration curve shows minimal deviation from the diagonal, indicating well-calibrated scores even at the highest risk ranges.

The model achieved an AUC of 0.936 and, at the model-specific top 5% risk threshold, a precision of 0.343, recall of 0.689, specificity of 0.966, and accuracy of 0.959, indicating strong performance in identifying individuals experiencing housing instability.

Subgroup performance showed particularly strong recall among American Indian or Alaska Native (81%), White (76%), and male (74%) populations. Recall remained above 61% across most other subgroups, including Black or African American (66%), Hispanic (62%), and individuals reporting English as their primary language (70%). Slightly lower precision and recall were observed among Spanish-speaking individuals, but specificity exceeded 99% in that group, as in several others.

6) Pediatrics Adverse Events (Physical Health)

The composite outcome for pediatric physical health adverse events had a prevalence of 6% in the pediatric population ($N = 4,549,042$). This composite was primarily composed of all-cause inpatient admissions (3%) and all-cause emergency department visits (3%). Less common but clinically significant components included diagnoses of common chronic illnesses (1%) and new onset morbidity ($<1\%$). Together, these outcomes reflect a range of acute and chronic physical health burdens relevant to pediatric populations.

The final model selected was XGBoost, with `n_estimators = 1671`, `learning_rate = 0.0359`, and `max_depth = 10`.

The model for physical health adverse events in pediatrics displayed excellent calibration. Predicted probabilities closely tracked observed event rates across all bins, indicating strong reliability of the risk scores for guiding tier assignment.

The model achieved an AUC of 0.775 and, at the model-specific top 5% risk threshold, a precision of 0.341, recall of 0.261, specificity of 0.965, and accuracy of 0.919, reflecting solid overall model performance and discrimination in identifying children at highest risk for adverse physical health outcomes.

Subgroup analysis revealed consistent performance across most demographic groups. Recall was slightly higher among new Medicaid enrollees (36%), English-speaking (27%) and White children (27%), while performance across sex, ethnicity, and language groups remained balanced. Spanish-speaking and Asian subgroups showed slightly lower precision, potentially reflecting variation in prevalence or care patterns. Specificity remained high ($\geq 91\%$) across all subgroups.

7) Pediatrics Adverse Events (Behavioral Health)

The composite outcome for pediatric behavioral health adverse events had a prevalence of 6% in the pediatric population. The most common components included new diagnoses of mental illness (5%) and emergency department visits for psychiatric reasons (1%). Less frequent but critical components included new diagnoses of developmental delay ($<1\%$), substance use disorders (1%), psychiatric admissions ($<1\%$), and rare events such as intentional self-harm and drug overdose (both $<1\%$). These outcomes capture both early identification and escalation of behavioral health needs in pediatric care.

The final model selected was XGBoost, with `n_estimators = 1050`, `learning_rate = 0.0359`, and `max_depth = 12`.

The model exhibited moderate overcalibration, particularly in the higher predicted risk ranges. As shown in the calibration curve (see appendix), predicted probabilities above 0.5 consistently overestimate the true outcome frequency. Still, model predictions remain directionally correct and actionable for prioritizing high-risk individuals.

The model achieved an AUC of 0.754 and, at the model-specific top 5% risk threshold, a precision of 0.274, recall of 0.218, specificity of 0.961, and accuracy of 0.915, reflecting a focused ability to identify pediatric patients at risk for emerging or severe behavioral health needs.

Subgroup analysis showed balanced performance across most demographic groups. Recall was slightly higher among male children (23%) and Black or African American

children (25%), with notably strong precision among American Indian or Alaska Native (27%) and White children (26%). Spanish-speaking and Asian subgroups showed lower recall (18% and 18%, respectively), though specificity remained high ($\geq 93\%$) across all subgroups.

8) Pediatrics Underutilization (Behavioral Health)

The composite outcome for pediatric behavioral health underutilization had a prevalence of 5% among the total pediatric strata ($N = 4,549,042$) and a prevalence of 78% among eligible children ($N = 299,096$). The remainder of metrics reported are among the eligible population. This composite reflects insufficient engagement with behavioral health services, including ambulatory mental health or substance use treatment (71%) and primary care visits with behavioral health or substance use diagnoses (20%). Less frequent components included metabolic screenings (3%), follow-up care after admissions or ED visits for behavioral health (3%), and ADHD-specific follow-up ($<1\%$). These outcome rates are calculated using a shared denominator, consistent with the composite inclusion criteria.

The final model selected was XGBoost, with $n_estimators = 2090$, $learning_rate = 0.0129$, and $max_depth = 10$.

The model's calibration curve shows that predicted probabilities are generally undercalibrated, particularly in the lower to mid-range. In these bins, observed outcome rates exceed predicted probabilities, indicating underestimation of risk. Calibration improves in the upper range, where predictions align more closely with actual event frequencies.

The model achieved an AUC of 0.782 and, at the model-specific top 5% risk threshold, a precision of 0.969, recall of 0.062, specificity of 0.993, and accuracy of 0.254, demonstrating high precision in identifying members with likely behavioral care gaps, albeit with limited sensitivity in a high-prevalence context.

Subgroup performance was generally consistent, with precision consistently above 95% across all groups. Recall was highest among new Medicaid enrollees (18.3%), American Indian or Alaska Native (11.7%), Native Hawaiian or Other Pacific Islander (8.8%), and Asian (6.8%) subgroups, possibly reflecting lower historical system contact. The model maintained high specificity ($\geq 98\%$) across all subgroups.

9) Pediatrics Underutilization (Physical Health)

The composite outcome for pediatric physical health underutilization had a prevalence of 20% among the total pediatric strata (N = 4,549,042) and a prevalence of 59% among eligible children (N = 1,559,019). The remainder of metrics reported are among the eligible population. This composite reflects unmet preventive and routine care needs, including adolescent and well-child visits (23%), chlamydia screening (15%), and topical fluoride or dental care (10%). Other components included immunization status (4%) and asthma medication indicators (5%). These outcomes were assessed against a common denominator, reflecting inclusion criteria applied at the composite level.

The final model selected was XGBoost, with `n_estimators = 1909`, `learning_rate = 0.0215`, and `max_depth = 12`.

The calibration curve shows that the model is slightly undercalibrated in the lower- to mid-range predicted probabilities, where observed event rates exceed the model's estimates. However, it becomes well-aligned above a predicted probability of 0.60 and closely tracks the diagonal through the mid-to-high ranges.

The model achieved an AUC of 0.763 and, at the model-specific top 5% risk threshold, a precision of 0.933, recall of 0.081, specificity of 0.992, and accuracy of 0.466, reflecting a conservative but precise strategy for identifying children at highest risk of care underutilization.

Subgroup analysis showed consistently high precision across all demographics, with modest variation in recall. Asian (13%), Native Hawaiian or Other Pacific Islander (11%), and female (10%) subgroups showed relatively higher recall, while male children (5.1%) and Black or African American children (5.8%) showed lower sensitivity. Specificity exceeded 98% for all subgroups, underscoring the model's reliability in filtering lower-risk populations.

10) Pediatrics Social Risk (Adverse Events)

The composite outcome for pediatric social risk adverse events had a prevalence of 1% in the pediatric population (N = 4,549,042). This outcome was defined entirely by the presence of housing instability.

The final model selected was XGBoost, with `n_estimators = 2090`, `learning_rate = 0.0129`, and `max_depth = 10`.

The model exhibits strong calibration across most of the prediction range, with slight underestimation of risk in mid-range (0.25–0.55) predicted probabilities. Calibration improves at both extremes, particularly for higher-risk members.

The model achieved an AUC of 0.923 and, at the model-specific top 5% risk threshold, a precision of 0.080, recall of 0.700, specificity of 0.954, and accuracy of 0.952, indicating strong discriminatory power for a rare outcome and suitability for flagging individuals with elevated social risk.

Subgroup performance remained consistent across most demographics. Recall exceeded 70% for English speakers, and both male and female children. Higher recall values were observed among American Indian or Alaska Native (91%), White (82%), and multiracial (80%) subgroups. In contrast, recall was lower among Spanish-speaking children (33%) and Asian children (51%). Specificity remained high ($\geq 93\%$) for most groups.

F. Tiering Analysis Results

After the completion of the modeling phase, members were assigned into risk tiers based on the framework decided upon by the Working Group. As mentioned in the [Tiering Analysis Evaluation Criteria](#) section above, the framework prioritized the following 5 principles, each with a set of quantitative metric(s) that the Work Group translated into a qualitative scorecard to weigh the 3 options from a policy and feasibility lens:

- » Technical Performance
- » Balance of Risk Types
- » Equity
- » Potential Benefits
- » Potential Costs and Impacts

A review of potentially comparable studies was conducted to determine an appropriate benchmark for AUC performance, and eight studies were summarized and used as part of the Working Group’s tiering decision-making process.

Table 8. Summary of Studies Used in Tiering Decision-Making Process

Study	Outcome	Population	Strata	Domain/ Subdomain	AUC
Yu et al. (2022)	Avoidable ED Visits (Psychiatric factors)	NHAMCS incl. 35% using Medicaid	peds; adults (mean age - 37.06)	Risk of Adverse Events / Behavioral Health	0.677 (Medicaid - only AUC)

