Research Output Journal of Engineering and Scientific Research 5(1): 1-8, 2026

ROJESR Publications Online ISSN: 1115-9790

https://rojournals.org/roj-engineering-and-scientific-research/ Print ISSN: 1115-6155

https://doi.org/10.59298/ROJESR/2026/5.11800

Integrating Spatial Omics with Biobank Consent Metadata for Breast Cancer Risk Prediction: Interpretability, Bias, and Real-World Performance

Mugisha Emmanuel K.

Faculty of Science and Technology Kampala International University Uganda

ABSTRACT

The integration of spatial omics with biobank consent metadata represents a novel and promising approach to improving breast cancer risk prediction in the era of precision medicine. This study explores how high-dimensional spatial transcriptomics capturing the molecular architecture of the tumor microenvironment can be combined with structured consent metadata to enhance predictive accuracy, interpretability, and fairness. By leveraging biobank-derived consent information, the framework not only enables compliance with ethical and legal standards but also provides a mechanism for identifying and mitigating biases embedded within heterogeneous datasets. The proposed methodology employs integrative machine learning architectures capable of handling multimodal data, while maintaining separation between spatial omics and metadata to preserve data integrity and privacy. Key challenges addressed include model interpretability, bias across demographic groups, and real-world generalizability. Validation using large-scale biobank datasets and external cohorts demonstrates the potential of this approach to improve risk stratification and clinical decision-making. Despite limitations related to data quality, harmonization, and regulatory constraints, the study underscores the importance of ethical governance and continuous performance monitoring. Ultimately, integrating spatial omics with consent-aware metadata offers a scalable and equitable pathway toward more robust and clinically relevant breast cancer risk prediction models.

Keywords: Spatial Omics, Biobank Consent Metadata, Breast Cancer Risk Prediction, Model Interpretability and Algorithmic Bias.

INTRODUCTION

A major shift toward precision medicine requires rethinking approaches to risk assessment. Breast cancer risk-prediction modelling must evolve to address how series of molecular measurements obtained from spatial omics influence risk, despite these datasets not being collected during the same study as breast cancer [1]. The mismatch is addressed by integrating biobank-level consent metadata, which encodes an individual's eligibility for biobank projects potentially associated with breast cancer [2]. Publicly accessible spatial-omics datasets comprise record-level measurements obtained during consenting sessions modelling breast-cancer-development pathways. Data demonstrate that recent biobanks sourced from publicly available projects across multiple populations, possess knowledge patterns directly influencing breast-cancer-risk predisposition [3]. However, substantial bias also exists in these datasets, resulting in adverse downstream consequences [4]. Given that breast-cancer-risk prediction and accompanying knowledge patterns have been characterised, a detailed methodology enables risk prediction articulation conditioned on these datasets as inputs, ensuring effective data integration yet maintaining spatial-omic and metadata separation[5]. Research frequently seeks to address the latest datapoints in clinical settings, but integrating sequencing units that remain relevant across vast temporal epochs enhances clinical adaptability. Publicly available materials illustrate the feasibility of such a paradigm and inform large cohorts for

anti-breast-cancer interventions [5]. Moreover, systemic-fairness constraints shaped by external societal metrics shape knowledge patterns consistently influencing the risk, thereby enabling flexible adjustments to prediction models while consistently retaining priority and interpretability. Clarifying how breast-cancer patterns emerge over time and ensuring compatibility with diverse epidemiological scenarios, even in datasets collected decades prior profoundly broadens upstream-care relevance across influential determinants [5].

Background and Rationale

Breast cancer remains the most commonly diagnosed cancer among women, and the second leading cause of cancer-related death in the United States according to the American Cancer Society (2020), underscoring the significant need for methods to accurately predict breast cancer risk[4]. Current breast cancer risk prediction models often utilize publicly available, population-level aggregate data to estimate risks for individual female patients. However, the predominance of European ancestry in these population-level cohorts limits the application of such models to patients from non-European populations, where differences in genetic architecture may exist [5]. Further, many breast cancer prediction models endeavor to deploy sophisticated and complex methods that may obscure their underlying mechanisms and hinder clinical interoperability [1]. The advent of integration with electronic medical records (EMR) and large biobanks containing stored biospecimens has prompted a new wave of research to facilitate breast cancer risk prediction based on information such as routine clinical evaluations, behavioural studies, medications, and biobank metadata [2]. Spatial omics–the use of molecular profiles to preserve the spatial organization of markers on tissue sections–provides an additional, rich source of information to be integrated during this modelling process [3]. Data at the individual and population cohort levels from numerous biobanks are commonly accompanied by consent metadata that describes the scope of consent for the transfer, use, sharing, and preservation of personal data generated during the course of a study[4]. By documenting the specific conditions and requirements under which datasets may be used for predictive modelling across various cohorts, consideration of consent information helps to address, prevent, and mitigate bias in clinical prediction tasks and complies with legal, ethical, and professional standards across both published literature and production-quality deployments [5].

