КУРОПАТКИНА Т.А.,
к.б.н., ФГБОУ ВО «Российский экономический университет имени Г.В. Плеханова», г. Москва, Россия,
e-mail: 0sylphide0@gmail.com

КИЛЬМИШКИН Н.В.,
ФГБОУ ВО «Российский экономический университет имени Г.В. Плеханова», г. Москва, Россия,
e-mail: kilmiskinn@gmail.com

ПАНТЕЛЕЕВ В.И.,
к.м.н., ФГБОУ ВО «Российский экономический университет имени Г.В. Плеханова», г. Москва, Россия,
e-mail: panteleev.vi@rea.ru

КУБРАКОВ Д.Д.,
ФГБОУ ВО «Российский экономический университет имени Г.В. Плеханова», г. Москва, Россия,
e-mail: kubrakoff.dmitry@yandex.ru

ИВАНОВА П.М.,
ФГБОУ ВО «Российский экономический университет имени Г.В. Плеханова», г. Москва, Россия,
e-mail: polianna_654@mail.ru

ТИТОВ Ю.П.,
к.т.н., ФГБОУ ВО «Российский экономический университет имени Г.В. Плеханова», г. Москва, Россия,
e-mail: kalengul@mail.ru

ВАЛЕНТЕЙ С.Д.,
д.э.н., профессор, ФГБОУ ВО «Российский экономический университет имени Г.В. Плеханова»,
г. Москва, Россия, e-mail: svalentey@yandex.ru

ШИМАНОВСКИЙ Н.Л.,
д.м.н., профессор, член-корреспондент РАН, ФГБОУ ВО «Российский экономический
университет имени Г.В. Плеханова», г. Москва, Россия, e-mail: shimann@yandex.ru

СРАВНЕНИЕ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ ТЕКСТОВ ИНСТРУКЦИЙ ЛЕКАРСТВЕННЫХ СРЕДСТВ ДЛЯ ОЦЕНКИ РИСКОВ ПОЛИФАРМАКОТЕРАПИИ С ПОМОЩЬЮ СЕТИ БАЙЕСА

DOI: 10.25881/18110193_2026_2_18

Аннотация. *Актуальность. На сегодняшний день не существует алгоритмов, способных учитывать взаимодействие трех и более одновременно применяемых лекарственных средств, что необходимо для снижения рисков проявления побочных эффектов при полипрагмазии. Для создания подобного алгоритма предлагается использовать байесовскую сеть — графическую вероятностную модель на основе семантического графа, который строится на информации, полученной из медицинских текстов. Однако для его построения необходимо проводить анализ обширного объема имеющихся медицинских текстов, что требует чрезвычайно больших временных и интеллектуальных затрат медицинских специалистов, ввиду чего существует необходимость автоматизации этого процесса.*

Цель. Сравнение различных архитектур обработки естественного языка и технологий глубокого обучения для создания инструмента автоматического анализа текстовой медицинской информации о фармакокинетике и фармакодинамике, побочных эффектах и взаимодействиях для создания семантических графов, использующихся впоследствии для построения байесовской сети анализа лекарственных взаимодействий.

Материалы и методы. Для автоматической разметки текстов медицинских инструкций можно использовать метод обработки естественного языка. В данной работе для выделения именованных сущностей в рамках исследования были протестированы модели глубокого обучения на основе архитектур BERT и T5 с использованием формата BIO.

Результаты. Наиболее высокую точность показали модели на базе BERT: RuBioBERT достигла показателя accuracy равного 0,9569±0,0052, немного опережая ruBERT, которая показала сопоставимые результаты. Выявлено, что предобучение на общих текстах дает преимущество на начальных этапах, тогда как специализированные модели, такие как RuBioBERT, становятся эффективнее при увеличении числа эпох обучения. Это связано с тем, что общеупотребительные тексты обеспечивают более универсальное представление языка, в то время как научные медицинские статьи отличаются своей спецификой, не полностью соответствующей структуре инструкций по медицинскому применению лекарственных средств.

Выводы. Модели BERT позволяют добиться качества разметки, близкого к экспертному, что делает их эффективным инструментом для построения семантического графа лекарственных взаимодействий. Автоматизация NER-разметки позволяет сократить трудозатраты экспертов и повысить точность обработки текстов. Таким образом, применение моделей BERT, таких как ruBERT и RuBioBERT, является перспективным методом для автоматизации анализа медицинских текстов и построения платформ для оценки сложных лекарственных взаимодействий.

Ключевые слова: *полифармация, байесовские сети, обработка естественного языка, искусственный интеллект, лекарственные взаимодействия, побочные эффекты.*

Для цитирования: *Куропаткина Т.А., Кильмишкин Н.В., Пантелеев В.И., Кубраков Д.Д., Иванова П.М., Титов Ю.П., Валентей С.Д., Шимановский Н.Л. Сравнение языковых моделей для извлечения именованных сущностей из текстов инструкций лекарственных средств для оценки рисков полифармакотерапии с помощью сети Байеса. Врач и информационные технологии. 2026; 2: 18-35. DOI: 10.25881/18110193_2026_2_18.*

KUROPATKINA T.A.,
PhD, Plekhanov Russian University of Economics, Moscow, Russia,
e-mail: 0sylphide0@gmail.com

KILMISHKIN N.V.,
Plekhanov Russian University of Economics, Moscow, Russia,
e-mail: kilmiskinn@gmail.com

PANTELEEV V.I.,
PhD, Plekhanov Russian University of Economics, Moscow, Russia,
e-mail: panteleev.vi@rea.ru

KUBRAKOV D.D.,
Plekhanov Russian University of Economics, Moscow, Russia,
e-mail: kubrakoff.dmitry@yandex.ru

IVANOVA P.M.,
Plekhanov Russian University of Economics, Moscow, Russia,
e-mail: polianna_654@mail.ru

TITOV YU.P.,
PhD, Plekhanov Russian University of Economics, Moscow, Russia,
e-mail: kalengul@mail.ru

VALENTEY S.D.,
DSc, Professor, Plekhanov Russian University of Economics, Moscow, Russia,
e-mail: svalentey@yandex.ru

SHIMANOVSKY N.L.,
DSc, Professor, Corresponding Member of the Russian Academy of Sciences, Plekhanov Russian University of Economics, Moscow, Russia, e-mail: shimann@yandex.ru

A COMPARATIVE STUDY OF LANGUAGE MODELS FOR NAMED ENTITY RECOGNITION FROM DRUG LABELS FOR POLYPHARMACY RISK ASSESSMENT USING BAYESIAN NETWORKS

DOI: 10.25881/18110193_2026_2_18

Abstract. *Background. Currently, there are no algorithms capable of accounting for interactions of three or more concomitantly administered medications, which is necessary to reduce the risks of side effects associated with polypharmacy. To develop such an algorithm, we propose using a Bayesian network — a graphical probabilistic model based on a semantic graph constructed using information extracted from medical texts. However, constructing such a network requires analyzing a vast amount of available medical texts, which is extremely time-consuming and intellectually demanding for medical professionals, necessitating the automation of this process.*

Objective. To compare various natural language processing architectures and deep learning technologies for the creation a tool for automatic analysis of medical text information on pharmacokinetics and pharmacodynamics, side effects, and interactions to create semantic graphs, subsequently used to build a Bayesian network for drug interaction analysis.

