[WORKING DRAFT]

1. What are the four model performance measures Bamboo Health makes available to users on the Overdose Risk Score (ORS) model?
    - **Precision** (aka positive predictive value, true positive rate): Precision is a measure of how many of the model's positive predictions are actually correct. In other words, precision tells us how reliable the model is when it indicates that an outcome of interest (e.g., unintentional death due to overdose) is more likely to occur (a 'positive' outcome). A higher precision value indicates a lower rate of false positives, suggesting that when the model predicts an instance as positive, it is usually correct. Conversely, a lower precision value implies a higher rate of false positives, indicating that the model may make more incorrect positive predictions.

Table 1: Example of Higher Precision (total N = 200 patients)

	Actual Positive	Actual Negative
Predicted Positive	30 True Positive	10 False Positive
Predicted Negative	5 False Negative	155 True Negative

precision = true positives / (true positives + false positives) = 30 / (30 + 10) = 0.75

Table 2: Example of Lower Precision (total N = 200 patients)

	Actual Positive	Actual Negative
Predicted Positive	10 True Positive	40 False Positive
Predicted Negative	25 False Negative	125 True Negative

precision = true positives / (true positives + false positives) = 10 / (10 + 40) = 0.20

In these examples, we can see how precision varies based on the number of true positive and false positive predictions in relation to the actual positive and negative instances.

- **Recall** (aka sensitivity): Recall helps assess the model's ability to avoid false negative predictions. A higher recall value indicates a lower rate of false negatives, implying that the model can effectively capture most of the positive instances. On the other hand, a lower recall value suggests a higher rate of false negatives, indicating that the model may miss a larger number of positive instances in the dataset.

Table 1: Example of Higher Recall (total N = 200 patients)

	Actual Positive	Actual Negative
Predicted Positive	30 True Positive	10 False Positive
Predicted Negative	5 False Negative	155 True Negative

recall = true positives / (true positives + false negatives) = 30 / (30 + 5) = 0.86

Table 2: Example of Lower Recall (total N = 200 patients)

	Actual Positive	Actual Negative
Predicted Positive	10 True Positive	40 False Positive
Predicted Negative	25 False Negative	125 True Negative

recall = true positives / (true positives + false negatives) = 10 / (10 + 25) = 0.29

In these examples, we can see how recall varies based on the number of true positive and false negative predictions in relation to the actual positive instances.

- Precision and recall often work against each other. That is, improving precision typically reduces recall and vice versa.
    a) If we decrease false positives, then precision increases and recall decreases.
    b) If we decrease false negatives, then recall increases and precision decreases.

Therefore, modelers need to decide which metric is more important for a given event/outcome.

- Precision is like being precise or accurate. It measures how many of the items our model predicted as positive are actually positive. So, it's about being careful not to make too many mistakes of saying something positive when it's not.
- Recall is about not missing things. It measures how many of the actual positive items our model managed to find. So, it's like capturing or recalling as many positive things as possible.

Think of precision as asking, "When the model says something is positive, how often is it correct?" And recall as asking, "Out of all the positive things, how many did the model manage to find?"

Imagine you're looking for your keys in a messy room:

Precision: If you only call out "keys" when you're absolutely sure you've found them, you have high precision.

Recall: If you manage to find most of the keys in the room, you have high recall.

In summary, precision is about being careful not to falsely claim something is positive, while recall is about ensuring you find as many positive things as possible.

- **Specificity** (aka true negative rate): Specificity is a measure of how well a model correctly identifies negative instances. It focuses on the accuracy of the model's negative predictions. In other words, specificity tells us how well the model can correctly identify outcomes as negative when they truly are negative. A higher specificity value indicates a lower rate of false positives, suggesting that when the model predicts an instance as negative, it is usually correct. Conversely, a lower specificity value implies a higher rate of false positives, indicating that the model may make more incorrect negative predictions.

