Predicting Cognitive Decline from Baseline Assessments: An OASIS-3 Study

Tanish Parida, Gary Shen, Yunong Wu
(Mentored by Dr. Marly Gotti)

MIT PRIMES Conference

October 25th, 2026

Table of Contents

1. The problem: predicting future cognitive decline
2. The OASIS-3 dataset, clinical vocabulary, and methodology
3. Classification models & Data Leakage
4. Results, limitations, and next steps

Section 1

The problem: predicting future cognitive decline

Motivation:

- Alzheimer's disease and other dementias place a substantial burden on patients, families, and healthcare systems.
- Identifying people at risk before substantial decline could support better treatments, and provide more time for families to plan.



Brain maps illustrating Alzheimer's-related changes over 18 months. Courtesy of Dr. Paul Thompson, UCLA School of Medicine, via ALZinfo.org.

Question

Can **baseline** assessments predict **reliable cognitive decline** over the next 2–4 years?

- In this study, we aim to classify whether a patient will experience future cognitive decline or not, using only non-imaging predictors.
- Such predictors include demographics, clinical impairment ratings, and cognitive test scores.

Finding

Baseline assessments provided meaningful prediction of future decline, establishing a non-imaging benchmark for testing whether imaging adds predictive value.

Question

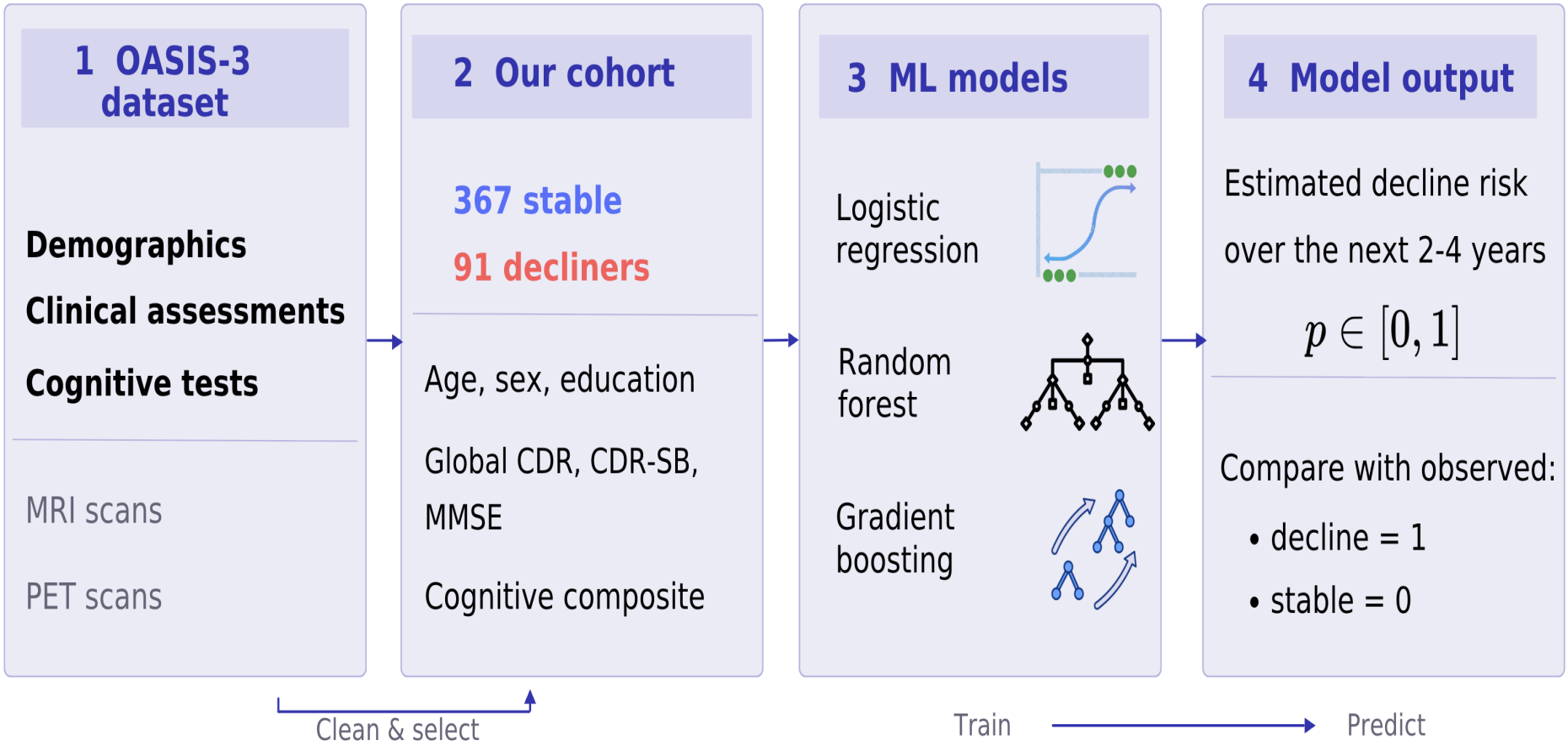
Can **baseline** assessments predict **reliable cognitive decline** over the next 2–4 years?