Methods. Natural language processing can be used to automatically tag medical instruction texts. In this study, deep learning models based on the BERT and T5 architectures were tested using the BIO format to extract named entities.

Results. BERT-based models showed the highest accuracy: RuBioBERT achieved an accuracy of 0.9569±0.0052, slightly outperforming ruBERT, which showed comparable results. It was found that pre-training on general texts provides an advantage in the initial stages, while specialized models such as RuBioBERT become more effective with an increasing number of training epochs. This is due to the fact that general texts provide a more universal language representation, while scientific medical articles have their own specific features which do not fully correspond to the structure of drug labels.

Conclusion. BERT models achieve near-expert-quality annotation, making them an effective tool for constructing semantic graphs of drug interactions. Automated NER annotation reduces expert workload and improves text processing accuracy. Therefore, the use of BERT models, such as ruBERT and RuBioBERT, is a promising method for automating medical text analysis and building platforms for assessing complex drug interactions.

Keywords: *polypharmacy, Bayesian networks, natural language processing, artificial intelligence, drug interactions, side effects.*

For citation: *Kuropatkina T.A., Kilmishkin N.V., Panteleev V.I., Kubrakov D.D., Ivanova P.M., Titov Yu.P., Valentey S.D., Shimanovsky N.L.A comparative study of language models for named entity recognition from drug labels for polypharmacy risk assessment using Bayesian networks. Medical doctor and information technology. 2026; 2: 18-35. DOI: 10.25881/18110193_2026_2_18.*

ВВЕДЕНИЕ

Полифармация представляет собой одновременное применение нескольких лекарственных средств (ЛС) и становится все более распространенной в условиях роста числа пациентов с полиморбидностью, особенно в пожилом возрасте [1, 2]. Для контроля сопутствующих заболеваний комбинированная терапия зачастую неизбежна, однако она повышает риск развития нежелательных взаимодействий и побочных эффектов [3]. В таких случаях применение большого числа ЛС требует тщательной оптимизации и анализа с целью предотвращения нежелательных лекарственных взаимодействий и снижения риска неблагоприятных исходов. Исследования, проведённые Н.В. Измозжеровой и соавторами, показали, что необоснованная и нерациональная полифармация встречается у 28–30% пожилых пациентов, наблюдаемых в амбулаторных условиях [4]. Это подчёркивает актуальность проблемы рационального назначения ЛС и необходимости разработки инструментов, способных автоматизировать анализ сложных лекарственных схем.

Современные платформы для анализа лекарственных взаимодействий, такие как Drugs.com [5], MedScape.com [6], DrugBank.com [7], WebMD.com [8], основаны преимущественно на данных клинических исследований парных комбинаций ЛС и не позволяют оценить риски при одновременном применении трех и более ЛС, а также не обеспечивают персонализацию. В связи с этим, разработка и внедрение систем, основанных на методах искусственного интеллекта (ИИ), становятся ценным инструментом для решения данной проблемы, поскольку они могут обучаться на текстовых данных, данных из баз международного фармаконадзора и клинических случаях, полученных от реальных пациентов, выявляя новые паттерны возможных взаимодействий и повышая адаптивность медицинских систем [9, 10].

Одним из перспективных методов создания математической модели (машинного алгоритма) являются байесовские сети — вероятностные модели, позволяющие учитывать множество факторов и их взаимосвязей [11]. Байесовские сети строятся на основе семантических графов [12], которые могут применяться для решения медицинских задач диагностики, прогнозирования течения болезни [13] и расчета вероятности возникновения конкретных побочных эффектов при комбинации различных ЛС. Качество построения байесовской сети и предсказаний напрямую зависят от качества составления семантических графов, которые требуют выявления из текста важной и специфической информации, установление связей между семантическими структурами (рис. 1).

Важным этапом для обеспечения адаптивности и точности предсказаний такой модели является обучение на реальных данных, включая информацию из текстов медицинских инструкций, профильных литературных источников, баз фармаконадзора и других источников достоверных медицинских материалов [9, 10, 14]. Это позволяет выявлять истинные закономерности и потенциальные взаимодействия, которые могут быть упущены практикующим врачом из-за огромного потока информации и ограниченности человеческой памяти.

Однако ключевым препятствием на пути создания байесовской сети является отсутствие высококачественных структурированных данных о фармакологических свойствах ЛС. Современные инструкции по медицинскому применению ЛС представляют собой объемные тексты, написанные на естественном языке с использованием свободных формулировок. Эти документы содержат критически важные сведения о фармакокинетике, фармакодинамике, метаболических путях, побочных эффектах и типах взаимодействий, но в форме, непригодной для прямого использования в вычислительных моделях. Ручное извлечение и аннотация такой информации требуют значительных временных ресурсов эксперта, что делает масштабирование системы весьма затруднительным. Кроме того, субъективность интерпретации и вариабельность формулировок между разными инструкциями существенно осложняют обработку текстовых данных.

В этих условиях автоматизация этапа подготовки текстовых данных становится актуальной задачей при разработки надежных ИИ-систем в фармакотерапии. Применение алгоритмов обработки естественного языка (NLP, Natural Language Processing) [15] позволяет автоматически извлекать именованные сущности (NER, Named Entity Recognition) из любых текстов, в

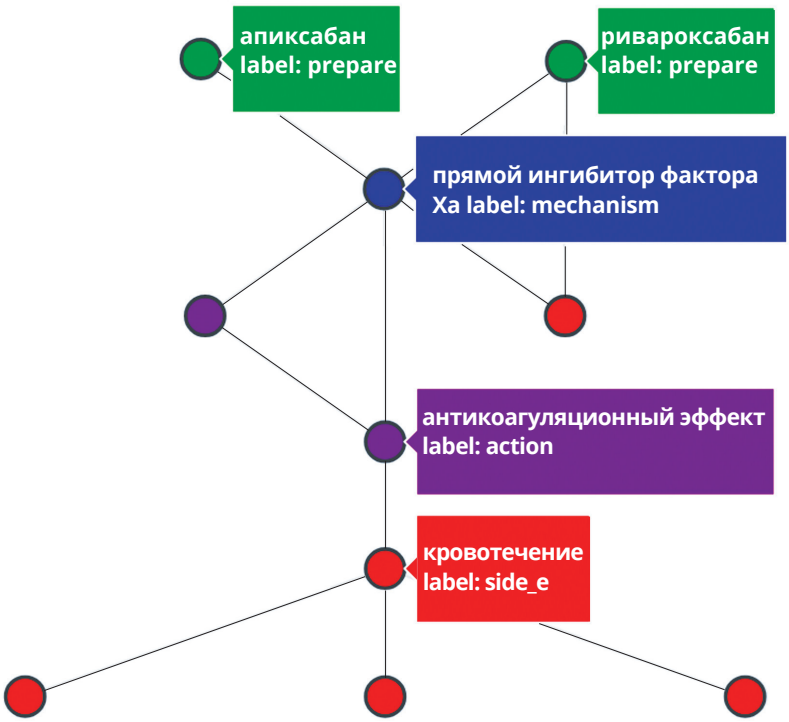

Рисунок 1 – Пример семантического графа для ЛС Апиксабан и Ривароксабан на основе инструкций по медицинскому применению ЛС.

