Т.В. Зарубина, С.Е. Раузина, П.А. Астанин, Ю.И. Королева, Л.В. Ронжин, А.А. Борисов, М.А. Афанасьева, А.В. Усова

Институт цифровой трансформации медицины Российского национального исследовательского медицинского университет имени Н.И. Пирогова, Москва, Российская Федерация

Создание базы медицинских знаний на основе национального метатезауруса для унификации разработки систем поддержки принятия клинических решений

Обоснование. *Стремительный рост объемов медицинских данных, широкие возможности информационных технологий, перевод медицинского документооборота в электронный формат порождают спрос на внедрение инструментов информационно-справочной помощи и систем поддержки принятия клинических решений (СППКР). Работа над СППКР в настоящее время объединяет экспертную деятельность врачей, специалистов в области информационных технологий, математической статистики, машинного обучения, инженеров по знаниям. Большинство разработок, предполагающих формирование баз знаний, создается изолированно, без использования универсальных подходов, позволяющих объединять различные решения. В основе любой базы медицинских знаний (БМЗ) лежит тезаурус, представляющий собой систематизированный словарь терминов, который позволяет стандартизировать терминологию и таким образом ускорить поиск и обмен информацией. Он включает в себя термины-концепты и связи между ними, а также синонимы и различные атрибуты.* ***Цель исследования*** *— создание национального медицинского метатезауруса, построенного по онтологическому принципу, и разработка на его основе БМЗ.* ***Методы.*** *Международный систематизированный словарь медицинских терминов UMLS (Unified Medical Language System); клинические рекомендации по 22 группам нозологий; справочники федерального портала нормативно-справочной информации Минздрава России; электронные медицинские карты – 330 тыс. (dataset MIMIC-IV); абстракты публикаций PubMed – 28 млн. Использовались семантические анализаторы SemRep (Semantic Repository) и MetaMap; методы оценки лексической схожести, связности, контекстной сочетаемости сущностей в подграфе, математической статистики.* ***Результаты****. Создана первая версия унифицированной национальной медицинской номенклатуры (УНМН). Показано, что онтологические модели являются эффективным способом представления структурированной информации. Созданы компоненты информационно-поисковых систем. Разработаны аналитические инструменты для работы с метатезаурусом.* ***Заключение.*** *На основе УНМН и созданных инструментов возможны автоматизированное формирование образа заболевания (базы знаний) и одноплатформенная разработка СППКР.*

Ключевые слова*: метатезаурус, база медицинских знаний, системы поддержки принятия клинических решений, медицинская онтология, семантический анализ медицинской информации*

Для цитирования*:* Зарубина Т.В., Раузина С.Е., Астанин П.А., Королева Ю.И., Ронжин Л.В., Борисов А.А., Афанасьева М.А., Усова А.В. Создание базы медицинских знаний на основе национального метатезауруса для унификации разработки систем поддержки принятия клинических решений. *Вестник РАМН.* 2024;79(2):175–192. doi: https://doi.org/10.15690/vramn17390

Проблематика

В последние годы наблюдается быстрое развитие рынка систем искусственного интеллекта, обусловленное в том числе финансированием со стороны государства, с чем связано создание новых подходов к разработке систем поддержки принятия клинических решений (СППКР). Создание современных систем искусственного интеллекта — междисциплинарное направление, в большинстве случаев нацеленное на анализ больших данных, интегрирующее для своего построения инструменты реляционной алгебры, элементы теории графов, методы математической статистики, инженерии знаний, алгоритмы машинного обучения. Наиболее развитой областью искусственного интеллекта, достигшей реального внедрения в клиническую практику, является обработка медицинских изображений [1–8], успешности развития которой способствовало наличие общепринятого стандарта обмена изображениями DICOM.

Ключевыми проблемами, препятствующими развитию современных технологий искусственного интеллекта в других областях медицины, в том числе получению сводных данных для анализа, являются формирование подавляющей доли клинической информации в неструктурированном виде, отсутствие общепринятой формализации при описании клинической картины пациента, использование единой медицинской терминологии по ключевым понятиям.

Решением данных проблем являются, с одной стороны, разработка структурированных электронных медицинских документов (СЭМД) на основе международных общепринятых стандартов, позволяющая исходно получать сопоставимую информацию, а с другой — совершенствование алгоритмов обработки неструктурированных медицинских текстов. Оба направления важны для развития цифровой трансформации медицины, являются трудо- и времязатратными, требуют привлечения специалистов по разработки унифицированной медицинской терминологии. Тезаурусы позволяют стандартизировать терминологию, использованную в медицинских записях, что значительно улучшает качество поиска и точность результатов, уменьшает возможность ошибок в интерпретации запросов пользователей в медицинских информационно-поисковых системах. Также тезаурусы служат основой построения баз знаний при решении интеллектуальных задач.

DOI: https://doi.org/10.15690/vramn17390

В настоящее время существует несколько организованных по онтологическому принципу масштабных метатезаурусов, которые контролируются экспертными сообществами, регулярно актуализируются и хранят накопленный теоретический и практический опыт из различных областей биологии и медицины. Например, наиболее используемым ресурсом в англоязычной среде являются UMLS (Unified Medical Language System) [9] — унифицированная медицинская языковая система, разрабатываемая Национальной медицинской библиотекой США с 1986 г., которая содержит ~ 4,6 млн уникальных понятий (концептов), 98 млн связей, и SemMed (база знаний библиотеки Pubmed) — ~ 420 тыс. понятий, 110 млн связей [10]. В свою очередь, UMLS агрегирует такие известные терминологические системы, как SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms — систематизированная номенклатура клинических терминов) — свыше 350 тыс. понятий, 1,5 млн связей; LOINC (Logical Observation Identifiers Names and Codes — кодификатор медицинских и лабораторных терминов) — ~ 218 тыс. понятий, 3,9 млн связей; RxNorm (медицинская терминология в области лекарственных средств) — ~ 170 тыс. понятий, 1,7 млн лекарственных взаимодействий и др. (около 150 различных справочников). Среди отечественных наиболее крупных примеров можно назвать Объединенную базу медицинских знаний UMKB (United Medical Knowledge Base) с более чем 2,5 млн терминов (без учета синонимов), более 4 млн родовидовых отношений, формирующих структуру классификаторов [11, 12], а также онтологическую базу медицинской терминологии и наблюдений, сформированную на платформе IACPaaS [13–15]. Перечисленные тезаурусы используются для создания единообразного языка при описании медицинских и биологических понятий, включая заболевания, симптомы, лекарства и т.д.

Большинство международных справочников исходно не представлено на русском языке. Адаптация исходя из количества терминов и связей трудоемка и потребует значительных временных затрат, привлечения экспертов, сопоставления и расширения на основе русскоязычных справочных источников, представленных в Федеральном реестре нормативно-справочной информации (ФР НСИ) Минздрава России. Также создание национального медицинского терминологического стандарта потребует учета национальной специфики медицины, основанной на данных из клинических рекомендаций, статей и записей реальной клинической практики — электронных медицинских карт пациентов. Несмотря на большую ресурсоемкость, данная разработка оправдана благодаря возможности приобретения единого систематизированного свода экспертно выверенных формулировок специализированных понятий для терминологического покрытия различных клинических областей.

Подавляющая часть медицинской документации продолжает формироваться в неструктурированном виде, что требует разработки и применения технологий выделения важных для последующего использования данных на основе обработки естественного языка (natural language processing, NLP). Обработка неструктурированной медицинской информации методами NLP — одна из наиболее сложных областей семантического анализа, целью которого является извлечение и последующее использование закономерностей, отражающих смысловое значение различных единиц языка. В последние годы развитие мето-

T.V. Zarubina, S.E. Rauzina, P.A. Astanin, Yu.I. Koroleva, L.V. Ronzhin, A.A. Borisov, M.A. Afanasyeva, A.V. Usova

Healthcare Digital Transformation Institute of the Pirogov Russian National Research Medical University (Pirogov Medical University), Moscow, Russian Federation

Creation of a Medical Knowledge Base for Unify the Development of Clinical Decision Support Systems Based on the National Metathesaurus

Background: *The rapid growth in the volume of medical data, the extensive possibilities of information technology, the transfer of medical document flow to electronic format generates a high demand for the introduction of information and reference assistance tools and clinical decision support systems (CDSS). Work on the creation of CDSS currently combines the expert activities of doctors with the work of information technology specialists, mathematical statisticians, data scientists, knowledge engineers. Most of the developments involving the formation of knowledge bases are created in isolation, without the use of universal approaches that allow combining various solutions. At the heart of any medical knowledge base (MKB) there is a thesaurus, which is a systematized dictionary of terms that helps to standardize terminology, which makes it possible to speed up the search and exchange of information. It includes concept terms and relationships between them, as well as synonyms and various attributes.* ***Aims*** *— creation of a national medical metathesaurus, built on the ontological principle and the development of MKB based on it.* ***Methods.*** *International systematized dictionary of medical terms UMLS (Unified Medical Language System); clinical recommendations for 22 groups of nosologies; reference books of the federal portal of normative reference information of the Ministry of Health of the Russian Federation; electronic medical records – 330 thousand (dataset MIMIC-IV); abstracts of PubMed publications-28 million. Semantic analyzers SemRep (Semantic Repository) and MetaMap were used; methods for evaluating lexical similarity, connectivity, contextual combinability of entities in a subgraph, and mathematical statistics.* ***Results.*** *The first version of the Unified National Medical Nomenclature (UNMN) has been created. It is proved that ontological models are an effective way of presenting structured information. Components of information search engines have been created. Analytical tools for working with metathesaurus have been developed.* ***Conclusions.*** *On the basis of the UNMN and the created tools, it is possible to automate the formation of a clinical picture of the disease (knowledge base) and single-platform development of the CDSS.*

Keywords*: metathesaurus, knowledge base, clinical decision support systems, biological ontologies, semantics analysis of health information*

For citation*:* Zarubina TV, Rauzina SE, Astanin PA, Koroleva YuI, Ronzhin LV, Borisov AA, Afanasyeva MA, Usova AV. Creation of a Medical Knowledge Base for Unify the Development of Clinical Decision Support Systems Based on the National Metathesaurus. *Annals of the Russian Academy of Medical Sciences*. 2024;79(2):175–192. doi: https://doi.org/10.15690/vramn17390

дов семантического анализа мощно ускорилось благодаря значительному увеличению вычислительных возможностей, появлению продвинутых машинных алгоритмов, автоматизации процессов обработки текста [16, 17].

Получаемая из медицинских текстов формализованная информация подлежит стандартизации и унификации для последующего использования и анализа по разным направлениям — от сопоставления различных ее источников, контроля и оценки проводимых лечебно-диагностических мероприятий, формирования сводных данных и Big Data до использования в информационно-поисковых средствах и иных решениях для интеллектуального сопровождения при работе с медицинскими и биологическими знаниями.

Применение систематизированных сводов понятий для работы языковых моделей, а также заключений, полученных в ходе экспертной деятельности специалистов, может стать одним из лучших подходов к структуризации данных и извлечению знаний из неструктурированных текстов. Сопоставление источников информации должно не только учитывать различные семантические основы (терминологии и классификации), но и предвидеть дальнейшую модификацию этой семантики в соответствии с местным контекстом [18–20].

Наиболее эффективным способом представления и анализа формализованных медицинских знаний являются семантические сети, организованные в виде множества разнородных вершин (концептов) и связей между ними. Организация данных в виде графовых информационных моделей имеет ряд преимуществ, среди которых следует выделить наличие больших возможностей для оптимизации аналитических операций и существование средств, обеспечивающих наглядную интерпретацию структуры знаний на пользовательском уровне [21, 22].

Автоматизированный анализ семантических сетей — сложная и нетривиальная задача. Несмотря на отсутствие универсальных подходов, существует большое количество отдельных независимых алгоритмов, позволяющих выделять скрытые аналитические паттерны при реализации задач выбора релевантной информации. Анализ исследований показывает высокий интерес и перспективность данного направления [23, 24]. Наиболее эффективное извлечение данных из графовой структуры достигается при комбинации разнородных аналитических алгоритмов, позволяющих учитывать не только конструкцию онтологического древа (смысловые связи), но и семантическую близость терминов, заключенных в ее узлах [25–27].

