

Synthetic Tau PET Model Card v1

Table of Contents:

[Model Name](#)

[Model Summary](#)

[Model Summary](#)

[Usage](#)

[Model Creators](#)

[System Type](#)

[Implementation Frameworks](#)

[Compute Requirements](#)

[Model Characteristics](#)

[Data Overview](#)

[Evaluation Results](#)

[Aggregate Evaluation Results](#)

[Subgroup Evaluation Results](#)

[Fairness Evaluation Results](#)

[Model Usage & Limitations](#)

[Terms of Art](#)

[Reflections on the Model](#)

Model Card Version: 1.0_2022
License: Apache 2.0

A classifier to distinguish Alzheimer’s Disease from other dementia etiologies and cognitively unimpaired individuals based on FDG PET scan and structural MRI, which returns a synthesized Tau PET scan as an intermediate product.

Synthetic Tau PET

This doc: go/mayo-neuro-documentation

Model Card Authors: Gerardo, Stephen, Jeyeon, Nick, Parker, Dave

Model Summary

Model Summary

MODEL ARCHITECTURE

U-Net to translate FDG-PET to Tau-PET scan, after using structural MRI for registration.

INPUT(S)

FDG-PET voxels, as registered by structural MRI.

OUTPUT(S)

Binary prediction of Alzheimer's based on thresholding of predicted SUVR, based on synthesized Tau-PET voxels.

Usage

APPLICATION

Where has this model been used, or where is it currently used?

- Currently a research project, establishing the viability of cross-modality neuroimaging synthesis.
- Eventually, for a clinician looking to distinguish whether a patient's symptoms are related to Alzheimer's pathophysiology, informing medical decision making.

BENEFITS

Why might users choose to use this model, relative to others?

- FDG-PET has utility across all potential etiologies of degenerative and non-degenerative dementia syndromes and is not restricted to Alzheimer's disease. Making it a general tool in a dementia practice that will be needed in many cases even if a direct tau biomarker is obtained
- Using CSF Tau to estimate Tau in the brain can be unreliable.
- FDG-PET is 5x cheaper than Tau-PET, doesn't require a cyclotron lab, is more widely available, has regulatory

KNOWN CAVEATS

Are there any known and preventable failures about this model?

- Based on a homogeneous population; mapping between FDG and Tau may not translate to other populations.
- Generalization to older generation scanners may not be as good.

	approval for clinical use and is reimbursed by medicare.	
Model Creators		
PERSON OF CONTACT	MODEL AUTHOR(S)	CITATION
<i>Who is the point of contact for questions about the model?</i>	<i>If different from previous, who all created the model?</i>	<i>If available, provide a citation to your model; else indicate unavailable.</i>
Jeyeon Lee, lee.jeyeon@mayo.edu	N/A	Manuscript in progress.
System Type		
SYSTEM DESCRIPTION	UPSTREAM DEPENDENCIES	DOWNSTREAM DEPENDENCIES
<i>Is this a standalone model, or intended to be used as part of a system with other models? Include links where necessary.</i>	<i>If the model requires specific inputs, where should they come from? Are there any specific preprocessing steps that should be applied? Include links where necessary.</i>	<i>If the model's outputs can be fed into another system, where should they go? Are there any specific post-processing steps that should be applied? Include links where necessary.</i>
Can be standalone. It can be incorporated into other models of FDG-PET that are planned but not yet built.	Requires FDG PET and registration based on structural MRI. Performance doesn't seem highly dependent on instrument type within the sample, but may not generalize well to others.	N/A
Implementation Frameworks		
HARDWARE & SOFTWARE FOR TRAINING	HARDWARE & SOFTWARE FOR DEPLOYMENT	
<i>Describe the hardware and software used for training the model.</i>	<i>Describe the hardware and software used for deploying the model.</i>	
Hardware (graphic processing unit): <ul style="list-style-type: none"> NVIDIA Tesla P100 The following packages were used: <ul style="list-style-type: none"> Tensorflow (v1.9.0) Conda (v4.8.2) 	<ul style="list-style-type: none"> N/A 	