том числе текстов медицинских инструкций, ключевую информацию, включая наименования ЛС, рекомендуемые дозы, метаболические пути, ферменты, целевые рецепторы, побочные эффекты и типы взаимодействий, лежащих в основе семантических графов, которые моделируют взаимосвязи между лекарственными веществами с учетом их влияния друг на друга через общие механизмы действия, метаболические пути и клинические проявления. Такой подход позволяет систематизировать и анализировать сложные фармакологические взаимодействия, способствуя более точному прогнозированию их последствий.

Целью настоящей работы была сравнительная оценка различных методов NLP и технологий глубокого обучения для последующего создания инструмента автоматического анализа текстовой медицинской информации о фармакокинетике и фармакодинамике, побочных эффектах, определяющих лекарственные взаимодействия для создания семантических графов (пример семантического графа представлен на рисунке 1), использующихся в последствии для построения байесовской сети анализа взаимодействий ЛС.

МАТЕРИАЛЫ И МЕТОДЫ

В качестве анализируемых текстов, которые вошли в основу разрабатываемого алгоритма, были использованы инструкции по медицинскому применению по 59 ЛС из государственного реестра лекарственных средств (ГРЛС) [16], отобранных на основании клинических рекомендаций для комплексной терапии хронической сердечной недостаточности [17]. Данная патология была выбрана ввиду того, что она имеет огромное социальное значение и зачастую характеризуется назначением множества ЛС. Международная классификация болезней (МКБ-10) [18] была использована для добавления наименований патологических состояний, их синонимов, названий синдромов, или симптомов, используемых в инструкциях. Работа с данными из МКБ-10 потребовала дополнительной обработки. Табличные данные были структурированы, а текстовые — нормализованы, что включало унификацию терминологии, удаление

неоднозначностей и преобразование в удобный для машинного анализа формат.

Подготовка набора данных для обучения осуществлялась с использованием реализованного алгоритма [19], автоматизирующего процессы очистки и форматирования текста, включая удаление скобок с их содержимым, неинформативных таблиц и метаданных, а также разбиение текста на отдельные предложения. Исключение данной информации позволило выделить ключевые смысловые элементы текста, а разбиение на предложения способствовало упрощению обработки, поскольку предложения представляют собой логически завершенные языковые конструкции, что облегчает анализ и структурирование данных.

Для реализации поиска именованных сущностей на основе правил был создан специализированный словарь именованных сущностей [20]. В данной задаче именованная сущность представляла собой определенный фрагмент текста, конкретную информацию, такую как международное непатентованное название ЛС, побочный эффект, механизм действия, реализуемый эффект, прочие медицинские термины и т.д.

В рамках обработки текстовых данных применялись процедуры сегментации на предложения и токенизации, а также лемматизации, направленной на нормализацию словоформ путем приведения их к исходной словарной форме [21]. Подготовка обучающего корпуса осуществлялась с помощью реализованного программного обеспечения на языке программирования Python [22], автоматизирующего ключевые этапы предобработки текстов, включая очистку и нормализацию формата исходных текстовых данных. Разработанные алгоритмы выполняли обработку исходных текстовых данных с устранением избыточной информации, такой как конструкции в скобках, второстепенные табличные данные и метаданные, не релевантные для последующего анализа. Использование собственных алгоритмических решений позволило оптимизировать качество обработки текстов на естественном языке.

Далее текст сегментировали на предложения с использованием python-библиотеки «razdel» (проект с открытым исходным кодом, разработанный Лабораторией анализа данных Александра Кукушкина [23]), которая обеспечивает высокую точность сегментации для русского языка. Это позволило сохранить структурированность данных, что критически важно для задач классификации и других этапов NLP. Выделение аббревиатур проводили с помощью разработанного алгоритма, который проверял слова на наличие двух или более заглавных букв, что позволяло автоматически классифицировать их как аббревиатуры (например, ВИЧ, НПВС).

Тексты были размечены с помощью BIO (Begin, Inside, Outside)-нотации (таблица 1) [24], которая является стандартным методом аннотирования текстов для обучения моделей распознавания именованных сущностей. Этот формат позволяет моделям точно определять границы и категории сущностей.

В представленной таблице 1 приведен пример разметки текста, где каждому слову или знаку препинания соответствует определенная

Таблица 1 – Разметки в нотации BIO

Слово	Разметка BIO
Периндоприл	B-prepare
оказывает	B-mechanism
сосудорасширяющее	I-mechanism
действие	I-mechanism
,	O
способствует	B-mechanism
восстановлению	I-mechanism
эластичности	O
крупных	B-noun
артерий	I-noun
и	O
структуры	B-noun
сосудистой	I-noun
стенки	I-noun
мелких	I-noun
артерий	I-noun
,	O
а	O
также	O
уменьшает	B-mechanism
гипертрофию	B-side_e
левого	I-side_e
желудочка	I-side_e

метка BIO. Например, слово "Периндоприл" помечено как B-prepare, что означает начало сущности категории "препарат". Следующие слова, такие как "оказывает" (B-mechanism) и "сосудорасширяющее" (I-mechanism), обозначают механизм действия, где "B-mechanism" — начало, а "I-mechanism" — продолжение этой категории. Слова, не относящиеся к выделенным сущностям, например, запятые или союзы, помечены как O.

В формате BIO каждому слову или токену — минимальной значимой единице, которая может быть не только словом, но и знаком препинания, числом или даже частью сложного слова в зависимости от языка в тексте, присваивается один из следующих тегов: B-XXX — начало сущности типа XXX; I-XXX — внутренняя часть сущности типа XXX; O — слово не является частью какой-либо сущности. Такой подход позволяет алгоритмам машинного обучения точно определять границы сущностей в тексте, даже если они состоят из нескольких слов.

Реализованный алгоритм [25] использовал библиотеки и инструменты NLP, такие как spaCy (библиотека для python компании Explosion AI, Германия [26,27]) для синтаксического анализа предложения и razdel для токенизации и разбиения текста на предложения на этапе подготовки структурированной коллекции текстовых данных, собранных для анализа и обучения моделей (Рисунок 2).

Алгоритм разметки включал несколько этапов: сначала проводился лексический анализ, при котором текст алгоритмом [28] разбивался на токены, определяя их текстовое представление и часть речи. Затем токены нормализовали с помощью лемматизации, чтобы исключить грамматическую вариабельность одного и того же слова. После этого каждому токену присваивали тег, отражающий его грамматическую и контекстуальную роль. Также была разработана функция для обработки отрицательной частицы "не", корректирующая её тег в зависимости от контекста.

В результате алгоритм выделял слова как показано на схеме, показанной ниже (Рисунок 3).

