SMU Data Science Review

Manuscript 1294

Predictive Modeling of Colorectal Cancer Risk: Leveraging Health, Demographic, and Socioeconomic Factors for Targeted Screening

Anish Bhandari

Shawn Deng

Michael Olheiser

Robert Slater

Predictive Modeling of Colorectal Cancer Risk: Leveraging Health, Demographic, and Socioeconomic Factors for Targeted Screening

Anish Bhandari[1], Michael Olheiser[1], Shawn Deng[1], Robert Slater[1]
[1] Master of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA

Abstract. Colorectal cancer (CRC) remains a significant public health concern, affecting millions in the United States and worldwide. This study investigates the risk factors associated with CRC using data from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial and aims to develop predictive models to identify high-risk individuals for targeted screening and increased awareness. The dataset integrates CRC incidence data from the National Cancer Institute with socioeconomic indicators from U.S. Census Bureau, linked by zip code. We employ Logistic Regression and Neural Network models to predict CRC risk, incorporating health, demographic, and socio-economic features. While the results suggest that factors such as age and address history are significant contributors to CRC risk, the inclusion of Census data had a marginal impact on model performance, likely due to limited geographic diversity. The study further explores the use of a user-facing risk calculator, designed to raise awareness and encourage screening, with a focus on accessibility and simplicity for users. These findings emphasize the importance of incorporating a broad range of factors in CRC risk prediction and the potential for improving outreach through user-friendly tools.

1 Introduction

In the United States alone, colorectal cancer is diagnosed approximately 150,000 times each year with over 50,000 deaths. Alarmingly, one in 24 individuals will be diagnosed with CRC in their lifetime with an average age at diagnosis of 66 (Siegel et al., 2023). The problem is not just evident in the United States either (Bray et al., 2024). CRC is the third leading cancer diagnosis worldwide and second in terms of cancer-related mortality with nearly 1 million deaths annually (Laversanne et al., 2022). Even more concerning is the recent trend in early onset CRC. Young adults (<50 years of age) are being diagnosed with colorectal cancer at increasing rates. In early 2023, the American Cancer Society (ACS) reported that nearly 20% of CRC diagnoses in 2019 occurred in patients under 55, a rate that has doubled since 1995 (ACS, 2023). This rise in early onset CRC is particularly concerning due to the higher rates of advanced disease and worse outcomes in younger individuals (McClelland et al., 2022). The statistics are clear – colorectal cancer is a prevalent and lethal disease for all ages but

there is hope. If the cancer is caught early, Stages I and II, the 5-year survival rate is 91% (ACS, 2023).

Early detection of CRC through screening is crucial to survival as it can reduce both incidence and mortality by identifying and removing precancerous lesions (polyps) and detecting the disease at earlier, more treatable stages (Zheng et al., 2023). Invasive screening methods like a colonoscopy or flexible sigmoidoscopy are the most effective screening methods but are a significant burden to patients given the preparation prior to the screening and the continued adherence to screening at least every 10 years. (Zhang et al., 2024) Approximately 1 in 3 individuals in the U.S. do not get appropriately screened for CRC (ACS, 2023). The gap in screening rates and continued morality of CRC has spurred interest in developing non-invasive and cost-effective methods for CRC risk assessment.

The advent of machine learning offers novel opportunities to analyze vast amounts of health data and advance our understanding of CRC risk. Among the various machine learning techniques, logistic regression is a widely used statistical model for classification problems and has shown success in Lung Cancer risk modeling using PLCO study data (Martin et al., 2014). Its interpretability of individual risk factors makes it a valuable tool for understanding the contribution of different variables to CRC risk. By identifying key predictors through logistic regression, it becomes possible to develop simplified risk assessment tools. To enhance public awareness and promote early detection, a user-facing risk calculator based on a robust prediction model can be an effective strategy that can provide individuals with a preliminary assessment of their CRC risk, encouraging them to seek appropriate medical advice and consider screening options.

This research paper aims to develop a predictive model for colorectal cancer using logistic regression based on PLCO study data. Furthermore, we present the development of a user-facing risk calculator derived from this model. The goal of this work is twofold: to contribute to the growing body of knowledge on data-driven CRC risk prediction and to provide a practical tool for increasing public awareness and potentially facilitating earlier detection of this prevalent and often preventable disease.

2 Literature Review

2.1 Colorectal Cancer Rates

Colorectal cancer (CRC) is the second most common cause of cancer death in the United States. In 2023, it was predicted that approximately 153,020 individuals would be diagnosed with CRC and 52,550 would die from the disease, including 19,550 cases and 3750 deaths in individuals younger than 50 years (Siegel et al., 2023). The incidence of early-onset colorectal cancer (CRC), which occurs in individuals less than 50 years of age, has been increasing worldwide and particularly in high-income countries. The reasons for this increase remain unknown but plausible hypotheses include greater exposure to potential risk factors, such as a Western-style diet, obesity,

physical inactivity and antibiotic use, especially during the early prenatal to adolescent periods of life (Akimoto et al., 2021). Studies have identified that early detection plays a crucial role in improving survival rates. In Brazil, research has shown that the mortality rate from CRC increased steadily from 2000 to 2019, with higher mortality observed among individuals aged over 45 years (Nascimento et al., 2022). Early detection through screening methods like colonoscopy is key to improving prognosis since CRC progresses slowly and can be treated more effectively when caught early. However, adherence to screening, particularly among high-risk populations, remains low in certain regions of Brazil (Nascimento et al., 2022).

2.2 Colorectal Cancer Screening

Screening for colorectal cancer (CRC) is pivotal in preventing rising incidence rates, particularly among younger adults who may overlook the necessity due to the invasive and time-consuming nature of procedures like colonoscopy. While colonoscopy remains the gold standard for CRC screening, randomized control trials have yet to conclusively favor one method over another. (Gupta, 2022). Screening completion rates vary significantly across age, race, socioeconomic status, and insurance access; however, insurance alone does not guarantee increased participation or follow-up on abnormal results. To address disparities, policies must promote equitable access to CRC screening. Although lowering the screening age has proven cost-effective and lifesaving on a population level, this strategy may not translate well for younger individuals, where the financial and health burdens of endoscopy-based methods outweigh potential benefits (Ladabaum et al., 2019). Any screening approach for young adults requires thorough risk analysis.

