1. DATA SUMMARY

1.1 Sample Overview

This data was collected from 997 patients in Greater Manchester, and the analysis aim at exploring the dataset to investigate the attributed risk factors of dementia. The dataset includes variables such as Age, HDL (High density Lipoprotein Cholesterol Level), Smoking status, Sleep duration, physical activity, irritability, True dementia status and Dementia diagnosis. No duplicate data or missing value was found in the dataset.

1.2 Dementia Prevalence

Of the 997 individuals who participated, I found out from the data that 52 participants have confirmed Dementia, which makes up about 5.22% of the total population, while 945 participants doesn't have dementia, of which makes up about 94.78% of the sample population.

1.3 Descriptive Statistics

Table 1.1: Summary statistics for the study Variables.

Variable	Category	n, %, mean ± sd	Range
Sample Size		997	
Age	years	55.4 ± 8.5	40 - 70
HDL	mmol/L	1.43 ± 0.38	0.23 - 3.84
Sleep	hours/day	7.1 ± 1.2	3 - 10
BMI	Kg/m²	26.8 ± 1.9	21.6 - 32.7
Smoking(Non)		584 (58.58 %)	
Smoking(Ex)		343 (34.40 %)	
Smoking (Current)		70 (7.02 %)	
Irritability	Yes	288 (28.89 %)	
Physical Activity	>15mins/day	963 (96. 59 %)	
Dementia	Yes	52 (5.22 %)	

1.4 Key Observations

Age Distribution: The Age range is from 40-70 with mean of 55.4 years, which implies there is a risk of dementia increase among the middle age to older Adults.

Clinical measurements: Measurements such as HDL has an average of 1.43mmol/L with a variation of (0.23-3.84)mmol/L, BMI has a mean of 26.8kg/m² which could be categorized in overweight category considering the range (21.6 - 32.7).

Lifestyle Factors: We have quite an high value of non-smokers of (58.58%) from our sample population, followed by the ex-smokers with a considerable percentage of (34.40%) with just (7.02%0 currently smoking. Also nearly all participants (96.59%) reported doing more than 15 minutes physical exercise per day, also the sleep duration is averaged at 7.1 hours within recommended ranges.

Behavioral Significance: 28.89% of the population reported having irritability

Data Quality: Overall, there was no missing values in the dataset and there are no duplicates, which is a good indicator of a strong data quality, even though relying on self report for behavioral factors could be biased and might affect analysis.

2. DO AVERAGE LEVELS OF HIGH-DENSITY LIPOPROTEIN DIFFER ACROSS SMOKING GROUPS?

2.1 Research Question and Hypothesis

This analysis investigates whether the average HDL levels differ across smoking status groups as seen in Table 1.1.

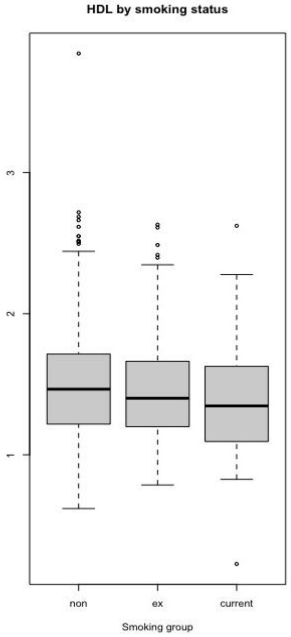
Null Hypothesis(H_0 : $\mu_1 = \mu_2 = \mu_3$)

Alternate Hypothesis (H_1 :At least one μ_i differs)

2.2 Exploratory Analysis.

I calculated a summary statistics for the HDL levels across all the smoking groups and I created a box-plot to visualize the distribution of the HDL levels by the Smoking status, and also a table to show the descriptive statistics for HDL by smoking Status.

Figure 1.1 & Table 2.1



Smoking status	n	Mean (mmol/L)	SD	Median	IQR	Min	Max
Non-Smokers	584	1.49	0.39	1.46	0.50	0.62	3.84
Ex-Smokers	343	1.46	0.36	1.40	0.46	0.79	2.63
Current-Smokers	70	1.38	0.39	1.35	0.51	0.23	2.62

2.3 Statistics Test Selection

A one-way ANOVA test is used as the test best fitted to tackle this question because

- The Variable HDL is a continuous variable
- The Variable which is the predictor (Smoking Status), is a categorical variable with three independent groups.
- The question will require us to compare means over against more than 2 groups.

