CLASSIFICATION OF RISKS OF USING ARTIFICIAL INTELLIGENCE SYSTEMS IN THE FIELD OF MENTAL HEALTH

Semenova NV ™, Martynyuk KL

Bekhterev National Medical Research Center of Psychiatry and Neurology, Saint Petersburg, Russia

The use of artificial intelligence systems (AIS) in healthcare is one of the promising solutions for improving access of citizens to modern medical technologies to prevent mental health disorders, stress-associated and psychosomatic diseases, to ensure early diagnosis and correction of mental disorders and their risk factors. However, an urgent problem is assessment of the risks of implementing AIS in the field of mental health and possibilities of managing such risks depending on various factors affecting clinical outcomes, as well as ethical aspects related to provision of this type of AIS-based medical care. The proposed risk classification system for the use of AIS in the field of mental health will expand the possibilities of introducing medical AIS into clinical practice while maintaining a high level of control over the risks of mental health disorders.

Keywords: mental health, psychological help, psychotherapy, artificial intelligence

Acknowledgements: the authors express their deep gratitude to Elena A. Volskaya, Candidate of Historical Sciences, Advisor to Director of N. A. Semashko National Research Institute of Public Health, Chairman of the Independent Interdisciplinary Committee for the Ethical Review of Clinical Trials, and Chairman of the Interuniversity Ethics Committee for supporting the idea of the article and advising on ethics issues.

Author contribution: Semenova NV — the idea of the article, problem statement, discussion of key substantive and ethical issues, planning and discussion of the article, editing the text of the article, its design; Martynyuk KL — participation in the discussion of the problem, study of world experience and literature on the topic, systematization and generalization of data, participation in discussion of the results, and writing the main text of the article.

Compliance with ethical standards: the ethics committee meeting was not held as the material for publication included theoretical provisions without involvement of patients; the article was prepared in compliance with ethical standards.

Correspondence should be addressed: Natalia V. Semenova

Bekhtereva St., 3, St. Petersburg, 191019, Russia; org@bekhterev.ru

Received: 22.04.2025 Accepted: 20.05.2025 Published online: 08.06.2025

DOI: 10.24075/medet.2025.005

КЛАССИФИКАЦИЯ РИСКОВ ПРИМЕНЕНИЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В СФЕРЕ ПСИХИЧЕСКОГО ЗДОРОВЬЯ

Н. В. Семенова [™], К. Л. Мартынюк

Национальный медицинский исследовательский центр психиатрии и неврологии имени В. М. Бехтерева, Санкт-Петербург, Россия

Применение систем на основе искусственного интеллекта (СИИ) в здравоохранении является одним из перспективных решений для улучшения доступа граждан к современным медицинским технологиям профилактики нарушений психического здоровья, стресс-ассоциированных и психосоматических заболеваний, ранней диагностики и коррекции психических нарушений и факторов риска их развития. Однако актуальной проблемой является оценка рисков внедрения СИИ в сфере психического здоровья и возможности управления такими рисками в зависимости от различных факторов, влияющих на клинические исходы, а также этические аспекты, связанные с оказанием этого вида медицинской помощи с применением СИИ. Предложенная система классификации рисков применения СИИ в сфере психического здоровья позволит расширить возможности внедрения медицинских СИИ в клиническую практику при сохранении высокого уровня контроля рисков нарушения психического здоровья.

Ключевые слова: психическое здоровье, психологическая помощь, психотерапия, искусственный интеллект

Благодарности: авторы выражают глубокую признательность кандидату исторических наук, советнику директора ФГБНУ «Национальный НИИ общественного здоровья имени Н. А. Семашко», председателю Независимого междисциплинарного комитета по этической экспертизе клинических исследований, председателю Межвузовского комитета по этике Елене Алексеевне Вольской за поддержку идеи статьи и консультации по вопросам этики.

Вклад авторов: Н. В. Семенова — идея статьи, постановка проблемы, обсуждение ключевых содержательных и этических вопросов, планирование и обсуждение статьи, редактирование текста статьи, ее оформление; К. Л. Мартынюк — участие в обсуждении проблемы, изучение мирового опыта и литературы по теме, систематизация и обобщение данных, участие в обсуждении результатов, написание основного текста статьи.