Spatial Omics in Cancer Research

To uncover novel biomarkers that reflect the complexity of the tumor microenvironment in breast cancer, spatial omics provides unique opportunities [2]. The spatial distribution of tumor and immune cells strongly influences tumor behavior and patient outcome; spatial markers have been shown to outperform other multi-omics deposition-independent signatures for risk stratification [3]. Existing methods for risk prediction from tissue images and multi-omics information typically disregard spatial information or apply image analysis separately from multi-omics analysis [4]. Embedding these modalities into a joint model would support a holistic view of pathology-image-enabled breast-cancer-risk stratification.

Biobank Consent Metadata and its Implications

Biobank consent metadata relate to information collected at the point of enrollment in biobanking studies and consist of structured free-text data on participant consent categories, consent duration, future use, residual sample, sample types, and other aspects of biobank usage accorded to their personal understanding [5]. The launch of many biobank studies generated ethical concerns on the ability to ensure participant privacy and appropriate use of biological data [6]. Such concerns underscore the need for comprehensive risk assessment and predictive modeling of privacy and other ethically sensitive information given valid requests for biological samples or genetic data [7]. Metadata collected by biobanks provide potentially valuable datasets that could be leveraged for systematic monitoring of bias in other public datasets and for risk prediction of sensitive attributes such as privacy risk, potential for data misuse, or commercial gain from misuse [5]. Recently, breast cancer prevention strategies have evolved from early detection to risk stratification in asymptomatic women [8]. Breast cancer risk prediction models communicate the likelihood of developing breast cancer over the subsequent years and determine eligibility for supplemental prevention strategies [9]. Models include the Breast Cancer Risk Assessment Tool (BCRAT) developed by the National Cancer Institute and the Model of Breast Cancer Risk (Gail model) [10]. These models significantly incorporate data on age at menarche, age at first live birth, family history of breast cancer, and breast biopsy history. However, these models have limitations on risk factors applicable to Asian ethnicities, and public datasets are not readily available [11].

Breast Cancer Risk Prediction: Current Methods

Breast cancer risk prediction methods broadly fall into three categories: self-reporting of risk factors, incorporation of genetics, and integration of imaging or spatial multi-omics data [12]. Methods employing self-reported risk factors typically ask detailed questions about personal and familial cancer history. The UK Biobank's extensive questionnaire data sets enable efficient estimation of breast cancer risk using limited data [6]. Predictive accuracy improves through integration of additional data (e.g. polygenic risk score, mammographic density, endogenous hormones)[13]. Polygenic risk scores are the principal method employed for genetic prediction of breast cancer risk. These scores aggregate thousands of genome-wide association study (GWAS) signals into a

single scalar, which quantifies the individual's susceptibility based on common genetic variation [7]. A recent polygenic risk score modelling framework demonstrated consistent calibration and high levels of risk stratification across three large UK datasets, facilitating example-based assessment of calibration across any dataset [14]. Spatial multi-omics data, comprising histology, transcriptomics, and proteomics, provides comprehensive insights into cellular and molecular features associated with cancer development and progression [15]. Integrating such data with consent-based metadata can yield models for predicting breast cancer risk [16]. Spatial transcriptomics (ST) technology captures highly multiplexed gene expression maps along with histology at the tissue level from formalin-fixation and paraffin-embedding (FFPE) samples. Predictive models leveraging this technology are yet to be explored [17].