2.4 Assumption Checking

Normality: A histogram plot was done to view how normally the data were distributed, our result shows an agreeable normally distributed data, but to further back this up, a Shapiro-Wilk test was conducted for each group and the result indicate a slight deviation from normality in non_smokers ($W = 0.964$, $p < 0.001$) and ex-smokers ($W = 0.969$, $p < 0.001$), while the current smoker shows normality ($W = 0.972$, $p = 0.115$), but ANOVA could handle minor normality deviations especially with larger sample size, making it the best for this analysis

Homogeneity of Variance: Levene's test confirmed homogeneity of variances across the groups, ($F(2,994)=0.67$,$P=0.514$). The largest to smallest standard deviation ratio was 1.08, which is below the rule of thumb 2.0 benchmark to further support equal variance.

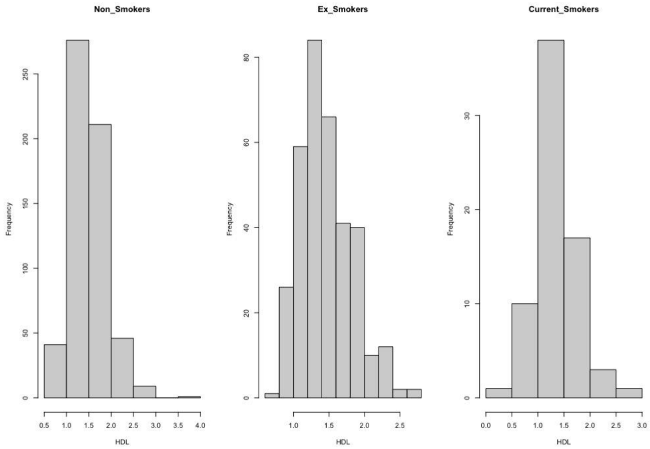


Figure 1.2: ***histogram of HDL across smoking groups***

2.5 ANOVA RESULTS

One-way ANOVA revealed a significant difference in HDL level across the smoking groups

```
> ## Run Anova test
> anova_test <- aov(HDL ~ Smoking, data = dim_study)
> summary(anova_test)
            Df Sum Sq Mean Sq F value Pr(>F)
Smoking       2    0.9  0.4476    3.16 0.0429 *
Residuals   994  140.8  0.1416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.6 Post -Hoc Analysis

Tukey's post-Hoc HSD test was conducted to identify specific pairwise differences:

```
> TukeyHSD(anova_test)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = HDL ~ Smoking, data = dim_study)