- Cuda Toolkit (v9.1.85)
- Cudnn (v7.0.5)
- Python (3.6.1)
- Numpy (v1.19.5)
- Scikit-learn (v0.22.2)
- Scipy (v1.4.1)

Supervised Learning Problem Setup

LIMITATIONS IN SCOPE	KNOWN OUT OF SCOPE	VALIDATION / DEPLOYMENT ENVIRONMENT SKEW
<i>What are the known technical limitations of the model? E.g. What kind(s) of data should the model be expected not to perform well on? What are the factors that might degrade model performance?</i>		<i>What is the size of the model? Include attributes like number of weights and layers.</i>
Potentially different, less effective equipment not used in Mayo training, although tests suggest that generalization between state of the art scanner types at Mayo is pretty good.	Trying on cognitively normal patients, and patients with strokes, trauma or other conditions that might affect FDG.	Validation environment reflects demographics of the upper midwest. The test set is national, but still skewed by patient recruitment and enrollment in aging studies.
METRICS	DECISION THRESHOLDS	BASELINE
<i>What are the relevant performance metrics for your model?</i>	<i>If decision thresholds are used, what are they, and why were those decision thresholds chosen?</i>	<i>If any, describe the techniques implemented to preserve privacy?</i>
AUROC for predicting over/under meta-ROI SUVR threshold for TAU PET. Also correlation and MAPE with per-ROI SUVR.	No. We use AUROC instead. However, there are four decision thresholds for the label (tau PET meta ROI SUVR) common in the literature, based on distribution in the cognitively normal. Of these 1.21 is the most common.	Compared to a straight sum of SUVR in other modalities, and to cortical thickness estimation from MRI, but not to a classifier trained directly on other modalities. Nor to behavioral diagnosis or structured non-imaging data.

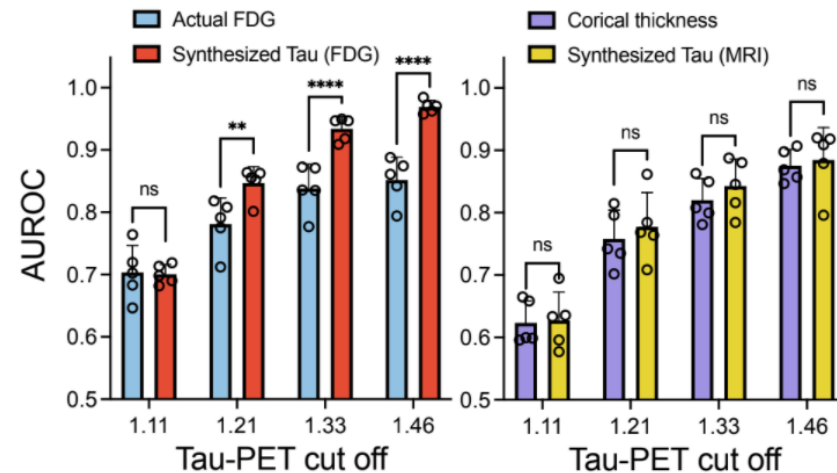
Data Overview		
TRAINING DATASET SOURCE	CHALLENGE TASK DATASET SOURCE	INSTRUMENTATION
<i>Describe the dataset used to train the model. If a requested detail is inapplicable, following guidance on N/A. Include links to additional table(s) with more detailed breakdowns in the caption.</i>	<i>What datasets distinct from the training / validation distribution are used to ensure generalization, and test for underspecification.</i>	<i>What instruments were used to collect or process the data? Describe any notable instrumentation requirements in the collection or preprocessing of data.</i>
<ul style="list-style-type: none"> The training data comes from a mix of two sources, both already used extensively for research: The Mayo Clinic Study of Aging, a longitudinal study of the aging process using a probability sample of older residents of Olmsted County, Minnesota, without any severe, acute medical conditions that interfere with neurological assessment. This study started in 2004 and data has been collected continuously since. The Mayo Clinic Alzheimer's Disease Research Center, which was founded in 1991 and collects data from existing Mayo Clinic patients who have Alzheimer's, FTD, DLB or related conditions. 	<ul style="list-style-type: none"> The Alzheimer's Disease Neuroimaging Initiative is a registry collecting a subset of data from studies conducted at various neurology research centers. Data from 288 patients is used to test out-of-sample generalization and underspecification performance. This data is more geographically dispersed, but coming from academic medical research centers, still has demographics that do not represent the US more broadly. 	<p>Collecting Data requires access to MRI for structural scans and registration, as well as PET scanners, and a nuclear medicine laboratory with a cyclotron to make the required radioactive isotopes. Tests show that, at least among the MRI types available at Mayo, the machine does not appear to significantly affect performance. However, lower field strength machines might have different results.</p>

