**АСТАНИН П.А.,**
ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России, Москва, Россия, e-mail: med_cyber@mail.ru

**РОНЖИН Л.В.,**
ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России, Москва, Россия, e-mail: levronzhin@gmail.com

**РАУЗИНА С.Е.,**
к.м.н., доцент, ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России, Москва, Россия,
e-mail: rauzina@mail.ru

# АЛГОРИТМ ОЦЕНКИ СПЕЦИФИЧНОСТИ ТЕРМИНОВ МЕТАТЕЗАУРУСА UMLS НА ПРИМЕРЕ АНАЛИЗА СЕМАНТИЧЕСКОЙ МОДЕЛИ ДЛЯ ДИФФЕРЕНЦИАЛЬНОЙ ДИАГНОСТИКИ АКСИАЛЬНОГО СПОНДИЛОАРТРИТА

**DOI:** 10.25881/18110193_2023_3_30

***Аннотация.*** *Актуальность. Ранняя диагностика аксиального спондилоартрита (акссА) является актуальной задачей, требующей поддержки принятия клинических решений (ППКР). В настоящее время значительная часть данных, применимых для разработки систем ППКР, представлена в неструктурированном виде. Семантический анализ медицинских текстов – сложная задача, требующая создания универсальных инструментов для извлечения именованных сущностей. Крупнейшим источником данных для аннотирования биомедицинских текстов является Unified Medical Language System (UMLS) – международный метатезаурус, включающий свыше 11 млн вариантов написания 4,6 млн терминов (концептов). Ключевой проблемой при использовании UMLS для анализа медицинских текстов является наличие большого количества неспецифичных (общих) терминов, не имеющих явного клинического смысла. Применение таких концептов приводит к значительному ухудшению результатов поиска, что указывает на необходимость создания инструментов автоматической оценки степени специфичности терминов UMLS.*

*Цель исследования. Разработка алгоритма для оценки степени специфичности терминов метатезауруса UMLS (на примере аксиального спондилоартрита).*

*Методы и материалы. В качестве источника информации для автоматического поиска клинических терминов использовались аннотации к англоязычным научным статьям. Тексты аннотаций извлекались с применением средств поисковой системы PubMed и помещались в единый электронный корпус, использованный для последующего извлечения концептов метатезауруса UMLS. Каждый из 24276 текстов корпуса имел однозначную метку принадлежности к одному из заболеваний дифференциального ряда для акссА. В общий свод включено 8260 концептов, каждый из которых получил экспертную бинарную метку относительной специфичности.*

*Результаты. Сформирован набор правил, основанных на сравнении средних длин иерархических цепей атомарных формулировок терминов, общего числа прямых связей, TF-IDF меры и числа связей «родитель-потомок» концептов UMLS. Данные правила включены в итоговый алгоритм оценки специфичности концептов, точность которого при попарном сравнении составила 99,1% для тестовой выборки. Однако точность модели при бинарной классификации всех концептов из выделенного свода терминов составила 74,2%, что является недостаточным для обоснования его применения при автоматическом сокращении терминологических сводов большого объема. Сформированы критерии и ограничения для использования разработанного алгоритма в процессе аннотации клинических документов.*

***Ключевые слова:*** *UMLS, метатезаурус, семантический анализ текста, граф, анализ естественного языка, аксиальный спондилоартрит.*

***Для цитирования:*** *Астанин П.А., Ронжин Л.В., Раузина С.Е. Алгоритм оценки специфичности терминов метатезауруса UMLS на примере анализа семантической модели для дифференциальной диагностики аксиального спондилоартрита. Врач и информационные технологии. 2023; 3: 30-43. doi: 10.25881/18110193_2023_3_30.*

**ASTANIN P.A.,**
Pirogov Russian National Research Medical University, Moscow, Russia, e-mail: med_cyber@mail.ru

**RONZHIN L.V.,**
Pirogov Russian National Research Medical University, Moscow, Russia, e-mail: levronzhin@gmail.com

**RAUZINA S.E.,**
PhD, Associate Professor, Pirogov Russian National Research Medical University, Moscow, Russia,
e-mail: rauzina@mail.ru

# ALGORITHM FOR UMLS METATHESAURUS CONCEPTS SPECIFICITY ESTIMATION USING EXAMPLE OF ANALYSIS OF THE SEMANTIC MODEL DESCRIBING AXIAL SPONDYLOARTHRITIS DIFFERENTIAL DIAGNOSTICS

**DOI:** 10.25881/18110193_2023_3_30

***Abstract.*** *Background. Early axial spondyloarthritis (axSpA) diagnostics is a difficult task requiring clinical decision support (CDS) making. Currently, there is a big unstructured data applicable in CDS systems development. Semantic data analysis is a complex issue to solve, and unified tools for named entity recognition are required. The biggest data source for biomedical text annotation is the Unified Medical Language System (UMLS) Metathesaurus. It includes more than 11 million atomic terms for writing of 4.6 million concepts. The main issue in UMLS using for medical text analysis is a presence of numerous unspecified (generic) terms without any clinical value. Their application leads to significant decrease of searching results. That is why tools for automatic specificity degree estimation are needed to be developed.*

*Aim. To develop an algorithm for specificity degree estimation for UMLS metathesaurus concepts (using example of axial spondyloarthritis).*

*Methods. English clinical abstracts have been used as data source for automatic UMLS named entity recognition. They have been extracted using free search engine PubMed followed by integration into single electronic corpus. Then each of 24276 texts in corpus has been labeled (affiliated with one of diagnosis in differential list for axSpA) and used for UMLS concepts mapping. A total of 8260 UMLS concepts have been recognized. Each term received an expert binary label of relative specificity.*

*Results. Rules for concepts specificity degree estimation have been developed based on comparison of 4 parameters: mean length of hierarchical chain, total count of direct relationships, TF-IDF score and count of hierarchical relationships with child concepts UMLS. These rules have been integrated into the total algorithm for UMLS concepts specificity degree estimation. Its accuracy was 99,1% for test data sample for paired comparisons. But its accuracy for solid comparison of all extracted concepts was 74,2%, which less than desirable for substantiation of this algorithm use for automatically terms big sets cutbacks. That is why some limitations for developed algorithm have been outlined.*

***Keywords****: UMLS, metathesaurus, semantics, graph, natural language processing, axial spondyloarthritis.*

***For citation:*** *Astanin P.A., Ronzhin L.V., Rauzina S.E. Algorithm for UMLS metathesaurus concepts specificity estimation using example of analysis of the semantic model describing axial spondyloarthritis differential diagnostics. Medical doctor and information technology. 2023; 3: 30-43. doi: 10.25881/18110193_2023_3_30.*

## ВВЕДЕНИЕ

Аксиальный спондилоартрит (аксСпА) — это хроническое воспалительное заболевание с преимущественным поражением позвоночника и крестцово-подвздошных суставов [1]. Важной особенностью аксСпА является вариабельность клинических признаков, приводящая к формированию обширного дифференциально-диагностического ряда [2]. Основным и наиболее частым проявлением аксСпА является боль в спине, сопровождающаяся ощущениями утренней скованности движений [3–4]. Нередко у пациентов с данным заболеванием могут наблюдаться внеаксиальные и внескелетные проявления [5–6]. К внеаксиальным поражениям относятся патогенетически связанные с аксСпА заболевания опорно-двигательного аппарата: периферические артриты и энтезиты [7]. Наиболее распространенными внескелетными проявлениями являются увеиты, поражения кожи, а также воспалительные заболевания кишечника, сердца и сосудов [8–10].