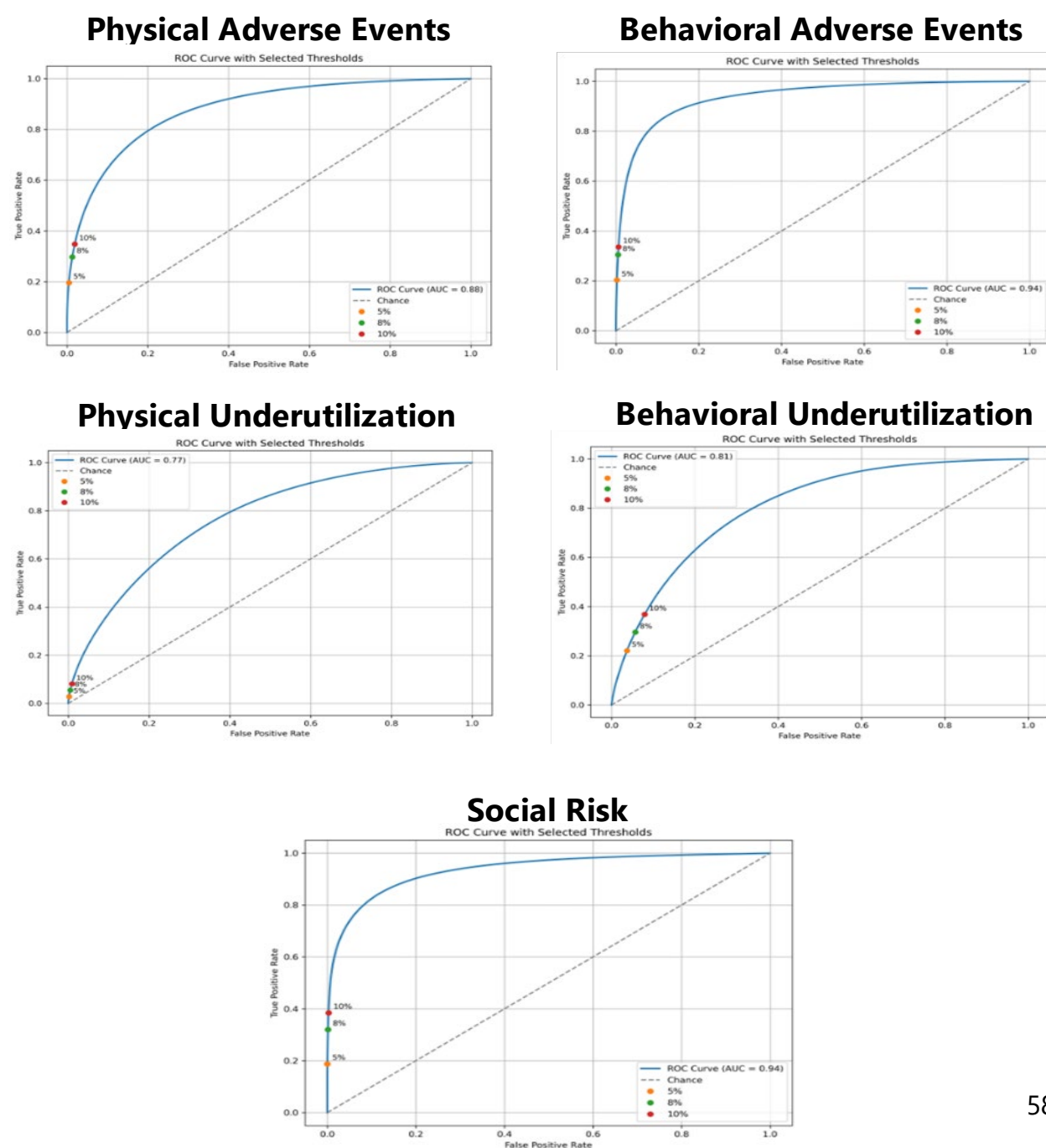
Study	Outcome	Population	Strata	Domain/ Subdomain	AUC
Rahman et al. (2022)	Intentional Self-Harm Prediction	NY Medicaid mental health clients	peds; adults (ages 10 to 64)	Risk of Adverse Events / Behavioral Health	0.86
Lo-Ciganic (2021)	Predicting risk of opioid overdose among Medicaid beneficiaries	Medicaid beneficiaries (n = 237,259) in Allegheny County, Pennsylvania	peds; adults (mean age - 38±18)	Risk of Adverse Events / Behavioral Health	0.885
Lo-Ciganic (2022)	Predict opioid overdose in Medicaid beneficiaries in two US states (Pennsylvania and Arizona)	Pennsylvania Medicaid beneficiaries	adults (18-64)	Risk of Adverse Events / Behavioral Health	0.828 (external valid.), 0.841 (internal valid.)
Gao et al (2021)	Predicting Opioid Use Disorder and Associated Risk Factors in a Medicaid Managed Care Population	Medicaid enrollment, medical, pharmacy, and care management administrative data from January 1, 2016, to December 31, 2018. Included data from the District of Columbia, Florida, Louisiana, Michigan, Pennsylvania, and South Carolina	adults (18+)	Risk of Adverse Events / Behavioral Health	0.914
Patel et al. (2024)	All-Cause Acute Care Visits	10 million Medicaid patients from 26 states and Washington DC	peds; adults (64.7% under 18)	Risk of Adverse Events / Physical Health	0.807 - 0.829
Pourat et al. (2023)	Homelessness identification via Medicaid admin data	Medicaid (California WPC program)		Social Risk / Adverse Events	0.95
Holcomb et al. (2022)	Prediction of health-related social needs incl. housing	Medicaid (85.4%) and Medicare (22.8)	peds; adults (mean age - 35.5)	Social Risk / Adverse Events	0.68

1) Adult Tiering: Technical Performance

All subdomain models were assessed for technical performance, equity, and appropriateness of predictive outcomes in the adult population. AUC values ranged from 0.879-0.938, with the highest discrimination statistic approaching 0.938.

Domain	Adverse Events		Underutilization		Social Risk
Subdomain	Physical Health	Behavioral Health	Physical Health	Behavioral Health	Adverse Social Events
AUC	0.879	0.938	0.769	0.805	0.936

None of the AUC curves demonstrated atypical patterns, with all five models exhibiting patterns consistent with standard model behavior:

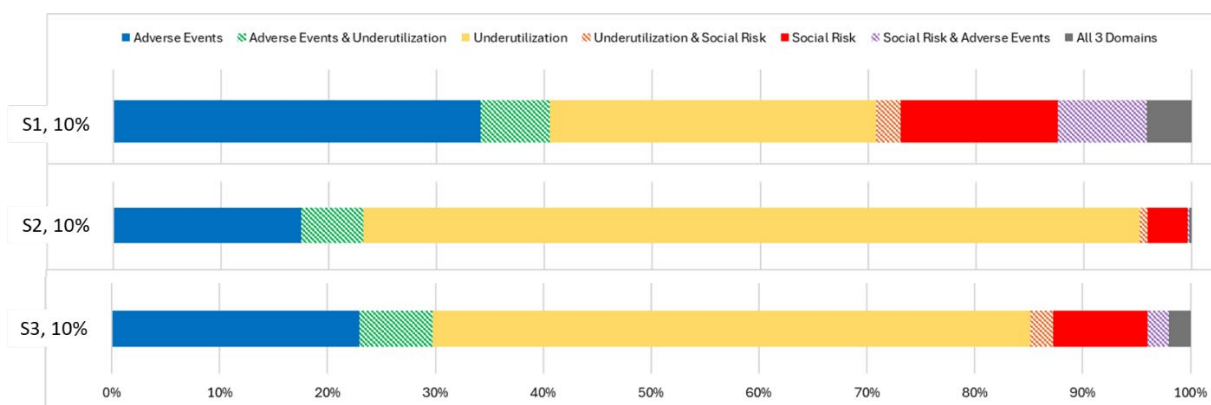


Finally, the Workgroup reviewed NNT and Recall for each of the 3 options. Option 2, which was ultimately selected, had overall recall of 0.23 and NNT of 1.75. In other words, approximately 1 in 5 of all adult state members who went on to have at least one RSST adverse outcome were flagged as high risk, and over half of members flagged as high risk went on to have at least one adverse RSST outcome. More details on these numbers can be found in the Equity section below.

Tiering Option (Adults)	Recall	NNT
Option 1	0.216	2.012
Option 2	0.226	1.751
Option 3	0.233	1.799

2) Adult Tiering: Balance of Risk Types

In determining the relative contribution of all risk types to the overall high-risk population, we calculated member counts by which domain (or combination of domains) drove the high-risk designation under each option. As expected, Option 1 was found to have the most even balance across all five models. Notably, Options 2 and 3 indexed heavily on the Underutilization models in determining who was High Risk. This was deemed appropriate given the extent to which many population subgroups have barriers to accessing care, and the priorities of DHCS leadership.



3) Adult Tiering: Equity Assessment

The goal of RSST is to ensure that the algorithm performs consistently across all member subgroups, without systematically under- or over-identifying risk for any particular population. DHCS prospectively considered equity by evaluating model

performance stratified by demographic factors such as race, ethnicity, gender, and new enrollment.

In practice, this was implemented by evaluating recall (sensitivity) for each subpopulation, with a target of ensuring that subgroup-specific recall was no worse than 80% of performance of overall statewide average recall (0.23), an approach and benchmark used in this field.



Representative but not actual tiering data on adult Medi-Cal members

In Option 1, primary Spanish-speakers and Asian subgroups fell below the threshold, with recall values of 0.13 and 0.11 respectively, compared to 0.22 overall. Similarly in Option 3, Spanish-speakers and Asian subgroups fell below the threshold, with recall values of 0.16 and 0.15 respectively, compared to 0.23 overall. In Option 2, however, almost all subgroups fell within 80% relative threshold, and this factored into DHCS decision to approve Option 2. New Medicaid Enrollees fell below the threshold in all three options, very likely due to lack of data history available on those members. The Work Group found this to be an acceptable limitation of the predictive models.

Note: Underutilization drives risk disproportionately among Spanish-speaking and Asian Members. Prioritization of equitable predictive performance resulted in relatively more people flagged as high-risk due to underutilization. Because many existing RSS methodologies largely focus on utilization and prevention of costly adverse events, DHCS's RSST approach flags people who may have been missed by existing approaches.

4) Adult Tiering Potential Benefit Assessment

Potential benefits were assessed via an estimation of net new assessments at a historic time point (January 2023), meaning the number of members who would receive a net new assessment (RSST results) who would not have received one otherwise per the current DHCS policy. We also evaluated the count of net new needs assessments on members who went on to have an event. Option 2 would have brought in the most net-new members for Needs Assessments (n=629,247) over and above current policy in

2023. Underutilization was the major driver of net new needs assessments: the vast majority (87-94%) of net-new members flagged here were "true positives" – in other words, these members went on to experience one of the RSST outcomes (adverse events, underuse, social) in the 12 months following the prediction (CY2023). This illustrates the performance of these models and factored into DHCS decisioning that RSST flags members in alignment with the goal to prioritize addressing Underutilization in the population.

5) Adult Tiering Potential Costs and Impacts

As part of the tiering analysis, the RSST Working Group considered the potential downstream impact of different tiering options on managed care plans. This included estimating how many members might qualify for additional assessments or transitional care services, and how those numbers—and associated costs—might vary across MCPs. While no new service requirements will be implemented at RSST V1 launch, this analysis helped inform understanding of how different design choices might translate into operational and resource implications across the state. These considerations supported decision-making but did not serve as the primary driver in option selection. Prior to policy being updated and enforced, additional analysis looking at benefits and feasibility including costs will be considered.

6) Adult Tiering Summary

The Working Group's evaluation of the different options resulted in agreement that Option 2 was the most appropriate path forward for the V1 launch in the adult population. It satisfied all evaluation criteria laid out by the group, with high technical performance and contribution across domains. It was the option with the strongest equity result, with only new enrollees falling below the 80% benchmark relative to the state average recall. The scorecard below was used by the Work Group to compare the relative performance of each of the 3 options on the 5 key evaluation principles we set.