Table 1: Example of Higher Specificity (total N = 200 patients)

	Actual Positive	Actual Negative
Predicted Positive	30 True Positive	10 False Positive
Predicted Negative	5 False Negative	155 True Negative

specificity = true negatives/ (true negatives + false positives) = 155 / (155 + 10) = 0.94

Table 2: Example of Lower Specificity (total N = 200 patients)

	Actual Positive	Actual Negative
Predicted Positive	10 True Positive	40 False Positive
Predicted Negative	25 False Negative	125 True Negative

specificity = true negatives / (true negatives + false positives) = 125 / (125 + 40) = 0.76

In these examples, we can see how specificity varies based on the number of true negative and false positive predictions in relation to the actual negative instances.

- **Negative Predictive Value (NPV)**: Negative Predictive Value helps assess the model's accuracy in predicting negative instances. A higher NPV value indicates a lower rate of false negatives, suggesting that when the model predicts an instance as negative, it is usually correct. On the other hand, a lower NPV value suggests a higher rate of false negatives, indicating that the model may miss negative instances more often.

Table 1: Example of Higher NPV (total N = 200 patients)

	Actual Positive	Actual Negative
Predicted Positive	30 True Positive	10 False Positive
Predicted Negative	5 False Negative	155 True Negative

NPV = true negatives/ (true negatives + false negatives) = 155 / (155 + 5) = 0.97

Table 2: Example of Lower NPV (total N = 200 patients)

	Actual Positive	Actual Negative
Predicted Positive	10 True Positive	40 False Positive
Predicted Negative	25 False Negative	125 True Negative

NPV = true negatives / (true negatives + false negatives) = 125 / (125 + 25) = 0.83

In these examples, we can see how the NPV varies based on the number of true negative and false negative predictions in relation to the actual negative instances. Please note that the NPV and Recall are two distinct metrics that measure different aspects of the model's performance. Counter to expectation, they are not always directly correlated, and changes in one metric do not necessarily guarantee changes in the other metric. Recall focuses on the ability of the model to correctly identify positive instances, while NPV measures the accuracy of the model in predicting negative instances. The two metrics can be influenced by different factors and considerations.

2. How useful are those metrics to evaluate a model that is intended to indicate the likelihood of a rare event (e.g., such as unintentional overdose death)?
    - When the outcome being modeled is rare, machine learning model metrics such as precision and recall can still be useful but need to be interpreted carefully:
        a) Precision: Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. In the case of rare events, precision can be informative as it emphasizes the model's ability to avoid false positives. For example, if a model is intended to detect rare diseases, a high precision implies that the model rarely misclassifies healthy individuals as positive cases.
        b) Recall (sensitivity): Recall measures the proportion of correctly predicted positive instances out of all actual positive instances. In the context of rare events, recall is often considered the most important metric because it focuses on capturing the actual positive cases. High recall indicates that the model is effective at identifying the rare occurrences correctly.
    - However, these metrics have limitations when dealing with imbalanced datasets and rare events:
        a) Imbalanced datasets: When the rare event is significantly outnumbered by the negative instances, the model can achieve high precision by predicting everything as negative. This would result in low recall, meaning the model would miss most of the positive instances. Therefore, evaluating the model's performance based on precision alone can be misleading. It can be important to consider recall or other metrics that provide a more balanced view.
        b) Sample size: With rare events, the number of positive instances available for training and evaluation might be limited. This scarcity can lead to increased uncertainty in estimating model performance. Cross-validation techniques or resampling strategies (e.g., oversampling or undersampling) can be employed to mitigate scarcity impact.

3. Should any metric be more of a focus than others for ORS?
    - All metrics should be taken in balance when considering a model that predicts a rare adverse event, such as unintentional overdose death. However, it ultimately depends on what is the greater focus in terms of patient outcomes. For example, in preliminary screening of patients for follow-up examinations for a rare but potentially fatal event, a modeler would likely desire a recall as close as reasonably possible to 1.0. The modeler will seek to identify as many patients as reasonably possible who actually need more medical attention — and can accept lower precision. In other words, the modeler would accept identifying some patients predicted to be in need of more targeted care who actually don't need it. This could lead to additional healthcare costs (more visits) and provider time for false positives but helps prevent a larger percentage of potential adverse events.

