Stats 211 Final Report

Sacha Robbins

3/21/2018

Abstract

Teams focused on researching Alzheimers disease (AD) prevention need its participants to commit to longitudinal studies oftentimes with a study partner, someone who spends a significant amount of time with the participant. Our questions of interest are: Is there an association between type of study partner and the odds of completing the long-term study? Does this association change at different levels of baseline-assessed dementia? Which patients are most likely to complete long-term trials? The data used to address these questions are demographics of the participants and their study partners, medical history of the participant, responses to cognitive questions, ratings and scores from functional and cognitive assessments, and an indicator of whether or not the participant completed the study. Our main conclusion is that our data provide strong evidence that supports no association between type of study partner and the odds of completing a long-term study. Furthermore, this null association does not change at different levels of baseline-assessed dementia. Potential confounding variables not measured in the data, such as reasons for incompletion or whether the type of study partner changed during the course of the study, may have led us to different conclusions.

Background/Introduction

Preventing Alzheimers disease is of great scientific interest. Common research in this field is longitudinal observational studies. And the common challenge with that is whether or not a subject stays in the study. If we can learn more about what characteristics of subject is associated with their commitment to study, we may be able overcome said challenge. It is hypothesized that the type of study partner might be related to the odds of a subject staying in a study. With the limitations of the data we have, there is no evidence that such an association truly exists. However, a good prediction model may still be of use; a prediction model can be used to predict the odds of whether or not a subject will stay. For those who are deemed less likely to stay, researchers can figure out ways to keep them involved.

The first fact we considered was the nature of our response variable, an indicator of whether or not a participant completed the study. Since it is a binary variable, we made an assumption that the values of this response variable are independent and identically distributed Bernoulli with probability of success (i.e. completing the study) μ and variance $\mu(1-\mu)$.

In this report, we aim to answer these questions: Is there an association between type of study partner and the odds of completing the long-term study? Does this association change at different levels of baseline-assessed dementia? Which patients are most likely to complete long-term trials? Based on our data, we will arrive at unsatisfactory responses to the first two questions. And the last question will be answered with a prediction model.

Our analysis aims to address two different types of questions: one of association and the other of prediction. When estimating associations, it is important to first brainstorm potential adjustment variables before even looking at the dataset. After looking at the dataset, limitations become a realistic concern. Not adjusting for confounding variables could affect the consistency and precision of estimated parameters, which consequently would affect inferences and conclusions. Alternatively, the process of building a prediction model differs in its goals: selecting a model that performs better than other models and assessing its performance in predicting new data. At this stage of the analysis, we mainly used “step” R-function in package “MASS” by B.D. Ripley. We also used the ROC software package by Michael C. Sachs, specifically the plotROC R-function.

Methods

Source of the Data

Our data was collected from an four-year observational study from 39 locations across the United States of N=644 participants paired with their study partners. Data regarding the participants medical history, cognitive status, demographics (also including those of the study partners), were collected at the start of the study. The Clinical Dementia Rating Scale (CDRS) and Modified Mini-Mental State Exam (MMMSE) were used to assess cognitive status. Responses from participants and study partners to questions from the Cognitive Function Instrument (CFI), a detection for early shifts in functional and cognitive abilities, were also included in the data. It is also known whether or not a participant completed the 48-month study, which we will use as our response variable.

Demographic information about the participant and the study partner was provided: age, gender, ethnicity, and education in years. Interestingly, there was 8.5% missing data (55/644) in the study partner's age; 14 of the 55 who had missing data here were study partners whose participants did not complete the study and 41 of the 55 had their participants complete the study. This is an interesting difference, but then again, 65% of the participants overall completed the study. This missing data should not intervene with our aims.

Ethnicity of the participant only had two levels ("white" and "other") unlike the ethnicity of the study partner with 5 levels. For cleaner comparison, we collapsed the later to fit the former's category levels. Participant's medical history in the form of indicator variables was also provided: alcohol abuse, drug abuse, smoking, cardiovascular disease, and cancer. Surprisingly, there was no missing data here, however, we were skeptical about response bias, as is typical with questions of this nature.

