ANNOTATION
Thesis for the PhD Doctor's Degree
in the educational program: 8D10101 - Public Health
Gauhar Abuevna Dunenova
"**Digital repository and machine learning of immunohistochemistry images as integral components of Precision Medicine in breast cancer**"

Relevance of research topic

Breast cancer is the most prevalent cancer among women worldwide, with over 2.3 million new cases and more than 680,000 deaths recorded in 2020 (Sung et al., 2021). Incidence rates are highest in high-income regions—such as Australia, New Zealand, Europe, and North America—where the rate exceeds 80 per 100,000 women, while lower rates of under 40 per 100,000 are observed in Central America, Africa, and South Central Asia (Heer et al., 2020). While high-income countries experience elevated incidence rates, they maintain comparatively low mortality rates (12–15 per 100,000) due to effective screening and treatment measures. By contrast, countries with transitioning economies have mortality rates that are 17% higher (Sung et al., 2021). Five-year survival rates also highlight these disparities: they exceed 90% in high-income countries, while in India and South Africa, survival rates fall to 66% and 40%, respectively (WHO). In Kazakhstan, breast cancer is the leading cancer among women, with a five-year overall survival rate of 57.1% (Kaidarova et al., 2023). Recent Globocan data (2022) reports 4,570 new cases and 1,574 deaths, with standardized incidence and mortality rates of 36.9 and 12.3 per 100,000, respectively (Ferlay et al., 2021).

Immunohistochemical (IHC) testing for receptor status has become a cornerstone of breast cancer diagnostics (Francis et al., 2022), laying the groundwork for precision medicine approaches in treatment. This testing enables classification of breast cancer into five primary subtypes, including HER2-positive, which is critical for guiding treatment and survival prognosis (Goddard et al., 2011; Gradishar et al., 2022; Hammond et al., 2010). Manual analysis of HER2-stained IHC images is complex and requires significant expertise, especially when faced with heterogeneous staining patterns (Palm et al., 2023). The situation is aggravated by the growing shortage of pathologists (Gross et al., 2023; Rozario et al., 2024): according to the literature, only 1.1% to 8.6% of medical school graduates choose this specialization as their main one (Masuadi et al., 2021).

Advancements in digital pathology and artificial intelligence (AI) have introduced new possibilities for diagnostic automation and precision. Whole-slide imaging (WSI) and machine learning (ML) algorithms can now process large datasets, uncover hidden patterns (Bhalla & Laganà, 2022; Yee, 2021), and improve diagnostic reproducibility. The accuracy of digital HER2 analysis, for example, has reached levels as high as 90% (Kabir et al., 2024; Yao et al., 2022).

In Kazakhstan, IHC testing for breast cancer has been in practice since 2012, but phenotype data remains fragmented and lacks integration into a centralized oncology registry, impeding comprehensive epidemiological analysis. The adoption of digital technologies is a priority in Kazakhstan's development strategies, as reflected in state programs aimed at reducing premature cancer mortality. Integrating biomedical data, AI, and digital repositories within precision medicine frameworks could significantly enhance breast cancer diagnosis and treatment in Kazakhstan. Digital repositories and AI-based diagnostic tools are pivotal for achieving precision medicine goals, improving economic efficiency (Smetherman et al., 2022), and supporting national health databases for evidence-based healthcare policy (Lauricella & Pêgo-Fernandes, 2022).

Given the absence of research efforts in Kazakhstan focused on developing and implementing digital repositories for automated breast cancer marker diagnosis using AI algorithms, this study is particularly timely. It addresses the pressing need for innovative diagnostic approaches and provides a foundation for future developments in digital healthcare, shaping the aims and objectives of this research.

Research aim

To develop and validate an algorithm for machine-assisted diagnostics of breast cancer, specifically utilizing the Human Epidermal Growth Factor Receptor 2 (HER2) marker, through the creation of a prototype digital repository of immunohistochemical images. This repository aims to enhance the

accuracy and effectiveness of breast cancer immunohistochemical diagnostics within the Republic of Kazakhstan.

Research objectives

1. To conduct a comprehensive analysis of breast cancer epidemiological indicators, including immunohistochemistry diagnostic data within the Republic of Kazakhstan, and evaluate disability-adjusted life years (DALY) specific to the city of Almaty.
2. To review and evaluate HER2 digital analysis algorithms, focusing on their performance characteristics and criteria for diagnostic efficacy (systematic review).
3. To assess the organizational structure of Kazakhstan's current pathomorphological service to establish prerequisites for implementing artificial intelligence and machine learning methods in immunohistochemical diagnostics.
4. To develop a machine-learning algorithm for HER2 breast cancer diagnosis, based on a prototype digital repository, and validate its effectiveness through training and testing on annotated image sets.