		Worst ← → Best				
Principle	Description	1	2	3	4	5
<div><div><div>Option 1</div><div>Equal Risk Threshold</div></div><div><div>Option 2</div><div>Balanced Sensitivity Across Domains</div></div><div><div>Option 3</div><div>Maximize Total Sensitivity Across Domains</div></div></div>	Technical Performance					<div><div></div><div></div><div></div></div>
	Balance of Risk Types			<div><div></div><div></div></div>		<div><div></div></div>
	Equity		<div><div></div><div></div></div>		<div><div></div></div>	
	Potential Benefits			<div><div></div></div>	<div><div></div><div></div></div>	
	Potential Costs and Impacts				<div><div></div><div></div><div></div></div>	

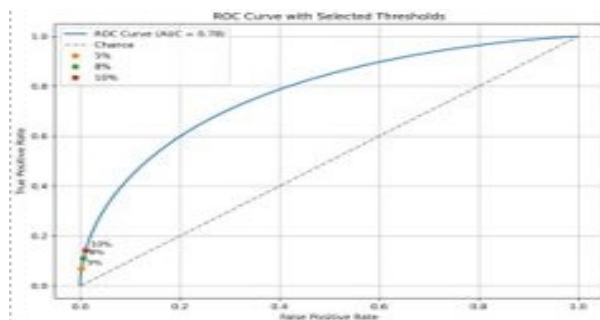
7) Pediatric Tiering: Technical Performance

All subdomain models were assessed for technical performance in the pediatric population. The models achieved sufficient levels of performance and effectively predicted member risk scores. AUC values ranged from 0.75-0.92, with the highest discrimination statistic approaching 0.92 in the Social Risk model. These results compared favorably to models predicting similar adverse events in similar populations (see above comparative studies).

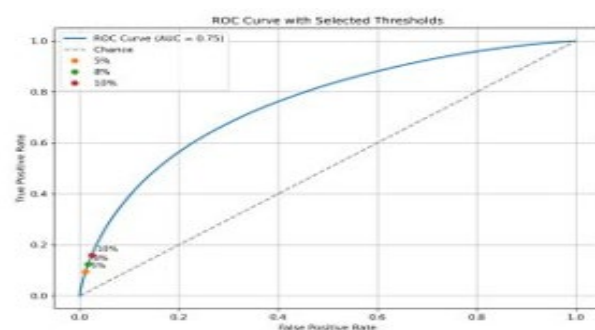
Domain	Adverse Events		Underutilization		Social Risk
Subdomain	Physical Health	Behavioral Health	Physical Health	Behavioral Health	Adverse Social Events
AUC	0.77	0.75	0.76	0.78	0.92

None of the AUC curves demonstrated atypical patterns, with all five models exhibiting patterns consistent with standard model behavior:

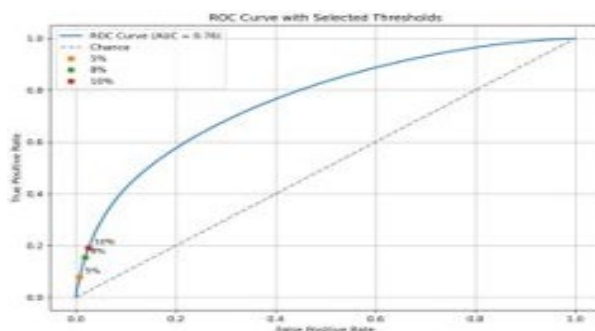
Physical Adverse Events



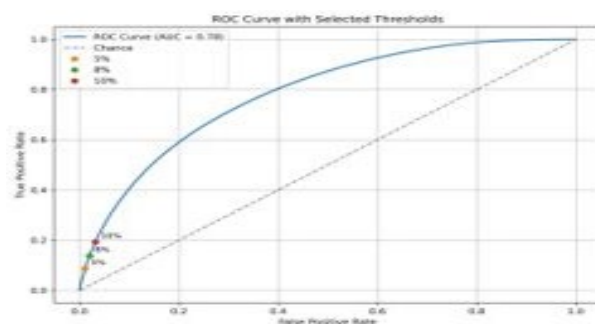
Behavioral Adverse Events



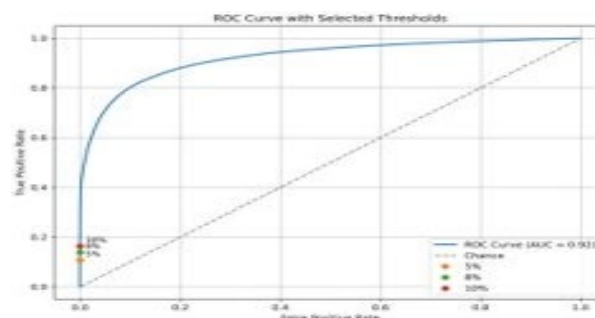
Physical Underutilization



Behavioral Underutilization



Social Risk

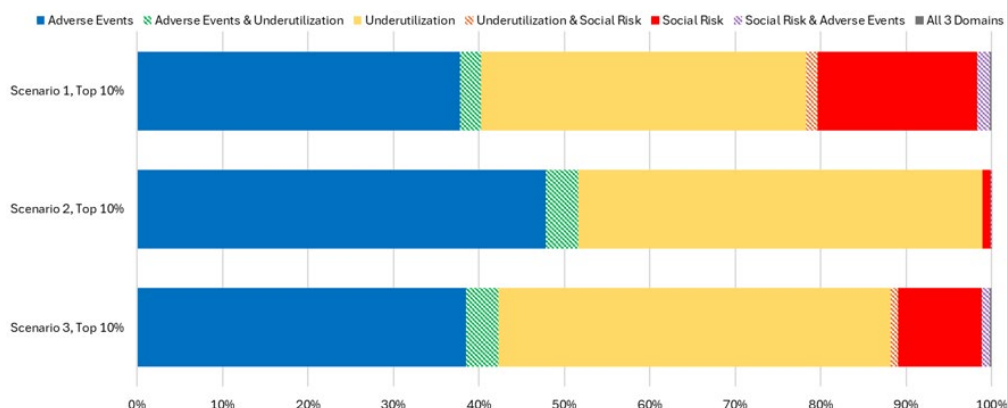


Finally, the Workgroup reviewed NNT and Recall for the overall RSST tier for the Pediatric strata, with a state average recall of 0.24 and a NNT of 1.97 (Option 2). In other words, approximately 1 in 4 of all pediatric state members who went on to have at least one RSST adverse outcome were flagged as high risk, and over half of pediatric members flagged as high risk went on to have at least one adverse RSST outcome.

Tiering Option (Pediatrics)	Recall	NNT
Option 1	0.215	2.207
Option 2	0.241	1.966
Option 3	0.245	2.001

8) Pediatric Tiering: Balance of Risk Types

The Pediatric models showed less balance of risk types, with Options 2 and 3 generating many members for Underutilization Domain due high level of underutilization in the Medi-Cal population, and an appropriately small number of members due to Social Risk. Option 1 was designed to balance the proportion of members flagged as high-risk at the subdomain level, which as expected resulted in greater balance than other scenarios at the domain level. However, the Workgroup determined that it over-emphasized Social Risk, which is based on only one rare Pediatric outcome (Housing Instability).



9) Pediatric Tiering: Equity Assessment

When comparing all subgroups to the statewide average, all three options nearly met the 80% relative recall threshold, with a few exceptions.

In Options 1 and 3, the American Indian or Alaska Native subgroup had higher-than-average recall (0.39 vs. 0.22 for Option 1 and 0.39 vs. 0.25 for Option 3), exceeding the

120% threshold. Likewise, in Option 2, the American Indian or Alaska Native subgroup had higher-than-average recall (0.30 vs. 0.24), exceeding the 120% threshold. The Workgroup agreed this positive outlier was acceptable given the small subgroup size. As in the adult models, New Medicaid Enrollees had lower than average recall (0.17 in Option 1 and 0.15 in Option 3), however this was not the case for Option 2.

Option 2 also showed lower recall for males (0.19 vs. 0.24), driven by the inclusion of chlamydia screening underuse among 16–17-year-old females in the Physical Health Underutilization composite. This outcome applied only to females and thus disproportionately improved recall for females in that model. A review of all other model outcomes found this to be the only such case. DHCS determined the result reflected real underutilization patterns and that sex-based differences in model performance were justified by the underlying data.

10) Pediatric Tiering: Potential Benefit Assessment

Potential benefits were assessed via an estimation of net new assessments at a historic time point (January 2023), meaning the number of members who would receive a net new assessment (RSST results) who would not have received one otherwise per the current DHCS policy. We also evaluated the count of net new needs assessments on members who went on to have an event. Option 2, which was ultimately selected, resulted in 351,177 net-new members being flagged, though fewer than Option 3 (n=364,0322). The majority (73–75%) of these net-new members flagged were “true positives” – in other words, they ended up having one of the RSST outcomes (adverse events, underuse, social) in the 12 months following the prediction (CY2023). While slightly lower than the adult model performance, this demonstrates meaningful predictive value in identifying pediatric members who may otherwise be missed. TCS impacts were excluded from this analysis since those are not net-new members flagged for care, even if there are net-new costs. Note that prior to policy being updated and enforced, additional analysis looking at benefits and feasibility including costs will be considered.

11) Pediatric Tiering: Potential Costs and Impacts

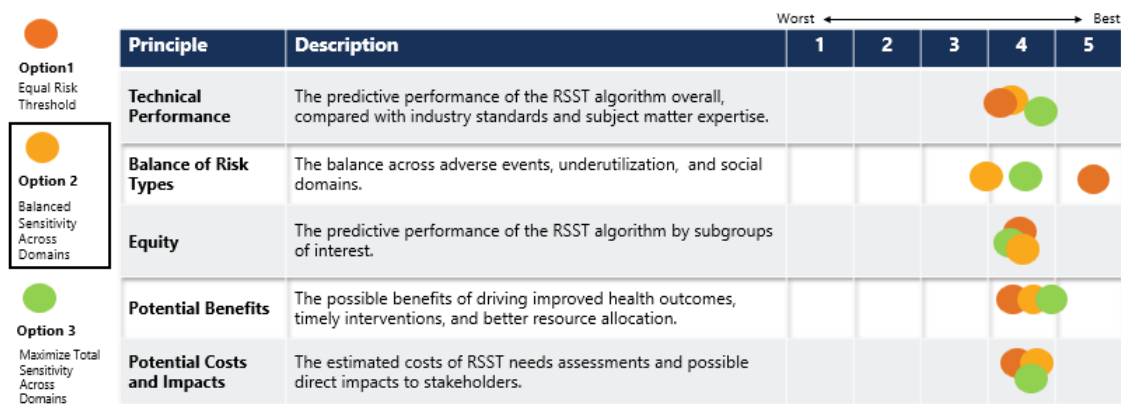
See [Adults section](#) above.

12) Pediatric Tiering Summary

The Working Group’s evaluation of the different options resulted in agreement that Option 2 was the most appropriate path forward for the V1 launch in the pediatric population. Both Option 2 and Option 3 performed well across the evaluation criteria,

and the decision between them was closer than in the adult analysis. Option 2 demonstrated recall and NNT values of 0.24 and 1.97, respectively similar to or higher than the other Options. It also performed best on the equity analysis, with all subgroups meeting or exceeding the 80% benchmark relative to the statewide average—aside from a single sex-specific underperformance in males, which was reviewed and determined to reflect real-world data patterns.