Автоматическая разметка на основе правил применялась в качестве первоначального этапа. Для получения более корректных результатов разметка была скорректирована экспертами, после чего было произведено дообучение на инструкциях для ЛС, размеченных алгоритмом, упомянутым ранее, по нотации BIO. Дообучение модели проводилось на архитектурах BERT (ruBERT, RuBioBERT) и T5 (ruT5 и ruT5-multitask). Архитектура BERT разработана компанией Google, США [29]. RuBERT (разработана лабораторией нейронных систем и глубокого обучения Московского физико-технического института (МФТИ) при поддержке Сбербанка и Национальной технологической инициативы (НТИ), Россия [30]), исходно предобученная на крупном корпусе общего русскоязычного текста, служит надёжной базой для широкого спектра NLP-задач на русском языке. RuBioBERT (разработана российскими исследователями для биомедицинских задач, Россия [31]), дополнительно дообученная на биомедицинских текстах (научные статьи, клинические протоколы, фармакологические описания) позволяет учесть предметную семантику, терминологию заболеваний, действующих веществ, побочных эффектов и дозировок. Модели T5 (разработана компанией Google, США [32]) и

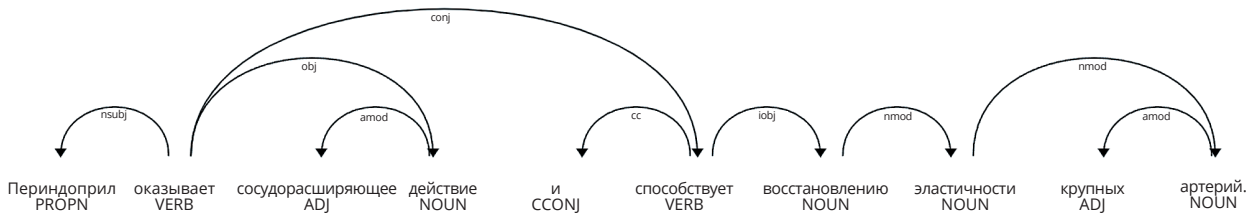

Рисунок 2 – Дерево зависимостей слов в предложении, построенное с помощью библиотеки «spaCy».

Рисунок 3 – Автоматическая разметка на основе алгоритма [25].

её русскоязычные версии ruT5 и ruT5-multitask (разработаны Sber AI, Россия [33, 34]) были выбраны для сравнения генеративного подхода к NER с традиционной токен-классификацией. Архитектура T5 позволяет формулировать задачу NER как генерацию текста, что особенно эффективно при работе с нестандартной, неполной или ограниченной разметкой, характерной для регуляторных и медицинских документов. Модель ruT5-multitask, дообученная на разнообразных мультизадачных корпусах русского языка, демонстрирует повышенную адаптивность к извлечению структурированной информации из неструктурированных текстов, что даёт возможность оценить вклад мультизадачного предобучения в точность распознавания медицинских сущностей.

Описание инструмента для ручной разметки данных экспертами

Для проведения медицинскими специалистами ручной разметки использовали Doccano, созданный командой разработчиков из Японии [35], — интуитивно понятный, настраиваемый и масштабируемый веб-интерфейс, который широко применяется для разметки данных в таких задачах, как NER [36].

Для определения иерархии именованных сущностей в системе будущего графа были выделены такие семантические группы, как «фармакологическая группа», «название препарата», «механизм действия», «реализуемый эффект», «побочное действие», «метаболизм», «распределение», «выведение» и «рекомендации». После определения принадлежности слова или фразы к конкретному токену вручную устанавливалась связь между ними.

Для примера взят текст из инструкции по применению препарата, действующим веществом которого является периндоприл (Рисунок 4).

Далее медицинскими специалистами были указаны последовательные связи, между выделенными фрагментами (Рисунок 5).

Обучение моделей

До начала обучения модели для NER специалистами были настроены гиперпараметры (Таблица 2), управляющие его процессом. От правильного подбора этих параметров зависит, насколько хорошо модель обучится и насколько быстро она достигнет приемлемого качества распознавания, а также будет ли она обобщать знания на новых данных. Оптимальные значения этих гиперпараметров были определены экспериментально авторами на основании метрик качества, что позволило достичь наибольшей эффективности модели.

Для оценки качества работы модели использовались метрики accuracy, precision, recall и F-score [37].

Периндоприл (название препарата) представляет собой ингибитор АПФ (фармакологическая группа) - фермента, превращающего ангиотензин I в ангиотензин II. АПФ (кининаза) является экзопептидазой, которая осуществляет как превращение ангиотензина I в сосудосуживающее вещество ангиотензин II, так и разрушение брадикинина, обладающего сосудорасширяющим действием, до неактивного гептапептида. Подавление АПФ (механизм действия) приводит к снижению содержания ангиотензина II (механизм действия) в плазме крови, в результате чего повышается активность ренина (механизм действия) в плазме крови (вследствие угнетения отрицательной обратной связи, которая препятствует высвобождению ренина) и снижается секреция альдостерона (механизм действия). Поскольку АПФ инактивирует брадикинин, подавление АПФ сопровождается повышением активности как циркулирующей, так и тканевой калликреин-кининовой системы, при этом активируется система простагландинов. Периндоприл (название препарата) уменьшает общее периферическое сосудистое сопротивление (ОПСС) (реализуемый эффект), что приводит к снижению артериального давления (АД) (реализуемый эффект), при этом периферический кровоток ускоряется без изменения частоты сердечных сокращений (ЧСС).

Рисунок 4 – Пример разметки текста медицинским специалистом с выделением семантических групп.

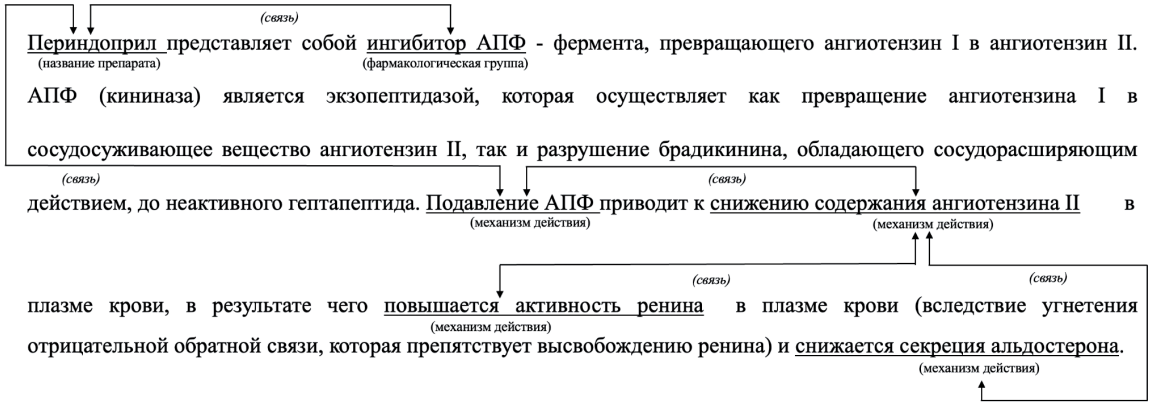

Рисунок 5 – Пример построения смысловых связей между выделенными семантическими группами.

