

Assessing the Correlation Between Genetic Predisposition and Health Insurance Rates

Grant Proposal

Jared Rosen

Massachusetts Academy of Math and Science at Worcester Polytechnic Institute

Worcester, MA

Abstract (RQ)

The overall aim of this project is to identify what information health insurers use to determine their rates. This project will create a correlative study between the rates of various health conditions and health insurance rates and compare the degrees of correlation between genetic and non-genetic conditions. This study will compile data from several public health databases such as the National Cancer Institute's State Cancer Profiles, the Institute for Health Metrics and Evaluation's health maps, and the Centers for Disease Control and Prevention's Interactive Atlas of Heart Disease and Stroke, along with data on health insurance rates via resources like the CDC's professional report on health insurance coverage and the Kaiser Family Foundation's Health Insurance Marketplace Calculator tool.

Once the appropriate data from the health condition and health insurance reports are collected, the two datasets will be compared and analyzed for their level of correlation with both quantitative and qualitative techniques. This project will need to determine a way to normalize between the two datasets and make sense of their combination in the context of other factors influencing them such that the level of correlation can be objectively measured. Ultimately, the level of correlation will be used to determine the likelihood of health insurance agencies basing their rates on genetic predisposition. The goal of this project is to prove that there is a strong likelihood that insurance providers do so.

Assessing the correlation between genetic variability by region to health insurance coverage rates

The Patient Protection and Affordable Care Act of 2010 (United States, 2010), also known as Obamacare, has been known to be a controversial policy by many Americans. However, it has had many undeniable impacts on the medical industry and health insurance providers, such as allowing for a steady decline to all-time lows in the uninsured rate by expanding consumers' access to healthcare and health insurance services (Becerra, 2023). The subsection of the law regarding policy relating to pre-

existing health conditions is considered by many to be the most impactful part of the law (Kirzinger, 2022). The section states that no individual with a pre-existing medical condition may be refused health insurance because of their pre-existing medical condition. Although aspects of these policies vary from state-to-state (Kominski, 2017), the section of the law describe above remains consistent throughout the nation.

Nevertheless, insurance costs remain significantly expensive for Americans across the nation, and the costs are only growing. In 2021, Americans spent about \$4.3 trillion on health care expenses – nearly 85% of these expenses being into Private Insurance Providers- meaning the average American spends \$10,976 a year on health insurance (Center for Medicare and Medicaid Services, 2021). Although the Affordable Care Act (ACA) aimed to combat these costs, in response to the policies put into place by the ACA, insurance agencies have many ways to combat these policies by taking advantage of loopholes. Insurance companies can do so by taking advantage of the poor definition of the phrase “refusing health insurance based on pre-existing medical conditions” (United States, 2010). Although insurance agencies cannot directly reject someone due to a pre-existing medical condition, they can determine the probability that someone has a pre-existing medical condition or will be likely to develop one in the future based on information from their demographics. After determining this probability, insurers can determine rates by incorporating disease probability into the calculation as a risk level. Because of the availability of this process, insurance companies can still effectively reject or raise someone’s health insurance rates because of a pre-existing medical condition without it being against the policies of the ACA.

While this accessibility to insurance providers has been a threat to both the ACA and individual clients (particularly those that fall into demographics with higher average rates of genetic predisposition) in the past, this threat is more prominent than ever. Genetic testing through genomic sequencing has become incredibly cheaper and more accessible to the public in the past decade (Bowen et al., 2018). Almost anyone can get their hands on a 23andMe test, and for much cheaper than it used to be. Because of this,

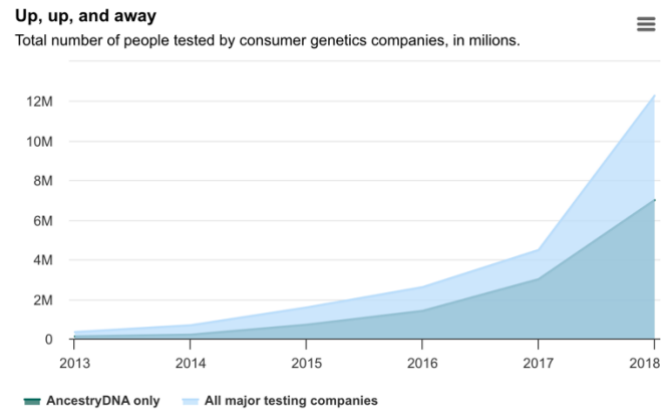


Figure 1. Graph showing rates of genetic testing from 2013-2018 (Regalado, 2018).