Given the importance of transparency and explainability to stakeholders, it was important to DHCS to select the same tiering option for both adult and pediatric populations, provided that no major equity or performance tradeoffs were present. Option 2 satisfied that condition, offering the clearest path for consistent statewide implementation in Version 1.



G. Validation Results

After model tuning, training, and tiering were approved by DHCS, several steps were taken to productionize the RSST predictive models (n=10) for routine monthly inference generation. The models were originally trained in the Training environment using historical data stored in an AWS Redshift SQL database. To support production go-live (target: July 2025), the models were migrated to run in the Production environment, with all data transformation and input files generated through the AWS Data Lake using Spark SQL.

Because this transition involved both code refactoring and changes to underlying data environments (including differences in available data by year), a structured series of validation exercises was conducted to ensure consistency, correctness, and readiness:

- » Tests 1–3 evaluated model alignment at the person-level across the full ~15 million-member population, verifying consistency in high-risk designations and risk scores.

- » Test 4 was conducted at the subdomain level, assessing broader differences in high-risk prevalence between historical and current populations.

Test 1: Input Code Refactor Validation – Redshift SQL to Spark SQL

Objective: Confirm that the model input logic was correctly refactored from Redshift SQL to Spark SQL when moving from Training to Production pipeline.

Approach:

- » Model inputs for IndexDate January 2023 were generated using Spark SQL in the AWS Data Lake (Training environment).
- » Outputs were compared at the PersonID level to the original model training inferences.
- » High-risk designations were used as the key comparison point across all 10 RSST models.

Results Summary:

Strata	Model Name	Percent Agreement
Adult	physical_adverse_events	99.2%
Adult	behavioral_adverse_events	99.8%
Adult	physical_underutilization	94.2%
Adult	behavioral_underutilization	92.6%
Adult	social_adverse_events	99.9%
Peds	physical_adverse_events	98.9%
Peds	behavioral_adverse_events	98.5%
Peds	physical_underutilization	95.7%
Peds	behavioral_underutilization	92.9%
Peds	social_adverse_events	100.0%

Findings:

- » Agreement between Redshift and Spark-generated inputs was high across models, confirming successful logic migration.
- » The behavioral_underutilization model showed the largest variation (92.6%), driven by changes in the bh_claims_12_months predictor, which relies on Short Doyle claims.

- » The discrepancy is expected: the Spark pipeline does not apply a received_date limit as in training. Additionally, Short Doyle claims are known to have long runout delays, which were communicated after model training was completed.

Test 2: Environment Shift Validation – Training to Production

Objective: Assess whether differences in model inference outputs arise when shifting from the Training to Production environment, using the same IndexDate and Production inference code.

Approach:

- » Inferences were generated using the Production codebase in both the Training and Production environments for IndexDate January 2023.
- » PersonIDs were converted back to CIN, and comparisons were made only for CINs present in both environments (11.56M adults, 5.03M pediatrics).

Results Summary:

Strata	Model Name	Percent Agreement
Adult	physical_adverse_events	99.3%
Adult	behavioral_adverse_events	99.8%
Adult	physical_underutilization	94.1%
Adult	behavioral_underutilization	87.6%
Adult	social_adverse_events	99.6%
Peds	physical_adverse_events	99.1%
Peds	behavioral_adverse_events	97.8%
Peds	physical_underutilization	97.0%
Peds	behavioral_underutilization	95.3%
Peds	social_adverse_events	99.9%

Findings:

- » Agreement across environments was generally high, particularly for adverse event models.
- » The behavioral_underutilization model again showed the lowest alignment, likely due to runout differences in Short Doyle claims.

- » Key known difference: Training includes claims through Dec 2023, while Production includes claims through Dec 2024.
- » The differences are consistent with expectations and considered acceptable.

Test 3: Inference Logic Validation – Training to Production

Objective: Confirm that the Production inference code produces consistent model outputs compared to the original training process when run on identical input data.

Approach:

- » Production inference code was run in the Training environment using January 2023 inputs—the same data used for model training.
- » Continuous risk scores were compared for all 10 models.

Findings:

- » Risk score outputs were effectively identical across all models.
- » Minor rounding differences (<0.01%) were observed but are not material.
- » This confirms the Production inference logic accurately reproduces the model training logic.

Test 4: Data Timeframe Shift – From Historical to Most Recent Data

Objective: Assess whether RSST model outputs differ when applied to the latest Medical population compared to the population used during development and determine whether any action was needed before go-live.

Approach:

- » Compared the percent of members flagged as high-risk in January 2023 (based on 2022 data used for Tiering Analysis) to December 2024 (the go-live month using 2024 data).
- » Examined trends at the subdomain, domain, and overall RSST tier level for both Adult and Pediatric strata.
- » Analysis was conducted using the fixed thresholds derived from model development, applying them to the updated 2024 population.

Findings:

- » Across most subdomains, percent of members flagged high-risk remained stable, with variation generally under 0.5%.
- » The exception was the Underutilization domain, where the proportion of members flagged increased by 2–3 percentage points, driving an increase in the overall RSSTTier prevalence from 10.0% to 12.1% in Adults and 10.1% to 11.5% in Pediatrics.

- » This increase was likely attributable to a combination of known limitations in behavioral health claims runout (e.g., Short Doyle data lag), and real population-level shifts in access patterns, potentially related to post-COVID care trends.

Decision and Action Taken:

To address this, the team performed a one-time rebasing of the subdomain-level thresholds. Rather than retrain the models, which was deemed unnecessary, the original top-K subdomain percentile thresholds from 2023 were reapplied to the December 2024 population. This realignment restored the RSSTTier high-risk prevalence to approximately 10% at go-live, consistent with DHCS's original guidance.

Decision Context:

In discussions with DHCS and RSST leads, consensus emerged that adjusting thresholds prior to go-live was preferable to allow RSST to launch close to the 10% target. Re-training was considered unnecessary, and this approach was viewed as a straightforward mitigation strategy that maintained alignment with the original tiering analysis. The team acknowledged that the prevalence of high-risk designations will still vary in future months and recommended continued monitoring (see next section).

Implications and Future Considerations:

- » The adjusted high risk threshold values will remain static for RSST Version 1, meaning future monthly prevalence may drift depending on population changes.
- » However, starting from a rebased 10% high-risk population ensures alignment with DHCS policy expectations at launch.
- » A broader re-tiering analysis is expected in 2025–2026 to revisit thresholds, cost impact, and equity results ahead of any formal policy mandate for MCPs to act on high-risk flags.
- » The table below summarizes the pre- and post-rebaselining results across strata and domains:

	Adults			Peds		
	Jan 2023 Tiering	Dec 2024 Prod	Diff	Jan 2023 Tiering	Dec 2024 Prod	Diff
RSSTTier	10.0%	12.1%	2.1%	10.1%	11.5%	1.4%
AdverseEventDomainTier	2.4%	2.1%	-0.3%	5.2%	4.8%	-0.4%
AdversePhysicalSubdomainTier	2.1%	2.0%	-0.1%	2.2%	2.0%	-0.2%
AdverseBehavioralSubdomainTier	0.5%	0.2%	-0.3%	3.5%	3.2%	-0.3%
UnderuseDomainTier	7.9%	10.4%	2.5%	5.2%	7.0%	1.8%
UnderusePhysicalSubdomainTier	6.7%	9.4%	2.7%	4.2%	6.1%	1.9%
UnderuseBehavioralSubdomainTier	1.6%	1.5%	-0.1%	1.0%	1.0%	0.0%
SocialRiskDomainTier	0.5%	0.3%	-0.2%	0.1%	0.1%	0.0%
SocialAdverseEventsSubdomainTier	0.5%	0.3%	-0.2%	0.1%	0.1%	0.0%

Conclusion

The four validation tests confirm that the RSST models approved by DHCS have been accurately migrated to the production environment. Tests 1–3 demonstrated strong consistency in logic, scoring, and data handling across environments and refactored code, with only minor, explainable differences. Test 4 identified a modest increase in high-risk prevalence—primarily in underutilization models—consistent with known data limitations and shifts in population-level care patterns.

Overall, the productionized RSST models remain aligned with those originally approved, and the system is ready for go-live. Outputs can be used with confidence, supported by ongoing monitoring and maintenance.

Looking ahead, future iterations of RSST models are expected to run within the same production environment and data pipelines, reducing the need for cross-environment validation. This streamlining will simplify future retraining and re-validation efforts while maintaining confidence in model continuity.

VII. MAINTENANCE AND MONITORING

The RSST Algorithm is operational in the production environment and supports a recurring monthly process for generating and distributing risk tier outputs across the Medi-Cal population. Ongoing activities include monthly delivery of member-level tier files, monitoring via a dashboard available to DHCS, and structured maintenance of the algorithm and its underlying models.

A. Monthly RSST Tier File Generation

Each month, a member-level RSST tier file is generated and delivered through secure channels for integration into the PHM Service. The tier file includes consistent, structured information to support use in MCP workflows, the longitudinal member record, and care management tools.

Each record contains:

- » Member identifiers and demographic fields
- » Stratification indicator (Adult or Pediatric)
- » Risk tier variables representing:
 - Five individual subdomain model outputs

- Aggregated domain-level tiers for Adverse Events, Underutilization, and Social Risk
- A single overall RSST tier value
- » Two binary indicators for pregnancy and postpartum status (used as flags only; no birthing model tiers are included at this stage)

Tier values are reported as integers (1 = Low, 2 = Medium, 3 = High). The file format and contents are consistent across months to support comparability and integration.

B. Dashboard-Based Monitoring (DHCS Access Only)

A secure monitoring dashboard is available to the RSST leadership team at DHCS. The dashboard allows for ongoing tracking of tier distribution patterns, system behavior, and potential signals of change in model performance.

Key features of the dashboard include:

- » Statewide and county-level tier distributions
- » MCP-level summaries
- » Subgroup breakouts (e.g., race/ethnicity, language, age group, dual eligibility)
- » Month-over-month trend tracking at the domain and subdomain levels
- » The dashboard also supports subgroup analysis using the same demographic variables examined in the Tiering Analysis, enabling DHCS to monitor monthly results for key populations and compare subgroup trends to statewide averages.

The dashboard is updated monthly and is designed to support oversight and decision-making. Unusual shifts in tiering patterns or subgroup variation may trigger further review and potential model updates.

C. Ongoing Maintenance of the RSST Algorithm

The RSST Algorithm will be maintained through a structured review and release process. Ongoing activities include:

- » Periodic retraining of underlying models using updated data
- » Adjustments to model logic, input features, or risk tiering methodology
- » Integration of additional models, including future iterations focused on birthing populations
- » Monthly review of outputs using the monitoring dashboard

- » Version tracking and documentation of changes over time

These processes ensure that the RSST Algorithm remains aligned with Medi-Cal program goals and continues to reflect changes in data, policy, and population characteristics.

VIII. ASSUMPTIONS AND LIMITATIONS

The RSST Algorithm operates within known constraints related to data availability, model design choices, and system infrastructure. These limitations should be considered when interpreting tier outputs and planning future development.