Таблица 2 – Гиперпараметры, использующиеся для обучения модели

Наименование гиперпараметра	Значение	Краткое описание
Скорость обучения (learning rate)	2e-5	Определяет размер шага при обновлении весов модели нейронной сети. Слишком высокий — может пропустить оптимум, слишком низкий — замедляет обучение.
Размер пакета данных для обучения (train batch size)	16	Влияет на стабильность и скорость обучения. Малые пакеты дают шумные градиенты, большие — требуют больше памяти и могут привести к переобучению.
Размер пакета данных для проверки (eval batch size)	16	Аналогично train batch size, но применяется при валидации модели.
Количество эпох (epochs)	10	Полные проходы по обучающим данным. Слишком мало – недообучение, слишком много – переобучение.
Коэффициент регуляризации (weight decay)	0.01	Штрафует большие веса, предотвращая переобучение. Слишком высокий – недообучение, слишком низкий – переобучение.

Для обучения модели использовался корпус текстов, размеченных медицинскими специалистами, который был строго разделён на три непересекающиеся подвыборки: обучающую (47 088 токенов), валидационную (5 886 токенов) и тестовую (5 887 токенов). Оценка качества проводилась исключительно на тестовой выборке, не участвовавшей в обучении и в настройке гиперпараметров.

РЕЗУЛЬТАТЫ

Самый высокий показатель accuracy оказался у RuBioBERT 0,9569±0,0052, сразу за ним был ruBERT 0,9534±0,0053. У моделей ruT5 и ruT5-multitask самые низкие значения этой метрики — 0,7441±0,0111 и 0,7792±0,0106, соответственно. RuBioBERT превзошел ruBERT по метрикам после 8 эпох обучения (Таблицы 2, 3).

Таблица 3 – Сравнение оценок качества работы ruBERT и RuBioBERT

Название модели	№ эпохи	Accuracy	Precision	Recall	F1
rubert	3	0.779528	0.466102	0.48246	0.474138
RuBioBERT	3	0.770284	0.430595	0.44444	0.43741
rubert	4	0.832592	0.528967	0.61404	0.568336
RuBioBERT	4	0.823348	0.55914	0.60819	0.582633
rubert	5	0.888394	0.666667	0.73684	0.7
RuBioBERT	5	0.868538	0.605744	0.67836	0.64
rubert	6	0.925368	0.777188	0.85673	0.815021
RuBioBERT	6	0.912701	0.721622	0.7807	0.75
rubert	7	0.942485	0.865385	0.92105	0.892351
RuBioBERT	7	0.939062	0.814516	0.88597	0.848739
rubert	8	0.949332	0.869919	0.9386	0.902954
RuBioBERT	8	0.951729	0.86921	0.93275	0.899859
rubert	9	0.953783	0.888283	0.95322	0.919605
RuBioBERT	9	0.953441	0.882834	0.94737	0.913963
rubert	10	0.953441	0.887978	0.95029	0.918079
RuBioBERT	10	0.956864	0.892562	0.94737	0.919149

Таблица 4 – Сравнение оценок качества работы ruT5 и ruT5- multitask

Название модели	№ эпохи	Accuracy	Precision	Recall	F1
rut5	3	0.365666	0.034612	0.107981	0.052422
rut5 multitask	3	0.398087	0.03681	0.112676	0.055491
rut5	4	0.41164	0.062689	0.194836	0.094857
rut5 multitask	4	0.702099	0.166234	0.150235	0.15783
rut5	5	0.715387	0.200846	0.223005	0.211346
rut5 multitask	5	0.725485	0.193478	0.20892	0.200903
rut5	6	0.731331	0.215352	0.237089	0.225698
rut5 multitask	6	0.738772	0.2103	0.230047	0.219731
rut5	7	0.735849	0.221766	0.253521	0.236583
rut5 multitask	7	0.749668	0.217039	0.251174	0.232862
rut5	8	0.735318	0.211503	0.267606	0.236269
rut5 multitask	8	0.75392	0.214936	0.276995	0.242051
rut5	9	0.744087	0.231618	0.295775	0.259794
rut5 multitask	9	0.766144	0.234676	0.314554	0.268806
rut5	10	0.744087	0.232558	0.305164	0.263959
rut5 multitask	10	0.779166	0.258467	0.340376	0.29382

Сравнительный анализ архитектур BERT и T5 в задаче автоматической разметки текстов для NER продемонстрировал явное превосходство моделей на основе BERT в условиях решения нашей задачи. Лишь к десятой эпохе обучения модели ruT5 и ruT5 multitask достигали приемлемого уровня производительности. При этом качество их результатов в финальных этапах обучения было сопоставимо с уровнем, достигнутым моделями ruBERT и RuBioBERT на более ранних этапах.

Модели BERT (ruBERT, RuBioBERT) обеспечили качество разметки, близкое к ручной, что подтверждается высокими значениями F1-score, поскольку этот параметр учитывает точное совпадение границ сущностей и их типов в сравнении с размеченными экспертами текстами. В таблице 4 показан наглядный пример автоматической разметки для NER моделями BERT и T5. Она вполне подтверждает значения метрики accuracy, показанные в таблицах 2 и 3.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В данной работе была использована модель, основанная на методе NLP в текстах инструкций по медицинскому применению для последующего составления семантического графа на его основе. Для нашей задачи рассматривались модели на основе архитектуры BERT и архитектуры T5: ruBERT, RuBioBERT, ruT5 и ruT5-multitask, загруженные из хранилища huggingface.

Модель BERT основана на архитектуре трансформера и использует исключительно механизм энкодера части нейронной сети, которая преобразует входные текстовые данные в семантическое представление, удобное для дальнейшей обработки. Его ключевая особенность — двунаправленное внимание, позволяющее учитывать контекст как предшествующих, так и последующих слов в предложении. Это обеспечивает глубокое понимание семантики текста, в связи с чем архитектура BERT особенно эффективна для задач классификации и извлечения информации. Модель ruBERT обучена на общих русскоязычных текстах, RuBioBERT — на медицинских статьях, что делает ее более подходящей для медицинского текста. Модели T5, такие как ruT5 и ruT5-multitask, решают широкий спектр задач NLP в формате «входной текст → выходной текст» и подходят для генеративных задач. Все четыре модели дообучались на инструкциях к ЛС, размеченных по нотации BIO, что позволяет применять их для автоматической разметки текста инструкций по медицинскому применению ЛС. Инструкции к ЛС были размечены по нотации BIO таким образом, чтобы каждому слову в предложении соответствовал свой тег BIO. То есть в результате получается размеченный текст в формате «слово» — «BIO-тег».

Сравнительный анализ ruBERT и RuBioBERT показал, что качество, оцениваемое по метрикам F1-score, точность (precision) и полнота (recall) работы, модели ruBERT выше, чем модели RuBioBERT. Результаты сравнительного анализа показаны в таблице 5. Интересно отметить, что, хотя модель RuBioBERT специально обучена для работы с медицинской терминологией при числе эпох 7, она уступает модели ruBERT по показателям F1-score, точность (precision) и полнота (recall) метрик качества. Вероятно, это связано с тем, что её обучение проводилось на медицинских статьях. Научные статьи сами по себе имеют определённую специфику, а тексты медицинских инструкций к ЛС — еще более узкую специфику. Эта особенность обуславливает различия в закономерностях, характерных для научных и медицинских текстов, по сравнению с общеупотребительными текстами, на которых была обучена модель ruBERT (таблица 3). Вместе с тем из результатов видно, что при увеличении числа эпох до 8 включительно и выше RuBioBERT приближается к ruBERT и немного превосходит его по части показателей.