Модель данных, построенную на основе тезауруса и использующую различные метрики значимости и связности в качестве атрибутов, можно назвать онтологической. Такая модель служит платформой для организации баз знаний, последовательно развиваемых и расширяемых по различным клиническим направлениям. Этот подход подразумевает участие экспертов для валидации качества базы знаний на более поздних сроках и в условиях значительной предварительной подготовки, что оправдано с точки зрения временных затрат на создание системы и использования интеллектуальных ресурсов.

Цель исследования — построение по онтологическому принципу национального медицинского метатезауруса и создание на его основе базы медицинских знаний для унификации разработок интеллектуального поиска медицинской информации и систем поддержки принятия клинических решений.

Методы

Исследование выполнено в рамках Программы стратегического академического лидерства «Приоритет–2030» на базе Института цифровой трансформации медицины РНИМУ им. Н.И. Пирогова Минздрава России.

В качестве источника для построения онтологически организованной системы медицинских терминов использовался метатезаурус UMLS версии 2022AB [19], исходно содержащий примерно 4,6 млн уникальных концептов и 11,2 млн их различных вариантов названий, составленных из 76 актуальных на данный момент словарей, а также более 90 млн уникальных связей. Причиной выбора данного ресурса явилось отсутствие описания с достаточным уровнем подробности в научных литературных источниках существующих отечественных номенклатур, а также их наличия в открытом доступе для сторонних разработчиков. Остается неясным вопрос, осуществлялось ли в них мапирование терминов с общепризнанными международными системами и с утвержденными Минздравом России для употребления в Российской Федерации справочниками ФР НСИ.

На первом этапе инструментом работы с UMLS стали аналитические панели системы Data Monitor компании «Авикомп Сервисез» (зарегистрирована в реестре отечественного ПО № 5609 от 26.07.2019) [28], позволяющие реализовать полнотекстовый и фасетный поиск по объектам, связям, источникам информации (словарям), а также по всем вариантам описаний медицинских понятий и сущностей (концептов) на английском и русском языках. Разработанные панели поддерживают возможность анализа структуры UMLS за счет отнесения медицинских и биологических концептов к одной из 15 (7 социально-экономических и 8 биомедицинских) категорий, которые, в свою очередь, разделяются на 127 тематических групп: растения, вирусы, бактерии, анатомические структуры, диагнозы и синдромы, симптомы, клинические находки, врожденные аномалии, антибиотики, гормоны, лекарственные препараты и иные семантические объединения терминов. Каждый из представленных в UMLS концептов (concept unified identifier, CUI) и тематических групп (type unique identifier, TUI) имеет уникальный идентификатор. У термина могут существовать синонимы, варианты формулировок на разных языках и аббревиатуры (atomic unified identifier, AUI), полученные из многочисленных словарей. Данные атрибуты однозначно связаны с термином, имеющим предпочтительное наименование (выбранное составителями UMLS).

Помимо обширной системы классификации терминов присутствует система классификации существующих связей между концептами по представленным типам и их атрибутам. Всего в семантической сети UMLS выделено 11 (9 базовых и 2 дополнительных) типов и 992 уникальных необязательных подтипа (уточнения) связей. Кроме того, аналитические панели позволяют применять сложную фильтрацию концептов по источникам знаний (справочникам) с использованием логических операторов конъюнкции («И»), дизъюнкции («ИЛИ») и дополнения («НЕ»).

Создание инструментов анализа информации осуществлялось на основе реляционного и графового представления метатезауруса UMLS. В процессе разработки использованы современные системы управления базами данных (СУБД): колоночная СУБД ClickHouse, объектно-реляционная СУБД PostgreSQL и графовая СУБД Neo4j. Реализация интерфейсных решений и интеграция

процессов взаимодействия с перечисленными СУБД проводились с применением программных платформ Django и Flask, а также специальных аналитических библиотек языка программирования Python 3.10. Для адаптации терминологии UMLS на русский язык использовались справочники Федерального реестра нормативно-справочной информации Минздрава России [29]. Для получения дополнительных прямых связей между концептами метатезауруса задействовался репозиторий семантических предикатов SemMedDB (Semantic Medline Database) — троек «субъект–предикат–объект», извлеченных из набора абстрактов библиотеки PubMed [30]. Для тех же целей были развернуты семантические анализаторы SemRep (Semantic Repository) и MetaMap, разработанные Национальной библиотекой США и находящиеся в открытом доступе [31], с помощью которых обрабатывались данные реальной клинической практики (dataset MIMIC-IV, содержащий более 330 тыс. электронных медицинских документов, каждый из которых имеет связь с кодами МКБ-9 и МКБ-10) [32]. С целью поиска методов семантического анализа, применимых для получения релевантных данных из UMLS, проведен обзор публикаций отечественных и зарубежных исследователей. Выделены три основные группы: 1) методы оценки лексической схожести сущностей (парное сравнение схожих участков наименований с использованием алгоритма локального выравнивания и определением степени близости между ними); 2) методы оценки связности сущностей в подграфе знаний с целью поиска семантически близких понятий из разных клинических областей и 3) методы оценки контекстной сочетаемости сущностей (анализ частоты совместной встречаемости с использованием их векторного представления) [25, 33–36].

Оценка потенциальной релевантности представленных в UMLS симптомов, синдромов, клинико-лабораторных находок осуществлялась по следующим нозологическим формам: ишемическая болезнь сердца (ИБС; I20–I25); острое нарушение мозгового кровообращения (ОНМК; I63, G45); желудочно-кишечное кровотечение (ЖКК; K92.2); рак молочной железы (РМЖ; C50); заболевания верхних и нижних дыхательных путей неопухолевого характера (аллергический ринит — J30; острые респираторные вирусные инфекции у взрослых – J00–J06, J20–J22; бронхиальная астма, хроническая обструктивная болезнь легких — J44; хронический бронхит — J40–J42; эмфизема — J43; внебольничная пневмония — J13–J16, J18); язвенная болезнь желудка (ЯБ желудка; K25) и двенадцатиперстной кишки (ЯБ ДПК; K26) с заболеваниями дифференциального ряда (хронический панкреатит — K86; рак желудка — C16; хронический холецистит — K81.1; хронический гастрит — K29.3, K29.4, K29.5). В качестве опорных критериев использовалась формализованная экспертным способом информация из утвержденных клинических рекомендаций (табл. 1) [37]. Выделенные из клинических рекомендаций фрагменты знаний анализировались на предмет качества представления ме-

Таблица 1. Перечень клинических рекомендаций для первичной валидации алгоритмов анализа графовой информационной модели UMLS

Наименование КР	ID	Год принятия
1. Аллергический ринит	КР261	2020
2. Острый синусит	КР313	2021
3. Острые респираторные вирусные инфекции у взрослых	КР724	2021
4. Хронический тонзиллит	КР683	2021
5. Острый тонзиллит и фарингит	КР306	2021
6. Острый обструктивный ларингит и эпиглоттит	КР352	2021
7. Паратонзиллярный абсцесс	КР664	2021
8. Грипп у взрослых	Проект	2021
9. Бронхиальная астма	КР359	2021
10. Внебольничная пневмония у взрослых	КР654	2021
11. Хроническая обструктивная болезнь легких	КР603	2021
12. Хронический бронхит	КР655	2021
13. Эмфизема легких	КР656	2021
14. Идиопатический легочный фиброз	КР677	2021
15. Новая коронавирусная инфекция (COVID-19)	Врем. МР	2022
16. Стабильная ишемическая болезнь сердца	КР155	2020
17. Ишемический инсульт и транзиторная ишемическая атака у взрослых	КР171	2021
18. Гастрит и дуоденит	КР708	2021
19. Язвенная болезнь	КР277	2020
20. Хронический панкреатит	КР273	2020
21. Рак желудка	КР574	2020
22. Рак молочной железы	КР379	2021

Примечание. КР — клинические рекомендации.

дицинских терминов на русском языке, степени прямой связности искомых концептов в метатезаурусе, полноты смыслового покрытия предметной области, возможности построения базы медицинских знаний по данным клиническим группам.

Создание унифицированной национальной медицинской номенклатуры

Исследование возможности использования международного метатезауруса UMLS при создании унифицированной национальной медицинской номенклатуры

Был осуществлен анализ полноты покрытия концептами UMLS образов заболеваний, построенных на основе формализации клинических рекомендаций в части описания клинической картины, данных анамнеза и факторов риска развития заболеваний 22 групп нозологий (см. табл. 1), раздел клинических рекомендаций «Особенности кодирования заболевания», содержащий формулировки и коды МКБ-10, и подраздел «Жалобы и анамнез». Разработанный словарь включил 579 терминов. Сопоставление с концептами UMLS проводилось экспертно с использованием аналитических панелей и языка графовых запросов Cypher. Примеры сопоставления терминов клинических рекомендаций с концептами UMLS представлены в табл. 2.

Практически все клинически значимые термины для постановки диагноза тем или иным образом были найдены. Большинство из них имеет аналогичное или полностью схожее наименование в UMLS (табл. 3). Концепты, которые имеют частичное совпадение (16%), специфичны. Например, для термина из клинических рекомендаций «усиление кашля в холодное время года» в UMLS были обнаружены концепты «усиление кашля» и «холодное время года». Точный вариант данной формулировки не найден. Причиной может быть специфичное отображение таких симптомов в англоязычной практике. Например, при эмфиземе наблюдается перкуторный

Таблица 2. Соответствие терминов клинических рекомендаций концептам UMLS

Симптом из клинических рекомендаций	Код концепта UMLS (CUI)	Русскоязычный вариант концепта в UMLS
Боли в подложечной области	C0232493	Боль в эпигастрии
Тошнота	C0027497	Тошнота
Лихорадка	C0015967	Лихорадка
Общая слабость	C0746674	Общая мышечная слабость
Недомогание	C0231218	Недомогание
Мышечная боль	C0231528	Миалгия
Суставная боль	C0003862	Артралгия
Головная боль	C0018681	Головная боль
Кашель	C0010200	Кашель
Насморк	C1260880	Ринорея
Чихание	C0037383	Чихание

Таблица 3. Оценка степени покрытия терминов из клинических рекомендаций в части симптомов и факторов риска заболеваний концептами UMLS

Аналитический показатель	Нозологическая форма						
	ОНМК	РМЖ	ИБС	ЖКК	Заболевания дыхательных путей*	ЯБ**	Все
Всего терминов в клинических рекомендациях, *n*	132	27	16	23	209	172	579
Найдено концептов в UMLS, *n* (%) Из них:	108 (82)	20 (74)	16 (100)	21 (91)	209 (100)	150 (87)	524 (91)
с точным совпадением термина	95 (88)	13 (65)	10 (63)	21 (100)	167 (80)	123 (72)	429 (74)
с частичным совпадением термина	13 (12)	7 (35)	6 (37)	0	42 (20)	27 (16)	95 (16)
Связаны напрямую с нозологией (из найденных), *n* (%)	35 (32)	4 (20)	3(19)	11 (52)	36 (17)	0	89 (17)
Не найдено соответствующих концептов в UMLS, *n* (%)	24 (18)	7 (26)	0	2 (9)	0	22 (13)	55 (9)

* Неопухолевые заболевания верхних и нижних дыхательных путей (данные без пневмонии).

** Язвенная болезнь желудка и двенадцатиперстной кишки с заболеваниями дифференциального ряда.

Примечание. ОНМК — острое нарушение мозгового кровообращения; РМЖ — рак молочной железы; ИБС — ишемическая болезнь сердца; ЖКК — желудочно-кишечное кровотечение.

коробочный звук, однако аналогичной формулировки в UMLS нет. В ходе детального изучения патологических перкуторных звуков в англоязычной литературе удалось установить, что при повышенной воздушности легочной ткани такой звук называется «Hyperresonance». Данный концепт успешно найден в метатезаурусе, а также обнаружена его связь с эмфиземой. Термины, которые не удалось сопоставить с концептами (9%), ожидаемо также присутствуют в UMLS. Однако ввиду наличия их сложного с медицинской точки зрения наименования, вероятно, иного звучания в англоязычной среде необходимы дополнительные, более глубокие знания предметной области. Трудности сопоставления концептов возникли также с именными клиническими признаками: правосторонний симптом Мюсси–Георгиевского, болезненность в зоне Губергрица–Скульского, симптом Тужилина и т.д.