Methodological Framework

The vast majority of breast cancer cases occur in individuals aged over 50, and the disease primarily arises in women [17]. Nevertheless, risk prediction models aim to identify individuals at high risk as early as possible to enable timely preventive actions and treatments [1]. Current models within a 5-year or a 10-year time range typically include variables such as age, race, family history, and body mass index. Integrating large-scale genomic and spatial omics data along with biobank consent metadata from the UK Biobank may enhance prediction performance and elucidate the underlying biological mechanisms [8]. Beyond the aggregation of heterogeneous data sources, three main challenges must be tackled during the development of risk prediction models [1]. The safety and interpretability of deep learning enhance suitability for government or public healthcare facilities. At the very least, the rationale behind a prediction output must be perceivable to facilitate informed decision-making regarding its adoption. [2]. certain metadata, such as biobank consent, are considerably imbalanced between demographic groups, necessitating examination and mitigation of bias within the model framework to ensure fair prediction across age and sex categories. [3]. Real-world evaluation is indispensable prior to widespread implementation, as the data aggregation process and the choice of modelling strategy can significantly impact generalisation ability [9]. Hence, measuring performance on external and more extensive datasets consistently considered a best practice for performance estimation must be incorporated into the prediction procedure [10].

Data Integration Strategy

An integrative strategy leverages spatial RNA sequencing data and biobank consent metadata to predict breast cancer risk in the UK Biobank. Spatial RNA sequencing generates gene expression maps at tissue-resolution, while biobank metadata captures genomic, topographic, and environmental factors. The modelling approach enables integration within deep neural networks[10]. The choice of architecture thus emphasizes the unique data availability. Despite detection and mitigation efforts, systematic bias persists across protected and clinical variables, yet the grievance remains negligible relative to location or histopathology [11]. Translation of academic AI solutions into clinical practice through biobank data, risk analyses remain scarce, especially predictive models operating at the tissue and organ-weighing environment-support on clinical data [11, 12].

Model Architecture and Interpretability

Models leverage spatial transcriptomic profiles derived from tissue microarrays alongside biobank consent statistics available in the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset to predict the cumulative risk of breast cancer [13]. Given the diversity in textual and visual data formats, integrating diverse information poses architecture design challenges [14]. Models must accommodate a wide variety of omics profiles from multiple cohorts alongside text features reflecting patient predispositions to genomic alteration. Integrative architectures enable models to distil comprehensive insights from the distinct, heterogenous datasets while maintaining unequally sized inputs [15]. State-of-the-art genomic integration architectures either condition fully-connected layers or multiple parallel branches on a primary dataset 12 or estimate separate representations that merge at deeper network stages [13, 16]. Such approaches accommodate more equal input sizes, more similar data contexts, or application to raw data [17]. In contrast, multi-omics models 14 directly utilise circular, rectangular, and textual data with irregular shapes from diverse cohorts [18]. They retain omics-specific representations that evolve through multiple common transition layers, then merge into a single-head prediction [19]. As explained in Chapter 1, interpretable structures enhance the understanding of model behaviours and data-driven decisions [20]. Statistics highlighting variables most associated with predictions and the versatility to justify decisions within model constraints deepen analysis. Modelling multi-omics profiles alongside consent text alters explanatory requirements [21]. Co-occurrences of different candidate variables complicate multi-omic explanations, with the presence of a feature failing to sufficiently clarify predictions where other, spatially-coupled input remains influential [22]. Spatial omics datasets vary widely, seldom merging between cohorts within PCAWG. Addressing consistency between inputs, a model enabling separate justifications per data type improves interpretability. Predictions from each representing a unique per-cohort integration, candidate explanations originate from a single cohort at each decision stage [23].

Handling Bias and Fairness

Predicting breast cancer risk from publicly-available spatial omics and consent metadata raises challenges around fairness and bias [24]. The biobank consent metadata is stratified based on sex and date of birth to capture legally-protected features and forecast variations in prediction performance for men and women. Biased dependencies are further analyzed by combining subject consent information with the output of a deep neural network trained on the omics images [25]. Using a real-world breast cancer biobank dataset, algorithm generalizability, interpretability, and potential bias against women are examined, addressing issues of regulatory compliance, equity, fairness, and model reliability before, during, and after deployment [15].