Соблюдение этических стандартов: заседание этического комитета не проводилось, так как материалом для публикации послужили теоретические положения без участия пациентов; подготовка статьи проводилась с соблюдением этических норм.

Для корреспонденции: Наталия Владимировна Семенова

ул. Бехтерева, д. 3, г. Санкт-Петербург, 191019, Россия; org@bekhterev.ru

Статья поступила: 22.04.2025 Статья принята к печати: 20.05.2025 Опубликована онлайн: 08.06.2025

DOI: 10.24075/medet.2025.005

High demand for medical, psychological and psychotherapeutic care among citizens, especially those living in metropolitan cities [1], which is accompanied by insufficient appealability of these types of medical aid in public health institutions due to persistent public stigma, contributes to stable inequality in access to

advanced medical technologies for prevention of mental health disorders, stress-associated and psychosomatic diseases, and non-pharmacological intervention in mental disorders.

Late treatment leads to late care for a patient, usually with more pronounced disorders, increasing the requirements for

volume, complexity and cost of interventions with less potential for functioning restoration and quality of prognosis. It ultimately results in professional burnout of qualified personnel with increased disability and decreased productivity, deterioration of health and a general decrease in life quality of employees. Meanwhile, mental disorders (MD) have been the leading factor in the global burden of diseases for more than 20 years regarding the number of years of life with disability [2, 3] and maintaining a high burden on social funds.

The use of artificial intelligence systems (AIS) is one of the promising solutions for improving access of citizens to modern medical technologies for prevention of mental health disorders, stress-associated and psychosomatic diseases, early diagnosis and correction of mental disorders and risk factors for their development, as well as expanding the possibilities of psychotherapeutic interventions in case of mental disorders when medical, psychological and psychotherapeutic types of care are limitedly available due to the continuing shortage of appropriate specialized personnel. The first experience of its application is being actively discussed in scientific literature [4–12].

In recent years, researchers were optimistic about Al capabilities, as Al models are increasingly able to simulate a real interaction with humans. This can contribute not only to treatment delivery, but also to treatment adherence as compared to other forms of e-health. For example, conversational Al-based apps have proven their effectiveness in reducing symptoms of depression and anxiety, preventing stress, general distress, and negative affects, and improving well-being [4, 8, 13]. Meanwhile, an uncontrolled growth in the number of Al applications and difficulties in tracking the empirical results of using AlS are considered as significant negative factors [10–12].

The most discussed areas of AI application in the field of mental health include as follows:

- Supporting decisions taken during diagnosis of mental disorders and choosing the treatment strategy;
- The function of a "supervisor" and a "medical assistant" increasing commitment of people to preventive, diagnostic and therapeutic measures;
- Dynamic monitoring of the condition of patients suffering from mental disorders;
- Predicting the risk of mental disorder exacerbation;
- Non-drug control of symptoms of mental disorder using an "Al therapist";
- Correction of emotional problems using AI models and methods of psychotherapy and psychocorrection;
- Prevention of emotional disorders, involving the use of AI, aimed at development of emotional intelligence and stress tolerance.

Taking into consideration high vulnerability of the citizens when they turn to specialists in mental health, limited practical use of technical devices in the clinical process by mental health professionals, and the emerging regulation of access of Al-based medical devices to circulation (according to Decree of the Government of the Russian Federation dated December 27, 2012 No. 1416 (as amended on November 24, 2020), and the fact that the Al-based MD belong to the 3rd (maximum) risk class) [14], it seems relevant to consider the risks of implementing AlS in the field of mental health and possibilities of managing such risks depending on various factors affecting clinical outcomes, as well as ethical aspects related to provision of this type of medical care using AlS.