$Smoking
                  diff         lwr           upr     p adj
ex-non      -0.03334793 -0.09344258  0.0267467250 0.3940731
current-non -0.11264530 -0.22437867 -0.0009119397 0.0476351
current-ex  -0.07929738 -0.19515601  0.0365612538 0.2433131
```

Current smoker shows a lower HDL levels compared to non-smokers and there is no Significant difference between ex-smokers and non-smokers.

2.7 Interpretations and Conclusions:

We reject the Null Hypothesis. There is difference in HDL levels across smoking groups.The difference in current smokers and non-smokers has a clinical significance, because current smoking is a known factor to suppress HDL levels through inflammatory mechanism and altered lipid metabolism.

Ex-smokers' HDL levels were not so much different from non-smokers as seen from the R code above, which suggests recovery of HDL level after quitting smoking. So quitting smoking habit is an intervention for cardio vascular risk improvement.

Lastly, These findings further reiterate the detrimental effect of active smoking on lipid profiles, and quitting it will help improve the HDL level return to its normal level. This is a strong indicator for Cardio Vascular health interventions and dementia prevention strategy on the basis of the established links between CV health and Cognitive function.

3. Exploring Risk Factors of Dementia.

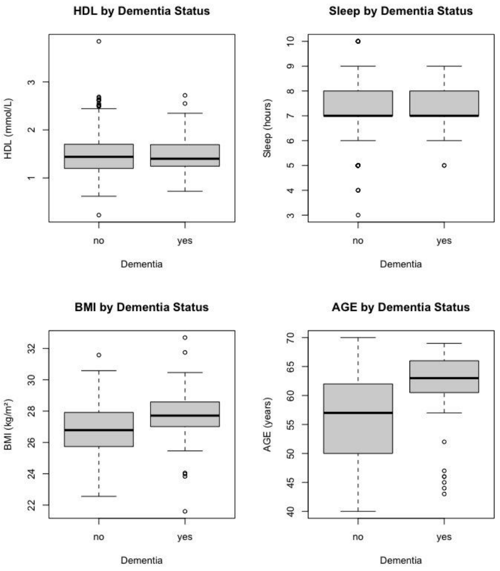
This analysis investigates potential risk factors linked with dementia using Logistic Regression modeling to find which of the variables significantly predict dementia occurrence in our sample population.

3.1 Methods

As earlier stated, 52 (5.2%) of our study population have true dementia. Initial exploratory analysis showed potential risk factors which include HDL, Cholesterol, Sleep duration, BMI, Age, Irritability, Smoking status and physical activity levels.

3.2 Descriptive Analysis

The mean of the continuous variables are presented in the Table 3 below. Also a visual exploration using box-plot shows a relationship between dementia status and some variables particularly Age and BMI.



Variables	**No Dementia (n=945)**	**Dementia (n=52)**
HDL(mmol/L)	1.47	1.52
Sleep (hours)	7.15	7.17
BMI (kg/m²)	26.8	27.7
Age (years)	55.8	61.4

Correlation analysis among these continuous variables shows from low to moderate correlations ($r < 0.3$), suggesting minute multicollinearity worries. The strongest of the correlation was between Sleep and BMI ($r=0.280$).

3.3 Statistical Modeling:

A logistic regression model is used to identify risk factors associated with Dementia. The initial predictors were seven, which were HDL, Sleep, BMI, Age, Irritability, Smoking status, and Physical activity. Backward stepwise selection based on Akaike Information Criterion(AIC) was used to identify the most fitting model.

The stepwise selection in sequence removed the non-significant predictors which are HDL, Sleep, Physical activity and Irritability. Which gave us a final model with three predictors which are:BMI, Age and Smoking status (AIC=365.0, compared to the initial AIC which was 370.9).

3.4 Results

Final Model Summary

The final logistic regression model shows that Age, BMI and Smoking Status were significant predictors of Dementia risk.

```
model <- glm(Dementia ~ HDL + Sleep + BMI + Age + Irritability + Smoking +
             Physical_activity_15mins, data = dim_study, family = binomial())

summary(model)
Call:
glm(formula = Dementia ~ HDL + Sleep + BMI + Age + Irritability +
    Smoking + Physical_activity_15mins, family = binomial(),
    data = dim_study)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -21.29075    3.42440  -6.217 5.06e-10 ***
HDL                            0.01011    0.37968   0.027  0.97877
Sleep                         -0.19511    0.15923  -1.225  0.22046
BMI                            0.43383    0.10682   4.061 4.88e-05 ***
Age                            0.12518    0.02529   4.950 7.44e-07 ***
Irritability-yes              -0.21933    0.35160  -0.624  0.53275
Smokingex                      0.02597    0.32722   0.079  0.93673
Smoking-current                1.47658    0.46177   3.198  0.00139 **
Physical_activity_15minsgreater   0.46844    1.06411   0.440  0.65978
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 408.4  on 996  degrees of freedom
Residual deviance: 352.9  on 988  degrees of freedom
AIC: 370.9

Number of Fisher Scoring iterations: 7

-- sorted_model <- step(model, direction = "backward")