there is now an outstanding influx in the percentage of the population taking genetic testing (See Appendix I). As an effect of this increase, there is now an incredibly large amount of data available to genetic testing providers. By taking a genetic test, an individual gives their provider access to not only all the information relating to their genome, but also information on notable statistics such as medical conditions, medical history, age, sex, ethnicity, and geographical location. All this data is collected and inputted into a database. While these databases have existed for a substantial amount of time, they have not had nearly as much data available as they do now. Because of this, the data available can be compiled into numerous GWAS, which are ever more accurate and developed than before.

However, with the increase in data from GWAS, that information is now also accessible to health insurance agencies. One important thing to note is that the ways in which insurers determine their rates are generally unknown. Because of this, health insurance providers could be using the information provided by GWAS to calculate their rates, and it would be unknown to the public. This project aims to determine whether insurance agencies are performing the procedure. By correlating incidence rates of multiple health conditions and health insurance premiums by county, the level of

correlation among genetic conditions will reveal the extent to which insurance providers use the information from GWAS to determine their rates.

Section II: Specific Aims

The long-term goal of this proposal is to provide significant evidence for a strong correlation between the rates of genetic health conditions, which have been found to have a strong correlation with risk of developing conditions, and corresponding health insurance rates.

Specific Aim 1: Find at least two datasets, one on incidence rates of health conditions by county and the other on health insurance rates, that can be used to conduct a correlative study. The two will be related by a demographic such as location.

Specific Aim 2: To effectively determine the correlation between the datasets, the information between them will need to be normalized. There are many other factors that affect health insurance costs, such as income, or if the data varies by location, policies that vary from state-to-state will cause different health insurance rates as well.

Specific Aim 3: After normalizing between the two sets, the level of correlation between them can be analyzed. If both Specific Aim #1 and #2 are completed properly, the expected outcome is to show clear evidence for a correlation between the two sets.

The expected outcome is to provide significant evidence for a correlation between the rates of genetic conditions and corresponding health insurance rates by demographic. Ultimately, such a

correlation will provide significant evidence that Health Insurance Providers are largely using data on genetic risk of developing conditions to determine their rates.

Section III: Project Goals and Methodology

Methodology: This project will compile data on the rates of health conditions by demographic, with data on health insurance by the same demographic, and run a correlative study on the two datasets. The information between the datasets will need to be normalized before running the study, and outside factors will need to be considered as contributors to the results and any possible error.

Preliminary Evidence: In 2023, a report by Rodriguez-Rincon et al. in association with the RAND Europe corporation titled “Assessing the Impact of Developments in Genetic Testing on Insurers' Risk Exposure” was done for health insurance agencies to predict the possible impact the genetic testing industry has on the health insurance industry. The report noted concerns associated with the growth of the industry and possible applications (Rodriguez-Rincon et al., 2023). The presence of this study suggests that health insurance companies are considering the existence of genetic testing databases and looking into their data as an option for determining their risk exposure and rates.

Data Compilation: There are multiple ways that this project can go about accessing two types of datasets on rates of genetic predisposition and health insurance rates, which can be correlated to one another as there are many datasets accessible on both. However, this can be done in two categories: applying a GWAS to health insurance rates or applying data on health insurance rates to a GWAS.

Technique 1: The first way in which data needed for the correlative study can be accessed is by applying the data from a report on incidence rates of health conditions and attempting to find a corresponding dataset on health insurance rates. For instance, State Cancer Profiles include interactive atlases with data of incidence rates by county over five-year periods (National Cancer Institute, 2020). Then, a dataset on health insurance rates that correspond to this set would need to be found. For this hypothetical case, a dataset on the average health insurance cost in each of these cohorts could be used to relate to the genetic data. This technique is useful in the sense that it allows the health insurance information accessed to be directly correlated to the health condition data. However, one key downside is that health condition data is often not updated to recent years, requiring the health insurance data found to be from earlier dates. Therefore, the results derived from such a correlation will not be as accurate for today.