The experience of medical use of AIS described in publications [2–18] helps us identify common sources of risks:

those related to the technological features of development and operation of Al models and systems based on them; specific for use in the relevant clinical field; ethical and social aspects covering the basic principles of civil rights and freedom, fairness, confidentiality, security and transparency [15-18]. Meanwhile, the latter go through the entire MD life cycle: "The manufacturer must establish, document and maintain a continuous process of MD-associated identifying hazards, identifying and evaluating associated risks, managing these risks and monitoring such management throughout the MD life cycle (starting from design, including scientific research, and to decommissioning according to A.2.1 Scope of application) in accordance with the requirements of GOST ISO 14971" [2]. Manufacturers of medical AIS have even stricter obligations if they are guided by the Code of Ethics of Artificial Intelligence in National Healthcare [15]: "AIS developers must adhere to the ethical obligations and values followed by medical personnel in their actions towards patients in clinical practice, including the Code of Professional Ethics of a Doctor of the Russian Federation." (Article 3 of the draft Code), which seems justified taking into account the specifics of AIS development and the inherent limitations of the ability to control individual risks during the operation of such systems in real clinical practice, especially when it is independently used by the patient.

Generalized sources of AI risks are given in PNS 840-2023 "Artificial Intelligence. An overview of ethical and social aspects" [17]. They include as follows:

- unauthorized means or methods of collecting, processing, or disclosing personal data;
- obtaining and using biased, inaccurate or unrepresentative data for AIS training;
- non-transparent machine learning (ML) decision-making or insufficient documentation, commonly referred to as lack of explainability;
- lack of tracking capability (iterative inaccuracy of AI models, working in an open contextual environment with unforeseen events and conditions);
- insufficient understanding of technology social impact after its introduction.

Specific sources of risks of using AIS in the field of mental health are as follows:

- Limited application of non-harm principle in case of incomplete fault tolerance, insufficient accuracy of operation or effectiveness of models for real clinical practice, or unpredictability of AIS when working under borderline conditions;
- Lack of transparency regarding the nature of the services provided by AIS and its representation as an assistant that uses therapeutic methods;
- Lack of control over the reviews and recommendations that users receive from offline AIS based on generative models;
- Limited understanding of clinical process components implemented by AIS, which form the basis for psychotherapeutic intervention effectiveness;
- AIS influence on autonomy and free will: formation of AIS-dependent response/behavior of patients due to over-accessibility, developing attachment (anthropomorphization) or excessive trust in AIS and loss of contact quality with a specialist, or development of anxiety, stress, and hypochondria due to AIS constant and frequent use;
- Increased resilience of the stigma associated with mental health disorders due to encouraging users to

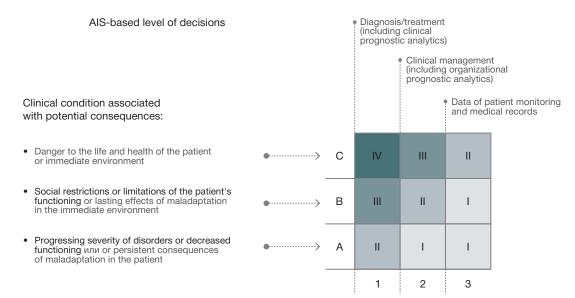


Fig. Classification of risks of using AIS in the field of mental health

use AIS or in the face of actual lack of alternatives, and expanding the possibilities of risky self-treatment when the system response is interpreted in an incorrect way.

Thus, to model the risks of using AIS specific to the field of mental health, it is necessary to determine the semantic space in the following areas: the goals of using AIS, continuum of relevant clinical conditions and outcomes, safety of software packages, and ethical certainty.

Currently, when technologies in the field of mental health are being developed, AIS are applicable for the following purposes (based on GOST R 59525-2021 Health informatics. Intelligent methods of medical data processing. Main provisions" [19–20]):

- to diagnose, prevent, observe, treat or relieve disorders;
- to support the vital activity (if the patient's autonomy is included into the term as well);
- functioning of medical decision support systems;
- predicting the appearance and/or development of diseases based on genetic data.

Meanwhile, the last point in the field of mental health can also be interpreted based on bio-psycho-social factors. Thus, the main patterns of reaction/behavior, as a rule, are unknowingly borrowed by the child from his family and form a characteristic pattern, which further becomes a part of the child's personality, and can no less affect resistance to stress and likelihood of developing stress-related diseases.

