**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

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

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

**ЗАРУБИНА Т.В.,**
член-корр. РАН, д.м.н., профессор, ФГАОУ ВО РНИМУ им. Н.И. Пи-рогова Минздрава России, Москва, Россия, e-mail: zarubina@rsmu.ru

# АВТОМАТИЗИРОВАННАЯ СИСТЕМА ИЗВЛЕЧЕНИЯ АББРЕВИАТУР ТЕРМИНОВ УНИФИЦИРОВАННОЙ НАЦИОНАЛЬНОЙ МЕДИЦИНСКОЙ НОМЕНКЛАТУРЫ ИЗ ТЕКСТОВ НАУЧНЫХ СТАТЕЙ

**DOI:** 10.25881/18110193_2023_4_24

***Аннотация.*** *Унифицированная национальная медицинская номенклатура (УНМН) разрабатывается с 2022 г. с использованием международного метатезауруса Unified Medical Language System (UMLS) и других источников. УНМН является терминологической системой, организованной по онтологическому принципу и потенциально применимой для аннотирования медицинских текстов на русском языке. В настоящее время словари и справочники УНМН наполняются различными вариантами возможных формулировок медицинских терминов автоматизированным и экспертным способами. В медицине часто используются аббревиатуры, которые позволяют в сокращенной форме выразить смысл используемых понятий. Однако их распознавание в неструктурированном тексте является нетривиальной задачей. Разработка программного инструмента для автоматического извлечения аббревиатур из текста научных статей позволит обогатить УНМН и ускорить создание систем поддержки принятия клинических решений на её основе.*

*Цель исследования. Создание алгоритма автоматического извлечения аббревиатур терминов УНМН из текста научных статей на русском языке.*

*Материалы и методы. Для валидации и тестирования алгоритма использовались неструктурированные тексты аннотаций к научным статьям на русском языке, полученные из информационнопоисковой системы eLIBRARY. Полнотекстовые расшифровки извлеченных аббревиатур корректировались с применением билингвального перевода (на русский язык и обратно).*

*Результаты. Разработанный на основе семантических правил алгоритм позволил обеспечить извлечение аббревиатур и их полнотекстовых расшифровок с ~93% чувствительностью и ~99% специфичностью. Для большинства (~87%) терминов с использованием билингвального перевода удавалось скорректировать орфографические ошибки и выполнить приведение к начальной форме. Половина (~49%) аббревиатур со 100% точностью сопоставлялась с терминами УНМН. Обработка текстов аннотаций к научным статьям (168 тыс.) с использованием разработанного алгоритма позволила сформировать основу для создания Единого справочника медицинских аббревиатур, сопоставленных с терминами УНМН (свыше 6,6 тыс. уникальных записей).*

***Ключевые слова:*** *УНМН, UMLS, обработка естественного языка, регулярные выражения, Regex, Googletrans, API, семантический анализ текста.*

***Для цитирования:*** *Астанин П.А., Ронжин Л.В., Федоров А.А., Раузина С.Е., Зарубина Т.В. Автоматизированная система извлечения аббревиатур терминов унифицированной национальной медицинской номенклатуры из текстов научных статей. Врач и информационные технологии. 2023; 4: 24-35. doi: 10.25881/18110193_2023_4_24.*

**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

**FEDOROV A.A.,**
Pirogov Russian National Research Medical University, Moscow, Russia, e-mail: fdrv_rnrmu@mail.ru

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

**ZARUBINA T.V.,**
Corresponding Member of the RAS, DSc, Prof., Pirogov Russian National Research Medical University, Moscow, Russia, e-mail: zarubina@rsmu.ru

# AUTOMATED ABBREVIATIONS RECOGNITION SYSTEM FOR UNIFIED NATIONAL MEDICAL NOMENCLATURE FILLING WITH USING RUSSIAN LANGUAGE UNSTRUCTURED TEXT OF ARTICLES

**DOI:** 10.25881/18110193_2023_4_24

***Abstract.*** *The Unified national medical nomenclature (UNMN) has been under development since 2022 with using the Unified Medical Language System (UMLS) Metathesaurus and other sources. UNMN is a terminological system based on ontological approach and potentially applicable in Russian language medical text annotating. Currently, terms from different clinical branches are being added to UNMN utilizing both automatized and expert ways. Often in medicine abbreviations allow expressing the meaning of the concepts in a rapid way. However, their recognition in unstructured text is not trivial issue. The development of software for automated abbreviations recognition from research articles could enrich UNMN and accelerate clinical decision support systems development.*

*The aim of this study was to create the automated algorithm for UNMN terms abbreviations recognition from text of Russian language research articles.*

*Methods. Validation and testing dataset included unstructured abstracts of Russian language research articles aggregated from eLIBRARY. Fulltext wordings of extracted abbreviations have been corrected with bilingual (RussianEnglish and EnglishRussian) translation.*

