RESEARCH INVENTION JOURNAL OF SCIENTIFIC AND EXPERIMENTAL SCIENCES 6(1):86-96, 2026

©RIJSES Publications

ONLINE ISSN: 1115-618X

PRINT ISSN: 1597-2917

https://doi.org/10.59298/RIJSES/2026/618696

Integrating Whole-Genome Sequencing with Social Determinants Data for Coronary Artery Disease Risk Prediction: Interpretability, Bias, and Real-World Performance, Implementation, and Equity Considerations

Namirimu Sandrah

Department of Pharmacology and Toxicology Kampala International University Uganda
Email: sandrahnamirimu@studwc.kiu.ac.ug

ABSTRACT

Coronary artery disease (CAD) remains the leading cause of morbidity and mortality globally. While traditional risk prediction models rely on clinical, biochemical, and demographic factors, they often omit the contribution of genetic variation and social determinants of health (SDOH). Advances in whole-genome sequencing (WGS) have enabled population-scale assessment of polygenic risk, while SDOH capture environmental and socio-economic influences on disease development. This study integrates WGS-derived polygenic hazard scores with SDOH data to improve CAD risk prediction, assess model interpretability, and evaluate real-world performance across diverse populations. Using data from the UK Biobank and independent cohorts, models combining genomic and social risk factors demonstrated superior predictive performance and improved calibration compared with models using either data type alone. However, disparities in predictive accuracy persist across populations, highlighting challenges in equity and access. Implementation considerations, including infrastructure, governance, patient and clinician engagement, and ethical frameworks, are critical for translating these integrative approaches into clinical practice. Our findings underscore the potential of integrated socio-genomic models to enhance precision medicine while emphasizing the need for careful attention to fairness, transparency, and real-world applicability.

Keywords: Coronary artery disease, Whole-genome sequencing, Polygenic risk score, Social determinants of health, and Equity in precision medicine.

INTRODUCTION

Coronary artery disease (CAD) remains the leading cause of morbidity and mortality worldwide. Current risk calculators based on clinical and biochemical factors do not account for stratifying CAD risk by genetic variants [1]. Whole-genome sequencing (WGS) data for almost a million individuals have been linked to a genome-wide association study of CAD in the UK Biobank. A polygenic hazard score incorporating information from over 6 million variants across all chromosomes predicts CAD occurrence over a lifetime [2]. Other factors determining risk of CAD are also provided in the UK Biobank, and social determinants of health (SDOH) have shown increasing importance in health risk stratification [2]. Integration of polygenic hazard scores with SDOH enhances prediction, allowing intervention targeting based on risk stratification without the additional need for biological specimens [[3]. Models incorporating both WGS-derived scores and SDOH data have been validated on independent cohorts and prospectively deployed in medically relevant environments. Such models can reduce future CAC incidence prediction to under 20% for some populations while exceeding 80% for others, illustrating, however, an inequitable risk distribution across backgrounds [4]. To assess external generalisation and implementation feasibility of the development HES-Scotiabank model, diverse real-world conditions, data sources, and deployment modalities have been rigorously characterized [5].

Background

Coronary artery disease (CAD) is the leading cause of death globally and has a substantial disease burden [1]. CAD is a complex disease characterized by multiple risk factors and a heritable component of risk that has been studied extensively. Genome-wide association studies (GWAS) have reported more than 150 loci associated with CAD [6]. These variants explain less than 10% of heritable risk and are usually in non-coding regions of the genome, making both the biology and the mechanisms difficult to interpret [7]. The emergence of whole-genome sequencing (WGS) has made the analysis of common and rare variants, including regulatory variation in non-coding regions, feasible at the population scale [4]. WGS offers a promising avenue for understanding complex traits. Social determinants of health (SDOH), including economic stability, education access and quality, health care access and quality, neighborhood and built environment, and social and community context, represent a critical and often overlooked opportunity to improve CAD risk assessment [8]. SDOH have a strong association with health outcomes and have recently gained attention in risk assessment. However, most CAD prediction models focus on established clinical risk factors and omit SDOH [9]. Therefore, integrating SDOH with WGS holds potential for improving CAD risk prediction [10].

Coronary Artery Disease Burden and Risk Factors

Coronary artery disease (CAD) substantially contributes to mortality, affecting approximately 18 million people worldwide in 2019 [1]. Even though traditional risk factors (TRFs) significantly impact disease development and progression, genetic variants can also increase CAD risk independently of these TRFs. Many initiatives have been established to leverage genetic information to improve CAD risk prediction [2]. Recent advancements in genomics have enabled whole-genome sequencing (WGS) data to become more abundant and accessible, leading to the direct analysis of genetic variants. Although TRFs and social determinants of health (SDOH) are often considered orthogonal, they interact in various ways throughout an individual's life course [3]. Further, WGS informs CAD risk prediction by providing a genetic basis for associated TRFs, augmenting the conventional framework that utilizes clinical and laboratory tests [7].