4. Is the metric performance different at certain score bands?
    - The metrics made available by Bamboo Health for ORS are based on how well the given set of PDMP features predict unintentional overdose death (a binary outcome) in the logistic regression model.
    - While the metrics from the logistic model are not amenable to the evaluation of multiple cross-category comparisons, we can get a sense of how well the model performs at each 100-point score category by examining the Odds Ratio 95% confidence intervals (95% CI) for each category. The narrower the 95% CI, the more confidence can be attributed to the Odds Ratio.

ORS Category	Non-decedent	Decedent	Odds Ratio	OR 95% CI
000-199	193,636	1,108	1	ref
200-299	258,183	1,669	1.12	1.05-1.22
300-399	47,698	938	3.43	3.15-3.75
400-499	16,357	702	7.5	6.81-8.25
500-599	6,200	438	12.35	11.01-13.82
600-699	2,347	259	19.29	16.72-22.19
700-799	634	101	27.87	22.29-34.53
800-999	192	32	29.25	19.66-42.15

*The content in the above table relates to the ORS training dataset.

5. Are there considerations to be mindful of when reviewing validations of rare event models?

    - It is relatively common for logistic regression model metrics to be lower when the outcome being modeled is rare and the sample size is small. Several factors contribute to this phenomenon:
        - Imbalanced classes: When the outcome of interest (in the case of ORS, decedents) is rare compared to the other class (non-decedents), it leads to class imbalance. Logistic regression models are sensitive to imbalances in the cells in the 2x2 tables in the above examples, and the metrics tend to be lower when the outcome is rare. The model may end up predicting the outcome that occurs more frequently in the data set, leading to lower sensitivity (true positive rate) and higher specificity (true negative rate) for the rare class.
        - Few positive examples: In small sample sizes, the number of positive examples may be limited, making it more challenging for a model to learn from them effectively. With limited positive instances, a model may not capture the underlying patterns of the rare outcome as well as it would with more positive instances, resulting in lower metrics.
        - Overfitting: With a small sample size, there is a higher likelihood of overfitting, which occurs when a model memorizes ‘noise’ or specific patterns from the training data rather than learning generalizable patterns. Overfitting can lead to lower performance when the model is applied to new, unseen data.
        - Uncertainty in estimates: Small sample sizes can lead to wider confidence intervals for model estimates, which means that the model's coefficients and predictions may be less precise and have higher uncertainty.
    - When the sample size is small, evaluation metrics like accuracy, precision, and recall can become less reliable indicators of model performance. The metrics can make the model's performance appear lower than it actually is due to the limited amount of data available for evaluation.
    - To address these challenges and improve the performance of a logistic regression model when dealing with a rare outcome and a small sample size, modelers can consider the following approaches:
        - Collect more data, if possible, to increase the size of the dataset and improve the model's ability to learn from the rare class.
        - Employ techniques to handle class imbalance, such as oversampling, undersampling, or using class-weighted approaches during model training.
        - Create relevant features that can potentially provide more information to the model and help it better discriminate between the classes.
        - Regularization: Apply regularization techniques to prevent overfitting and improve the generalization of the model.

        - Consider alternative modeling approaches, such as tree-based models or ensemble methods, to address imbalanced data.
    - It is important to remember that the success of any approach will depend on the specific characteristics of the data and the challenge the modeler is trying to address. It is important to carefully assess the performance of the model using appropriate evaluation metrics and cross-validation techniques to confirm effectiveness and generalization to new data.

6. How have the metrics changed for ORSv1 from the training dataset to additional validation (% increase / decreases)?

Metric	ORSv1 training (y1)	ORSv1 additional validation (y2)	% change	% difference
Precision	0.753377	0.516892	(31.4)	37.2
Recall	0.571320	0.630228	10.3	9.8
Specificity	0.813162	0.411616	(49.4)	65.6
Negative Predictive Value	0.655029	0.527052	(19.5)	21.7

- Percent change: (y2 – y1 / y1) * 100
- Percent difference: abs(y2 – y1) / ((y2 + y1) / 2) * 100
- Analytic note: It is important to note that the recall increased from the first set of inputs to the second set, indicating that the model improved in capturing positive instances. However, the negative predictive value decreased in the second set, suggesting a decrease in the model's ability to predict negative instances. This situation can occur when there is a class imbalance in the dataset or when the distribution of true positives and true negatives differs between the two sets of inputs. For example, if the second set of inputs contains more challenging negative instances that are more difficult for the model to predict, it can lead to a decrease in the negative predictive value despite the improvement in recall. To better understand the relationship between the metrics, it is important to analyze the underlying characteristics of the data and consider other evaluation metrics, such as precision and specificity, for a comprehensive assessment of the model's performance.