2.3 Geographic Variation

The geographic variation of colorectal cancer (CRC) incidence provides critical insights to enable targeted screening and awareness. For example, a 2020 study identified hot spot counties in the Southern U.S. where men experienced significantly worse mortality outcomes, even after adjusting for confounding variables such as age, race, and cancer progression (Rogers et al., 2020). These findings highlight the importance of integrating geographic analysis with sociodemographic and health-related data to address disparities.

Geographic Information Systems (GIS) is a commonly used tool to study CRC variability. A particular study utilized the Maryland Cancer Registry and Baltimore neighborhood-level data to map cancer hot and cold spots using ArcGIS. Their regression analysis revealed significant variation in CRC rates driven by neighborhood-level factors, including race, income, and age. (Torres et al., 2018). The study highlighted the need for intervention strategies to consider community characteristics, aligning with the observed geographic disparities. This geospatial methodology is relevant to young adult-onset CRC studies, providing a robust framework for exploring cancer trends.

Another study that leveraged ArcGIS examined CRC screening disparities at the ZIP-code level, leveraging data from the CDC's 500 Cities and PLACES projects alongside the American Community Survey. The authors constructed a Social

Deprivation Index (SDI) to quantify socioeconomic disadvantage and identified that high SDI, along with race, income, and healthcare access, significantly drove screening rates. In particular. Black and Hispanic populations, as well as individuals in high-deprivation areas, are less likely to have gone through CRC screening, reinforcing the role of geographic and social determinants in cancer prevention (Parraga, 2023).

Risk factors for CRC vary by geographic region depending on the age of onset according to a recent study in 2023 that explored geographic variation in early versus late-onset CRC using random forest analysis to evaluate community-level risk factors (Dong et al., 2023). The study utilized the SEER database, County Health Rankings and Roadmaps, and the Health Resources and Services Administration data. The results showed significant geographic variation between early and late-onset CRC, with risk factors like diabetes and sedentary lifestyle impacting both but with different relative importance. Notably, some regions exhibited high early-onset but low late-onset CRC rates, again highlighting the need for community level risk assessments and interventions.

These studies demonstrate the critical role of geographic analysis in understanding early-onset CRC and its associated risk factors. By integrating GIS tools, neighborhood-level data, and machine learning methods, researchers can uncover nuanced patterns and inform targeted, equitable interventions to reduce disparities in CRC outcomes.

2.4 Health Risk Factors

The American Cancer Society identifies several health risks that are associated with CRC incidence. These factors include obesity, diabetes, smoking, alcohol consumption, and diets high in processed meats and low in fiber. (ACS, 2023). A meta-analysis that modelled the incidence of Early-Onset Colorectal Cancer (EOCRC) found the top risk factors associated with the recent increase in EOCRC was due to family history of CRC, obesity, hyperlipidemia, and alcohol use. Smoking was suggestive but not significant across the multitude of studies. (O'Sullivan et al., 2022). This study aims to close that gap by integrating environmental exposures such as industry, agriculture, and pollution along with the community health care availability in each county.

2.5 Social Determinants of Health: Education, Gender, Income, Race

In addition to medical and technological advancements, sociodemographic factors also play a role in CRC mortality. For instance, in the US, the average annual overall incidence rate during 2015 through 2019 was 33% higher in men (41.5 per 100,000) than in women (31.2 per 100,000) (Siegel et al., 2023). A separate study also showed higher risk in Men in addition to African American ethnicity (McClellan, 2021). Additionally, the same study found that certain age groups had lower risk of survival, specifically, 20's, 30's and 70's.

In Brazil, higher mortality rates were observed among certain racial groups with higher mortality rates in the South compared to the North and Northeast regions (Nascimento et al., 2022). These findings suggest that both biological factors and access to healthcare services may influence survival outcomes, underlining the need for targeted public health interventions to address these inequalities.

2.6 Machine Learning Methods

Machine Learning has become an emerging tool in predicting cancer outcomes, including CRC survival rates. In a study utilizing data from São Paulo, Brazil, machine learning algorithms, such as Random Forest and XGBoost, achieved high accuracy in predicting survival outcomes for CRC patients (Cardoso et al., 2023). These models highlighted clinical staging, age, and recurrence as the most important factors influencing survival, which aligns with other epidemiological studies on CRC prognosis. Another study utilized a large dataset including CRC patients from multiple continents and created nine supervised and unsupervised machine learning algorithms that were evaluated on the aggregated dataset. The optimal was an Artificial Neural Network (ANN) with false negative rate of 1% and false positive rate of 3% (Rahman et al., 2023). Geospatial analysis may be combined with machine learning methods. Another study used a random forest to measure the variable importance of community level risk factors comparing early onset to late onset (Dong et al., 2023). By leveraging machine learning, the authors were able to predict outcomes and guide treatment decisions for each community, further emphasizing the potential of integrating data-driven models into clinical practice.

2.7 Logistic Regression Risk Modeling

Logistic regression is a well-established statistical technique frequently used in clinical risk modeling due to its interpretability, efficiency, and effectiveness in binary classification problems. These qualities make it an ideal choice for evaluating individual-level risk using user-provided data inputs.

A study utilizing data from the Physicians' Health Study (PHS) modeled CRC risk in men using logistic regression. The final model achieved an area under the curve (AUC) of 0.70, with body mass index (BMI), alcohol use, smoking history, and age as the most significant predictors (Driver et al., 2007). These results highlight both the success and challenges of CRC risk modeling, particularly when constrained by a limited set of health and demographic features.

