УДК 615.19:004.89
https://doi.org/10.30895/1991-2919-2025-15-6-630-641
Обзор | Review

М.А. Ярошинский[1], М.В. Андреева[1], Е.И. Балакин[1] ✉, А.Ю. Савченко[2], А.С. Павлов[4], В.И. Пустовойт[3]

Нейросетевые технологии в регистрации лекарственных средств: методология машинного анализа документов и интеллектуальных систем реального времени

[1] *Акционерное общество «Фарм-Синтез», Верейская ул., д. 29, стр. 134, Москва, 121357, Российская Федерация*

[2] *Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский ядерный университет «МИФИ» (НИЯУ МИФИ), Каширское ш., д. 31, Москва, 115409, Российская Федерация*

[3] *Федеральное государственное бюджетное учреждение «Государственный научный центр Российской Федерации — Федеральный медицинский биофизический центр имени А.И. Бурназяна» Федерального медико-биологического агентства, ул. Живописная, д. 46, корп. 8, Москва, 123098, Российская Федерация*

[4] *Федеральное государственное бюджетное учреждение высшего образования «Российский химико-технологический университет им. Д.И. Менделеева», Миусская пл., д. 9, Москва, 125047, Российская Федерация*

✉ ***Балакин Евгений Игоревич***; *ebalakin@pharm-sintez.ru*

РЕЗЮМЕ

ВВЕДЕНИЕ. Методы подготовки документов, используемые в ходе разработки лекарственных средств, характеризуются высокими временными затратами (40–60% рабочего времени специалистов), высокой частотой ошибок, вносимых в документацию, и ограниченной интероперабельностью данных. Повышение эффективности подготовки документов возможно при использовании нейросетевых технологий и переходе к комплексной автоматизации процедур жизненного цикла регистрационного досье.

ЦЕЛЬ. Оценка возможности использования систем искусственного интеллекта (ИИ) и машинного анализа при подготовке регистрационного досье лекарственного препарата в процессе разработки лекарственного средства.

ОБСУЖДЕНИЕ. Модели обработки естественного языка (NLP) показывают высокую эффективность в области обработки технической и регуляторной документации. Системы распознавания именованных сущностей (NER) с точностью извлечения 89–96% позволяют сократить время обработки (подготовки и последующей проверки) производителем материалов при формировании электронного общего технического документа на 64%, однако возможность обработки информации ограничена трудностями интерпретации морфологически сложных терминов и требует использования наборов аннотированных данных. Следует отметить, что генеративные модели типа GPT-4 без дополнительной настройки при использовании в архитектуре генерации, дополненной поиском (RAG), могут формировать фактологически некорректную информацию. Предиктивные системы на основе графовых нейросетей и ансамблей XGBoost демонстрируют высокую точность (ROC AUC до 0,88) при прогнозировании одобрения препаратов, но имеют недостаток в виде невозможности интерпретации решений и систематических смещений в данных. Разработка документоцентричных платформ с NLP позволяет сократить время подготовки досье на 60%, однако при внедрении

автоматизированной процедуры формирования разделов досье требуется экспертная верификация.
ВЫВОДЫ. Концепция интегрированных ИИ-систем подтверждает свою эффективность, сокращая сроки обработки документов производителем и повышая точность решений, что способствует ускорению вывода препаратов на рынок. Перспективы внедрения цифровых технологий связаны с преодолением различий в терминах через унифицированные онтологии. Для практической реализации требуется разработка единых стандартов валидации ИИ-алгоритмов и адаптивных систем.

Ключевые слова: регистрационное досье; искусственный интеллект; обработка естественного языка; предиктивное моделирование; документоцентричные платформы; BioBERT; стандарты ALCOA+; IDAAPM-база; валидация алгоритмов; эОТД-досье; общий технический документ

Для цитирования: Ярошинский М.А., Андреева М.В., Балакин Е.И., Савченко А.Ю., Павлов А.С., Пустовойт В.И. Нейросетевые технологии в регистрации лекарственных средств: методология машинного анализа документов и интеллектуальных систем реального времени. *Регуляторные исследования и экспертиза лекарственных средств.* 2025;15(6):630–641. https://doi.org/10.30895/1991-2919-2025-15-6-630-641

Финансирование. Работа выполнена без спонсорской поддержки.

Потенциальный конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Milan A. Yaroshinsky[1], **Maria V. Andreeva**[1], **Evgenii I. Balakin**[1] ✉, **Alla Yu. Savchenko**[2], **Alexander S. Pavlov**[4], **Vasily I. Pustovoit**[3]

Neural Network Technologies in Drug Registration: Computerised Analysis of Documents and Real-Time Systems

[1] *Pharm-Sintez AO,*
29/134 Vereyskaya St., Moscow 121357, Russian Federation

[2] *National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),*
31 Kashirskoe Hwy, Moscow 115409, Russian Federation

[3] *State Research Center — Burnasyan Federal Medical Biophysical Center of Federal Medical Biological Agency,*
46 Zhivopisnaya St., Moscow 123098, Russian Federation

[4] *Dmitry Mendeleev University of Chemical Technology of Russia,*
9 Miusskaya Sq., Moscow 125047, Russian Federation

✉ ***Evgenii I. Balakin****; ebalakin@pharm-sintez.ru*

ABSTRACT

INTRODUCTION. Current methods of handling medicine regulatory documents are associated with high time cost (40–60% of labour hours), frequent documentation errors, and limited data interoperability. Neural network technologies have enabled the enhanced document preparation and a transition to full automation of the registration dossier life cycle.
AIM. This study aimed to evaluate the possibility of using artificial intelligence (AI) systems in preparing a drug registration dossier.
DISCUSSION. Natural language processing (NLP) models demonstrate high efficiency for the regulatory documentation. Named entity recognition (NER) systems with 89–96% entity extraction accuracy rate reduces the processing (preparation and quality review) time for documents within electronic Common Technical Document (eCTD) by 64%, but face limitations in interpreting morphologically complex terms and require annotated datasets. Without additional fine-tuning, generative models such as GPT-4, are prone to generating inaccurate facts when used in the Retrieval-Augmented Generation (RAG) architecture. Predictive systems based on graph neural networks and XGBoost ensembles demonstrate high accuracy (ROC AUC up to 0.88) when predicting drug approval; however, they cannot interpret

decisions and data systematic biases. Developing document-centric platforms with NLP reduces the dossier preparation time by 60%, still, implementing an automated procedure for generating dossier sections requires an expert verification.
CONCLUSIONS. The concept of integrated AI systems proves its effectiveness by reducing the document handling time by manufacturers and increasing the accuracy of decisions, which in turn speeds up the market launch of medicinal products. The prospects of introducing digital technologies are associated with overcoming definition differences through unified ontologies. Practical implementation requires the development of unified standards for the validation of AI algorithms and adaptive systems.