Важно отметить, что в метатезаурусе UMLS при контекстном поиске термина или его синонима для многих часто встречающихся медицинских понятий присутствует не один концепт. Так, термин «хрипы» может присутствовать и в семантической группе T184 «Симптомы и признаки» — C0043144 («свистящее дыхание, хрипы»), и в семантической группе T033 «Клинические находки» — C0035508 («хрипы при аускультации»), которая указывает на обнаружение признака в процессе медицинского осмотра пациента. Аналогичный пример можно привести с термином «анемия», который может иметь место в качестве клинического признака (T033), профиля лабораторного исследования (T059) и его результата, синдрома (T047) — C0475143, C4554633, C2210818, C0002871. В подобных контекстах этот термин имеет несколько разное смысловое значение, и, соответственно, с ними будут связаны отличные медицинские термины. Таким образом, при работе с тезаурусом очень важно не только рассматривать совпадения по названию, но и всегда учитывать семантическую составляющую поиска. Различные варианты звучания термина также хорошо представлены в UMLS, их может быть до нескольких десятков вариантов. Здесь речь идет не о комбинаторике или морфемах, а именно о различных названиях одного и того же. Например, «насморк», «шмыгать носом», «повышенная носовая секреция», «выделения из носа» — варианты звучания одного и того же концепта с предпочтительным названием «ринорея» (C1260880).

Многие концепты имеют уточненные разновидности, важные для конкретного заболевания и, соответственно, его дифференциальной диагностики. Так, кашель (C0010200) в зависимости от наличия мокроты может быть продуктивный (C0239134) и непродуктивный (C0850149); от времени суток — утренний (C0240351) и ночной (C0231912); от условий — в покое (C0231914), при физической нагрузке (C0586750) и после еды (C0231916); по длительности — острый (C0742857), персистирующий (C0562483) и хронический (C0010201).

Таким образом, сопоставление терминов проводилось с учетом всех возможных особенностей. Результаты сопоставления, представленные в табл. 3, позволяют предположить, что метатезаурус UMLS может быть использован для создания унифицированной национальной терминологической базы (найдено более 90% искомых терминов). Однако процесс реализации данной задачи сопровождается рядом проблем, требующих разработки средств для их решения.

Первая проблема заключается в сложности поиска концептов с использованием автоматических алгоритмов анализа неструктурированных текстов из-за низкой доли русскоязычных понятий. Исследуемая версия UMLS содержит только 304 тыс. исходных русских переводов, что составляет 2,87% общего объема актуальных терминов (11,2 млн).

Вторая проблема состоит в наличии выраженной неравномерности структуры семантической сети UMLS. Число прямых связей между концептами, отражающее сложность структуры внутри исследуемой подобласти знаний, колеблется от 1 до 91 730. Подобная неоднородность структуры знаний приводит к смещению любых математических оценок при попытке анализа связности концептов и снижению релевантности получаемой информации при автоматизированной обработке. Это требует разработки специализированных взвешенных аналитических метрик, учитывающих плотность структуры связей вокруг исследуемого концепта.

Третья проблема связана с недостатком прямых связей между концептами-заболеваниями и концептами-симптомами (17%). Это порождает необходимость создания и реализации математических алгоритмов поиска непрямых путей между корневыми и концевыми узлами семантической сети. Данная закономерность согласуется с результатами проведенных ранее исследований [38]. Также возможным решением этой задачи может стать создание прямых связей на основе анализа медицинских текстов и данных реальной клинической практики.

Исходно говорилось о присутствии семантической организованности метатезауруса, однако из-за его масштабности, широкого охвата медицинских направлений, использования для составления словарей (созданных разными разработчиками) в UMLS присутствует масса неточностей и ошибок, в том числе существенных для автоматизации процесса поиска информации. Это говорит еще об одной проблеме и необходимости реализации собственных аналитических решений путем добавления атрибутов и связей.

Вопросы переводов и пути их решения

Первоначально все не представленные на русском языке понятия метатезауруса были переведены при помощи нейросети-трансформера. Результаты показали, что встречалось довольно много грубых ошибок, искажающих смысл термина или некорректно звучащих (табл. 4). Далее были предприняты пробные попытки перевода части понятий при помощи популярных онлайн-переводчиков,

Таблица 4. Примеры некорректных переводов нейросети

CUI	Исходная формулировка	Перевод нейросети	Корректный перевод
C5192202	Tingling pain	Тиннинг боли	Ощущение покалывания / Парестезия
C0001125	Lactic acidosis	Молочная ацидоза	Лактатацидоз / Молочнокислый ацидоз
C5235233	Pigmentary iris degeneration	Пигментная дегенерация ириса	Пигментная дегенерация сетчатки
C5574737	Hand–Schueller–Christian syndrome	Христианский синдром Рук–Шуллера	Болезнь Хенда–Шюллера–Крисчена
C0578896	Osler's node of hand	Узел руки Ослера	Узелок Ослера на руке

таких как Google Translate, Yandex Translate и PROMT. Результат показал, что медицинская лексика является крайне сложной для полной автоматизации процесса и на текущий момент ни один из перечисленных переводчиков не показал удовлетворительное качество. Эта проблема связана с такими особенностями медицинских терминов, как метонимия, терминологическая изменчивость и синонимия, распространенность применения англицизмов и латинизмов, эпонимов, многообразие вариантов сокращений и аббревиатур, различия в организации системы здравоохранения в разных странах [39].

Одним из возможных в настоящее время решений пока остается создание технической версии автоматического перевода, которую далее необходимо уточнять экспертно для групп наиболее клинически значимых терминов (семантические группы концептов «Симптомы и признаки», «Клинические находки», «Заболевания и синдромы»). На данный момент именно таким образом уточнено более 191 тыс. понятий.

Другой важный способ расширения русскоязычной локализации UMLS — сопоставление справочников или отдельных терминов, содержащихся в метатезаурусе, с русскоязычными словарями, представленными в ФР НСИ Минздрава России. Таким образом был полностью мапирован справочник международной классификации болезней МКБ 10-го пересмотра с его английской версией UMLS, что позволило уточнить 16 095 концептов русскоязычных формулировок. Также ряд справочников портала ФР НСИ (табл. 5) содержит в качестве атрибута уникальные идентификаторы таких известных международных тезаурусов, как SNOMED CT, LOINC, RadLex, которые входят в состав UMLS. Это позволяет обоснованно использовать экспертную формулировку из отечественного справочника применительно к концепту из метатезауруса UMLS, имеющему аналогичные атрибуты. На данный момент из Федеральных справочников лабораторных исследований удалось соотнести 6096 (32%) понятий, которые имеют кодировку LOINC; из Федерального справочника инструментальных и диагностических исследований (ФСИДИ) — 476 (28%) понятий, имеющих кодировку LOINC; из справочника анатомических локализаций — 1114 (90%) понятий, имеющих кодировку в SNOMED CT. Следует отметить, что Федеральные справочники лабораторных и диагностических исследований составлены с участием соответствующих экспертных сообществ и покрывают практически весь перечень лабораторных и инструментальных диагностических исследований нашей страны. Потенциальную возможность для сопоставления терминологии имеют следующие справочники: Параметры клинических шкал и опросников (OID 1.2.643.5.1.13.13.11.1515), Международная классификация болезней — Онкология (3-е изд.), Морфологические коды (OID 1.2.643.5.1.13.13.11.1486), Единицы измерения (OID 1.2.643.5.1.13.13.11.1358), Иммунобиологические лекарственные препараты (OID 1.2.643.5.1.13.13.11.1078), Действующие вещества лекарственных препаратов (OID 1.2.643.5.1.13.13.11.1367) и многие другие.

Таким образом удалось добавить в метатезаурус UMLS более 215 тыс. экспертно верифицированных переводов, что вместе с исходно русскоязычными понятиями покрывает большую часть используемой медицинской терминологии в области описания заболеваний и их признаков. Как будет показано далее, оценка клинической значимости концептов метатезауруса дает примерно 1,5 млн потенциально применимых уникальных терминов. Планируется добавление новых экспертных переводов и сопоставление с новыми справочниками. Не исключено, что работа с отечественными справочниками и другими источниками данных приведет к необходимости добавления в метатезаурус новых терминов и связей.

Анализ закономерностей (паттернов) организации метатезауруса для построения базы медицинских знаний

Получение релевантных данных из метатезауруса UMLS на первом шаге было апробировано путем использования исходных закономерностей в семантической организации терминов и связей. Все термины объединены в тематические группы в соответствии с иерархическим принципом организации (рис. 1).

Также имеет значение принцип организации связей между концептами (табл. 6). Они объединены в 11 групп и свыше 980 их подтипов.

Метатезаурус представляет собой ориентированный граф, т.е. связи в нем имеют направленность. Выделяют симметричные связи (SY, SIB, RO, RL и RQ), когда одна и та же связь между концептами присутствует в прямом и обратном отношениях, обратные (CHD и PAR, RB и RN, AQ и QB), когда в одном направлении пара концептов связана противоположными по смыслу отношениями, и однородные (PAR и RB, CHD и RN) отношения (рис. 2).

Выявление паттернов по семантическим типам терминов и связей при составлении образа нозологии в части определения предварительного диагноза осуществлялось на примере язвенной болезни желудка, двенадцатиперстной кишки и их дифференциального ряда (рис. 3), а также заболеваний верхних дыхательных путей с использованием соответствующих клинических

Таблица 5. Перечень мапированных с терминами UMLS справочников Федерального реестра нормативно-справочной информации

Название	Версия	OID (уникальный идентификатор)
1. Федеральный справочник лабораторных исследований. Справочник лабораторных тестов	3.42	1.2.643.5.1.13.13.11.1080
2. Федеральный справочник лабораторных исследований. Профили лабораторных исследований	3.36	1.2.643.5.1.13.13.11.1437
3. Федеральный справочник инструментальных диагностических исследований	2.34	1.2.643.5.1.13.13.11.1471
4. Анатомические локализации	4.10	1.2.643.5.1.13.13.11.1477
5. Выявленные патологии	2.17	1.2.643.5.1.13.13.11.1473
6. Тип патологии	1.1	1.2.643.5.1.13.13.99.2.840