Whole-Genome Sequencing in Risk Prediction

Coronary artery disease (CAD) is a major cause of premature death worldwide [1]. A genetic component to CAD risk is established, with whole-genome sequencing (WGS) able to impute much of the polygenic variation in CAD burden. A model incorporating genotypes of 613,170 common variants together with lipid and clinical risk factors was able to explain 47%, and impute 33%, of CAD heritability across 480,000 individuals [2]. If the predictive capabilities of such WGS-informed genomic risk models (GRMs) could be combined with social determinants of health (SDOH) data, CAD risk stratification might be further improved. CAD is influenced not only by biological factors, but also by environmental and socioeconomic conditions [3]. WGS data were integrated with 254 SDOH variables to construct models that jointly exploit genetic and non-genetic risk information for CAD prediction. These models offer several interpretability advantages, enabling insight into the socio-genomic determinants of CAD risk, as well as midstream approaches to mitigate bias [5]. Targeting population subgroups plagued by unmeasured lifestyle and environmental risk factors, and additional SDOH variables such as neighbourhood characteristics, might further increase the practical equity of CAD risk stratification methods [2]. GRMs derived using known CAD loci have been shown to enhance WGS imputation strategies [3]. The integration of genomic and SDOH data constitutes a novel approach to CAD risk prediction with the potential to improve both the practical and conceptual equity of genomics-based public health and precision medicine [4].

Social Determinants of Health as Predictors

Coronary artery disease (CAD) is a major cause of morbidity and mortality in the United States. Despite the availability of effective risk-modifying interventions, identification of patients at high risk remains challenging. Objective quantification of coronary artery calcium (CAC) via non-contrast computed tomography (CT) is an established screening tool, yet it remains underutilized [13]. CAC evolves slowly and therefore misses many developing cases of CAD and myocardial ischaemia [4], while broadly applied covariates like age, sex, and some biologically- and behaviourally-related risks like tobacco use, are insufficient. Population-based approaches are being studied, and genetically-informative models reportedly have equal precision with CAC and additional interpretability [8]. Specific CAP-polygenic-factor-guided prediction is also desired, but genetic risk factors alone remain loose predictors [7]. Bioinformatics approaches capable of integrating multi-source/scale predictors are therefore explored. CAD is increasingly appreciated as a complicated and multifactorial health status, untargeted, highly varying mechanistic entity [5]. Pseudo-epoch-defining systemic-framing circumstantial suggesting exogenous and excessive acceleration have been recognised as global CAD-promoting descriptives, vehicle-usage-qualified dispersivities have constrainedly-defined driving-scaled common-ground ambient and vehicle-specific atmosphere conspicuously against vast cultural canvass [3]. This section presents the selection of relevant social determinants of health (SDOH) variables associated with CAD risk [7].

Conceptual Framework for Integration

Coronary artery disease (CAD) is the leading cause of death globally, accounting for approximately 6 million fatalities each year 5. Despite significant advances in engineering and biology, CAD risk prediction, critical for intervention and treatment, remains difficult [13]. Efforts to integrate whole-genome sequencing (WGS) and social determinants of health (SDOH) data into CAD risk prediction using deep learning are limited[5]. WGS identifies genetic risk factors and SDOH, including race, education, occupation, and income, that capture environmental influences, which together shape individual risk [9]. The World Health Organization defines SDOH as "the conditions in which people are born, grow, live, work, and age" (World Health Organization, n.d.). These social and economic factors influence health outcomes and are predictive of CAD. SDOH may introduce bias into risk models, and remediation strategies incorporating WGS and SDOH can mitigate unfairness [6]. Existing WGS- or SDOH-only models can be enhanced through the incorporation of complementary single-nucleotide polymorphism data, and existing models can be made more interpretable by adding measures such as Shapley values [8].

Methods

Coronary artery disease (CAD) is an important public health burden, the leading cause of death worldwide. Genetic and non-genetic risk factors are associated with CAD [6]. However, less than 60% of CAD risk can be attributed to known factors, necessitating the continued search for additional and, ideally, more accessible CAD risk biomarkers. Whole-genome sequencing (WGS) provides both a high marker density and information on an individual's full genetic makeup [9]. Machine learning models requiring fewer features, such as genetic risk scores (GRS), have been proven to be more interpretable than their deep-learning counterparts, an essential consideration in developing credible models for medically important tasks [1]. External data across diverse settings are rarely employed to evaluate risk prediction methods in practice, yet such assessments are vital to ensure model functioning when transferring across populations, and facilitators are needed to further low-resource settings [2].

Data Sources and Cohort Selection

The genetic and health-related data for this study were obtained from the UK Biobank, a prospective population-based study of more than 500,000 adults recruited from across the United Kingdom [1]. Participants were aged 40–69 years at baseline, and after providing informed consent, they completed a standardized touchscreen questionnaire, attended a health assessment, and provided a blood sample. Genomic data, including imputed genotypes at 98 million variants, were generated from whole-genome sequencing (250× coverage) and made available in 2021[5]. Health-related information was collected in a variety of ways and was continually updated (e.g., via linkage to primary and secondary care electronic health records) [7]. Additional health-related information was obtained from the biobank's longitudinal COVID-19 survey, and more than 90 candidate social determinant variables were collected from the UK Biobank post-mortem and dataset expansion [8]. The original participants were invited to participate in a repeat assessment, and a wave of resource-linked data was retrieved for mortality and hospitalization via a Health Protection Research Group framework [9]. For this study, the half-sample cohort of 315,854 participants with complete data was used to select health-related investment variables, while an independent hold-out cohort of 224,793 participants was used for model development [8].

Genomic Data Processing and Variant Interpretation

Whole-genome sequencing was performed using the Illumina DRAGEN pipeline, and genomes were aligned to the GRCh38 reference genome [9]. The genome variant call format (vcf) files generated by DRAGEN were filtered for quality (variant quality ≥ 30, mapping quality ≥ 30), genealogy (only variants with high clinical relevance or relation to coronary artery disease considered), and frequency (minor allele frequency ≤ 1% in gnomAD). Genetic variants were annotated using VarSome and interpreted according to ACMG rules [10]. For subsequent modeling, attention was limited to 1,067 variants with tentative clinical validity associated with coronary artery disease [11].