Выбор тактики ведения пациентов с неустановленным диагнозом аксСпА на уровне первичного звена системы здравоохранения является сложной междисциплинарной проблемой, требующей персонифицированного подхода [11–13]. В значительной мере, своевременному выявлению аксСпА препятствуют системный характер и отсутствие патогномоничных признаков данного заболевания [14]. Период времени от начала возникновения первых клинических проявлений аксСпА до окончательного подтверждения диагноза, по разным оценкам, составляет 5–8 лет [15]. Определенные надежды на повышение качества ранней диагностики аксСпА возлагают на разработку и внедрение систем поддержки принятия клинических решений (СППКР).

Создание СППКР для дифференциальной диагностики системных заболеваний требует использования больших данных, в полной мере описывающих широкое признаковое пространство [16–18]. Значительная часть клинических данных, применимых при разработке подобных систем, хранится в электронных медицинских картах и представлена в неструктурированном (текстовом) виде [19–20]. Ключевой этап разработки СППКР, основанных на анализе неструктурированного текста, заключается в создании терминологического свода, описывающего изучаемую клиническую область [21–24]. Обязательным требованием служит строгое соответствие используемых понятий общепринятым согласованным номенклатурам [25–26].

В настоящее время крупнейшим систематизированным сводом терминов, предназначенным для описания различных областей биологии и медицины, является Unified Medical Language System (UMLS) [27]. Актуальная версия метатезауруса UMLS (2022AB) содержит свыше 4,6 млн концептов — клинических и параклинических понятий, представленных в 76 справочниках. Каждый концепт UMLS имеет собственные варианты написания, а также внутреннюю систему приоритетов и статусов терминов. Общее число англоязычных и русскоязычных атомарных формулировок написания для всех концептов UMLS превышает 11,2 млн.

Практически каждый концепт UMLS связан как минимум с одним другим концептом, что позволяет представить данный свод терминов в виде гигантского ориентированного мультиграфа [28]. Разработка инструментов автоматического анализа графовой информационной модели UMLS внесет значительный вклад в создание неспецифичных универсальных средств обработки естественного языка (NLP). В свою очередь, использование технологий NLP обеспечит возможность решения клинических задач с использованием анализа неструктурированных текстов — основной формы представления и хранения данных в современных медицинских информационных системах [29–31].

Анализ метатезауруса UMLS — сложная и нетривиальная задача. Основной проблемой, связанной с извлечением терминов UMLS из текста, является наличие обобщающих понятий, вносящих существенный вклад в ухудшение результатов поиска. Разработка алгоритма оценки специфичности концептов метатезауруса UMLS позволит создать инструменты автоматической элиминации обобщающих терминов при извлечении именованных сущностей из текста. Использование указанных инструментов позволит значительно улучшить качество извлечения релевантной информации из графовой модели UMLS.

Целью настоящего исследования является разработка алгоритма оценки относительной специфичности концептов метатезауруса UMLS

(на примере терминов для дифференциальной диагностики аксиального спондилоартрита).

## МАТЕРИАЛ И МЕТОДЫ

Исследование проведено в рамках программы стратегического академического лидерства «Приоритет — 2030» на базе Института цифровой трансформации медицины (ИЦТМ) ФГАОУ ВО «Российский национальный исследовательский медицинский университет имени Н.И. Пирогова» Минздрава России. Используемый дифференциально-диагностический ряд для аксСпА включал 9 заболеваний костно-мышечной системы: стеноз поясничного отдела позвоночного канала (M48.0, M99.5–M99.7), спондилолистез поясничного отдела (M43.1), спинальная нестабильность (M53.2), миофасциальный болевой синдром (M79.1), инфекционные поражения структур позвоночника (M49.0–M49.2, M49.3, M86), новообразования в области поясничного и крестцового отделов позвоночника (D16.6, D16.8), болезнь Форестье (M48.1), болезнь Педжета (M88) и анкилозирующий спондилит (M45) [32–35].

Для извлечения именованных сущностей применялись международные клинические справочники, представленные в актуальной версии метатезауруса UMLS (2022AB). В качестве источника информации для автоматического поиска концептов использованы тексты 24276 аннотаций к англоязычным статьям по исследуемой клинической области. При формировании корпуса текстов для всех кодов МКБ-10, заявленных для перечисленных ранее заболеваний, извлекались все англоязычные формулировки из UMLS. Затем в поисковой системе PubMed автоматически создавались запросы с использованием в качестве ключевых слов полученных вариантов написания вышеперечисленных нозологий. На основании привязки к ключевым словам в PubMed каждый извлекаемый текст аннотации получал собственную метку связи с определенным заболеванием из дифференциального ряда для аксСпА. Тексты аннотаций сохранялись на жесткий диск, подвергались предобработке и извлечению именованных сущностей. Подробное описание данной части исследования дано в одной из предыдущих работ [36].

В ходе обработки всего корпуса текстов сформированы пересекающиеся своды терминов, описывающих клинические особенности заболеваний из дифференциального ряда. В общей сложности в созданную номенклатуру понятий включено 8260 концептов UMLS. Доля понятий, присущих только одному заболеванию (условно-патогномоничных в рамках дифференциально-диагностического ряда), составила ~41% от общего числа извлеченных терминов.

Каждому извлеченному термину экспертным способом присвоена бинарная метка относительной специфичности. В зависимости от значения данной метки были сформированы две статистические группы терминов. В первую группу включены относительно специфичные (частные) термины, во вторую — относительно неспецифичные (общие). Примеры пар относительно частных и общих терминов представлены в таблице 1.

**Таблица 1 — Примеры пар относительно неспецифичных и относительно специфичных терминов UMLS**

| № | Неспецифичный (общий) термин | Специфичный (частный) термин |
|---|---|---|
| 1 | Боль в спине неуточненной локализации | Боль в пояснице |
| 2 | Артрит | Артрит крестцово-подвздошного сустава |
| 3 | Симптом | Симптомы заболеваний желудочно-кишечного тракта |
| 4 | Системные заболевания | Болезнь Рейтера |
| 5 | Заболевания мягких тканей | Инфекции мягких тканей |
| 6 | Нарушение баланса электролитов | Гипокалиемия |
| 7 | Некроз | Сухая гангрена стопы |
| 8 | Неуточненное заболевание молочной железы | Мастопатия |
| 9 | Отсутствие признака | Отсутствие аппетита |
| 10 | Признаки | Признаки поражения кожи |
| 11 | HLA-антигены | Положительный результат теста на HLA-B27 |
| 12 | Аксиальный спондилоартрит | Анкилозирующий спондилит |

Сформулирован перечень из пяти количественных параметров для сравнительной оценки степени специфичности терминов UMLS. Первый параметр отражает плотность структуры связей концепта в графовой модели и определяется как количество прямых связей между исследуемым и соседними узлами.