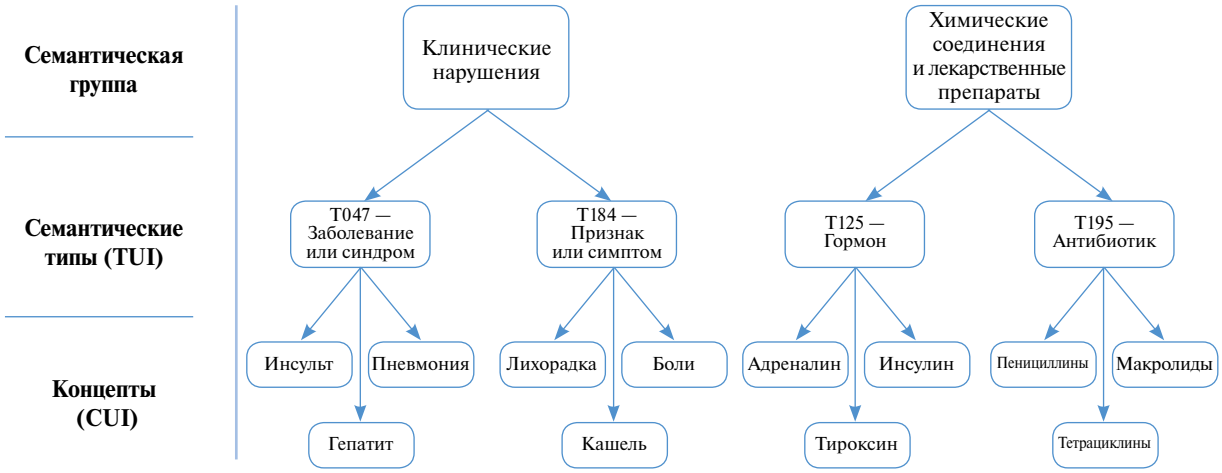

Рис. 1. Организация тематических групп терминов в UMLS

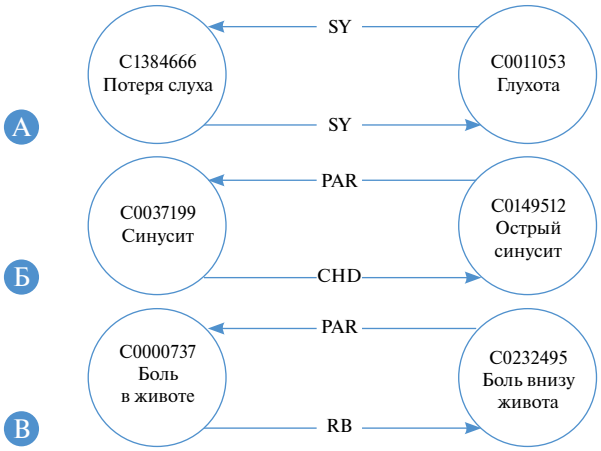

Рис. 2. Примеры симметричной (А), обратной (Б) и однородной (В) связей

Таблица 6. Описание связей UMLS

Класс связей	Группа	Смысловой перевод	Структура группы
Синонимичные	SY (Synonymous relationships)	Синонимичные связи	Содержит 17 подтипов. Например: same_as
Иерархические	CHD (Child relationships)	Связь с уточненным (дочерним) термином	Содержит 6 подтипов. Например: isa, part_of
	PAR (Parent relationships)	Связь с обобщающим (родительским) термином	Содержит 6 подтипов. Например: inverse_isa, has_part
	RB (Broader relationships)	Связь с более широким по смыслу термином	Содержит 15 подтипов. Например: inverse_isa, has_part
	RN (Narrower relationships)	Связь с более узким по смыслу термином	Содержит 15 подтипов. Например: isa, part_of
	SIB (Sibling relationships)	Связанные концепты имеют общий родительский термин (концепт более высокого класса) внутри одного справочника	Содержит 4 подтипа. Например: sib_in_isa, sib_in_part_of
Ассоциативные	AQ (Allowed qualifier)	Качественное уточнение	Содержит 3 подтипа. Например: actual_outcome_of, expected_outcome_of, modifies
	QB (Qualified by)	Качественное обобщение	Содержит 3 подтипа. Например:has_actual_outcome, has_expected_outcome, modified_by
	RO (Other relationships)	Другие связи, отличные от синонимичных и иерархических	Содержит 919 подтипов. Например: component_of, consists_of
	RL (Relationships as «like»)	Связь с аналогичным или подобным термином	Содержит 2 подтипа. Например: mapped_from, mapped_to
	RQ (Related and possibly synonymous relationships)	Связанные и возможно синонимичные концепты	Содержит 32 подтипа. Например: associated_with

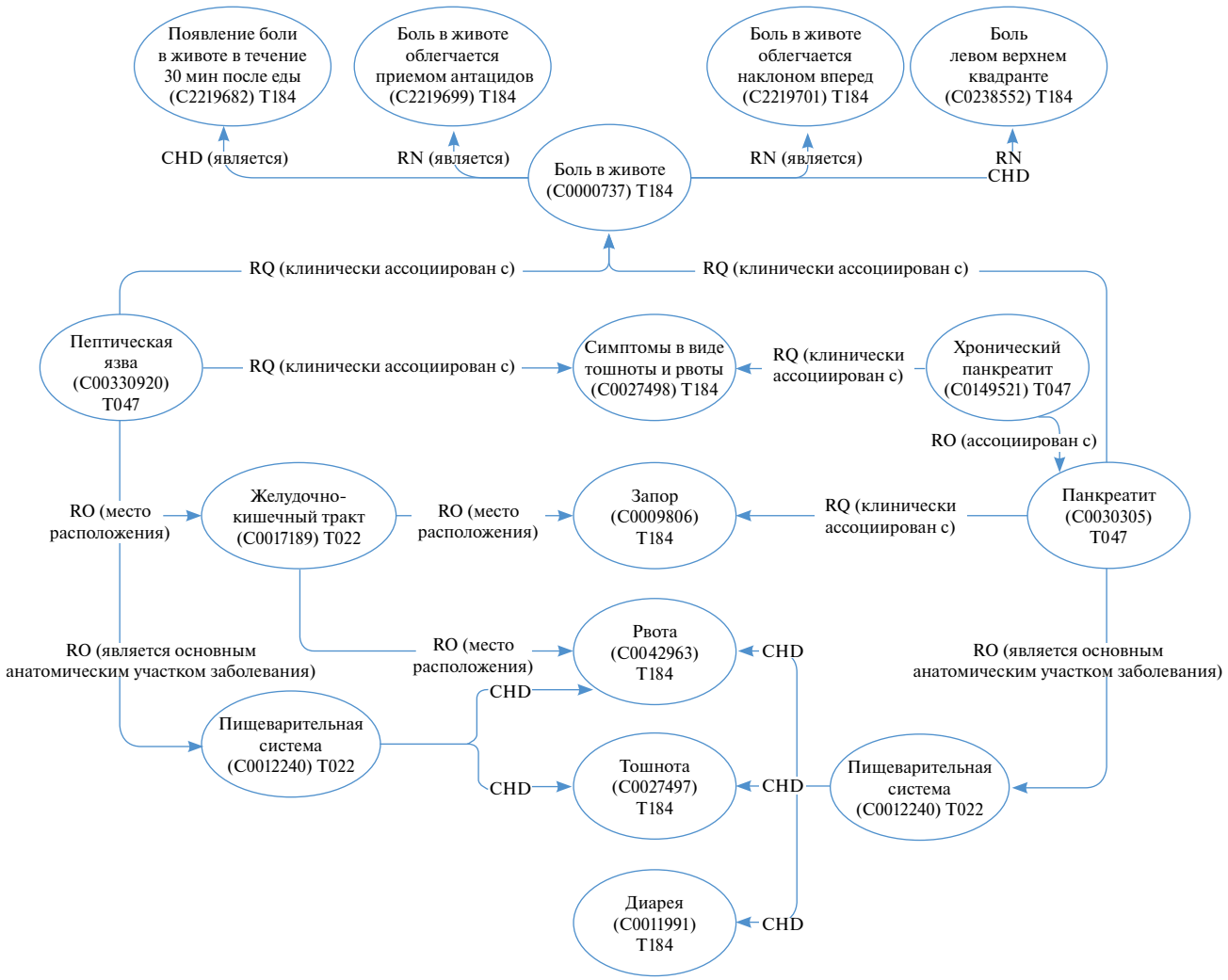

Рис. 3. Фрагмент графического представления симптоматического «образа» язвенной болезни и хронического панкреатита

рекомендаций. Для получения результата применялись аналитические панели и язык запросов Cypher к графовой БД.

Выявлены следующие закономерности: корневой термин (нозологический концепт) относится к семантической группе терминов T047 «Заболевания или синдромы», концевой концепт (симптом, синдром, симптомокомплекс, фактор риска, анамнестический признак) — к группе T184 «Симптомы и признаки» или T033 «Клинические находки». Все искомые концевые концепты были найдены максимум через два промежуточных узла. В качестве промежуточных узлов могли выступать концепты из практически любых семантических типов группы «Disorders – Клинические нарушения», однако преимущественно — уже описанные ранее T047, T184, T033 и T046 «Патологические функции», а также T022 «Анатомические и функциональные системы», T023 «Части тела, органы или части органов» из группы «Anatomy – Анатомия». Наименьшее количество связей между искомыми концептами составляли прямые. В подавляющем большинстве это была связь RQ (с подтипом clinically_associated_with). Связи через один и два промежуточных узла также имели закономерности. Если промежуточный концепт принадлежал семантическому типу «Симптомы и признаки», то концевой термин связан с ним по типу CHD или RN. Если промежуточные концепты были из групп T022, T023, то в основном это была связь RO (с подтипом finding_site_of). Обобщая результат по частоте встречаемости связей, можно сказать: связь RO встретилась в 52% случаев, CHD или RN — в 26%, RQ — в 21%, все остальные связи составили только 1%.

Аналогично были проанализированы закономерности в отношении инструментальной и лабораторной диагностики, рекомендуемой при выбранных патологиях. На рис. 4 видно, что результат инструментального исследования (термин чаще всего принадлежит семантическому типу T033) может быть напрямую связан через связь RN. При ее отсутствии самый оптимальный путь в графе проходит через промежуточный концепт, относящийся к типу с идентификатором T023.

На рис. 5 показано, что если в основе метода лабораторного исследования лежит определение возбудителя (например, Helicobacter pylori при язвенной болезни), то промежуточный концепт относится к семантическому типу T007 «Бактерии». В других ситуациях в качестве связующего звена выступает термин, характеризующий системы организма (T022).

Проделанная работа по анализу метатезауруса и поиску закономерностей его организации крайне ресурсозатратна, трудоемка и потребовала большого количества времени. Однако была необходима как непременное условие для поиска способов автоматизированного получения требуемых данных из UMLS.

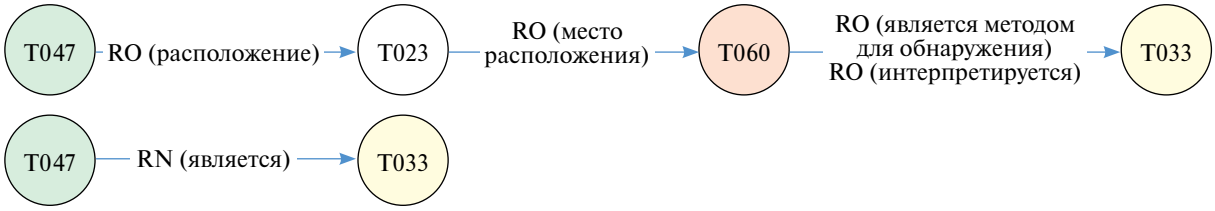

Рис. 4. Закономерности при анализе графа в отношении инструментальной диагностики заболеваний (T060 «Диагностические процедуры»)

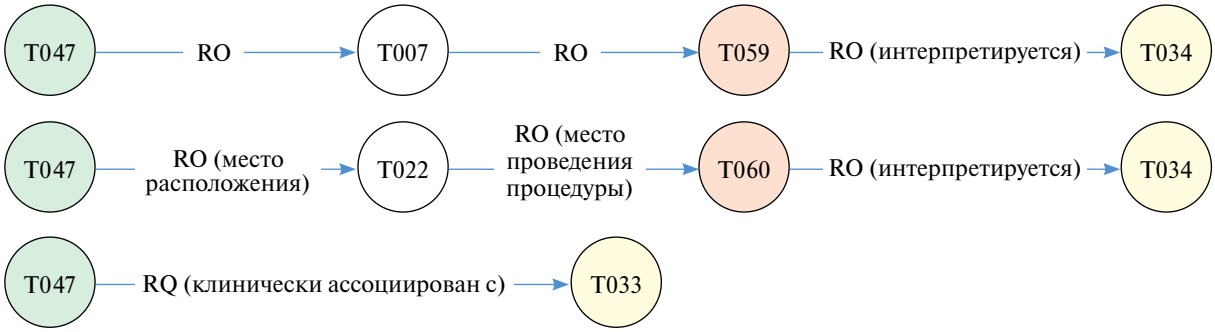

Рис. 5. Закономерности при анализе графа в отношении лабораторной диагностики заболеваний
Примечание. T007 — «Бактерии»; T022 — «Анатомические и функциональные системы»; T033 — «Клинические находки»; T034 — «Результаты лабораторных тестов»; T047 — «Заболевания или синдромы»; T059 — «Лабораторные процедуры»; T060 — «Диагностические процедуры».

Автоматизация получения релевантной информации из метатезауруса

Для воспроизведения релевантной информации из UMLS автоматизированным способом с помощью запросов к графовой БД использование только описанных выше закономерностей организации метатезауруса по семантическим принципам оказалось недостаточным. В выгрузке присутствовали далеко не все значимые концепты предметной области, при этом попадало огромное количество не имеющих отношения к искомой задаче концептов. Осуществлен поиск дополнительных подходов для автоматической выгрузки клинически значимых терминов.

На основе данных литературы, модификации и адаптации существующих алгоритмов семантического анализа были разработаны методы, которые учитывают способы организации метатезауруса как графовой информационной модели, основанные не только на конструкции онтологического древа (смысловые связи), но и на семантической близости терминов, заключенных в ее узлах:

- метод поиска кратчайшего пути, который позволяет найти оптимальный путь, связывающий корневой и концевой концепты и дает возможность определить связи, группы связей и концептов, вносящие наибольший вклад в построение подграфа (рис. 6) [40, 41];
- методы поиска релевантного окружения, базирующиеся на оценке связности концептов в графе и оценке их контекстной сочетаемости [42].