7. What are the potential reasons for the changes in the metrics?
    - Sample size: The original model was trained using data from a state in the Midwest with 5,247 cases (decedents) and 525,247 age- and gender-matched controls (non-decedents). The additional validation dataset was obtained from a different state than the training dataset, with ~ 400 decedents and ~ 32,000 non-decedents[1]. The smaller sample size resulted in a very small number of patients with a score >= 700.
    - Prescribing patterns: Prescriber and patient behavior began to evolve in the mid-2010's due to many factors, such as legislation that was passed in each state to reduce improper prescribing behaviors and doctor-shopping. The original training data spanned the years 2013-2016. In contrast, the additional validation data is from 2017 through 2023. Thus, the more recent PDMP data should reflect the aforementioned changes in prescriber and patient behavior compared to the older

[1] Rounded to the nearest 100.

PDMP data. Moreover, there has been an uptick in the use of MOUDs to treat chronic pain since 2018 when the FDA approved a buccal film and transdermal patch for severe pain that requires management with opioids. Buprenorphine formulations have historically had very high MMEs, so any MOUD dispensation fed to the ORSv1 model would have influenced model performance and score calculations even if the MOUD was used appropriately to treat a chronic pain condition. In other words, ORSv1 includes MOUDs in all MME-related feature/predictor calculations. This is important to note since two out of the top four predictors in the ORSv1 model use some form of MME. 23% of the decedents in the additional validation data were prescribed a MOUD, while only 7% of the decedents in the original training data were prescribed a MOUD. This is a nontrivial swing in prescribing behavior, especially in light of knowing that MOUDs were comingled in every MME calculation in ORSv1. This could significantly dilute the separation in scores across all categories when a decedent has a MOUD dispensation for the treatment of chronic pain.

- PDMP features for ORSv1 upon which the scores are based: The utilization of PDMPs during the past decade (which includes the use of ORSv1 in many states across the U.S.) has reduced dangerous prescribing (e.g., polypharmacy, multiple provider use, etc.) (1). However, the overall landscape of the opioid epidemic evolved during this same period, and the drivers of unintentional overdose death have shifted away from prescription opioids to illicit opioids (2). The U.S. is now in the 3rd wave of the epidemic, and models using data primarily obtained from PDMPs may not consider certain data that would help convey risk linked to illicit opioids.

8. What is the predictive value of ORS when considered alongside the individual PDMP content it evaluates?
    - Daily MME is an example of individual PDMP content that might be assumed to provide meaningful value as a predictor of unintentional overdose death. Indeed, daily MME is a popular opioid exposure metric that is often used in the clinical setting. However, MME alone does not convey overall controlled substance utilization and has limited power in predicting unintentional overdose death. Understanding the likelihood of the rare event of unintentional overdose death calls for a multi-factor approach. This is why the ORS model incorporates many other PDMP features into assessing patient exposure. These data include drug seeking behavior (i.e., the count of pharmacies and prescribers), simultaneous overlapping drug usage of the same drug type, simultaneous overlapping drug usage of different drug types (e.g., opioids and sedatives), abrupt changes in MME, and usage patterns across a variety of time windows (i.e., near term versus long-term). As shown in Figure 1 below, increasing MME during the last 60 days in our training data was not a significant predictor of unintentional drug overdose death compared to increasing ORS.

References

1. Rhodes, E., Wilson, M., Robinson, A. *et al.* The effectiveness of prescription drug monitoring programs at reducing opioid-related harms and consequences: a systematic review. *BMC Health Serv Res* **19**, 784 (2019). https://doi.org/10.1186/s12913-019-4642-8.
2. Overdose Death Rates. National Institute on Drug Abuse. Published January 20, 2022. https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates#:~:text=Opioid%2Dinvolved%20overdose%20deaths%20rose.