Второй параметр характеризует положение концепта в иерархии графовой модели UMLS. Расчет значений данного параметра осуществляется с использованием специальной таблицы Mrhier оригинального метатезауруса. Указанная таблица содержит информацию об источнике концепта, его положении в локальной иерархии терминов, формулировке родительского термина и уточнении связей. Каждая запись в таблице Mrhier позволяет определить, к какой иерархической цепочке привязана отдельно взятая формулировка UMLS. Положение концепта в иерархии графовой модели метатезауруса рассчитывается как среднее значение длин иерархических цепей (или числа связей), соединяющих корневое понятие с концевым концептом, включающим атомарные формулировки соответствующего понятия. Простейший пример иерархических цепей для термина «боль в нижней части спины» представлен на рисунке 1.

Согласно данным, представленным на рисунке 1, термин «боль в нижней части спины» является концевым элементом трех иерархических цепей. Так, для цепи с корневым термином «симптомы и признаки» количество связей в иерархической цепи равно трём. Для цепи с корневым термином «Анатомические образования» длина иерархической цепи будет равна двум. Для цепи с корневым термином «Клинические находки» длина цепи будет равна единице. Таким образом, среднее значение длин иерархических цепей для термина «боль в нижней части спины» равняется двум. Подобным образом значения количественной меры, отражающей положение концепта в иерархической структуре, можно рассчитать для всех терминов UMLS.

Третьим параметром для оценки специфичности терминов UMLS является рассчитанное значение TF-IDF меры. Математический смысл TF-IDF меры состоит в том, что вес токена (отдельного слова, фразы или термина) прямо пропорционален TF (term frequency) — частоте употребления в документе и обратно пропорционален IDF (inverse document frequency) — частоте

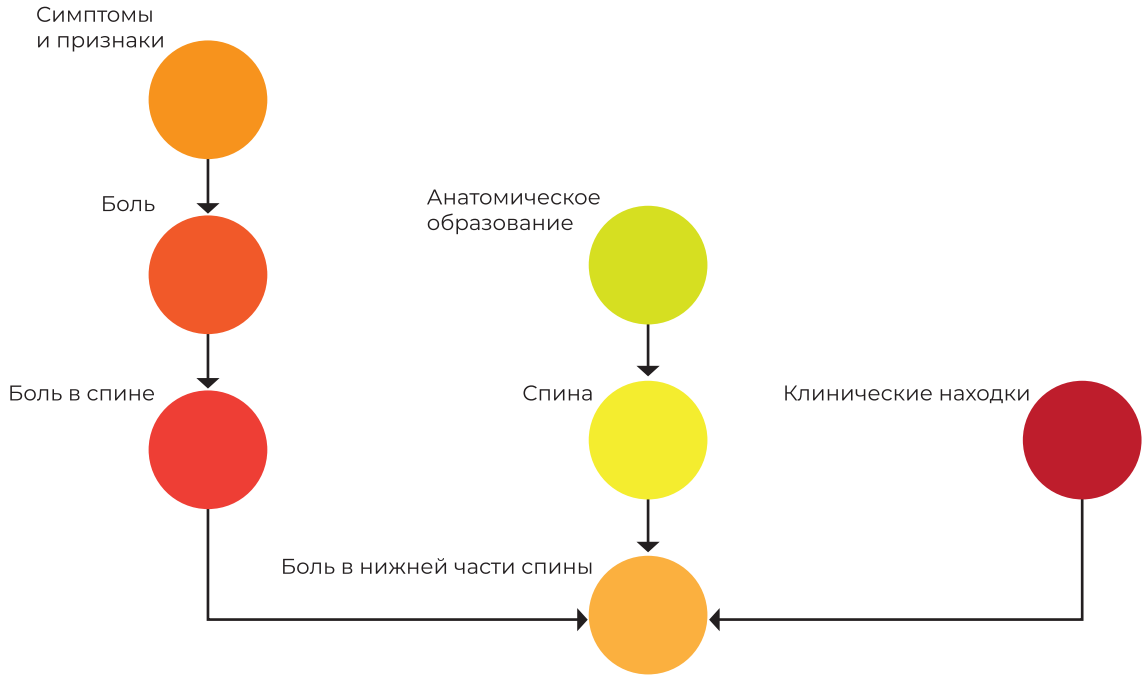



**Рисунок 1 — Примеры иерархических цепей для термина «боль в нижней части спины».**

использования во всех документах корпуса [37]. TF-IDF мера, использованная в данном исследовании для определения степени специфичности терминов UMLS, рассчитывалась по следующей формуле:

$$TF - IDF_i = \frac{n_i}{m_i} \cdot \ln \frac{D}{d_i} \quad (1),$$

где: $n_i$ — количество токенов i в корпусе документов; $m_i$ — общее число токенов в документах, в которых токен i встретился хотя бы один раз; D — общее количество документов в корпусе; $d_i$ — количество документов, в которых токен i встретился хотя бы один раз.

Основным источником информации, необходимой для расчета значений TF-IDF, является NLM (U. S. National Library of Medicine) — Национальная Медицинская Библиотека США. Одним из важнейших инструментов, разработанных NLM, является SemRep (Semantic Repository) — средство для автоматического извлечения связей между концептами UMLS из неструктурированного текста. Одним из результатов работы SemRep является свободно распространяемая база данных SemMedDB (Semantic Medline Database), созданная на основе машинной обработки массивных корпусов текстов на английском языке. Данные, использованные в настоящем исследовании для определения значений TF-IDF мер, содержались в базе данных SemMedDB в таблице Predication. Каждой записи в указанной таблице сопоставлен уникальный номер публикации (PMID), служащей в качестве источника извлеченной связи. Для расчета TF-IDF мер агрегированы записи со значениями идентификаторов корневого и концевого концептов, а также PMID соответствующих англоязычных статей из поисковой системы PubMed.

Четвертый параметр позволяет оценить количество дочерних терминов изучаемого концепта. Для расчета данного параметра определяется число прямых связей, относящихся к типам CHD — child relationships (связи родительских терминов с дочерними) и RN — narrower relationships (связи более широких терминов с более узкими) [36]. Предполагается, что указанные типы вертикальных иерархических связей соединяют более общие и неспецифичные термины с более частными и специфичными.

Пятый параметр определяет семантическую сложность термина. В качестве значения данного параметра в настоящем исследовании использовалось среднее число слов, содержащихся во всех вариантах написания концепта. Параметр определялся, исходя из предположения, что более специфичные термины содержат больше языковых единиц.