Participant's baseline results from cognitive and functional assessments were included in the dataset. Assessments were the Clinical Dementia Rating Scale or CDRS (0 = Normal, 0.5 = Very Mild Dementia, 1 = Mild Dementia, 2 = Moderate Dementia, 3 = Severe Dementia) and the Modified Mini-Mental State Exam (MMMSE) that has scores ranging from 0 to 100 with greater values signifying better cognition. The dataset also includes participant and study partner responses from six questions in the Cognitive Function Instrument (CFI), used to assess early stages of functional and cognitive decline. Study partners needed to answer these questions based on what they observed from the participants. Here, we noticed some missing data, as little as 2 and at most 5 participants had these variables blank. Table 1 shows information on the participant's with at least 1 missing entry out of the 12 CFI questions (6 similar questions for the participant and his/her study partner). We will continue with complete-case analysis by excluding those 6 observations.

Table 1: Participant's Info from Those with Missing Data

ID	Age	Gender (Male=1)	Ethnicity	Education	Study Partner's Age	No. of Missing Data
1366	84	0	White	19	43	6
1413	79	1	White	18	73	1
1450	84	1	White	20	NA	7
1734	83	0	White	12	NA	7
1447	79	0	Other	12	74	12
1744	76	0	Other	5	NA	13

Lastly, we also have data on the type of study partner (our predictor of interest), the estimated measure of weekly time spent with the participant, whether or not they lived together, and if the study partner noticed some decline in his/her memory.

For our association analysis, it is important to brainstorm potential confounding variables, causally related to our response variable and associated with our predictor of interest. Our predictor of interest is the type of study partner. These were the potential confounding variables that came to mind: survival, location changes, socioeconomic status, depression, study partner's gender, participant's gender, and study partner's level of involvement.

Out of these potential confounders, some were not measured in the dataset. Whether or not a participant stays alive throughout the study is an obvious confounder to consider. Unfortunately, our data did not provide reasons why participants did not complete the study. If we had data on survival, it would be considered a precision variable since it does not seem related to our predictor of interest. If a participant moves locations during the study, they are less likely to complete the study, especially if transportation is an issue. Our dataset does not include this information. However, location would also be associated with our predictor of interest, making it a confounding variable we cannot adjust for.

However, some of our variables in the dataset do measure the remaining potential confounders. Socioeconomic status can be implied from a few measured variables in the dataset: education and ethnicity. Higher socioeconomic status is usually associated with longer marriages, stronger health, higher education, and of "white" ethnicity. It is reasonable to assume that individuals from this subpopulation are more inclined to stick with long-term commitments. Therefore, it is reasonable to think that a more educated participant is likely to survive and complete the study.

If a participant suffers from depression, he or she may be less likely to complete the study. People with depression usually suffer from low motivation. Two CFI questions in our dataset may help indicate depression. However, we will choose not to adjust for these variables since it is possible that a participant's depression may affect the type of study partner. This would make variables that measure depression a mediator in our model; we tend to avoid adjusting for mediators since they may derail us from the question of interest.

It's generally assumed that there are more female caretakers. But it is also worth exploring the idea that if you have a female caretaker, you may have a higher chance of completing the study. Continuing with possible gender effects, it is also generally assumed that more females would be willing to commit to longer studies involving their health. And, of course, if a study partner is highly involved with the participant, he or she may be more likely to complete the study.

Statistical Methods

Association Model

We started by fitting an unadjusted generalized logistic regression model of response variable and predictor of interest. After conducting an $\alpha = 0.05$ level Wald test, all regression coefficient estimates except the intercept had a p-value less than α, which shows evidence that our predictor of interest may not have an association with our response variable.

Next, we slowly adjusted for all the confounders measured : participant's ethnicity, participant's education, and study partner's level of mental decline (see descriptive statistics for more detail). At each step we observed regression parameter estimates, confidence intervals, and p-values. After each additional adjustment, these figures for our predictor of interest did change, but still deemed mostly statistically insignificant. The estimated regression coefficients that proved significant were those of our confounders. At each step, we also compared each model with its previous "nested/reduced" model using a likelihood ratio test (LRT). The null hypothesis when comparing the nested model M_0 with model M_1 is that all parameters in model M_1 that are not in model M_0 are zero (i.e. statistically insignificant, M_0 is a better fit). The alternative hypothesis is that at least one parameter in model M_1 that is not in model M_0 is nonzero (i.e. significant, M_1 is a better fit).

Since LRT depends on our variance assumption, it's important to check whether that assumption is reasonable with empirical evidence. To check if the variance assumption for binary data ($V(\mu) = \mu(1 - \mu)$) is reasonable, we plotted the pearson-residuals-squared against fitted probabilities of our association model and prediction model. After using a smoother that essentially averages over all the x values, we obtained another curve. If our variance assumption is true, then that curve should be approximately a horizontal line at 1. We did not run into any issues, proving that our variance assumption for binary data was adequate.

