IAA Journal of Biological Sciences 14(1):87-97, 2026 www.iaajournals.org

©IAAJOURNALS ISSN:2636-7254

https://doi.org/10.59298/IAAJB/2026/1418797 IAAJB1411

Integrating Long-Read Sequencing with Imaging Biomarkers for Stroke Risk Prediction: Interpretability, Bias, and Real-World Performance with Implementation and Equity Considerations

Serunjogi Ruth

Department of Clinical Pharmacy Kampala International University Uganda
Email: ruth.serunjogi@studwc.kiu.ac.ug

ABSTRACT

Ischemic stroke remains a leading cause of mortality and long-term disability worldwide, with risk prediction constrained by incomplete characterization of biological heterogeneity and limited integration of high-dimensional data. Advances in long-read genomic sequencing and neuroimaging now enable comprehensive profiling of structural variants, haplotypes, epigenetic signals, and imaging biomarkers directly linked to cerebrovascular pathology. This paper examines the integration of long-read sequencing data with imaging biomarkers for stroke risk prediction, with particular emphasis on interpretability, bias, real-world performance, and equity-oriented implementation. We review methodological foundations for multimodal data fusion, feature extraction from long-read sequencing, and quantitative imaging biomarker analysis, highlighting the complementary biological insights each modality provides. We further analyze modeling strategies that balance predictive performance with clinical interpretability, including fairness-aware machine-learning approaches to address demographic and technical biases. Real-world validation challenges, external generalizability, calibration, and clinical utility are critically assessed, alongside governance, privacy, and regulatory considerations. By situating genomic imaging integration within healthcare system constraints and equity frameworks, this work outlines a translational pathway for responsible deployment. We conclude that integrative genomic imaging models hold substantial promise for improving stroke risk stratification, provided that interpretability, bias mitigation, and equitable access are embedded throughout model development and implementation.

Keywords: Ischemic stroke, Long-read sequencing, Imaging biomarkers, Multimodal risk prediction, Health equity and fairness

INTRODUCTION

Ischemic stroke is a leading cause of morbidity and mortality worldwide, and one of the major underlying causes is atherothrombotic disease on the background of Cardiometabolic conditions [2]. Effective prevention strategies, including the use of antihypertensive, atherosclerosis-lowering, or antidiabetic medications, offer the opportunity to avert the course of disease progression, leading to cardiac ischemia, renal failure, and eventually aortic dissection or ischemic stroke [9]. For stroke prevention to be effective, a refined prediction of the 5-year vascular risk focusing on the transitional state from cardiometabolic to known clinical endpoints is, however, critical [1]. The translational pipeline for Genomic longitudinal low-pass and long-read sequencing of the human genome provides the opportunity to tap the potential beyond currently available services, as the median turnaround time is 24 hours for prognostic imaging biomarkers, clinical biochemical tests, and transcriptome services[2]. In parallel, acquiring the imaging databank, extracting clinically necessary prognostic brain MRI-VBM directly correlating to

cerebrovascular risk and outcomes, and directly predicting the 5-year independent prognosis further substantiates the feasibility of risk model updating [1].

Background and Rationale

Genomic sequencing devices employing long-read strategies (≥10 kb) enhance the discovery of clinically relevant structural variants (SVs)[1]. These include off-targets in the CFTR or SMN1 genes, Mendelian mutations causative for familial stroke such as COL4A1, and haplotypes covering variant-rich loci, e.g., APOE. Such information, currently unacquirable via short reads, applies to stroke prediction, thereby motivating the integration of long-read data with imaging markers [2]. Moreover, the 3D genomics of cell types implicated in cardiovascular pathology, such as endothelial ancestors, pericytes, and microglia, remains to be deciphered. Long-read sequencing can elucidate the regulatory networks governing their differentiation from pluripotent cells, thereby refining the description of stroke-relevant 3D organization [2]. Imaging biomarkers improve risk stratification for future cerebrovascular events and are common in machine-learning studies. A deep-learning model jointly considering radiological and clinical variables surpassed clinical-parameter-only approaches in predicting future vascular events after transient Ischaemic Attacks [1]. Models integrating genome-wide genotype arrays and transcriptomic profiles containing cerebrovascular loci also achieve competitive performance [2]. Risk-prediction tasks generally combine heterogeneous data using either multi-input architectures or separate unconnected models. Both strategies offer advantages: multi-modal integration can enhance representation through richer joint feature extraction, while separate systems afford interpretability through modality-specific models [7]. Long-read genomic information, typically donut-shaped statistics capturing sequencing coverage from different genomic segments, lends itself well to separate genomic pipelines [18]. Such protocols are thus favoured to describe the demographic characteristics and loss-of-function spectrum of long-read cohorts within the alimentary stomodaeum of organoid-derived gastric tissue [19].

Long-Read Sequencing in Risk Prediction

Long-read sequencing has emerged as a promising tool in stroke-risk prediction due to its unique ability to capture large structural variants linked to cerebrovascular disease and its relevance to complex regions harbouring transposable elements and methylation patterns associated with stroke [14]. Genomic data can be extracted from two large long-read cohorts containing more than 1,000 individuals across diverse populations, with extensive imaging measurements, accessible to qualified investigators and managed by a long-read data commons [13]. The availability of validated imaging protocols and surrogates, such as standardised MRI morphometry, further facilitates the estimation of multiple imaging biomarkers linked to cerebrovascular risk and outcomes, as well as their combined longitudinal dynamics from annual follow-up datasets across a 10-year period [15]. Thus, using advanced fusion modelling enables the combination of long-read and imaging data to predict multifaceted stroke risks while alleviating the challenges of feature extraction [16].