- » **Data availability:** The models are trained using Medicaid administrative claims and eligibility data. These sources provide broad statewide coverage but do not include clinical records (e.g., EHR data) or many non-health social factors. Social risk features are currently limited in scope. EHR data is not available and is not expected to be included in future RSST versions. Medicare claims and encounter data is not included in any of the models.
- » **Excluded populations:** Members with only partial Medi-Cal coverage (e.g., GHPP, Family PACT, CCS) are excluded from RSST tiering due to data limitations, and dual-eligible members (enrolled in both Medicaid and Medicare) are excluded from receiving Underutilization tier values as Medicare claims are not accessible and that model is particularly sensitive to the lack of Medicare data. These exclusions are detailed in the Population Definition section.
- » **Claims lag:** RSST models were intentionally designed to account for the natural delays in claim submission and processing. During model training, inputs were constructed using a “what was known when” approach: only data with a processing date prior to the index date was included as a predictor. If a claim had not yet been processed at the prediction time, it was excluded—even if the service occurred earlier. This mirrors real-world data availability conditions and helps ensure that model performance reflects how the algorithm would function in practice.
- » **Data transfer lag within PHM Service:** In addition to claims runout, data must pass through several systems within the PHM Service before it is available for tiering. These delays were not quantified at the time of model development and were not incorporated into Version 1. Future retraining is expected to account for this lag explicitly.

- » **Fixed thresholds:** High-risk tier thresholds were set using January 2023 predicted scores—the most recent month with complete follow-up data. After a one-time correction to a more current population, these thresholds will remain fixed for consistency, but the share of members flagged as high risk may vary over time as populations shift. This will require active monitoring, a cadence for re-training when population shifts occur, and/or a change to tiering methodology.
- » **Model generalizability over time:** Models are based on historical data and may become less predictive as care patterns, access, or population characteristics change. Ongoing monitoring and periodic retraining will be necessary to maintain performance.
- » **Birthing model exclusion:** Models for birthing populations were trained alongside adult and pediatric models but did not meet performance standards for inclusion in Version 1. These models will be revised, retrained, and incorporated into a future RSST release.
- » **Assumptions around service capacity:** The 10% high-risk flagging rate was informed by estimates of what managed care plans could reasonably manage under future policy scenarios. These assumptions may not align with actual MCP capacity and will be revisited over time, at which point additional analytics will be conducted to support a new high-risk rate.
- » **Medium-risk tiering:** Medium-risk thresholds were set using a uniform 50% recall cutoff across subdomains for simplicity. These thresholds do not currently influence service requirements and were not optimized for operational outcomes. They are to be considered pre-decisional placeholders while the team refines the analytical approach to medium threshold setting.
- » **Evaluation scope:** The Tiering Analysis included key performance and equity metrics, but not all possible subgroup or policy simulations. The analysis is meant to inform policy choices, not prescribe them.
- » **Rare subgroups:** For smaller or underrepresented populations—such as certain race/ethnicity categories or low-volume counties—model performance and equity statistics may be less stable due to limited sample size.

IX. APPENDIX

A. Glossary of Key Policy Terminology Used

Admission, discharge, and transfer (ADT) feed is a standardized, real-time data feed sourced from a health facility, such as a hospital, that includes members' demographic and healthcare encounter data at time of admission, discharge, and/or transfer from the facility.

Assessment is a process or set of questions for defining the nature of a risk factor or problem, determining the overall needs or health goals and priorities, and developing specific treatment recommendations for addressing the risk factor or problem. Health assessments can vary in length and scope.

Basic Population Health Management (BPHM) is an approach to care that ensures that needed programs and services are made available to each member, regardless of their risk tier, at the right time and in the right setting. BPHM includes federal requirements for care coordination (as defined in 42 C.F.R. § 438.208).

Care manager is an individual identified as a single point of contact responsible for the provision of care management services for a member.

Care Management Plan (CMP) is a written plan that is developed with input from the member and/or their family member(s), guardian, authorized representative, caregiver, and/or other authorized support person(s), as appropriate, to assess 86 strengths, risks, needs, goals, and preferences, and make recommendations for service needs.

Complex Care Management (CCM) is an approach to care management that meets differing needs of high-and rising-risk members, including both longer-term chronic care coordination and interventions for episodic, temporary needs. Medi-Cal Managed care plans (MCPs) must provide CCM in accordance with all NCQA CCM requirements.

Early and Periodic Screening, Diagnostic, and Treatment (EPSDT) is a federal entitlement that states are required to provide to all children under age 21 enrolled in Medicaid. This includes any Medicaid-coverable service in any amount that is medically necessary, regardless of whether the service is covered in the state plan.³

³ EPSDT in Medicaid. Medicaid and CHIP Payment and Access Commission. DHCS specific requirements on EPSDT is outlined in APL 23-010.

Enhanced Care Management (ECM) is a whole-person, interdisciplinary approach to care that addresses the clinical and nonclinical needs of high-cost and/or high-need members who meet ECM Populations of Focus eligibility criteria through systematic coordination of services and comprehensive care management that is community-based, interdisciplinary, high-touch, and person-centered.

Health Information Form (HIF)/Member Evaluation Tool (MET) is a screening tool that is required to be completed within 90 days of MCP enrollment for new members. It fulfills the federal initial screening requirement.⁴

Health Risk Assessment (HRA) is an assessment required for Seniors and Persons with Disabilities. Effective January 1, 2023, HRA assessment requirements for Seniors and Persons with Disabilities are simplified, while specific member protections are kept in place.

Initial Health Appointment(s) previously called Initial Health Assessment, now refers to appointment(s) required to be completed within 120 days of MCP enrollment for new members and must include a history of the member's physical and behavioral health, an identification of risks, an assessment of need for preventive screens or services and health education, and the diagnosis and plan for treatment of any diseases.⁵

Long-Term Care (LTC) includes specialized rehabilitative services and care provided in a Skilled Nursing Facility, subacute facility, pediatric subacute facility, or Intermediate Care Facilities (ICFs).⁶

Long-Term Services & Supports (LTSS) includes services and supports designed to allow a member with functional limitations and/or chronic illnesses the ability to live or work in the setting of the Member's choice, which may include the Member's home, a worksite, a Provider-owned or controlled residential setting, a nursing facility, or other institutional setting. LTSS includes both LTC and HCBS and includes carved-in and carved-out services.⁷

Risk stratification and segmentation (RSS) is the process of separating member populations into different risk groups and/or meaningful subsets using information

⁴ 42 CFR 438.208(b)(3)-(4)

⁵ These required Initial Health Appointment(s) elements are specified in 22 C.C.R. § 53851(b)(1).

⁶ 2024 Re-Procurement. Exhibit A, Attachment I, Definitions and Acronyms

⁷ 2024 Re-Procurement. Exhibit A, Attachment I, Definitions and Acronyms

collected through population assessments and other data sources. RSS results in the categorization of members with care needs at all levels and intensities.

Risk tiering is the assigning of members to standard risk tiers (i.e., high, medium- rising, or low), with the goal of determining appropriate care management programs or specific services.

Population Health Management (PHM) is a whole-system, person-centered, population-health approach to ensuring equitable access to health care and social care that addresses member needs. It is based on data-driven risk stratification, analytics, identifying gaps in care, standardized assessment processes, and holistic care/case management interventions.

The Population Health Management (PHM) Service collects and links Medi-Cal beneficiary information from disparate sources and performs risk stratification and segmentation (RSS) and tiering functions, conducts analytics and reporting, identifies gaps in care, performs other population health functions, and allows for multiparty data access and use in accordance with state and federal law and policy.

DHCS Population Health Management Strategy (PHM) Deliverable is an annual deliverable that MCPs submit to DHCS to demonstrate that it is responding to identified community needs, to provide other updates on its PHM program as requested by DHCS, and to inform the DHCS quality assurance and Population Health Management program compliance and impact monitoring efforts.

Screening is a brief process or questionnaire for examining the possible presence of a particular risk factor or problem to determine whether a more in-depth assessment is needed in a specific area of concern.

Social drivers of health (SDOH) are the environments in which people are born, live, learn, work, play, worship, and age that affect a wide range of health functioning and quality-of-life outcomes and risk factors.

Transitional care services (TCS) are services provided to all members transferring from one institutional care setting or level of care to another institution or lower level of care (including home settings).

Wellness and prevention programs are programs that aim to prevent disease, disability, and other conditions; prolong life; promote physical and mental health and efficiency; and improve overall quality of life and well-being.

B. Glossary of Technical Terms Used

Area Under the Curve (AUC) A performance metric used to evaluate how well a model distinguishes between members who will and won't experience an outcome. AUC measures the trade-off between sensitivity (recall) and specificity (true negative rate) across thresholds. Higher values indicate better discrimination. AUC ranges from 0.5 (random) to 1.0 (perfect).

Buffer window A time gap added between the end of the predictor period and the start of the outcome observation window to avoid overlap between inputs and labels. This helps prevent label leakage and ensures that predictions are based only on past information.

Calibration A measure of how well predicted probabilities align with actual outcomes. A calibrated model assigns higher risk scores to members who are more likely to experience an outcome. This matters when risk scores are interpreted directly (e.g., used for prioritization or thresholding).

Dask A Python computing library used to process large-scale healthcare data during RSST model development. It enables distributed, memory-efficient operations across multiple processors or machines.

Distributed computing A computing setup where processing is split across multiple machines or processors. This approach was used during RSST model development to handle large Medi-Cal datasets efficiently.

Early stopping / pruning A technique used during model training or optimization to stop processing when performance no longer improves. In RSST, pruning was used in tuning to reduce unnecessary computation.

Feature A variable used by a model to predict outcomes. RSST features include demographics, claims-based indicators, encounter data, pharmacy use, and social service history.

Feature importance A metric that indicates how much each feature contributes to a model's predictions. Feature importance helps users interpret which factors the model relies on most.

Gradient boosting A machine learning method that builds a predictive model in stages, with each stage correcting errors from the previous one. RSST used gradient boosting frameworks such as XGBoost and LightGBM.

Hyperparameter tuning The process of selecting model settings—such as learning rate or tree depth—to improve performance. RSST used automated hyperparameter tuning with Optuna to find the best configurations.

Label leakage When a feature inadvertently contains information about the label (outcome), which can distort model training and performance. RSST used buffer windows and feature checks to prevent this.

LightGBM A gradient boosting model optimized for speed and scalability. LightGBM was one of the modeling methods used in RSST to predict future outcomes from complex Medi-Cal data.

Number Needed to Treat (NNT) A measure of how many members must be flagged as high risk to prevent one adverse event. A lower NNT means more efficient targeting. For example, an NNT of 5 means 5 members need to be assessed to prevent one event.

Optuna A software library used to automate model tuning by searching for the best hyperparameters. RSST used Optuna to efficiently improve performance during model training.

Outcomes The specific adverse events or service gaps the RSST models are trained to predict. Examples include hospitalization, overdose, or failure to refill a medication. Each outcome is binary and tied to a defined time window. Outcomes in the RSST make up the overall subdomain composite that models aim to predict.

