Multi-level Cardiovascular Disease Risk

Natasha Carpio Castellanos, Shawn Li, Jiayu Zhao

Executive summary

Cardiovascular disease (CVD) has long been one of the world's leading health threats. It causes roughly 18 million deaths worldwide each year (32% of global mortality) and drives over $200 billion in direct U.S. healthcare spending on hospitalizations, emergency care, and rehabilitation (WHO, 2021; Centers for Disease Control and Prevention [CDC], 2024).

Over 30% of heart attacks and strokes occur in patients with no prior noticeable symptoms, resulting in delayed treatment, higher complication rates, and greater costs (Centers for Disease Control and Prevention [CDC], 2024).

Existing risk models rely predominantly on clinical or self-reported data, overlooking socioeconomic and environmental drivers that limit prediction accuracy and generalizability.

This project integrates data from diverse sources and implements a data engineering pipeline to construct a unified, structured database. The pipeline includes data extraction, cleaning, normalization, and transformation processes to ensure consistency and quality. The final product is an accessible and interpretable dataset designed to support scalable analysis and facilitate multi-level risk assessment for cardiovascular disease.

Business case

This project proposes an integrated approach that combines individual-level and state-level data to enable a comprehensive, multi-level risk assessment framework, which better captures the complex and layered factors contributing to CVD. For example, a wellness platform could use our risk stratification to design personalized intervention plans for its high-risk members. Public health institutions could also make use of it for better identifying the needs and risks of their population.

Platform users

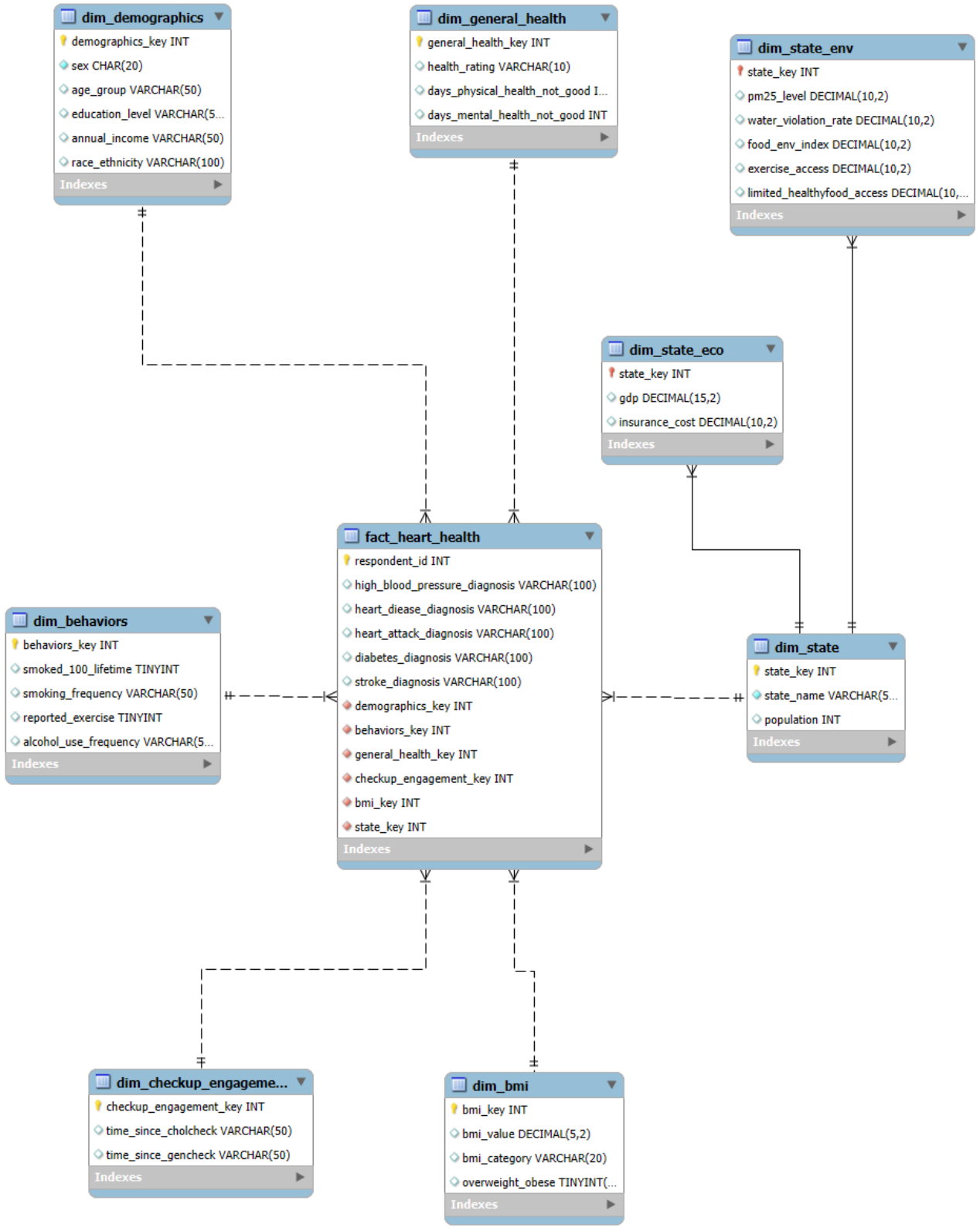
Stakeholder	Pain Points	Value Proposition
Health Insurers	High claims costs; difficulty distinguishing truly high-risk members	Lower acute event payouts through early identification; optimize premiums and reserves
Care Management Providers	Limited resources; challenge in prioritizing interventions	Ranked risk lists to focus care on those who need it most, reducing readmissions
Telemedicine & Digital Health	Low patient engagement; few high-value paid services	Offer premium risk reports and tailored coaching to boost retention and ARPU
Public Health & Government	Uneven resource allocation; difficulty quantifying health disparities	Targeted policies based on risk "hot spots" to improve outcomes in vulnerable counties