Группа методов для оценки связности концептов подразумевает расчет графовых метрик для ранжирования узлов по степени их значимости в контексте решаемой поисковой задачи на основании числа геометрических контуров (семантических треугольников или ромбов) в извлекаемом подграфе для каждого узла (рис. 7).

Методы оценки контекстной сочетаемости предполагают анализ частоты совместной встречаемости семанти-

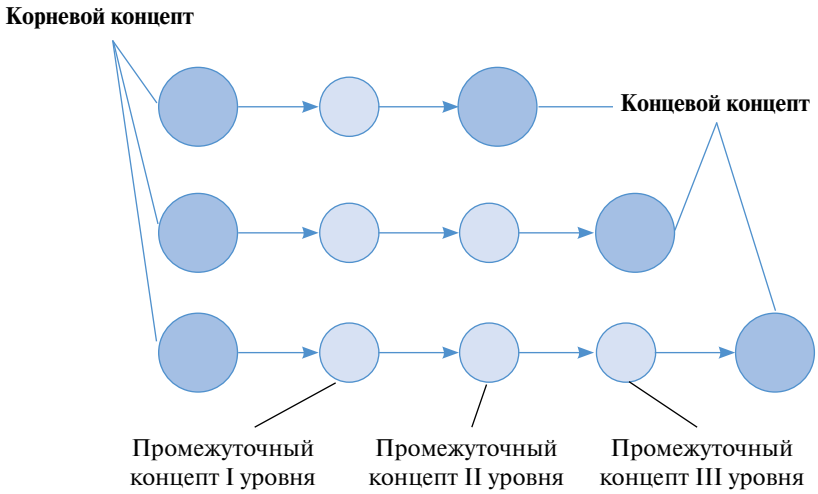

Рис. 6. Варианты графовых путей с включением корневого, концевого и промежуточных концептов UMLS

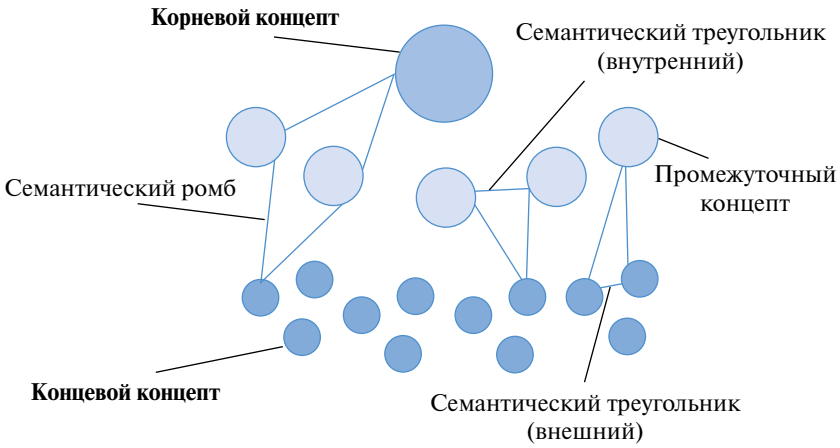

Рис. 7. Графовые контуры в окружении корневого концепта

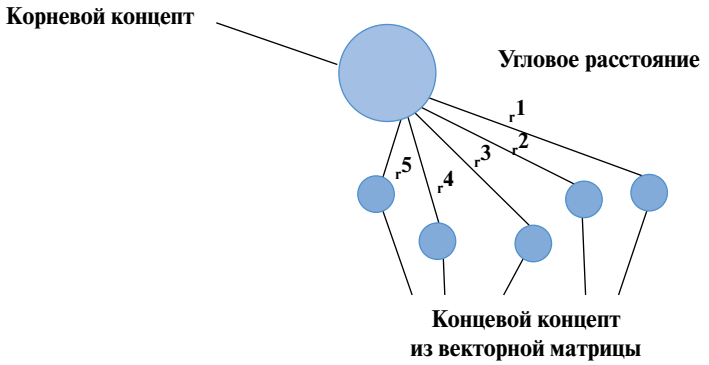

Рис. 8. Схема оценки углового расстояния между концептами

ческих единиц с использованием их векторного представления. В качестве меры близости терминов используются угловое расстояние и последующее ранжирование концептов по возрастанию (рис. 8).

Полученные модели легли в основу аналитической системы с пользовательским интерфейсом[1], работа с которой предлагает настройки для автоматического формирования комплекса запросов к СУБД Neo4j, в которой разворачивается семантическая сеть UMLS, с использованием языка программирования Cypher.

Входными параметрами являются такие значения, как глубина графового поиска, идентификаторы корневых и концевых концептов, их тематические группы, типы и атрибуты связей, выходными данными — перечень путей с указанием всех необходимых атрибутов для анализа кратчайшего пути и перечень узлов с указанием значений метрик ранжирования для оценки релевантного окружения. Также предусмотрено формирование частотного среза по типам встречающихся связей (и их атрибутов) в путях.

Верификация работы полученных модулей осуществлялась с использованием представленных выше клинических рекомендаций (см. табл. 2), раздел «Жалобы и анамнез». В «карту симптомов» вошло 579 уникальных формулировок (симптом, синдром, фактор риска, условия появления). Коды заболеваний по МКБ-10 выбирались из раздела клинических рекомендаций 1.4 «Особенности кодирования заболевания или состояния (группы заболеваний или состояний) по Международной статистической классификации болезней и проблем, связанных со здоровьем». Для поиска закономерностей проведено ранжирование типов связей и тематических групп терминов по частоте их встречаемости на каждом уровне. Установлено, что свыше 95% промежуточных и концевых симптоматических концептов относится к следующим группам терминов: T184 «Симптомы и признаки», T033 «Клинические находки», T046 «Патологические функции», T047 «Заболевания или синдромы». Выделены наиболее часто встречающиеся в графовых цепях типы связей: SIB — «концепты из одного справочника с общим родительским термином»; пары PAR–CHD и RB–RN — «иерархические связи»; RO — отношения, отличные от синонимичных и иерархических; AQ–QB — уточняющие связи (табл. 7).

Полученные метрики автоматизированной работы с графовой моделью данных дают сопоставимый с ручным анализом результат (см. рис. 4) по группам концептов, наиболее часто встречающимся связям при поиске симптомов, анамнестическим признакам и факторам риска заболеваний. Данный факт позволяет сформировать пул оптимальных настроек для автоматизированного получения образа заболеваний при решении информационно-поисковых задач и задач поддержки принятия клинических решений.

Дальнейшая апробация всех полученных как ручным способом, так и с помощью программных модулей закономерностей продемонстрировала точность выдачи релевантных данных около 70%. То есть треть всех характерных для конкретных заболеваний симптомов/синдромов, анамнестических признаков, факторов риска ана-

[1] Свидетельства о регистрации программы для ЭВМ № 2022684714 от 16.12.2022 и № 2022684715 от 16.12.2022.

Таблица 7. Структура прямых и непрямых связей между корневыми нозологиями и концептами-симптомами

	Нозологическая форма						
	ОНМК	**РМЖ**	**ИБС**	**ЖКК**	**Заболевания дыхательных путей***	**ЯБ****	**Все**
Всего связей, *n* Из них:	237	26	33	40	849	1164	2349
RO, *n* (%)	34 (14)	5 (19)	5 (15)	2 (5)	201 (24)	683 (59)	930 (40)
SIB, *n* (%)	65 (27)	8 (31)	3 (9)	14 (35)	379 (45)	—	469 (20)
CHD, RN, *n* (%)	14 (6)	6 (23)	12 (36)	7 (18)	89 (10)	275 (24)	403 (17)
RQ, *n* (%)	39 (16)	7 (27)	10 (30)	10 (25)	57 (7)	101 (9)	224 (10)
PAR, RB, *n* (%)	76 (32)	0 (0)	1 (3)	7 (18)	60 (7)	—	144 (6)
Другие, *n* (%)	9 (4)	0 (0)	2 (6)	0 (0)	63 (7)	105 (9)	179 (8)

*Неопухолевые заболевания верхних и нижних дыхательных путей (данные без пневмонии).
**Язвенная болезнь желудка и двенадцатиперстной кишки с заболеваниями дифференциального ряда.
Примечание. ОНМК — острое нарушение мозгового кровообращения; РМЖ — рак молочной железы; ИБС — ишемическая болезнь сердца; ЖКК — желудочно-кишечное кровотечение.

литическая система не выбирает при условии, что только 9% потенциально не удается найти в метатезаурусе. Также в выгрузки попадает множество концептов, не относящихся к искомой задаче, но искажающих результаты поиска каждого последующего уровня. Анализ показал, что из-за последней причины глубина рассматриваемого подграфа не должна превышать двух вложенных уровней, иначе доля полезной информации будет теряться в объеме несущественной.

В связи с этим возникает необходимость рассматривать дополнительные способы увеличения точности. Ожидаемо, что наиболее существенный вклад внесет наложение прямых связей, которые можно получить при обработке медицинской литературы и данных клинической практики. Как уже говорилось ранее (см. табл. 3), исходно только 17% симптоматических терминов связано напрямую с нозологией.

Использование прямых связей между концептами UMLS, полученными на 28 млн абстрактов публикаций PubMed, предоставленных в открытом доступе Национальной медицинской библиотекой США (SemMed), не позволило кардинально улучшить результат, а привело к значительному увеличению объема обрабатываемой информации (20 млн дополнительных связей) и, соответственно, времени выполнения запросов. Это может быть связано с тем, что в медицинских статьях довольно редко приводится полноценное описание картины заболеваний с перечнем всех важных признаков, а в основном представлены результаты научных исследований по какому-либо направлению.

Получение прямых связей на основе данных реальной клинической практики в части описания клинической картины заболеваний возможно при обработке неструктурированной медицинской информации. Полноценных больших наборов российских медицинских данных в открытом доступе пока не найдено. Для этих целей использован dataset MIMIC-IV из хранилища свободно доступных данных медицинских исследований, которым управляет лаборатория вычислительной физиологии Массачусетского технологического института (PhysioNet), на английском языке, содержащий 330 тыс. электронных деперсонифицированных медицинских документов, каждый из которых имеет связь с кодом МКБ-10. Семантический анализ (выделение концептов, их отнесение к классу заболеваний и разделу оказания медицинской помощи) проводился с использованием инструментов SemRep и MetaMap. Семантический анализатор SemRep, который позволяет извлекать из текстов триплеты (два концепта и связь между ними) на основе UMLS, ощутимо не улучшил качество выгрузки «клиническая картина — диагноз». Использование настроек и их сочетаний для семантического анализатора MetaMap позволяет выделять из текста электронных медицинских карт пациентов необходимые концепты, каждый из которых получает привязку к нозологии, таким образом формируя клинический «образ» заболевания. MIMIC-IV также дает возможность учитывать клиническую направленность выделяемых терминов (симптомы, анамнез, диагностика, лечение, риск развития и т.д.). В настоящее время на серверных мощностях РНИМУ им. Н.И. Пирогова завершена обработка dataset MIMIC-IV с помощью MetaMap. Результаты апробируются.

Перспектива видится и в результатах анализа российских медицинских публикаций и данных электронных медицинских карт пациентов, работы над которыми ведутся в настоящее время.

Таким образом, использование адаптированных переводов UMLS, внедрение справочников ФР НСИ, добавление дополнительных прямых связей между концептами, полученных на основе данных реальной клинической практики, дают возможность составления унифицированной национальной медицинской номенклатуры терминов (УНМН), а дополнительное использование полученных метрик и закономерностей работы с УНМН как графа — реализовать онтологическую модель для получения автоматизированным способом клинического «образа» заболеваний на высоком качественном уровне.

Построение онтологической модели данных на основе национального метатезауруса

Онтологии отличаются от простых терминологий тем, что определяют отношения между понятиями таким образом, который допускает вычислительные логические взаимодействия, позволяя делать выводы из связанных утверждений [20].