Imaging Biomarkers for Stroke

The presence of even a single imaging biomarker before an incident cardiovascular event, such as a heart attack or stroke, significantly increases the likelihood of such an event occurring in the subsequent one to three years [4]. Reliable detection of these biomarkers could enable prioritization of individuals for further investigation. Carotid plaques that are heterogeneous and/or have a lipid core, or microbleeds (the only intracerebral biomarker directly correlated with stroke risk and a determinant of eligibility for thrombolytic therapy) constitute unequivocal imaging biomarkers [6]. In patients with atherosclerotic carotid artery disease, whole-brain CT and MRI, advanced machine-learning, and artificial-intelligence approaches can extract imaging biomarkers that non-invasively help to stratify the risk of future stroke and non-stroke vascular events [3]. Deep-learning architectures enable the extraction of CT-based imaging biomarkers indicating expected recovery post-ischemic stroke in a clinical setting [9]. The lack of prior knowledge on the categorical, pixel-based output labeling of unconstrained volumetric data in radiology poses a significant challenge, compounded by competing independent variables that determine long-term functional outcome. Assessment of learned representations reveals links to well-established biomarkers [4].

Integrative Frameworks for Genomic and Imaging Data

Integrating genomic long-read sequencing data with imaging biomarkers has the potential to provide unique insights into complex diseases at an unprecedented scale and precision [17]. Multi-site, multi-modal studies such as the Alzheimer's Disease Neuroimaging Initiative database have demonstrated advantages of jointly modeling imaging and genetic data for neurodegenerative disorders [5]. Formulating clinical challenges as risk-prediction problems using integrated genomic and imaging data is attractive and widely adopted. For large-scale, complex traits such as stroke, however, existing precision medicine approaches struggle: the majority of studies consider either genetic or imaging data as inputs, and seldom jointly with multi-modal data directly [15]. Any integrative approach must accommodate diverse data types: longitudinal variants or haplotypes extracted directly from long-read sequence data, and quantifiable imaging information such as volumes, shapes, or other characteristics from serial imaging biomarkers on clinical MRI data [16]. Integrative frameworks can allow the fusion of different

types of data from multiple sources to enhance risk-prediction models [6]. More precisely, complementary information derived from diverse genomic sources can contribute additional a priori evidence to stroke models; nonetheless, based only on quantifiable imaging characteristics [7]. When considering data generated by long-read sequencing, integrating rather than separately analyzing genetic and imaging data offers the opportunity to leverage structural variants or haplotypes distinct attributes that remain poorly characterized in stroke genomics and are likely to yield larger gains in predictive performance [2]. Such long-read analyses are essential given that stroke-associated variants are distributed evenly across the genome: 59% outside regulatory regions, 40% intergenic, 44% within genes, and 70% outside common gene-promoter–associated haplotype blocks [1].

Methodological Foundations

Stroke ranks among the most pressing health challenges and the second leading cause of death worldwide 1. Globally, stroke accounts for about 12% of deaths and approximately 9% of disability-adjusted life years [13]. Effective prevention relies on assessing an individual's stroke risk [18]. Accurate predictions enable healthcare professionals to prioritize preventive measures, optimize patient management, and enhance clinical decision-making. As age, hypertension, diabetes, and the presence of carotid stenosis are some clinical factors associated with cerebrovascular disease risk, creating predictive models that include those variables may facilitate early interventions to prevent acute ischemic stroke [7]. The predictive modelling community has recognized the necessity of integrating clinical–radiological features in prediction models. Such integration processes to obtain accurate predictions are not trivial, however, and the quantitative integration of multimodal data addresses the vital need for that process [4]. Genomic data from long-read sequencing has emerged as a valuable source of information to inform stroke-risk stratification, particularly concerning patients with atrial fibrillation (AF) and those outside standard targets. Long-read WGS enables the comprehensive characterization of genomic variation over complex regions, facilitates the analysis of large structural variants, and enhances haplotype-resolved genomes [9]. In tandem with long-read genomic data, imaging biomarkers derived from standard brain magnetic resonance imaging (MRI) sequences have been associated with the risk and prognosis of ischemic stroke. Integrative modelling that leverages both long-read genomic and imaging data presents an opportunity to bolster stroke-risk prediction [18]. Enabling the simultaneous modelling of diverse input features as either fused or separate data allows the joint use of complete complementary genomic and imaging information [20].

Data Acquisition and Cohort Design

Stroke is a time-sensitive, multifactorial disease requiring rapid diagnosis and treatment, yet most patients arrive at hospitals through emergency services, which cannot acquire patient histories early [11]. Clinicians, therefore, remain unaware of individual patients' risks for stroke during the critical minutes after symptom onset [15]. Especially in suspicion of stroke, patients who receive medications to mitigate damage from thrombosis, hemorrhage, or other lesions incur significant medical costs [16]. Informatics broadly considers two approaches to enable analysis of otherwise unavailable information in a time-sensitive manner: the information-theoretic approach, which focuses on modeling various types of data, and the causal approach, which emphasizes the understanding of data distribution not captured in the data-generating process [13]. Each approach carries its advantages, but consideration of the nature of the available data guides the decision-making [12]. Stroke risk constitutes an auxiliary clinical problem, which demands joint consideration of various data with incomplete supervision because such risk appears at the onset of the disease, when prior events remain unavailable. The framing leads to the possibility of exploring specific spelling and implementation strategies for the auxiliary problem [8].