 Df Deviance    AIC
<none>        355.00 365.00
- Smoking  2   363.96 369.96
- BMI      1   371.84 379.84
- Age      1   387.08 395.08
```

Predictor	Co-efficient	SE	z-value	P-value	OR	95% CI
Intercept	-21.35	3.27	-6.52	<0.001	-	-
BMI	0.40	0.10	4.00	<0.001	1.50	1.23- 1.83
Age	0.12	0.02	4.99	<0.001	1.13	1.08 - 1.19
Ex-smoking	-0.02	0.32	-0.05	0.960	0.98	0.51 - 1.85
Current smoking	1.45	0.46	3.19	0.001	4.28	1.66-10.17

Model fit: Null deviance = 408.4 (df=996); Residual deviance = 355.0 (df=992): AIC=365.0

3.5 BMI: The body mass index appeared to as a strong risk factor for dementia with (OR=1.50, 95%CI:1.23-1.83, p <0.001). Each increase in one unit BMI is a 50% increase in the dementia odds after adjusting for age and smoking status. This is clinically meaningful and significant.

Smoking Status: The smoking status shows a strong effect on dementia risk.
Current_smokers showed the strongest association with dementia with (OR=4.28, 95%CI:1.66-10.17, p=0.001). Current smokers have over 4 times the odd of dementia compared to non-smokers thus representing a 32% increase in risk after considering the effect of BMI and Age also.
Ex-Smokers showed no significant difference from non-smokers (OR = 0.98%, 95%CI:0.51-1.85, P=0.960). The OR which appears to be near 1.0, shows that people who have quit smoking tend to return to the little or no level of dementia risk.

3.6 DISCUSSION

This analysis observed three significant risk factors for dementia risk, which are BMI, Age and current Smoking Status.
Age came out as a robust predictor and has a consistent role as a primary UN-modifiable risk factor for dementia. This finding points towards neuron-degeneration and reduced cognitive function over the human lifespan.
BMI showed a clear positive association with dementia risk, which supports the growing evidence that links obesity to the cognitive decline. This presents a modifiable risk factors that can be adjusted with lifestyle and habits intervention.
Smoking Status provided us with the strongest evidence for intervention proposals. The contrast between current smokers and ex-smokers shows that quitting smoking habit will reduce Dementia risk. The toxicity of smoking can be reversed or amended for upon quitting.
The absence of significant effect for HDL, Sleep , physical activity and irritability deserves some extra consideration. Perhaps these variables have weaker independent effects when accounting for BMI, Age AND Smoking status. Measurement limitation could also be the reason why there was detection of smaller effects.
We will properly address the limitations attached to this analysis at the question 7 of this report, where we can check for plausible reasons why we might have some little deviation from a 100% accuracy.

4. INVESTIGATING THE REALIBILITY OF DEMENTIA DIAGNOSIS

This analysis evaluates the reliability of Dementia diagnosis by comparing the diagnosis against the actual dementia status which will serve as our reference standard using measures of statistical accuracy.

4.1 **Methods:** A 2 by 2 contigency table was created to cross-tabulate the clinical diagnosis (Dementia_diag) against the True dementia status (Dementia) for all the 997 participants. From this confusin matrix, we were able conduct a accuracy test, specificity, positive predictive value (PPV), and a negative predictive value (NPV).
4.2

4.2 Results

Confusion Matrix
The table 4 below shows a confusion matrix comparing the clinical diagnosis with the actual dementia status

	Dementia_Diag (NO)	**Dementia_Diag (YES)**	**Total**
True-Dementia (NO)	927 (TN)	18 (FP)	945
True-Dementia (YES)	4 (FN)	48 (TP)	52

Total	931	66	997

Where TN is True Negative, TP is True Positive, FP is False Positive, FN is False Negative.

4.3 Diagnostic Performance Metrics:

Table 2 below summarize the key performances tested for the dementia diagnosis.

Metric	Formula	Vaalue	95% CI	Interpretation
Accuracy	(TP+TN) / Total	97.79%	-	Overall correctness of diagnosis
Sensitivity	TP/(TP + FN)	92.31%	81.5 - 97.9%	Ability to correctly identify dementia cases
Specificity	TN / (TN + FP)	98.10%	97.1 - 98.8%	Ability to correctly identify non-dementia cases
PPV	TP/(TP + FP)	72.73%	60.4-83.0%	Accuracy when diagnosis is positive
NPV	TN/(TN + FN)	99.57%	98.9-99.9%	Accuracy when diagnosis is negative

4.4 Interpretation in detail:

The overall accuracy of the diagnosis is 97.79% which implies that 975 out of 997 diagnosis were accurately captured. This high accuracy shows that the dementia diagnosis data is-very reliable for this population study.

Sensitivity (92.31%), 52 participants have an actual dementia, and of those 52, 48 were accurately captured during the diagnosis process, while just 4 cases were missed, this brings us to a sensitivity percentage of 92.31%. Vast majority of the cases were accurately captured so we have a relatively low miss rate.