Consequently, the use of AIS is justified in accordance with the principles of evidence-based medicine. It is limited to conditions ranging from subclinical disorders recognized by diagnostic methods validated for the Russian population to threats to life and health of the patient or people around the patient (the latter is relevant for a number of chronic and long-term mental disorders with severe persistent or frequently aggravated painful manifestations).

Taking into account the principle of classification of the software (including AIS) used while providing medical care, in terms of safety, a class is assigned according to the risk of harm to the patient, user or other persons, based on a dangerous situation to which the program system (PS) can contribute in the worst-case scenario [2]:

 Class A: PS can contribute to a dangerous situation that does not lead to an unacceptable risk;

- Class B: PS can contribute to a dangerous situation that leads to an unacceptable risk, and the resulting possible harm is not a serious injury;
- Class C: PS can contribute to a dangerous situation that leads to an unacceptable risk, and as a result, possible harm can include death or serious injury.

It should be clarified that in case of mental health disorders, the concept of an unacceptable risk that does not result in a serious injury or death is most consistent with consequences in the form of social restrictions or limited functioning of patients, as well as the persistent consequences of maladaptation associated with close interaction with such a patient in the immediate environment (cohabiting relatives, especially minors in the process of personality formation, and persons who have been informal caregivers for a long time). The boundary condition for gradation within the acceptable risk range should include progression of the severity of disorders or stable consequences of maladaptation in the patient, since management of such risks is fully available within the framework of a high-quality clinical process (adherence to treatment, therapeutic alliance, compliance with clinical recommendations, scientifically based innovative methods of intervention).

To expand the possibilities of AI implementation in supporting and auxiliary processes, it is advisable to take into account the importance of information for making medical decisions, which is processed by AIS, and to varying degrees may affect the level of clinical risks [18] in the range of:

- data (including AIS interpreted data) for diagnosis or treatment (including clinical predictive analytics);
- data for clinical management (including organizational predictive analytics);
- data of patient monitoring and medical records (including those entered into AIS by the patient or caregiver).

It is acceptable to reduce the risk for each step of the above gradation of information significance, as this corresponds to a reduction in the impact of data on clinical decisions, due to the possibility of maximum control of individual risks within the framework of a qualitative clinical process.

Thus, it is possible to propose the following gradation of the risk of using AIS in the field of mental health n the form of a 2-dimensional matrix (Fig.):

 A) regarding safety (clinical condition associated with potential consequences):

- Risk category IV danger to the life and health of the patient or immediate environment;
- Risk category III social restrictions or limitations of the patient's functioning or lasting effects of maladaptation in the immediate environment;
- Risk category II is the progressing severity of disorders or persistent consequences of maladaptation in the patient;
- B) regarding the process significance (the level of decisions based on the AIS provided information):
- Risk category N diagnosis or treatment (including clinical predictive analytics);
- Risk category N-1 clinical management (including organizational predictive analytics);
- Risk category N-2 ≥ 1 data of patient monitoring and medical records.

The proposed risk classification system for using AIS in the field of mental health will expand the possibilities of introducing medical AIS into clinical practice while maintaining a high level of control over both individual and public (micro- and macrosocial) risks of mental health disorders and development of stress-associated diseases, as well as increasing availability of qualified care and, accordingly, earlier interventions for mental disorders with a high potential for restoring functioning and maintaining the quality of life of patients. At the same time, it provides the possibility of high-quality ethical risk management for the specialized use of AIS throughout the entire life cycle (from design, including scientific research, post-registration monitoring of quality/effectiveness and recalibration of Al models, to decommissioning) of an Al-based medical device at all levels of regulatory and industry requirements.