For comparison, a more comprehensive model for lung cancer risk prediction published in the New England Journal of Medicine demonstrated superior performance. By integrating data from the National Lung Screening Trial (NLST) and the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, their logistic regression model achieved an AUC greater than 0.80 (Tammemägi et al., 2007). The enhanced predictive power was attributed to the richness and diversity of risk factors available across both datasets. Two models from that study were ultimately deployed as user-facing risk calculators, including the simplified web-based model (PLCOm2012) which requires only five input variables—primarily age and smoking-related factors.

Inspired by this approach, the current study aims to develop a similar logistic regression-based model for predicting CRC risk, leveraging PLCO study data. The goal is to translate this model into a practical, user-facing risk calculator to support early detection and screening decisions.

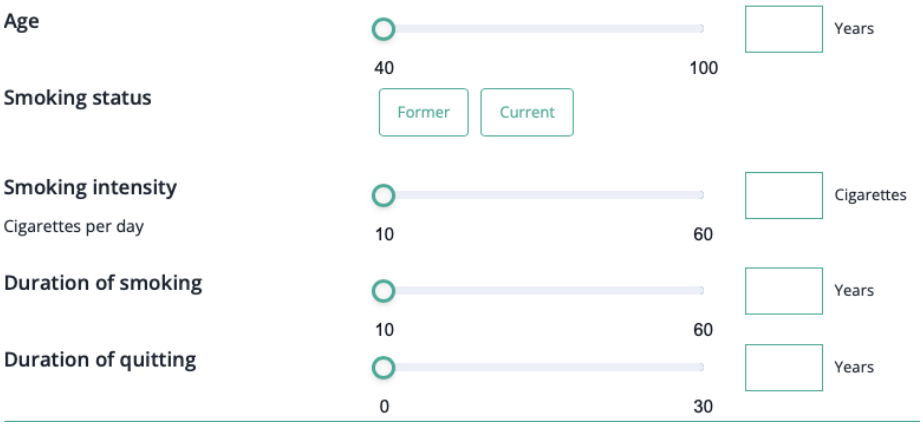

Fig 1. PLCOm2012 Simplified Screening Model for Lung Cancer

3 Methods

This study employs machine learning classification techniques to assess and predict the risk factors associated with colorectal cancer (CRC). By leveraging comprehensive datasets that integrate cancer incidence data from the National Cancer Institute and sociodemographic data from the U.S. Census, the methodology aims to uncover patterns in health and community-level variables that contribute to CRC risk.

The primary objective is to develop a predictive model capable of classifying individuals into high- or low-risk categories, forming the foundation for a user-facing risk calculator. To ensure broad usability and practical application, the modeling process emphasizes simplicity, efficiency, and interpretability.

3.1 Data

The data used for machine learning modeling was obtained from the National Cancer Institute's Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. This dataset includes information on over 95,000 colorectal cancer cases in the United States, spanning the years 1990 to 2020. To enrich the analysis with contextual and environmental variables, the PLCO data was supplemented with U.S. Census data linked via geographic codes.

The integrated dataset contains over 300 features, including sociodemographic and community-level variables such as unemployment rate, poverty level, housing characteristics, internet access, educational attainment, health insurance coverage, and more. This comprehensive dataset enables a multifaceted exploration of both individual and structural risk factors associated with colorectal cancer.

3.2 Preprocessing

Data cleaning began by removing features missing most data, retaining only those with sufficient data for reliable modeling. Approximately 40 features directly indicative of CRC incidence—such as mortality, age of onset, and study-specific identifiers—were excluded to prevent data leakage.

To avoid bias in the predictive model, 3% of participants with a prior history of colorectal cancer were removed from the dataset. Categorical variables, such as race, were one-hot encoded. All remaining numerical features were scaled using Scikit-Learn's Standard Scaler to improve model performance and speed up feature selection. Missing values in health-related variables were minimal (affecting less than 1% of the data) and were imputed using the median value of each respective feature. Census-derived variables had no missing values and required no imputation.

To ensure robust and unbiased model evaluation, Stratified K-Fold Cross-Validation was employed. This technique maintains the proportion of high- and low-risk cases in each fold, preserving the class distribution across training and validation subsets. A 5-fold split was used.

3.3 Feature Selection

Feature selection was conducted through a combination of manual curation and automated techniques to identify the most informative variables while maintaining model interpretability for a user-facing application. Census-derived and non-Census health features were separated prior to selection to enable the development of distinct models: one leveraging publicly available geographic and sociodemographic data, and another based on individual-level health and lifestyle inputs. This distinction reflects practical considerations, as users are more likely to know personal attributes such as height and weight than local poverty or unemployment rates.

Exploratory Data Analysis (EDA) was first conducted to understand the relationships among variables. Pearson correlation coefficients were calculated to identify multicollinearity, while t-tests and associated p-values were used to assess statistical differences between risk groups. Effect sizes were also evaluated to determine the strength of these associations. Insights from this analysis helped guide manual feature curation that ultimately were used in the final risk calculator model.

In addition to manual methods, several automated feature selection techniques were applied to both datasets independently, including Recursive Feature Elimination (RFE) and Variance Inflation Factor (VIF) analysis. These approaches helped prioritize features that contributed most to model performance while minimizing redundancy. The resulting curated feature sets were then passed to the classification model for evaluation.

3.4 Classification Modeling

The primary objective of classification modeling in this study is to distinguish between individuals at high versus low risk for developing colorectal cancer (CRC), while also identifying and explaining key contributing risk factors. Logistic Regression was selected as the baseline model due to its simplicity, efficiency, and interpretability. Its regression coefficients offer direct insight into the influence of each feature on CRC risk, making it an ideal candidate for user-facing risk assessment tools.

However, the dataset presented a significant class imbalance, with only 1.5% of participants developing CRC during the study period. To address this imbalance, we employed three complementary strategies: Stratified K-Fold Cross-Validation ensure that each fold preserved the original class distribution and reduced variance in performance estimates., Class-balanced model weights within the Logistic Regression to penalize misclassification of the minority class more heavily. Adjustable prediction probability threshold to optimize sensitivity and specificity, with the ideal cutoff based on maximizing the AUC ROC.