Materials and methods of research

The study was conducted using the following methodological framework:

Study design:

1. Epidemiological analysis: addressing the first objective, this study involved an epidemiological investigation into breast cancer phenotypes, with a comprehensive analysis and calculation of the BC burden.
2. Systematic review: for the second objective, a systematic review was conducted to evaluate the efficacy of digital analysis methods for the HER2 marker, specifically focusing on machine learning applications.
3. Descriptive and analytical research: in line with the third objective, the study employed descriptive and analytical methods, including content analysis, to assess data patterns and thematic insights.
4. Experimental research: the fourth objective involved experimental research, testing the performance and accuracy of proposed methodologies under controlled conditions.

Object of research

This study analyzes breast cancer incidence reports in the Republic of Kazakhstan for the period 2012–2021, based on form No. 7 of the Ministry of Health of the Republic of Kazakhstan ("Report on Malignant Diseases", C50) (n=44,331). Data on immunohistochemistry diagnostics for breast cancer were collected from the EROB database for Almaty and its region for the period 2020–2022 (n=2783, C50, C50.0-C50.9). IHC indicators were extracted from reports of the State Enterprise 'Almaty Oncology Center' and the State Enterprise 'Almaty Regional Multidisciplinary Clinic, and combined into a unified database for breast cancer patients for the years 2020–2022. Statistical analysis of these data was performed using the IBM SPSS Statistics 26 software. Further, the study includes statistical data on Disability-Adjusted Life Years (DALY) for breast cancer in Almaty for 2017–2021. To calculate Years of Life Lost (YLL) and Years Lived with Disability (YLD), aggregated data on prevalence and mortality due to breast cancer, disaggregated by disease code, age group, gender, and year of death in Almaty, were sourced from the Center for Disease Control and Prevention. The Global Burden of Disease (GBD) 2017 life expectancy table and GBD 2013 disability weights were applied for analysis. Additionally, the study includes a systematic review of current algorithms for digital HER2 analysis, as well as a review of normative legal acts (NPA) governing the operations of Kazakhstan's pathomorphology services. This encompasses telemedicine infrastructure, current workload, and computational resources available for implementing AI methodologies in digital IHC image analysis. Finally, digital images of the HER2 marker, collected from pathology laboratories, were compiled for developing the machine learning algorithm.

Subject of research

This study examines key epidemiological metrics related to breast cancer, including standardized incidence and mortality rates, and stage distribution across cases in the Republic of Kazakhstan for the period 2012–2021. It also explores the prevalence of IHC phenotypes of breast cancer in Almaty and its region for the years 2020-2022, as well as the DALY (Disability-Adjusted Life Years) index for breast cancer in Almaty between 2017 and 2021. Furthermore, the study assesses global algorithms for digital

HER2 analysis that utilize ML techniques, evaluating their accuracy, reproducibility, and potential for clinical integration. The research reviews current standards and methodologies in morphological and IHC diagnostics for breast cancer. Additionally, it identifies organizational and technical prerequisites essential for implementing AI algorithms within Kazakhstan's pathomorphological laboratories, specifically considering the capabilities of the existing national portal for pathomorphological research. The study involves digitized HER2 IHC images of breast cancer samples (n=419), a prototype system architecture, and a structured algorithm for digital image analysis, with a focus on validating the criteria for diagnostic accuracy.

Key findings and contributions for defense

1. In the Republic of Kazakhstan for the period from 2012 to 2021, the incidence of breast cancer increased, and the mortality rate decreased. The standardized DALY indicator in Almaty decreased from 2017 to 2017, which indicates an improvement in survival, despite the significant burden of the disease. IHC diagnostics of breast cancer is routinely used in Almaty and Almaty region for HER2 screening, which creates potential conditions for the formation of large-scale databases in the digital IHC repository for epidemiological analysis, health resources planning and improvement of digital diagnostics algorithms.
2. Digital analysis of HER2 is an actively developing and studied method in the world, it demonstrates high experimental accuracy and potential clinical significance.
3. The oncomorphological service in Kazakhstan is characterized by significant differences in the distribution of workload and resources across the regions, as well as a chortage of specialists. Integrating AI technologies into IHC workflows, based on a national portal for pathomorphological studies, holds potential as a supplementary tool to assist pathologists.
4. The architecture of the prototype of the digital repository allows to structure the existing practice of using the national portal of pathomorphological studies, ensuring the development of clinical, educational, scientific directions in the pathomorphology of the oncological service of the Republic of Kazakhstan. Algorithm of digital analysis of HER2, developed on the Kazakh data set, has an accuracy of 75%.