Next, we assessed whether or not the association between study completion and type of study partner changed for different levels of CDRS Baseline Score (i.e. effect modifier). First, we ploted a histogram of these reported dementia ratings stratified by those who completed the study and those who did not. Due to risk of sparcity in the data, we regrouped the results into four categories (versus the original six): 0 = normal, 0.5 = very mild dementia, 1 = moderate dementia, 1+ = moderate to severe dementia). There seems to be a slight interaction between completion and these baseline dementia ratings.

Although there was not much evidence, we fitted models with CDRS Baseline Score as a continuous variable, as a main effect, and as an interaction with type of study partner. We also fit it with the original unadjusted model and the final association model with the other confounding variables. We even fit the model without the predictor of interest. All models rejected the null hypothesis for wald tests. In conclusion, CDRS Baseline Score is not an effect modifier.

Prediction Model

Lastly, we built a prediction model. We needed to deal with missing data before conducting model selection algorithms. As mentioned before, there were six observations that we elmininated in order to do a complete-case study. However, the data on the study partner's age had 8.9% missing data. For simplicity, we eliminated the variable entirely, leaving 29 candidate covariates.

We used stepwise regression algorithms with AIC and BIC criterions in combinations of forward, backward, and both direction. Out of the six combinations, there was only three unique models in the end to compare.

We conducted a goodness of fit tests (Pearson Chi-Squared) in order visualize their differences. Hosmer and Lemeshow suggests a post-hoc grouping of data to evaluate the goodness of fit. Table 4 shows a summary of each model that includes method, AIC/BIC score, number of covariates in the model, chi-squared statistic, degrees of freedom, and p-values.

In order to choose the final model, we compared the nested models using LRT. Model 1 was the largest model, so first we compared that with our second largest model, Model 3. The LRT test resulted in a p-value of 0.9921, which means that under an $\alpha = 0.05$, we fail to reject the null. Our reduced model, Model 3, is a better fit. We do another LRT between Model 3 and nested Model 2, which results in a p-value of 0.0074. So, we reject the null: Model 2 is not as good of a fit as Model 3. So, we accepted Model 3 as our final prediction model.

Lastly, we analyzed the model's predictive performance by plotting a receiver operating characteristic (ROC), which plots the sensitivity and specificity of the model. Figure 3 shows the ROC curve. The area under the curve (AUC) is a measurement that indicates predictive performance. The model's AUC was approximately 0.70 where 0.5 is as good as guessing and 1 is perfect predicability.

Results

Descriptive Statistics

Table 2 and 3 provides an overview of the data. The tables decompose the data into three groups: all participants, those who completed the study, and those who did not. From these tables, we can identify potential confounding variables.

Table 2 shows that participants who failed to complete the study were more likely to have history of cardiovascular disease, which means a higher risk for myocardial infarction. Survival is a strong potential confounder. It would be helpful for our association model to know how many of those who did not complete the study died.

Table 2 shows that there is a slightly higher education average in those who completed the study compared with those who did not. Figure 1 investigates this association between a participant's education and type of study partner with density plots. If you have a college degree (> 16 years of education), it is more likely that a friend or spouse would be your study partner than a child or "other" (e.g. hired caretaker). If you have a high school diploma (~12 years of education), it is more likely that a friend or "other" would be your study partner than a spouse or child. So, participant's education is a confounding variable and should be adjusted for.

Table 2 shows a strong difference in participant's ethnicity between those who completed and those who do not. Figure 2 shows a bar chart that stratifies participant's ethnicity based on type of study partner. Notably, white participants are much more likely to have their spouse as their study partner compared with other participants. So, participant's ethnicity is a confounding variable and should also be adjusted for.

There are a few surprising finds that seem to contrast with our original assumption: females tend to be more involved with their health than males. Table 3 shows that there is not much of an effect of study partner's gender or participant's gender on whether or not the participant completes the study. A possible reason for this surprising result is that females survive longer than males, making females more likely to survive the study. And males are usually tended by females spouses, who probably hold them more accountable to commiting to such a study. These effects probably balance each other out so that there appears to be marginally no effect of gender across completion.

Table 3 shows data on whether or not a partner lived with the participant and how many days per week he/she was involved with the participant. Another surprising result: there does not seem to be much difference in the data between those who complete and those who do not. So, we will not adjust for it in our association model. However, Table 3 does show possible interaction between study partner's mental decline and the response variable. This may be another potential confounder. It's reasonable to think that if your study partner is also developing dementia, the odds to sticking through a 4-year study may decrease. So, we adjusted for it in our association model.