This dataset has also been continuously collected to this day.		
DATA PRE-PROCESSING	CHALLENGE TASK DEMOGRAPHICS	
Describe any augmentation methods used during pre-processing to attain the requisite format. Are there any criteria that data points must satisfy to be included in the training set?	Does the data contain any labeled** groups, or attributes that suggest <u>demographic group membership</u> ? Describe any demographic groups considered when evaluating distributions in the data.	
Individuals who are missing one or more scanner modalities are dropped. PET scans are registered to the structural MRI and mapped to a brain template. Technicians perform quality control and discard low quality images.	Gender Race / Ethnicity Age	57% Male 43% Female 86% White 6% Black 5% Hispanic 1% Asian 2% More than one race 86% Over 65 14% Under 65
Challenge tasks test implicit claims about model reliability. They come in three major flavors: 1) Stratification 2) Testing on a different data set 3) Transformations on individual examples		
• In this case, we have evaluated performance of the Alzheimer's classification model on a different dataset collected in a different manner, and broken down performance by age, gender and race/ethnicity. • We have not evaluated the effect of systematic transformations on input voxels on the quality of the output generated.		
**If there are groups that may be present, but are not labeled in the training data, please note this in the Ethical Considerations section below.		

Evaluation Results	
Aggregate Evaluation Results	
Document your aggregate or overall model performance evaluation.	
EVALUATION PROCESS	EVALUATION RESULTS
Describe any notable factors in your process for evaluating your model's overall performance.	Summarize and link to evaluation results for this analysis.

Since clinical definitions of Alzheimer's are based on actual Tau-PET, the goal is to evaluate the ability of the synthesized Tau-PET to come to the same prediction. For the sake of giving a binary outcome several thresholds referenced in the literature were tested (see next cell). However, for the purpose of stress testing, we used the threshold of 1.21.

The model trained to predict Tau-PET scans substantially outperformed a model that directly used FDG-PET scans as a prediction of Tau-PET. However, it also substantially outperformed a model evaluating cortical thickness from structural MRI.