- In this study, we aim to classify whether a patient will experience future cognitive decline or not, using only non-imaging predictors.
- Such predictors include demographics, clinical impairment ratings, and cognitive test scores.

Finding

Baseline assessments provided meaningful prediction of future decline, establishing a non-imaging benchmark for testing whether imaging adds predictive value.

Study Workflow



Section 2

The OASIS-3 dataset, clinical vocabulary, and methodology

About the OASIS-3 Dataset

- OASIS-3 brings together anonymized data from long-running studies at the Washington University Knight ADRC.
- The dataset used in this study includes **1,378 participants**, ages 42–95, with different levels of cognitive impairment.
- **Longitudinal**: Repeated visits let us connect an earlier assessment with cognitive performance several years later.
- Scope: **demographic, clinical, and cognitive data** (non-imaging), MRI and PET (imaging) data were not used.

About the OASIS-3 Dataset

- OASIS-3 brings together anonymized data from long-running studies at the Washington University Knight ADRC.
- The dataset used in this study includes **1,378 participants**, ages 42–95, with different levels of cognitive impairment.
- **Longitudinal**: Repeated visits let us connect an earlier assessment with cognitive performance several years later.
- Scope: **demographic, clinical, and cognitive data** (non-imaging), MRI and PET (imaging) data were not used.

About the OASIS-3 Dataset

- OASIS-3 brings together anonymized data from long-running studies at the Washington University Knight ADRC.
- The dataset used in this study includes **1,378 participants**, ages 42–95, with different levels of cognitive impairment.
- **Longitudinal**: Repeated visits let us connect an earlier assessment with cognitive performance several years later.
- Scope: **demographic, clinical, and cognitive data** (non-imaging), MRI and PET (imaging) data were not used.

About the OASIS-3 Dataset

- OASIS-3 brings together anonymized data from long-running studies at the Washington University Knight ADRC.
- The dataset used in this study includes **1,378 participants**, ages 42–95, with different levels of cognitive impairment.
- **Longitudinal**: Repeated visits let us connect an earlier assessment with cognitive performance several years later.
- Scope: **demographic, clinical, and cognitive data** (non-imaging), MRI and PET (imaging) data were not used.

Values in the dataset

We list some of the variables in the dataset that were useful for prediction.

Variable	Type	Possible values	Meaning
Age	demographic	42–95	age of participant at visit
Global CDR	clinical	0, 0.5, 1, 2, 3	derived value; higher score = more severe dementia
MMSE	clinical	integer in $[0, 30]$	lower = more cognitive impairment
Logical Memory	cognitive	integer in $[0, 25]$	verbal episodic memory; higher = better
Craft Story	cognitive	integer in $[0, 44]$	verbal episodic memory; higher = better
Boston Naming	cognitive	integer in $[0, 60]$	visual object naming abilities; higher = better
Animal Fluency	cognitive	nonnegative integer	semantic memory; higher = better
Trail Making B (s)	cognitive	0–300	executive function; lower = better

Clinical Dementia Rating

Definition

The **Clinical Dementia Rating (CDR)** measures the severity of symptoms associated with Alzheimer's disease and other dementias.

- Assigns a score of 0, 0.5, 1, 2, and 3 to six areas: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care.
- An algorithm combines these ratings into a **global CDR score**. Higher scores indicate greater impairment.

Definition

CDR-SB is the sum of the six individual category scores.

Clinical Dementia Rating

Definition

The **Clinical Dementia Rating (CDR)** measures the severity of symptoms associated with Alzheimer's disease and other dementias.

- Assigns a score of 0, 0.5, 1, 2, and 3 to six areas: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care.
- An algorithm combines these ratings into a **global CDR score**. Higher scores indicate greater impairment.

Definition

CDR-SB is the sum of the six individual category scores.

Clinical Dementia Rating

Definition

The **Clinical Dementia Rating (CDR)** measures the severity of symptoms associated with Alzheimer's disease and other dementias.