Онтологический принцип организации данных предполагает возможность выбирать релевантную информацию автоматизированным способом. Закономерности семантической сети и графовые метрики являются шагом к построению онтологии. Однако точность, которую они позволяют достичь, недостаточна для решения задач на клиническом уровне (информационно-поисковых систем и СППКР). Повышение качества выгрузки информации возможно путем создания системы весовых коэффициентов для узлов и связей UMLS. Весовые коэффициенты позволяют ранжировать узлы графа по степени потенциальной значимости в контексте решаемых задач и в настоящее время представлены совокупностью валидированных аналитических разработок [43].

В первую очередь можно говорить о создании универсальной аналитической метрики для оценки связности узлов в извлекаемом подграфе. Одной из существенных проблем семантической сети UMLS является неравномерность прямых связей между концептами, приводящая к смещению любых математических оценок при попытке анализа связности концептов и снижению релевантности получаемой информации при автоматизированной обработке. Это требует необходимости создания специализированных взвешенных аналитических метрик, учитывающих плотность структуры связей вокруг исследуемого концепта. Указанная графовая метрика называется взвешенным коэффициентом кластеризации (ВКК) и обеспечивает учет количественной относительной меры плотности структуры связей и степень неоднородности последней. Расчет ВКК производится по формуле

$$\text{ВКК}_n = C_n \cdot R_n^{-1,29},$$

где C_n — число геометрических контуров или незамкнутых путей, в образовании которых участвует узел n; R_n — число прямых связей между узлом n и любыми другими вершинами графа.

ВКК подразумевает расчет числа графовых контуров или суммы весов связей для каждого концепта из извлекаемого множества произвольного объема. Учет неоднородности структуры UMLS производится с использованием эмпирически выведенного закона, устанавливающего степенную зависимость между числом прямых связей и долей концептов UMLS с количеством прямых связей выше указанного. Ближайшим аналогом данной закономерности является закон Ципфа, используемый в компьютерной лингвистике для оценки семантических характеристик различных элементов текста [44].

С использованием ВКК было автоматически размечено свыше 300 тыс. клинически значимых концептов из семантической группы «Disorders — Клинические нарушения» метатезауруса по степени их принадлежности к профилям заболеваний. Для этого для каждого исследуемого концепта определялась его связность с концептами из справочников ICD-10 и ICD-10-CM, в свою очередь однозначно сопоставимыми с МКБ-10. На выходе получался ранжированный по данной метрике ряд. Точность разметки составила 91%, что позволило внедрить данный инструмент в прототип информационно-поисковой системы для автоматического определения наиболее вероятного класса заболеваний при составлении дифференциально-диагностического ряда произвольного клинического образа.

Следующий элемент системы весовых коэффициентов связан с оценкой клинической значимости концептов, которая определялась с использованием модели машинного обучения, анализирующей различные характеристики терминов из метатезауруса UMLS (тематическая группа, актуальность термина, предпочтительное название и др.). Значения метрик точности, чувствительности и специфичности логистической регрессионной модели составили соответственно 91, 90 и 91% для валидационной выборки [43]. При оценке клинической значимости концептов UMLS сделан вывод о том, что для описания заболеваний и их признаков потенциально применимо свыше 1,5 млн уникальных терминов. Важно отметить, что подобная разметка не только позволила сузить круг концептов для использования при поиске релевантной информации, но и обеспечила возможность создания перечня приоритетных справочников и групп терминов UMLS, требующих экспертного перевода и адаптации на русский язык.

Завершена работа по созданию алгоритма оценки относительной специфичности терминов UMLS. Данный алгоритм позволяет решить проблему, связанную с извлечением из текста неспецифичных (обобщающих) понятий, существенно ухудшающих результат поиска. Примерами таких понятий являются сущности «симптом», «признак», «заболевание», «пациент». В ходе исследования было установлено, что наилучшими прогностическими характеристиками для выявления относительно специфичных и относительно неспецифичных концептов обладают правила, основанные на сравнении общего числа прямых связей и числа связей «родитель–потомок». Важно отметить, что наиболее применимыми для UMLS являются правила, подразумевающие оценку общего числа прямых связей и среднюю длину иерархических цепей для атомарных формулировок концептов. Точность итогового алгоритма составила 99,1% при попарном сравнении концептов тестовой выборки. Допускается использование алгоритма только для ранжирования небольших наборов терминов UMLS, связанных прямыми ассоциативными связями.

Разработка информационно-поисковой системы

Основная цель любой информационно-поисковой системы (ИПС) заключается в том, чтобы удовлетворить информационную потребность пользователя, предоставив всю важную информацию и при этом не перегружая лишней. К ИПС предъявляется ряд требований, таких как полнота, достоверность и актуальность получаемой информации, высокая скорость и удобство поиска, возможность распознавания свободного текста.

Информационно-поисковая система — это совокупность алгоритмов, обеспечивающих выбор необходимой информации с помощью специальных баз данных. Описанная выше онтологическая модель метатезауруса может являться основой для реализации информационно-поисковых задач для врача, пациента, исследователя в медицине. Это возможность, с одной стороны, получить информацию из метатезауруса о диагнозах и их причинах, патогенезе и рисках возникновения, диагностических исследованиях и лечении и др., а с другой — создав алгоритмы разметки отечественной медицинской литературы по принципу библиотеки PubMed и систем семантического анализа медицинских текстов SemRep и MetaMap, иметь тематическую привязку к актуальным научным источникам знаний.

Важнейшей задачей при создании ИПС является необходимость обработки свободного текста пользователя с целью автоматического распознавания концептов метатезауруса и на их основании — поиска необходимой

информации, например, дифференциального ряда заболеваний по выделенным симптомам, признакам, анамнестическим данным. В настоящее время создан прототип ИПС, в котором реализована задача симптомчекера. Упрощенный алгоритм его действий выглядит следующим образом.

Предварительно концепты из всех возможных для данной задачи тематических групп и их варианты звучания, присутствующие в словаре, подготовлены в виде нормализованных концептов-лемм. Для задачи симптомчекера это составило около 400 тыс. записей (без учета перестановки слов). Вводимый в ИПС текст подвергается сегментации, токенизации и лемматизации. В случае обнаружения во вводимом тексте совпадения с нормализованной леммой из метатезауруса ему присваивается максимальный ранг. Если лемма вводимого текста является подстрокой леммы нормализованного, то термину присваивается более низкий ранг. Все концепты, имеющие максимальный ранг, выводятся пользователю как «выбранные симптомы», остальные предлагаются в качестве возможных (рис. 9). Также подбираются концепты, тесно связанные с выбранными симптомами, полученные на основе рассчитанных весовых коэффициентов. Из общего перечня удаляются неспецифичные (обобщенные) концепты. Пользователь может скорректировать полученный симптоматический образ, перенеся признак из списка возможных в список выбранных. Поиск диагностического ряда осуществляется на основании суммирования и ранжирования всех используемых метрик каждого выбранного концепта.

Разработка базы медицинских знаний

Все описанные подходы, а также получение прямых связей на данных реальной клинической практики позволяют автоматизированным способом извлекать максимально возможную информацию из онтологического представления метатезауруса. Предварительные результаты говорят приблизительно о 80–90%-й точности данных, что все же недостаточно для создания основанных на знаниях интеллектуальных систем и требует разработки полноценных баз знаний для каждой клинической области, а также учета коморбидности. Необходимо привлечение экспертов, использование медицинской литературы, результатов математического анализа больших данных, накопленных в медицине, а также проектирование архитектуры для их построения. Крайне важен тот факт, что в основе должна лежать единая унифицированная медицинская терминология, организованная по онтологическому принципу. С точки зрения организации база знаний представляет собой надстройку национальной медицинской номенклатуры для каждого разрабатываемого клинического направления, содержащую в качестве атрибутов информацию о концептах и их связях на основе исходных закономерностей метатезауруса, графовых метрик, весовых коэффициентов, полученных из верифицированных источников информации и/или экспертных знаний, триггерных событий и других необходимых для работы решателя моментов.

Подход к построению прототипа экспертной системы описанным способом был апробирован на примере язвенной болезни и ее дифференциального ряда (хронический гастрит, хронические панкреатит, хронический холецистит, рак желудка). Автоматизированным способом из метатезауруса извлечен 71% необходимых жалоб, анамнестических данных и факторов риска для исследуемых нозологий. Для этого в СУБД PostgreSQL сформирована таблица, содержащая уникальные коды связанных между собой концептов, типы связи между ними, силу связи между нозологическим концептом и признаком заболевания. Оставшиеся признаки после перевода полученных структур в графовую модель были размечены вручную с использованием языка запросов Cypher (добавлены концепты и связаны прямой связью с нозологиями).

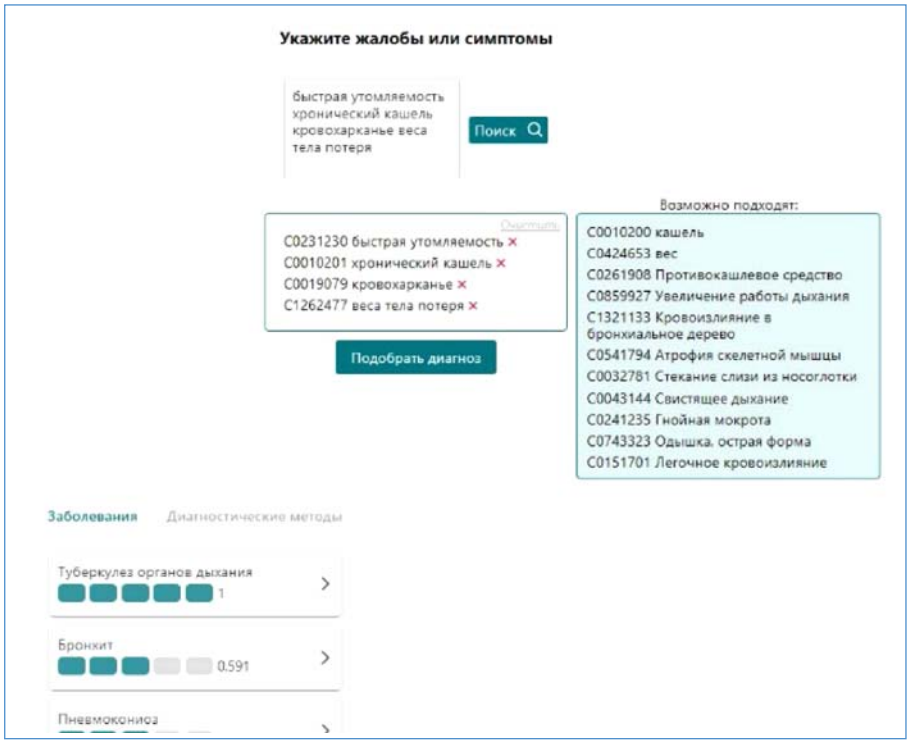

Рис. 9. Фрагмент веб-интерфейса рабочего варианта прототипа информационно-поисковой системы

В связи с тем, что данный этап разработки был начальным подходом, сила связи определялась без привлечения экспертов на основе верифицированных источников информации (клинические рекомендации, медицинская литература). Все термины были разделены на четыре группы по убыванию значимости (ведущий, характерный, возможный, редко встречающийся) в отношении постановки диагноза для каждой исследуемой нозологии, где первая группа характеризует ведущие симптомы, получившие силу связи 0,8, а четвертая — наименее важные для диагностики признаки (0,2). Сумма весовых коэффициентов всех введенных жалоб, анамнестических признаков и имеющихся факторов риска позволяет сформировать ранжированный список заболеваний. Предварительная проверка алгоритма на историях болезни, которые используются в качестве учебных материалов на кафедре факультетской терапии РНИМУ им. Н.И. Пирогова и соответствуют необходимым нозологиям, продемонстрировала его потенциальную применимость. Только в одном из семи случаев верный диагноз попал на 3-е место, в остальных случаях занимал 1-е место в ранжированном списке заболеваний. В одной из рассматриваемых ситуаций у пациента имелись два состояния в качестве диагноза (хронический холецистит и хронический панкреатит), обе нозологии были выведены на 1-е и 2-е место.