Feature Extraction from Long-Read Sequencing

Extracting informative genomic, epigenomic, and transcriptomic features from long-read sequencing data enhances understanding of stroke pathobiology and supports risk prediction. Structural variants, phase information, or haplotypes are essential for interpreting the effects of genetic variants on transcriptome and phenotype [3]. Targeted methylation detection in stroke-relevant regions, alongside full-length isoforms, elucidates controls of exon skipping, alternative splicing, and transcriptional bursts [2]. Combined with clinical or multi-omic data, these long-read features contribute to integrative risk models capturing stochasticity and non-linear processes with potentially higher predictive performance [5]. Complete and diverse methylation signals, transcriptomic profiles, and image features from longitudinal or multi-modal studies further augment risk predictions. Intact methylation patterns, transcript compositions, and isoform expression influence long-term outcomes post-stroke, shaping recovery trajectories [7]. Time-series models predict sub-acute and chronic outcomes, facilitating timely intervention planning. The absence of clear predictive solutions in the literature motivates the exploration of these genotypic and phenotypic markers across a range of temporal profiles [4].

Imaging Biomarker Quantification

Longitudinal multimodal data can capture different disease stages, aid trajectory modeling, and enhance treatment outcome predictions with low signal loss [19]. Irrespective of multi-omics or image types, initial stage determination remains essential for data collection design and temporality decisions, such as synchronizing with

the first infusion in therapy response studies. Extensive datasets are being proliferated across various biobanks to exploit these synergies [18]. To predict therapy response duration in HI, considered singular due to a clear initial timeline, 233,000 longitudinal samples are integrated, unifying TCR, transcript, and Epigenome datasets across millions of cells and biospecimens [13]. As a trade-off between granularity and volume, the largest single-cell TCR from the Human Cell Atlas comprises >124 M cells yet only 55 k TCRs [4]. Quantifying imaging biomarkers from non-contrast CT can provide interpretable insights into stroke risk and outcomes by identifying structural characteristics such as infarct volume and lesion distribution [9]. Biomarker selection targeted a set of standard size and shape parameters, enabling comparison across datasets and applications to screening uncertainties [6].

Following initial quality assurance, the extraction procedure comprises three steps. First, ASPECTS is computed using a contour-based method on the supplied mask [8]. The mass at non-zero locations is accumulated, including T2, T2_fl, DWI, and DWI_fl, allowing, for example, the evaluation of recanalization gained with respect to the mask at R. A wide parameter search indicated a ten-fold upper limit is generally effective to minimize sensitivity to acquisition conditions [3]. These pre-defined metrics have been associated with structural recovery and prognosis in literature; therefore, they can serve as relevant indicators of how structural changes relate to clinical outcomes [10].

Modeling Strategies and Interpretability

In predictive modeling, interpretability remains a fundamental concern [9]. Models with complex joint genomic-imaging inputs should produce explanations compatible with domain expertise. For model-agnostic explanation, feature importance metrics or SHAP-like explanations elucidate how changes in each feature shift the predicted risk [8]. When the exposure or baseline risk for a designated population group varies widely across demographic attributes, such as age or sex [7]. Fairness-Aware Machine Learning (FAM-L) implementation aids in selecting training inputs and shaping explanations [11].

Bias, Fairness, and Equity Considerations

Estimating stroke risk via medical images and long-read genomic sequencing has the potential to fill an urgent clinical need for preventive care [10]. Imaging biomarkers derived from computerized tomography (CT) and magnetic resonance imaging (MRI) scans are linked to cerebrovascular risk, yet few systems incorporate such features into machine-learning models, and none combine them with long genomic reads [13]. Long-read sequencing resolves structural variants and haplotypes that influence stroke biology but is rarely incorporated into neural datasets [18]. Integrating data from these modalities may enhance risk prediction by leveraging genomic and imaging information in tandem [16].

Population Representativeness

Stroke is the second leading cause of death worldwide, causing 1 in 11 deaths globally (World Health Organization, 2022) [11]. Identifying individuals at risk of future stroke is an important public health priority to enable patients to adopt preventive measures early (e.g., treat hypertension, abstain from smoking, adopt healthier diets) and/or to receive timely diagnostic imaging and treatment following warning signs [12]. Stroke represents severe interruptions in blood supply to part of the brain that can lead to permanent disability or death, resulting from either blood vessel rupture (haemorrhagic) or vessel blockage (ischaemic) [13]. Though computation and the digital technologies required to enable its widespread adoption have substantially evolved over the past two decades, none of the initial efforts to introduce thousands of biological and imaging biomarker datasets have yet progressed to clinical practice [14]. Integrating whole-genome long-read sequencing with imaging biomarkers of cerebrovascular risk and ischaemic stroke on the same platform has the potential to substantially improve both longitudinal characterisation and lifetime prediction of future vascular burden [15].

Technical and Measurement Biases

Bias can adversely affect the clinical use of risk stratification equations. Such equations are often derived from observational studies in which risk predictors are subject to technical and measurement bias [4]. Preprocessing and feature extraction pipelines can introduce biases that are unrepresentative of the general population [3]. Patient heterogeneity coupled with variable compliance with imaging protocols, leads to inconsistencies in risk predictor measurement [2]. These issues are likely to be exacerbated in existing clinical implementations, where the equation may already comprise a specific subset of predictors [13]. Long-read genomic sequencing techniques can suffer from several technical and measurement biases regardless of the platform. At the technical level, sequencing data may not cover the full complement of biomarkers, while measurement noise can vary across individuals and influence model calibration [5]. At the measurement level, the genomic content of arteries may differ from that of other tissues or fluid types and remain unaccounted for. Such biases can trigger clinical inequities in the risk assessment of stroke or other diseases with high societal costs, including cardiovascular disease and certain types of cancer [14]. The capacity to identify these and other potential biases depends on understanding the relevant technical and procedural requirements for both genomic and imaging data. Emerging

interpretability strategies compatible with high-dimensional inputs can measure biases and disentangle their influences on final predictions [15].