Figure 1: Density Curves Stratified by Type of Study Partner

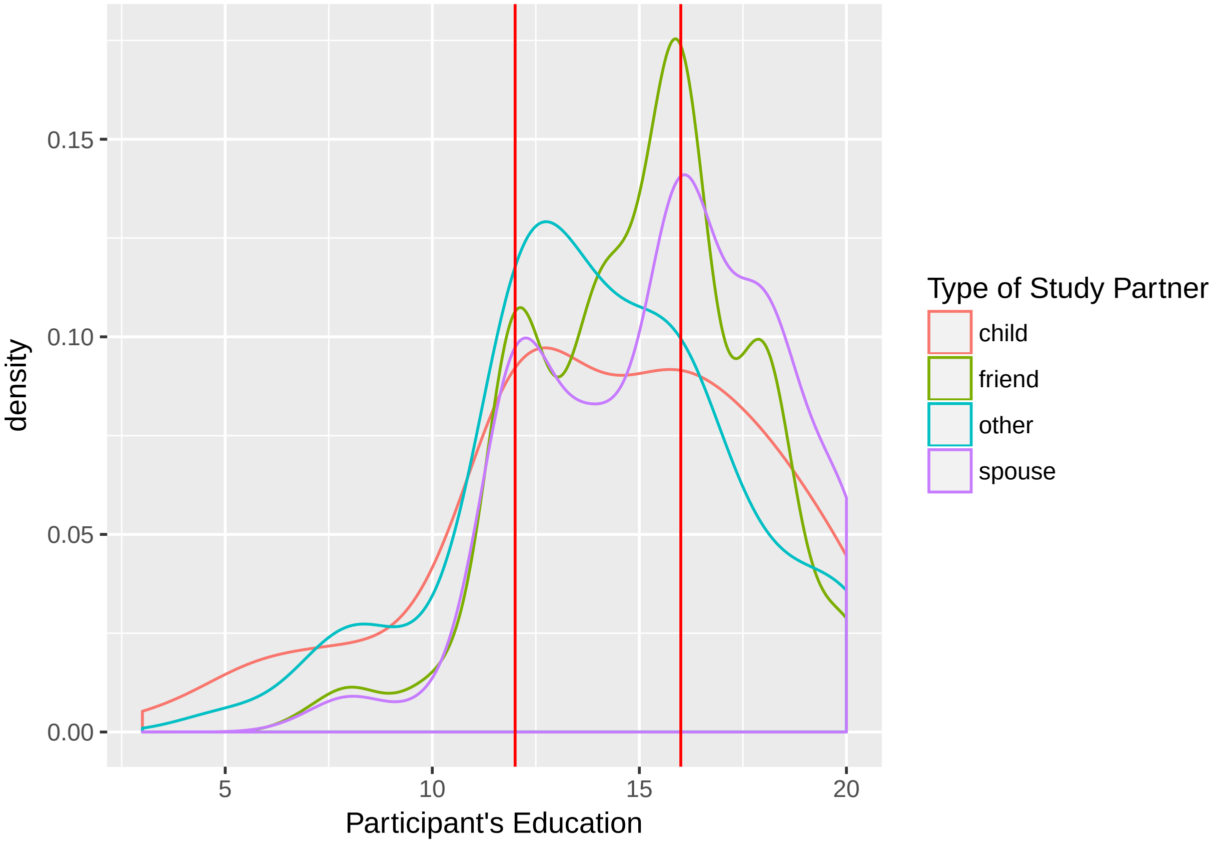


Figure 2: Histogram of Participant's Ethnicity Across Type of Study Partner

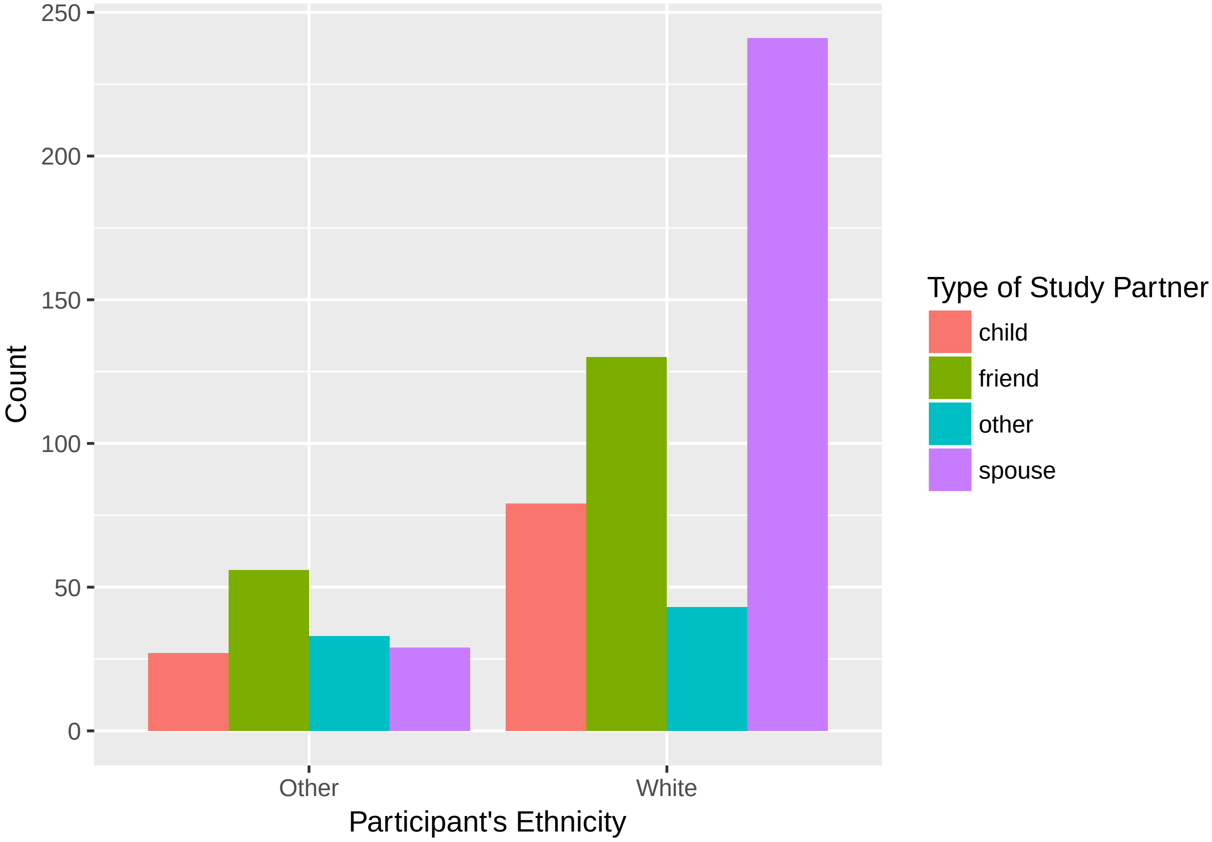


Table 2: Summary Statistics for Study Participants in Averages or Proportions

Characteristic	Overall	Completed	Failed to Complete
Age	79.51	79.1	80.25
Gender	male: 41.8%	male: 41.1%	male: 42.9%
Ethnicity	white: 77.2% other: 22.8%	white: 81.6% other: 18.4%	white: 69.0% other: 31.0%
Education in years	14.96	15.28	14.36
Alcohol Abuse History (Yes)	3.4%	3.1%	4.0%
Drug Abuse History (Yes)	0%	0%	~0%
Smoking History (Yes)	38.8%	38.0%	40.3%
Cardiovascular Disease History (Yes)	65.5%	63.4%	69.5%
Cancer History (Yes)	3.1%	25.3%	26.1%
Baseline MMMSE (0-100)	95.28	95.85	94.23
Baseline CDRS	0.31	0.25	0.42

Table 3: Summary Statistics for Study Partners in Averages or Proportions

Characteristic	Overall	Completed	Failed to Complete
Relation to Participant	child: 16.8% friend: 29.3% other: 12.0% spouse: 42.0%	child: 17.9% friend: 28.5% other: 10.5% spouse: 43.1%	child: 14.6% friend: 31.0% other: 14.6% spouse: 40.0%
Age	68.68	68.39	69.19
Gender	female: 77.8%	female: 74.0%	female: 75.7%
Ethnicity	white: 74.5%	white: 81.6%	white: 70.8%
Education in years	14.89	15.11	14.48
Days per Week with Participant	5.6	5.6	5.7
Live with Participant? (Yes)	48.9%	49.3%	48.2%
Show signs of Mental Decline? (Yes)	30.1%	27.8%	36.3%

Models

Unadjusted Model:

$$E[I_{complete\ study}] = \beta_0 + \beta_1 I_{partner=friend} + \beta_2 I_{partner=other} + \beta_3 I_{partner=spouse}$$