To compare the baseline performance with a more complex model, we also implemented a Sequential Neural Network using TensorFlow. The architecture consisted of dense layers interspersed with dropout layers to mitigate overfitting. Due to the computational burden of training on a highly imbalanced dataset, we employed a hybrid sampling approach combining SMOTE (Synthetic Minority Oversampling Technique) with under sampling of the majority class and over sampling of the minority class.

3.5 Model Evaluation

Given the rarity of colorectal cancer (CRC) cases in the dataset and the study's goal of identifying individuals at elevated risk, the model is not expected to perfectly predict CRC incidence. Instead, the emphasis is placed on recall, prioritizing the correct identification of high-risk individuals to encourage proactive screening. At a minimum, the model should perform better than random guessing, which corresponds to an AUC of 0.5.

Model performance was primarily evaluated using two key metrics: Area Under the ROC Curve (AUC) and Recall: to measure the model's effectiveness. Logistic Regression and Neural Network models were compared on these metrics. In addition to performance comparison, feature importance was interpreted using the regression coefficients from the Logistic Regression model, offering insight into which risk factors most strongly influence CRC prediction. Balancing interpretability with predictive performance further supports the development of a user-facing risk calculator that is both meaningful and actionable.

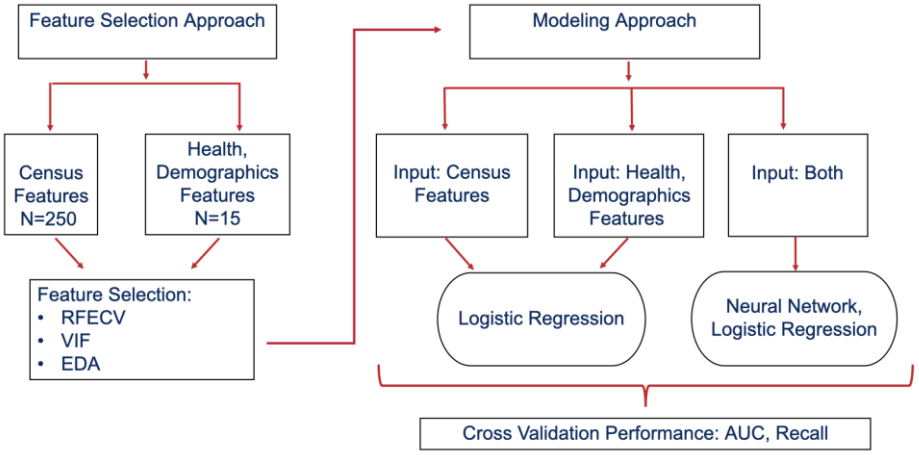

Fig. 2. High level Model Pipeline

4 Results

4.1 Model Comparison and Features

The combined model, which included both Census and Health features, achieved the highest performance in terms of AUC and Recall, followed closely by the Health-only model (Figure 3). This suggests that while Census data can offer slight improvements, the added complexity may not be justified for a user-facing application.

As expected, Age was the most influential feature in the Logistic Regression Health model, with secondary factors including the number of addresses, days at current address, height, weight, and education (Figure 4). Notably, even when Age was excluded, the model maintained similar AUC and Recall, indicating these secondary features hold substantial predictive value on their own.

Models using only Census features performed worse. This may be due to a lack of heterogeneity in the census-level risk factors: the Census data in this study covers only 10 study center areas, with most participants likely living near urban centers. As a result, there's limited variability in socioeconomic and environmental exposures across participants, making it difficult to distinguish high-risk from low-risk individuals using Census data alone.

The feature correlation heatmap (Figure 3) showed minor correlations among smoking-related variables, but overall multicollinearity was minimal, supporting the robustness of the feature set.

The Neural Network model did not outperform Logistic Regression, despite incorporating sampling strategies to address class imbalance. Given this, the Health-only Logistic Regression model stands out as the most effective and practical option for a user-facing risk calculator, offering strong performance alongside simplicity and interpretability.

Base Feature Set	Model	AUC	Recall	Precision	Accuracy
Health & Demographics 1	LR	64.2%	60.7%	2.4%	60.1%
Health & Demographics 2	LR	64.0%	62.2%	2.4%	59.2%
Health & Demographics 2 w/o Age	LR	60.4%	51.3%	2.4%	65.3%
Health & Demographics All	LR	64.2%	66.2%	2.3%	54.5%
Census 1	LR	59.6%	53.1%	2.3%	62.1%
Census 1 + Health & Demographics 2	LR	63.8%	65.3%	2.4%	56.0%
Health & Demographics 2	NN	58.9%	47.9%	2.1%	63.9%
Census 1 + Health & Demographics 2	NN	56.3%	46.0%	2.0%	62.4%