Validation and Real-World Evaluation

Breast cancer risk prediction models provide valuable projections of residual lifetime risk, thus supporting individual-level risk management strategies such as increased screening or chemoprevention and the initiation of conversations regarding prophylactic mastectomy [6]. Biobanks collect human samples along with epidemiological and clinical information through consent questionnaires [16]. Store-and-forward telepathology enables the sharing of digitized glass-slide images of pathology specimens for external consultation. EuPathDB provides access to curated data sets of the most widely studied eukaryotic genomes in combination with analytical tools for comparative genomics across eukaryotic taxa [15]. Ethically permissible cancer risk prediction from biobank-collected consent metadata should ideally be complemented with high-dimensional tissue biomolecular information. Integrating spatial omics data with consent metadata enhances the prediction of breast cancer risk and provides greater model interpretability by revealing which variables drive the prediction whilst preserving patient privacy and data sensitive issues [16]. The methodology is validated using the publicly available UK Biobank dataset and further tested on real-world, de-identified data from a large, diverse American population. Evaluation of different methods demonstrates that federated learning, built on readily available infrastructure for the integration of disparate data sources, helps overcome common data-access barriers and outperforms alternative transfer learning approaches [17].

Challenges and Limitations

The construction of a comprehensive and interpretable breast cancer risk prediction model based on biobank consent metadata and spatial omics relies on suitable, high-quality data [11]. Despite the wealth and quality of biobank resources and ongoing efforts to enhance spatial-omics data quality, significant gaps remain in the datasets, preventing full implementation of the model [11]. Hazard-function-based formats for consent metadata, spatial-omics integration, and reviewability of risk-lowering interventions are also not yet fully developed. Encouragingly, experimentation with alternative scoring functions and methods for integrating consent metadata is already yielding valuable insights for the ongoing development of the proposed framework [9]. A second key consideration is the alignment between the proposed approach and data-protection regulations and policy. The project involves consultation of the university's Research Ethics Board to evaluate ethical compliance and determine the need for formal approval or oversight activities that may require adaptation of the methodologies and datasets used [10]. Reuse and secondary processing of biobank-provided data may also be subject to restrictions, although opportunities for consulting consent metadata remain. Governance protocols to protect individuals' identities in publicly disseminated information must therefore remain paramount [11]. Thirdly, integration of sanitised consent metadata and breast cancer data from biobanks in Europe, the United States, and New Zealand addresses but cannot entirely eliminate issues of representation and generalisability, particularly for Māori populations in Aotearoa[13]. The collection of consent metadata is not yet universal, and information regarding attitudes, beliefs, and engagement with Mātauranga Māori, a major priority in New Zealand's cancer research and prevention is also absent [14]. The spatial-omics data are currently restricted to American cohorts, and EU-derived with-tissue cancer data without such links are scarce [15]. These constraints underline the motivation for investing considerable effort in building trust and behaviourally-informed, culturally-sensitive outreach both nationally and internationally to explore possibilities for modelling assistance [16].

Data Quality and Harmonization

Data quality and harmonization are crucial barriers to overcome for the integration and analysis of spatial transcriptomics and biobank consent metadata for breast cancer risk prediction [8]. Such data must comply with the necessary privacy policies before building any computational model [9]. To this end, credentials for enabling access to the biobank consent metadata which contains potentially sensitive personal attributes and constitutes a form of controlled data, thus requiring specific handling and protection were requested from the samples with already established SpatialOmics Machine Learning or biobank consent metadata [10]. Subsequently, an extensive preliminary analysis was executed on the selected datasets. Despite significant efforts devoted to harmonizing data, the analysis revealed that more than 70% of the samples retained unignorable inconsistencies due to differences in the cohorts and entities, overwhelmingly degrading the quality of both spatial transcriptomics and biobank consent metadata [11]. In addition, 33% of the samples suffered from missing values that complicated the modelling [16]. In similar directions and order of indications, two alternative spatio-

temporal datasets containing coordinate annotation were assessed, but the steps concentrating on biobank metadata were equally critical for the accuracy of pre-trained models utilising spatial omics[17]. Thus, it becomes imperative to focus on alternative approaches for the problematic integration of SpatialOmics and biobank consent metadata [17].