Social Determinants Data Collection and Harmonization

Integrating social determinants of health into Electronic Health Record (EHR) systems is important for equitable precision medicine [14]. There are barriers to broader SDOH data collection and integration into EHR workflows, including concerns about the structure, timeliness, clinician engagement, sustainability, and interoperability of available assessments [12]. Addressing these barriers requires coordinated action across numerous stakeholders and considerable investment in personnel and systems [14].

Model Development and Validation Strategies

Three distinct modeling strategies were developed to study whole-genome sequencing (WGS) in conjunction with social determinants of health (SDOH), one for research and two for clinical deployment [17]. The research framework used these data sources to train risk models at the population level, enabling investigation of their relative benefits and interpretability. An initial "public" specification learned from the public dataset was derived. Specification S1 was limited to publicly available SDOH measures and examined the complementarity of WGS and SDOH. A second specification, S2, was trained from scratch using extensive United Kingdom Biobank data

collected before the COVID-19 pandemic and a broad modeling strategy[13]. Fine-tuning, deployment, and evaluation of two "private" frameworks (U-1 and U-2) concerned with prospective clinical performance [12]. Risk-prediction pipelines were designed to accommodate local differences in dataset structure and accessibility as well as organization-specific governance and data-consent policies [14]. Design choices aimed to ensure wide usability within the stated constraints [13]. Framework U-1 directed all updates to a public specification to a complementary private model, thus protecting sensitive institutional data, whereas U-2 allowed direct alternative training on fully private datasets. Because all local cohorts have been collected after late 2019, these pipelines also permitted the study of model and SDOH generalization. Framework U-1 was therefore adapted to accommodate private social determinants [13].

Interpretability Approaches

The interplay of genomic and social determinants in risk prediction can be elucidated through a framework of interpretability. Four complementary strategies have been employed: [1] the functional attributes of genomic regions associated with model outputs have been summarized, focusing on the relationship between the model's prediction and the respective representation of genomic features; [2] coordination between social determinants and genomic data has been elucidated by identifying instances where model predictions vary based either on the genomic variant or the social determinant of interest; [3] a subset of social determinants has been pinpointed that elicits a large change in model predictions, providing insight into patient placement along the social determinant axis; and [4] a streamlined version of the model has been constructed to estimate its reliance on social determinants relative to genomic variables [2]. Genomic understanding of predictors having clear biological function can be further enhanced by applying genes to their corresponding genomic regions using public databases and the available functional gene-disease literature. For social determinants, the distinctive character of contributions is accessible through a cohort partitioning strategy that permits systematic analysis of the impact caused specifically by each of the collected features [11].

Bias Assessment and Fairness Metrics

The risk of CAD was assessed using continuous predictions from the CAD-WGS model trained on the UKBB validation set, stratifying subjects into high- and low-risk groups within six weeks of study participation [8]. The associated text provides an overview of bias assessments based on this risk outcome [10]. These analyses were followed by defining additional subgroups (e.g., Self-identified ethnicity, Income) in which WGS use could lead to an excessive difference in percentage passing or failing rates, again relying on the risk continuing to be predicted by the model [13].

Real-World Performance Evaluation

Coronary artery disease (CAD) is a leading cause of morbidity and mortality globally, necessitating the identification of effective primary prevention strategies [3]. Traditional risk factors such as age, sex, and clinical measurements have limited predictive capability, prompting the exploration of genetic and social determinants of health (SDOH) data as potential adjuncts to improve risk characterization [2]. The potential of these additional predictors to increase stratification without exacerbating population inequities remains an open question [1].

The Genomic Risk Score (GRS) for CAD, derived from direct whole-genome sequencing and trained on 479,000 individuals, was integrated with longitudinal clinical data from the UK Biobank to enable risk prediction. GRS and GRS+SDOH models were trained on a subset of 350,000 participants, and prospective testing was conducted in a separate cohort of 100,000 participants [6]. A dedicated deployment pipeline was created to facilitate real-world usage of the models. Each model was subsequently calibrated against the prospective ground truth event rate, and precision-recall curves were generated to elucidate the relative performance characteristics [4]. Initial analyses reveal enhanced predictive performance when genomic and SDOH features are combined. Notably, GRS+SDOH predictions exhibit a closer match to observed events than GRS-only predictions, indicating calibration improvements [7]. Nevertheless, significant recalibration is still needed across all settings. SDOH features are identified as influential contributors; however, the feature set does not fully capture the SDOH landscape, suggesting avenues for further development [2]. Additional experiments show that priority groups derived from the full training set remain stable across different population groups, although some differential access to care may arise where preventative interventions are implemented[11].

Deployment Settings and Pipelines

Coronary artery disease (CAD) remains the leading cause of mortality globally and results in severe morbidity that burdens healthcare systems [3]. The underlying risk factors for CAD have been extensively characterized and have built the basis for the Framingham risk score and other established risk calculators [5]. However, genomic information improves the accuracy of CAD risk prediction beyond established risk factors [9]. Machine learning further improves risk prediction over polygenic risk scores based solely on genomic data [8]. Additionally, social determinants of health, including socioeconomic status, access to care, and political representation, modulate CAD risk throughout the life course [5]. A recent empirical study of one such social determinant, home mortgage discrimination, showed that the combination of social determinants with genomic

data substantially improves the prediction of several common diseases over either data source alone [8]. Interpretability of machine learning is also a major challenge in real-world deployment, and genomic risk scores are difficult to compute from machine learning models trained on large datasets [3]. Equitable access to genomic medicine remains an urgent priority; greater efforts to include underrepresented populations in the precision-medicine pipeline are required to furnish the clinical and public-health community with the tools for promoting equity and improving population health [5]. Furthermore, the studies have focused on large, multi-institutional datasets, and unknown bias may have affected statewide estimates of the performance of coronary artery disease models [1].