- Assigns a score of 0, 0.5, 1, 2, and 3 to six areas: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care.
- An algorithm combines these ratings into a **global CDR score**. Higher scores indicate greater impairment.

Definition

CDR-SB is the sum of the six individual category scores.

Clinical Dementia Rating

Definition

The **Clinical Dementia Rating (CDR)** measures the severity of symptoms associated with Alzheimer's disease and other dementias.

- Assigns a score of 0, 0.5, 1, 2, and 3 to six areas: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care.
- An algorithm combines these ratings into a **global CDR score**. Higher scores indicate greater impairment.

Definition

CDR-SB is the sum of the six individual category scores.

Building the Prediction Dataset

1. **Choose an eligible starting visit.**
   Participants are at least 60 years old, with global CDR ≤ 0.5 and available cognitive data.
2. **Find a follow-up 24–48 months later.**
   Choose the qualifying visit closest to 36 months.
3. **Create one record per participant.**
   Combine the available baseline demographic, clinical, and cognitive measures.

Given data from the baseline visit, can we predict the cognitive decline that will occur 2-4 years later at the follow-up visit?

Cognitive Composite

- A single cognitive test may provide an incomplete picture. We combine several tests into one summary of cognitive performance.
- The composite covers **five domains**: immediate memory, delayed memory, naming, fluency, and executive function.
- Scores are standardized against a clinically stable reference group, then averaged across domains. **Higher means better**.
- We require at least **3 of 5 domains** to have valid data at both baseline and follow-up.

Cognitive Composite

- A single cognitive test may provide an incomplete picture. We combine several tests into one summary of cognitive performance.
- The composite covers **five domains**: immediate memory, delayed memory, naming, fluency, and executive function.
- Scores are standardized against a clinically stable reference group, then averaged across domains. **Higher means better**.
- We require at least **3 of 5 domains** to have valid data at both baseline and follow-up.

Cognitive Composite

- A single cognitive test may provide an incomplete picture. We combine several tests into one summary of cognitive performance.
- The composite covers **five domains**: immediate memory, delayed memory, naming, fluency, and executive function.
- Scores are standardized against a clinically stable reference group, then averaged across domains. **Higher means better**.
- We require at least **3 of 5 domains** to have valid data at both baseline and follow-up.

Cognitive Composite

- A single cognitive test may provide an incomplete picture. We combine several tests into one summary of cognitive performance.
- The composite covers **five domains**: immediate memory, delayed memory, naming, fluency, and executive function.
- Scores are standardized against a clinically stable reference group, then averaged across domains. **Higher means better**.
- We require at least **3 of 5 domains** to have valid data at both baseline and follow-up.

From Cognitive Change to an Outcome Label

- Raw score differences do not distinguish genuine decline from expected effects of baseline performance, demographics, time, and repeated testing.
- We estimate expected follow-up cognition in a **clinically stable reference group** using regression, accounting for baseline cognition, demographics, and time between visits.
- We compare actual performance with that expectation. A more negative standardized residual, R, means **worse than expected**.

Outcome	Rule
Decliner	$R \leq -1.645$
Stable	$R > -0.5$ and CDR-SB increase < 1 point
Intermediate	Excluded from the primary binary model

The primary task compares the clearest extremes.

From Cognitive Change to an Outcome Label

- Raw score differences do not distinguish genuine decline from expected effects of baseline performance, demographics, time, and repeated testing.
- We estimate expected follow-up cognition in a **clinically stable reference group** using regression, accounting for baseline cognition, demographics, and time between visits.
- We compare actual performance with that expectation. A more negative standardized residual, R, means **worse than expected**.

Outcome	**Rule**
Decliner	$R \leq -1.645$
Stable	$R > -0.5$ and CDR-SB increase < 1 point
Intermediate	Excluded from the primary binary model

The primary task compares the clearest extremes.


Cognitive composite: stable vs. decliners
Baseline
Follow-up
Cognitive composite (z-score)
2
1
0
−1
−2
−3
−4
−5
Stable
(n=367)
Decliners
(n=91)
Stable
(n=367)
Decliners
(n=91)

Final Analytic Cohort

Total OASIS-3 participants with non-imaging data (1378)

↓

Eligible baseline and follow-up visit for prediction dataset (916)

↓

Compute cognitive composite and reliable-change labels (643)

↓

Decliner (91) / Stable (367) / Intermediate (185)

↓

Final analytic sample for primary binary classification (458)

One entry per participant with relevant demographic, clinical, and cognitive data.

Section 3

Classification models & Data Leakage

Understanding Data Leakage

Definition

Data leakage occurs when a model uses information that would not be available at prediction time, artificially inflating its performance.