На основе анализа этих результатов можно сделать вывод, что на ранних эпохах на силу предсказательной способности моделей большее влияние оказывают структура и специфика текстов инструкций для ЛС, особенно если он не очень большой. Из этого следует, что для более быстрого получения убедительного качества работы модели, лучше дообучать модель, предобученную на общеязыковых текстах. Однако при увеличении объёма текстовых данных для обучения, состоящего из корректно размеченных медицинскими специалистами текстов, увеличения эффективности модели от общеязыковых паттернов отмечено не было.

Следовательно, для приближения закономерностей, свойственных текстам инструкций для применения ЛС, больше подходят

Таблица 5 – Сравнение результатов работы rubert-base-cased, RuBioBERT, rut5-base, rut5-base-multitask

RuBERT		RuBioBERT		ruT5-base		ruT5-base-multitask		Мнение эксперта	
[CLS]	O	[CLS]	O	Периндоприл	B-ргераге	Периндоприл	B-ргераге	[CLS]	O
Периндоприл	B-ргераге	Периндоприл	B-ргераге	оказывает	O	оказывает	O	Периндоприл	B-ргераге
оказывает	B-action	оказывает	B-mechanlsm	сосудорасширя-ющее	B-action	сосудорасширя-ющее	B-mechanlsm	оказывает	B-action
сосудорасширя-ющее	I-action	сосудорасширя-ющее	I-mechanism	действие	O	действие	O	сосудорасширя-ющее	I-action
действие	I-action	действие	I-mechanism	,	O	,	O	действие	I-action
,	O	,	O	способствует	B-action	способствует	B-action	,	O
способствует	B-action	способствует	B-mechanism	восстановлению	I-mechanism	восстановлению	I-mechanism	способствует	B-action
восстановлению	I-mechanism	восстановлению	I-mechanism	эластичности	I-action	эластичности	I-mechanism	восстановлению	I-action
эластичности	I-mechanism	эластичности	I-mechanism	крупных	I-action	крупных	I-mechanism	эластичности	I-action
крупных	I-mechanism	крупных	I-mechanism	артерий	O	артерий	O	крупных	I-action
артерий	I-action	артерий	I-mechanism	и	O	и	O	артерий	I-action
и	I- mechanism	и	I-mechanism	структуры	I-mechanism	структуры	I-mechanism	и	O
структуры	I-mechanism	структуры	I-mechanism	сосудистой	I-mechanism	сосудистой	I-mechanism	структуры	I-action
сосудистой	I-mechanism	сосудистой	I-mechanism	стенки	I-mechanism	стенки	I-mechanism	сосудистой	I-action
стенки	I-mechanism	стенки	I-mechanism	мелких	I-action	мелких	I-mechanism	стенки	I-action
мелких	I-mechanism	мелких	I-mechanism	артерий	O	артерий	O	мелких	I-action
артерий	I-action	артерий	I-mechanism	,	O	,	O	артерий	I-action
,	O	,	O	а	O	а	o	,	O
а	O	а	O	также	O	также	O	а	O
также	O	также	O	уменьшает	B-action	уменьшает	B-action	также	O
уменьшает	B-mechanism	уменьшает	B-mechanism	гипертрофию	I-action	гипертрофию	I-mechanism	уменьшает	B-action
гипертрофию	I-action	гипертрофию	I-mechanism	левого	I-action	левого	I-action	гипертрофию	I-action
левого	I-action	левого	I-mechanism	желудочка	I-action	желудочка	I-action	левого	I-action
желудочка	I-action	желудочка	I-mechanism	.	O	.	O	желудочка	I-action
.	O	.	O	Одновременное	O	Одновременное	O	.	O

RuBERT		RuBioBERT		ruT5-base		ruT5-base-multitask		Мнение эксперта	
[CLS]	O	[CLS]	O	Периндоприл	B-ргерагe	Периндоприл	B-ргерагe	[CLS]	O
Одновременное	B-action	Одновременное	O	применение	O	применение	O	Одновременное	O
применение	I-action	применение	O	тиазидных	B-group	тиазидных	B-group	применение	B-action
тиазидных	B-group	тиазидных	B-group	диуретиков	I-group	диуретиков	I-group	тиазидных	B-group
диуретиков	I-group	диуретиков	I-group	group	I-group	group	I-group	диуретиков	I-group
group	O	group	O	усиливает	B-action	усиливает	B-action	group	O
усиливает	B-action	усиливает	B-action	выраженность	O	выраженность	O	усиливает	B-action
выраженность	I-action	выраженность	I-action	антигипертензи-вного	I-action	антигипертензи-вного	I-action	выраженность	I-action
антигипертен-зивного	I-action	антигипертен-зивного	I-action	эффекта.	I-action	эффекта.	I-action	антигипертензив-ного	I-action
эффекта	I-action	эффекта	I-action	Сердечная	B-illness	Сердечная	B-illness	эффекта	I-action
.	O	.	O	недостаточность	I-illness	недостаточность	I-illness	.	O
Сердечная	B-illness	Сердечная	B-illness	Периндоприл	B-prepare	Периндоприл	B-prepare	Сердечная	B-illness
недостаточность	I-illness	недостаточность	I-illness	prepare	O	prepare	O	недостаточность	I-illness
Периндоприл	B-prepare	Периндоприл	B-prepare	нормализует	B-action	нормализует	B-action	Периндоприл	B-prepare
prepare	I-prepare	prepare	O	работу	O	работу	O	prepare	O
нормализует	B-action	нормализует	B-action	сердца	I-action	сердца	I-mechanism	нормализует	B-action
работу	I-action	работу	I-action	,	O	,	O	работу	I-action
сердца	I-action	сердца	I-action	снижая	B-action	снижая	I-action	сердца	I-action
,	O	,	O	преднагрузку	I-action	преднагрузку	I-action	,	O
снижая	B-action	снижая	B-action	и	O	и	O	снижая	B-action
преднагрузку	I-action	преднагрузку	I-action	постнагрузку.	I-action	постнагрузку.	I-mechanism	преднагрузку	I-action
и	I-action	и	I-action	У	O	У	O	и	O
постнагрузку	I-action	постнагрузку	I-action	пациентов	O	пациентов	O	постнагрузку	I-action
.	O	.	O	с	O	с	O	.	O
У	O	У	O	хронической	B-Illness	хронической	B-illness	У	O

закономерности, характерные для общих текстов, поскольку они отражают более универсальные свойства естественного языка. В отличие от специализированных текстов эти общие закономерности применимы к любым разновидностям языка и дают более точные результаты.

Сравнительный анализ результатов работы моделей показал, что метрики качества однозадачных моделей ruT5 ниже, чем у многозадачной модели ruT5 (таблица 4). Сравнительный анализ архитектур BERT и T5 в задаче автоматической разметки текстов для NER продемонстрировал явное превосходство моделей на основе BERT в условиях решения нашей задачи. Лишь к десятой эпохе обучения модели ruT5 и ruT5 multitask достигли приемлемого уровня производительности. При этом качество их результатов в финальных этапах обучения было сопоставимо с уровнем, достигнутым моделями ruBERT и RuBioBERT на более ранних этапах.

