THE CONCEPT OF PREDICTABLE HARM IN DEVELOPMENT OF AI-POWERED MEDICAL DEVICES

Begishev IR ✉, Shutova AA

Kazan Innovative University named after VG Timiryasov, Kazan, Russia

The article reviews the concept of predictable harm as a methodological tool for a comprehensive risk assessment while developing and implementing AI-powered medical devices. The study is relevant due to exponential growth of using AI-powered technologies in healthcare and lack of unified approaches to prediction of potential negative consequences of their usage. Existing regulatory approaches to risk assessment, including Russian regulatory documents and international standards, have been analyzed. A multidimensional classification of types of predictable harm is proposed considering the entire life cycle of medical AI systems. Special attention is given to ethical aspects of using artificial intelligence in medicine, including the principles of patient autonomy, equity, non-harm and transparency of algorithms. An expanded matrix for assessing predictable harm has been developed. It integrated technological, clinical and ethical parameters for each stage of development and implementation of AI systems in medical practice. The results of the study can be used as a methodological framework for developers of medical AI systems, regulatory authorities and medical organizations in assessing safety and effectiveness of introducing intelligent technologies into clinical practice.

Keywords: artificial intelligence, medical devices, predictable harm, ethics of artificial intelligence, regulation, patient safety, risk management

Financing: the work was carried out under the grant from the Academy of Sciences of the Republic of Tatarstan, provided to young candidates of sciences (postdoctoral fellows) so that they could defend their doctoral dissertation, perform research and labor functions in scientific and educational organizations of the Republic of Tatarstan within the framework of the Scientific and Technological Development of the Republic of Tatarstan State Program.

Acknowledgements: the authors express their sincere gratitude to the Academy of Sciences of the Republic of Tatarstan for financial support of this study based on the results of a competitive selection (grant No. 153/2024-PD dated December 16, 2024) for ‘Ensuring the technological sovereignty of the healthcare system by criminal means’ research project, as well as deep gratitude to the staff of the Scientific Research Institute of Digital Technologies and Law of Kazan Innovative University named after Timiryasov VG for valuable consultations and constructive discussion of the conceptual provisions of the study.

Author contribution: Begishev IR — research conceptualization; development of predictable harm theoretical foundations; analysis of specific risks associated with using artificial intelligence in medical devices; systematization of predictable harm typology; research of legal framework for regulating AI systems in healthcare with various jurisdictions; formulation of research conclusions; preparation of the manuscript initial version; Shutova AA — development of a methodology for predicting and preventing predictable harm; creation of a matrix for assessing predictable harm at various stages of medical AI system life cycle; developing recommendations for practical implementation of predictable harm concept; analysis of literature sources; editing and critical revision of the manuscript with introduction of valuable intellectual content; visualization of research (making tables).

Compliance with ethical standards: a meeting of the ethics committee was not required, since this study is theoretical and methodological in nature and analyzes open literature sources and regulatory legal documents, without conducting experiments involving humans or animals and without using personal patient data.

✉ **Correspondence should be addressed:** Ildar R. Begishev
42 Moskovskaya St., Kazan, Russia, 420111; begishev@mail.ru

Received: 03.05.2025**Accepted:** 16.05.2025 **Published online:** 20.06.2025

DOI: 10.24075/medet.2025.006

КОНЦЕПЦИЯ «ПРЕДСКАЗУЕМОГО ВРЕДА» ПРИ РАЗРАБОТКЕ МЕДИЦИНСКИХ ИЗДЕЛИЙ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

И. Р. Бегишев ✉, А. А. Шутова

Казанский инновационный университет имени В. Г. Тимирясова, Казань, Россия

В статье представлена концепция «предсказуемого вреда» как методологический инструмент для комплексной оценки рисков при разработке и внедрении медицинских изделий на основе искусственного интеллекта. Актуальность исследования обусловлена экспоненциальным ростом применения ИИ-технологий в здравоохранении при одновременном отсутствии унифицированных подходов к прогнозированию потенциальных негативных последствий их использования. Проведен критический анализ существующих регуляторных подходов к оценке рисков, включая отечественные нормативные документы и международные стандарты. Предложена многомерная классификация типов предсказуемого вреда с учетом всего жизненного цикла медицинских ИИ-систем. Особое внимание уделено этическим аспектам применения искусственного интеллекта в медицине, включая принципы автономии пациента, справедливости, непричинения вреда и прозрачности алгоритмов. Разработана расширенная матрица оценки предсказуемого вреда, интегрирующая технологические, клинические и этические параметры для каждого этапа разработки и внедрения ИИ-систем в медицинскую практику. Результаты исследования могут служить методологической основой для разработчиков медицинских ИИ-систем, регуляторных органов и медицинских организаций при оценке безопасности и эффективности внедрения интеллектуальных технологий в клиническую практику.