Technique 2: The second technique of obtaining the two necessary datasets for this project is the inverse of the first, by using data on health insurance rates by a certain demographic and applying that to find health condition incidence rate reports which can be correlated to this data. There are many published datasets already in existence providing data on the average health insurance cost for variables like age and income (Masterson, 2023). However, another way data on health insurance rates can be easily accessible is using the price estimators provided by the insurance companies. The program requires the user to input their date of birth and Zip Code (Kaiser Family Foundation, 2023), as well as the number of members in their family, and then provides an estimate on their monthly health

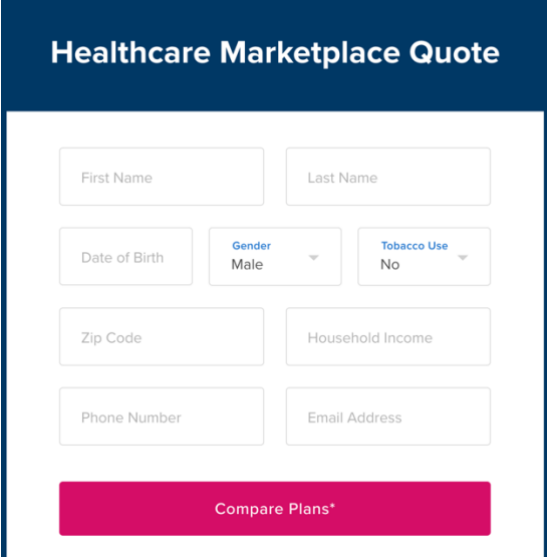


Figure 2. Price estimation tool provided by insurance companies (Rivelli, 2023)

insurance cost using their plan. Because the program asks for information on age and location, this program can be used to obtain data on how health insurance costs vary by age and location. With this access to data, all that is needed is a health condition incidence rate report with data on the rates of genetic and non-genetic conditions by age and/or location, and the two datasets needed will be obtained for a correlative study.

Normalizing Between Datasets: Either way that the two data sets are accessed, the data will first need to be normalized prior to running the correlative study. The study will need to account for how factors, such as differences in income and state-by-state policy, will affect health insurance rates in addition to the genetic data. Although it is difficult, eliminating these factors, it will improve the overall quality of the correlative study and lessen the probability of error. For example, eliminating variation in income could be done using data on the average health insurance cost by income (Urban Institute) and the average income by the selected demographic. Then, rather than assuming that the average health insurance rates by demographic are the same in the null hypothesis, the study will assume that the rates are based on the average health insurance cost by income using a ratio between the costs by each type of the demographic. By performing this process, the study will rule out important factors that affect health insurance rates so that the data will provide a more accurate assessment of how health insurance rates are affected by Genome-Wide Association Studies.

Correlative Study: The final step following accessing and normalizing the data is to run a correlative study between the two data sets. This process will be done both qualitatively as well as quantitatively. Qualitative correlations can be assessed by overlaying the two datasets onto one another and observing whether positive and negative trends between the data align between the sets. Quantitative correlations can be assessed by running significance tests, such as two-variable t-tests

between the rates of each gene and the corresponding health insurance rates to see how much each genotype corresponds with health insurance costs. Qualitative and quantitative observations will allow for informative conclusions to be made from the data regarding the correlation between the two datasets.

Preliminary Study: For a preliminary study, data was collected on health insurance premiums from 2024 for 10 states using the Kaiser Family Foundation's Health Insurance Marketplace Calculator.

This was done by finding the estimated premium for a 40-year-old individual who does not use tobacco

and does not receive any financial

aid and varying the location. Then,

this was correlated with data from

the same states on rates of alcohol

abuse by county from the Centers

for Medicare & Medicaid Services'

annual report on Chronic Conditions

(Centers for Medicare & Medicaid Services, 2018).