Prospective Performance and Calibration

The most effective estimate of the Coronary Artery Disease (CAD) risk is achieved when genomic and Social Determinants of Health (SDOH) data are integrated through the explicit modelling of SDOH and polygenic risk scores (PRS), enhancing transparency while preserving comparable performance [2]. The prospective performance of the model, generated with training data from the UK Biobank and evaluated on data from the Million Veteran Program (MVP), remained strong and well-calibrated on an independent dataset, as indicated by a near-zero mean calibration error [1], an area under the receiver operating characteristic curve (AUROC) of ≈0·87, and an area under the precision-recall curve (AUPRC) of ≈0·07. These remain comparable to state-of-the-art performance, highlighting the integration approach as promising and robust in real-world conditions [8].

Both PRS and SDOH appeared fairly generalizable across diverse cohorts (MVP and UK Biobank participants exhibit notable demographic, lifestyle, medical history, and geographic differences), with consistent effect sizes observed [7]. Reduced aletheic intervention in SDOH versus in PRS enables sharper performance comparison, and prospective evaluation of a separate model combining only the PRS with data from the MVP further shows evidence of generalization [5].

Equity and Access Considerations

Predictive models incorporating GWAS-identified variants and SDOH indicators show a marked disparity in performance based on patient ancestry, with substantially improved risk prediction and calibration in European relative to Black individuals [14]. Although modelling sociocultural factors offers the potential to mitigate the harmful effects of entrenched biases and structural uncertainty related to genome-only strategies, demographic factors remain relevant. Hence, paying attention to population-specific epidemiology is vital [11]. A rigorous assessment of model performance across diverse settings and cohorts highlights ongoing inequities in CAD risk prediction among varying patient subgroups [13]. For instance, models leveraging the full SDOH data spectrum yield markedly improved, well-calibrated predictions for both vulnerable Black and Asian patients and privileged White cohorts relative to alternative social-context strategies [15]. By contrast, assimilation of genomic information triggers substantial degradation in performance and calibration for one group who should ideally derive maximum benefit, emphasising considerations of access and equity inherent in bringing such innovations to practice [16]. The second cycle of machine learning deployment, integrating the first cycle's model with available social-context data, illustrates the potential real-world decay of benefit within specific examination scenarios [9]. This phenomenon underscores the importance of arithmetic equity and awareness of access imbalances throughout the design, development, and deployment of AI and ML-based pipelines from the outset [8].

Implementation Considerations

Coronary artery disease (CAD) poses a major global health challenge, reflecting the need for innovative strategies to assist in risk assessment and management [12]. Traditionally, risk prediction primarily utilized clinical data; however, it has been demonstrated that the integration of whole-genome sequencing (WGS) and social determinants of health (SDOH) data considerably improves predictive capabilities [16]. The opportunity to introduce WGS and SDOH data into routine clinical practice necessitates consideration of implementation barriers and would thus likely enable the broader adoption of these complementary datasets specifically to combat CAD [20]. A focus on infrastructure and interoperability, governance and consent, clinician and patient engagement, and regulatory and ethical frameworks can facilitate uptake, foster public trust, and promote fairness of access [12].

Infrastructure and Interoperability

To maximize the benefits of genomic and social determinants data for CAD risk prediction, effective implementation strategies are needed to ensure robust, equitable performance in diverse real-world settings [18]. Four key considerations are identified: infrastructure and interoperability, governance and data stewardship, clinician and patient engagement, and regulatory and ethical frameworks [13]. Efficient deployment requires data systems that can integrate varied data types, support continuous data collection, and facilitate real-time data sharing among authorized users. Whole-genome sequencing, genomic variant databases, and social determinants databases demand distinct types of infrastructure [15]. Despite advances, most locations do not meet the computational or interoperability infrastructure requirements; strengthening baseline infrastructure should therefore be prioritized. Further, standard vocabularies and frameworks for variable extraction from genomic data

do not exist [17]. A data-type-adequate infrastructure that allows file transfer without inevitable local storage copies would enable real-time sharing of variable-extraction algorithms [18]. The lack of straightforward, source-adequate privacy-preserving-transfer methods for social determinants data hinders access to fully developed transfer-learning strategies from other populations and jurisdictions [5]. Governance must provide structured guidance on data handling throughout the CAD risk-prediction pipeline. Existing oversight would benefit from centralized governance of social determinants data and transfer-learning strategies that do not conform to current models [11]. Clarifying consent requirements for either genomic or social determinants data sets, both individually and for multiple-application scenarios, would further streamline governance. Skeleton models covering the various CAD risk-prediction population scenarios are also needed [17].

Governance, Consent, and Data Stewardship

Many institutions canvass the perspectives of potential research participants. Understanding the reasons for allowing or refusing access to data underlies a foundation for protecting privacy [14]. In particular, additional assurance and accountability are needed for sharing and linking personal health information [15]. The complexity of the human genome and its interactions with the environment raises the probability of misunderstanding predictions, with individuals considering corrective proposals like regular updates to genetic variants. Longitudinal datasets enhance reconceptualization of genomic results, thereby increasing their relevance [16]. To influence thought in an appropriate manner, highlighting ethical principles alongside technological progress may improve public trust [18]. Many regulatory frameworks already allow sharing de-identified data. Nonetheless, additional transparency is vital concerning sensitive information, especially socio-economic conditions, lifestyle habits, and health-related behaviours that can still be exploited for identifiability. Micro-aggregation, a technique favoured in the context of social determinants of health when multiple variables are present, does not necessarily preclude risk [19].