References

- Hołownia-Voloskova M, Tarbastaev A, Golicki D. Population norms of health-related quality of life in Moscow, Russia: the EQ-5D-5L-based survey. Quality of Life Research. 2021; 30(3): 831–840.
- Global Burden of Disease Study 2019 (GBD 2019) Available from URL: https://ghdx.healthdata.org/gbd-2019 (accessed: 11.03.2025).
- Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al. Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019. The Lancet. 2020; 396.
- Espejo G, Reiner W, Wenzinger M. Exploring the Role of Artificial Intelligence in Mental Healthcare: Progress, Pitfalls, and Promises. Cureus. 2023; 15(9): e44748.
- Groot J, MacLellan A, Butler M, Todor E, Zulfiqar M, Thackrah T, Clarke C, Brosnan M, Ainsworth B. The Effectiveness of Fully Automated Digital Interventions in Promoting Mental Well-Being in the General Population: Systematic Review and Meta-Analysis. JMIR Ment Health. 2023; 10: e44658.
- Grote T, Berens P. A paradigm shift? On the ethics of medical large language models. Bioethics. 2024; 38: 383–390.
- 7. Jain G, Pareek S, Carlbring P. Revealing the source: How awareness alters perceptions of Al and human-generated mental health responses. Internet Interv. 2024; 36: 100745.
- 8. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, Paulus MP, Krystal JH, Jeste DV. Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging. 2021: 6(9).
- Rahsepar Meadi M, Bernstein JS, Batelaan N, van Balkom AJL, Metselaar S. Does a lack of emotions make chatbots unfit to be psychotherapists? Bioethics. 2024; 1–8.
- Rogan J, Bucci S, Firth J. Health Care Professionals' Views on the Use of Passive Sensing, Al, and Machine Learning in Mental Health Care: Systematic Review With Meta-Synthesis. JMIR Ment Health. 2024; 11: e49577.
- Tornero-Costa R, Martinez-Millana A, Azzopardi-Muscat N, Lazeri L, Traver V, Novillo-Ortiz D. Methodological and Quality Flaws

- in the Use of Artificial Intelligence in Mental Health Research: Systematic Review. JMIR Ment Health. 2023; 10: e42045.
- 12. Vandersluis R, Savulescu J. The selective deployment of Al in healthcare. Bioethics. 2024; 38: 391–400.
- He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, Hou X. Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. J Med Internet Res. 2023; 2: e43862.
- 14. Postanovlenie Pravitel'stva RF ot 27.12.2012 № 1416 (s izmenenijami na 24.11.2020) «Ob utverzhdenii Pravil gosudarstvennoj registracii medicinskih izdelij». Available from URL: http://government.ru/docs/all/85645/ (accessed: 12.03.2025). Russian.
- Al'jans v sfere iskusstvennogo intellekta. Kodeks jetiki v sfere iskusstvennogo intellekta. Available from URL: https://ethics.a-ai. ru/ (accessed: 11.03.2025). Russian.
- Nacional'nyj centr razvitija iskusstvennogo intellekta pri Pravitel'stve Rossijskoj Federacii. Informacionno-analiticheskaja spravka «Zarubezhnaja praktika ustanovlenija trebovanij k bezopasnosti II» (2023). Available from URL: https://ai.gov.ru/ (accessed: 10.03.2025). Russian.
- PNST 840-2023. Iskusstvennyj intellekt. Obzor jeticheskih i obshhestvennyh aspektov. Available from URL: https://nd.gostinfo. ru/document/7475760.aspx (accessed: 11.03.2025). Russian.
- 18. International Medical Device Regulators Forum. Software as a Medical Device: 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: 10.03.2025).
- GOST ISO 14971-2021. Izdelija medicinskie. Primenenie menedzhmenta riska k medicinskim izdelijam. Available from URL: https://nd.gostinfo.ru/document/5307378.aspx (accessed: 11.03.2025). Russian.
- GOST R 59525-2021. Informatizacija zdorov'ja. Intellektual'nye metody obrabotki medicinskih dannyh. Osnovnye polozhenija. Available from URL: https://nd.gostinfo.ru/document/6856334. aspx (accessed: 11.03.2025). Russian.