Fig. 3. Model Results

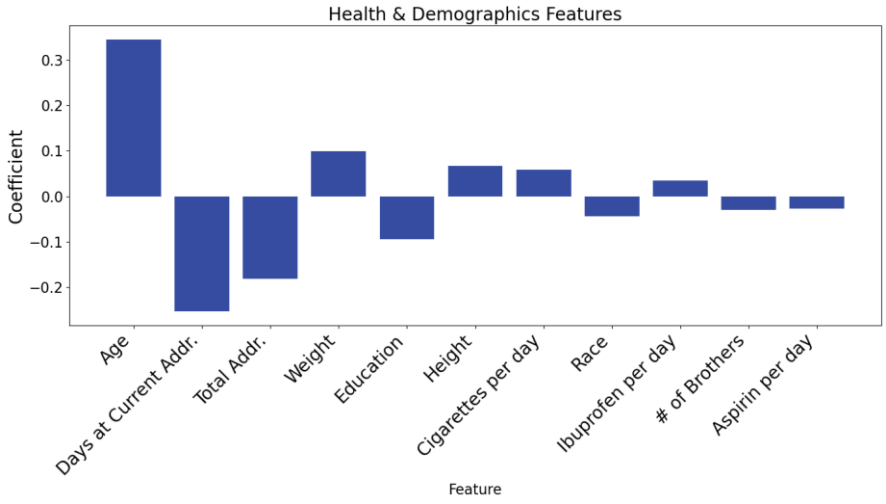

Fig. 4. Feature Importances of Logistic Regression

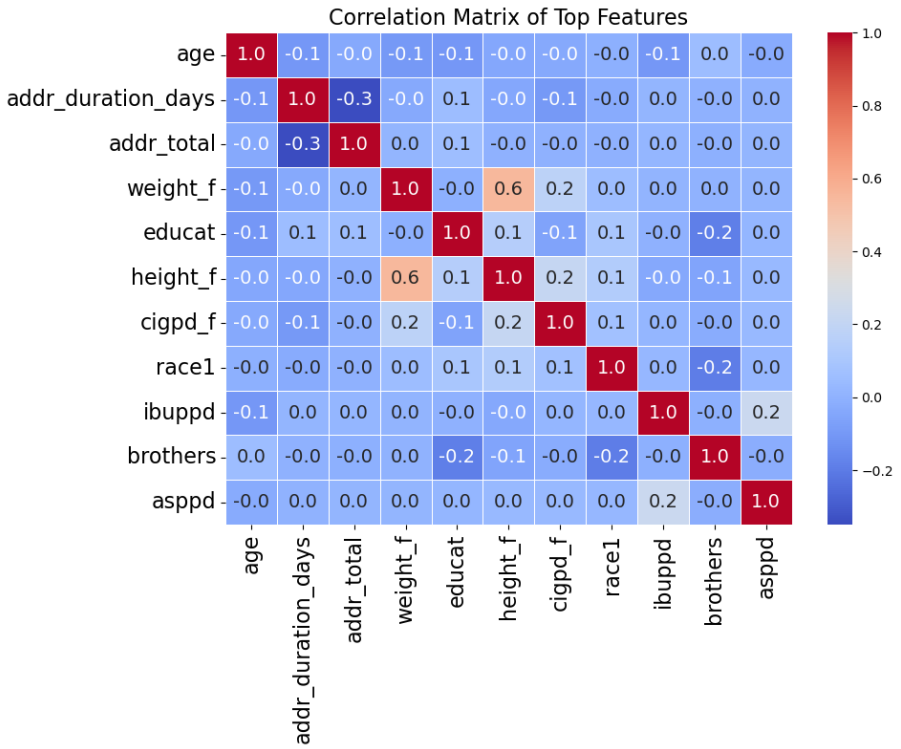

Fig. 5. Correlation Heatmap of Features in Health Model

4.2 Risk Calculator

The second health and demographics feature set was selected for the back-end model of the risk calculator dashboard. This feature set achieved the highest recall and the second-highest AUC, making it ideal for a screening tool focused on identifying high-risk individuals. Additionally, this set was curated using exploratory data analysis to identify and mitigate multicollinearity particularly among variables like cigarette pack-years and BMI.

Although BMI is typically used as a composite feature of height and weight, feature selection indicated a slight performance improvement when height and weight were included separately. This also supports transparency and ease of input for users. Logistic Regression was chosen as the final classification model for the calculator due to its simplicity, interpretability, and strong performance. The model takes the user's answers to a brief set of questions and outputs a binary classification: high risk or low risk of developing colorectal cancer. Each question maps directly to one of the features used in the model, ensuring alignment between the input data and model predictions. Figure 6 displays the user interface of the risk calculator with corresponding input fields for each feature