Ключевые слова: искусственный интеллект, медицинские изделия, предсказуемый вред, этика искусственного интеллекта, регулирование, безопасность пациентов, управление рисками

Финансирование: работа выполнена за счет гранта Академии наук Республики Татарстан, предоставленного молодым кандидатам наук (постдокторантам) с целью защиты докторской диссертации, выполнения научно-исследовательских работ, а также выполнения трудовых функций в научных и образовательных организациях Республики Татарстан в рамках Государственной программы Республики Татарстан «Научно-технологическое развитие Республики Татарстан».

Благодарности: авторы выражают искреннюю признательность Академии наук Республики Татарстан за оказанную финансовую поддержку данного исследования по результатам конкурсного отбора (грант № 153/2024-ПД от 16 декабря 2024 г.) для научно-исследовательского проекта «Обеспечение технологического суверенитета системы здравоохранения уголовно-правовыми средствами», а также глубокую благодарность коллективу Научно-исследовательского института цифровых технологий и права Казанского инновационного университета имени В. Г. Тимирясова за ценные консультации и конструктивное обсуждение концептуальных положений исследования.

Вклад авторов: И. Р. Бегишев — концептуализация исследования; разработка теоретических основ понятия «предсказуемого вреда»; анализ специфики рисков, связанных с применением искусственного интеллекта в медицинских изделиях; систематизация типологии предсказуемого вреда; исследование нормативных правовых основ регулирования ИИ-систем в здравоохранении в различных юрисдикциях; формулирование выводов исследования; подготовка первоначального варианта рукописи; А. А. Шутова — разработка методологии прогнозирования и превенции предсказуемого вреда; создание матрицы оценки предсказуемого вреда на различных этапах жизненного цикла медицинских ИИ-систем; формирование рекомендаций по практической имплементации концепции предсказуемого вреда; анализ источников литературы; редактирование и критический пересмотр рукописи с внесением ценного интеллектуального содержания; визуализация исследования (разработка таблиц).

Соблюдение этических стандартов: заседание этического комитета не требовалось, поскольку данное исследование носит теоретико-методологический характер и основано на анализе открытых литературных источников и нормативных правовых документов, без проведения экспериментов с участием людей или животных и без использования персональных данных пациентов.

Для корреспонденции: Ильдар Рустамович Бегишев
ул. Московская, д. 42, Республика Татарстан, г. Казань, 420111; begishev@mail.ru

Статья поступила: 03.05.2025 **Статья принята к печати:** 16.05.2025 **Опубликована онлайн:** 20.06.2025

DOI: 10.24075/medet.2025.006

Integration of artificial intelligence into medical devices offers unprecedented opportunities to improve diagnostic processes, personalize therapeutic approaches, and optimize clinical solutions. However, rapid introduction of AI systems into healthcare is associated with specific risks that require systematic analysis and proactive management. In this context, the concept of predictable harm is gaining crucial significance as a methodological tool for preventive identification and minimization of potential negative consequences of using AI-powered medical devices. A critical analysis of existing regulatory approaches to assessing the risks of AI systems in healthcare, as well as integration of ethical principles into the process of forecasting and preventing possible harm is of particular importance.

The relevance of the study is determined by exponential growth of the market for AI solutions in healthcare and lack of unified approaches assessing their safety. According to the Grand View Research analytical report, the global market for artificial intelligence in medicine will reach 120.2 billion US dollars by 2028 with an annual increase of about 41.8%. It shows the scope of challenges in the field of patient safety [1].

The purpose of this study is to form a methodological framework for identifying and minimizing predictable harm when developing and implementing AI-powered medical devices.

To achieve this goal, the following tasks have been set:

1. To conceptualize the term of predictable harm in the context of medical AI technologies;
2. To analyze the specifics of the risks associated with the use of artificial intelligence in medical devices;
3. To investigate existing approaches to regulation of safety of AI systems in healthcare and compare them with the author's concept;
4. To develop a methodology for predicting and preventing potential harm when creating medical AI systems with detailed ethical elaboration.