Risks

Risk	Mitigation
Data Quality & Latency	Implement robust data-QA processes, SLAs with vendors, and regular validation
Model Explainability	Integrate SHAP/LIME to generate business-readable factor reports
Compliance & Privacy Changes	Employ differential privacy/anonymization and monitor regulations in real time
Market Adoption & Education	Run pilot programs with leading insurers/telemedicine partners; publish white papers and case studies for rapid feedback and iteration

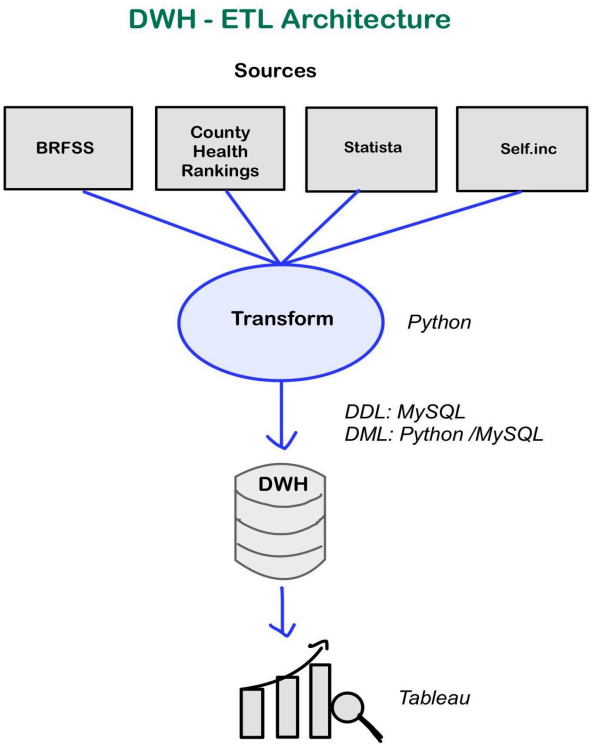
Data Model

Dimensional Modeling ERD



Methodology and various tools used in the process

- Data cleaning and preprocessing via Python
- Schema (DDL) managed in MySQL
- Data loading/manipulation (DML) done via Python scripts and direct MySQL commands



(Pipeline)

Data Sources

Behavioral Risk Factor Surveillance System. (2023). *BRFSS 2023: Annual data*. Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/annual_data/annual_2023.html