Понятно, что продемонстрированный результат имеет исключительно поисковый характер, требующий серьезной доработки как в техническом направлении (разработка архитектуры системы, интерфейса эксперта, разработчика базы знаний, пользователя), так и предметной части (экспертная верификация, пороговые величины для принятия решений, диалог с пользователем, проверка на реальных клинических данных и др.). Однако данная работа позволила оценить перспективы построения базы знаний на основе онтологического представления УНМН, спроектировать архитектуру системы для разработки алгоритмов принятия клинических решений, основанных на знаниях.

Заключение

Таким образом, в ходе настоящего исследования на основе метатезауруса UMLS с использованием экспертных переводов, федеральных медицинских справочников, прямых связей, полученных на абстрактах статей и реальных клинических данных, создана первая версия унифицированной национальной медицинской номенклатуры. Необходимо ее дальнейшее пошаговое развитие.

Одним из параллельных результатов, которые можно получить в данном направлении, является разработка новых национальных словарей, например группы справочников, имеющих отношение к симптомам и клиническим находкам (синонимы, характер, сезонность, периодичность, условия возникновения, распространенность, источники и методы получения и др.), которые могут быть использованы в медицинской информационной системе для формализации ведения медицинских документов и однозначно мапированы с УНМН.

Показано, что онтологические модели знаний являются эффективным способом представления формализованной и структурированной информации в медицине. Преобразование знаний в графовые информационные модели позволяет на их основе создавать компоненты информационно-поисковых систем. Формируемые с применением соответствующих метрик подграфы семантической сети, добавление к узлам и связям рангов и коэффициентов значимости (полученных как математическим, так и экспертным путем) дают возможность создать образ заболевания (базу знаний) и подойти к задаче разработки систем поддержки принятия клинических решений.

Разработаны аналитические инструменты для работы с метатезаурусом, реализующие низкоуровневые запросы к графовой модели. К перспективам исследовательских задач в этой части необходимо отнести создание метрик ранжирования узлов и связей на основе данных реальной клинической практики, а также модификацию и оптимизацию алгоритмов извлечения именованных сущностей из неструктурированных текстов.

В качестве перспективных проектных задач, в основе которых лежит УНМН, следует назвать необходимость создания полноценных информационно-поисковых систем, рассчитанных на пользователя — врача, пациента и исследователя в медицине, а также системы для разработки СППКР, в которой будут реализованы автоматизированные рабочие места эксперта и инженера по знаниям.

Дополнительная информация

Источник финансирования. Настоящее исследование финансируется из средств Программы стратегического академического лидерства «Приоритет–2030». Номер субсидии 075-15-2021-1325 (от 30 сентября 2021 г.).

Конфликт интересов. Авторы данной статьи подтвердили отсутствие конфликта интересов, о котором необходимо сообщить.

Участие авторов. Т.В. Зарубина — идеология построения исследования, общее руководство, разработка концепции статьи, редактирование текста; С.Е. Раузина — руководство направлениями исследования, постановка задач, контроль выполнения, анализ результатов, написание основного текста статьи; П.А. Астанин — проведение исследовательских работ, создание аналитических метрик, разработка программного обеспечения, реализация запросов к базе данных, анализ результатов, оформление статьи, перевод на английский язык; Ю.И. Королева — руководство направлениями исследования, постановка задач, контроль выполнения, анализ результатов, оформление статьи; Л.В. Ронжин — проведение исследовательских работ, создание аналитических метрик, разработка программного обеспечения, реализация запросов к базе данных, анализ результатов, оформление статьи, перевод на английский язык; А.А. Борисов — проведение исследовательских работ, создание аналитических метрик, разработка программного обеспечения, реализация запросов к базе данных, анализ результатов, оформление статьи, перевод на английский язык; М.А. Афанасьева — участие в проведении исследования, подготовка таблиц и рисунков; А.В. Усова — участие в проведении исследования, подготовка таблиц и рисунков. Все авторы статьи внесли существенный вклад в поисково-аналитическую работу, прочли и одобрили окончательную версию рукописи перед публикацией.

Выражение признательности. Выражаем слова благодарности всем сотрудникам кафедры медицинской кибернетики и информатики им. С.А. Гаспаряна и Института цифровой трансформации медицины РНИМУ им. Н.И. Пирогова, принявшим участие в проведении данного исследования.

ЛИТЕРАТУРА

1. Указ Президента Российской Федерации от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации». [Decree of the President of the Russian Federation No. 490 of 10 October 2019 “O razvitii iskusstvennogo intellekta v Rossijskoj Federacii”. (In Russ.)] Available from: http://www.kremlin.ru/acts/bank/44731 (accessed: 02.10.2023).
2. Гусев А.В., Владзимирский А.В., Шарова Д.Е., и др. Развитие исследований и разработок в сфере технологий искусственного интеллекта для здравоохранения в Российской Федерации: итоги 2021 года // *Digital Diagnostics*. — 2022. — Т. 3. — № 3. — С. 178–194. [Gusev AV, Vladzymyrskyy AV, Sharova DE, et al. Evolution of research and development in the field of artificial intelligence technologies for healthcare in the Russian Federation: results of 2021. *Digital Diagnostics*. 2022;3(3):178–194. (In Russ.)] doi: https://doi.org/10.17816/DD107367
3. Искусственный интеллект поможет московским терапевтам в постановке диагнозов // *Коммерсантъ*. 08.09.2023. [Iskusstvennyj intellekt pomozhet moskovskim terapevtam v postanovke diagnozov. *Kommersant*. 08.09.2023. (In Russ.)] Available from: https://www.kommersant.ru/doc/6199795 (accessed: 02.10.2023).
4. Гусев А.В., Астапенко Е.М., Иванов И.В., и др. Принципы формирования доверия к системам искусственного интеллекта для здравоохранения // *Вестник Росздравнадзора*. — 2022. — № 2. — С. 25–33. [Gusev AV, Astapenko EM, Ivanov IV, et al. Principles for building confidence in artificial intelligence systems for healthcare. *Vestnik Roszdravnadzora*. 2022;2:25–33. (In Russ.)]
5. Поповьян Р.А., Будкова Н.Н., Раковский А.А., и др. Составление тезауруса — важный шаг по настройке работы информационной системы в любой области медицины // *Врач и информационные технологии*. — 2005. — № 4. — С. 57–61. [Popovyan RA, Budkova NN, Rakovskij AA, et al. Sostavlenie tezaurusa-vazhnyj shag po nastrojke raboty informacionnoj sistemy v lyuboj oblasti mediciny. *Vrach i informacionnye tekhnologii*. 2005;4:57–61. (In Russ.)]
6. Морозов С.П., Владзимирский А.В., Шулькин И.М., и др. Исследование целесообразности применения технологий искусственного интеллекта в лучевой диагностике // *Врач и информационные технологии*. — 2022. — № 1. — С. 12–29. [Morozov SP, Vladzymyrskyy AV, Shulkin IM, et al. Feasibility of using artificial intelligence in radiation diagnostics. *Medical doctor and information technology*. 2022;1:12–29. (In Russ.)] doi: https://doi.org/10.25881/18110193_2022_1_12
7. Нейросеть поможет врачам одновременно определять до 10 патологий на КТ-снимках // *mos.ru. Новости*. 27.07.2023. [Nejroset` pomozhet vracham odnovremenno opredelyat` do 10 patologij na KT-snimkakh. *mos.ru. News*. 27.07.2023. (In Russ.)] Available from: https://www.mos.ru/news/item/125792073/ (accessed: 02.10.2023).
8. Компьютерное зрение поможет врачам в 29 направлениях лучевых исследований // *mos.ru. News*. 19.03.2022. [Komp`yuternoe zrenie pomozhet vracham v 29 napravleniyakh luchevy`kh issledovanij. *mos.ru. News*. 19.03.2022. (In Russ.)] Available from: https://www.mos.ru/news/item/103860073/?utm_source=search&utm_term=serp (accessed: 02.10.2023).
9. Jing X. The Unified Medical Language System at 30 Years and How It Is Used and Published: Systematic Review and Content Analysis. *JMIR Med Inform*. 2021;9(8):e20675. doi: https://doi.org/10.2196/20675
10. National Library of Medicine, USA, Unified Medical Language System (UMLS). Available from: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html (accessed: 02.10.2023).
11. Румянцев П.О., Бледжянц Г.А., Туманов Н.А., и др. UMKB-технология для создания «интеллектуальных» систем в области медицины // *Здравоохранение*. — 2015. — № 11. — С. 82–89. [Rumyanczev PO, Bledzhyancz GA, Tumanov NA, et al. UMKB-tekhnologiya dlya sozdaniya “intellektualnykh” sistem v oblasti mediczinу. *Zdravookhranenie*. 2015;11:82–89. (In Russ.)]
12. Maslova AY, Mishvelov AE, Dudusheva MJ, et al. Using a Decision Tree with a Feedback Function to Select Therapeutic Tactics for Viral Infection of the Respiratory Tract in the Medical Expert System. *International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies*. 2022;13(8):1–10. doi: https://doi.org/10.14456/ITJEMAST.2022.157
13. Грибова В.В., Окунь Д.Б. Онтологии для формирования баз знаний и реализации лечебных мероприятий в медицинских интеллектуальных системах // *Информатика и системы управления*. — 2018. — Т. 57. — № 3. — С. 71–80. [Gribova VV, Okun DB. Ontologies for the formation of knowledge bases about disease treatment in medical intelligent systems. *Informatika i sistemy upravleniya*. 2018;57(3):71–89. (In Russ.)] doi: https://doi.org/10.22250/isu.2018.57.71-80
14. Переволоцкий В.С., Грибова В.В. Подход к автоматическому формированию баз знаний на основе онтологий // *Научный аспект*. — 2023. — Т. 2. — № 2. — С. 213–221. [Perevolockij VS, Gribova VV. Podhod k avtomaticheskomu formirovaniyu baz znanij na osnove ontologij. *Nauchnyj aspekt*. 2023;2(2):213–221. (In Russ.)]
15. Москаленко Ф.М., Окунь Д.Б., Петряева М.В. База терминов для интеллектуальных медицинских сервисов // *Системный анализ в медицине (САМ 2016)*: материалы X Международной научной конференции, 22–23 сентября 2016 г. — Благовещенск, 2016. — С. 155–158. [Moskalenko PM, Okun DB, Petryaeva MV. Terminology base for intelligent medical services. *Sistemnyj analiz v medicine (SAM 2016):* Materialy X mezhdunarodnoj nauchnoj konferencii; 2016 Sep. 22–23; Blagoveshchensk; 2016. P. 155–158. (In Russ.)]
16. Мусаев А.А., Григорьев Д.А. Обзор современных технологий извлечения знаний из текстовых сообщений // *Компьютерные исследования и моделирование*. — 2021. — Т. 13. — № 6. — С. 1291–1315. [Musaev AA, Grigoriev DA. Extracting knowledge from text messages: overview and state-of-the-art. *Computer Research and Modeling*. 2021;13(6):1291–1315. (In Russ.)] doi: https://doi.org/10.20537/2076-7633-2021-13-6-1291-1315
17. Катенко Ю.В. Применение методов машинного обучения для анализа текстовой информации // *Охрана, безопасность, связь*. — 2019. — № 4-3. — С. 90–94. [Katenko YuV. Application of machine learning methods for text information analysis. *Ohrana, bezopasnost’, svyaz’*. 2019;(4 Pt 3):90–94. (In Russ.)]
18. Донитова В.В., Киреев Д.А., Титова Е.В., и др. Методы обработки естественного языка для извлечения факторов риска инсульта из медицинских текстов // *Труды Института системного анализа Российской академии наук*. — 2021. — Т. 71. — № 4. — С. 93–101. [Donitova VV, Kireev DA, Titova EV, et al. Natural language processing models for extraction of stroke risk factors from electronic health records. *Proceeding of the Institute for Systems Analysis of the Russian Academy of Science*. 2021;71(4):93–101. (In Russ.)] doi: https://doi.org/10.14357/20790279210410
19. Тучкова П.А. Аннотирование документов и применение речевых технологий для решения задач обработки естественного языка в медицине // *Наукосфера*. — 2021. — № 9 (2). — С. 69–73. [Tuchkova PA. Text annotation and using speech technologies for solving natural language processing tasks in medicine. *Naukosfera*. 2021;(9Pt2):69–73. (In Russ.)] doi: https://doi.org/10.5281/zenodo.5531131