Specificity (98.10%), Among the 945 participants who doesn't have dementia, 927 were correctly captured in the diagnosis as dementia-free.This high specificity of 98.10% shows that the diagnosis rarely mis-classify health issues and thus this data set in that regard has a good quality with only 1.9% of non-dementia cases gt a false positive diagnosis.

Positive Predictive Value (72.73%), when the positive dementia diagnosis is made, it is correct 72.73% of all the time of the 66 cases, 48 were true and 18 were false, while this appears low compared to other tests, it still appears high and indicates about 3/4 of the positive diagnosis are accurate.

Negative Predictive Value (99.57%), when the diagnostic process shows that there is no dementia., it is correct by 99.57% of the time. Of the 931 negative diagnoses, only 4 were not accurate. This is still very much high and accurate.

4.5 Discussion:

The dementia diagnosis shows an overall strong reliability with impressive specificity and negative predictive vale.

The Sensitivity although strong shows room for better diagnosis next time, as the four cases of missed diagnosis raise clinical concern on the patient's health if they do not receive appropriate treatment.

The moderate positive predictive value needs a very careful interpretation because roughly 1/4 cases were incorrect, the false positives could lead to anxiety and make the patient use drugs that could cause side effects and affect their health in other ways, psychological and social impact also.

There is also an observable balance between sensitivity ad specificity which appears conservative and normal for dementia diagnosis. It shows that the avoidance of false negative were prioritized during the diagnosis process which is a reasonable approach given the psychological and social implications attached to having dementia.

Overall, the dementia process shows sufficient reliability for clinical use with just few cases missed. The pattern of result also shows diagnostic conservative approach which is appropriate for a life-altering diagnosis like dementia.

5. SUGGESTED INTERVENTIONS TO PREVENT DEMENTIA

Based on the logistic regression analysis from question 3, three risk factors were identified which are BMI, Age and current smoking habit. This intervention will be tailored to address those effect.

Recommendations;

1. **With smoking** being the strongest relation to dementia from the analysis, I will advise Smoking cessation programs and I propose

- Structured smoking cessation programs with habitual and behavioral counselling
- Pharmacology support and Nicotine replacement therapy
- Tobacco control policy like ban or high taxation on tobacco production companies

These impact will potentially reduce dementia risk while the smoking habit which is a strong cause is being addressed at a social and clinical level.

2. Weight Check and Obesity Prevention:

From the analysis, I was able to deduce that 1kg increase in BMI gives a 50% odd of increase in dementia risk. And I propose that

- Since BMI is modifiabe and isa biological risk factor, community based weiht managemet programs
- Nutritional Councelling emphasy
- Physical activity promotion
- Clinical weight loss in very Obese individual

Reducing BMI from 30 to 25 could decrease dementia by 60%

3. Possible Life-style intervention for Aging population

Age is non-modifiable, but has high-risk groups that requires intervention, so I suggest

- Integrated programs targeting muliplr tisk factors in Older adults
- Cognitive training and mental stimulation activities
- Cardiio Vascular risk managements
- Regular health screening and check ups.

Conclusion

These intervention, identifies two modifable risk factors which are smokers and high BMi individuals with smoking having the most impactful intervention, and weight managemnt offers moderate and significant intervention, and Sge although not modifiable but can be actively monitored by targeted intervention to prevention, early detection and adequate management at a very early stage that could possible lead to recovery by cognitive stimulations.

6. **ACCOUNTING FOR A SURPRISE RISK FACTOR**

6.1 Introduction:

Our analysis identifies BMI, Age, and Current Smoking as significant risk factors for dementia. In case any of these-for example-is reported not to be associated with dementia previously according to literature, it should be investigated and analytical adjustments made accordingly.

6.2 Possible Causes and Remediation Strategies

1. Check for Confounding

Problem: The unexpected association may be confounded by unmeasured or uncontrolled variables.

Adjustments:

- Other co-variates that need to be included in the model are education, cardiovascular disease, diabetes, and alcohol consumption.
- Balance groups on observed confounders using propensity score matching.

Explanation: Confounding produces spurious associations. Adequate control may eliminate the surprising association.

2. Investigate Reverse Causation