Keywords: registration dossier; artificial intelligence; natural language processing; predictive modelling; document-centric platforms; BioBERT; ALCOA+ standards; IDAAPM database; algorithm validation; eCTD dossier

For citation: Yaroshinsky M.A., Andreeva M.V., Balakin E.I., Savchenko A.Yu., Pavlov A.S., Pustovoit V.I. Neural network technologies in drug registration: Computerised analysis of documents and real-time systems. *Regulatory Research and Medicine Evaluation.* 2025;15(6):630–641. https://doi.org/10.30895/1991-2919-2025-15-6-630-641

Funding. The study was performed without external funding.

Disclosure. The authors declare no conflict of interest.

ВВЕДЕНИЕ

Использование искусственного интеллекта (ИИ) регуляторными медицинскими органами и производителями лекарственных средств (ЛС) растет высокими темпами, в том числе увеличивается количество одобренных регуляторными органами инструментов для использования в процессе разработки ЛС, подготовки и анализа регистрационного досье [1]. Ключевыми причинами являются сокращение времени экспертизы ЛС и усиление борьбы с фальсифицированными препаратами [1–3]. Так, опыт Национального агентства по надзору за здравоохранением Бразилии (Brazilian Health Regulatory Agency, Anvisa) показывает, что применение алгоритмов ИИ сократило сроки регистрации на 45–60% [1]. Помимо операционной эффективности, внедрение процедур использования ИИ обусловлено введением в обращение биотехнологических лекарственных средств (биоаналоги, генотерапевтические препараты), что требует работы с большим массивом данных и, как следствие, большого количества времени. Процесс внедрения инструментов на основе ИИ поддерживается появлением технологий машинного и глубокого обучения, роботизированной автоматизации процессов (Robotic Process Automation, RPA) и обработки естественного языка. Исследования подтверждают, что такая интеграция значительно ускоряет некоторые процедуры в процессе жизненного цикла лекарственного препарата, в том числе сокращая время клинических испытаний, анализа и составления документации на лекарственный препарат [6–9].

Интеграция ИИ также трансформирует регуляторные практики, чему способствуют объективные потребности в оптимизации ресурсов, повышении точности принимаемых решений и оперативном предоставлении новейших препаратов пациентам. Широко используются инструменты прогностической аналитики для работы с клиническими данными, автоматизированные системы внутренней проверки документов и алгоритмы мониторинга безопасности. Проверка становится эффективнее: применение ИИ позволяет сократить время анализа регистрационного досье на 20–30% и усиливает инструменты выявления риска несоответствия документов нормативным требованиям и реальным технологическим процессам производства лекарственных средств [10–12]. На ручную обработку документов, главным образом поиск и проверку данных, приходится 40–60% времени, что серьезно замедляет процесс. Ошибки, вносимые в досье электронного общего технического документа из-за человеческого фактора (15–30% случаев), ведут к дополнительным запросам экспертов регуляторного органа и затратам времени. Например, некорректное заполнение фармакокинетических разделов или данных об эффективности, с учетом приостановки экспертизы на срок ответа и необходимости внесения изменений в документы регистрационного досье, могут задержать регистрацию препарата до 6 мес. Системные ошибки при вводе данных в модулях качества и несвоевременное обновление ссылок на нормативную документацию повышают риски отказа в регистрации, подчеркивая преимущества использования цифровых инструментов. До 25% ситуаций с затягиванием момента вывода лекарств в обращение связаны с исправлением ошибок ручной подготовки документации [4, 6, 9, 11]. Длительность задержек растет из-за необходимости повторной оценки

при подачах в несколько стран, ежегодно снижая глобальную операционную эффективность [1, 5, 10]. Наличие единой информационной среды позволит найти различия как в регуляторных требованиях между странами, так и указать документы регистрационного досье, требующие корректировок.

Статический формат хранения данных для документоориентированной модели предполагает менее эффективную (отношение объема проверенных документов к затраченным ресурсам) систему подготовки и проверки документов. К примеру, бумажный архив производственной и регистрационной документации предполагает, что для оперативного внесения изменений должна присутствовать контролируемая электронная версия документа в редактируемом формате, либо измененный документ должен каждый раз создаваться на основании действующей версии. При этом наличие валидированной системы электронного документооборота не обеспечивает должной связанности ввиду отсутствия семантических конструкций и обновляющихся блоков внутри самого документа. Перечисленное выше ограничивает возможность каскадного обновления документов регистрационного досье, повторное использование информации и интеграцию новых источников [3, 7, 8, 12].

Растущие объемы данных доклинических и клинических исследований и усложнение регуляторных норм затрудняют верификацию безопасности, эффективности и качества лекарственных средств. Одной из задач при внедрении нейросетевых технологий является валидация процесса автоматизированного аудита и прогностической аналитики систем ИИ и машинного обучения, где терминологические и методологические различия в регуляторных подходах становятся барьером.

Основные технологические решения представлены системами обработки естественного языка (NLP-модели), в том числе терминов, касающихся биомедицинских моделей, для анализа регуляторных документов, прогностическими системами оценки эффективности и безопасности, а также платформами фармаконадзора для выявления нежелательных явлений в реальном времени после регистрации препарата [1, 4, 5].

Цель работы — оценка возможности использования интеллектуальных систем при подготовке регистрационного досье лекарственного препарата в процессе разработки лекарственного средства.