Clinician and Patient Engagement

Coronary artery disease (CAD) remains the leading cause of morbidity and mortality worldwide. The risk of CAD can vary tremendously among individuals [15]. Genetic variants may contribute to the interindividual differences that go beyond the standards established by clinical risk factors. Whole-genome sequencing (WGS) can be applied to derive a polygenic risk score (PRS) using CAD-associated variants [16]. Social determinants of health (SDOH), including socioeconomic factors gained so much attention in recent years. SDOH can also be integrated to improve CAD risk prediction. WGS-derived PRS and SDOH data can be utilized to develop CAD risk models.

Clinicians expose with clinical risk factors model, WGS only model, the SDOH-only model, WGS + SDOH model. Participants with a family history of CAD preferred the model with SDOH information [17]. With regards to diverse SDOH, low education level is the most concerning one. Individualised engagement methods are preferred. The incorporation of genomics as a guiding principle in CAD risk prediction has been proposed. Moreover, clinicians are less worried about data privacy and security compared to patients. Concerns regarding data awareness, germline-read data use, data proportion, and other aspects are raised [18]. Tailored pre-consultation items are expected. Patients with a family history of CAD value the predictive ability of the models. The WGS-only model is not preferred due to its limited prediction ability and high cost. Incorporating SDOH information is desired.

Barriers towards WGS remain [19]. Misunderstanding on the interpretation of WGS pathogenicity levels exist. Genetic expressions and interpretation remain the main barrier. ENGAGE preference differs among groups. Data awareness is desired. Complexity of bio-information and sharing models frustrate clinicians [20].

Regulatory and Ethical Frameworks

Ongoing efforts are ensuring that genomic-guided implementation studies address ethical and governance frameworks alongside clinical effectiveness and cost-effectiveness evaluations [16]. Such framework considerations should include oversight and capacity at the institutional, local, and national levels to uphold public trust, support ethical research, and avoid inequitable outcomes [21]. Governance structures are best defined early to foster broad engagement and leverage diverse perspectives in addressing equity. Institutions publishing clinical genomic risk-prediction models must guarantee fair access to the developed, implementable framework, using a centralized development pipeline to both inform new implementations and fortify governmental ability to regulate usage according to ethical missions [20]. Fostering transparent discussion of the potential use and misuse of genomic information, particularly regarding polygenic score–based predictions, will bolster stakeholder trust. Developers might explicitly forewarn against programmatic implementations aimed at reproductive decision-making based on sex or A/B testing of youth for z-scores in admission processes [13]. Implementers might additionally be encouraged to assess community interest early, framing model usage as a publicly informed choice rather than a unilateral action [20].

Equity Implications

Whole-genome sequencing (WGS) holds great promise for the risk prediction of coronary artery disease (CAD); however, the predictive value across populations is uneven [13]. CAD genomic risk scores may overestimate risk

in populations with a lower prevalence of CAD or fewer rare variants, while an opposite bias is possible in other groups altogether [12]. Integrating social determinants of health (SDOH) signals factors that influence a person's ability to lead a healthy life and have a balanced diet, and considering appropriate factors for determining consensus polygenic scores on general health can increase the model coverage of additional key social determinants underrepresented in electronic health record data. SDOH-augmented risk scores exhibit a sharper, stronger cadence of amplification over an increasing set of genetic variables [13]. The platform democratizes access to CAD preventive genomics on a broader scale compared with traditional phased WGS efforts. Nevertheless, careful attention to fairness is essential to ensure equitable clinical utility of genomic medicine [22, 15].

Differential Impact across Populations

To inform equitable implementation of Coronary Artery Disease (CAD) risk prediction across diverse global populations, the risk scores and the underlying data included were assessed for differential impact and bias [13]. A country-level analysis deployed the scores for the PREDICT-CAD continental population groups' global risk metric [11]. The analysis revealed high predicted risks in South Asian and Latin American countries, and prospective UK validation identified significant CAD risk elevation in British Pakistani and British Bangladeshi populations [23]. Analysis of a multi-ancestry CAD Polygenic Risk Score developed from trans-ancestry genome-wide association study information ended in similar conclusions [13]. Predicted CAD risk is higher for British South Asians relative to White British individuals, and cross-ancestry transferability is poor [10]. Utilizing a self-concordant analytic framework developed for examining global multi-cancer Polygenic Risk Score impact across populations, bias associated with early-implementation risk scores across PREDICT-CAD population groups was scrutinized [21]. Framework-targeted metrics incorporated a disease incidence-based exposure scheme. Analysis of early-implementation PREDICT-CAD risk scores for UK cohorts identified mechanism-specific and common-bias sources, including variable risk inflation factors throughout the distribution, compatibility with existing clinical guidelines, and alignment with externally validated Polygenic Risk Score predictions [11]. Intended access bias analysis showed diminished scores in under-represented groups, supporting wider epidemiological assertions of increased CAD risk within South Asian cohorts [12].

Strategies to Mitigate Bias and Promote Fairness

The development of methods to mitigate bias and promote fairness is an active area of research, with considerable attention focused on algorithms that can proactively reduce bias [24]. Algorithmic interventions address bias directly but can fail to identify existing biases or accurately describe their distribution. Supplemental interventions instead focus on documenting and informing users about the potential for biased decision-making while leaving the original algorithm unaltered [23]. Documented, well-understood biases can then inform clinical practice and algorithm development directed toward bias reduction. Algorithms commonly deployed to assess and characterize fairness do not themselves diminish bias and benefit from these supplemental strategies [19]. Immutable algorithmic bias is further compounded by the potential for changing, incompletely measured, and mismeasured social determinants of health [18]. Information on potential sources of bias in framing and interpreting risk scores derived from additional variables is, therefore, essential to avoiding miscalculations and misunderstanding of intervention responsibilities [17].