Societal and Ethical Implications

Large-scale sequencing technologies, machine learning, and neural networks fuel unprecedented progress in genomics and biomedical research [1]. A combination of genomics long-read data and digital imaging has opened new perspectives in stroke preventive medicine [8]. Multi-modality input data promises higher outcome prediction accuracy for risk stratification and better preventive treatment [7]. Recent research uncovered that stroke can occur without prior transient symptoms. Large-scale biomarker data still provides a substantial reduction in false positive rates. Gradually increasing data allows a reduction in the individualized experimental threshold for liabilities. Low-cost imaging on already-installed scanners can be leveraged for both convenience and cost-effectiveness [11]. Long-read assembly suggests that genetic predispositions can be decoupled from lifestyle limitations. Phasing and structural variations for variant candidates in a vast amount of pelosomedicis record strengthening causal inference for long-read assembly. Default genomic samples already collected serve as a solid groundwork [12]. A detailed explanation of how these biomarkers inform end-users remains imperative for evoking trust among specialists and transference from the regulator [13].

Real-World Performance and Validation

Machine learning (ML) and deep learning models have been extensively applied to construct predictive models of ischemic stroke due to their ability to unmask complex relationships in large datasets. However, ML models are rarely incorporated in real-world clinical stroke workflows [7]. One solution to overcome the challenges faced by ML models is transfer learning, where one model learns from another pre-trained model to address similarities in data probability distributions [6]. Transfer learning techniques like domain adaptation and domain generalization allow the source domain to differ from the target domain in the feature space. Transfer learning has been used with imaging datasets, but it remains unexplored for structured tabular datasets characterized by partial feature overlap. Stroke prediction on tabular datasets remains dominated by conventional methods or interpretable ML using lightweight, readily available models [8]. A benchmarking study evaluating modern ML techniques 16 found that XGBoost outperformed deep learning models in terms of consistency, precision, and robustness on multiple datasets with structured data [9]. Concerning open tools for medical imaging, FastAi achieves high generalization with Low-Dose-Kidney computed tomography data acquired under systematic noise. However, no open-source tools are available for tabular datasets [10].

External Validation in Diverse Settings

Turning risk factors into risk prediction is critical for improving disease prevention. For stroke, the implementation of such models is hindered by the limited availability, poor synchrony, and low predictive performance of existing models [7]. Integrating genomic and medical imaging data into a single prediction enables more accurate stroke risk estimation than is possible with either data type alone [16]. Yet, such a combination poses technical, ethical, and social challenges that need to be addressed before implementation. Extensive performance evaluation against established standards will afford insights into the strengths and weaknesses of the integrated genomic-imaging approach relative to the use of either data type alone [5]. External validation in diverse clinical settings and populations will provide evidence of its generalizability [4]. Additional measures of performance calibration, discrimination, clinical utility via decision-curve analysis, and consideration of resource use and embedding into clinical workflows will quantify its readiness for prospective evaluation. Methods of analysis will account for the special conditions and impact of the population selected; any measure that strictly indicates calibration, discrimination, or clinical utility will be retained, regardless of its wider applicability [17].

Calibration and Clinical Utility

Forecasting the future state of stroke prevention, risk stratification, and management inspires health equity considerations in biomedical research using genomic and imaging data [3]. These data types provide stroke-relevant brain cell signatures informing pathology, potential therapeutic targets, and connections to equity, especially long-read sequencing targeting structural variation [5]. Growing imaging biomarkers also assess stroke severity, risk, and prognosis through automatic quantification. Integration of such diverse data offers a potential solution with expected gain in predictive [6]. Longitudinal data from existing cohorts and prospective cohorts with long-read sequencing (Pacific Biosciences or Oxford Nanopore Technologies) and MRI standards of imaging acquisition for informatics are being used to test [4]. Long-read sequencing generation with focused-stroke data [6]. The unique potential of long-read sequencing to reveal stroke-relevant structural variation targeting diverse genomic regions relevant to stroke (genes, enhancers), combined with relevant MRI-based and further PET-based imaging (and their potential combined temporal) modeling in diverse health-system-representative settings and populations, merits consideration [18].

Implementation Challenges in Healthcare Systems

Adoption of new prediction models in healthcare systems poses significant hurdles [16]. Integration of long-read sequencing with imaging biomarkers for stroke-risk estimation will necessitate consideration of practical challenges known to hinder the uptake of innovative methods [19]. Systemic barriers include effective workflow incorporation and user interest issues that could obstruct valuable approaches, even when interpretability, bias, and external-validation criteria have been addressed [14].

Implementation and Equity Considerations

Risk prediction tools have the potential to improve stroke prevention by identifying high-risk individuals who may benefit from preventive interventions [4]. The long-term objective is to develop and deploy stroke risk prediction models based on genomic long-read sequencing and imaging biomarkers through a collaborative effort between engineering and biomedical research teams [5]. The research focuses on bridging the algorithmic and clinical gaps by addressing three key questions: Is it possible to combine genomic long-read data with brain imaging features in a unified stroke risk prediction framework? If so, can such integrative models outperform their genomic-only counterparts? How can model predictions achieve a level of interpretability that allows for cross-domain knowledge transfer and incorporation of medical insight and expertise? The proposed evaluation framework includes complementary checks for model interpretability and data fairness, with interpretability metrics suggested by the IEEE Standards Association and multiple fairness benchmarks introduced in prior work specifically defined for genomic and imaging data [7]. The study involves an analysis of the scientific literature and operational procedures of publicly funded, peer-reviewed human-genome-sequencing projects to construct a transparent translational pathway from research to implementation that meets the needs of public health and government stakeholders [9].

