Group Member Information

Assignment will be completed in a group of 4-5 people. These will be the same people you work with for the final project. This assignment will help get you started on the final project.

First Name	Last Name	PID
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

Question

What are the top five risk factors that contribute to the development of Alzheimer's Disease, ranked from largest risk to smallest risk?

Hypothesis

Write down your group's hypothesis to your question. Provide justification how you came to this hypothesis. (What background information or instinct led you to that hypothesis?).

We predict that the 5 most influential risk factors for Alzhemier's Disease ordered from most influential to least influential are…

1. Age
2. Head trauma
3. Genetics/inheritance
4. Comorbidities
5. Diet

Research indicates that an individual's risk of developing neuro-degenerative disease rises significantly with age. Anecdotally, patients who develop Alzheimer's Disease at a younger age, that is, patients younger than 65 years old with an Alzheimer's diagnosis tend to have a history of blunt force trauma to the head, suggesting that head trauma is also a key risk factor. What is less known is the impact of a patient's genetic history or familial risk; Alzheimer's has been observed in patients as young as thirty in subpopulations with a family history of neurodegenerative disease.

An individual's risk of developing Alzheimer's Disease is closely related to their overall level of health - in particular their level of brain and heart health. We predict that patients with a history of other neurological disorders such as patients with a history of strokes, hypertension, diabetes, or other comorbidities are also more susceptible to dementia.

Diets also play a more indirect role in determining whether or not a person develops a neurodegenerative disease, because they also provide information about a person's cholesterol levels, a proxy for vascular health.

Background Information

Include a few paragraphs of background research and information on your topic. This should include at least 2 citations to work from others. Including hyperlinks to reputable sources is fine.

According to the Alzheimer's Association[1], Alzheimer's Disease is a neurodegenerative brain disease which is a primary cause of dementia. A patient's risk of developing Alzheimer's is based on a combination of risk factors which contribute to their risk over time. Some risk factors researchers have found include non-modifiable risk factors such as age, family history, and genetics (such as the APOE-e4 gene) as well as modifiable risk factors such as cardiovascular health, diet, and a patient's exposure to traumatic brain injury (TBI).[2] [3]Common symptoms among neurodegenerative diseases include memory loss, confusion, mood changes, anxiety, difficulty with language, and coordination.[4]

Currently, we cannot measure an individual's risk of developing Alzheimer's Disease directly. A correct diagnosis relies on multiple sources of information - a medical evaluation which takes into account imaging studies, bloodwork, family history, and cognitive tests. However, we can take indirect measurements of an individual's Alzheimer's risk factors by measuring quantifiable information including their demographic information (age), social history (history of smoking), physical activity (frequency of exercise), and diet (levels of cholesterol).

More recent research has shown that sleep disorders, particularly sleep disorders that interfere with the brain's ability to clear toxic proteins such as beta-amyloid plaques, are also important risk factors in neurodegenerative diseases, primarily Alzheimer's Disease. However, since sleep

[1] https://www.alz.org/media/documents/facts-and-figures-2018-r.pdf
[2] https://www.uptodate.com/contents/clinical-features-and-diagnosis-of-alzheimer-disease
[3] https://alzheimer.ca/en/Home/About-dementia/Alzheimer-s-disease/Risk-factors
[4] https://www.alz.org/alzheimers-dementia/what-is-alzheimers/causes-and-risk-factors

disorders are difficult to measure directly and do not have clear baselines established for all patients, understanding the role of sleep disorders falls outside the scope of this project.[5][6]

Data

Include a description of the perfect dataset you would need/want to answer this question. How many observations would you need? What variables would you collect? Explain the perfect dataset that you would want to answer this question.

Then, look online for available datasets. Find a dataset that could be used to answer this question. Describe how many observations are included and what variables have been collected. Discuss the datasets limitations and how it differs from your ideal dataset.

Our ideal dataset would include at least 10000 observations from people of various backgrounds, lifestyles, and socioeconomic statuses. The dataset should include metrics, such as scores on standardized memory and cognitive tests. But it should also include variables like age, history of related illnesses (hypertension, diabetes, strokes), history of blunt force trauma localized to the head, family history of vascular or neurodegenerative disease, genetic information, and diet. Certain variables, such as age, family history, genetic data, and history of related illnesses we could find by asking for official documents such as photo identification, medical records, and genetic records. Other variables rely heavily on patient self-reported information such as diet and history of traumatic brain injury. Many of these variables are not direct measures for how likely a person is to develop Alzheimer's, but instead serve as proxies of things that are more indicative of neurodegenerative diseases. For example, diets that are high in cholesterol may increase a patient's risk of developing brain or heart disease while family history of vascular disease is predictive for a patient's elevated risk at developing similar vascular illnesses in later life.