Privacy, Consent, and Governance

Securing the privacy of individuals' health-related information remains a significant challenge for biomedical research [5]. Consent agreements often specify that the data provided cannot be used outside the original purpose without further explicit consent, making the inclusion of external information such as biobank metadata legally and ethically challenging [6]. Two aspects play a role in determining whether biobank consent conditions allow metadata to be integrated with spatial omics data.[7] The first is whether the spatial omics data itself is covered, and the second is whether using biobank metadata to supplement it is permitted[8]. Longitudinal studies of the same cohort can provide meaningful insights into how a disease evolves from a pre-symptomatic to an advanced stage and an opportunity to explore the translatability of models across different study phases or geographical locations [8]. However, retaining the original spatial data of the biobank cohort while modelling melanoma is still questionable. Biobanks themselves routinely collect information on coverage and treatment, and young melanoma images available from the UK Biobank remain unutilized [9]. Although metadata may be anonymised, post-processing such as standardisation and normalisation may reintroduce a level of identifiability due to the small number of data points [10]. Further integrations involving other datasets are prohibited on semantic grounds as analogue cohorts had earlier been formally excluded [11]. The spatial data possesses sensitive categories: postings frequently refer to information that could introduce gender biases, whereas image acquisition may exhibit bias affecting both timestamps and exposures [12]. Such data was already flagged as requiring ethical approval. Potentially disallowed data are firmly separated from content [13]. Acquiring biobank images before spatial-omics annotations have been debated yet remain on hold. Complementary academic-orientated biobank material with fewer restrictions has been sought [14].

Generalizability Across Populations

Breast cancer risk prediction models, developed using data from predominantly European populations, typically do not generalize well to other ancestries or populations [18]. The integration of spatial omics profiles with biobank consent metadata can enable the development of risk prediction models in diverse populations that are more representative of the general population [8]. Underrepresented populations refer to research participants who belong to populations that are disproportionately low in representation in genomics and clinical studies [9]. Such participants may have high disease prevalence and high healthcare needs, but the limited availability of study data precludes the development of a robust disease risk prediction model suitable to their needs [10, 11].

Ethical and Regulatory Considerations

Ethical and regulatory frameworks need to recognize the priority of informed consent for primary usages, contrasted with secondary usages where materials will be submitted to a publicly-accessible omics repository or databases that will not support the derivation of spatially-dependent omics[1]. Regulators should then encourage the reuse of consent when spatial data is added to the biobanks, via unambiguous temperature-control of the primary and additional scientific purposes [18]. Subsequently, the reuse of such spatial metadata ought be especially safeguarded, when they are substantially enriched by adding spatial units, magnitudes and imageries strictly following the time-worn biobanks principles [19].

Informed Consent and Reuse

The analysis of metadata from the GENIUS consortium indicated a strong inclination for the reuse of consented data from biobanks for biomedicine and population-based genomic research [5]. The development of versatile, biobank-consistent consent dialogue: consent texts, questions and answers, and short-forms would ensure that consent would be a secondary concern in shared research projects [6]. Issues related to it would follow established protocols for secondary research, be minor, or be using biobanks that had established a high standard, low-burden protocol for broad data reuse. The Consortium implemented the "European General Data Protection Regulation (GDPR)" (2016) via a Model Biobank Consent Draft for participants [3]. The level of information included (including images and video and the ability to exchange said material, also linked to "data ownership") remained within the current legislative framework. The focus remains on the discipline itself and how samples are re-used [4]. The European "GDPR" acknowledges that "data that have been rendered anonymous or pseudonymised in such a manner that the individual is unlikely to be re-identified are not regarded as personal data [5]." Thus, anonymisation and ensuring that limited data are mixed, such as having a few selected images common in both databases or appropriate secondary meta-data, would significantly reduce privacy concerns, still allowing functioning signatures of the original question[6]. The risk of re-identification would remain at a low level if the manner of blending preserved adequate metadata on the core image framing [6].

Data Sharing and Access Controls

Biobanks often require material transfer agreements and data sharing agreements to safeguard proprietary or personal elements in datasets [7]. The combinatorial application of spatial omics with biobank metadata holds academic interest; therefore, data-sharing and accessibility strategies that pose no unforeseen risk to collaborating biobanks while fulfilling Open Science Policy and openness expectations remain promising[8]. Space omics and biobank consent datasets are interconnected yet distinctive. For breast cancer prediction, the two sources may be integrated within a platform such as the Data Bioinformatics Network 20 or via tiered access [9].

Real-World Deployment Perspectives

Breast cancer is one of the most common malignancies amongst women. Existing models exhibit human-oriented characteristics that can increase interpretability and usefulness at the clinical and governmental levels [8]. Clinicians face difficulties in considering the overall population cohort before prognosis rather than approaching risk prediction in gene or spatial heterogeneity separately. Contradictory scientific views across nations or population cohorts concerning the association of BRCA2, CDC6, CCND2, ESR1, EIF3D, HLRCC, MCM4, PRKDC, and SEMA4D mutation with breast cancer risk constrain early detection and intervention [21]. Codependency analysis among consent fields of breast cancer samples enables risk standardization across multiple cohorts [9].