Задачи:

- сравнительный анализ эффективности NLP-моделей (BioBERT, SpaCy, GPT-4+RAG) в отношении обработки регуляторной документации;
- оценка возможности предиктивных моделей (графовые нейросети, ансамбли XGBoost) прогнозировать исход регуляторной экспертизы;
- систематизация рисков применения генеративного ИИ в RAG-архитектурах и подходов к их минимизации;
- обобщение опыта и возможности внедрения документоцентричных платформ и систем реального времени в регуляторные процессы.

Методы исследования. Проведен обзор научной литературы за период 2016–2025 гг. в библиографических базах данных (PubMed, IEEE Xplore) и базах данных регуляторных органов: Управление по санитарному надзору за качеством пищевых продуктов и медикаментов (Food and Drug Administration, FDA) и Европейское агентство лекарственных средств (European Medicines Agency, EMA). Для поиска использовали также поисковую систему (Google Scholar). Поиск осуществлялся с использованием следующих запросов:

- (“Artificial Intelligence” [Mesh]) AND (“Drug Approval” [Mesh]);
- allintitle: (“regulatory affairs” OR “drug registration”) AND (“AI” OR “NLP” OR “computer vision”).

В анализ включали работы, содержащие количественные метрики результатов использования нейросетевых технологий и исследования с валидацией в реальных условиях регуляторных агентств или фармкомпаний.

Критерии исключения. Обзоры, комментарии, тезисы конференций и публикации без полного текста. Исследования, не описывающие параметры или архитектуру цифровых систем. Работы на языках, отличных от английского и русского (если не удалось получить перевод).

Процедура отбора исследований. Первоначально был проведен отбор по заголовкам и аннотациям, после чего полный текст был проанализирован для окончательного включения. Каждое исследование оценивалось двумя авторами на соответствие критериям включения и исключения. Для оценки достоверности данных, извлеченных из отобранных исследований, использовались метрики валидации, приведенные в оригинальных публикациях (accuracy (точность), ROC AUC (кривая ошибок), F1-оценка

(F1-score, метрика оценки эффективности классификации)). Сравнительный анализ технологий проводился на основе качественного синтеза заявленных в источниках результатов. Из 142 идентифицированных публикаций отобрано 28 статей, соответствующих критериям.

ОСНОВНАЯ ЧАСТЬ

Модели ИИ-ассистированной экспертизы

Модели NLP. При обработке текстовых документов в фармацевтике выбор архитектуры NLP-систем требует решения проблем морфологической сложности медицинских терминов, включая синонимию, традицию активного использования аббревиатур и вариативность формулировок замечаний экспертов. В работе S. Viswanath и соавт. [13] предложен алгоритм Calibrated Quantum Mesh (CQM, метод квантово-инспирированной обработки естественного языка для снижения неоднозначности терминов), который использует ансамблевый подход для обработки естественного языка без обременительного аннотирования, достигая точности извлечения сущностей 89% на наборе из 65 тыс. документов. Этот метод позволяет достичь однозначности трактовки терминологии за счет контекстуализации слов через «квантовые состояния» и сетевые корреляции, что критично для анализа химических названий и клинических описаний. Экспериментально подтверждены трудности оптического распознавания символов (OCR-распознавание) специализированных обозначений (например, названий химических соединений) и необходимость использования онтологий для уточнения реакций растворителей с компонентами лекарственного препарата при формировании запроса к нейросети.

В работе [14] показано, что тонкая настройка моделей, таких как BioBERT, на аннотированных данных регуляторных документов позволяет достичь высокой точности извлечения сущностей (субъекты, объекты, действия). Точность достигает 90–96% при F1-оценке 0,91 после оптимизации иерархии онтологических признаков и применения семантических правил, что вполне может быть использовано в фармацевтической отрасли для документов электронного общего технического документа. Обученные на клинических данных BERT-модели (ClinicalBERT) наиболее адаптированы к решению задач распознавания смысловой эквивалентности вопросов благодаря адаптации к морфологическим особенностям медицинских текстов — синонимам и аббревиатурам [14].

Проблемы морфологической сложности, включая вариативность формулировок замечаний экспертов регуляторных органов, требуют рекурсивной верификации связанных сущностей между группой документов через онтологические графы. Технически это реализуется вычислением сходства между регуляторными утверждениями и процессами на основе трехуровневой метрики: сходство тем, ядерных и вспомогательных сущностей [15]. В свою очередь, генеративные модели типа GPT-4 при использовании в архитектуре генерации, дополненной поиском (Retrieval-Augmented Generation, RAG), склонны к формированию фактологически некорректной информации («галлюцинациям») при отсутствии тонкой настройки на наборы данных, подаваемых в регуляторные органы, тогда как при использовании spaCy Clinical возможны ошибки в извлечении числовых параметров (дозировок) из-за слабой адаптации к контекстным условиям.

Преимущества использования модели BioBERT при подготовке регистрационных досье заключаются в возможности более корректной трактовки терминов (в том числе распознавания синонимичных конструкций и аббревиатур) благодаря предварительному обучению на биомедицинских данных (на базе PubMed), а также интеграции с онтологическими графами (например, расширение «Базовая онтология для обмена нормативно-правовой информацией» (Legal Knowledge Interchange Format — Core) через семантическую модель регуляторных требований, что позволяет рекурсивно анализировать связанные документы через семантическое сопоставление параметров. Однако ограничением остается зависимость от наличия аннотированных данных для тонкой настройки, а также величины вычислительной мощности при обработке больших объемов текстов [15].