Первичный анализ качества выделенных параметров осуществлялся с использованием методов непараметрической статистики. Для оценки типа распределения количественных признаков применялся критерий Колмогорова-Смирнова. Однако, поскольку вид распределения всех проверяемых признаков отличался от нормального, для описания количественных данных рассчитывались медиана и межквартильный размах (Me [$Q_1$; $Q_3$]). Для оценки структуры исследуемых групп определялись доли и ошибки долей с последующим выражением в процентах (p±M%).

Для валидации алгоритма оценки специфичности терминов UMLS и его составных частей использованы дополнительные данные, представленные 424 предварительно размеченными парами концептов, не пересекающихся с номенклатурой терминов для описания заболеваний из дифференциально-диагностического ряда для аксСпА. Основным условием для формирования пар было наличие пересечений хотя бы по одному из токенов. В качестве примера пары терминов, подходящих для ранжирования по степени специфичности, можно выделить «неуточненную боль в области живота» и «боль в правом боку» с общим токеном «боль». В каждой паре тремя независимыми экспертами выделено по одному относительно специфичному и относительно неспецифичному концепту. Было предложено исключать пары терминов, не подлежащих ранжированию, однако ни один из экспертов не посчитал необходимым элиминировать хотя бы одну из пар в автоматически сформированном своде.

Качество экспертных оценок подтверждалось с применением непараметрического критерия $X^2$ Пирсона. Количественная оценка согласованности экспертных мнений производилась с использованием коэффициента ассоциации по следующей формуле:

$$K_a = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c} \quad (2),$$

где: a — число истинно положительных оценок, b — число ложноположительных оценок, c — число ложноотрицательных оценок, d — число истинно отрицательных оценок.

Экспертные оценки, полученные в ходе семантического анализа пар концептов, характеризовались высокой степенью согласованности (таблица 2). Итоговое решение об используемом варианте разметки пар терминов UMLS принималось по принципу наибольшего количества голосов.

Анализ прогностической способности правил, основанных на использовании выделенных параметров для сравнительной оценки степени специфичности терминов, осуществлялся путем вычисления стандартных метрик бинарной классификации: точности, чувствительности и специфичности. Статистически значимыми считались результаты проверки гипотез при уровне значимости $p<0,05$.

## РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

При оценке межгрупповых различий по параметрам для определения степени специфичности терминов UMLS в экспертно размеченных группах статистически значимые различия выявлены лишь по четырём из пяти в сформированном перечне (таблица 3).

Согласно данным, представленным в таблице 3, медиана общего числа прямых связей концептов UMLS оказалась выше в группе относительно неспецифичных терминов. Схожая тенденция наблюдалась по числу иерархических связей, соединяющих родительские термины с дочерними. Медиана по средним значениям длин иерархических цепей оказалась выше в группе относительно специфичных терминов. Аналогичная закономерность характерна для TF-IDF меры. Все полученные закономерности оправдали изначальные ожидания и подтвердили возможность использования данных параметров для оценки специфичности терминов UMLS. Крайне важно отметить, что не были выявлены статистически значимые различия по среднему числу слов в атомарных формулировках концептов. Вопреки ожиданиям, данный параметр не продемонстрировал наличия потенциальных статистических закономерностей и, в

**Таблица 2 — Матрица согласованности экспертных мнений ($K_a$)**

| Номер эксперта | 1 | 2 | 3 |
|---|---|---|---|
| 1 | | $K_a$ = 0,730; p<0,001 | $K_a$ = 0,744; p<0,001 |
| 2 | $K_a$ = 0,730; p<0,001 | | $K_a$ = 0,803; p<0,001 |
| 3 | $K_a$ = 0,744; p<0,001 | $K_a$ = 0,803; p<0,001 | |

**Таблица 3 — Оценка межгрупповых различий по параметрам для определения степени специфичности терминов UMLS**

| Краткое описание параметра | Группа терминов UMLS (Me [Q1; Q3]) | | P |
|---|---|---|---|
| | Относительно неспецифичные (общие) термины (n = 4099, 49,6%) | Относительно специфичные (частные) термины (n = 4161, 50,4%) | |
| Общее количество прямых связей концепта | 52,0 [35,0; 158] | 44,0 [24,0; 121] | 0,026 |
| Среднее значение длин иерархических цепей атомарных формулировок концепта | 4,60 [4,00; 5,83] | 5,95 [4,74; 7,37] | <0,001 |
| Значение TF-IDF меры для концепта | 1,80 [1,41; 2,33] | 1,97 [1,57; 2,43] | <0,001 |
| Количество иерархических связей «родительский термин – дочерний термин» концепта | 15,0 [2,00; 81,0] | 12,0 [3,00; 42,0] | <0,001 |
| Среднее число слов в атомарных формулировках концепта | 2,43 [1,85; 2,97] | 2,42 [1,89; 3,04] | 0,468 |

связи с этим, не рассматривался при последующем анализе.

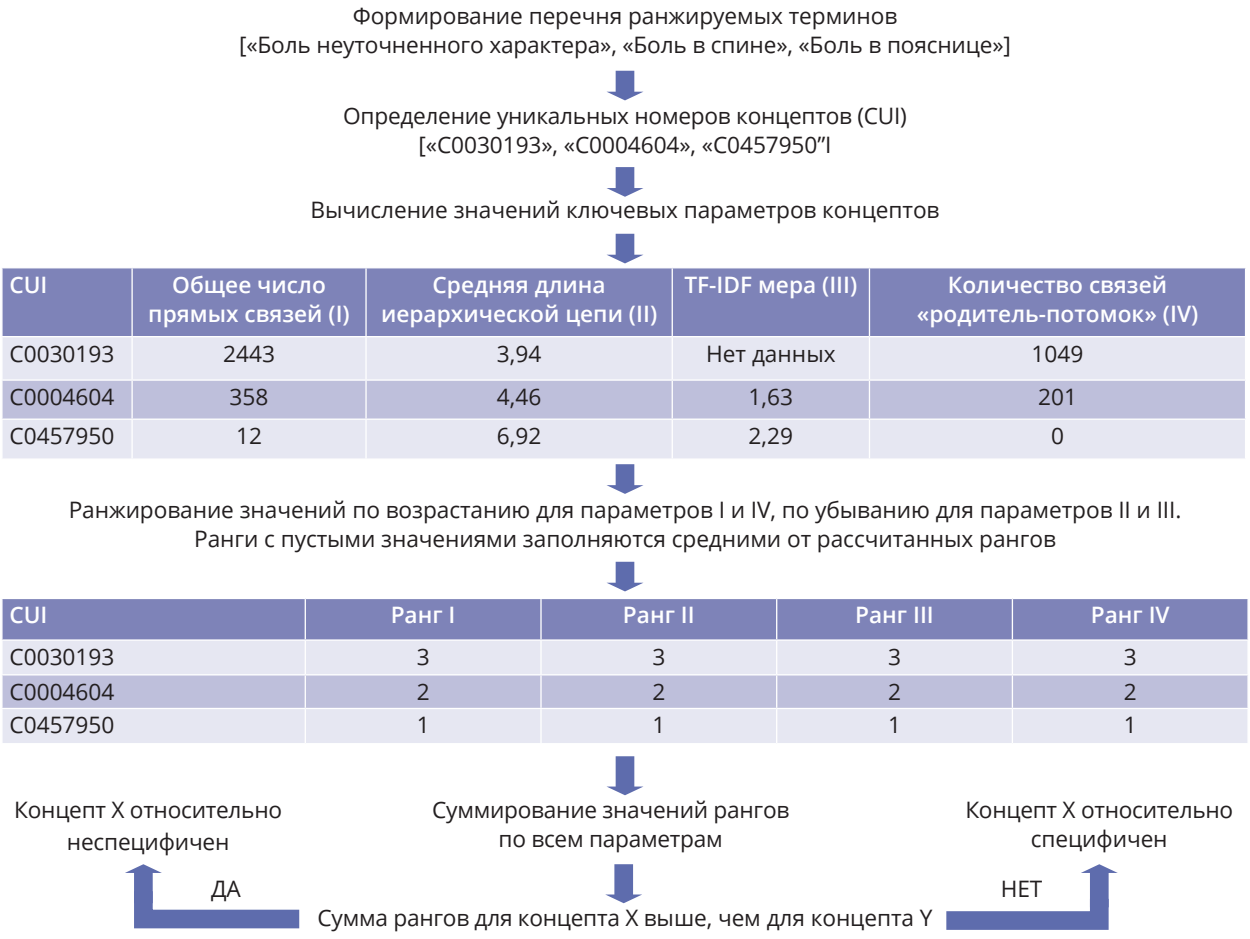
На основании результатов из таблицы 3 сформулированы бинарные правила для оценки степени относительной специфичности терминов UMLS. При анализе качества выделенных правил определялись следующие количественные характеристики: точность, чувствительность, специфичность, а также доля концептов, для которых является возможным применение соответствующего правила. Метрики бинарной классификации рассчитывались для набора размеченных экспертным способом пар терминов. Доли концептов, для которых возможно применение правил, рассчитывались для всего метатезауруса UMLS.