Subgroup Challenge Task Results

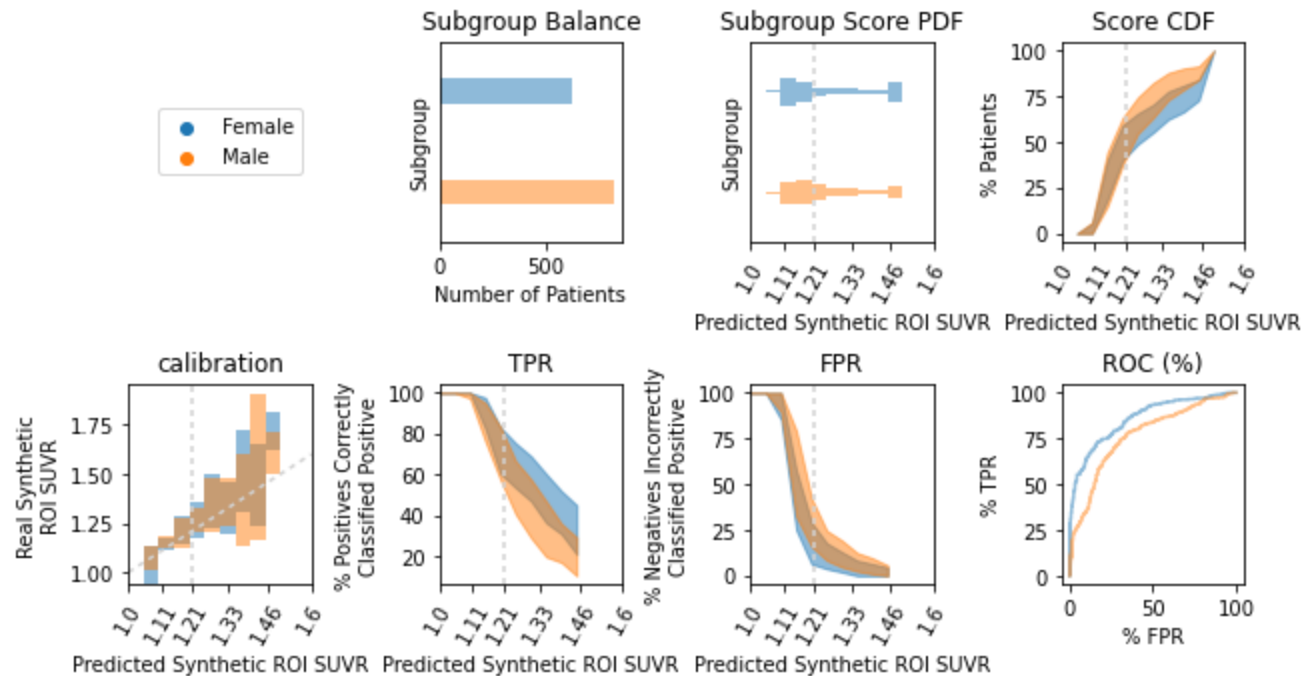
Compare all three major types of fairness metrics across subgroups:

- 1) Calibration - are members of different subgroups with the same model prediction, actually similar?
 - 2) Demographic Parity - do the distributions of scores for different subgroups match?
 - 3) Equalized Odds - Conditional on real disease state, what are the odds of correct classification for different groups?
- [read this off a vertical slice of the TPR and FPR charts]

It's important to consider not only different performance on different groups, but underspecification - whether irrelevant factors like a random seed used in training can radically change performance outside the training set. Quantify uncertainty across both random initializations, and different instantiations of the testing set.