Рекурсивная проверка данных в связанных документах, например технических отчетах и электронных лабораторных журналах, может быть реализована через интеграцию CQM в рабочие процессы исследователей, что позволяет обрабатывать взаимосвязанные группы файлов с сохранением целостности данных. Подобные архитектуры удобны при анализе документов благодаря возможности работать со структурированными и неструктурированными данными параллельно, сокращая время подготовки документов на 64% по сравнению с ручной обработкой [13]. Внедрение стандартов «Ресурсы для быстрого взаимодействия в здравоохранении (Fast Healthcare Interoperability Resources, FHIR)» дополняет NLP-методики, обеспечивая

совместимость данных для прогностической аналитики, однако требует решения проблемы их согласованности между различными системами [16]. Интеллектуальные системы ускоряют анализ за счет поиска на естественном языке и извлечения информации из изображений, но имеют ограничения, связанные с интерпретацией морфологически сложных конструкций и необходимостью ручной верификации для исключения ошибок [13, 17]. Современные подходы, такие как иерархическая кластеризация при помощи специализированной языковой модели, предобученной на биомедицинских текстах для улучшения понимания медицинской терминологии (Self-alignment pretraining for BERT, SapBERT), дают возможность уменьшить текстовой объем на 37% при сохранении ключевых сущностей [17].

Сравнительный анализ технологий *(табл. 1)* основан на данных из различных исследований с отличающимися наборами документов и условиями валидации. Поэтому прямое сопоставление их эффективности имеет ограниченную применимость и служит лишь для иллюстрации потенциальных возможностей. Для объективного сравнения необходимы контролируемые эксперименты на идентичных задачах и данных.

Предиктивные модели. Эффективным инструментом для разработки предиктивных моделей могут являться системы полного цикла, в которых интегрированы молекулярные дескрипторы и результаты клинических исходов, данных Абсорбция–Распределение–Метаболизм–Выведение–Токсичность (Absorption–Distribution–Metabolism–Excretion–Toxicity, ADMET), что подтверждается возможностями интегрированной базы данных ADMET и побочных реакций для предиктивного моделирования (IDAAPM, Integrated Database of ADMET and Adverse Effects of Predictive Modeling), объединяющей свойства, включенные в ADMET, нежелательные явления и биоактивность одобренных препаратов. Данные IDAAPM включают 1629 молекулярных структур с дескрипторами (логарифм коэффициента распределения, молекулярная масса, количество водородных связей), 36 963 описания взаимодействий «лекарство–мишень» и 2,5 млн данных о нежелательных явлениях, структурированных по системно-органным классам медицинского регуляторного словаря (Medical Dictionary for Regulatory Activities System Organ Classes, MedDRA SOC), что дает возможность значительно расширить входную матрицу для предиктивных моделей [18, 19]. Так, в моделях, которыми пользовалась компания Novartis при проверке концепции прогноза вероятности одобрения лекарственного препарата или наличия у него нового фармакологического действия, входные данные включали состав препарата, текущий статус показания, перечень биологических мишеней, механизм действия, данные клинических испытаний (фазы II–III), включая параметры количества пациентов, отобранных для исследования, длительности исследований и др. Результат работы модели — предположение о вероятности одобрения регулирующими органами с доверительным интервалом, рассчитанным методом бутстрепа[1] для оценки неопределенности. Ключевые параметры валидации точности ответа модели соответствовали следующим значениям: оценка ROC AUC (кривая ошибок) на независимом тестовом наборе (требование AUC > 0,8), где лучшие решения достигли AUC 0,88, превосходя базовые модели (AUC 0,78). Переобучение контролировалось при помощи сравнения публичной и приватной частей

Таблица 1. *Оценка эффективности цифровых моделей на основе искусственного интеллекта для работы с электронной версией общего технического документа*

Table 1. *Comparative table of digital model effectiveness for artificial intelligence-based electronic common technical document*

Модель *Model*	Точность извлечения сущностей, % *Entity Extraction Accuracy, %*	Скорость, стр/с *Speed, pps*	Проблемы *Troubles*
BioBERT	89–96 (субъекты/действия) *(subjects/actions)*	3,2	Низкая полнота в сложных контекстах *Low completeness in complex contexts*
spaCy Clinical	75–82	8,5	Ошибки в дозировках *Dosage errors*
GPT-4+RAG	78–85	0,8	Галлюцинации *Hallucinations*

Таблица составлена авторами по данным источников литературы / The table was adapted by the authors from the literature

[1] Dikta G, Scheer M. Bootstrap methods: With applications in R. Cham: Springer International Publishing; 2021. https://doi.org/10.1007/978-3-030-73480-0

тестовой выборки (так называемых «лидербордов»[2]) [20].

Внедрение интеллектуальных систем сталкивается с такими проблемами, как, например, неполнота данных о связывании активных молекул с мишенями или ограниченная интерпретируемость сложных ансамблевых моделей, что может затруднять принятие решений в отношении сформированных моделью ответов. Возможности расширения области применения интеллектуальных систем включают улучшение прогнозирования идиосинкразической токсичности за счет анализа сетевого взаимодействия белков (положение и функция конкретного белка внутри сложной сети) и функционального воздействия препаратов, что особенно актуально для раннего скрининга [21, 22].

Для задач прогнозирования сложных конечных точек, таких как орган-специфичная токсичность или полиморфный метаболизм, могут использоваться глубокие нейронные сети благодаря способности выявлять нелинейные взаимодействия в высокоразмерных данных[3] [23]. Графовые нейросети предпочтительны для задач, требующих анализа биологических взаимодействий (например, прогнозирование токсичности), бустинговые ансамбли[4] (XGBoost) более эффективны для задач классификации и регрессии на основе табличных данных результатов испытаний благодаря обработке разнородных признаков и высокой точности (AUC 0,84–0,88 против 0,73 в случае специализированных моделей токсичности) [20, 22]. Таким образом, решения на основе графов дают возможность интерпретировать данные в рамках конкретных биологических механизмов, но требуют значительных вычислительных ресурсов, в то время как бустинговые ансамбли демонстрируют робастность в решении прогностических задач на основе имеющихся данных, но ограничены в выявлении новых механизмов. Ключевые проблемы применения состоят в систематических смещениях обучающих данных при тонкой настройке, обусловленных ограниченной репрезентативностью выборок или культурными стереотипами [23], а также рисках экстраполяции вне облака данных стандартных электронных баз, содержащих сведения о свойствах химических соединений.