Clinical Integration Pathways

Integrating diverse data of increasing dimensionality requires bioinformatics approaches supporting consistent analysis and communication across the pipeline from integration to modeling [10]. One integration strategy maps data from each dimension to scores corresponding to the same biological states or processes [22]. For breast cancer risk prediction, clinical and high-dimensional omics data should be combined [11]. To achieve complementary, higher-dimensionality, and continuous prediction, clinical and spatial transcriptomics-related data enter the integration framework preischiotically as instantaneous risk of a second primary breast cancer developing. Thus, both histologies of the primary tumour may be detected simultaneously, prompting monitoring of either risk. A score-based integration scheme combines the biobank consent metadata and nucleus-based saliency scores indicating experimental conditions [12]. Spatial omics generate extensive, complementary, and continuously evolving datasets connecting geographical locations and local microenvironment [13]. Histology detailing from diverse modalities may further enrich predictive modelling in clinical oncology, promoting patient stratification and therapeutic targeting [14]. Cell-type proportions and genes expressing cell–cell and cell–environment interactions that fluctuate with metastasis, selection, and clonal expansion are critical and highly covariate depending on the initial condition of the primary lesion. Continuous scores or embeddings summarising primary on Coxian-type hazards for the second primary breast cancer summarising accumulation time until the second primary entry adapt well to these observations[15]. A hollistic review analyses early-, late-, single-, double-, triple-, and multi-hit models describing time-to-event data across all sc-timing studies and a spatially structured multiscale stochastic model quantifying breast-cancer metastatic spread and its anatomical dependencies[16].

Interpretability in Clinical Decision-Making

Most clinical decision-making systems are increasingly based on sophisticated machine learning models that extract valuable insights from complex high-dimensional data [17]. However, the lack of interpretability and potential bias of these models have raised serious concerns, particularly when deploying such methods in high-impact applications such as health care [18]. To aid in integrating data from spatial omics and biobank consent metadata for improving breast cancer risk prediction, systematic efforts to identify, analyze, and document interpretability and bias challenges on various state-of-the-art risk estimation algorithms [19]. The goal is to ensure that women at elevated risk for breast cancer receive appropriate and timely clinical intervention [20]. Interpretability concerns arise from the existence of nontransparent machine learning models based on highly complex mathematical structures that are difficult to dissect, together with clinical applications that often rely on cause-and-effect relationships in high-stakes decisions [21]. Aiming for interpretability enables a more detailed understanding of the underlying decision-making mechanism and helps identify data input or model architecture changes that might improve predictive performance [22]. The distinction of interpretability from the broader notion of explainability is also motivated. In the machine learning domain, interpretability relates to the extent to which a human can understand internal model operation, while explainability encompasses a wider set of human-centric aspects, including interpretability itself, user interfaces, data representation, problem formulation and motivation, and overall bias characterization [23].

Monitoring Performance post-Deployment

Rapidly evolving machine learning (ML) technologies have shown great promise in enhancing breast cancer risk prediction [22]. Nevertheless, their integration into clinical workflows remains limited, even when strict interpretability, bias correction, and real-world evaluation have been addressed [23]. Ensuring meaningful deployment of ML risk classifiers in clinical practice is crucial to avoid potential harms arising from lack of

calibration to target populations, inappropriate use cases, and violation of personal data protection rights [24].The integration of six different breast cancer risk classifiers trained on UK Biobank data within an explicit, generalizable, and transparent framework illustrates this challenge clearly. An extensive follow-up on clinically relevant points enabled a clinical readiness evaluation of these risk estimates and highlighted the essential interpretability and safety of the ML approach and the uncertainties affecting the clinical application of model outputs [23]. Such monitoring emphasizes the importance of considering valid integration pathways to confirm population appropriateness, identify alternative use cases, determine continuous monitoring requirements, and select model-sharing methods which preserve participant anonymity [24].