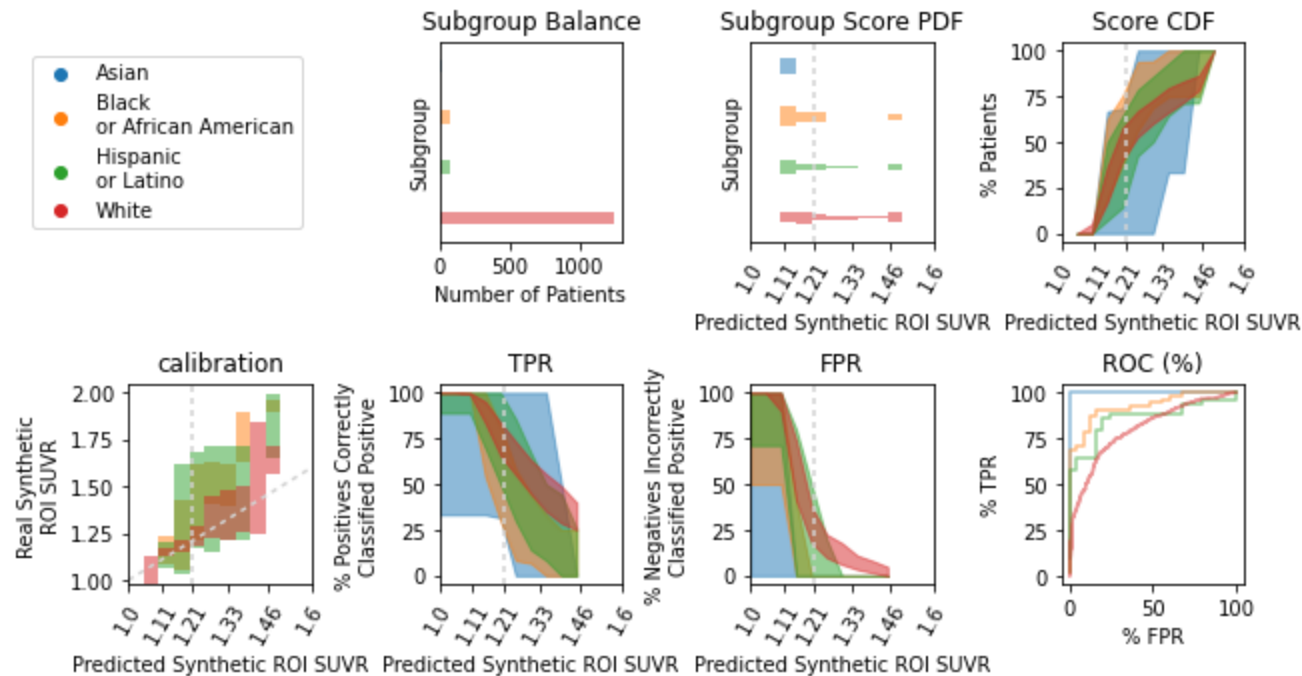
In this case, we've truncated scores to the clinically significant range of 1-1.5.

GENDER



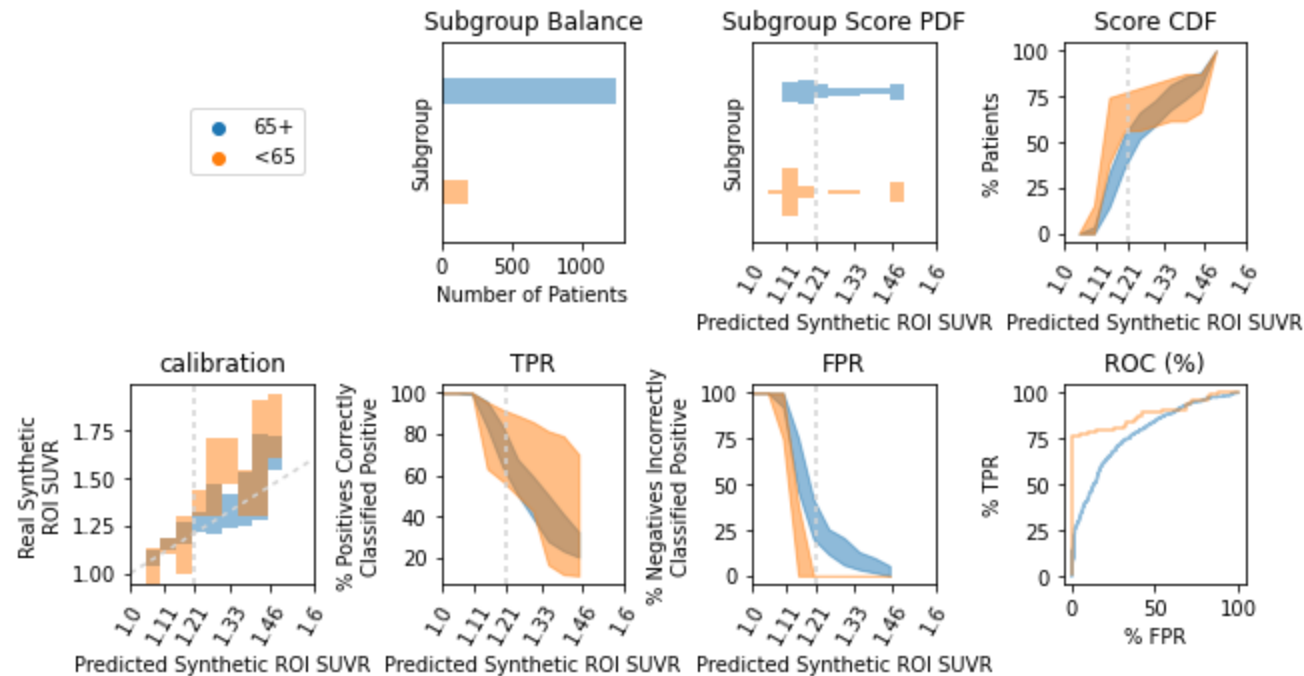
In this case it seems likely there are notably more female patients than male with very high predicted scores. It also seems likely the model performs better for both female patients with and without Alzheimers, although the confidence intervals overlap using an SUVR threshold of 1.21. There's no obvious large discrepancy in calibration between the subgroups.