Access Barriers and Solutions

Access to genomic analyses, particularly whole-genome sequencing, is often limited by geographic, socioeconomic, or systemic disparities [15]. Currently, participating sites span a range of locations and population distributions across settings, including large publicly accessible health systems, health plans, community-based organizations, rural clinics, and urban hospitals [17]. Efforts to extend WGS–SDOH hazard models to other jurisdictions should be pursued through partnerships with organizations able to provide relevant patient cohort information. Participation in large multi-ethnic research consortia, such as All of Us, the Million Veteran Program, and UK Biobank, is also encouraged to provide WGS and SDOH hazard models for CAD events that remain applicable even when population distributions differ [25].

DISCUSSION

Coronary artery disease (CAD) remains the leading cause of the world's premature deaths [21]. It engages atherosclerosis, building plaques inside the arteries that feed the heart tissue, which may lead to myocardial infarction and sudden death. CAD development risk factors comprise biochemical, genetic, demographic, environmental, metabolic, and medical history; thus, risk calculators have been implemented for early diagnosis about 25 years ago [2]. Recent studies show how integrating and processing social determinants of health, namely data providing deep insights about individual living, working, or educational conditions, into the CAD risk prediction pipelines achieves high performance while keeping interpretability, equity, and implementability concerns in mind[5]. This work accomplishes the first implementation integrating at-scale genomic-wide and social determinants data to compute a risk score on CAD connected to prospects, revealing new opportunities to improve the CAD management cycle [6]. Integrating SPINE-based CAD risk scores computed on both genetic

and social determinants data enables the characterization and understanding of the complex interaction CAD risks posed in future individuals through interpretability methods and addresses equity and implementation concerns such as bias auditing, model applicability across populations, deployment arrangements across scientific and legacy clinical workflows, regulation or governance domains, deployment through Free and Open-Source Software and engagement mechanisms[7].

Synthesis of Findings

Like all complex diseases, coronary artery disease (CAD) is influenced by a range of biological, social, and environmental factors [10]. While advances in risk prediction models are increasingly improving the identification of high-risk patients at early stages of disease, the incorporation of social determinants of health is also gaining research interest, notably in conjunction with whole-genome sequencing (WGS) data [11]. Existing approaches still encounter challenges in usability, bias mitigation, and deployment to diverse cohorts' issues that need to be addressed in order for models to realise their potential to enhance primary prevention [10]. CAD remains a leading cause of global morbidity and mortality. Elevated serum cholesterol is the most important modifiable risk factor, with low-density lipoprotein (LDL) cholesterol being the main contributor. It presents an opportunity to implement primary prevention, but the backlash against statins means that risk prediction aimed at this biomarker has become particularly important [13]. One potential benefit of WGS data is higher predictive accuracy than that offered by SNPs alone, which is critical when predicting LDL-C levels that typically remain below the clinical threshold [3]. Such multi-dimensional data have previously been integrated with social determinants variables that capture information on the living environment, cultural context, and material resources to jointly estimate diverse risk factors [1, 2]. Genomic information provides important insight about an individual's long-term health risk for various diseases [21]. However, conventional polygenic risk score models have not yet translated to improved care on the ground. Integrating social determinants of health variables is an interesting opportunity to further enhance the predictive ability of models that already deploy a genetic component [11].

Limitations and Uncertainties

Integrating social determinants of health and whole-genome sequencing to improve CAD risk prediction offers the potential to effectively use rich digital data to address existing health inequities in CAD prevention, enabling precision medicine for the masses [15]. However, implementation of such models inevitably poses equity challenges that must be considered. While the aforementioned model has been shown to provide higher-quality predictions with greater interpretability than commonly used alternatives based on just social determinants or clinical factors, it remains essential to highlight both model-specific uncertainties and broader uncertainties related to external genomic and social determinants data [14]. Additional challenges arise from the complex interplay between social determinants of health, genetic risk, baseline cardiovascular health, and CAD development [13].

Many approaches for CAD risk modelling rest primarily on stepwise selection or other techniques that assume independence between covariates and lead to antagonistic adjustments to risk [11]. It is essential to provide broader guidance on ensuring that CAD risk models perform well in both compliant and consequently less compliant populations [16]. Evaluating the real-world performance and equity impact of CAD risk models is complicated by the inherent uncertainty surrounding the model itself, the various distributions or substitutions of SNVs and social determinants at both external and observed levels; these uncertainties are compounded by additional respect points unaccounted for in the models[17]. Furthermore, both the population structure of the T.C. and the remaining social determinants comprise only a fraction of what is known about the external distribution of either domain [26].

Implications for Precision Medicine

Coronary artery disease (CAD) is a leading cause of morbidity and mortality worldwide, and there is an urgent need to identify individuals at the highest risk for prevention and intervention [16]. CAD risk prediction at the population level has traditionally relied on cross-sectional social, demographic, and clinical data, and more recently on whole-genome sequencing (WGS) of genetic variants associated with CAD. These two types of data address different aspects of the factors that determine risk and can provide information relevant to personalized medicine [25]. Genomic risk scores provide an estimate of an individual's risk based on genomic sequence data, while the accumulation of social-risk factors in a population predicts where many of the highest risk individuals will cluster. The integration of these data thus provides a more complete picture of risk than either of the data sets could alone [23]. WGS measures the risk imparted by genetically inherited variants, which is permanent and unchanging, while social determinants of health (SDOH) capture the contemporary influence of the environment on individuals and communities [22]. SDOH can be expected to change, sometimes rapidly, with substantial implications for the planning and conduct of interventions, community engagement, and the provision of suitable infrastructures and support [21]. Moreover, specific genome-variant information permits individuals to observe online, in real-time, the provisional assessment of genomic risk and to engage with the underlying basis of that assessment [20]. Such

interactions testify to the potential value and range of approaches that integrated multi-data sets can offer to a community facing the pervasive and complex challenge of disease [26].