Resource Allocation and Access

Despite having potentially devastating consequences, stroke prediction remains a non-trivial problem that only a handful of studies have investigated within the large and heterogeneous stroke domain [5]. When integrating information across multiple modalities, scarce access to multimodal datasets further complicates the design of integrative prediction models [4]. Being able to use one modality as a reference, usually the one with the broadest applicability, would therefore considerably increase the feasibility of the combination. When such reference modality possesses additional predictive power regarding the outcome of interest, the complementarity principle indicates that a predictive model combining both modalities should be more informative than either of the two subsets alone [9]. This leads to the biologically-grounded hypothesis that genomic and imaging biomarkers contain complementary information for the prediction of cerebrovascular risk and outcomes after Stroke, thus enabling models trained under the two-modality combination to outperform those trained solely on either genomic or imaging data [7]. Acquisition protocols for long-read genomic datasets tailored to cerebrovascular risk have been established; image-based stroke-specific biomarkers objectively characterizing the disease state have been identified; multiple publicly available datasets are accessible for training and validation; and several academic, industrial, and governmental end-users have expressed demand for such models with stated intentions for pilot studies, early adoptions, and customisation attempts[11]. The methods taken to answer these scientific questions thus present a compelling, relevant, and impactful integration of long-read genomic data onto an imaging stroke risk model that broadens the factors used for the prognosis from one modality to two without sacrificing generalizability, along a translational pathway with precedents of large-scale interest allowing rapid deployment of short-read genomic data using existent administrative data to tackle other common yet still complex genotypes[12]. Stroke is one of the leading diseases affecting public health, including morbidity and mortality, along with a significant burden on health-care systems in many countries. Genomic and imaging modalities, tied to physiological factors influencing the disease severity, have also garnered much attention as the basis for predictive models, and prediction tasks are even more challenging given the complexity of both modalities and the relatively restricted number of data collected from the community [7]. The area of study thereby aims to investigate whether and how the addition of imaging datasets from the modeled population enhances predictive accuracy and robustness of genomic models, towards addressing the aforementioned gap within the community [6].

Infrastructure and Workforce

Integrating large-scale imaging and genomic datasets provides an opportunity to tackle pressing translational neuroscience questions [20]. Multi-site studies have emerged in recent years to take advantage of the extensive collection of neuroimaging data and associated metadata made available via open-access resources. Such initiatives improve the reproducibility of studies and collaborative works by offering a reliable and safe environment for sharing and reusing datasets [12]. For example, the development of dedicated stroke imagery banks has led to an ever-growing collection of stroke-related images under well-defined conditions [11]. Expanded access to neuro-imaging datasets contributes to enhanced basic/clinical research in neurosciences and, therefore, offers a better understanding of the serious impact of brain disorders in today's world [10]. Furthermore, these collaborative

research resources turn out to be essential in academic research networks where malignant brain-related disorders such as ischemic stroke and brain tumours are prevalent [9].

Governance, Privacy, and Informed Consent

Integration of long-read sequencing data with imaging biomarkers for stroke risk prediction raises governance, privacy, and informed consent issues requiring careful attention [3]. Research proposals should provide specific details on the steps undertaken to protect data subject rights while remaining consistent with the overarching aims of the project [21]. Research on stroke risk prediction using long-read genomic data and imaging markers may involve communication of sensitive health information about an individual's stroke risk, including whether risk is low or high[22]. When addressing the motivation for stroke prevention through biomedical research and innovation, it is vital to implement sufficient safeguards to protect personal data rights and dignity [13]. Research on the genetic and genomic underpinnings of diseases and conditions of public health concern is intrinsically valuable, but it is fundamental to govern such research on human beings, their bodies, and their biological samples responsibly. Public health risks and burdens imposed by preventive measures should not be amplified by exposure to research [22].

Interpretability and Clinical Decision-Making

Clinician decision-making hinges on interpretability, necessitating a deep understanding of the relationship between predictors and predicted outcome [2]. Well-established biomarkers facilitate risk communication, guiding treatment discussions and enabling patient education. Risk-management options at the individual level hinge on knowledge of modifiers influencing future risk [3]. The prediction of substantial vascular-event risk is divorced from clinical action if the corresponding combined risk is absent. Stakeholder requirements are met by targeting the clinical question of future vascular-event risk, amenable to the provably explainable class of influence functions [23]. Modeling clarifies the pathway to risk and enables group-by-group stratification of expected permutations.

Risk melting, founded on counterfactual mechanics for precise human risk evaluation, enables risk communication and shared decision-making [22]. Health is phrased probabilistically to disclose the genuine quality of routes to maximal future health, forming consensus among diverse stakeholder perspectives [13]. The integration of genomic and imaging materials mandates multifactorial behavior and coupling, with the approach chosen reflecting the minimum necessary rank-complexity augmentation to accommodate the additional material. The intersection of the two modalities improves fidelity, and freely accessible, openly documented computational implementations invite shared scrutiny [7].

Explainable Models for Genomic-Imaging Integration

Stroke prevention strategies depend on the identification of high-risk individuals who would benefit from medical therapy [14]. Despite numerous risk factors and an array of clinical tools for predicting individuals' chances of stroke, the observed stroke rate remains low, raising questions regarding the effectiveness of these prediction algorithms at the single-subject level [13]. High-dimensional genomic data may provide complementary, mechanistic-level information that could boost prediction performance and assist in designing preventive strategies tailored to an individual's biology [12]. Long-read sequencing technology shows promise for uncovering structural variants (SVs), determining haplotypes, and analysing repeat regions that may act as stroke-relevant regulatory elements of large-effect genes, all of which are poorly captured by short-read sequences [10]. Imaging biomarkers, meanwhile, offer a phenotypic heritable signature tightly linked to underlying genomic variation. Integrating long-read genomic sequencing and imaging biomarkers thus represents a novel, potentially impactful approach to high-dimensional and multimodal risk prediction [24], Jiang et al., 2014.