Итоговый алгоритм оценки специфичности терминов UMLS представлен на рисунке 2.

Результат проверки качества отдельных правил и итогового алгоритма представлены в таблице 4.

Из данных таблицы 4 следует, 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 формулировок сравниваемых концептов. Точность итогового алгоритма составила 99,1% для тестовой выборки.

Предпринята попытка бинарной классификации терминов в исходно сформированном




| CUI | Общее число прямых связей (I) | Средняя длина иерархической цепи (II) | TF-IDF мера (III) | Количество связей «родитель-потомок» (IV) |
|---|---|---|---|---|
| C0030193 | 2443 | 3,94 | Нет данных | 1049 |
| C0004604 | 358 | 4,46 | 1,63 | 201 |
| C0457950 | 12 | 6,92 | 2,29 | 0 |

| CUI | Ранг I | Ранг II | Ранг III | Ранг IV |
|---|---|---|---|---|
| C0030193 | 3 | 3 | 3 | 3 |
| C0004604 | 2 | 2 | 2 | 2 |
| C0457950 | 1 | 1 | 1 | 1 |

**Рисунок 2 — Алгоритм оценки относительной специфичности терминов метатезауруса UMLS.**

**Таблица 4 — Правила для оценки степени относительной специфичности терминов метатезауруса UMLS**

| № | Логическое правило | Метрики качества бинарной классификации, % | | | Число и доля (n, %) концептов, для которых возможно применение правила |
|---|---|---|---|---|---|
| | | А | Б | В | |
| 1 | Количество прямых связей неспецифичного концепта больше, чем у специфичного | 97,2 | 97,7 | 96,6 | 4548967 (98,6%) |
| 2 | Длина иерархической цепи неспецифичного концепта меньше, чем у специфичного | 91,3 | 92,7 | 89,8 | 3600628 (78,1%) |
| 3 | TF-IDF мера неспецифичного концепта меньше, чем у специфичного | 83,5 | 85,8 | 81,1 | 300459 (6,51%) |
| 4 | Количество связей «родитель-потомок» неспецифичного концепта больше, чем у неспецифичного | 97,9 | 97,7 | 98,1 | 864833 (18,8%) |
| 5 | Итоговый алгоритм | 99,1 | 99,1 | 99,0 | 864833 (18,8%) |

**Примечание**: А — Точность, Б — Чувствительность, В — Специфичность.

терминологическом своде для описания клинических особенностей заболеваний из дифференциально-диагностического ряда для аксСпА. Точность модели при бинарной классификации всех 8260 концептов составила 74,2%, что является недостаточным для обоснования его применения при автоматическом сокращении терминологических сводов большого объема.

Таким образом, внедрение разработанного алгоритма для использования при аннотации клинических документов требует предварительного формирования перечня критериев и ограничений. Одним из таких критериев может стать наличие пересечений токенов в извлеченных терминах. Также допускается использование алгоритма для ранжирования небольших наборов терминов UMLS, связанных прямыми ассоциативными связями.

## ОБСУЖДЕНИЕ

Автоматическое аннотирование неструктурированных медицинских текстов представляет большой практический интерес для современной системы здравоохранения [38]. На сегодняшний день крупнейшим источником биомедицинских терминов, пригодных для использования при обработке клинических текстов, является метатезаурус UMLS. Однако процесс извлечения именованных сущностей с применением UMLS сопровождается ключевой проблемой, связанной с агрегацией избыточных и обобщающих понятий, не относящихся к признакам исследуемых заболеваний. В нашей работе доля таких терминов составила около 50%. Использование подобных терминов при решении информационно-поисковых задач без предварительной обработки и ранжирования приводит к значительному ухудшению качества извлекаемой информации.

В настоящем исследовании была предпринята попытка создания универсального алгоритма для ранжирования терминов по степени их относительной специфичности. Выделены основные параметры, используемые при создании правил ранжирования и базируемые на анализе семантических характеристик терминов UMLS. Необходимо отметить, что наименьшую эффективность продемонстрировал параметр, основанный на оценке среднего числа слов, входящих в термины исследуемого концепта. Данный факт указывает на низкое качество отбора специфичных концептов при использовании правил, основанных только на подстрочном поиске. Остальные параметры, описанные в настоящей работе (общее количество прямых связей, средняя длина иерархической цепи, TF-IDF мера и количество связей «родитель-потомок») демонстрировали высокую точность и были использованы при создании итогового алгоритма. При попарном сравнении концептов точность данного алгоритма составила 99,1%, однако по мере увеличения числа ранжируемых концептов

качество оценки относительной специфичности заметно снижается из-за увеличения накопленной ошибки. В связи с этим применение разработанного алгоритма для элиминации обобщающих терминов из сводов большого объема не представляется возможным. Предполагается, что максимальная эффективная работа алгоритма будет достигнута при ранжировании небольших наборов терминов. В настоящее время разработанный алгоритм внедрен в аналитические сервисы информационно-поисковой системы, разрабатываемой на базе Института цифровой трансформации медицины РНИМУ им. Н.И. Пирогова.