Литература

- Hołownia-Voloskova M, Tarbastaev A, Golicki D. Population norms of health-related quality of life in Moscow, Russia: the EQ-5D-5L-based survey. Quality of Life Research. 2021; 30 (3): 831–840.
- Global Burden of Disease Study 2019 (GBD 2019) Available from URL: https://ghdx.healthdata.org/gbd-2019 (accessed: 11.03.2025).
- Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al. Global Burden of 369 Diseases and Injuries in 204 Countries
- and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019. The Lancet. 2020; 396.
- Espejo G, Reiner W, Wenzinger M. Exploring the Role of Artificial Intelligence in Mental Healthcare: Progress, Pitfalls, and Promises. Cureus. 2023; 15(9): e44748.
- Groot J, MacLellan A, Butler M, Todor E, Zulfiqar M, Thackrah T, Clarke C, Brosnan M, Ainsworth B. The Effectiveness of Fully Automated Digital Interventions in Promoting Mental Well-Being

ORIGINAL RESEARCH

- in the General Population: Systematic Review and Meta-Analysis. JMIR Ment Health. 2023; 10: e44658.
- Grote T, Berens P. A paradigm shift? On the ethics of medical large language models. Bioethics. 2024; 38: 383–390.
- 7. Jain G, Pareek S, Carlbring P. Revealing the source: How awareness alters perceptions of Al and human-generated mental health responses. Internet Interv. 2024; 36:100745.
- 8. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, Paulus MP, Krystal JH, Jeste DV. Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging. 2021; 6(9).
- Rahsepar Meadi M, Bernstein JS, Batelaan N, van Balkom AJL, Metselaar S. Does a lack of emotions make chatbots unfit to be psychotherapists? Bioethics. 2024; 1–8.
- Rogan J, Bucci S, Firth J. Health Care Professionals' Views on the Use of Passive Sensing, AI, and Machine Learning in Mental Health Care: Systematic Review With Meta-Synthesis. JMIR Ment Health. 2024; 11: e49577.
- Tornero-Costa R, Martinez-Millana A, Azzopardi-Muscat N, Lazeri L, Traver V, Novillo-Ortiz D. Methodological and Quality Flaws in the Use of Artificial Intelligence in Mental Health Research: Systematic Review. JMIR Ment Health. 2023; 10: e42045.
- 12. Vandersluis R, Savulescu J. The selective deployment of Al in healthcare. Bioethics. 2024; 38: 391–400.
- He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, Hou X. Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. J Med Internet Res. 2023; 25: e43862.
- 14. Постановление Правительства РФ от 27.12.2012 № 1416 (с изменениями на 24.11.2020) «Об утверждении Правил государственной регистрации медицинских изделий». Режим

- доступа: [Электронный ресурс] URL: http://government.ru/docs/all/85645/ (дата обращения: 11.03.2025).
- Альянс в сфере искусственного интеллекта. Кодекс этики в сфере искусственного интеллекта. Режим доступа: [Электронный ресурс] URL: https://ethics.a-ai.ru/ (дата обращения: 11.03.2025).
- 16. Национальный центр развития искусственного интеллекта при Правительстве Российской Федерации. Информационно-аналитическая справка «Зарубежная практика установления требований к безопасности ИИ» (2023). Режим доступа: [Электронный ресурс] URL: https://ai.gov.ru/ (дата обращения: 11.03.2025).
- ПНСТ 840-2023. Искусственный интеллект. Обзор этических и общественных аспектов. Режим доступа: [Электронный ресурс] URL: https://nd.gostinfo.ru/document/7475760.aspx (дата обращения: 11.03.2025).
- International Medical Device Regulators Forum. Software as a Medical Device: Framework for Risk Categorization and Corresponding Considerations. Режим доступа: [Электронный ресурс] URL: https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/ imdrf-tech-140918-samd-framework-risk-categorization-141013. pdf (дата обращения: 11.03.2025).
- ГОСТ ISO 14971-2021. Изделия медицинские. Применение менеджмента риска к медицинским изделиям. Режим доступа: [Электронный ресурс] URL: https://nd.gostinfo.ru/ document/5307378.aspx (дата обращения: 11.03.2025).
- ГОСТ Р 59525-2021. Информатизация здоровья. Интеллектуальные методы обработки медицинских данных. Основные положения. Режим доступа: [Электронный ресурс] URL: https://nd.gostinfo.ru/document/6856334.aspx (дата обращения: 11.03.2025).