Precision The proportion of members flagged as high risk who actually experience the predicted outcome. Higher precision means fewer false positives. Precision helps assess how efficiently the model targets people who benefit from intervention.

Predictors Also known as features. The variables used to generate risk scores. In RSST, predictors come from administrative data sources including claims, pharmacy, encounters, and social service interactions.

Probabilistic predictions Numeric predictions ranging from 0 to 1 that represent the estimated likelihood of an outcome occurring.

Recall Also known as sensitivity. The percentage of members who experienced an outcome and were correctly flagged as high risk. Higher recall means the model is better at capturing true cases. For example, a recall of 0.65 means 65% of true outcomes were identified.

Regularized Logistic Regression A simple, interpretable modeling method that includes penalty terms to reduce overfitting.

Risk score A numeric value representing a member's predicted probability of experiencing a composite of the outcomes of interest in the next 12 months. Risk scores range from 0 (lowest risk) to 1 (highest risk) and are used to assign members to tiers.

SHAP (SHapley Additive exPlanations) A method for interpreting model predictions by calculating how much each feature contributed to a given prediction. RSST used SHAP values to improve model transparency and explainability.

Specificity The percentage of members who did not experience the outcome and were correctly not flagged as high risk.

Supervised learning A machine learning approach where models are trained on input data paired with known outcomes.

Test set A dataset used to evaluate model performance after training and tuning are complete.

Validation set A portion of data used during model development to evaluate different model parameters and select the best learner for a subdomain model. Validation helps ensure that models generalize well to unseen data.

XGBoost A popular machine learning framework used for gradient boosting. It was one of the primary model types used in RSST to generate predictions for several subdomains.

C. Additional Model Detail

1) Tables

Table A.1. Adult Composite and Outcome Prevalence⁸

Outcome	Shorthand	Prevalence	Denominator	Prevalence Rate	Strata-Specific Prevalence Rate
Adults physical health adverse events	adult_adverse_ph_composite	703337	8875329	8%	-
All cause ip admit	o1	717488	8875329	8%	-
All cause ed visit	o2	532217	8875329	6%	-
Mortality	o5	89048	8875329	1%	-
New morbidity/ci	o9	81854	8875329	1%	-
Adults behavioral health adverse events	adult_adverse_bh_composite	98127	8875329	1%	-
Psych admit	o3	51918	8875329	1%	-
Psych ed	o4	109876	8875329	1%	-
Intentional self-harm	o6	11990	8875329	<1%	-
Drug overdose	o7	26038	8875329	<1%	-
Injection related adverse event	o8	31006	8875329	<1%	-
All cause ed visits (3+) and short-doyle (1+)	o83	78718	8875329	1%	-
Adults behavioral underutilization	adult_under_bh_composite	618277	861901	72%	7%
Pcp visits with bh/sud diagnosis	o12	207604	861901	24%	2%
Ambulatory mh/sud visits	o13	452473	861901	52%	5%
Antidep med mgmt	o19	141461	861901	16%	2%
Antipsych for schiz	o20	43112	861901	5%	<1%
Opioid for oud	o21	13294	861901	2%	<1%

⁸ Outcome prevalence is calculated at the composite and individual outcome levels using the full eligible population (denominator) and the January 2023 index date from the test set. Underutilization models applied additional inclusion criteria that excluded some members within their respective strata. As a result, denominators for underutilization models are smaller than those used in adverse event models. For underutilization models, prevalence rates for individual outcomes share the same denominator as the composite. This reflects inclusion criteria applied at the composite level rather than for each outcome individually.

Outcome	Shorthand	Prevalence	Denominator	Prevalence Rate	Strata-Specific Prevalence Rate
Adults physical health underutilization	adult_under_ph_composite	2713059	4290776	63%	31%
Pcp visit	o10	772829	4290776	18%	9%
No claims	o99	746422	4290776	17%	8%
Dental care	o11	345051	4290776	8%	4%
Chlamydia screen	o15	594083	4290776	14%	7%
Pharm for htn	o18_a	326312	4290776	8%	4%
Pharm for dm	o18_b	227529	4290776	5%	3%
Asthma med ratio	o41	123667	4290776	3%	1%
Adults social risk adverse events	adult_social_ae_composite	227505	8875329	3%	-
Housing insecurity	o22	227505	8875329	3%	-

Table A.2. Pediatrics Composite and Outcome Prevalence⁹

Outcome	Shorthand	Prevalence	Denominator	Prevalence Rate	Strata-Specific Prevalence Rate
Pediatrics physical health adverse events	peds_adverse_ph_composite	284,785	4,549,042	6%	-
All cause ip admit	o1	141,464	4,549,042	3%	-
All cause ed visit	o2	147,980	4,549,042	3%	-
Diagnosis of common chronic illness	o26	28,143	4,549,042	1%	-
Morbidity	o30	18,457	4,549,042	< 1%	-
Pediatrics behavioral health adverse events	peds_adverse_bh_composite	281,656	4,549,042	6%	-
New diagnosis of mental illness	o27	227,941	4,549,042	5%	-
New diagnosis of developmental delay	o28	13,294	4,549,042	< 1%	-
New diagnosis of SUD	o29	24,956	4,549,042	1%	-
Psych admit	o3	13,019	4,549,042	< 1%	-
Psych ed	o4	44,542	4,549,042	1%	-
Intentional self-harm	o6	4,537	4,549,042	< 1%	-
Drug overdose	o7	2,141	4,549,042	< 1%	-

⁹ See Table C.1 – Adult Composite and Outcome Prevalence footnote

Outcome	Shorthand	Prevalence	Denominator	Prevalence Rate	Strata-Specific Prevalence Rate
Pediatrics behavioral underutilization	peds_under_bh_composite	233,355	299,096	78%	5%
Pcp visits with bh/sud diagnosis	o12	59,690	299,096	20%	1%
Ambulatory mh/sud visits	o13	211,182	299,096	71%	5%
Metabolic Screenings	o39	7,574	299,096	3%	<1%
Follow-up after admission or ED visit for mental illness/SUD	o42	7,851	299,096	3%	<1%
ADHD follow-up care	o43	1,196	299,096	<1%	<1%
Pediatrics physical health underutilization	peds_under_ph_composite	914,777	1,559,019	59%	20%
Chlamydia screen	o15	226,628	1,559,019	15%	5%
Well child visits (first 30 months)	o31	126,884	1,559,019	8%	3%
Well child and adolescent visits	o32	352,677	1,559,019	23%	8%
Topical fluoride and/or Dental care	o33	151,667	1,559,019	10%	3%
Immunization status	o36	62,626	1,559,019	4%	1%
Asthma medication indicator	o41	74,408	1,559,019	5%	2%
Pediatrics social risk adverse events	peds_social_ae_composite	27,081	4,549,042	1%	-
Housing instability	o22	27,081	4,549,042	1%	-

Table A.3. Best Hyperparameters per Model¹⁰

Model	n_estimators	learning_rate	max_depth
Adults Behavioral Health Adverse Events	2833	0.0599	6
Adults Physical Health Adverse Events	1210	0.0359	12

¹⁰ All hyperparameters are for XGBoost. All XGBoost models were trained using a common set of default parameters unless otherwise specified. These included: objective='binary:logistic', eval_metric='auc', tree_method='hist', device='cuda', sampling_method='gradient_based', max_delta_step=1, and a fixed random_state=123 for reproducibility. Learning rate and maximum tree depth were tuned per model.

†These models were not the top performers in terms of AUC but were retained due to their close performance to the best model (LightGBM) and the scalability benefits of XGBoost, which supports multi-GPU training and inference at production scale.

Adults Social Risk Adverse Events [†]	1868	0.0359	10
Adults Behavioral Underutilization	552	0.0359	16
Adults Physical Health Underutilization	1671	0.0359	10
Peds Behavioral Health Adverse Events	1050	0.0359	12
Peds Physical Health Adverse Events	1671	0.0359	10
Peds Social Risk Adverse Events [†]	2090	0.0129	10
Peds Behavioral Underutilization	2090	0.0129	10
Peds Physical Health Underutilization	1909	0.0215	12

Table A.4A. Adults Behavioral Health Adverse Events ¹¹

Accuracy	Precision	Recall (Sensitivity)	Specificity	Person- months	Subgroup
0.9549	0.1733	0.6966	0.9582	108261397	Overall
0.9736	0.1429	0.5316	0.9768	9920799	new_medicare_enrollee_12_months
0.9481	0.1731	0.7117	0.9515	47948906	sex_m
0.9603	0.1735	0.6815	0.9634	60312491	sex_f
0.9392	0.1747	0.7131	0.9430	74711414	primary_language_eng
0.9891	0.1521	0.4856	0.9908	23857940	primary_language_spa
0.9672	0.1663	0.6460	0.9702	42500094	ethnicity_category_hispanicorlatino
0.9469	0.1760	0.7171	0.9503	65761303	ethnicity_category_nothispanicorlatino
0.9099	0.1702	0.7323	0.9141	446431	race_category_american_indian_or_alaska
0.9911	0.1527	0.5452	0.9922	11256276	race_category_asian
0.9213	0.1829	0.7348	0.9255	8113512	race_category_black_or_african_american
0.9789	0.1700	0.6209	0.9811	1378639	race_category_native_hawaiian_or_other_
0.9650	0.1692	0.6666	0.9679	47710767	race_category_other
0.9444	0.1658	0.6912	0.9482	6204729	race_category_two_or_more_races
0.9514	0.1748	0.6964	0.9549	10472088	race_category_unknown
0.9317	0.1762	0.7255	0.9356	22678955	race_category_white

Table A.4B. Adults Physical Health Adverse Events

Accuracy	Precision	Recall (Sensitivity)	Specificity	Person- months	Subgroup
0.9284	0.5947	0.3669	0.9779	108261397	Overall
0.9354	0.5651	0.3023	0.9826	9920799	new_medicare_enrollee_12_months

¹¹ Table A.4A – J -- Subgroup Performance at the Top 5% Threshold per Model

0.9313	0.5942	0.3792	0.978	47948906	sex_m
0.9261	0.5951	0.3578	0.9778	60312491	sex_f
0.9294	0.604	0.3972	0.9768	74711414	primary_language_eng
0.9295	0.5684	0.3086	0.9807	23857940	primary_language_spa
0.9356	0.5757	0.3219	0.982	42500094	ethnicity_category_hispanicorlatino
0.9237	0.6032	0.3902	0.9752	65761303	ethnicity_category_nothispanicorlatino
0.9027	0.5827	0.4249	0.9621	446431	race_category_american_indian_or_alaska
0.9432	0.5844	0.2754	0.9871	11256276	race_category_asian
0.9088	0.6064	0.4302	0.9664	8113512	race_category_black_or_african_american
0.9256	0.589	0.3511	0.9778	1378639	race_category_native_hawaiian_or_other_
0.9332	0.5815	0.3355	0.9808	47710767	race_category_other
0.9347	0.5684	0.3203	0.9815	6204729	race_category_two_or_more_races
0.9277	0.6105	0.3941	0.9768	10472088	race_category_unknown
0.9171	0.6087	0.4187	0.971	22678955	race_category_white