Генеративный ИИ. Модели, основанные на архитектуре типа RAG, могут быть успешно использованы для автоматизации поиска благодаря интеграции семантических эмбеддинговых технологий, что позволяет преодолевать ограничения традиционных методов, связанных с ключевыми словами или классификационными кодами. Применение таких моделей обеспечивает сопоставимую с ручным поиском точность идентификации релевантных документов, однако риски генерации недостоверной информации, включая ложные приоритеты, остаются существенными и могут приводить к ошибочным решениям [24]. Примеры ошибок ИИ, такие как присвоение ложного приоритета или некорректная интерпретация химических структур, подчеркивают необходимость внедрения компенсационных механизмов, включающих экспертизу человеком на заключительном этапе и специализированные модули проверки данных, которые верифицируют соответствие сгенерированных выводов оригинальным патентным источникам [24, 25].

В *таблице 2* обобщены ключевые параметры безопасности, которые могут быть использованы при разработке больших языковых моделей для обработки регуляторной документации. Генеративные нейросети подходят для задач семантического поиска и систематизации документов благодаря способности анализировать контекстную информацию и выявлять скрытые паттерны (устойчивая, систематически повторяющаяся схема или модель) в текстовых данных, что особенно актуально для обработки сложной технической терминологии в регуляторных документах, однако их внедрение требует проведения валидации точности ответов модели для минимизации рисков.

Прикладные решения для экспертных систем

Документоцентричные платформы. При помощи систем анализа и оценки документов на основе регуляторных требований реализуются процессы автоматической валидации данных по принципам ALCOA+ (attributable (прослеживаемость), legible (читаемость), contemporaneous (своевременность), original (подлинность),

[2] В контексте валидации моделей машинного обучения «лидерборд» используется для предотвращения переобучения, когда тестовая выборка часто разделяется на публичную и приватную части. Модель оптимизируется и проходит первоначальную проверку на публичном лидерборде, результаты которого видны разработчикам. Окончательная же валидация и сравнение моделей проводятся на приватном лидерборде, данные которого остаются скрытыми до окончания анализа.

[3] Высокоразмерные данные — данные, где на каждое исследуемое соединение (наблюдение) приходится огромное количество измеренных характеристик.

[4] Техника машинного обучения для задач классификации и регрессии, которая строит модель прогнозирования в форме ансамбля моделей с низкой точностью прогнозирования, обычно деревьев решений.

Таблица 2. *Параметры безопасности больших языковых моделей*

Table 2. *Security parameters of large language models in regulatory documents*

Параметр безопасности ***Security parameter***	**Описание** ***Description***	**Метод обеспечения** ***Method***
Точность классификации *Classification accuracy*	Доля корректно идентифицированных объектов относительно общего объема выборки *Proportion of correctly identified objects relative to the total sample size*	Валидация на независимых тестовых наборах *Validation on independent test kits*
Уровень ложноположительных результатов *False positive rate*	Частота ошибочного отнесения объектов к целевой категории *Frequency of erroneous object rating as target category*	Калибровка порогов классификации *Calibration of classification thresholds*
Уровень ложноотрицательных результатов *False negative rate*	Пропуск релевантных документов или данных в них *Omitting relevant documents or data*	Оптимизация полноты модели *Model completeness optimisation*
Устойчивость к галлюцинациям *Resistance to hallucinations*	Склонность модели к формированию неподтвержденных данных *Model propensity to generate unconfirmed data*	Интеграция механизмов проверки на достоверность *Integration of fact-checking mechanisms*
Контролируемость решений *Auditability of solutions*	Возможность ретроспективного анализа оснований для принятия решения *Possible retrospective analysis of the grounds for decision-making*	Регистрация входных данных и выходных прогнозов *Logging of input data and output forecasts*
Зависимость от качества обучения *Dependence on the quality of education*	Чувствительность к репрезентативности и объему обучающих данных *Sensitivity to representativeness and volume of training data*	Использование диверсифицированных источников *Use of diversified sources*

Таблица составлена авторами / The table was prepared by the authors

accurate (точность), complete (полнота), consistent (последовательность), enduring (устойчивость), available (доступность)), семантической организации архивов, интеллектуальной классификации запросов и кросс-документного выявления несоответствий, что обеспечивает снижение трудозатрат на 40%, повышение глубины проверки документации в хода аудита и оперативное выявление рисков несоблюдения GxP [14, 15, 26]. Для установления связей между процессом производства и формированием регистрационного досье немаловажное значение играет автоматизация обработки рукописных записей лабораторных журналов средствами оптического распознавания символов, преобразующая неструктурированные данные в цифровой формат для обеспечения атрибутивности и прослеживаемости изменений.

При этом динамическая структуризация ведомственных/отраслевых руководств по обращению лекарственных средств, а также внутрикорпоративных документов в XML-формат с тегированием компонентов позволяет автоматически связывать разделы документов, например изменения требований к очистке оборудования, с данными журналов и актуализировать пересмотр связанных валидационных отчетов в онтологических системах типа OntoReg[5], устраняя необходимость в ручном поиске и поддерживая актуальность существующих документов регуляторным требованиям и действующих версий документов архива при подготовке или внесении изменений в регистрационное досье [15].

Другие преимущества для пользователей раскрываются в возможности классификации запросов через модели распознавания семантического сходства на основе Clinical BERT, сопоставляющие вопросы пользователей с информационной базой знаний с точностью 90,66%, как, например, при направлении запроса о пределах стабильности препарата в категорию «Информация о продукте», что ускоряет обработку и стандартизирует ответы [14], тогда как поиск несоответствий реализует алгоритмы сравнения характеристик между взаимосвязанными документами, включая случаи расхождений заявленных параметров в валидационных

[5] Онтологическая информационная система, разработанная Engineering Science Department in the University of Oxford. Информация опубликована в статье Sesen MB, Suresh P, Banares-Alcantara R, Venkatasubramanian V. An ontological framework for automated regulatory compliance in pharmaceutical manufacturing. *Comput Chem Eng.* 2010;34(7):1155–69. https://doi.org/10.1016/j.compchemeng.2009.09.004

отчетах и реальных измерений в аналитических листах, выявляя нарушения ALCOA+ или противоречия в регистрационных досье, где нарушения могут маркироваться системой как критическое отклонение от требований нормативно-правового акта [14].