20. Chute CG. Clinical classification and terminology: some history and current observations. *J Am Med Inform Assoc*. 2000;7(3):298–303. doi: https://doi.org/10.1136/jamia.2000.0070298
21. Слепов А.Э. Модуль построения семантических запросов к онтологическим базам знаний // *Наука настоящего и будущего*. — 2022. — Т. 2. — С. 188–191. [Slepov AE. Module for constructing semantic queries to ontological knowledge bases. *Nauka nastoyashhego i budushhego*. 2022;2:188–191. (In Russ.)]
22. Кукарцев В.В., Колмакова З.А., Мельникова О.Л. Системный анализ возможностей по извлечению именованных сущностей с применением технологии Text Mining // *Перспективы науки*. — 2019. — Т. 120. — № 9. — С. 18–20. [Kukarcev VV, Kolmakova ZA, Melnikova OL. System Analysis of Possibilities to Retrieve Essentials Using Text Mining Technology. *Science Prospects*. 2019;120(9):18–20. (In Russ.)]
23. Орлова Д.Е., Падалко А.В. Использование аппарата семантических сетей для интеллектуальной поддержки принятия решений // *Вестник Воронежского института высоких технологий*. — 2021. — Т. 36. — № 1. — С. 61–65. [Orlova DE, Padalko AV. Using the semantic network apparatus for intelligent decision support. *Bulletin Voronezh institute of high technologies*. 2021;36(1):61–65. (In Russ.)]
24. Кайда А.Ю., Савельев А.О. Применение графовой визуализации данных в системах поддержки принятия решений для решения задач автоматизации проведения исследований // *Электронные средства и системы управления*: материалы докладов международной научно-практической конференции. — 2020. — № 1-2. — С. 137–139. [Kajda AYu, Savelev AO. Primenenie grafovoj vizualizacii dannyh v sistemah podderzhki prinyatiya reshenij dlya resheniya zadach avtomatizacii provedeniya issledovanij. *Elektronnye sredstva i sistemy upravleniya*. Materialy dokladov mezhdunarodnoj nauchno-prakticheskoj konferencii. 2020;(1 Pt 2):137–139. (In Russ.)]
25. Szárnyas G, Kővári Z, Salánki A, et al. Towards the characterization of realistic models: evaluation of multidisciplinary graph metrics. Proceedings of the ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems (MODELS'16); 2016 Oct 2. New York, NY; 2016. P. 87–94. doi: https://doi.org/10.1145/2976767.2976786
26. Varghese J, Sünninghausen SS, Dugas M. Standardized Cardiovascular Quality Assurance Forms with Multilingual Support, UMLS Coding and Medical Concept Analyses. *Stud Health Technol Inform*. 2015;216:837–841.
27. Мосалов О.П. Векторные представления ребер графа онтологии как инструмент для анализа и генерации новых данных // *Информационно-технологический вестник*. — 2021. — Т. 27. — № 1. — С. 93–101. [Mosalov OP. Edge embedding of ontology graphs as a tool for analysis and generation of new data. *Informacionno-Tekhnologicheskij Vestnik*. 2021;27(1):93–101. (In Russ.)]
28. Семантическая платформа DataMonitor. [Semanticheskaya platforma DataMonitor. (In Russ.)] Available from: http://avicomp.ru/services/datamonitor (accessed: 02.10.2023).
29. Портал нормативно справочной информации Министерства здравоохранения Российской Федерации. [Portal normativno spravochnoj informacii Ministerstva zdravoohraneniya Rossijskoj Federacii. (In Russ.)] Available from: https://nsi.rosminzdrav.ru (accessed: 02.10.2023).
30. Kilicoglu H, Rosemblat G, Fiszman M, et al. Broad-coverage biomedical relation extraction with SemRep. *BMC Bioinformatics*. 2020;21(1):188. doi: https://doi.org/10.1186/s12859-020-3517-7
31. Lang F, Mork JG, Demner-Fushman D, et al. Increasing UMLS Coverage and Reducing Ambiguity via Automated Creation of Synonymous Terms: First Steps toward Filling UMLS Synonymy Gaps. American Medical Informatics Association Annual Symposium; 2018. P. 1–26.
32. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1. doi: https://doi.org/10.1038/s41597-022-01899-x
33. Daniel G, Sunyé G, Cabot J. UMLtoGraphDB: Mapping Conceptual Schemas to Graph Databases. *Lecture Notes in Computer Science*. 2016;9974:430–444. doi: https://doi.org/10.1007/978-3-319-46397-1_33
34. Масликова У.В., Супильников А.А. Технологии разработки программы содействия принятию решения в диагностике заболеваний системы крови с использованием сверточных искусственных нейронных сетей // *Вестник медицинского института «Реавиз»: реабилитация, врач и здоровье*. — 2020. — Т. 47. — № 5. — С. 138–150. [Maslikova UV, Supilnikov AA. Technologies for developing decision support systems for the diagnosis of blood disorders using convolutional neural networks. *Bulletin of Medical University Reaviz*. 2020;47(5):138–150. (In Russ.)] doi: https://doi.org/10.20340/vmi-rvz.2020.5.16
35. Berger B, Waterman MS, Yu YW. Levenshtein Distance, Sequence Comparison and Biological Database Search. *IEEE Trans Inf Theory*. 2021;67(6):3287–3294. doi: https://doi.org/10.1109/ tit.2020.2996543
36. Черкашин А.М. Применение метода расстояния Левенштейна библиотеки Fuzzywuzzy языка Python для исправления данных // *Постулат*. — 2021. — Т. 66. — № 4. — С. 11. [Cherkashin AM. Applying the Levenshtein distance method of the Python Fuzzywuzzy library to correct data. *Postulat*. 2021;66(4):11. (In Russ.)]
37. Рубрикатор клинических рекомендаций Министерства здравоохранения Российской Федерации. [Rubrikator klinicheskih rekomendacij Ministerstva zdravoohraneniya Rossijskoj Federacii. (In Russ.)] Available from: https://cr.minzdrav.gov.ru/ (accessed: 02.10.2023).
38. Астанин П.А. Применение автоматизированного анализа семантической сети UMLS для решения задачи поиска релевантных знаний о ревматических заболеваниях // *Математическое моделирование систем и процессов*: сборник материалов Международной научно-практической конференции. — Псков, 2022. — С. 6–12. [Astanin PA. Primenenie avtomatizirovannogo analiza semanticheskoj seti UMLS dlya resheniya zadachi poiska relevantnyh znanij o revmaticheskih zabolevaniyah. *Matematicheskoe modelirovanie sistem i processov:* Sbornik materialov Mezhdunarodnoj nauchno-prakticheskoj konferencii. Pskov; 2022. P. 6–12. (In Russ.)] doi: https://doi.org/10.37490/978-5-00200-102-6-6-12
39. Ширинян М.В., Шустова С.В. Трудности медицинского перевода и способы их преодоления при обучении студентов неязыковых вузов // *Язык и культура*. — 2018. — № 43. — С. 295–316. [Shirinyan MV, Shustova SV. Trudnosti medicinskogo perevoda i sposoby ih preodoleniya pri obuchenii studentov neyazykovyh vuzov. *Yazyk i kul'tura*. 2018;43:295–316. (In Russ.)] doi: https://doi.org/10.17223/19996195/43/18
40. Астанин П.А., Ронжин Л.В., Королева Ю.И. *Модуль поиска кратчайшего пути между концептами семантической сети UMLS*. Свидетельство о регистрации программы для ЭВМ № 2022684714. 16.12.2022. [Astanin PA, Ronzhin LV, Koroleva YuI. *Modul' poiska kratchajshego puti mezhdu konceptami semanticheskoj seti UMLS*. Certificate of computer program registration RUS No. 2022684714. 16.12.2022. (In Russ.)] Available from: https://www.elibrary.ru/download/elibrary_49979662_35579965.PDF (accessed: 02.10.2023).
41. Близнякова Е.А., Куликов А.А., Куликов А.В. Сравнительный анализ методов поиска кратчайшего пути в графе // *Архитектура, строительство, транспорт*. — 2022. — № 1. — С. 80–87. [Bliznyakova EA, Kulikov AA, Kulikov AV. Comparative analysis of methods for finding the shortest distance in a graph. *Architecture, construction, transport*. 2022;1:80–87. (In Russ.)] doi: https://doi.org/10.31660/2782-232X-2022-1-80-87

42. Астанин П.А., Ронжин Л.В., Раузина С.Е. *Модуль поиска релевантного окружения концептов семантической сети UMLS*. Свидетельство о регистрации программы для ЭВМ № 2022684715. 16.12.2022. [Astanin PA, Ronzhin LV, Rauzina SE. *Modul' poiska relevantnogo okruzheniya konceptov semanticheskoj seti UMLS*. Certificate of computer program registration RUS No. 2022684715. 16.12.2022. (In Russ.)] Available from: https://www.elibrary.ru/download/elibrary_49979663_14353005.PDF (accessed: 02.10.2023).
43. Астанин П.А., Раузина С.Е., Зарубина Т.В. Автоматизированная система извлечения клинически релевантных терминов UMLS из текстов англоязычных статей на примере аксиального спондилоартрита // *Электронный научный журнал «Социальные аспекты здоровья населения»*. — 2023. — Т. 69. — № 3. — С. 14. [Astanin PA, Rauzina SE, Zarubina TV. Automated system for recognizing clinically relevant UMLS terms in texts of the englishlanguage articles exemplified by axial spondyloarthritis. *Social'nye aspekty zdorov'ya naseleniya [serial online]*. 2023;69(3):14. (In Russ.)] doi: 10.21045/2071-5021-2023-69-3-14. Available from: http://vestnik.mednet.ru/content/view/1491/30/lang.ru (accessed: 02.10.2023).
44. Astanin PA, Rauzina SE, Zarubina TV. *Digital Tools in UMLS Metathesaurus Knowledge Processing*. Studies in Health Technology and Informatics; 2023 Jun 29; Athens, Greece; 2023;305:186–189. doi: https://doi.org/10.3233/SHTI230458

КОНТАКТНАЯ ИНФОРМАЦИЯ

Раузина Светлана Евгеньевна, к.м.н., доцент [***Svetlana E. Rauzina***, MD, PhD, Associate Professor];
адрес: 117997, Москва, ул. Островитянова, д. 1 [**address**: 1 Ostrovitianov str., Moscow, 117997, Russia];
e-mail: rauzina@mail.ru, **SPIN-код**: 1164-3516, **ORCID**: https://orcid.org/0000-0002-9535-2847

Зарубина Татьяна Васильевна, д.м.н., профессор, член-корреспондент РАН [***Tat'yana V. Zarubina***, MD, PhD, Professor, Corresponding Member of the RAS]; **e-mail**: t_zarubina@mail.ru, **SPIN-код**: 4354-3290, **ORCID**: https://orcid.org/0000-0002-4403-8049

Астанин Павел Андреевич, аспирант [***Pavel A. Astanin***, PhD Student]; **e-mail**: med_cyber@mail.ru, **SPIN-код**: 2658-1189, **ORCID**: https://orcid.org/0000-0002-1854-8686

Королева Юлия Ивановна, к.м.н., доцент [***Julia I. Koroleva***, MD, PhD, Associate Professor];
e-mail: koroleva_iui@rsmu.ru, **SPIN-код**: 4950-1854, **ORCID**: https://orcid.org/0000-0003-3560-312X

Ронжин Лев Вячеславович, аналитик [***Lev V. Ronzhin***, Analyst]; **e-mail**: izopropylbenzol@gmail.com, **SPIN-код**: 5502-6792, **ORCID**: https://orcid.org/0000-0002-4653-1611

Борисов Александр Александрович, аналитик [***Alexsandr A. Borisov***, Analyst]; **e-mail**: aleksandrborisov10650@gmail.com, **SPIN-код**: 4294-4736, **ORCID**: https://orcid.org/0000-0003-4036-5883

Афанасьева Мария Александровна [***Maria A. Afanasyeva***]; **e-mail**: usovanastasija@gmail.com, **ORCID**: https://orcid.org/0009-0002-2680-6291

Усова Анастасия Владимировна [***Anastasia V. Usova***]; **e-mail**: usovanastasija@gmail.com, **SPIN-код**: 9109-8782, **ORCID**: https://orcid.org/0009-0001-0625-1105