Самый высокий показатель точности (accuracy) был достигнут моделью RuBioBERT — 0,9569±0,0052, тогда как у модели ruBERT этот показатель составил 0,9534±0,0053. Значения погрешностей получены как половина ширины 95% доверительных интервалов, рассчитанных на основе биномиального распределения для текстовой выборки объёмом N = 5887. Несмотря на несколько более высокую точечную оценку у RuBioBERT, результаты сопоставимы ($p>0,05$, парный t-тест), что не позволяет утверждать о статистически значимом превосходстве одной модели над другой при уровне значимости $\alpha = 0,05$. В то же время обе BERT-подобные модели (RuBioBERT и ruBERT) значительно превосходят архитектуры на основе T5: у ruT5 accuracy 0,7441±0,0111, у ruT5-multitask — 0,7792±0,0106 ($p<0,05$, t-тест для независимых выборок), что подтверждает преимущество подхода на основе токен-классификации (BERT) над генеративным (T5) для данной задачи извлечения именованных сущностей из текстов медицинских инструкций.

Таким образом, разметка текста, выполненная разновидностями модели BERT (ruBERT и RuBioBERT), близка к ручной разметке, выполненной экспертом по предметной области, и может быть использована для автоматической разметки текстов инструкций по медицинскому применению ЛС для NER. Дальнейшим развитием технологии может стать реализация и применение самообучения для регулярного самообновления моделей с целью адаптации к новым паттернам и закономерностям, присутствующим в новых данных. Кроме того, агрессивное обучение может использоваться для адаптации моделей к спорным данным и разрешения неоднозначных ситуаций, когда модель не уверена в правильности своего ответа.

В контексте обработки большого объёма неструктурированных инструкций по медицинскому применению ЛС использование дообученных моделей машинного обучения позволяет автоматически и с высокой точностью извлекать именованные сущности, включая препараты, активные вещества, показания, противопоказания и побочные эффекты. Такой подход не только минимизирует необходимость ручной аннотации, но и обеспечивает масштабируемость процесса предобработки данных, что особенно критично при работе с постоянно расширяющимися корпусами медицинской документации. После автоматического извлечения именованных сущностей осуществляется их интеграция в единую структуру на основе существующей базы знаний. Эта стадия позволяет установить семантические отношения между выделенными элементами, отражая как фармакологические взаимодействия, так и клинические корреляции. В результате формируется единый семантический граф (Рисунок 6), который интегрирует информацию из множества источников и обеспечивает структурированное представление знаний о ЛС и их взаимодействиях.

На основе данного графа строится байесовская сеть, предназначенная для вероятностной оценки рисков возникновения побочных эффектов при полифармакотерапии. Архитектура такой сети позволяет учитывать сложные зависимости между различными факторами, включая фармакокинетические и фармакодинамические параметры, индивидуальные особенности пациента и комбинационные эффекты ЛС. Автоматизация этапа извлечения и структурирования информации повышает эффективность подготовки данных и способствует созданию более надёжных и интерпретируемых моделей для оценки лекарственных рисков в условиях реальной клинической практики.

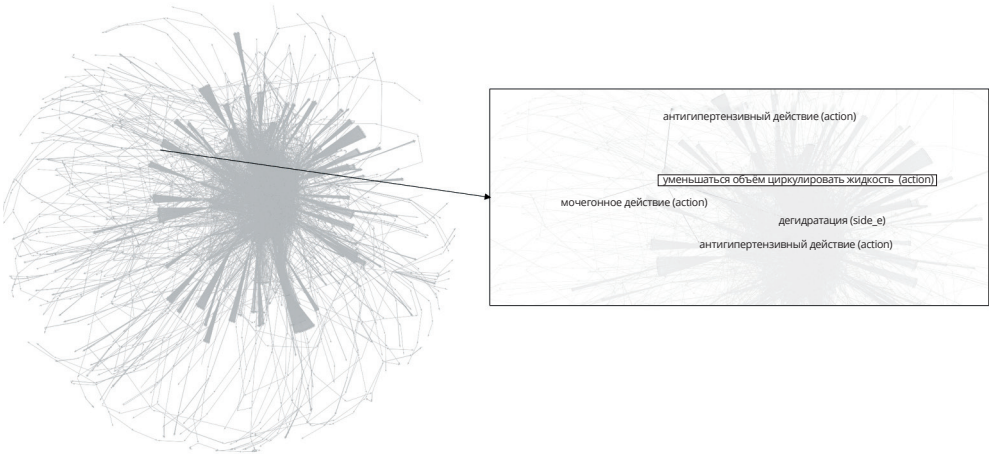

Рисунок 6 – Семантический граф, использующийся для построения сети Байеса.

ЗАКЛЮЧЕНИЕ

В исследовании описана разработка автоматизированного анализа текстов медицинских инструкций, которая является одним из этапов для создания платформы анализа рисков проявления побочных эффектов при полифармакотерапии. Модели на основе архитектуры BERT, в частности ruBERT и RuBioBERT, достигают высокой точности (accuracy 0,9534±0,0053 и 0,9569±0,0052 соответственно) при NER, демонстрируя результаты, сопоставимые с экспертной оценкой. В то же время модели семейства T5 показали более низкую эффективность (accuracy 0,7441±0,0111 и 0,7792±0,0106 соответственно), особенно на начальных стадиях обучения. Результаты проведённого исследования подтверждают, что применение методов автоматической разметки текстов позволяет существенно сократить временные и ресурсные затраты на подготовку обучающих данных, необходимых для построения семантического графа, выступающего в качестве основы функционирования байесовской сети.

Источники финансирования

Исследование выполнено коллективом авторов за счет гранта Российского научного фонда № 23-75-30012.

Конфликт интересов

Авторы подтверждают отсутствие конфликта интересов.

ЛИТЕРАТУРА/REFERENCES

1. Королева М.В., Ильницкий А.Н., Кудашкина Е.В., Коршун Е.И., Шарова А.А., Полев А.В. Современные направления фармакотерапии гериатрических пациентов: полиморбидность — полипрагмазия — депрескрайбинг // Современные проблемы здравоохранения и медицинской статистики. — 2019. — №3. — С.150-171. [Koroleva MV, Ilinitskiy AN, Kudashkina EV, Korshun EI, Sharova AA, Polev AV. Sovremennye napravleniya farmakoterapii geriatricheskikh pacientov: polimorbidnost' — polipragmaziya — depreskraibing. Modern directions in pharmacotherapy for geriatric patients: multimorbidity — polypharmacy — deprescribing. Sovremennye problemy zdorov'ya i meditsinskoy statistiki. 2019; 3: 150-171. (In Russ.)]
2. Cucinotta D. Multimorbidity and polypharmacy: a risk factor for older patients. Acta Biomed. 2022; 93(2): e2022137.
3. Сычев Д.А., Отделенов В.А., Краснова Н.М., Ильина Е.С. Полипрагмазия: взгляд клинического фармаколога // Терапевтический архив. — 2016. — №12. — С.94-102. [Sychev DA, Otelenov VA, Krasnova NM, Ilina ES. Polipragmaziya: vzglyad klinicheskogo farmakologa. Polypharmacy: a clinical pharmacologist's perspective. Terapevticheskii arkhiv. 2016; 12: 94-102. (In Russ.)]