## ЗАКЛЮЧЕНИЕ

Разработанный алгоритм для оценки относительной специфичности терминов метатезауруса UMLS продемонстрировал высокую точность (99,1%) при попарном сравнении концептов с пересечением токенов, что указывает на широкие возможности использования данного алгоритма при аннотировании клинических текстов. Одним из перспективных направлений для продолжения настоящего исследования является создание системы поддержки принятия решений для дифференциальной диагностики аксиального спондилоартрита с использованием анализа неструктурированного текста.

**Конфликт интересов:** Авторы заявляют об отсутствии конфликта интересов.

**Финансирование:** Исследование выполнено в рамках федеральной программы «Приоритет 2030».

**Благодарность.** Авторы выражают признательность сотрудникам Института цифровой трансформации медицины (ИЦТМ), Ядгаровой П.А., Николаиди Е.Н., Потаповой И.И., за экспертную работу по разметке валидационного набора данных. Авторы выражают отдельную благодарность директору ИЦТМ, Зарубиной Т.В., за помощь в общей корректуре и редактуре текста данной статьи.

## ЛИТЕРАТУРА/REFERENCES

1. Моисеев С.В., Новиков П.И., Гуляев С.В. и др. Анкилозирующий спондилит: подходы к диагностике и клиническая эффективность упадацитиниба // Клиническая фармакология и терапия. — 2021. — Т.30. — №4. — С.62-70. [Moiseev S, Novikov P, Gulyaev S, et al. Ankylosing spondylitis: diagnostic challenges and efficacy of upadacitinib. Klinicheskaja farmakologija i terapija. 2021: 30(4): 62-70. (In Russ.)] doi: 10.32756/0869-5490-2021-4-62-70.
2. Кричевская О.А., Дубинина Т.В., Ильиных Е.В. и др. Боль в спине и оценка активности анкилозирующего спондилита на фоне гестации: симптомы, отражающие обострение заболевания, и беременность // Современная ревматология. — 2022. — Т.16. — №5. — С.60-65. [Krichevskaya OA, Dubinina TV, Ilinykh EV, et al. Back pain and assessment of ankylosing spondylitis activity during gestation: symptoms reflecting exacerbation of the disease and pregnancy. Modern Rheumatology Journal. 2022: 16(5): 60-65. (In Russ.)] doi: 10.14412/1996-7012-2022-5-60-65.
3. Дубинина Т.В., Демина А.Б. Методы лучевой диагностики как инструмент мониторинга аксиального спондилоартрита в реальной клинической практике // Современная ревматология. — 2022. — Т.16. — №1. — С.91-96. [Dubinina TV, Demina AB. Radiologic methods as a tool for monitoring axial spondyloarthritis in real clinical practice. Modern Rheumatology Journal. 2022: 16(1): 91-96. (In Russ.)] doi: 10.14412/1996-7012-2022-1-91-96.
4. Варавин Н.А., Верткин А.Л. Боль в спине в терапевтической практике // Лечащий врач. — 2022. — Т.25. — №7-8. — С.52-56. [Varavin NA, Vertkin AL. Back pain in therapeutic practice. The Attending Physician. 2022: 25(7-8): 52-56. (In Russ.)] doi: 10.51793/OS.2022.25.8.008.
5. Каратеев Д.Е., Степанова Е.А., Лучихина Е.Л. Методические рекомендации по рентгенологическим методам исследования при ревматоидном артрите и анкилозирующем спондилите // Эффективная фармакотерапия. — 2022. — Т.18. — №18. — С.12-25. [Karateev DE, Stepanova EA, Luchikhina EL. Practical guidelines for radiological investigation methods in rheumatoid arthritis and ankylosing spondylitis. Effective Pharmacotherapy. 2022: 18(18): 12-25. (In Russ.)] doi: 10.33978/2307-3586-2022-18-18-12-25.

6. Гайдук А.С., Железняк И.С., Тыренко В.В. и др. Цифровой томосинтез и другие методы визуализации в ранней диагностике аксиальных спондилоартритов: обзор литературы // Лучевая диагностика и терапия. — 2022. — №2(13). — С.25-35. [Gaiduk AS, Zheleznyak IS, Tyrenko VV, et al. Digital tomosynthesis and other visualization methods in the early detection of axial spondyloarthritis: a review. Diagnostic radiology and radiotherapy. 2022: 2(13): 25-35. (In Russ.)] doi: 10.22328/2079-5343-2022-13-2-25-35.
7. Гараева А.Р., Лапшина С.А., Анисимов В.И. и др. Клинико-инструментальная диагностика ранних изменений атлантоаксиальной области при анкилозирующем спондилите // Практическая медицина. — 2023. — Т.21. — №3. — С.68-73. [Garaeva AR, Lapshina SA, Anisimov VI, et al. Clinical and instrumental diagnostics of early changes in the atlantoaxial area in ankylosing spondylitis. Practical medicine. 2023: 21(3): 68-73. (In Russ.)] doi: 10.32000/2072-1757-2023-3-68-73.
8. Иванова Л.В., Акулинушкина Е.Ю., Лапшина С.А., Абдулганиева Д.И. Ранняя диагностика воспалительных заболеваний кишечника у пациентов со спондилоартритами // Практическая медицина. — 2023. — Т.21. — №2. — С.54-57. [Ivanova LV, Akulinushkina EYU, Lapshina SA, Abdulganieva DI. Early diagnosis of inflammatory bowel diseases in patients with spondyloarthritis. Practical medicine. 2023: 21(2): 54-57. (In Russ.)] doi: 10.32000/2072-1757-2023-2-54-57.
9. Моисеев С.В., Буланов Н.М. Аутоиммунитет, аутовоспаление и почки // Клиническая фармакология и терапия. — 2022. — Т.31. — №4. — С.7-17. [Moiseev S, Bulanov N. Autoimmunity, autoinflammation and kidney. Klinicheskaya farmakologiya i terapiya. 2022: 31(4): 7-17. (In Russ.)] doi: 10.32756/0869-5490-2022-4-7-17.
10. Пономарева М.Н., Карпова Д.А., Петров И.М. Анкилозирующий спондилит: гипотезы патогенеза, новые биомаркеры и особенности терапии // Современные проблемы науки и образования. — 2021. — №6. — С.188. [Ponomareva MN, Karpova DA, Petrov IM. Ankylosing spondilitis: hypotheses of pathogenesis, new biomarkers and features of therapy. Modern problems of science and education. 2021: 6: 188. (In Russ.)] doi: 10.17513/spno.31264.
11. Мартюшев-Поклад А.В., Гулиев Я.И., Казаков И.Ф. и др. Персоноцентрированные инструменты цифровой трансформации здравоохранения: пути совершенствования // Врач и информационные технологии. — 2021. — №S5. — С.4-13. [Martyushev-Poklad AV, Guliev YI, Kazakov IF, et al. Person-centered instruments in digital transformation of healthcare: ways to improve. Medical doctor and information technologies. 2021: S5: 4-13. (In Russ.)] doi: 10.25881/18110193_2021_S5_4.
12. Батудаева Т.И., Павлова А.Б., Лобышева Е.А., Арзуманян Э.А. Анализ стационарной помощи пациентам с анкилозирующим спондилитом // Вестник Бурятского государственного университета. Медицина и фармация. — 2022. — №1. — С.7-14. [Batudaeva TI, Pavlova AB, Lobysheva EA, Arzumanyan EA. Analysis of hospital care for patients with ankylosing spondylitis. Vestnik buryatskogo gosudarstvennogo universiteta. Meditsina i farmatsiya. 2022: 1: 7-14. (In Russ.)] doi: 10.18101/2306-1995-2022-1-7-14.
13. Шостак Н.А., Правдюк Н.Г., Новикова А.В. Боль в спине у молодых: алгоритм ведения в практике врача первичного звена // Клиницист. — 2022. — Т.16. — №3. — С.48-57. [Shostak NA, Pravdyuk NG, Novikova AV. Back pain in young people: algorithm of management in practice of primary physician. Analysis of hospital care for patients with ankylosing spondylitis. The Clinician. 2022: 16(3): 48-57. (In Russ.)] doi: 10.17650/1818-8338-2022-16-3-K674.
14. Лила А.М., Дубинина Т.В., Древаль Р.О. и др. Медико-социальная значимость и расчет экономического бремени аксиального спондилоартрита в Российской Федерации // Современная ревматология. — 2022. — Т.16. — №1. — С.20-25. [Lila AM, Dubinina TV, Dreval RO, et al. Medical and social significance and calculation of the economic burden of axial spondyloarthritis in the Russian Federation. Modern Rheumatology Journal. 2022: 16(1): 20-25. (In Russ.)] doi: 10.14412/1996-7012-2022-1-20-25.
15. Астанин П.А., Наркевич А.Н. Цифровые технологии в оценке течения заболеваний с выраженным болевым синдромом на примере анкилозирующего спондилита // Российский