Scientific novelty: As a result of the conducted study, new data on epidemiological indicators of BC in the Republic of Kazakhstan, IHC phenotypes of BC at the national level in the period 2020-2022 were obtained. For the first time, the indicator DALY from the hospital in Almaty was calculated for the period 2017-2021 (Author's certificate #52072 dated 02.12.2024). A systematic review of the algorithms of digital analysis of HER2, their characteristics, which influence their effectiveness from the point of view of clinical evaluation, has been carried out. For the first time, a dataset of digital images of IHC marker HER2 of breast cancer (n=419) was created (Author's certificate #48772 dated 01.08.2024). For the first time, a prototype architecture was developed (Author's certificate No. 32121 dated 30.01.2023) and an algorithm for digital analysis of images of the IHC marker HER2 was proposed, tested on our own dataset and demonstrated 75% accuracy results.

Determined possibilities of application of AI methods in IHC on the platform Qazhisto.com.

Theoretical significance of research

The obtained data on the epidemiology of IHC phenotypes of breast cancer allow the development of methodological recommendations for improving the oncological register, including the integration of IHC diagnostics with current patient data. Using DALY to estimate the total burden of breast cancer (mortality and morbidity) helps to assess the effectiveness of strategies, optimize resources and adjust programs to reduce the burden of breast cancer at the national level.

A systematic review of world algorithms for digital analysis of HER2 forms the basis for the development and organization of components of a new algorithm with clinical evaluation.

Data on the structure of the pathomorphological service will help to develop recommendations for the training of oncomorphologists for complex studies. Additionally, this data may serve as a scientific foundation for proposing legislative amendments and initiating the development of normative legal acts that regulate the application of artificial intelligence in medicine.

The proposed architecture supports the creation of large digital libraries of slides, accessible to pathologists regardless of location, which will enrich future research.

The developed algorithm of digital analysis of HER2 contributes to the improvement of the biomedical data collection system for further research.

Practical significance of research

A labeled clinical dataset of digital images of IHC HER2 of breast cancer has been established, as documented in the implementation act by Department of Epidemiology, Biostatistics, and Evidence-Based Medicine, Faculty of Medicine and Health Care, KazNU. This dataset facilitates the development and implementation of algorithms for digital analysis of these images, fostering scientific collaboration to advance research in oncology and information technologies.

The dataset created on the Qazhisto platform can be utilized for educational purposes, training students and pathologists, and can serve as a supplementary tool for obtaining a "second opinion" among practicing oncomorphologists.

The proposed architecture provides a foundation for the development and potential enhancement of the capabilities of the Qazhisto platform, as evidenced by the implementation act from the Almaty oncology center molecular research laboratory.

Methodological recommendations have been developed for conducting teleconsultations on tumor biospecimens at the national level. These recommendations encompass the organizational structure and workflow of telepathology utilizing the Qazhisto platform, as detailed in the implementation acts from the KazIOR, AOC and six regional oncocenters.

The scientific data obtained from this research can serve as a basis for integrating artificial intelligence methods in immunohistochemical analysis within the current framework of the Qazhisto pathomorphological service in the Republic of Kazakhstan.

The proposed recommendations regarding the algorithm's application are intended to enhance the efficient use of resources on the Qazhisto platform and facilitate data accumulation.

Personal contribution to the doctoral student

Includes development of theoretical and methodological research program, participation in all stages of work, creation of database and dataset, statistical processing of data, writing of thesis sections, interpretation of data and development of practical recommendations. Task setting and discussion of results were carried out together with scientific consultants. The thesis was completed on an initiative basis.