Final Model for Association:

$$E[I_{complete\ study}] = \beta_0 + \beta_1 I_{partner=friend} + \beta_2 I_{partner=other} + \beta_3 I_{partner=spouse}$$

$$+\beta_5 I_{white} + \beta_6 * Education + \beta_7 I_{partner'smental\ decline}$$

Table 4: Summary of Fitted Model for Association

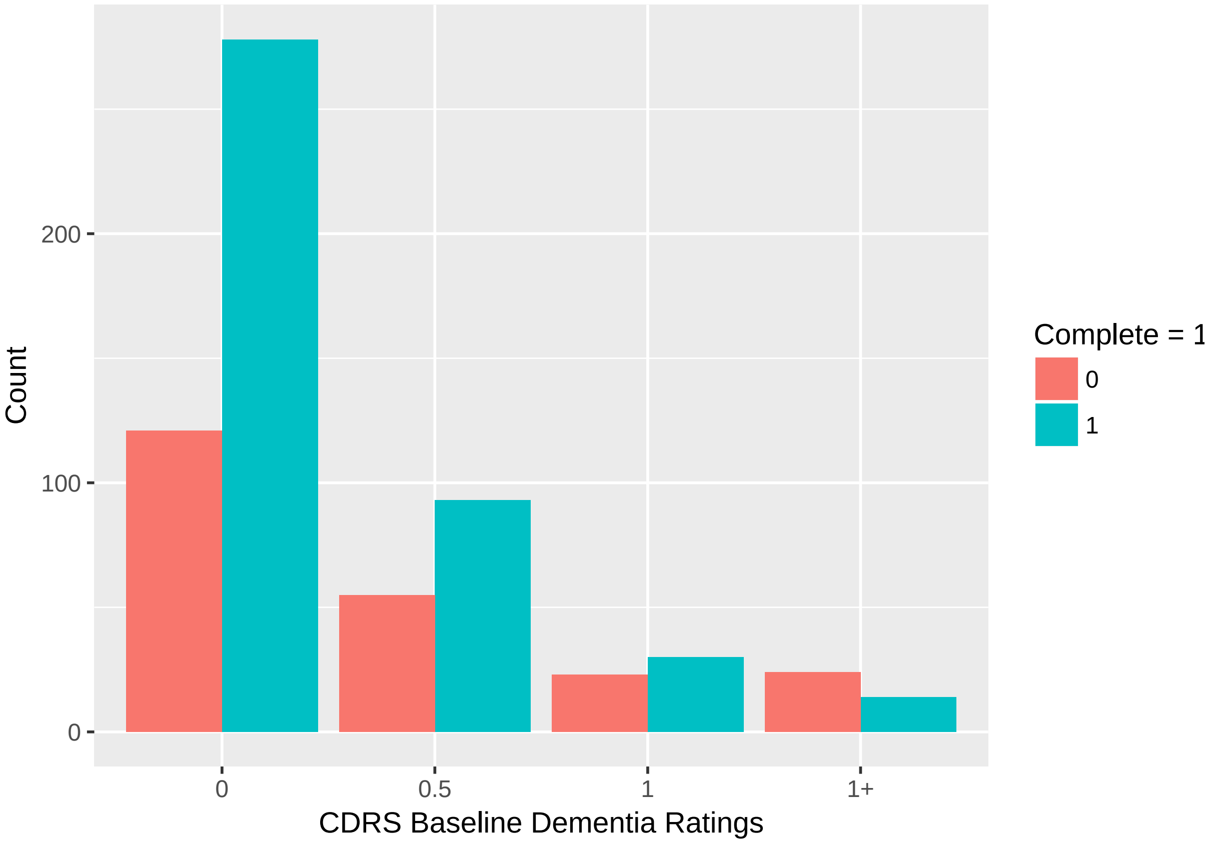
Covariate	exp(Est)	95% Confidence Interval	P-value
Intercept	0.5217	(0.2193, 1.2412)	0.1412
Indicator Study partner, Friend	0.7334	(0.4341, 1.2390)	0.2465
Indicator Study partner, Other	0.6616	(0.3520, 1.2436)	0.1995
Indicator Study partner, Spouse	0.7915	(0.4735, 1.3232)	0.3725
Participant's Ethnicity, White	1.7542	(1.1653, 2.6407)	0.0071
Participant's Education	1.0868	(1.0267, 1.1504)	0.0041
Indicator, Study partner mental decline, Yes	0.6189	(0.4312, 0.8882)	0.0092

Each of these p-values are based on $\alpha = 0.05$ level Wald test. Since the first four regression coefficients have p-values that fall in the rejection region on a standard normal distribution, we will interpret the other estimated parameters.