Некоторые ограничения, например снижение точности распознавания рукописно введенных сложных химических формул при оптическом распознавании (менее 70%), системы компенсируют прогностической аналитикой, в том числе автоматической генерацией подсказок по аудиту электрофоретических данных через анализ соответствия требованиям ICH Q2, что демонстрирует их роль в трансформации регуляторных процессов [14, 15]. Внедрение системы интеллектуальной подготовки документов (Intelligent Machine for Document Preparation, IMDP) в компании Eli Lilly продемонстрировало существенное сокращение временных затрат при подготовке документов: точность при создании документов достигла 89%, а скорость извлечения необходимой информации выросла в 3,6 раза [13]. Другой пример сокращения временных затрат при внедрении систем ИИ-контроля можно увидеть у компании Janssen, которая получила одобрение FDA на переход от серийного к непрерывному производству и сократила время от выходного анализа до выпуска серии с 30 до 10 дней [27].

Подобные результаты могут быть достигнуты благодаря автоматизации извлечения и структурирования данных, поддержке стандартов ALCOA+ (атрибутивность, прослеживаемость, современность), а также внедрению алгоритмов обработки естественного языка для анализа рукописных лабораторных записей, что устранило необходимость ручного ввода и верификации [7]. Например, при необходимости перекрестной проверки данных регистрационных карт, содержащихся в приложениях к отчету. Однако полная автоматическая генерация разделов документов по безопасности (например, разделы 2.3, 2.6 и 2.7 общего технического документа) сохраняет риски, связанные с интерпретацией контекстуальных нюансов и неструктурированных данных, требующих экспертного контроля для обеспечения регуляторной точности.

Исследования S. Viswanath и соавт. подчеркивают эффективность интеллектуальных систем, таких как IMDP, использующих NLP для автоматизированного поиска и интеграции данных из неструктурированных источников (текстовые документы) в шаблоны документов. Алгоритм CQM обрабатывает мультимодальные данные, включая тексты, таблицы и изображения, что позволяет сократить время подготовки документов в 3,6 раза при точности до 89%, превосходя традиционные методы [13]. Автореферирование и генерация текстовых разделов осуществляются через иерархическую кластеризацию семантически связанных фрагментов и последующее суммирование кластеров с помощью моделей типа BERT, обеспечивая релевантность контента [17].

Важной задачей является минимизация ошибок при интеграции данных: инструменты на базе SapBERT генерируют контекстные векторные представления, идентифицирующие ключевые сущности (например, медицинские термины), а проверка качества включает комплексную оценку по нескольким метрикам (например, ROUGE-L для полноты, BERTScore для семантической согласованности), позволяющую контролировать сохранение смысла оригинальных источников. Для оптимизации алгоритмической проверки предлагается трехуровневая схема: 1) автоматическая верификация через сравнение с эталонными онтологиями ADMET; 2) экспертный аудит происхождения данных с аннотированием источников; 3) динамическая коррекция на основе обратной связи [18].

Проблемы обработки и интерпретации данных, такие как обработка нестандартных форматов данных (химические структуры, графики) и валидация онтологической точности, частично решены через OCR и NLP, однако требуют ручной калибровки в случае сложных элементов. Интеграция с базами знаний типа IDAAPM (Integrated Database of ADMET and Adverse Effects of Predictive Modeling) предоставляет структурированные фармакологические данные для автоматического заполнения шаблонов, сокращая рутинные операции.

Контроль в реальном времени. Системы мониторинга в реальном времени находят применение в фармацевтическом производстве для непрерывного контроля критических процессов, особенно в асептических условиях, обязательным является соблюдение правил надлежащей производственной практики (GMP). Эти системы позволяют, например, автоматизировать обнаружение нарушений герметичности упаковки, отсутствие маркировки или несоответствие внешнего вида продукции, что значительно снижает риски, связанные с человеческим фактором, и повышает общую эффективность контроля качества [16, 28].

Интеграция платформ с системами компьютерного зрения для реального мониторинга в асептических условиях производства требует алгоритмов, обеспечивающих высокую точность детекции аномалий и соответствие требованиям GMP, где применение систем реального времени на базе компьютерного зрения (YOLOv8) демонстрирует значительные преимущества благодаря высокой скорости обработки изображений, достигающей 30 и более кадров в секунду (FPS), что критично для обнаружения отклонений в динамических производственных процессах, таких как визуальный контроль укупорки флаконов или перемещения персонала в зонах чистых помещений [28].

Технические требования к таким системам: разрешение камер не менее 1920×1080 пикселей для детализации объектов; частота кадров не менее 30 FPS для обеспечения плавного анализа движущихся элементов; уровень освещенности от 500 лк для минимизации шумов; точность детекции аномалий ≥95% для корректной идентификации несоответствий, как, например, микродефекты упаковки или нарушения асептических процедур; калибровка алгоритмов под специфику производственных линий, включая адаптацию к изменениям освещения, конфигурации оборудования и типам продукции, что позволяет повысить надежность системы, хотя это сопряжено с необходимостью масштабной предварительной настройки и обучением моделей на репрезентативных наборах данных [16]. Так, системы компьютерного зрения могут автоматически сверять выполняемые оператором на производстве действия с последовательностью этапов, зафиксированных в технологических инструкциях и спецификациях, являющихся частью регистрационного досье. Результаты параметров контроля в режиме реального времени могут быть сопоставлены с утвержденными значениями и использованы при формировании отчета.

Преимуществами подобных систем являются снижение количества ручных проверок, повышение преемственности данных и возможность непрерывного аудита; к недостаткам можно отнести высокие первоначальные затраты на внедрение, необходимость постоянного дообучения моделей и риски, связанные с ложными срабатываниями. Для валидации предлагается использовать многоуровневую схему тестирования, включающую синтетические данные, контролируемые нарушения на испытательных стендах и постепенное внедрение в реальные производственные циклы с последующей корректировкой на основе обратной связи [16, 28]. Результаты обработки систем компьютерного зрения могут автоматически фиксироваться, структурироваться и сопоставляться с регуляторными требованиями, заложенными в онтологических системах типа OntoReg и утвержденными в регистрационном досье спецификациями, и технологическими процессами.