Conclusion

1. Prevalence of BC in Kazakhstan for the period 2012–2021 increased from 314.4 in 2012. do 444.3 in 2021 per 100 thousand female population (average growth rate - 3.92%). Between 2020 and 2022, the most prevalent phenotype in Almaty and the Almaty region was Luminal B, HER2 negative, accounting for 42.4% of cases. The HER2-enriched phenotype, which may be suitable for inclusion in the digital repository of HER2 breast cancer IHC images, represented 16.7% of cases. The phenotype with the highest mortality risk was triple-negative breast cancer, which comprised 11.9% of cases and was associated with a sevenfold increased risk of death ($p = 0.001$). This group also exhibited the highest proportion of women under 45 years of age (24.9%). Stage II was the most commonly diagnosed stage, accounting for 58% of all cases. The predominant histological subtype was invasive ductal carcinoma (NST), which accounted for 95.9% of cases. There was a decrease in standardized DALY for the studied period from 638.9 to 489.5, with some increase in YLD (6.7% vs 6.0%).
2. In world practice, the accuracy of algorithms of digital analysis of IHC HER2 artificially created with the use of computational algorithms on the basis of publicly available resource Warwick had reached 98.8%.
3. In 2021, significant regional differences in the workload of the pathomorphology service were observed in Kazakhstan: the volume of IHC studies varied from 283 in the Taldykorgan Oncology center to 9050 in the AOC. The workload of staff units varied from 54 to 2262.5 per study unit, and for individuals — from 81 to 2262.5. The average staffing level was 64%. A potential solution to this issue may involve the targeted registration of residents and specialists in the field of pathomorphology, which could serve as a foundation for optimizing the planning and regional distribution of human resources. The existing Qazhisto platform, which leverages machine learning algorithms for diagnostic and educational purposes, offers an environment conducive to creating optimal conditions for training

specialists and centralizing information on the available human resources within the field of pathomorphology.

4. An experimental study of the digital analysis of HER2 using MO showed an average accuracy model (Accuracy) of 75%, with an accuracy of Precision 1.0 for the HER2 "3+" class.

Practical recommendations

To the Ministry of Health of the Republic of Kazakhstan: Consider the creation of a reference center for the centralized collection and analysis of data on breast cancer, including digital IHC images, to improve resource planning, epidemiological monitoring and screening. Include DALY indicators every 5 years to estimate disease burden and resource allocation. Implement the Qazhisto platform for data storage and analysis, available to profile institutions. Increase the number of pathologists through reforming the residency program, using the digital repository in training as an educational platform for training specialists in the field of digital pathology. Consider categorizing residents and pathology specialists separately in reporting forms to facilitate more accurate tracking of their numbers and better monitor workforce dynamics. Create an expert group to establish the standards of workload of the pathologist. Initiate the creation of a specialized expert group with the aim of developing scientifically based recommendations on the use of AI technologies in medical practice.

Healthcare managers (head of pathomorphological laboratories oncomorphological service): Introduce regular training for medical personnel and pathologists in modern methods of digital pathology, organize webinars and seminars with the involvement of local and international experts.

For the Ministry of Education and Science and educational institutions: Create an interdisciplinary group for the development of joint research with international centers, which will allow to adapt and implement advanced technologies of digital pathology in local practice. Support the development of grant programs to stimulate research and development in the field of digital pathology and AI.

Publications on the topic of thesis: 4 works were published on the subject of the thesis, of which 2 articles were published in journals indexed in the Scopus database: Cancers (Basel) (Scopus, 5-year IF: 4.9, ISSN 2072-6694, CiteScore 8,0, Rank#81/404, 79 percentile in category «Oncology»); Scientific Reports (Scopus, 5-year IF: 4.3, ISSN:2045-2322, CiteScore 7,5, , Rank#14/171, 92 percentile in category «Multidisciplinary»), 1 article was published in journals recommended by the Committee for Control in Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan, and 1 methodological recommendations (ISBN 978-601-7548-27-8).

Approbation thesis at conferences: KazNU interdisciplinary conference "Automation and Machine Learning in Medicine and Health Organization", 05.02.2021, report on the topic "Digital repository and machine diagnostics of immunohistochemical preparations as integral elements of Precision Medicine in breast cancer". Scientific forum "Intellectual potential of independent Kazakhstan: 30 years of formation and development", section of young scientists "Artificial intelligence, big data & medicine: challenges and prospects", 09.09.2021, report on the topic "Machine diagnostics of IHC images. Perspectives of development".

Innovation patents, copyright certificates: Author's certificate #32121 "Architecture of support of digital analysis in immunohistochemical diagnosis of breast cancer" 30.01.2023. Author's certificate #48772 "ADEL Dataset. Dataset of digital images of HER2 breast cancer" 01.08.2024. Author's certificate #52072 «Adapted methodology for studying the burden of cancer in the Republic of Kazakhstan» 02.12.2024.

Based on the results of the thesis, the following have been developed: Methodological Recommendations for conducting teleconsultations of tumor biospecimens at the national level, which include guidelines on the organizational structure and workflow of telepathology using the Qazhisto platform. Implementation Acts for integrating the outcomes of scientific research into the practices of healthcare organizations.

Structure and the scope of the thesis: The thesis is presented in 178 pages and comprises an introduction, three sections, a conclusion, findings, practical recommendations, and appendices. The work is illustrated with 22 tables and 27 figures. The bibliography includes 41 domestic and 235 international sources.