- **Temporal leakage:** using a variable from a future visit to make a baseline prediction.
- **Target leakage:** allowing the outcome, or information that directly reveals it, into the predictors.

Our Prediction Setup

Model Inputs come from information available by the prediction visit. Follow-up cognition determines the **outcome only**.

Understanding Data Leakage

Definition

Data leakage occurs when a model uses information that would not be available at prediction time, artificially inflating its performance.

- **Temporal leakage:** using a variable from a future visit to make a baseline prediction.
- **Target leakage:** allowing the outcome, or information that directly reveals it, into the predictors.

Our Prediction Setup

Model Inputs come from information available by the prediction visit. Follow-up cognition determines the **outcome only**.

Understanding Data Leakage

Definition

Data leakage occurs when a model uses information that would not be available at prediction time, artificially inflating its performance.

- **Temporal leakage:** using a variable from a future visit to make a baseline prediction.
- **Target leakage:** allowing the outcome, or information that directly reveals it, into the predictors.

Our Prediction Setup

Model Inputs come from information available by the prediction visit. Follow-up cognition determines the **outcome only**.

Understanding Data Leakage

Definition

Data leakage occurs when a model uses information that would not be available at prediction time, artificially inflating its performance.

- **Temporal leakage:** using a variable from a future visit to make a baseline prediction.
- **Target leakage:** allowing the outcome, or information that directly reveals it, into the predictors.

Our Prediction Setup

Model Inputs come from information available by the prediction visit. Follow-up cognition determines the **outcome only**.

Prediction Models

We compared four models using information available at prediction time.

Model	**Purpose**
Clinical-only logistic regression	Establish a simple baseline.
Clinical + cognitive logistic regression	Test the added value of the baseline cognitive composite.
Random forest	Combine decision trees to capture nonlinear patterns.
Gradient boosting	Build trees sequentially to improve earlier predictions.

Tabular neural networks did not meet our pre-specified sample-size requirements, so we did not train them.

Validation and Leakage Control

- **Evaluate on held-out participants.**
  Five outer folds estimate how well the models generalize.
- **Tune within the training data.**
  Five inner folds select model settings without using the held-out evaluation fold.
- **Fit preprocessing within training folds.**
  Missing-value imputation and scaling use training data only.

Validation and Leakage Control

- **Evaluate on held-out participants.**
  Five outer folds estimate how well the models generalize.
- **Tune within the training data.**
  Five inner folds select model settings without using the held-out evaluation fold.
- **Fit preprocessing within training folds.**
  Missing-value imputation and scaling use training data only.

Validation and Leakage Control

- **Evaluate on held-out participants.**
  Five outer folds estimate how well the models generalize.
- **Tune within the training data.**
  Five inner folds select model settings without using the held-out evaluation fold.
- **Fit preprocessing within training folds.**
  Missing-value imputation and scaling use training data only.

Section 4

Results, limitations, and next steps

Primary Results: Model Performance

AUROC measures how well a model ranks future decliners above stable participants. (e.g. AUROC $= 0.5$ suggests random ranking)

Model	AUROC	95% CI
Clinical-only logistic regression	0.841	0.789–0.889
Clinical + cognitive logistic regression	0.859	0.814–0.900
Random forest	0.855	0.804–0.902
Gradient boosting	0.838	0.786–0.886

- All models' bootstrap 95% confidence intervals overlap substantially.
- No model showed a clear advantage.