County Health Rankings & Roadmaps. (n.d.). *National data documentation 2010–2023*. https://www.countyhealthrankings.org/health-data/methodology-and-sources/data-documentation/national-data-documentation-2010-2023

Self Financial. (n.d.). *Health insurance costs by state*. https://www.self.inc/info/health-insurance-costs-by-state/

Statista. (n.d.). *U.S. gross domestic product (GDP) by state*. https://www.statista.com/statistics/248023/us-gross-domestic-product-gdp-by-state/

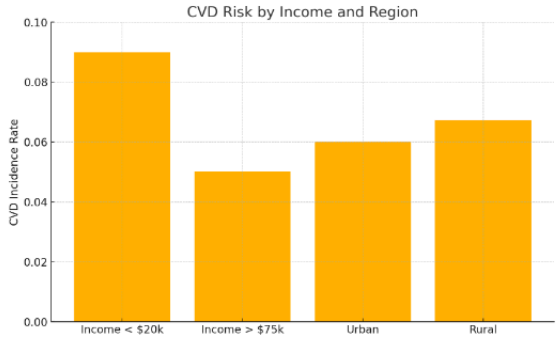
Insights

Tableau Report
https://public.tableau.com/app/profile/zizhan.li8282/viz/CVDRisk_17483262004220/Dashboard1?publish=yes

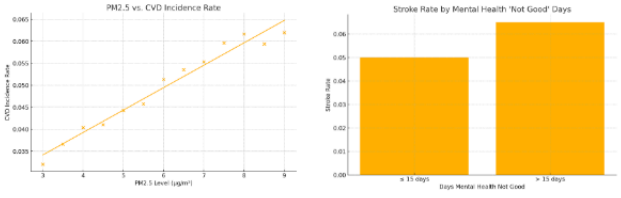
Conclusion

What we found

Significant socio-economic differences

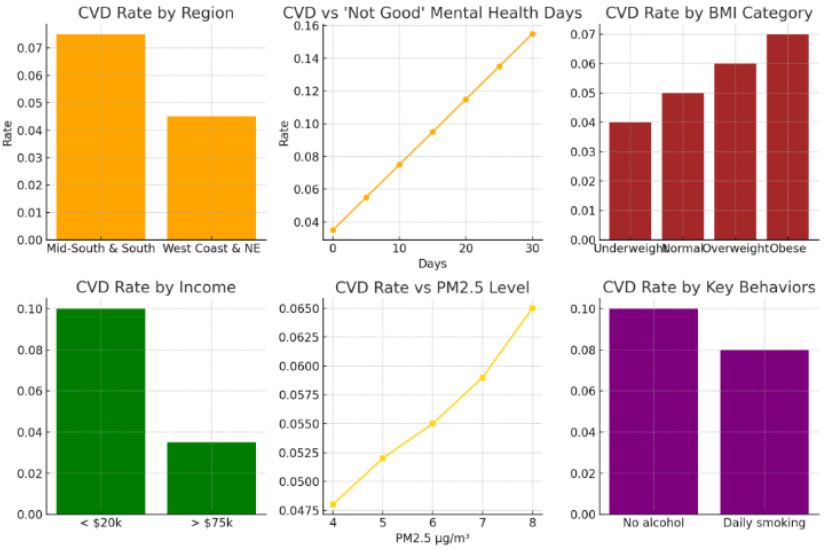


Driven by both the environment and behavior



Conclusion

What we found



Mid-South states (MO, SC, PA) have stroke rates >8%, vs. ~4.5% on West Coast/NE.