Problem: The risk factor might be due to dementia, rather than vice versa; for instance, early dementia behavioral changes affecting either BMI or smoking.

Adjustments: Exclude participants who are newly diagnosed if timing data is available.
Perform sensitivity analysis, excluding borderline cases.

Reason: Since cross-sectional data cannot establish temporal sequence, unusual associations may be due to reverse causation.

3. Investigate Effect Modification (Interactions)

Problem: The risk factor may be related to dementia in subgroups that have not been studied up to now.

Adjustments: Test interaction terms (BMI × Age, Smoking × Age, BMI × Sex)
Stratification of analyses by age group, sex, or other relevant characteristics
Report stratified odds ratios to identify vulnerable sub-populations

Reason: A true but context-dependent association might only manifest in some populations.

4. Assess Measurement Error or Mis-classification

Problem: The way in which the variable is operationalized, or measured, may be different compared to the previous research.

Adjustments: Review measurement methods in detail and report.
Sensitivity analysis with alternative definitions; for instance, categorical versus continuous BMI
Check for systematic bias in measurement, for instance, recall bias and measurement error.
Compare measurement protocols with previous literature

Reason: Different operationalizations of the same construct might result in different findings.

5. Consider Sample Characteristics

Problem: Our sample may be systematically different from those studied in the past research. **Adjustments:** Describe in detail sample characteristics, such as demographics, method of recruitment, and geographic region. Compare prevalence rates and distributions with those from published studies. Perform subgroup analysis that matches characteristics of previous research. Recognize limitations of generalization

Reason: The choice of the sample, inclusion/exclusion criteria, and/or the characteristics of the population may account for inconsistent findings.

7. LIMITATIONS OF STUDY AND FUTURE IMPROVEMENTS

7.1 Cross-Sectional Design

Limitation: Data collected at a single time-point prevents establishment of temporal relationship or causality between the risk factors and dementia.

Limitations Impact: Unable to establish whether risk factors BMI, smoking-preceded the onset of dementia or were consequences of early decline in cognitive function. Reverse causation. Associations do not imply causation.

Improvement: Do a prospective cohort study where dementia-free individuals aged 50+, followed for 10-20 years, have their risk factors measured at baseline and their dementia incidence monitored. This establishes temporal sequence and allows for causal inference.

7.2 Self-Reported Measures

This is a limitation, as smoking, sleep, and physical activity depended on self-report and were thus subject to recall bias and social desirability bias.

Impact: Participants likely over-reported physical activity, as 96.6% >15 mins/day is unrealistically high, and under-reported smoking. Measurement error weakens true associations and may account for sleep and physical activity being nonsignificant despite literature support.

When appropriate, utilize objective measures: cotinine test to confirm smoking status, actigraphy to monitor sleep/activity, and medical records to confirm. For subjects with cognitive impairment, data from caregivers should be obtained.

7.3 Limited Generalisability

Limitation: Sample restricted to Greater Manchester, n=997; possible selection bias in the recruitment.

Impact: The results may not be generalizable to UK regions with different environmental risk profiles, ethnic groups, or international populations. Geographic-specific environmental factors, socioeconomic composition, and healthcare access patterns limit the external validity of possible findings.

Improvement: Multi-site recruitment across various UK regions with population-based random sampling from GP registries. Stratification will be by ethnicity and socioeconomic status, increasing the sample size for adequate power and representativeness to n = 5,000-10,000.

7.4 Omitted Confounders Limitation: Dataset lacked key variables: education, socioeconomic status, APOE genotype, medical history hypertension, diabetes, alcohol consumption, and diet quality.

Impact: Any observed associations may be due to confounding. Example: Smoking may be associated with low education, a protective factor. The smoking effect might be overestimated, OR = 4.28. Unable to adjust for established dementia risk factors identified in literature. **Improvement:** Collect comprehensive data on education levels, genetic testing (APOE-ε4), complete medical history via NHS record linkage, validated dietary assessment (Mediterranean Diet Score), and cognitive reserve measures. Use propensity score methods for sensitivity analyses of unmeasured confounding.