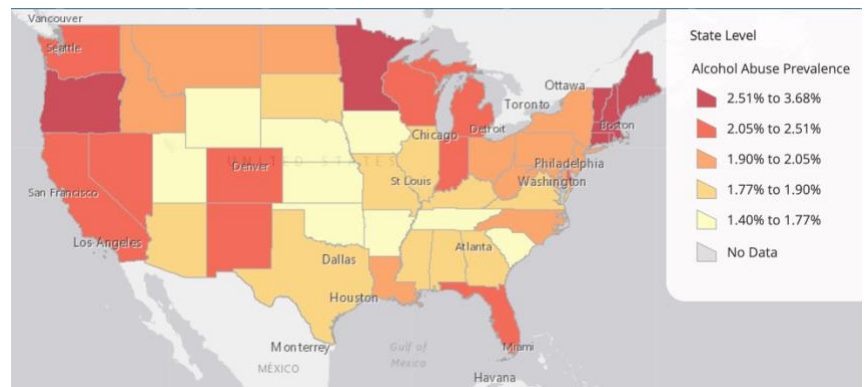


Figure 3. Screenshot from the Centers for Medicare & Medicaid Services' Interactive Atlas on Chronic Conditions (Centers for Medicare & Medicaid Services, 2018).

Statistical Testing:

The Pearson Correlation Assessment was used as a statistical test for this study. A Pearson correlation is a form of a linear regression model which assesses the degree to which two sets have a correlation with each other, reflected by a coefficient (R) from -1 to 1. A Pearson Correlation was run for each health condition in relation to the average health insurance premium by county.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Equation 4. Pearson Correlation Coefficient Formula.

- x_i = values of the x-variable in a sample
- \bar{x} = mean of the values of the x-variable
- y_i = values of the y-variable in a sample
- \bar{y} = mean of the values of the y-variable

Results: The result of this Pearson Correlation was a Pearson Correlation Coefficient of 0.5, showing a moderately positive correlation between the two samples. This proves that there is a statistical correlation between health insurance rates and alcohol abuse rates by county. A similar correlation done on health conditions such as Breast Cancer or heart disease will likely yield similar results once enough data has been collected for such a correlation.

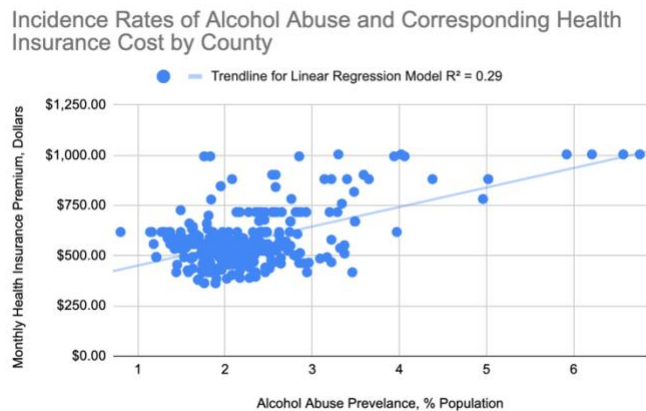


Figure 4. Linear Regression Model Displaying Pearson Correlation between Alcohol Abuse and Health Insurance Premiums by County

Conclusion: Once the data is analyzed and the correlation between the two datasets is assessed, the study will provide significant evidence on whether insurance agencies use information on an individual’s probability of genetic predisposition to determine their coverage rates. The expected result

of the study is that there will be a higher correlation among genetic conditions and corresponding health insurance costs than non-genetic ones. If this expected result is achieved, it will provide valuable information to governmental agencies, genetic testing companies, and the public regarding the methods insurers use to determine their rates. Individuals interested in taking genetic testing will be informed of the associated risk, should health insurance agencies be able to access their information. Similarly, genetic testing agencies will be informed of this risk and develop techniques to combat insurance providers from accessing their data to protect the privacy of their customers. Perhaps most significantly, this study will provide important information to governmental agencies that create and enforce policies on how health insurance companies determine their rates. With this information, these governmental agencies may be able to expand the details and restraints listed in their policies of the Affordable Care Act so that they will be able to better enforce them. This result will assure more affordable and equitable health insurance rates for consumers across the United States. On the contrary, if this result is not achieved, it will still provide significant information to the public that insurance agencies do not use the data on rates of genetic predisposition to determine their costs, but rather other data sources. This information will be useful for future studies as it will eliminate one option for how insurance providers are determining their rates and allow for other options to be investigated in a similar process.