ESSENTIAL PART

The concept of predictable harm for AI-powered medical devices is a methodological construct that integrates the principles of predictive risk analysis, proactive safety management, and iterative reassessment of the potential harmful effects of the technology. The fundamental difference of this concept from traditional approaches to risk assessment is in the shift of focus from reactive incident response to preventive forecasting of possible scenarios of adverse events caused by specific functioning of AI systems.

In the reviewed context, the terminological definition of predictable harm can be formulated as a set of potential negative effects of using medical AI systems. It is possible to identify and minimize them systematically analyzing characteristics of the technology, the context of its application and possible trajectories of evolution of the system during operation. The key attributes of this definition include predictive nature of assessment, a systematic approach to risk analysis, and consideration of the dynamic nature of AI technologies.

At the present stage, there are some regulatory approaches to risk assessment in the use of AI-powered medical devices.

Thus, Order No. 686n of the Ministry of Health of the Russian Federation dated July 7, 2020 [2] and letter No. 02I-297/20 of the Federal Service for Healthcare Supervision dated February 13, 2020 [3] provide for risk rating. According to it, all AI-powered medical devices are classified as Class III MD before their application and at the stage of state registration. This approach is aimed at centralized regulation and a priori high-risk classification of all AI systems in medicine.

The International Forum of Medical Device Regulators (IMDRF, 2014) offers a differentiated classification of potential risks of AI-powered medical devices, depending on clinical application and possible impact on the treatment process (Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations) [4]. This classification takes into account both the severity of potential harm to the patient and the role the AI system plays in the clinical process.

The industry appendix to the Code of Ethics for Artificial Intelligence of Alliance for Artificial Intelligence provides details for gradation of risks depending on the severity of errors associated with the use of artificial intelligence systems and focuses on consequences of incorrect medical decisions [5].

Specific risks associated with the use of artificial intelligence in medical devices are due to a number of unique characteristics of these technologies: autonomous functioning, potential non-transparency of the decision-making process (the black box problem), capacity to self-education and adaptation, and high dependence on source data quality [6]. These features constitute a multidimensional risk profile that needs a differentiated approach to their identification and management.

The proposed concept of predictable harm is characterized by a multidimensional structure focused on the entire life-cycle of AI systems. It includes as follows:

- proactive focus on risk identification;

Table 1. Typology of predictable harm for medical AI systems

Harm category	Distinguishing feature	Examples	Identification and minimization methods
Algorithmic	It is associated with defects in mathematical models and decision-making logic	False positive/ false negative results, classification errors	Validation of representative samples, testing of boundary cases
Data-centric	Caused by problems in the training data	Systematic bias, inapplicability to certain groups of patients	Audit of data, stratification of samples, control of representativeness
Integrative	It occurs when the AI system interacts with clinical processes	Violations of protocols, conflicts with existing systems	Process modeling, simulation of clinical scenarios
Interpretative	Associated with incorrect interpretation of results by users	Overestimation/underestimation of AI recommendations, omission of critical information	Interface optimization, user training
Evolutionary	It is driven by the change in the system behavior during self-education	Concept drift, accuracy degradation	Performance monitoring, periodic recertification

– differentiated approach to types of harm (algorithmic, data-centric, integration, etc.);
– multilevel stratification of responsibilities of participants;
– iterative risk assessment and adaptation to the evolution of AI systems;
– integration of technological and clinical aspects of quality.

Table 1 systematizes the types of predictable harm for medical AI systems.

Safety regulation of AI-powered medical devices is characterized by significant heterogeneity of approaches in different jurisdictions. The European Union is implementing a structured regulatory system through the Medical Devices Regulation (MDR 2017/745) [7] and the Artificial Intelligence Regulation (AI Act) [8], which classifies medical AI systems as high-risk ones and sets strict requirements for their transparency and validation.

The US Food and Drug Administration (FDA) implements an adaptive approach which is based on the Pre-Certification Program focusing on evaluation of development processes and quality culture of the developer [9]. This approach involves continuous monitoring of system performance under real operating conditions and iterative reassessment of the risk-benefit profile.

Based on the analysis of existing approaches and regulatory requirements, an integrated methodology for predicting and preventing predictable harm when developing AI-powered medical devices is proposed, which includes the following components: a multi-level risk assessment model, an inclusive validation system for heterogeneous populations of patients, mechanisms for ensuring interpretability of algorithms, an infrastructure for continuous performance monitoring, and iterative safety reassessment processes.

The proposed methodology can be implemented in practice through the matrix of predictable harm assessment presented in Table 2.