Future Research Directions

A notable gap in genomics research concerns the inclusion of diverse populations underrepresented in the scientific literature [25-28]. The absence of whole-genome sequencing variation and polygenic risk score data in Māori, Pacific, Indian, Asian, and Middle Eastern communities creates an opportunity to implement the established modelling approaches for factoring social and ancestral types into genome-wide association studies. The feasibility of conducting research in non-European groups is enhanced by the increased connectivity (phone, internet) of these populations [24]. The recent inception of comparable genomic, lifestyle, and social determinant datasets in New Zealand offers a unique advantage for ensuring that research actively encompasses underrepresented groups [29-31].

CONCLUSION

Integrating whole-genome sequencing with social determinants of health represents a promising approach to improve coronary artery disease risk prediction and enhance precision medicine. Such integration provides richer, multidimensional insights into individual and population-level risk, improves predictive performance, and enables more targeted prevention strategies. However, disparities in predictive accuracy across diverse populations underscore the need for equitable model design, careful evaluation of bias, and inclusion of underrepresented groups in research. Real-world implementation requires robust infrastructure, governance, ethical oversight, and active clinician and patient engagement to ensure transparency, accessibility, and trust. Future research should focus on expanding genomic and SDOH datasets to diverse global populations, refining bias mitigation strategies, and translating predictive models into actionable clinical and public health interventions. Collectively, these efforts can advance equitable, data-informed strategies for CAD prevention and management.

REFERENCES

1. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, Lai FY, Kaptoge S, Brozynska M, Wang T, Ye S. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. Journal of the American College of Cardiology. 2018 Oct 16;72(16):1883-93.
2. Riveros-Mckay F, Weale ME, Moore R, Selzam S, Krapohl E, Sivley RM, Tarran WA, Sørensen P, Lachapelle AS, Griffiths JA, Saffari A. Integrated polygenic tool substantially enhances coronary artery disease prediction. Circulation: Genomic and Precision Medicine. 2021 Apr;14(2):e003304.
3. Jostins L, Levine AP, Barrett JC. Using genetic prediction from known complex disease loci to guide the design of next-generation sequencing experiments. PloS one. 2013 Oct 18;8(10):e76328.
4. Ugwu CN, Ugwu OP, Alum EU, Eze VH, Basajja M, Ugwu JN, Ogenyi FC, Ejemot-Nwadiaro RI, Okon MB, Egba SI, Uti DE. Sustainable development goals (SDGs) and resilient healthcare systems: Addressing medicine and public health challenges in conflict zones. Medicine. 2025 Feb 14;104(7):e41535.
5. Howell CR, Zhang L, Yi N, Mehta T, Garvey WT, Cherrington AL. Race versus social determinants of health in COVID-19 hospitalization prediction. American Journal of Preventive Medicine. 2022 Jul 1;63(1):S103-8.
6. Ugwu OP, Alum EU, Ugwu JN, Eze VH, Ugwu CN, Ogenyi FC, Okon MB. Harnessing technology for infectious disease response in conflict zones: Challenges, innovations, and policy implications. Medicine. 2024 Jul 12;103(28):e38834.
7. Li S, Cai T, Duan R. Targeting underrepresented populations in precision medicine: A federated transfer learning approach. The annals of applied statistics. 2023 Oct 30;17(4):2970.
8. Goldstein BA, Knowles JW, Salfati E, Ioannidis JP, Assimes TL. Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example. Frontiers in genetics. 2014 Aug 1;5:254.
9. Manemann SM, St Sauver JL, Liu H, Larson NB, Moon S, Takahashi PY, Olson JE, Rocca WA, Miller VM, Therneau TM, Ngufor CG. Longitudinal cohorts for harnessing the electronic health record for disease prediction in a US population. BMJ Open. 2021 Jun 1;11(6):e044353.
10. Ongesa TN, Ugwu OP, Ugwu CN, Alum EU, Eze VH, Basajja M, Ugwu JN, Ogenyi FC, Okon MB, Ejemot-Nwadiaro RI. Optimizing emergency response systems in urban health crises: A project management approach to public health preparedness and response. Medicine. 2025 Jan 17;104(3):e41279.
11. Agrawal S, Klarqvist MD, Emdin C, Patel AP, Paranjpe MD, Ellinor PT, Philippakis A, Ng K, Batra P, Khera AV. Selection of 51 predictors from 13,782 candidate multimodal features using machine learning improves coronary artery disease prediction. Patterns. 2021 Dec 10;2(12).
12. Mehandziska S, Stajkovska A, Stavrevska M, Jakovleva K, Janevska M, Rosalia R, Kungulovski I, Mitrev Z, Kungulovski G. Workflow for the implementation of precision genomics in healthcare. Frontiers in Genetics. 2020 Jun 30;11:619.