Clinician and Patient Interfaces

Longitudinal biomedical data are useful for quantifying a wide range of health-related factors and for predicting clinical outcomes [25]. Using a combination of genomic, imaging, and electronic health record (EHR) data, observables such as social history or prior illnesses relevant to cerebrovascular health can be selected. Visualizations are important to present information in a clear and efficient manner, facilitating understanding and interpretation both for the health and medical area, and the general public [14]. Time-based information allows for the monitoring of the evolution of certain biomarkers with respect to time, which is particularly important for temporal-data-viewing software [13]. Moreover, it aids in selecting the cohort population to focus on for long periods of time without intervention for an accurate prediction of strokes [11]. The risk based on the predictive model is generated and presented in simple terms to the user at once, without the need to scroll through a lengthy report, which saves the user time. An information-retrieving process and a graphical software solution are proposed to bridge time-varying datasets and prediction status [25].

Risk Communication and Shared Decision-Making

Clinical risk assessments inform disease management strategies and engage patients in shared decision-making 26. When prognostic models are employed within screening or treatment contexts, individuals are ideally notified of their risk estimates and counselled regarding proactive measures or treatment options [15]. This discussion varies

in complexity based on the intended application [16]. An initial scenario involves models that leverage easily interpretable data, allowing for transparent communication of risk without hindering shared decision-making. A second scenario pertains to advanced models that necessitate complex data or integration of multiple modalities, complicating the communication process [13]. These challenges are particularly relevant when modelling long-read genomic sequencing alongside imaging biomarker data to project stroke risk. In these instances, sophisticated algorithms enable the extraction of substantial biological information without the demand for auxiliary multimodal data [13].

Policy and Regulation Implications

Implementing a stroke-risk-prediction approach integrating long-read sequencing data and imaging biomarkers raises several policy and regulatory considerations [18]. These include the establishment of standards for using clinically relevant integrated biomarkers, their validation and reporting, and guidelines for data stewardship. Consistent with the priority of the NIH's All of Us program, which endeavors to increase equity in biomedical research, special emphasis is placed on these aspects [17]. Standards for integrated-biomarker use are necessary to underpin clinical deployments. A growing set of foundational components, across different data modalities including data-acquisition, processing, and analytical methods are becoming established, many of which conform with the FIND framework for guiding responsible genomic data generation, analysis, and dissemination [18]. Nevertheless, only a handful of end-to-end systems are publicly available to facilitate entire biomedical machine-learning workflows, yet even these seldom provide comprehensive guidance on development and embedding. Correspondingly, complementary meta-level standards for integrated-biomarker reporting are also lacking [15].

Standards for Integrated Biomarker Use

Existing guidelines governing prediction model development, validation, and reporting lack recommendations for the integrated use of clinical, imaging, and genomic biomarkers, even as the number of studies incorporating combined biosensors and data types increases [22]. Models predicting functional outcome after acute ischemic stroke have been trained on clinical and imaging markers [7], while hazard ratios for paper-based questionnaires remain unchanged when incorporating imaging features, though biomarker integration remains opaque and groups may differ in observed risk [21]. The ongoing integration of many different data modalities combining clinical with imaging, imaging with genomic, and so on, still lacks shared standardization [20]. Further, healthcare regulations tend to prioritize data security and privacy in the use of artificial-intelligence (AI) technology while leaving unaddressed the equally vital matters of equity, fairness, and algorithmic bias that need further guidance [19].

Data Stewardship and Equity Requirements

Unregulated use of AI-based technology in public services raises significant concerns regarding equity and widespread discriminatory practices against underserved racial, ethnic, and socio-economic groups [27]. A typical artificial intelligence tool combines observations to produce an action via statistical inference representing a data-generating distribution. The use of AI exhibits widespread bias; unregulated usage amplifies pre-existing social inequities into the potential environment for decisions on health care to promote more equitable group opportunities [28]. Current regulations on artificial intelligence use in private organisations, especially in monitoring for safety and risks associated with decision (service) suggestions in artificial intelligence. Ethical mitigation of social-race-biased approaches occurring in algorithm decision-making has also been under discussion [24]. However, such mitigation cannot resolve equity concerns if wider sources of influence exist, such as differential access or stress, health choices before care, or continued influence of social position on health. A conceptual framework was developed based on ethical model properties to assess and measure quantitatively fairness and equity in algorithms quantitatively [23] quantitatively.

Future Directions and Research Priorities

Integrating long-read sequencing and imaging biomarkers for stroke risk prediction directly addresses a pressing clinical need [29-33]. Stroke remains a leading cause of death and disability worldwide, with major implications for individual and population health. The incidence of cerebrovascular events is rising in low- and middle-income countries due to an increase in risk factors [3]. Online risk scoring tools produce estimates clinically relevant for preventive interventions. Risk calculators can guide decisions on management in asymptomatic patients with carotid artery stenosis and inform targeted therapeutic interventions for secondary prevention [28]. The goal is to develop a model that combines genomic long-read sequencing data with imaging biomarkers of plaque characteristics and burden to predict the 5-year risk of incident symptomatic vascular disease [27]. The long-read strategy enables the characterization of structural variants and complex regions relevant to cerebrovascular disease, such as the 9p21 locus, that are inaccessible using short-read sequencing [26]. Non-invasive imaging of the carotid arteries facilitates risk stratification and guides clinical management for patients with asymptomatic stenosis through the assessment of plaque features, vessel wall volume, lipids, fibrous cap, and intraplaque hemorrhage that are difficult to quantify by other means. Integrating long-read genomic data and plaque imaging features is hypothesized to enhance stroke risk prediction, due to the distinct but complementary biological

insights obtained and multimodal data-driven synthesis of such information [26]. Models will be evaluated for interpretability, potential to inform clinical decision-making, and fairness across population subgroups defined by ethnicity, sex, geography, and socio-economic status [25].