Primary Results: Model Performance

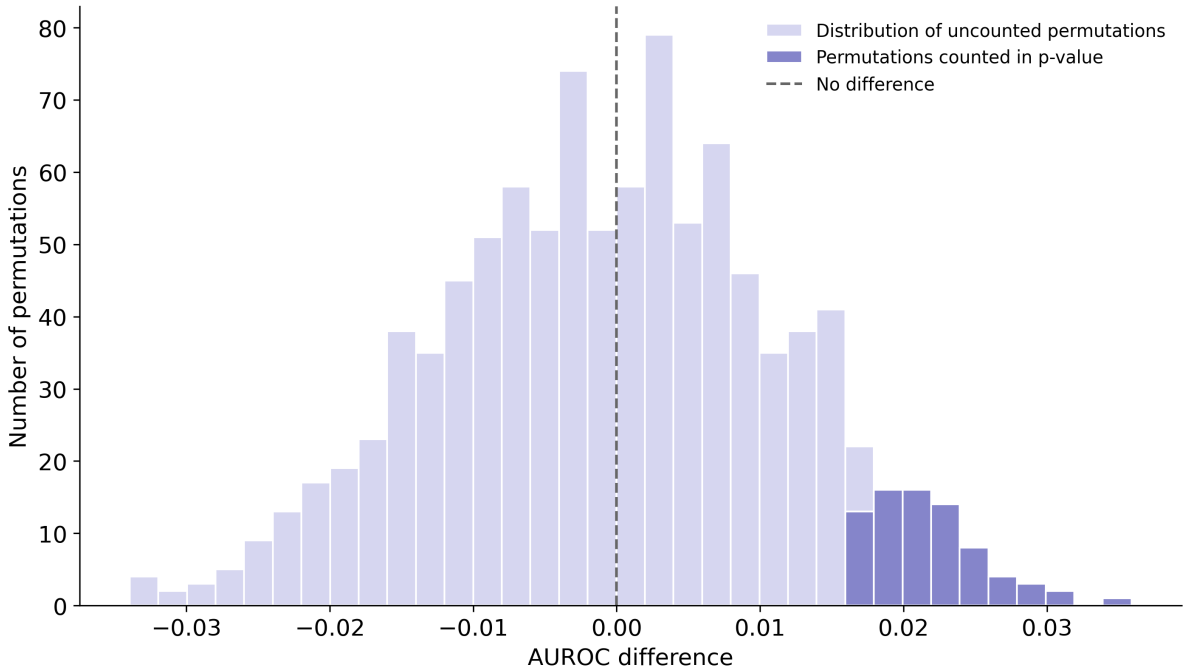
AUROC measures how well a model ranks future decliners above stable participants. (e.g. AUROC $= 0.5$ suggests random ranking)

Model	AUROC	95% CI
Clinical-only logistic regression	0.841	0.789–0.889
Clinical + cognitive logistic regression	0.859	0.814–0.900
Random forest	0.855	0.804–0.902
Gradient boosting	0.838	0.786–0.886

- All models' bootstrap 95% confidence intervals overlap substantially.
- No model showed a clear advantage.

Does the Cognitive Composite Add Predictive Value?

Model	AUROC	95% CI
Clinical-only logistic regression	0.841	0.789–0.889
Clinical + cognitive logistic regression	0.859	0.814–0.900



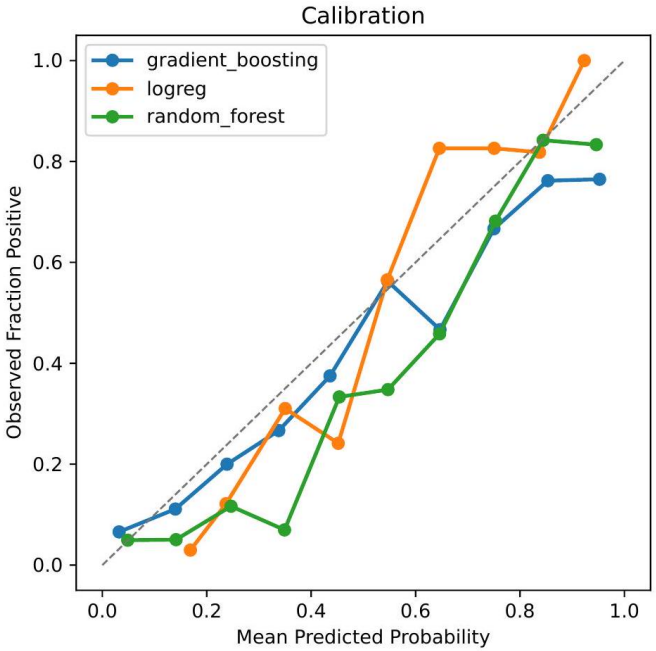
- Permutation-test $p = 0.07$.
- No statistically significant improvement.

Interpretation: Factors of Influence

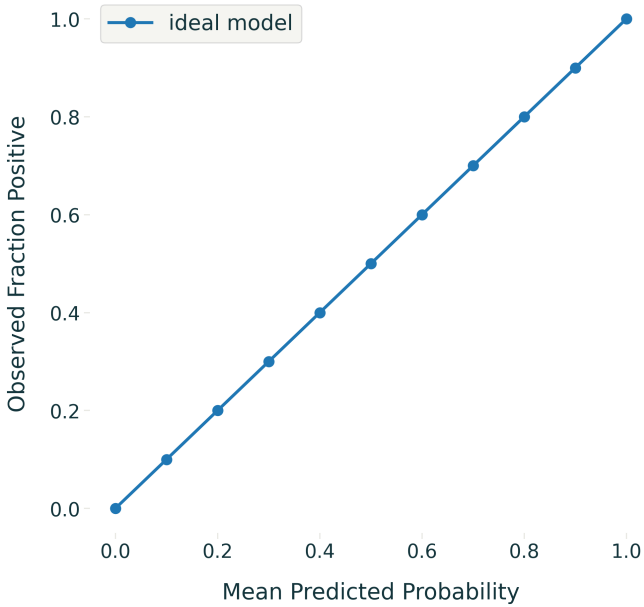
Baseline measure	Group difference Stable vs. future decliner	Model coefficient Clinical + cognitive LR
Global CDR = 0.5	1.236	+0.4735
CDR-SB	1.402	+0.4353
MMSE	1.018	−0.3571
Cognitive composite	0.997	−0.3364
Age	0.379	+0.1922