Justification and Feasibility: The results of this study will be a quantifiable level of correlation between the rates of genetic health conditions and health insurance costs. If there is a significant correlation between the two datasets, it will prove that health insurance agencies are, to a significant extent, using the information on rates of genetic predisposition (or another dataset highly correlated to rates of genetic predisposition) when determining their rates, because there is already an existing strong correlation between rates of genetic conditions and their corresponding genetic mutations. External

factors such as income and state-by-state differences in policy will be ruled out when normalizing between the two datasets to rule out other error factors.

Section IV: Resources/Equipment

This project will not need any outside equipment. Data will be accessed from Genome-Wide-Association Studies such as ones performed by 23andMe, or the 1000 Genomes Project, as well as price estimators provided by health insurance companies.

Section V: Ethical Considerations

Many consider the proposal that insurance providers are basing their rates off genetic predisposition or similar data to be controversial because an individual's degree of "genetic predisposition" is determined by associative data. Someone more genetically predisposed to a disease will not necessarily be guaranteed to develop that condition in the future but simply has a higher likelihood of developing that condition. However, this study suggests that although it is speculative for insurers to base their rates on genetic predisposition data, the largely increasing amount of data gathered through GWAS enables this data to provide an accurate and reliable prediction on an individual's genetic predisposition to developing a condition. Ultimately, although it is a controversial suggestion, this study remains assured that it is very much so a possibility that insurance agencies are enabled to base their rates on data on genetic predisposition.