CONCLUSION

The integration of long-read genomic sequencing with imaging biomarkers represents a transformative opportunity for advancing stroke risk prediction beyond the limitations of single-modality approaches. Long-read sequencing uniquely captures structural variants, haplotypes, and regulatory regions that are inaccessible to short-read technologies, while imaging biomarkers provide phenotypic signatures of cerebrovascular disease that are directly relevant to clinical decision-making. Together, these modalities offer complementary insights into stroke pathobiology and enable more precise, biologically grounded risk stratification. However, improved predictive performance alone is insufficient for clinical translation. Interpretability remains essential for clinician trust, regulatory acceptance, and meaningful patient engagement. Models integrating genomic and imaging data must provide transparent explanations that align with established clinical knowledge while remaining robust to demographic and technical variability. Addressing bias and fairness is particularly critical, as disparities in data representation, imaging access, and genomic reference panels risk amplifying existing health inequities if left unmitigated. Real-world deployment further requires rigorous external validation across diverse populations, careful calibration, and demonstration of clinical utility within constrained healthcare workflows. Governance frameworks must safeguard privacy, ensure informed consent, and establish standards for integrated biomarker use, while policy guidance should extend beyond data security to explicitly address equity and algorithmic accountability. In conclusion, integrative genomic imaging approaches have the potential to substantially improve stroke prevention and long-term vascular risk management. Achieving this potential will depend on embedding interpretability, fairness, and implementation feasibility into every stage of model development. When responsibly designed and validated, such models can support equitable, data-driven decision-making and contribute meaningfully to reducing the global burden of stroke.

REFERENCES

1. Lip GY, Tran G, Genaidy A, Marroquin P, Estes C, Landsheft J. Improving dynamic stroke risk prediction in non-anticoagulated patients with and without atrial fibrillation: comparing common clinical risk scores and machine learning algorithms. European Heart Journal-Quality of Care and Clinical Outcomes. 2022 Sep;8(5):548-56.
2. Li J, Han M, Chen Y, Wu B, Wu Y, Jia W, Liu J, Luo H, Yu P, Tu J, Kuang J. Prediction of recurrent ischaemic stroke using radiomics data and machine learning methods in patients with acute ischaemic stroke: protocol for a multicentre, large sample, prospective observational cohort study in China. BMJ open. 2023 Oct 1;13(10):e076406.
3. 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.
4. Donners SJ, Toorop RJ, de Kleijn DP, de Borst GJ. A narrative review of plaque and brain imaging biomarkers for stroke risk stratification in patients with atherosclerotic carotid artery disease. Annals of Translational Medicine. 2021 Aug;9(15):1260.
5. Oliveira G, Fonseca AC, Ferro J, Oliveira AL. Deep learning-based extraction of biomarkers for the prediction of the functional outcome of ischemic stroke patients. Diagnostics. 2023 Dec 5;13(24):3604.
6. Edyedu I, Ugwu OP, Ugwu CN, Alum EU, Eze VH, Basajja M, Ugwu JN, Ogenyi FC, Ejemot-Nwadiaro RI, Okon MB, Egba SI. The role of pharmacological interventions in managing urological complications during pregnancy and childbirth: A review. Medicine. 2025 Feb 14;104(7):e41381.
7. Hao X, Yao X, Yan J, Risacher SL, Saykin AJ, Zhang D, Shen L, Alzheimer's Disease Neuroimaging Initiative. Identifying multimodal intermediate phenotypes between genetic risk factors and disease status in Alzheimer's disease. Neuroinformatics. 2016 Oct;14(4):439-52.
8. Mohammed S, Bharath K, Kurtek S, Rao A, Baladandayuthapani V. RADIOHEAD. The Annals of Applied Statistics. 2021 Dec 1;15(4):1808-30.
9. Jo H, Kim C, Gwon D, Lee J, Lee J, Park KM, Park S. Combining clinical and imaging data for predicting functional outcomes after acute ischemic stroke: an automated machine learning approach. Scientific reports. 2023 Oct 7;13(1):16926.
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. Ni Y, Alwell K, Moomaw CJ, Woo D, Adeoye O, Flaherty ML, Ferioli S, Mackey J, De Los Rios La Rosa F, Martini S, Khatri P. Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis. PloS one. 2018 Feb 14;13(2):e0192586.