Main Limitation

- **Further validation needed before clinical application.**
  Our model does not yet produce accurate risk percentages despite strong ranking performance.



Observed calibration
Brier score = 0.111



Ideal calibration
Brier score = 0

Conclusion

- Routine baseline assessments helped identify who would decline over the next 2–4 years.
- Clinical information already carried much of the predictive signal.
- More complex non-imaging models did not show a clear advantage in ranking future decliners.

Central Contribution

We established a non-imaging benchmark for testing whether imaging data adds useful predictive information.

Conclusion

- Routine baseline assessments helped identify who would decline over the next 2–4 years.
- Clinical information already carried much of the predictive signal.
- More complex non-imaging models did not show a clear advantage in ranking future decliners.

Central Contribution

We established a non-imaging benchmark for testing whether imaging data adds useful predictive information.

Selected References



LaMontagne PJ, Benzinger TLS, Morris JC, et al.

OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease.



Morris JC.

The Clinical Dementia Rating (CDR): current version and scoring rules.



Morris JC, Weintraub S, Chui HC, et al.

The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers.



Weintraub S, Besser L, Dodge HH, et al.

Version 3 of the Alzheimer Disease Centers' Neuropsychological Test Battery in the Uniform Data Set (UDS).



Gavett BE, Ashendorf L, Gurnani AS.

Reliable change on neuropsychological tests in the Uniform Data Set.