Ethical aspects form an integral part of the predictable harm concept and are reflected at all stages of AI-powered medical device life cycle. Let's look at the main dimensions of ethical responsibility:

1. Patient autonomy — it is critically important to make sure that implementation of AI systems does not diminish the role of the patient in the decision-making process. The risks of excessive automatic trust of clinicians in AI advice and quality of informed consent should be taken into account.
2. Justice means preventing algorithmic discrimination and providing access to AI technologies to various groups. It is necessary to concentrate on data-centric risks and data representativeness.
3. Non-harming takes into account the possibility of delayed and systemic consequences associated with evolutionary changes and self-learning algorithms.
4. Transparency and explainability means ensuring interpretability of decisions and audit opportunities for both specialists and patients; overcoming the black box effect.
5. Mandatory ethical audit is analysis of compliance of artificial intelligence used with medical ethics, and regular revision of the risk matrix taking into account vulnerability of certain categories of patients and long-term consequences.

These provisions are shown in the matrix of foreseeable harm assessment and presented in details (see Table 2).

The use of the matrix allows to structure the process of identifying and minimizing predictable harm at all stages of AI-powered medical device life cycle, providing an integrated approach to risk management and compliance with regulatory and ethical requirements.

While developing medical AI systems, forming a culture of transparency is crucial for effective implementation of the predictable harm concept. This aspect includes open communication regarding technological limitations, active involvement of clinical specialists at all stages of product creation, and use of the safety through design principle involving integration of safety mechanisms directly into the system architecture. In contrast to existing regulatory approaches that focus primarily on technical characteristics and preliminary risk classification, the proposed concept assumes mandatory integration of ethical audits at each stage of the life cycle of a medical AI system. This requires multidisciplinary collaboration between developers, clinicians, ethicists, and patient community representatives to prevent algorithmic discrimination, preserve patient autonomy, and maintain equitable access to the benefits of technology. The matrix of predictable harm assessment, including ethical and clinical parameters, becomes not a simple documentation tool, but a platform for continuous dialogue between all participants in the process of introducing AI into clinical practice.

SUMMARY AND CONCLUSIONS

The conducted research allows us to formulate the following main conclusions:

1. The predictable harm concept is an effective tool for improving the safety of introducing artificial intelligence into medicine, which proactively identifies and minimizes risks.

Table 2. Matrix assessing predictable harm to medical AI systems

Stage	Key assessment issues	Tools	Responsible parties	Clinical aspects	Ethical aspects
Conceptualization and design	Compliance with the target application, coverage of clinical scenarios, technical feasibility	Ethical audit, analysis of clinical scenarios, review of the evidence base	Developers, clinical experts, ethics committees	Assessing clinical significance of the problem being solved, potential changes in clinical practice, and the risk/benefit ratio	Compliance with the values of the medical profession, ensuring patient autonomy, compliance with the principles of medical ethics
Development and training	Data representativeness, algorithm validity, resistance to extreme cases	Statistical analysis, algorithmic audit, simulation of boundary cases	Developers, data scientists, ML engineers	Coverage of diverse clinical situations, inclusion of rare cases, consideration of comorbid conditions	Prevention of discrimination and bias based on gender, age, ethnicity, socio-economic status
Validation and verification	Accuracy, specificity, sensitivity, robustness, productivity	Cross-validation, boundary case testing, external validation	Independent experts, clinicians, regulators	Assessing the impact on clinical decisions and treatment outcomes, comparison with the gold standard of diagnosis	Transparency and explainability of results, the possibility of challenging, protection from automated discrimination
Implementation and integration	Protocol compatibility, impact on clinical decisions, easy use	Simulation of working processes, testing in real conditions, audit of clinical pathways	Medical organizations, IT specialists, and clinicians	Assessing the impact on care provision process, decision-making time, integration into existing clinical protocols	Impact on doctor-patient relations, level of trust, preservation of clinical autonomy of the doctor
Post-marketing monitoring	Undesirable phenomena, production drift, unforeseen consequences	Real-world data analytics, incident reporting system, regular audit	Manufacturers, regulators, medical professionals, patient communities	Monitoring of deviations of clinical outcomes from expected ones, long-term impact on the quality of care, detection of rare complications	Considerating patients' experience, psychosocial consequences of AI use, assessment of the impact on medical care availability

2. The unique risks of using artificial intelligence require a differentiated management approach where integration of ethical aspects is mandatory.
3. According to the comparative analysis, the author's concept complements and expands existing regulatory approaches, providing a multidimensional, continuous and ethically supported harm assessment.
4. The prospects for further work are associated with universal methodologies and standardization of risk assessment practices for creation and application of artificial intelligence in healthcare.