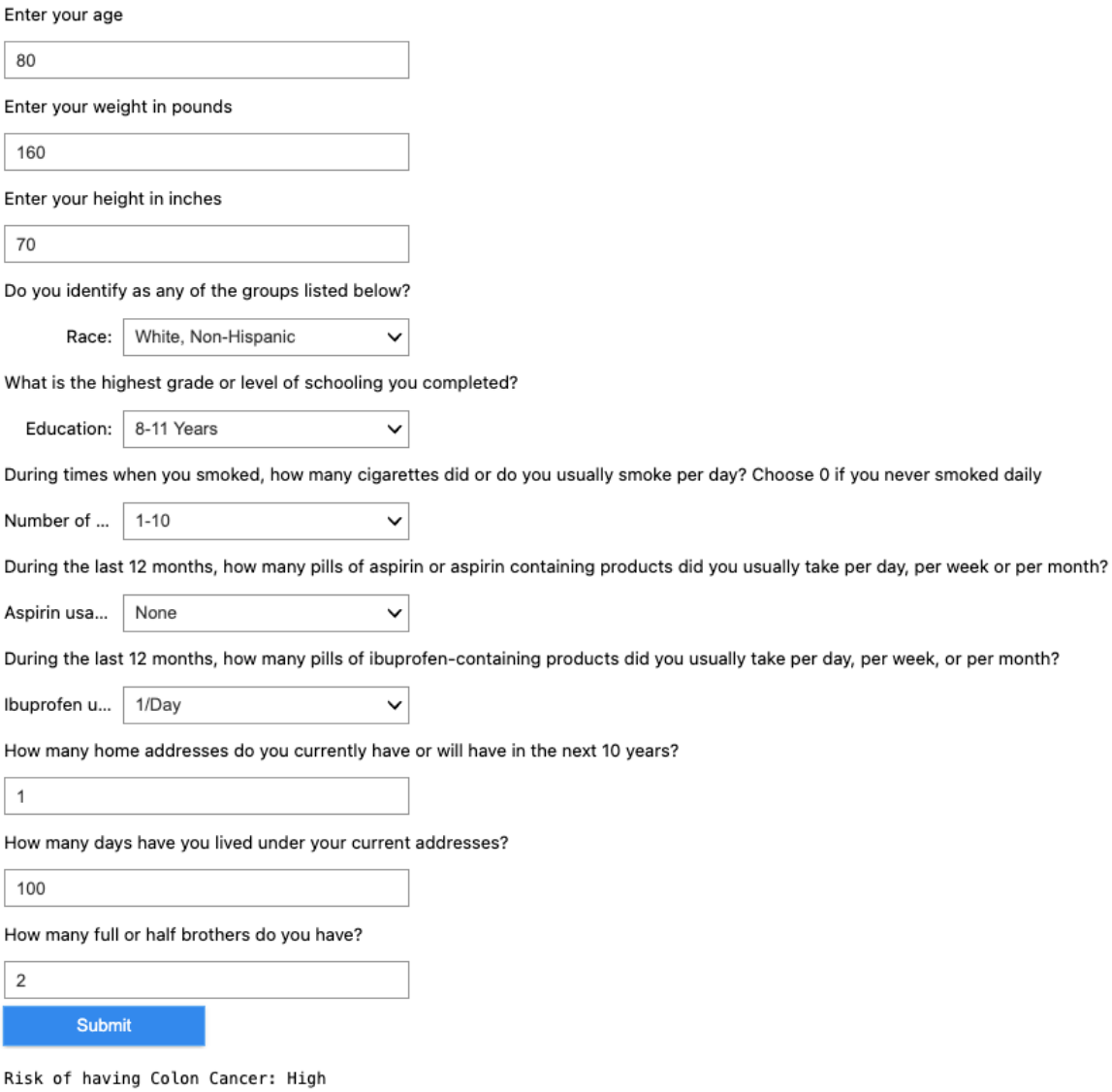

Fig 6. Risk Calculator Questionnaire

5 Discussion

5.1 Screening Implications

The model demonstrates a fair ability to identify individuals at high risk for colorectal cancer, with performance clearly exceeding that of random chance. While encouraging widespread screening for everyone is not feasible due to the associated healthcare costs, a risk calculator like ours provides a data-driven approach to prioritize screening for individuals who are more likely to benefit from it.

In our validation set, the model successfully reduced the number of individuals recommended for screening from 9,498 to 3,975, a 58% reduction, while unfortunately missing 60 individuals who were later confirmed to have CRC (Figure 7). This highlights the trade-off between reducing unnecessary screenings and the risk of false negatives. To illustrate, out of a sample of 200 participants, the model typically flags 80 as high risk, with 2 individuals in that group actually having CRC. However, 1 individual with CRC was labeled as low risk, which underscores a significant challenge in cancer screening models, that is False positives may strain healthcare systems through unnecessary procedures. False negatives carry more severe consequences, including potential death or legal liabilities.

Cancer risk models should be implemented responsibly. In this study, the model is intended strictly for awareness and early screening encouragement, not for clinical diagnosis. It serves as a tool to prompt individuals to consider medical advice or screening, especially in populations where screening uptake is historically low.

Confusion Matrix	**Predicted No**	**Predicted Yes**
Actual No	5619	3879
Actual Yes	60	96

Fig. 7. Confusion Matrix for Risk Calculator Model

5.2 Key Risk Factors

As anticipated, age was the most influential factor in the CRC risk model. This aligns with the CRC risk increases significantly with age. Yet, given the rising incidence of CRC in younger populations, it's notable that a model excluding age still performed better than random, underscoring the role of other health and demographic factors.

One such predictor is address history, which initially appeared to suggest that frequent relocation could increase CRC risk. However, further analysis clarified that this variable reflects the number of unique addresses a participant has owned, not the frequency of moves. Interestingly, the data show that having more addresses is actually associated with decreased CRC risk. This could be interpreted as a proxy for affluence or stability—individuals who own multiple properties may have better access to healthcare, healthier lifestyles, or reduced exposure to certain environmental stressors.

On the other hand, participants who had lived at the same address for a prolonged period also demonstrated lower risk, suggesting that residential stability—whether through long-term residence or the ability to own multiple homes—may both be protective in different ways. Overall, these findings point to the multifaceted nature of socioeconomic factors in CRC risk, and how seemingly simple variables like address

count can capture deeper insights into a person's economic status, healthcare access, and living environment.

5.3 Limitations and the Increase in Early-Onset CRC

A major limitation of this study is its focus on individuals aged 55 to 75, which reflects the eligibility criteria of the PLCO study. This age range aligns with standard colorectal cancer screening guidelines but excludes younger adults, who are increasingly affected by CRC. According to the American Cancer Society, nearly 20% of CRC diagnoses in 2019 occurred in individuals under 55, a rate that has doubled since 1995 (ACS, 2023). These younger patients often present with more advanced disease and poorer outcomes, making their exclusion from risk modeling particularly concerning.

Because age is such a dominant feature in our model, its applicability to younger populations is likely limited. The absence of younger participants in the dataset prevents us from evaluating how risk manifests earlier in life or from identifying non-age-related predictors that could be useful for early detection. To address this critical gap (Katella, 2024), future studies should focus on young-onset CRC as a separate research domain. One promising direction is to use the SEER database, which includes a broader age range and could allow researchers to study CRC incidence among younger adults in greater depth (Abbott, 2024). Furthermore, classification models specifically trained on younger populations could help develop early-warning tools to support targeted outreach and intervention. These models might rely more heavily on lifestyle, family history, or geographic indicators—rather than age—and serve as a basis for awareness campaigns or screening efforts tailored to underserved or high-risk youth populations.

In particular, the integration of geospatial analysis with sociodemographic and lifestyle data could help identify environmental and systemic risk factors that contribute to early-onset CRC. Linking incidence rates to factors such as housing stability, food access, healthcare availability, and socioeconomic status may reveal patterns not captured in traditional clinical datasets. Although our study used the PLCO trial as a proof-of-concept for risk modeling, future efforts must extend beyond this demographic. Expanding research to younger cohorts is essential not only for addressing the alarming rise in early-onset CRC but also for informing prevention strategies and screening policies.