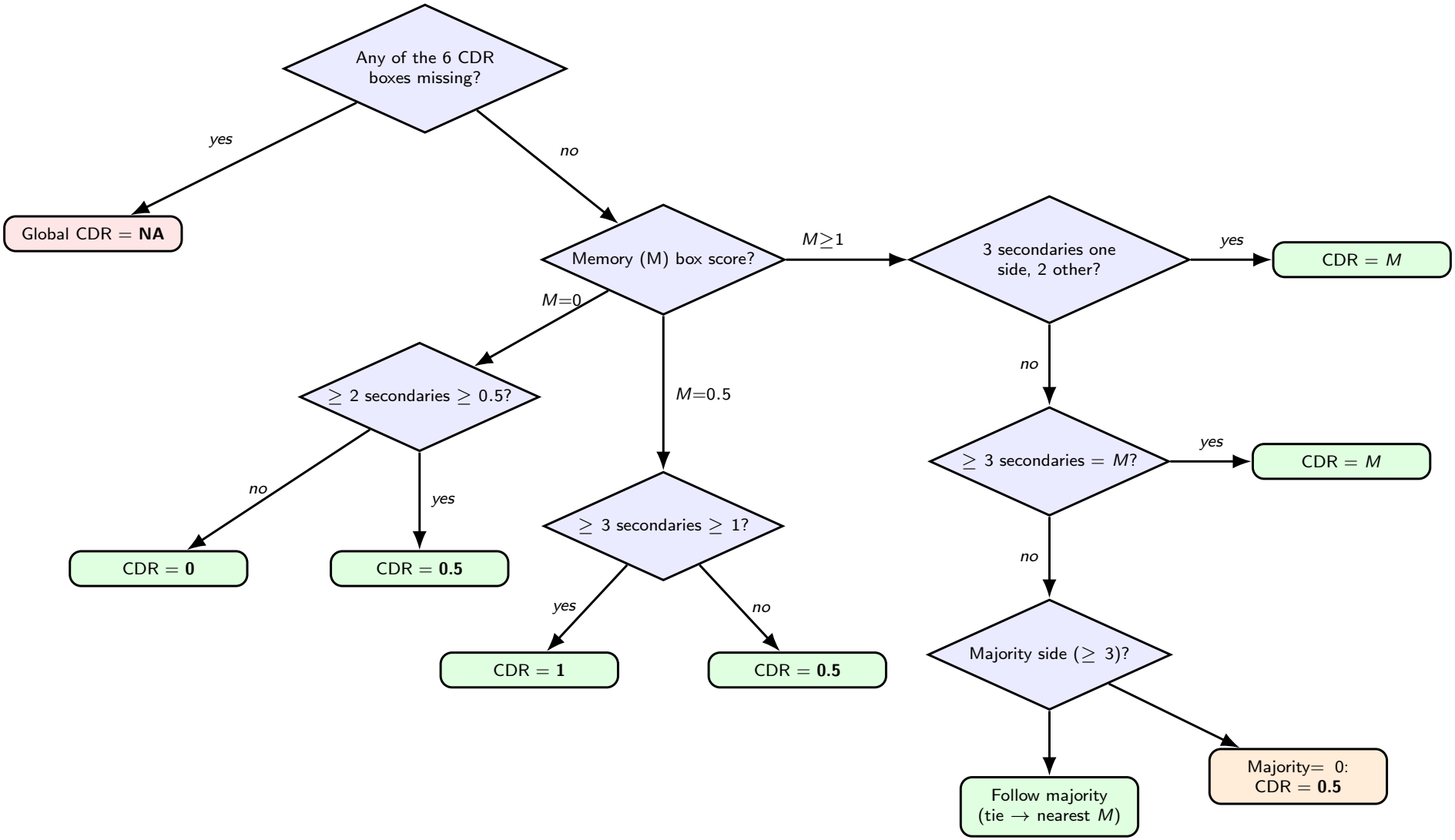
Acknowledgements

We would like to thank

- Dr. Marly Gotti for her guidance and mentorship throughout this research
- Dr. Willy Rodriguez and Dr. Felix Gotti for their advice and feedback as we prepared this presentation
- The PRIMES community for fostering a supportive environment and making this project possible
- The OASIS-3 investigators and study participants for making this research possible through their collection and sharing of data

Thank you!

Appendix A: Global CDR (Washington University/Morris Algorithm)



Appendix B: Clinically Stable Reference Group

The reference group was defined using clinical criteria, independently of cognitive-composite change:

- Baseline global CDR ≤ 0.5.
- Global CDR does not increase across the window.
- CDR-SB does not increase by ≥ 1 point.
- No worsening diagnostic conversion under the NACC hierarchy.

This avoids selecting the reference group using the cognitive change that we are trying to measure.

Appendix C: Constructing the Cognitive Composite

Step 1: Standardize each raw score into a z-score

$$Z_{ijv} = \frac{X_{ijv} - \mu_j}{\sigma_j}$$

X_{ijv} raw score for participant i, test j, visit v

$\mu_j,\ \sigma_j$ reference-group mean/SD for test j, estimated at baseline

Step 2: Average within domain, then across domains

$$C_{iv} = \frac{1}{D_i} \sum_{d=1}^{D_i} Z_{idv}$$

D_i number of available domains for participant i

Completeness rule

Requires ≥ 3 of 5 domains (otherwise composite is `NA`); an instrument only counts if valid at **both** baseline and follow-up.

Appendix D: Practice-Adjusted Reliable Change

Estimate expected follow-up cognition in the clinically stable reference group using regression:

$$C_{i,\text{follow}} = \beta_0 + \beta_1 C_{i,\text{index}} + \beta_2 \text{Age}_i + \beta_3 \text{Sex}_i + \beta_4 \text{Education}_i + \beta_5 \Delta t_i + \epsilon_i.$$

Compute the standardized residual:

$$R_i = \frac{C_{i,\text{follow}} - \widehat{C}_{i,\text{follow}}}{\widehat{\sigma}_\epsilon}.$$

- Negative residuals indicate worse follow-up cognition than expected.
- Decliner: $R_i \leq -1.645$.
- Stable: $R_i > -0.5$ and $\Delta\text{CDR-SB}_i < 1$.

Appendix E: Model Settings and Preprocessing

Model	**Selected settings**
Logistic regression	L2 regularization, $C = 0.01$, unweighted.
Random forest	200 trees, unrestricted depth, minimum leaf size 10, square-root feature sampling, balanced weights.
Gradient boosting	Learning rate 0.03, 100 iterations, L2 penalty 1, maximum 15 leaves, minimum leaf size 20, balanced weights.

- Training-fold preprocessing: median/mode imputation, missingness indicators, and scaling.
- Neural-network gate: $N \geq 500$, at least 100 decliners, and at least 10 events per encoded feature. The cohort did not qualify.