Thus, it is essential to develop and implement the predictable harm concept in development and implementation of AI-powered medical devices as it can ensure an optimal balance between the innovative potential of these technologies and patient safety. A comparative analysis with existing regulatory approaches shows the advantages of the proposed concept in terms of multidimensional risk assessment and integration of ethical principles at all AI life cycle stages. The matrix of predictable harm, which includes parameters of clinical consequences and ethical assessment, allows us to proceed from formal risk management procedures to a systematic approach that takes into account both technological and humanitarian aspects of using artificial intelligence in healthcare. Promising trends of further research in this area include development of standardized risk assessment methodologies for various categories of AI systems, creation of validated ethical audit tools for medical AI solutions, and formation of unified regulatory requirements that synthesize technological standards with the principles of medical ethics and focus on long-term social consequences of introducing intelligent technologies in healthcare.

References

1. Grand View Research. Artificial Intelligence in Healthcare Market Size Report, 2021–2028. Available from URL: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market (accessed: 01.05.2025).
2. Prikaz Ministerstva zdravookhraneniya Rossiyskoy Federatsii ot 7 iyulya 2020 g. № 686n «O vnesenii izmeneniy v prilozheniya № 1 i № 2 k prikazu Ministerstva zdravookhraneniya Rossiyskoy Federatsii ot 6 iyunya 2012 g. № 4n «Ob utverzhdenii nomenklaturnoy klassifikatsii meditsinskikh izdeliy» Ofitsial'nyy internet-portal zakona informatsii. Elektron. dan. Available from URL: https://www.pravo.gov.ru, svobodnyy. № publikatsii: 0001202008100015 (accessed: 10.05.2025). Russian.
3. Pis'mo Federal'noy sluzhby po nadzoru v sfere zdravookhraneniya ot 13 fevralya 2020 g. № 02I-297/20 «O programmnom obespechenii». [Tekst pis'ma opublikovan ne byl]. Russian.
4. «Software as a Medical Device»: Possible Framework for Risk Categorization and Corresponding Considerations. Available from URL: https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf (accessed: 10.05.2025).
5. Kodeks etiki v sfere II v meditsine i zdravookhranenii. Available from URL: https://ethics.a-ai.ru/ethics-of-medicine (accessed: 10.05.2025). Russian.
6. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Available from URL: https://www.who.int/publications/i/item/9789240029200 (accessed: 01.05.2025).
7. European Parliament. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002

and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.). Available from URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng (accessed: 01.05.2025).

8. European Commission. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act). Available from URL: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (accessed: 01.05.2025).

9. U. S. Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan. Available from URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (accessed: 01.05.2025).

Литература

1. Grand View Research. Artificial Intelligence in Healthcare Market Size Report, 2021–2028. Available from URL: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market (accessed: 01.05.2025).

2. Приказ Министерства здравоохранения Российской Федерации от 7 июля 2020 г. № 686н «О внесении изменений в приложения № 1 и № 2 к приказу Министерства здравоохранения Российской Федерации от 6 июня 2012 г. № 4н «Об утверждении номенклатурной классификации медицинских изделий» Официальный интернет-портал правовой информации. Электрон. дан. Режим доступа: [Электронный ресурс] URL: https://www.pravo.gov.ru, свободный. № опубликования: 0001202008100015 (дата обращения: 10.05.2025).

3. Письмо Федеральной службы по надзору в сфере здравоохранения от 13 февраля 2020 г. № 02И-297/20 «О программном обеспечении». [Текст письма опубликован не был].

4. «Software as a Medical Device»: Possible Framework for Risk Categorization and Corresponding Considerations. Available from URL: https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf (accessed: 10.05.2025).

5. Кодекс этики в сфере ИИ в медицине и здравоохранении. Режим доступа: [Электронный ресурс] URL: https://ethics.a-ai.ru/ethics-of-medicine (accessed: 10.05.2025).

6. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Available from URL: https://www.who.int/publications/i/item/9789240029200 (accessed: 01.05.2025).

7. European Parliament. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.). Available from URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng (accessed: 01.05.2025).

8. European Commission. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act). Available from URL: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (accessed: 01.05.2025).

9. U. S. Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan. Available from URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (accessed: 01.05.2025).