5.3 Ethical Considerations

Race was included in the model using one-hot encoding which may result in certain racial categories appearing more predictive of CRC risk than others. This model is not a fair representation of all races as 86% of the participants in this data set reported White – Non-Hispanic ethnicity. It is crucial to emphasize that while colorectal cancer incidence and outcomes do vary across racial and ethnic groups, these associations are not causal and the scope of this study cannot infer CRC race disparities. This model should never be used to justify excluding or prioritizing individuals for screening based solely on race.

The dataset used in this study is fully anonymized, and all geocodes were randomized to prevent participant identification. In the risk calculator dashboard, user input is

entirely voluntary, and no historical medical data is required. To maintain privacy and comply with HIPAA standards, the tool does not store or track any user responses. Any future implementations of the tool that involve data storage must undergo strict data governance review to ensure compliance.

One of the most significant ethical concerns is the risk of false negatives which are individuals incorrectly classified as low risk who may go on to develop CRC. Such outcomes could delay screening and diagnosis, potentially leading to serious consequences. While no predictive model is perfect, any model used in public health or clinical settings must undergo rigorous validation and oversight to ensure safety, fairness, and effectiveness.

The tool developed in this study is intended solely for awareness and educational purposes, not as a diagnostic tool. Individuals, especially those over age 45, with a family history of CRC, or other risk factors, are strongly encouraged to undergo regular screening based on medical guidelines, regardless of their risk calculator results.

5.4 Further Reasearch

This study found that incorporating Census-based sociodemographic features into colorectal cancer (CRC) risk models did not significantly improve predictive performance. A key limitation was the lack of geographic and socioeconomic diversity in the PLCO dataset, which was limited to 10 urban study centers. As a result, many participants shared similar environmental and social risk profiles, reducing the utility of Census features for differentiating high-risk vs. low-risk individuals.

Future research should prioritize datasets with broader geographic representation, including rural, suburban, and underserved regions. This would allow for a more nuanced exploration of how social determinants of health such as income inequality, food access, housing stability, and healthcare availability, contribute to CRC risk. Linking cancer incidence data with high-resolution geospatial and environmental datasets (e.g., from SEER or other national cancer registries) could significantly expand the model's ability to detect community-level patterns in CRC risk.

Additionally, while this study used logistic regression for its interpretability and simplicity, more complex, non-linear models may be better suited to uncover subtle patterns within high-dimensional Census data. Techniques such as gradient boosting machines, random forests, or deep learning could identify interactions between variables that were not captured here. However, interpretability must remain a priority if the goal is to integrate these models into user-facing tools for public health screening.

Another critical direction is the investigation of early-onset colorectal cancer, which is rising at an alarming rate. Because the PLCO dataset only includes individuals aged 55 to 75, this study was unable to explore risk factors among younger populations. Future work should leverage datasets like SEER, which includes a wider age range, to build models tailored to predicting CRC risk in individuals under 50. Such models could be invaluable for raising awareness and promoting early screening among younger adults.

Finally, future studies should also examine how lifestyle, environmental exposures, and social instability (e.g., frequent relocation or housing insecurity) interact with biological risk factors to influence cancer risk. These insights could inform the design of more holistic screening interventions and public health strategies.

6 Conclusion

This study developed a logistic regression model to predict colorectal cancer (CRC) risk. While the model performs better than random guessing, it still presents a substantial risk of false negatives, as it incorrectly labels some high-risk individuals as low risk. This highlights the limitations of the model in accurately identifying all at-risk individuals, underlining the importance of complementing risk prediction models with actual clinical screening programs.

The key factors contributing to CRC risk in this model include Age and Address History. Interestingly, even when Age was excluded from the model, the remaining factors (e.g. height, weight, education) still yielded a comparable performance, suggesting that non-age-related factors also play some role in predicting CRC risk.

Incorporating Census data did not significantly improve the model's performance, which may be due to the limited geographic diversity of the study centers in the PLCO study. This indicates that Census features may require more sophisticated feature engineering or a more complex model to uncover their potential value. Geospatial diversity in future datasets may help to better capture the socio-economic and environmental factors that influence CRC risk.

The CRC Risk Calculator developed as part of this study, while imperfect, provides a useful tool for raising awareness and encouraging individuals to seek screening. However, it should not be considered a clinical diagnostic tool. The model's design underscores the trade-offs between simplicity, interpretability, and accuracy, aiming to make risk assessment more accessible while ensuring that individuals who are at risk are encouraged to undergo professional screening.

Future research should focus on enhancing model performance through more diverse datasets, advanced modeling techniques, and greater consideration of lifestyle factors to improve predictions, especially for younger individuals where CRC rates are rising.

7 Appendix

7.1 Feature Sets

Feature Sets Used in LR and NN Models			
Health & Demographics 1	**Health & Demographics 2**	**Health & Demographics 3**	**Census 1**
asppd	asppd	asppd	census_pct_f_unemploy_1990
cig_years	cigpd_f	cigpd_f	census_pct_hh_pub_asst_2000
pack_years	height_f	cig_years	census_pct_nh_blacks_2000
height_f	ibuppd	pack_years	census_hisp_black_2010
weight_f	weight_f	height_f	census_housing_rental_acs5
brothers	brothers	ibuppd	census_pct_crowding_acs5
educat	educat	weight_f	census_pct_medicare_ins_acs5
age	race1	brothers	census_pct_tricare_ins_acs5
bmi_curr	age	educat	census_b01001_009e_acs5
addr_duration_days	addr_duration_days	race1	census_b19056_002e_acs5
addr_total	addr_total	age	
		bmi_curr	
		addr_duration_days	
		addr_total	

Fig 8: Selected Feature sets from PLCO Study data

Acknowledgments - The authors thank the National Cancer Institute for access to NCI's data collected by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial". (PLCO1677)