We estimate that participants who are white ethnically have 75.4% higher odds in completing the 4-year study than those labeled “other” ethnically and all other covariates held constant (95% confidence interval: (1.1653, 2.6407)).

We estimate that for every 1-year increase in education, the difference in the odds of completing the 4-year study comparing two populations increases by 8.9% and all other covariates held constant (95% confidence interval: (1.0267, 1.1504)).

We estimate that participants whose study partners reported a recently developed mental decline at the start of the study are 38.1% less likely to complete the 4-year study than those whose study partners reported no mental decline and all other covariates held constant (95% confidence interval: (0.4312, 0.8882)).



Prediction Model

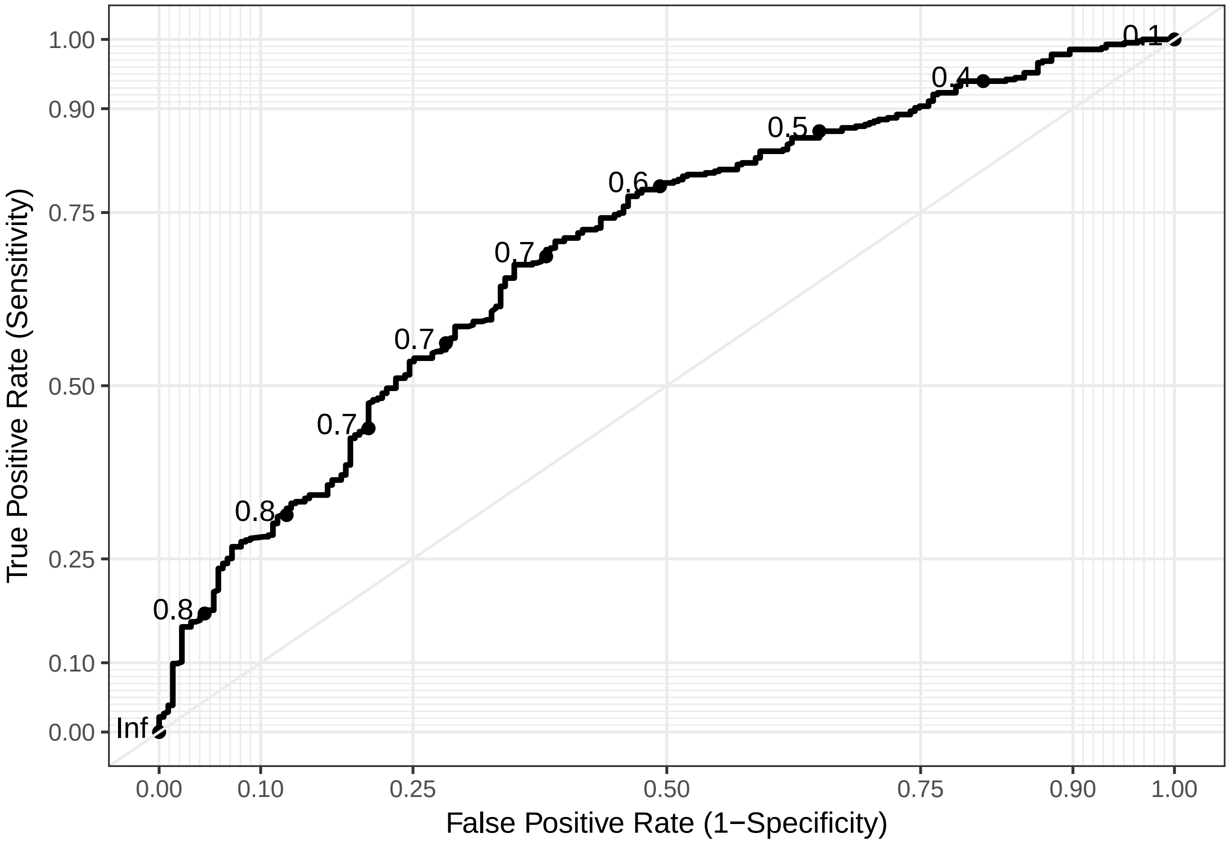
$$E[I_{complete\ study}] = \beta_0 + \beta_1 * Participant'sAge + \beta_2 I_{white} + \beta_3 * Education$$
$$+\beta_4 I_{MMMScore} + \beta_6 * I_{mpsocial} + \beta_7 * I_{mpapplia}$$
$$+\beta_8 * I_{mshelp} + \beta_9 I_{partner'smental\ decline}$$

Note: Mpsocial is a binary variable to represents a participant's response to whether or not they have been feeling social. Mpapplia is a binary variable to represents a participant's response to whether or not they need help with home appliances. Mshelp is a binary variable to represents a partner's response to whether or not the participant needs help from others with appointments, remembering occasions, etc.