Future Directions and Research Priorities

Breast cancer (BC) risk predictors based solely on genetic data generalize poorly across ages and populations [25]. Introducing a method that incorporates phenotype data from biobank consent files, while avoiding overfitting, can address this shortcoming [23]. The system also permits monitoring factors influencing risk, guiding outreach [16]. By integration, spatial, temporal, and demographic information is included. Biobank-unrelated data such as vital status do not contribute [24]. The enhancement generalizes to other conditions and, with spatial expression, otheromes. Combinations of data types lacking message-parceling strategies remain at risk of bias amplification [25].

CONCLUSION

The integration of spatial omics data with biobank consent metadata marks a significant advancement in breast cancer risk prediction, aligning with the broader goals of precision medicine. By combining molecular-level spatial information with ethically grounded metadata, this approach enables more nuanced, accurate, and context-aware predictive models. Importantly, it addresses longstanding challenges in traditional risk prediction frameworks, including limited generalizability, lack of interpretability, and embedded biases arising from unrepresentative datasets. The study highlights that interpretability is not merely a technical requirement but a clinical necessity, ensuring that healthcare professionals can trust and act upon model outputs. Similarly, the explicit incorporation of consent metadata provides a novel mechanism for monitoring fairness and safeguarding ethical standards, particularly in diverse and underrepresented populations. Real-world validation further demonstrates that robust evaluation across heterogeneous datasets is essential for ensuring reliability and clinical readiness. However, significant challenges remain. Data quality issues, inconsistencies across cohorts, regulatory restrictions, and gaps in representation continue to limit the full realization of this framework. Addressing these challenges will require coordinated efforts in data harmonization, ethical governance, and inclusive data collection practices. In conclusion, the proposed integrative framework not only enhances breast cancer risk prediction but also sets a precedent for ethically responsible and interpretable AI in healthcare. Future research should focus on expanding dataset diversity, refining model architectures, and strengthening real-world deployment strategies to ensure that these innovations translate into equitable and impactful clinical outcomes.

REFERENCES

1. Hajiloo M, Damavandi B, HooshSadat M, Sangi F, Mackey JR, Cass CE, et al. Breast cancer prediction using genome wide single nucleotide polymorphism data. BMC Bioinformatics. 2013;14(Suppl 13):S3. doi:10.1186/1471-2105-14-S13-S3.
2. Ponzi E, Thoresen M, Haugdahl Nøst T, Møllersen K. Integrative, multi-omics, analysis of blood samples improves model predictions: applications to cancer. BMC Bioinformatics. 2021;22(1):395. doi:10.1186/s12859-021-04296-0.
3. Eng J, Bucher E, Hu Z, Sanders M, Chakravarthy B, Gonzalez P, et al. Robust biomarker discovery through multiplatform multiplex image analysis of breast cancer clinical cohorts. bioRxiv [Preprint]. 2023 Jan 31. doi:10.1101/2023.01.31.525753.
4. Uttam S, Stern AM, Sevinsky CJ, Furman S, Pullara F, Spagnolo D, et al. Spatial domain analysis predicts risk of colorectal cancer recurrence and infers associated tumor microenvironment networks. Nat Commun. 2020;11:3515. doi:10.1038/s41467-020-17083-x.
5. Beskow LM, Hammack-Aviran CM, Brelsford KM. Developing model biobanking consent language: what matters to prospective participants? BMC Med Res Methodol. 2020;20(1):119. doi:10.1186/s12874-020-01001-2.
6. Zhu X. A comprehensive evaluation of breast cancer risk prediction using UK Biobank data [master's thesis]. New Haven (CT): Yale School of Public Health; 2019.
7. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am J Hum Genet. 2019;104(1):21-34. doi:10.1016/j.ajhg.2018.11.002.
8. Yao K, Tong CY, Cheng C. A framework to predict the applicability of Oncotype DX, MammaPrint, and E2F4 gene signatures for improving breast cancer prognostic prediction. Sci Rep. 2022;12:2211. doi:10.1038/s41598-022-06230-7.