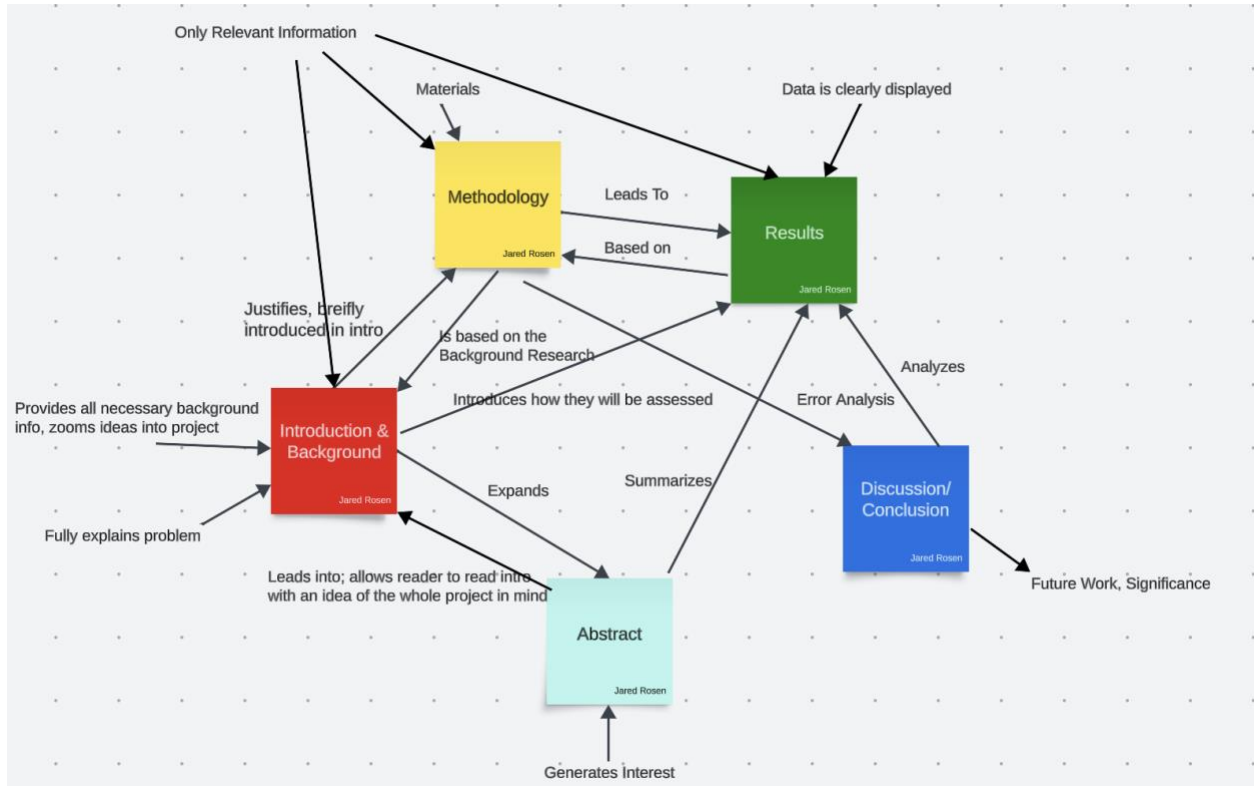
References

- Maier, Robert M, et al. "Improving Genetic Prediction by Leveraging Genetic Correlations among Human Diseases and Traits." *Nature Communications*, U.S. National Library of Medicine, 7 Mar. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC5841449/.
- Masterson, Les. "How Much Does Health Insurance Cost in 2023?" *Forbes*, Forbes Magazine, 23 Aug. 2023, www.forbes.com/advisor/health-insurance/how-much-does-health-insurance-cost/.
- Rivelli, Elizabeth. "PPO Insurance: What Is It?" *Forbes*, Forbes Magazine, 23 Nov. 2023, www.forbes.com/advisor/health-insurance/ppo-health-insurance-plans/.
- Becerra, X. (2023, August 8). *New HHS report shows national uninsured rate reached all-time low in 2023 after record-breaking ACA enrollment period*. HHS.gov. <https://www.hhs.gov/about/news/2023/08/03/new-hhs-report-shows-national-uninsured-rate-reached-all-time-low-2023-after-record-breaking-aca-enrollment-period.html#:~:text=These%20gains%20build%20on%20the,the%20uninsured%20rate%20in%202023.>
- Centers for Disease Control and Prevention. (2021). *Interactive Atlas of Heart Disease and Stroke*. <https://nccd.cdc.gov/dhdspatlas/reports.aspx>
- Institute for Health Metrics and Evaluation. (2019). *US health map*. <https://vizhub.healthdata.org/subnational/usa>
- Kirzinger, A., & Montero, A. (2023, November 28). *5 charts about public opinion on the Affordable Care Act*. KFF. <http://www.kff.org/health-reform/poll-finding/5-charts-about-public-opinion-on-the-affordable-care-act-and-the-supreme-court/>
- Kominski, G. F. (2015, December 15). *The Affordable Care Act's impacts on access to insurance and health ... The Affordable Care Act's Impacts on Access to Insurance and Health Care for Low-Income Populations*. <http://www.annualreviews.org/doi/pdf/10.1146/annurev-publhealth-031816-044555>
- Masterson, L. (2024, January 3). *How much does health insurance cost in 2024?*. Forbes. <https://www.forbes.com/advisor/health-insurance/how-much-does-health-insurance-cost/>
- National Cancer Institute. (2020). *State Cancer Profiles*. <https://statecancerprofiles.cancer.gov/index.html>
- Regalado, A. (2020, April 2). *2017 was the year consumer DNA testing blew up*. MIT Technology Review. <https://www.technologyreview.com/2018/02/12/145676/2017-was-the-year-consumer-dna-testing-blew-up/>

Rodriguez-Rincon, D. (2022, August 31). *Assessing the impact of developments in genetic testing on ...*
Assessing the impact of developments in genetic testing on insurers' risk exposure.
https://www.rand.org/content/dam/rand/pubs/research_reports/RRA1200/RRA1209-1/RAND_RRA1209-1.pdf

United States. (2010). *Read the affordable care act, Health Care Law.* , Health Care Law | HealthCare.gov. <http://www.healthcare.gov/where-can-i-read-the-affordable-care-act/>

Section VI: Appendix



Systems Diagram Chart