RACE AND ETHNICITY



Ethnic minority representation is very low in the test set, leading to very wide intervals. We cannot reasonably establish whether the model is fair to ethnic minorities. For example, the one asian american woman without alzheimer's in the sample consistently ranked lower than the two asian american women with Alzheimers across all 5 random seeds. It's difficult to generalize from this. Calibration for Black and Hispanic groups seems likely worse, but again the intervals are too wide to know. Minority representation is inadequate to break down further intersections, such as race x gender.

AGE



The set of patients under 65 is too small to be sure. It seems to have few patients with very high tau protein levels, and model calibration looks like it may be worse. A patient under 65 without alzheimers seems to be much less likely to be incorrectly classified as having alzheimers than one over 65.

Model Usage & Limitations

SENSITIVE USE

Are there any use cases where deployment of this model would be considered sensitive? In particular, consider those listed in the [Sensitive Topics Consult documentation](#).

The determination of Alzheimer's disease state is sensitive, but the

LIMITATIONS

What factors might limit the performance of the model? What conditions must be satisfied to use the model?

- The model has not been tested on MRI machines with significantly lower field

PATIENT INVOLVEMENT

How were patient groups involved in the collection of data? Was data directly collected, and were patients able to consent? Can this consent be revoked?

- Data was collected directly from patients on Mayo's campus (or in the ADNI case at other

data would require a patient to have already consented to an FDG-PET scan. If this scan were taken for another reason, then the decision to also infer Alzheimer's status would be sensitive, and require further patient consent.

strength than those available at Mayo Clinic or in research laboratories contributing to the Alzheimer's Disease Neuroimaging Initiative.


- The model has also not been tested on groups who were not well represented in the training or stress test sets.

research facilities). Given the study subject, informed consent will in some cases have been provided by a legal designee, rather than the study subject themselves. Patients and their designees can opt out of the study at any time.

- Advocacy groups have been involved in designing questionnaires, but not directly in imaging.

Terms of Art

Concepts and Definitions referenced in this Model Card

 Use this space to include the expansions and definitions of any acronyms, concepts, or terms of art used across the Model Card. Use standard definitions where possible (e.g. [MLCC Glossary](#)). Include the source of the definition where indicated. If you are using an interpretation, adaptation, or modification of the standard definition for the purposes of your Model Card or model, include your interpretation as well.

FDG PET

PET scan based on glucose metabolism byproducts, gives a sense of brain activity.

Tau PET

PET scan that directly identifies tau protein deposition (i.e. tangles) in the brain - only approved for research use in the past two years, and still quite expensive.

CSF

Cerebrospinal fluid. Tau protein can accumulate here, but relationship to levels in the brain is not simple, so estimates may be unreliable.

SUVr

Standardized Uptake Value Ratio - the degree of radiotracer uptake in a target region of interest with respect to a reference region.

FTD

Frontotemporal dementia, a non-Alzheimer's cause of neurodegeneration and cognitive impairment.

DLB

Dementia with Lewy Bodies, another non-Alzheimer's source of neurodegeneration and cognitive impairment.