13. Ramos EM, Din-Lovinescu C, Berg JS, Brooks LD, Duncanson A, Dunn M, Good P, Hubbard TJ, Jarvik GP, O'Donnell C, Sherry ST. Characterizing genetic variants for clinical action. InAmerican Journal of Medical Genetics Part C: Seminars in Medical Genetics 2014 Mar (Vol. 166, No. 1, pp. 93-104).
14. Paul-Chima UO, Ugwu CN, Alum EU. Integrated approaches in nutraceutical delivery systems: optimizing ADME dynamics for enhanced therapeutic potency and clinical impact. RPS Pharmacy and Pharmacology Reports. 2024 Oct;3(4):rqae024.
15. Amendola LM, Coffey AJ, Lowry J, Avecilla J, Malhotra A, Chawla A, Thacker S, Taylor JP, Rajkumar R, Brown CM, Golden-Grant K. Development of a comprehensive cardiovascular disease genetic risk assessment test. Genetics in Medicine Open. 2025 Dec 5:103482.
16. Berg K, Doktorchik C, Quan H, Saini V. Automating data collection methods in electronic health record systems: a Social Determinant of Health (SDOH) viewpoint. Health Systems. 2023 Oct 2;12(4):4Castela Forte J, Folkertsma P, Gannamani R, Kumaraswamy S, Mount S, de Koning TJ, van Dam S, Wolffenbuttel BH. Development and validation of decision rules models to stratify coronary artery disease, diabetes, and hypertension risk in preventive care: cohort study of returning UK Biobank participants. Journal of Personalized Medicine. 2021 Dec 7;11(12):1322.72-80.
17. McClellan KA, Avard D, Simard J, Knoppers BM. Personalized medicine and access to health care: potential for inequitable access?. European Journal of Human Genetics. 2013 Feb;21(2):143-7.
18. Ugwu OP, Ogenyi FC, Ugwu CN, Ugwu MN. Gut microbiota-derived metabolites as early biomarkers for childhood obesity: A policy commentary from urban African populations. Obesity Medicine. 2025 Sep 1;57:100641.
19. Blizinsky KD, Bonham VL. Leveraging the learning health care model to improve equity in the age of genomic medicine. Learning health systems. 2018 Jan;2(1):e10046.
20. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics. 2018 Sep;50(9):1219-24.
21. Menestrel TL, Craig E, Tibshirani R, Hastie T, Rivas M. Using Pre-training and Interaction Modeling for ancestry-specific disease prediction in UK Biobank. arXiv preprint arXiv:2404.17626. 2024 Apr 26.
22. Rivas Velarde MC, Tsantoulis P, Burton-Jeangros C, Aceti M, Chappuis P, Hurst-Majno S. Citizens' views on sharing their health data: the role of competence, reliability and pursuing the common good. BMC Medical Ethics. 2021 May 18;22(1):62.
23. Minari J, Brothers KB, Morrison M. Tensions in ethics and policy created by National Precision Medicine Programs. Human genomics. 2018 Apr 17;12(1):22.
24. Sweet K, Gordon ES, Sturm AC, Schmidlen TJ, Manickam K, Toland AE, Keller MA, Stack CB, García-España JF, Bellafante M, Tayal N. Design and implementation of a randomized controlled trial of genomic counseling for patients with chronic disease. Journal of personalized medicine. 2014 Jan 8;4(1):1-9.
25. Ugwu OP, Ogenyi FC, Ugwu CN, Basajja M, Okon MB. Mitochondrial stress bridge: Could muscle-derived extracellular vesicles be the missing link between sarcopenia, insulin resistance, and chemotherapy-induced cardiotoxicity?. Biomedicine & Pharmacotherapy. 2025 Dec 1;193:118814.
26. Smith JL, Tcheandjieu C, Dikilitas O, Iyer K, Miyazawa K, Hilliard A, Lynch J, Rotter JI, Chen YD, Sheu WH, Chang KM. Multi-ancestry polygenic risk score for coronary heart disease based on an ancestrally diverse genome-wide association study and population-specific optimization. Circulation: Genomic and Precision Medicine. 2024 Jun;17(3):e004272.
27. Moore CB, Dolan DD, Yarmolinsky R, Cho MK, Soo-Jin-Lee S. The ELSI Virtual Forum, 30 Years of the Genome: Integrating and Applying ELSI Research. Journal of Law, Medicine & Ethics. 2023 Sep;51(3):661-71.
28. Huang QQ, Sallah N, Dunca D, Trivedi B, Hunt KA, Hodgson S, Lambert SA, Arciero E, Wright J, Griffiths C, Trembath RC. Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals. Nature communications. 2022 Aug 9;13(1):4664.
29. Wastvedt S, Snoke J, Agniel D, Lai J, Elliott MN, Martino SC. De-biasing the bias: methods for improving disparity assessments with noisy group measurements. Biometrics. 2024 Dec;80(4):ujae155.
30. Chappell E, Arbour L, Laksman Z. The inclusion of underrepresented populations in cardiovascular genetics and epidemiology. Journal of Cardiovascular Development and Disease. 2024 Feb 5;11(2):56.
31. Dogan MV, Beach SR, Simons RL, Lendasse A, Penaluna B, Philibert RA. Blood-based biomarkers for predicting the risk for five-year incident coronary heart disease in the Framingham Heart Study via machine learning. Genes. 2018 Dec 18;9(12):641.

CITE AS: Namirimu Sandrah (2026). Integrating Whole-Genome Sequencing With Social Determinants Data for Coronary Artery Disease Risk Prediction: Interpretability, Bias, and Real-World Performance, Implementation, and Equity Considerations. RESEARCH INVENTION JOURNAL OF SCIENTIFIC AND EXPERIMENTAL SCIENCES 6(1):86-96. https://doi.org/10.59298/RIJSES/2026/618696