Table A.4C. Adults Social Risk Adverse Events

Accuracy	Precision	Recall (Sensitivity)	Specificity	Person- months	Subgroup
0.9594	0.3430	0.6889	0.9663	108261397	Overall
0.9728	0.3410	0.6113	0.9792	9920799	new_medicare_enrollee_12_months
0.9437	0.3438	0.7368	0.9509	47948906	sex_m
0.9719	0.3416	0.6170	0.9784	60312491	sex_f
0.9438	0.3422	0.7040	0.9523	74711414	primary_language_eng
0.9930	0.3528	0.4303	0.9959	23857940	primary_language_spa
0.9748	0.3259	0.6219	0.9802	42500094	ethnicity_category_hispanicorlatino
0.9494	0.3480	0.7099	0.9572	65761303	ethnicity_category_nothispanicorlatino
0.8744	0.4152	0.8077	0.8814	446431	race_category_american_indian_or_alaska
0.9948	0.3348	0.4635	0.9967	11256276	race_category_asian
0.9149	0.3116	0.6570	0.9277	8113512	race_category_black_or_african_american
0.9845	0.3803	0.5989	0.9889	1378639	race_category_native_hawaiian_or_other_
0.9739	0.2999	0.6146	0.9791	47710767	race_category_other
0.9464	0.3610	0.7163	0.9546	6204729	race_category_two_or_more_races
0.9588	0.3359	0.6580	0.9665	10472088	race_category_unknown
0.9312	0.3791	0.7562	0.9397	22678955	race_category_white

Table A.4D. Adults Behavioral Underutilization

Accuracy	Precision	Recall (Sensitivity)	Specificity	Person- months	Subgroup
0.3156	0.9612	0.0658	0.9928	11064665	Overall
0.2898	0.9570	0.0617	0.9915	746013	new_medicare_enrollee_12_months
0.2860	0.9673	0.0790	0.9909	4663338	sex_m
0.3371	0.9542	0.0551	0.9938	6401327	sex_f
0.3119	0.9586	0.0629	0.9926	9379020	primary_language_eng
0.3121	0.9735	0.0889	0.9926	1330548	primary_language_spa
0.2883	0.9611	0.0667	0.9914	3709292	ethnicity_category_hispanicorlatino
0.3293	0.9612	0.0653	0.9934	7355373	ethnicity_category_nothispanicorlatino
0.2564	0.9725	0.0866	0.9894	82388	race_category_american_indian_or_alaska
0.3435	0.9619	0.0471	0.9959	420040	race_category_asian
0.2785	0.9597	0.0787	0.9882	1077127	race_category_black_or_african_american
0.3513	0.9661	0.0739	0.9940	88309	race_category_native_hawaiian_or_other_
0.3010	0.9632	0.0687	0.9922	4314101	race_category_other
0.2939	0.9538	0.0601	0.9913	744561	race_category_two_or_more_races
0.3221	0.9587	0.0675	0.9923	1172935	race_category_unknown
0.3475	0.9608	0.0589	0.9946	3165204	race_category_white

Table A.4E. Adults Physical Health Underutilization

Accuracy	Precision	Recall (Sensitivity)	Specificity	Person- months	Subgroup
0.4206	0.9406	0.0754	0.9921	52216732	Overall
0.4345	0.9337	0.0442	0.9955	5862856	new_medicare_enrollee_12_months
0.3896	0.9393	0.0739	0.9909	20421256	sex_m
0.4405	0.9414	0.0764	0.9928	31795476	sex_f
0.4138	0.9422	0.0771	0.9919	37499791	primary_language_eng
0.4267	0.9350	0.0707	0.9922	11708380	primary_language_spa
0.4233	0.9390	0.0659	0.9932	22048178	ethnicity_category_hispanicorlatino
0.4185	0.9415	0.0821	0.9913	30168554	ethnicity_category_nothispanicorlatino
0.3978	0.9506	0.1158	0.9874	237088	race_category_american_indian_or_alaska
0.4202	0.9431	0.0627	0.9939	4069022	race_category_asian
0.4220	0.9410	0.1002	0.9889	4281011	race_category_black_or_african_american
0.4255	0.9440	0.0875	0.9913	589895	race_category_native_hawaiian_or_other_
0.4250	0.9394	0.0676	0.9931	24307995	race_category_other
0.4287	0.9388	0.0687	0.9930	3269442	race_category_two_or_more_races
0.4172	0.9401	0.0795	0.9914	5004166	race_category_unknown
0.4090	0.9419	0.0856	0.9905	10458113	race_category_white

Table A.4F. Peds Behavioral Health Adverse Events

Accuracy	Precision	Recall (Sensitivity)	Specificity	Person- months	Subgroup
0.9147	0.2735	0.2183	0.9612	58509970	Overall
0.9563	0.2553	0.1114	0.9879	5608602	new_medicare_enrollee_12_months
0.914	0.2904	0.2259	0.9617	29959997	sex_m
0.9153	0.2553	0.2097	0.9607	28549973	sex_f
0.9092	0.2744	0.2383	0.956	38558675	primary_language_eng
0.9213	0.2706	0.1753	0.9695	17764789	primary_language_spa
0.916	0.274	0.195	0.9649	32429813	ethnicity_category_hispanicorlatino
0.913	0.2729	0.2481	0.9566	26080157	ethnicity_category_nothispanicorlatino
0.8712	0.2751	0.2659	0.9309	179927	race_category_american_indian_or_alaska
0.9491	0.2733	0.1807	0.9804	3277630	race_category_asian
0.9033	0.2906	0.247	0.9537	3718155	race_category_black_or_african_american
0.9472	0.2819	0.1926	0.9792	518424	race_category_native_hawaiian_or_other_
0.9188	0.2757	0.196	0.9662	32894021	race_category_other
0.9064	0.2673	0.2173	0.9566	3808178	race_category_two_or_more_races
0.9231	0.2713	0.2074	0.9664	6136898	race_category_unknown
0.8851	0.2654	0.2949	0.9333	7976737	race_category_white

Table A.4G. Peds Physical Health Adverse Events

Accuracy	Precision	Recall (Sensitivity)	Specificity	Person- months	Subgroup
0.9187	0.3411	0.2613	0.9647	58509970	Overall
0.8583	0.3225	0.3573	0.915	5608602	new_medicare_enrollee_12_months
0.9173	0.3447	0.2608	0.9644	29959997	sex_m
0.9201	0.3372	0.2618	0.9649	28549973	sex_f
0.9154	0.3343	0.275	0.961	38558675	primary_language_eng
0.9223	0.3592	0.2374	0.9703	17764789	primary_language_spa
0.9145	0.3464	0.2625	0.9631	32429813	ethnicity_category_hispanicorlatino
0.9238	0.3337	0.2595	0.9666	26080157	ethnicity_category_nothispanicorlatino
0.9114	0.3216	0.2679	0.9586	179927	race_category_american_indian_or_alaska
0.9549	0.3362	0.1933	0.9849	3277630	race_category_asian
0.9194	0.339	0.2636	0.9646	3718155	race_category_black_or_african_american
0.9377	0.3476	0.2236	0.9769	518424	race_category_native_hawaiian_or_other_
0.9119	0.3471	0.2671	0.9614	32894021	race_category_other

0.9326	0.3348	0.2137	0.975	3808178	race_category_two_or_more_races
0.9153	0.3358	0.2608	0.9627	6136898	race_category_unknown
0.9263	0.3215	0.2734	0.9654	7976737	race_category_white

Table A.4H. Pediatrics Social Risk Adverse Events

Accuracy	Precision	Recall (Sensitivity)	Specificity	Person- months	Subgroup
0.9523	0.08	0.6998	0.9537	58509970	Overall
0.9324	0.0712	0.707	0.934	5608602	new_medicare_enrollee_12_months
0.9527	0.0807	0.7022	0.9542	29959997	sex_m
0.9518	0.0792	0.6974	0.9533	28549973	sex_f
0.9311	0.0801	0.7356	0.9327	38558675	primary_language_eng
0.9926	0.0762	0.3306	0.9936	17764789	primary_language_spa
0.9712	0.0754	0.6004	0.9726	32429813	ethnicity_category_hispanicorlatino
0.9288	0.0822	0.7558	0.9302	26080157	ethnicity_category_nothispanicorlatino
0.722	0.1114	0.9103	0.7146	179927	race_category_american_indian_or_alaska
0.9947	0.0914	0.5052	0.9952	3277630	race_category_asian
0.9348	0.0927	0.6654	0.9374	3718155	race_category_black_or_african_american
0.9678	0.0565	0.6185	0.9689	518424	race_category_native_hawaiian_or_other_
0.9795	0.0737	0.5278	0.9808	32894021	race_category_other
0.891	0.0823	0.8053	0.892	3808178	race_category_two_or_more_races
0.918	0.0691	0.7164	0.9197	6136898	race_category_unknown
0.8906	0.0844	0.8162	0.8915	7976737	race_category_white

Table A.4I. Pediatrics Behavioral Underutilization

Accuracy	Precision	Recall (Sensitivity)	Specificity	Person- months	Subgroup
0.2541	0.9692	0.0615	0.9925	3798653	Overall
0.3231	0.9770	0.1830	0.9798	149361	new_medicare_enrollee_12_months
0.2649	0.9664	0.0587	0.9928	1978536	sex_m
0.2423	0.9720	0.0644	0.9922	1820117	sex_f
0.2616	0.9693	0.0600	0.9931	2657721	primary_language_eng
0.2326	0.9724	0.0635	0.9919	1077401	primary_language_spa
0.2401	0.9714	0.0626	0.9922	2072126	ethnicity_category_hispanicorlatino
0.2709	0.9665	0.0601	0.9929	1726527	ethnicity_category_nothispanicorlatino
0.2485	0.9900	0.1167	0.9934	20030	race_category_american_indian_or_alaska
0.2766	0.9570	0.0684	0.9895	126575	race_category_asian

0.2643	0.9713	0.0668	0.9927	290791	race_category_black_or_african_american
0.2692	0.9609	0.0878	0.9859	19282	race_category_native_hawaiian_or_other_
0.2415	0.9711	0.0633	0.9921	2008188	race_category_other
0.2581	0.9690	0.0644	0.9922	281977	race_category_two_or_more_races
0.2693	0.9672	0.0629	0.9925	326484	race_category_unknown
0.2722	0.9643	0.0487	0.9942	725326	race_category_white