Mental health >20 "bad" days/month → stroke rate ~8%; Mental health ≤5 days → ~3.5%

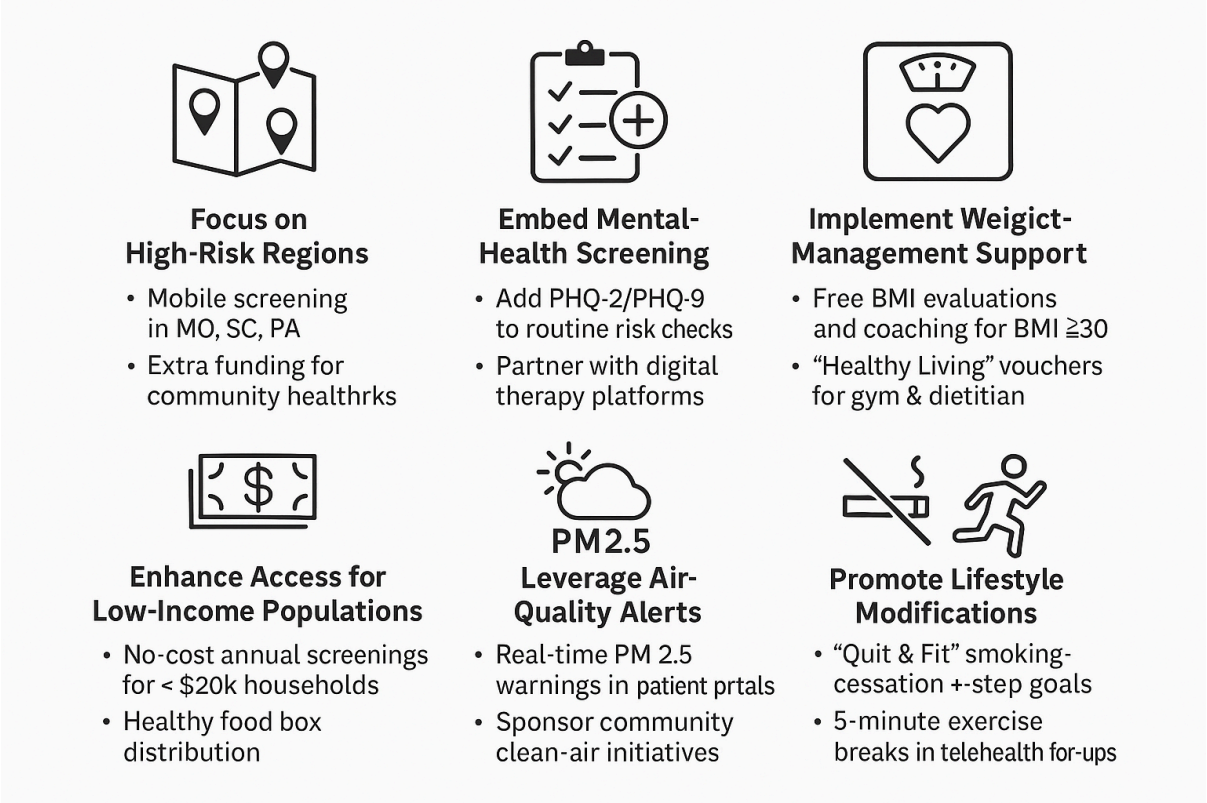
Obese individuals have the highest stroke rate at ~7.5%, Overweight and Normal-weight groups follow at ~6.3% and ~4.8%; Underweight individuals show the lowest rate at ~5.5%.

Income <$20k/year → >10% stroke rate; Income >$75k → ~3%

For air quality, PM2.5 ↑ from 4 to 9 µg/m³ → stroke rate ↑ from ~4.6% to ~6.5%

No exercise → ~8% risk; daily exercise → ~4.7%

Recommendations



References

Behavioral Risk Factor Surveillance System. (2023). *BRFSS 2023: Annual data*. Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/annual_data/annual_2023.html

Centers for Disease Control and Prevention. (n.d.). *Heart disease facts and statistics*. https://www.cdc.gov/heart-disease/data-research/facts-stats/?CDC_AAref_Val=https://www.cdc.gov/heartdisease/facts.htm

County Health Rankings & Roadmaps. (n.d.). *National data documentation 2010–2023*. https://www.countyhealthrankings.org/health-data/methodology-and-sources/data-documentation/national-data-documentation-2010-2023

Self Financial. (n.d.). *Health insurance costs by state*. https://www.self.inc/info/health-insurance-costs-by-state/

Statista. (n.d.). *U.S. gross domestic product (GDP) by state*. https://www.statista.com/statistics/248023/us-gross-domestic-product-gdp-by-state/

World Health Organization. (2021). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)