журнал боли. — 2021. — Т.19. — №2. — С.38-41. [Astanin PA, Narkevich AN. Digital technology for estimation of course of diseases with acute pain syndrome on the example of ankylosing spondylitis. Russian Journal of Pain. 2021: 19(2): 38-41. (In Russ.)] doi: 10.17116/pain20211902138.

**16.** Киселев К.В., Ноева Е.А., Выборов О.Н. и др. Разработка алгоритма работы логического решателя интеллектуальной системы поддержки принятия врачебных решений для инструментальной диагностики стенокардии // Медицинские технологии. Оценка и выбор. — 2019. — №1(35). — С.32-42. [Kiselev KV, Noeva EA, Vyborov ON. Development of a reasoning solver algorithm for instrumental diagnostics of angina pectoris in intelligent clinical decision support system. Medical Technologies. Assessment and Choice. 2019: 1(35): 32-42. (In Russ.)] doi: 10.31556/2219-0678.2019.35.1.032-042.

**17.** Кобринский Б.А., Благосклонов Н.А., Демикова Н.С., и др. Компьютерная система для дифференциальной диагностики лизосомных болезней накопления на основе методов искусственного интеллекта // Бюллетень сибирской медицины. — 2022. — Т.21. — №2. — С.67-73. [Kobrinskii BA, Blagosklonov NA, Demikova NS, et al. An artificial intelligence computer system for differential diagnosis of lysosomal storage diseases. Bulletin of Siberian Medicine. 2022: 21(2): 67-73. (In Russ.)] doi: 10.20538/1682-0363-2022-2-67-73.

**18.** Орлова Н.В., Суворов Г.Н., Горбунов К.С. Этика и правовое регулирование использования больших баз данных в медицине // Медицинская этика. — 2022. — Т.10. — №3. — С.4-9. [Orlova NV, Suvorov GN, Gorbunov KS. Ethics and legal regulation of using large databases in medicine. Medical ethics. 2022: 10(3): 4-9. (In Russ.)] doi: 10.24075/medet.2022.056.

**19.** Шарова Д.Е., Михайлова А.А., Гусев А.В. Анализ мирового опыта в регулировании использования медицинских данных для целей создания систем искусственного интеллекта на основе машинного обучения // Врач и информационные технологии. — 2022. — №4. — С.28-39. [Sharova DE, Mikhailova AA, Gusev AV. An analysis of global experience in regulations on the use of medical data for artificial intelligence systems development based on machine learning. Medical doctor and information technologies. 2022: 4: 28-39. (In Russ.)] doi: 10.25881/18110193_2022_4_28.

**20.** Гусев А.В., Зингерман Б.В., Тюфилин Д.С., Зинченко В.В. Электронные медицинские карты как источник данных реальной клинической практики // Реальная клиническая практика: данные и доказательства. — 2022. — Т.2. — №2. — С.8-20. [Gusev AV, Zingerman BV, Tyufilin DS, Zinchenko VV. Electronic medical records as a source of real-world clinical data. Real-World Data & Evidence. 2022: 2(2): 8-20. (In Russ.)] doi: 10.37489/2782-3784-myrwd-13.

**21.** Гурдаева Н.А. Специальная лексика современного русского языка в свете теории функционально-семантического поля // Вестник Таганрогского государственного педагогического института. — 2012. — №2. — С.15-19. [Gurdaeva NA. Spetsial'naya leksika sovremennogo russkogo yazyka v svete teorii funktsional'no-semanticheskogo polya. Vestnik Taganrogskogo gosudarstvennogo pedagogicheskogo instituta. 2012: 2: 15-19. (In Russ.)]

**22.** Зарубина Т.В. Медицинская информатика: учебник / Под ред. Т.В. Зарубиной, Б.А. Кобринского. — Москва: ГЭОТАР-Медиа, 2022. [Zarubina TV. Medical informatics. 2nd ed. Moscow: GEOTAR-Media, 2022. (In Russ.)] doi: 10.33029/9704-6273-7-TMI-2022-1-464.

**23.** Осмоловский И.С., Зарубина Т.В., Шостак Н.А. и др. Разработка структуры базы знаний в области диагностики подагры // Сибирский журнал клинической и экспериментальной медицины. — 2022. — Т.37. — №3. — С.149-158. [Osmolovsky IS, Zarubina TV, Shostak NA. Development of knowledge base structure for gout diagnosis. The Siberian Journal of Clinical and Experimental Medicine. 2022: 37(3): 149-158. (In Russ.)] doi: 10.29001/2073-8552-2022-37-3-149-158.