Table A.4J. Pediatrics Physical Health Underutilization

Accuracy	Precision	Recall (Sensitivity)	Specificity	Person- months	Subgroup
0.4662	0.9335	0.0810	0.9921	19527583	Overall
0.4711	0.9591	0.0708	0.9960	4065079	new_medicaid_enrollee_12_months
0.5053	0.9167	0.0507	0.9950	9322099	sex_m
0.4306	0.9398	0.1037	0.9887	10205484	sex_f
0.4520	0.9339	0.0753	0.9924	13784517	primary_language_eng
0.5008	0.9307	0.0875	0.9923	5130519	primary_language_spa
0.4909	0.9240	0.0709	0.9930	9979071	ethnicity_category_hispanicorlatino
0.4405	0.9405	0.0904	0.9910	9548512	ethnicity_category_nothispanicorlatino
0.4443	0.9536	0.0872	0.9935	62934	race_category_american_indian_or_alaska
0.4656	0.9414	0.1328	0.9871	986614	race_category_asian
0.4165	0.9169	0.0585	0.9915	1429960	race_category_black_or_african_american
0.4378	0.9487	0.1071	0.9903	179909	race_category_native_hawaiian_or_other_
0.4884	0.9275	0.0717	0.9932	10404749	race_category_other
0.4590	0.9307	0.0791	0.9917	1216306	race_category_two_or_more_races
0.4524	0.9424	0.0790	0.9930	2431728	race_category_unknown
0.4271	0.9421	0.1048	0.9888	2815383	race_category_white

2) Figures

Figure A.1. Grid of Calibration Plots (Adults)¹²

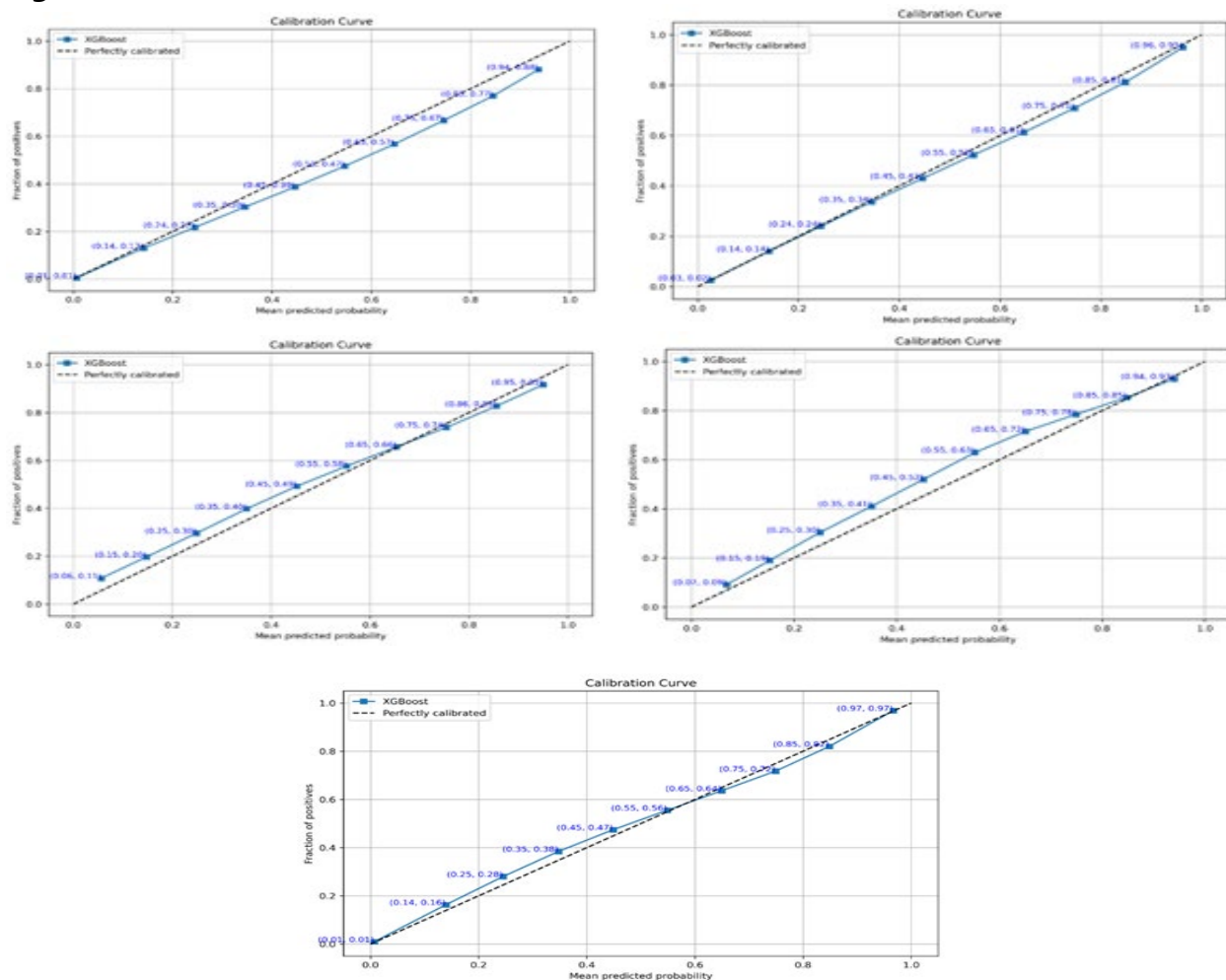
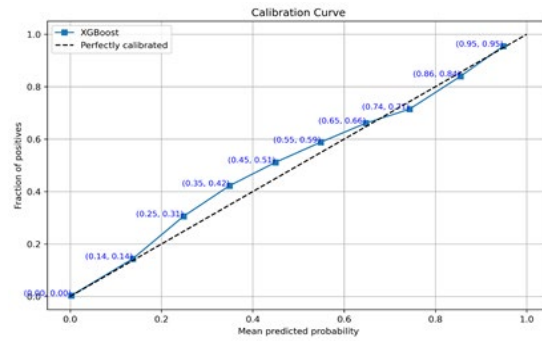
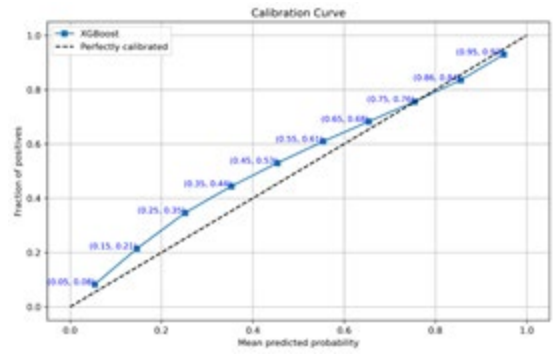
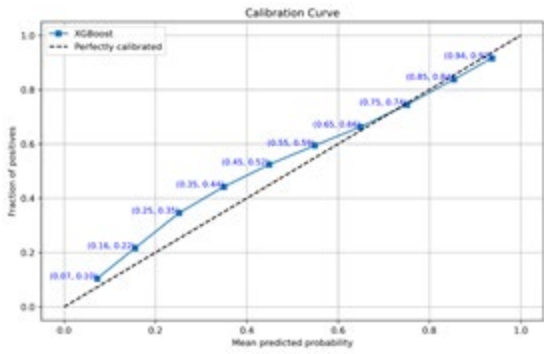
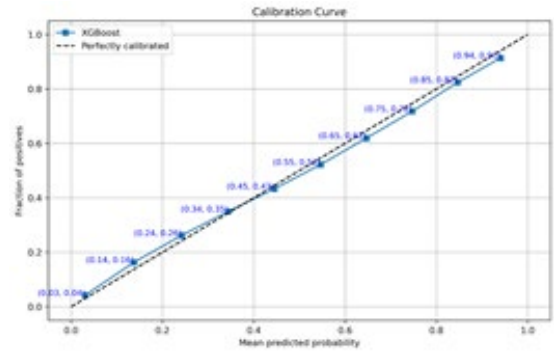
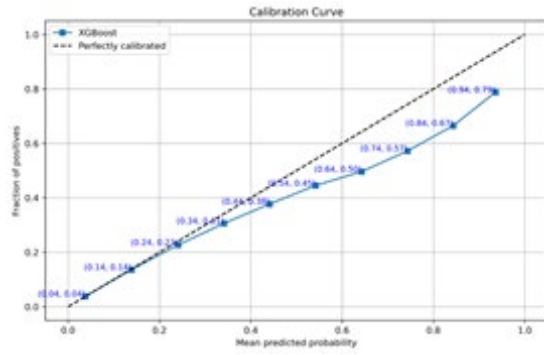


Figure A.2. Grid of Calibration Plots (Pediatrics)¹³

¹² The grid of calibration plots for adult models is ordered left to right, top to bottom as follows: (1) Behavioral Health Adverse Events, (2) Physical Health Adverse Events, (3) Behavioral Health Underutilization, (4) Physical Health Underutilization, and (5) Social Risk Adverse Events. Each plot compares predicted probabilities to observed event rates across deciles of risk.

¹³ See Figure A.1 – Grid of Calibration Plots (Adults) footnote



D. References

References used for performance comparisons:

Gao, J., Reyes, E. M., Scherer, J. A., and Cooper, K. M. 2021. "Predicting Opioid Use Disorder and Associated Risk Factors in a Medicaid Managed Care Population." *American Journal of Managed Care* 27 (10): e328–e336.

Holcomb, R., Murali, M., Dalton, D., and Shade, L. 2022. "Predicting health-related social needs using machine learning in Medicaid and Medicare populations." *Scientific Reports* 12: Article 5085.

Lo-Ciganic, W.-H., Huang, J. L., Zhang, H. H., Weiss, J. C., Kwoh, C. K., Donohue, J. M., and Cochran, G. 2021. "Using machine learning to predict risk of opioid overdose among Medicaid beneficiaries in the United States: A cross-sectional study." *PLOS ONE* 16 (4): e0248360.

Lo-Ciganic, W.-H., Huang, J. L., Zhang, H. H., Weiss, J. C., Donohue, J. M., Kwoh, C. K., and Cochran, G. 2022. "Machine learning prediction of opioid overdose risk among Medicaid beneficiaries in two US states: A retrospective cohort study." *The Lancet Regional Health – Americas* 9: 100181.

Patel, M., Callison-Burch, C., Stewart, E., Le, Q., and Shah, N. H. 2024. "Predicting all-cause acute care visits among Medicaid beneficiaries using machine learning." *Scientific Reports* 14: Article 7183.

Pourat, N., Wallace, S. P., Slusser, W., and Kominski, G. F. 2023. "Identifying people experiencing homelessness in Medicaid data: A validation study using California's Whole Person Care program." *Health Services Research* 58 (1): 21–30.

Rahman, M. M., Yu, S. W. Y., Lightman, M., and Mezuk, B. 2022. "Machine learning to predict suicide-related events among Medicaid beneficiaries with mental illness." *Journal of Affective Disorders* 304: 27–35.

Yu, J., Brenner, P. R., Kedia, S. K., DeMik, R. J., and Mackey, K. 2022. "Predicting potentially avoidable psychiatric emergency department visits using machine learning: A national study." *Psychiatry Research* 312: 114558.