References

1. Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A., & Jemal, A. (2023). Colorectal cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73(3), 233–254. https://doi.org/10.3322/caac.21772
2. American Cancer Society (ACS). Colorectal Cancer Facts & Figures 2023-2025. Atlanta: American Cancer Society, Inc. 2022.
3. McClelland, P. H., Liu, T., & Ozuner, G. (2022). Early-Onset Colorectal Cancer in Patients under 50 Years of Age: Demographics, Disease Characteristics, and Survival. Clinical Colorectal Cancer, 21(2), e135–e144. https://doi.org/10.1016/j.clcc.2021.11.003
4. Akimoto, N., Ugai, T., Zhong, R., Hamada, T., Fujiyoshi, K., Giannakis, M., Wu, K., Cao, Y., Ng, K., & Ogino, S. (2021). Rising incidence of early-onset colorectal cancer - a call to action. Nature Reviews. Clinical Oncology, 18(4), 230–243. https://doi.org/10.1038/s41571-020-00445-1
5. O'Sullivan, D. E., Sutherland, R. L., Town, S., Chow, K., Fan, J., Forbes, N., Heitman, S. J., Hilsden, R. J., & Brenner, D. R. (2022). Risk factors for early-onset colorectal cancer: A

systematic review and meta-analysis. Clinical Gastroenterology and Hepatology, 20(6). https://doi.org/10.1016/j.cgh.2021.01.037

6. Katella, K. (2024, January 17). Colorectal cancer: What millennials and gen zers need to know. Yale Medicine. https://www.yalemedicine.org/news/colorectal-cancer-in-young-people
7. Ladabaum, U., Mannalithara, A., Meester, R. G. S., Gupta, S., & Schoen, R. E. (2019). Cost effectiveness and national effects of initiating colorectal cancer screening for average-risk persons at age 45 years instead of 50 Years. Gastroenterology, 157(1), 137–148. https://doi.org/10.1053/j.gastro.2019.03.023
8. Dong, W., Kim, U., Rose, J., Hoehn, R. S., Kucmanic, M., Eom, K., Li, S., Berger, N. A., & Koroukian, S. M. (2023). Geographic variation and risk factor association of early versus late onset colorectal cancer. Cancers, 15(4), 1006. https://doi.org/10.3390/cancers15041006
9. Rogers CR, Moore JX, Qeadan F, Gu LY, Huntington MS, Holowatyj AN. Examining factors underlying geographic disparities in early-onset colorectal cancer survival among men in the United States. Am J Cancer Res. 2020 May 1;10(5):1592-1607. PMID: 32509399; PMCID: PMC7269786.
10. Gupta, S. (2022). Screening for Colorectal Cancer. Hematology/Oncology Clinics of North America, 36(3), 393–414. https://doi.org/10.1016/j.hoc.2022.02.001
11. Cardoso, L. B., Parro, V. C., Peres, S. V., Curado, M. P., Fernandes, G. A., Filho, V. W., & Toporcov, T. N. (2023). Machine learning for predicting survival of colorectal cancer patients. Scientific Reports, 13(1), 8874–8874. https://doi.org/10.1038/s41598-023-35649-9
12. Abdul Rahman, H., Ottom, M. A., & Dinov, I. D. (2023). Machine learning-based colorectal cancer prediction using global dietary data. BMC Cancer, 23(1), 144–144. https://doi.org/10.1186/s12885-023-10587-x
13. Nascimento, A. Q., Dantas, D. B., Melo, G. S., Gomes, F. de C., & de Melo Neto, J. S. (2022). Impact of sociodemographic factors and screening, diagnosis, and treatment strategies on colorectal cancer mortality in Brazil: A 20-year ecological study. PloS One, 17(9), e0274572–e0274572. https://doi.org/10.1371/journal.pone.0274572
14. Torres, A. Z., Phelan-Emrick, D., & Castillo-Salgado, C. (2018). Evaluating neighborhood correlates and geospatial distribution of breast, cervical, and colorectal cancer incidence. Frontiers in Oncology, 8. https://doi.org/10.3389/fonc.2018.00471
15. Abbott, B., & Blesener, S. (2024, January 25). More young people are getting cancer, leaving doctors alarmed and baffled. The Wall Street Journal. https://www.wsj.com/story/more-young-people-are-getting-cancer-leaving-doctors-alarmed-and-baffled-0540eacc
16. Driver, J. A., Gaziano, J. M., Gelber, R. P., Lee, I.-M., Buring, J. E., & Kurth, T. (2007). Development of a risk score for colorectal cancer in men. The American Journal of Medicine, 120(3), 257–263. https://doi.org/10.1016/j.amjmed.2006.05.055
17. Zheng, S., Schrijvers, J. J., Greuter, M. J., Kats-Ugurlu, G., Lu, W., & de Bock, G. H. (2023). Effectiveness of Colorectal Cancer (CRC) screening on all-cause and CRC-specific mortality reduction: A systematic review and meta-analysis. Cancers, 15(7), 1948. https://doi.org/10.3390/cancers15071948
18. Zhang, Y., Sheng, C., Fan, Z., Liu, Y., Liu, X., Duan, H., Dai, H., Lyu, Z., Yang, L., Song, F., Song, F., Huang, Y., & Chen, K. (2024). Risk-stratified screening and colorectal cancer incidence and mortality: A retrospective study from the prostate, Lung, colorectal, and Ovarian Cancer Screening trial. Preventive Medicine, 187, 108117. https://doi.org/10.1016/j.ypmed.2024.108117
19. Tammemägi, M. C., Katki, H. A., Hocking, W. G., Church, T. R., Caporaso, N., Kvale, P. A., Chaturvedi, A. K., Silvestri, G. A., Riley, T. L., Commins, J., & Berg, C. D. (2013). Selection criteria for Lung-cancer screening. New England Journal of Medicine, 368(8), 728–736. https://doi.org/10.1056/nejmoa1211776

20. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229–263. https://doi.org/10.3322/caac.21834