**24.** Будыкина А.В., Тихомирова Е.В., Киселев К.В. и др. Формализация знаний о желудочно-кишечном кровотечении неясного генеза для использования в интеллектуальных системах поддержки принятия врачебных решений // Вестник новых медицинских технологий. — 2020. — Т.27. — №4. — С.98-101. [Budykina AV, Tikhomirova EV, Kiselev KV, et al. Formalization of knowledge about gastrointestinal bleeding of unknown origin for use in intelligent clinical decision support systems. Journal of New Medical Technologies. 2020: 27(4): 98-101. (In Russ.)] doi: 10.24411/1609-2163-2020-16741.

**25.** Колесникова О.И., Лопатина Е.В., Соколова В.В. Терминологические соответствия при переводе экономических терминов с английского на русский язык // Международный научно-исследовательский журнал. — 2021. — №1-3(103). — С.153-157. [Kolesnikova OI, Lopatina EV, Sokolova VV. Terminological correspondence in the translation of economic terms from English to Russian. Mezhdunarodnyi nauchno-issledovatel"skii zhurnal. 2021: 1-3(103): 153-157. (In Russ.)] doi: 10.23670/IRJ.2021.103.1.083.

**26.** Зацман И.М. Формы представления нового знания, извлеченного из текстов // Информатика и ее применения. — 2021. — Т.15. — №3. — С.83-90. [Zatsman IM. Forms representing new knowledge discovered in texts. Informatics and Applications. 2021: 15(3): 83-90. (In Russ.)] doi: 10.14357/19922264210311.

**27.** Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004; 32: 267-270. doi: 10.1093/nar/gkh061.

**28.** Астанин П.А. Применение автоматизированного анализа семантической сети UMLS для решения задачи поиска релевантных знаний о ревматических заболеваниях // Математическое моделирование систем и процессов. 2022. — С.6-12. [Astanin PA. Primenenie avtomatizirovannogo analiza semanticheskoi seti UMLS dlya resheniya zadachi poiska relevantnykh znanii o revmaticheskikh zabolevaniyakh. Matematicheskoe modelirovanie sistem i protsessov. 2022: 6-12. (In Russ.)] doi: 10.37490/978-5-00200-102-6-6-12.

**29.** Тучкова П.А. Применение методов обработки естественного языка для анализа текстовых и речевых данных в медицине // Наукосфера. — 2021. — №5-1. — С.174-179. [Tuchkova PA. Application of natural language processing methods for analysing of text and speech data in medicine. Naukosfera. 2021: 5-1: 174-179. (In Russ.)] doi: 10.5281/zenodo.4771893.

**30.** Сбоев А.Г., Селиванов А.А., Рыбка Р.Б. Современные методы экстракции связанных именованных сущностей на примере биомедицинских текстовых данных // Вестник Военного инновационного технополиса «Эра». — 2022. — Т.3. — №1. — С.57-67. [Sboev AG, Selivanov AA, Rybka RB. Sovremennye metody ekstraktsii svyazannykh imenovannykh sushchnostei na primere biomeditsinskikh tekstovykh dannykh. Vestnik voennogo innovatsionnogo tekhnopolisa «Era» (In Russ.)] doi: 10.56304/S2782375X22010193.

**31.** Гусев А.В., Владзимирский А.В., Голубев Н.А., Зарубина Т.В. Информатизация здравоохранения Российской Федерации: история и результаты развития // Национальное здравоохранение. — 2021. — Т.2. — №3. — С.5-17. [Gusev AV, Vladzimirskii AV, Golubev NA, Zarubina TV. Informatization of healthcare in the Russian Federation: history and results of development. National Health Care (Russia). 2021: 2(3): 5-17. (In Russ.)] doi: 10.47093/2713-069X.2021.2.3.5-17.

**32.** Никитина Н. М., Юпатова М. И., Ребров А. П. Проблемы остеопороза у пациентов с анкилозирующим спондилитом (обзор литературы) // Медицинский алфавит. — 2023. — №9. — С.40-45. [Nikitina NM, Yupatova MI, Rebrov AP. Problems of osteoporosis in patients with ankylosing spondylitis (literature review). Medical alphabet. 2023: 9: 40-45. (In Russ.)] doi: 10.33667/2078-5631-2023-9-40-45.

**33.** Годзенко А.А., Черемушкина Е.В., Димитрева А.Е., Урумова М.М. Сочетание анкилозирующего спондилита и ревматоидного артрита: клинические наблюдения и обзор литературы // Современная ревматология. — 2021. — Т.15. — №4. — С.72-80. [Godzenko AA, Cheremushkina EV, Dimitreva AE, Urumova MM. Combination of ankylosing spondylitis and rheumatoid arthritis: clinical observations and literature review. Modern Rheumatology Journal. 2021: 15(4): 72-80. (In Russ.)] doi: 10.14412/1996-7012-2021-4-72-80.

**34.** Нурполатова С.Т., Косымбетова А.Б., Джуманазарова Г.У. Боль в спине, как одна из проблем медицины // Бюллетень науки и практики. — 2021. — Т.7. — №6. — С.200-207. [Nurpolatova S, Kosymbetova A, Dzhumanazarova G. Back pain, as one of the problems of medicine. Bulletin of science and practice. 2021: 7(6): 200-207. (In Russ.)] doi: 10.33619/2414-2948/67/23.

**35.** Олюнин Ю.А., Лила А.М. Хроническая боль в спине глазами ревматолога // Современная ревматология. — 2022. — Т.16. — №5. — С.94-100. [Olyunin YuA, Lila AM. Chronic back pain from rheumatologist point of view. Modern Rheumatology Journal. 2022: 16(5): 94-100. (In Russ.)] doi: 10.14412/1996-7012-2022-5-94-100.

**36.** Астанин П.А., Раузина С.Е., Зарубина Т.В. Автоматизированная система извлечения клинически релевантных терминов UMLS из текстов англоязычных статей на примере аксиального спондилоартрита. Социальные аспекты здоровья населения. — 2023. — Т.69. — №3. — С.1. [Astanin PA, Rauzina SE, Zarubina TV. Automated system for recognizing clinically relevant UMLS terms in texts of the English-language articles exemplified by axial spondyloarthritis. Social Aspects of Population Health. 2023: 69(3): 1. (In Russ.)] doi: 10.21045/2071-5021-2023-69-3-14.

**37.** Валиев А.И., Лысенкова С.А. Применение методов машинного обучения для автоматизации процесса анализа содержания текста // Вестник кибернетики. — 2021. — №4(44). — С.12-15. [Valiev AI, Lysenkova SA. Application of machine learning methods for automation of the process of the text contents analysis. Proceedings in Cybernetics. 2021: 4(44): 12-15. (In Russ.)] doi: 10.34822/1999-7604-2021-4-12-15.

**38.** Зулкарнеев Р.Х., Юсупова Н.И., Сметанина О.Н. и др. Методы и модели извлечения знаний из медицинских документов // Информатика и автоматизация. — 2022. — Т.21. — №6. — С.1169-1210. [Zulkarneev R, Yusupova N, Smetanina O. Method and models of extraction of knowledge from medical documents. Informatics and Automation. 2022: 21(6): 1169-1210. (In Russ.)] doi: 10.15622/ia.21.6.4.