Appendix H: Evaluation metrics — The Confusion Matrix

Each participant receives a predicted probability of future reliable decline. At a chosen threshold, the probability becomes a binary prediction:

	Actual outcome	
Model prediction	**Decliner**	**Stable**
Decliner	**True positive (TP)**	**False positive (FP)**
Stable	**False negative (FN)**	**True negative (TN)**

Sensitivity/Recall

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

Among true decliners, the proportion correctly flagged.

Specificity

$$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

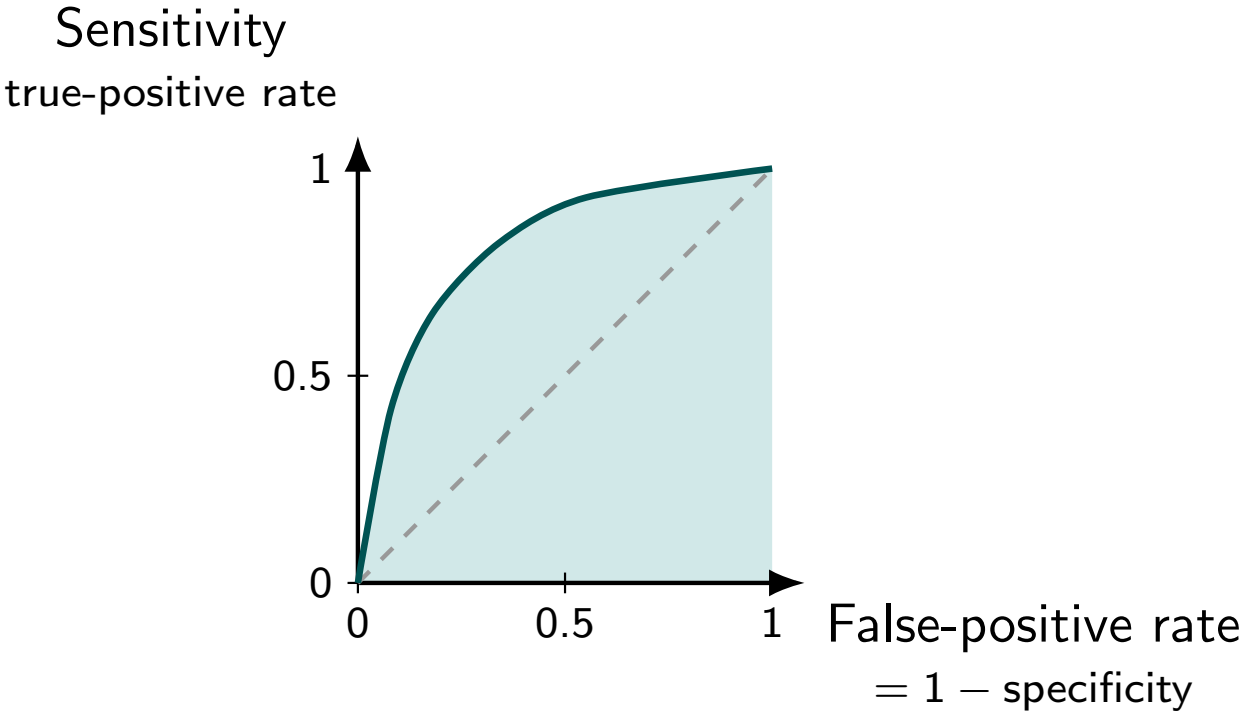
Among stable participants, the proportion correctly left unflagged.

Precision

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

Among participants flagged as decliners, the proportion who truly declined.

Appendix F: Evaluation metrics — AUROC



- Each point on the curve represents a different decision threshold.
- AUROC $= 0.50$: no better than random ranking.
- AUROC $= 1.00$: perfectly ranks every future decliner above every stable participant.

Definition

AUROC is the probability that a model assigns a higher predicted risk to a randomly chosen future decliner than to a randomly chosen stable participant.

Appendix G: Interpreting the Reported Results

- **AUROC 0.859:** a randomly chosen future decliner receives a higher risk score than a randomly chosen stable participant about 85.9% of the time. This is not classification accuracy.
- **Composite gain:** the reported increase is $+0.017$, with permutation-test $p = 0.157$. The study did not establish a statistically significant improvement.
- **Calibration:** predicted risks should agree with observed outcomes. The primary model's intercept was -0.219 and slope was 1.434, compared with ideal values of 0 and 1.

Appendix H: Future work

1. Test the models on an independent datasets (e.g. NACC).
2. Train a convolutional neural network (CNN) on imaging data such as MRI and PET scans and evaluate whether they provide additional predictive value.
3. Calibrate the model to develop a dementia prediction tool for clinicians.