4. Изможерова Н.В., Попов А.А., Курындина А.А., Гаврилова Е.И., Шамбатов М.А., Бахтин В.М. Полиморбидность и полипрагмазия у пациентов высокого и очень высокого сердечно-сосудистого риска // Рациональная фармакотерапия в кардиологии. — 2022. —№1. — С.20-26. [Izmozherova NV, Popov AA, Kuryndina AA, Gavrilova EI, Shambatov MA, Bakhtin VM. Polimorbidnost' i polipragmaziya u pacientov vysokogo i ochen' vysokogo serdechno-sosudistogo riska. Multimorbidity and polypharmacy in patients with high and very high cardiovascular risk. Ratsional'naya farmakoterapiya v kardiologii. 2022; 1: 20-26. (In Russ.)]
5. drugs.com. Available at: https://www.drugs.com. Accessed 07.06.2025.
6. medscape.com. Available at: https://www.medscape.com. Accessed 05.07.2025.
7. DrugBank: Drug Interaction Checker Available at: https://www.drugbank.com/clinical/drug_drug_interaction_checker. Accessed 05.07.2025.
8. WebMD. Available at: https://www.webmd.com. Accessed 06.07.2025.
9. Murali K, Kaur S, Prakash A, Medhi B. Artificial intelligence in pharmacovigilance: Practical utility. Indian J Pharmacol. 2019; 51(6): 373-376.
10. Li Z, Wu W, Kang H. Machine Learning-Driven Metabolic Syndrome Prediction: An International Cohort Validation Study. Healthcare. 2024; 12(24): 2527.
11. Rodrigues PP, Ferreira-Santos D, Silva A, Polónia J, Ribeiro-Vaz I. Causality assessment of adverse drug reaction reports using an expert-defined Bayesian network. Artif. Intell. 2018; 91: 12-22.
12. Wang L, Hao H, Yan X, et al. From biomedical knowledge graph construction to semantic querying: a comprehensive approach. Sci Rep. 2025; 15: 8523.
13. Polotskaya K, Muñoz-Valencia CS, Rabasa A, Quesada-Rico JA, Orozco-Beltrán D, Barber X. Bayesian Networks for the Diagnosis and Prognosis of Diseases: A Scoping Review. Machine Learning and Knowledge Extraction. 2024; 6(2): 1243-1262.
14. Arora P, Boyne D, Slater JJ, Gupta A, Brenner DR, Druzdzel MJ. Bayesian networks for risk prediction using real-world data: a tool for precision medicine. Value Health. 2019; 22(4): 439-445.
15. Sarem M, Saeed F, Alsaeedi A, et al. A survey on text preprocessing techniques based on NLP tasks. Multimedia Tools and Applications. 2022; 81(20): 28443-28470.
16. Единая государственная система регистрации лекарственных средств и медицинских изделий. [доступ от 07.07.2025]. Доступ по ссылке: https://grls.rosminzdrav.ru/Default.aspx.
17. Бойцов С.А., Бубнова М.Г., Васюк Ю.А. и др. Хроническая сердечная недостаточность. Клинические рекомендации 2024 // Российский кардиологический журнал. — 2024. — №29(11). — С.6162. [Boytsov SA, Bubnova MG, Vasyuk YA, et al. Khronicheskaya serdechnaya nedostatochnost'. Klinicheskie rekomendatsii 2024. Rossiiskii kardiologicheskii zhurnal. 2024; 29(11): 6162. (In Russ.)]
18. Международная классификация болезней 10-го пересмотра (МКБ-10) [интернет]. [доступ от 07.07.2025]. Доступ по ссылке: https://mkb-10.com.
19. export_from_pdf.py. GitHub. Available at: https://github.com/kalengul/graph_model_midicine/blob/main/doccano/data_jsonl_export/export_from_pdf.py — Emhyr713 / graph_model_midicine. Accessed 30.06.2025.
20. Finding. GitHub Available at: https://github.com/kalengul/graph_model_midicine/tree/main/semantic_graph/finding. — Emhyr713 / graph_model_midicine. Accessed 30.06.2025.
21. Tabassum A, Patil DR. A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. IRJET. 2020; 7(6).
22. Python 3.11 [интернет]. Версия 3.11 [доступ от 08.07.2025]. Доступ по ссылке: https://docs.python.org/3.11/

23. Кукушкин А.Е. Razdel — сегментация русскоязычного текста на токены и предложения [интернет]. [доступ от 11.07.2025]. Доступ по ссылке: https://github.com/natasha/razdel

24. Alshammari N, Alanazi S. The impact of using different annotation schemes on named entity recognition. Egypt Inform J. 2021; 22: 295-302.

25. Creater_graph. GitHub. Available at: https://github.com/kalengul/graph_model_midicine/blob/main/Creater_graph/mark_BIO_spacy.py . — Emhyr713 / graph_model_midicine. Accessed 30.06.2025.

26. Библиотека spaCy [интернет]. [доступ от 10.07.2025]. Доступ по ссылке: https://github.com/explosion/spaCy.

27. Кукушкин А.Е. Razdel — сегментация русскоязычного текста на токены и предложения [интернет]. [доступ от 11.07.2025]. Доступ по ссылке: https://github.com/natasha/razdel.

28. mark_sent_1.py. GitHub. Available at: https://github.com/kalengul/graph_model_midicine/blob/main/semantic_graph/full_pipeline/mark_sent_1.py . — Emhyr713 / graph_model_midicine. Accessed 30.06.2025.

29. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, eds. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis, MN: Association for Computational Linguistics; 2019: 4171-4186.

30. Kuratov Y, Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language. arXiv preprint arXiv: 1905.07213. 2019.

31. Bobrova EV, Makanov AZ, Osnovin SS, et al. Generating medical opinions and classification according to Bethesda using deep learning. Int J Open Inf Technol. 2023; 11(10): 119-129.

32. Ni J, Hernandez Abrego G, Constant N, Ma J, Hall K, Cer D, Yang Y. Sentence-T5: scalable sentence encoders from pre-trained text-to-text models. In: Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics; 2022: 1864-1874.

33. ruT5-base: модель для генерации текста на русском языке [интернет]. [доступ от 12.07.2025]. Доступ по ссылке: https://huggingface.co/ai-forever/ruT5-base.

34. ruT5-base-multitask: модель для выполнения нескольких текстовых задач на русском языке [интернет]. [доступ от 14.07.2025]. Доступ по ссылке: https://huggingface.co/cointegrated/rut5-base-multitask.

35. Doccano: инструмент с открытым исходным кодом для аннотации текстовых данных в задачах NLP [интернет]. [доступ от 14.07.2025]. Доступ по ссылке: https://github.com/doccano/doccano.

36. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticæ Investigationes. 2007; 30(1): 3-26.

37. Powers D. Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation. Machine Learning Technology. 2008; 2: 2229-3981.