This existing Kaggle dataset was created to answer a similar question to ours, and it uses publicly available data on brain scans, consisting of 150 observations of patients aged 60 to 96, grouped into three categories, demented, nondemented, and converted. Every patient had undergone at least one brain scan and they were all right-handed. The dataset has 15 columns, including variables like sex, age, number of years of schooling, a socioeconomic status score, a mental wellness score, a clinical dementia rating, an estimation of total intracranial volume, and normalized whole brain volume.

[5] https://www.npr.org/sections/health-shots/2019/10/31/775068218/how-deep-sleep-may-help-the-brain-clear-alzheimers-toxins

[6] https://www.ncbi.nlm.nih.gov/pubmed/28550253

Ethical Considerations

Read about Deon's data science ethics checklist and consider the topics discussed in lecture. Then, discuss what ethical considerations must be made when answering your specific data science question. Brainstorm and explain how you would address these considerations for each of the following categories in your specific project: Data Collection, Data Storage, Analysis, Modeling, Deployment. Feel free to write about additional ethical considerations you would make that aren't included on the checklist. Note that data privacy is NOT the only ethical consideration for a data science project. It is a piece, but there is a lot more that has to be considered.

Data Collection: Patients participating in any medical research could worry about how their data will be used. A primary concern may be how we maintain anonymity. We would need to ensure the accurate collection of data without putting a patient's personal information at risk. We would need to secure informed consent for the data. Additionally, we would need to ensure our data is coming from different hospitals, geographic areas, and incomes to prevent confounding and to ensure that our sample is representative of the population.

Data Storage: As soon as we have the data, we would need to anonymize it if that step had not already been completed. Our data would have to be stored in a secure location to ensure that only authorized researchers had access to this information for legitimate research purposes. We would also need to develop an automatic data removal tool in the case someone revokes their consent. This tool would remove the data from the initial set, but also from the modeling and analysis if that is already complete.

Analysis: It is important to capture direct and indirect measures of Alzheimer's risk as well as related variables to prevent confounding. Using demographic information as an example, while other datasets do not ask for socioeconomic and demographic variables, we would like to collect this information to reduce the risk of confounding and to further improve our model's generalizability. Using traumatic brain injury as an example, patients who have a history of playing contact sports could be similar in their diet, so not accounting for the possibility of variables with co-linear relationships could skew our results. Holistically, our analysis would need to be documented thoroughly so it can be audited.

Modeling: As Alzheimer's is extremely prevalent in older populations, it's pertinent that we make sure that the variables we choose are not variables that discriminate against the elderly. While age is a major factor in the development of Alzheimer's we need to be aware that it isn't the only factor, and that some mental functions naturally degrade with age and are not indicative of Alzheimer's development. It's important that we assess each participant the exact same way with little to no bias to ensure the study is done fairly. A good metric we can use to assess impacts of Alzheimer's is a simple memory test, where scores of participants are compared to a scores of people in similar situations who have no known history of Alzheimers. To ensure accuracy, fairness, and a lack of bias, multiple tests should be used and questions that are commonly missed or deemed as unfair/discriminatory by a related proxy to the participant should be removed after being put under review. A larger data set ensures more accuracy and less discrimination. It is pertinent all bias is communicated between the participant, any proxy related to the participant, and the researchers. This aids in providing clean, accurate, and ethical data.

Deployment: The effective introduction of this project could lead to widespread analysis to see who is at risk of neurodegenerative diseases. While at first glance this may seem beneficial, it could indirectly hurt impoverished individuals. It is not unlikely that this analysis could be used against the individuals, such as health insurance companies that decide that premiums should be increased to preemptively respond to the possibility of the disease in the future.

This could increase premiums for at-risk persons and thus could hurt their ability to pay for care in the future, resulting in an unintended harm which resulted from their participation in this study. Safeguards must be created to protect against these types of scenarios.