12. Rajashekar D, Hill MD, Demchuk AM, Goyal M, Fiehler J, Forkert ND. Prediction of clinical outcomes in acute ischaemic stroke patients: a comparative study. Frontiers in neurology. 2021 May 6;12:663899.
13. Gerbasi A, Konduri P, Tolhuisen M, Cavalcante F, Rinkel L, Kappelhof M, Wolff L, Coutinho JM, Emmer BJ, Costalat V, Arquizan C. Prognostic value of combined radiomic features from follow-up DWI and T2-FLAIR in acute ischemic stroke. Journal of cardiovascular development and disease. 2022 Dec 19;9(12):468.
14. Okon MB, Ugwu OP, Ugwu CN, Ogenyi FC, Swase DT, Anyanwu CN, Eze VH, Ugwu JN, Akinola SA, Mujinya R, Anyanwu EG. From pandemics to preparedness: harnessing AI, CRISPR, and synthetic biology to counter biosecurity threats. Frontiers in Public Health. 2025 Nov 26;13:1711344.
15. Su PY, Wei YC, Luo H, Liu CH, Huang WY, Chen KF, Lin CP, Wei HY, Lee TH. Machine learning models for predicting influential factors of early outcomes in acute ischemic stroke: registry-based study. JMIR Medical Informatics. 2022 Mar 25;10(3):e32508.
16. Giles D, Foulon C, Pombo G, Ruffle JK, Xu T, Jäger HR, Cardoso J, Ourselin S, Rees G, Jha A, Nachev P. Individualized prescriptive inference in ischaemic stroke. Nature Communications. 2025 Oct 16;16(1):8968.
17. Zingaro A, Ahmad Z, Kholmovski E, Sakata K, Dede' L, Morris AK, Quarteroni A, Trayanova NA. A comprehensive stroke risk assessment by combining atrial computational fluid dynamics simulations and functional patient data. Scientific reports. 2024 Apr 25;14(1):9515.
18. Luijken K, Song J, Groenwold RH. Quantitative prediction error analysis to investigate predictive performance under predictor measurement heterogeneity at model implementation. Diagnostic and Prognostic Research. 2022 Apr 7;6(1):7.
19. Zambrano Chaves JM, Wentland AL, Desai AD, Banerjee I, Kaur G, Correa R, Boutin RD, Maron DJ, Rodriguez F, Sandhu AT, Rubin D. Opportunistic assessment of ischemic heart disease risk using abdominopelvic computed tomography and medical record data: a multimodal explainable artificial intelligence approach. Scientific reports. 2023 Nov 29;13(1):21034.
20. Wang W, Rudd AG, Wang Y, Curcin V, Wolfe CD, Peek N, Bray B. Risk prediction of 30-day mortality after stroke using machine learning: a nationwide registry-based cohort study. BMC neurology. 2022 May 27;22(1):195.
21. Li Y, Salimi-Khorshidi G, Rao S, Canoy D, Hassaine A, Lukasiewicz T, Rahimi K, Mamouei M. Validation of risk prediction models applied to longitudinal electronic health record data for the prediction of major cardiovascular events in the presence of data shifts. European Heart Journal-Digital Health. 2022 Dec 22;3(4):535-47.
22. Meller R, Pearson AN, Hardy JJ, Hall CL, McGuire D, Frankel MR, Simon RP. Blood transcriptome changes after stroke in an African American population. Annals of clinical and translational neurology. 2016 Feb;3(2):70-81.
23. Dinsdale NK, Bluemke E, Sundaresan V, Jenkinson M, Smith SM, Namburete AI. Challenges for machine learning in clinical translation of big data imaging studies. Neuron. 2022 Dec 7;110(23):3866-81.
24. Danso SO, Job DE, Gonzalez DR, Dickie DA, Palmer J, Ure J, Bath PM, Sandercock PA, Wardlaw JM. developing an Integrated Image Bank and Metadata for Large-scale Research in Cerebrovascular disease: our experience from the stroke Image Bank Project. Frontiers in ICT. 2016 Dec 26;3:32.
25. Milne R, Brayne C. We need to think about data governance for dementia research in a digital era. Alzheimer's Research & Therapy. 2020 Jan 31;12(1):17.
26. 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.
27. Valente F, Paredes S, Henriques J. Personalized and reliable decision sets: enhancing interpretability in clinical decision support systems. arXiv preprint arXiv:2107.07483. 2021 Jul 15.
28. Mesinovic M, Watkinson P, Zhu T. Explainable AI for clinical risk prediction: a survey of concepts, methods, and modalities. arXiv preprint arXiv:2308.08407. 2023 Aug 16.
29. Osuala R, Li J, Arandjelovic O. Bringing modern machine learning into clinical practice through the use of intuitive visualization and human–computer interaction. Augmented Human Research. 2019 Dec;4(1):3.
30. Walsh CG, McKillop MM, Lee P, Harris JW, Simpson C, Novak LL. Risky business: a scoping review for communicating results of predictive models between providers and patients. JAMIA open. 2021 Oct 1;4(4):ooab092.
31. Carruthers R, Straw I, Ruffle JK, Herron D, Nelson A, Bzdok D, Fernandez-Reyes D, Rees G, Nachev P. Representational ethical model calibration. NPJ digital medicine. 2022 Nov 4;5(1):170.

32. List JM, Palevsky P, Tamang S, Crowley S, Au D, Yarbrough WC, Navathe AS, Kreisler C, Parikh RB, Wang-Rodriguez J, Klutts JS. Eliminating algorithmic racial bias in clinical decision support algorithms: use cases from the Veterans Health Administration. Health Equity. 2023 Nov 1;7(1):809-16.
33. Olaiya MT, Sodhi-Berry N, Dalli LL, Bam K, Thrift AG, Katzenellenbogen JM, Nedkoff L, Kim J, Kilkenny MF. The allure of big data to improve stroke outcomes: review of current literature. Current neurology and neuroscience reports. 2022 Mar;22(3):151-60.

CITE AS: Serunjogi Ruth (2026). Integrating Long-Read Sequencing with Imaging Biomarkers for Stroke Risk Prediction: Interpretability, Bias, and Real-World Performance with Implementation and Equity Considerations. IAA Journal of Biological Sciences 14(1):87-97. https://doi.org/10.59298/IAAJB/2026/1418797