Table 5: Summary of Best Stepwise Regression Models

	Method	AIC/BIC	Number of Covariates	chi-sq statistic	df	p-Value
goftestAICf	AIC,forward	808.1543	29	12.656	8	0.124
goftestBICf	BIC,forward	808.1543	29	12.656	8	0.124
goftestBICbk	BIC,backward	777.9904	4	11.59	8	0.17
goftestBICboth	BIC,both	777.9904	4	11.59	8	0.17
goftestAICbk	AIC,backward	772.0239	8	19.871	8	0.011
goftestAICboth	AIC,both	772.0239	8	19.871	8	0.011

Figure 4: ROC Curve

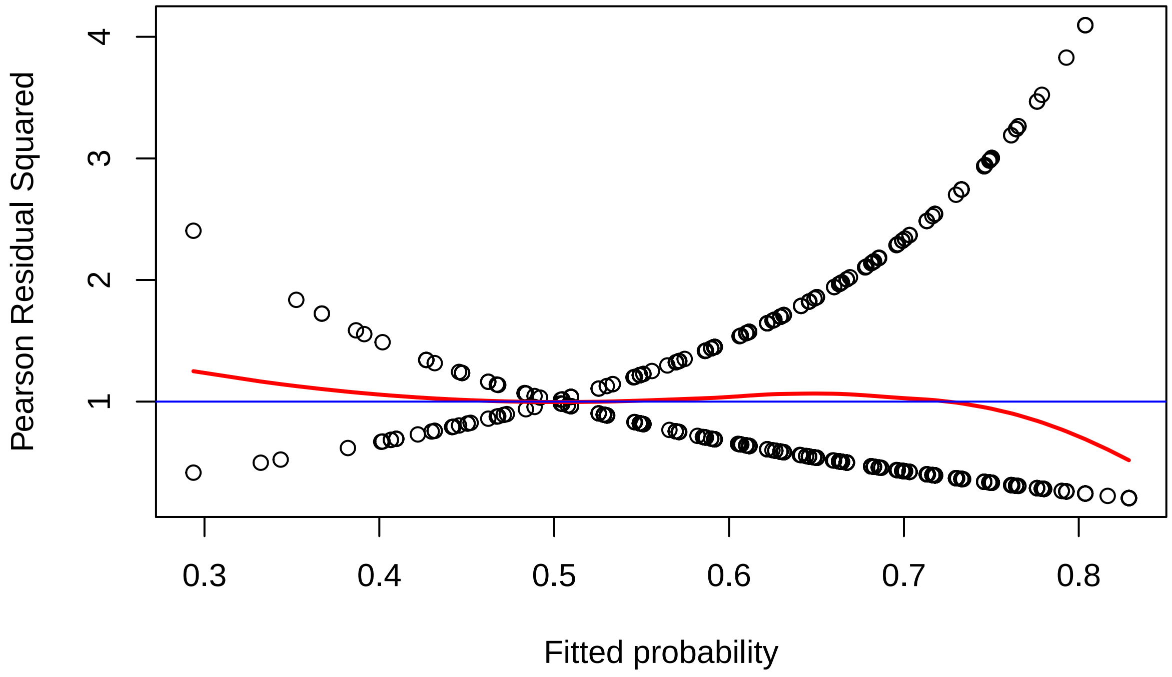


Discussion

In this report, we concluded that due to our limitations of our dataset, we could not find evidence of an association between type of study partner and odds of completing a study. Furthermore, this does not change when considering CRDS baseline dementia ratings as an effect modifier. We were able to create a prediction model that fit our data adequately. If we had data that noted reasons why some people did not complete the study, it would help us understand and prevent such challenges. Then, maybe we can make faster progress in Alzheimers prevention research.

Appendix

Figure 5: Diagnostic for Variance Assumption


Pearson Residual Squared
1
2
3
4
0.3
0.4
0.5
0.6
0.7
0.8
Fitted probability