Это позволит не только создать запись об инциденте в реальном времени, но и мгновенно сформировать соответствующий отчет об отклонении, обновить статусы рисков в связанных документах, а также инициировать корректирующие действия. Интеллектуальная система превращает единичное событие в структурированную, прослеживаемую и атрибутивную запись в документоцентричной платформе, напрямую связывая его с документами общего технического документа досье на лекарственный препарат. Это обеспечивает неразрывную связь между разрабатываемой/утвержденной документацией и производственной практикой, создавая надежную доказательную базу для регуляторных органов.

ЗАКЛЮЧЕНИЕ

Таким образом, переход от фрагментарной автоматизации к комплексному управлению жизненным циклом регистрационного досье обеспечивает значительные операционные и качественные улучшения. Системы NLP на базе специализированных архитектур, таких как BioBERT и Calibrated Quantum Mesh с точностью извлечения сущностей до 96%, способны сократить время обработки документов досье на 45–64%, минимизируя ошибки ручной подготовки и ресурсозатраты. Предиктивные ансамблевые модели и графовые нейронные сети с AUC до 0,88 повышают точность прогноза одобрения препаратов за счет интеграции множества баз данных. Документоцентричные платформы с поддержкой ALCOA+, интегрирующие NLP для работы с неструктурированными источниками, сокращают сроки подготовки досье, обеспечивая сквозную прослеживаемость и соответствие GxP-требованиям. Системы реального времени на базе компьютерного зрения (YOLOv8) с точностью детекции ≥95% позволяют автоматизировать контроль производственных процессов, снижая риски нарушений GMP, однако их максимальная эффективность раскрывается при интеграции с документоцентричными платформами.

Цифровая трансформация процессов подготовки документов регистрационного досье может способствовать сокращению сроков вывода

препаратов на рынок, повышению качества регистрационных документов. Однако внедрение цифровых систем сталкивается с методологическими проблемами, включая риск галлюцинаций генеративного ИИ, ограниченную интерпретируемость моделей и проблему валидации в условиях различий терминологических и нормативно-правовых требований между странами и регуляторными органами.

Для минимизации этих рисков необходимо развитие регуляторных стандартов валидации систем ИИ и машинного обучения, развертывание изолированных тестовых сред для критически значимых контекстов и внедрение многоуровневых схем аудита решений. Следует отметить, что обобщение эффективности таких технологий, как RAG, является предварительным, поскольку их результат критически зависит от реализации (состава векторной базы, стратегии поиска и ранжирования фрагментов).

Формирование окончательных выводов о преимуществах той или иной архитектуры требует их тестирования в единых регуляторных сценариях. Перспективные направления включают разработку унифицированного терминологического аппарата, создание самообучающихся систем пострегистрационного мониторинга безопасности.

ЛИТЕРАТУРА / REFERENCES

1. Niazi SK. Bridging the regulatory divide: a dual-pathway framework using SRA approvals and AI evaluation to ensure drug quality in developing countries. *Pharmaceuticals (Basel)*. 2025;18(7):1024. https://doi.org/10.3390/ph18071024
2. Cova T, Vitorino C, Ferreira M, et al. Artificial intelligence and quantum computing as the next pharma disruptors. *Methods Mol Biol*. 2022;2390:321–47. https://doi.org/10.1007/978-1-0716-1787-8_14
3. Patil RS, Kulkarni SB, Gaikwad VL. Artificial intelligence in pharmaceutical regulatory affairs. *Drug Discov Today*. 2023;28(9):103700. https://doi.org/10.1016/j.drudis.2023.103700
4. Ajmal CS, Yerram S, Abishek V, et al. Innovative approaches in regulatory affairs: leveraging artificial intelligence and machine learning for efficient compliance and decision-making. *AAPS J*. 2025;27(1):22. https://doi.org/10.1208/s12248-024-01006-5
5. Macdonald JC, Isom DC, Evans DD, Page KJ. Digital innovation in medicinal product regulatory submission, review, and approvals to create a dynamic regulatory ecosystem — are we ready for a revolution? *Front Med (Lausanne)*. 2021;8:660808. https://doi.org/10.3389/fmed.2021.660808
6. Nene L, Flepisi BT, Brand SJ, et al. Evolution of drug development and regulatory affairs: The demonstrated power of artificial intelligence. *Clin Ther*. 2024;46(8):e6–14. https://doi.org/10.1016/j.clinthera.2024.05.012
7. Sharma K, Manchikanti P. Regulation of artificial intelligence in drug discovery and health care. *Biotechnol Law Rep*. 2020;39(5):371–80. https://doi.org/10.1089/blr.2020.29183.ks
8. Fu L, Jia G, Liu Z, et al. The applications and advances of artificial intelligence in drug regulation: a global perspective. *Acta Pharmaceutica Sinica B*. 2025;15(1):1–14. https://doi.org/10.1016/j.apsb.2024.11.006
9. Gude S, Gude YS. The synergy of artificial intelligence and machine learning in revolutionizing pharmaceutical regulatory affairs. *Transl Regul Sci*. 2024;6(2):37–45. https://doi.org/10.33611/trs.2024-005
10. Higgins DC, Johner C. Validation of artificial intelligence containing products across the regulated healthcare industries. *Ther Innov Regul Sci*. 2023;57(4):797–809. https://doi.org/10.1007/s43441-023-00530-4
11. Kandhare P, Kurlekar M, Deshpande T, Pawar A. Artificial intelligence in pharmaceutical sciences: a comprehensive review. *Med Nov Technol Devices*. 2025;27:100375. https://doi.org/10.1016/j.medntd.2025.100375
12. Oualikene-Gonin W, Jaulent MC, Thierry JP, et al. Artificial intelligence integration in the drug lifecycle and in regulatory science: policy implications, challenges and opportunities. *Front Pharmacol*. 2024;15:1437167. https://doi.org/10.3389/fphar.2024.1437167
13. Viswanath S, Fennell JW, Balar K, Krishna P. An industrial approach to using artificial intelligence and natural language processing for accelerated document preparation in drug development. *J Pharm Innov*. 2021;16(2):302–16. https://doi.org/10.1007/s12247-020-09449-x
14. Saraswat N, Li C, Jiang M. Identifying the question similarity of regulatory documents in the pharmaceutical industry by using the recognizing question entailment system: evaluation study. *JMIR AI*. 2023;2:e43483. https://doi.org/10.2196/43483
15. Sapkota K, Aldea A, Younas M, et al. Automating the semantic mapping between regulatory guidelines and organizational processes. *Serv Oriented Comput Appl*. 2016;10(4):365–89. https://doi.org/10.1007/s11761-016-0197-2
16. Anderson C, Algorri M, Abernathy MJ. Real-time algorithmic exchange and processing of pharmaceutical quality data and information. *Int J Pharm*. 2023;645:123342. https://doi.org/10.1016/j.ijpharm.2023.123342
17. Dalal A, Ranjan S, Bopaiah Y, et al. Text summarization for pharmaceutical sciences using hierarchical clustering with a weighted evaluation methodology. *Sci Rep*. 2024;14(1):20149. https://doi.org/10.1038/s41598-024-70618-w
18. Legehar A, Xhaard H, Ghemtio L. IDAAPM: Integrated database of ADMET and adverse effects of predictive modeling based on FDA approved drug data. *J Cheminformatics*. 2016;8:33. https://doi.org/10.1186/s13321-016-0141-7
19. Whitehead TM, Strickland J, Conduit GJ, et al. Quantifying the benefits of imputation over QSAR methods in toxicology data modeling. *J Chem Inf Model*. 2024;64(7):2624–36. https://doi.org/10.1021/acs.jcim.3c01695
20. Siah KW, Kelley NW, Ballerstedt S, et al. Predicting drug approvals: the Novartis data science and artificial intelligence challenge. *Patterns (NY)*. 2021;2(8):100312. https://doi.org/10.1016/j.patter.2021.100312
21. Boonsom S, Chamnansil P, Boonseng S, Srisongkram T. ToxSTK: A multi-target toxicity assessment utilizing molecular structure and stacking ensemble learning. *Comput Biol Med*. 2025;185:109480. https://doi.org/10.1016/j.compbiomed.2024.109480
22. Lysenko A, Sharma A, Boroevich KA, Tsunoda T. An integrative machine learning approach for prediction of toxicity-related drug safety. *Life Sci Alliance*. 2018;1(6):e201800098. https://doi.org/10.26508/lsa.201800098