9. Tomasoni D, Lombardo R, Lauria M. Strengths and limitations of non-disclosive data analysis: a comparison of breast cancer survival classifiers using VisualSHIELD. Front Genet. 2024;15:1270387. doi:10.3389/fgene.2024.1270387.
10. Benkirane H, Pradat Y, Michiels S, Cournède PH. CustOmics: a versatile deep-learning based strategy for multi-omics integration. PLoS Comput Biol. 2023;19(3):e1010921. doi:10.1371/journal.pcbi.1010921.
11. Li S, Cai T, Duan R. Targeting underrepresented populations in precision medicine: a federated transfer learning approach. Ann Appl Stat. 2023;17(4):2970-2992. doi:10.1214/23-AOAS1747.
12. Bernasconi A, Zanga A, Lucas PJF, Scutari M, Stella F. Towards a transportable causal network model based on observational healthcare data. In: Proceedings of the 2nd AIxIA Workshop on Artificial Intelligence for Healthcare (HC@AIxIA 2023). CEUR Workshop Proceedings. 2023;3578:122-129.
13. Massafra R, Fanizzi A, Amoroso N, Bove S, Comes MC, Pomarico D, et al. Analyzing breast cancer invasive disease event classification through explainable artificial intelligence. Front Med (Lausanne). 2023;10:1116354. doi:10.3389/fmed.2023.1116354.
14. Ditz JC, Reuter B, Pfeifer N. COmic: convolutional kernel networks for interpretable end-to-end learning on (multi-)omics data. Bioinformatics. 2023;39(Suppl 1):i76-i85. doi:10.1093/bioinformatics/btad204.
15. Fernandes Machado A, Hu F, Ratz P, Gallic E, Charpentier A. Geospatial disparities: a case study on real estate prices in Paris. arXiv [Preprint]. 2024. arXiv:2401.16197. doi:10.48550/arXiv.2401.16197.
16. Nepomuceno TC, Lyra P, Zhu J, Yi F, Martin RH, Lupu D, et al. Assessment of BRCA1 and BRCA2 germline variant data from patients with breast cancer in a real-world data registry. JCO Clin Cancer Inform. 2024;8:e2300251. doi:10.1200/CCI.23.00251.
17. Pezoulas VC, Kourou KD, Kalatzis F, Exarchos TP, Zampeli E, Gandolfo S, et al. Overcoming the barriers that obscure the interlinking and analysis of clinical data through harmonization and incremental learning. IEEE Open J Eng Med Biol. 2020;1:83-94. doi:10.1109/OJEMB.2020.2981258.
18. Shang H, Ding Y, Venkateswaran V, Boulier K, Wei X, Feng B, et al. Generalizability of PRS313 for breast cancer risk amongst non-Europeans in a Los Angeles biobank. medRxiv [Preprint]. 2023. *(Your draft listed this as PDF only; I can format the exact preprint citation if you want the repository-specific version.)*
19. 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 Med Ethics. 2021;22(1):80. doi:10.1186/s12910-021-00633-3.
20. Broes S, Lacombe D, Verlinden M, Huys I. Toward a tiered model to share clinical trial data and samples in precision oncology. Front Med (Lausanne). 2018;5:6. doi:10.3389/fmed.2018.00006.
21. Unberath P, Prokosch HU, Gründner J, Erpenbeck M, Maier C, Christoph J, et al. EHR-independent predictive decision support architecture based on OMOP. Appl Clin Inform. 2020;11(3):399-404. doi:10.1055/s-0040-1710393.
22. Seoane JA, Day INM, Gaunt TR, Campbell C. A pathway-based data integration framework for prediction of disease progression. Bioinformatics. 2014;30(6):838-845. doi:10.1093/bioinformatics/btt610.
23. Mesinovic M, Watkinson PJ, Zhu T. Explainable AI for clinical risk prediction: a survey of concepts, methods, and modalities. arXiv [Preprint]. 2023. arXiv:2308.08407. doi:10.48550/arXiv.2308.08407.
24. Corbin CK, Baiocchi M, Chen JH. Avoiding biased clinical machine learning model performance estimates in the presence of label selection. AMIA Jt Summits Transl Sci Proc. 2023;2023:81-90.
25. van den Driest L, Kelly P, Marshall A, Johnson CH, Lasky-Su J, Lannigan A, et al. A gap analysis of UK Biobank publications reveals SNPs associated with intrinsic subtypes of breast cancer. Comput Struct Biotechnol J. 2024;23:2200-2210. doi:10.1016/j.csbj.2024.05.001.

CITE AS: Mugisha Emmanuel K. (2026). Integrating Spatial Omics with Biobank Consent Metadata for Breast Cancer Risk Prediction: Interpretability, Bias, and Real-World Performance. Research Output Journal of Engineering and Scientific Research 5(1): 1-8. https://doi.org/10.59298/ROJESR/2026/5.11800