23. Gray M, Samala R, Liu Q, et al. Measurement and mitigation of bias in artificial intelligence: a narrative literature review for regulatory science. *Clin Pharmacol Ther.* 2024;115(4):687–97. https://doi.org/10.1002/cpt.3117
24. Ribeiro LC, Muzaka V. Needle in a haystack: Harnessing AI in drug patent searches and prediction. *PloS One.* 2024;19(12):e0311238. https://doi.org/10.1371/journal.pone.0311238
25. Williams J, Boyce D, Collu G, et al. Generative AI: A generation-defining shift for biopharma regulatory affairs. *Nat Rev Drug Discov.* 2025;24(9):651–2. https://doi.org/10.1038/d41573-025-00089-9
26. Sharma S, Mathure D, Dingankar S, Dhapte-Pawar V. In pursuit of software solutions for pharmaceutical regulatory affairs: Insights and trends. *Ann Pharm Fr.* 2025;83(6):1053–61. https://doi.org/10.1016/j.pharma.2025.05.005
27. Niazi SK. Regulatory perspectives for AI/ML implementation in pharmaceutical GMP environments. *Pharmaceuticals (Basel).* 2025;18(6):901. https://doi.org/10.3390/ph18060901
28. Syed FM, Faiza Kousar ES. AI in securing pharma manufacturing systems under GxP compliance. *International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence.* 2024;15(1):448–72.

Вклад авторов. Все авторы подтверждают соответствие своего авторства критериям ICMJE. Наибольший вклад распределен следующим образом: *М.А. Ярошинский* — концепция работы; *Е.И. Балакин, А.С. Павлов* — написание текста рукописи; *М.В. Андреева* — формулировка выводов; *А.С. Павлов* — работа с источниками литературы; *А.Ю. Савченко, В.И. Пустовойт* — утверждение окончательной версии рукописи для публикации.

Authors' contributions. All the authors confirm that they meet the ICMJE criteria for authorship. The most significant contributions were as follows. *Milan A. Yaroshinsky* conceptualised the study. *Evgenii I. Balakin*, *Alexander S. Pavlov* drafted the manuscript. *Maria V. Andreeva* formulated the conclusions. *Alexander S. Pavlov* worked with literature sources. *Alla Yu. Savchenko, Vasily I. Pustovoit* approved the final version for publication.

ОБ АВТОРАХ / AUTHORS

Ярошинский Милан Анатольевич / **Milan A. Yaroshinsky**
ORCID: https://orcid.org/0009-0008-5302-7609
Андреева Мария Владимировна / **Maria V. Andreeva**
ORCID: https://orcid.org/0009-0008-9805-2402
Балакин Евгений Игоревич, канд. мед. наук / **Evgenii I. Balakin**, Cand. Sci. (Med.)
ORCID: https://orcid.org/0000-0001-5545-135X
Савченко Алла Юрьевна, канд. мед. наук / **Alla Yu. Savchenko**, Cand. Sci. (Med.)
ORCID: https://orcid.org/0000-0003-2734-5036
Павлов Александр Сергеевич / **Alexander S. Pavlov**
ORCID: https://orcid.org/0009-0006-4636-0978
Пустовойт Василий Игоревич, д-р мед. наук / **Vasily I. Pustovoit**, Dr. Sci. (Med.)
ORCID: https://orcid.org/0000-0003-3396-5813

Поступила 04.09.2025
После доработки 28.10.2025
Принята к публикации 10.12.2025

Received 4 September 2025
Revised 28 October 2025
Accepted 10 December 2025