*Results. Final version of the algorithm based on semantic rules demonstrated ~93% sensitivity and ~99% specificity in abbreviations and their fulltext wordings extraction. Large percentage (~87%) of terms has been successfully corrected and presented in the initial form after bilingual translation. Half (~49%) of abbreviations has been mapped with 100% accuracy to UNMN terms. Processing of 168 000 abstracts using the developed algorithm lead to creation of the Unified medical abbreviations thesaurus with UNMN terms (exceeding 6600 unique entries).*

***Keywords:*** *UNMN, UMLS, NLP, regular expressions, Regex, Googletrans, API, text semantic analysis*

***For citation:*** *Astanin P.A., Ronzhin L.V., Fedorov A.A., Rauzina S.E., Zarubina T.V. Automated abbreviations recognition system for unified national medical nomenclature filling with using russian language unstructured text of articles. Medical doctor and information technology. 2023; 4: 24-35. doi: 10.25881/18110193_2023_4_24.*

## ВВЕДЕНИЕ

Онтологический подход является одной из моделей представления знаний для построения систем поддержки принятия клинических решений (СППКР) [1–3]. Данный подход предусматривает формализацию знаний об исследуемой предметной области в виде семантических сетей — терминологических сводов, представленных в сетевой (графовой) форме [4–6]. Достоинством онтологического подхода является возможность частичной автоматизации процесса разработки СППКР при существовании терминологических систем, обеспечивающих смысловое покрытие значительной части существующих клинических областей [2, 7]. Одной из крупнейших русскоязычных терминологических систем является Унифицированная национальная медицинская номенклатура (УНМН), разрабатываемая с 2022 г. с использованием международного метатезауруса Unified Medical Language System (UMLS) [8, 9]. В настоящее время УНМН активно наполняется русскоязычными терминами из различных областей медицины и является основой для создания инструментов обработки естественного языка (NLP — natural language processing) или текста — основного способа представления данных в медицинских информационных системах [10, 11].

Важной частью любой терминологической системы являются аббревиатуры (сокращения и акронимы), активно применяемые в реальной медицинской практике для описания клинической картины пациента [12]. Однако современные алгоритмы семантического анализа обладают низкой чувствительностью к аббревиатурам и не позволяют использовать их в процессе обработки неструктурированного текста. По этой причине для обеспечения качественного распознания аббревиатур в медицинских текстах требуется использование специальных справочников. На момент проведения настоящего исследования не было найдено открытой информации о существовании единого справочника медицинских аббревиатур (ЕСМА) на русском языке.

Главным источником аббревиатур и их полнотекстовых расшифровок служат теоретические разделы научных статей: аннотация, введение и обсуждение. Данные разделы содержат наибольшее количество вводной информации с первым упоминанием большинства ключевых терминов и их аббревиатур [13]. Разработка программного инструмента для автоматического извлечения аббревиатур позволит повысить чувствительность алгоритмов извлечения именованных сущностей при обработке данных реальной клинической практики.

**Целью** настоящего исследования является разработка, валидация и тестирование алгоритма автоматического извлечения аббревиатур терминов УНМН из текста научных статей на русском языке.

## МАТЕРИАЛ И МЕТОДЫ

Исследование проведено сотрудниками Института цифровой трансформации медицины (ИЦТМ) ФГАОУ ВО «Российский национальный исследовательский медицинский университет имени Н.И. Пирогова» Минздрава России в рамках программы стратегического академического лидерства «Приоритет-2030». На всех этапах исследования использовались неструктурированные тексты аннотаций к научным статьям на русском языке, полученные из информационнопоисковой системы eLIBRARY. Для агрегации текстов случайным способом было извлечено 728 ссылок на статьи из журналов, отнесённых к рубрике 76.00.00 «Медицина и здравоохранение» в соответствии с Государственным рубрикатором научнотехнической информации (ГРНТИ) [14]. Самая ранняя публикация в полученной выборке была размещена в eLIBRARY в 1999 г. Основная часть (n = 666, 91,5%) извлечённых текстов датировалась 20142023 гг. и являлась актуальной на момент исследования. Статьи принадлежали к различным областям медицины и находились в открытом доступе для зарегистрированного в eLIBRARY пользователя.

Предобработка текстов состояла в выполнении стандартных технических операций [15–17]. Производилось удаление технических символов, абзацных отступов и переводов строк, кавычек, квадратных скобок и косых черт. Затем все многоточия в тексте заменялись на точки, а двойные пробелы — на одинарные.

Для извлечения аббревиатур созданы семантические правила, представленные в виде регулярных выражений. Написание регулярных выражений осуществлялось с использованием синтаксиса библиотеки Regex языка

программирования Python [18]. Составленные семантические правила были предназначены для решения двух ключевых задач: поиска паттернов, указывающих на наличие аббревиатуры в тексте, и проверки наличия полнотекстовой расшифровки найденной аббревиатуры в её окрестностях. Основным критерием успешного извлечения аббревиатуры было её непосредственное извлечение с нахождением правильной полнотекстовой расшифровки на русском или английском языках [19]. Примеры находимых аббревиатур и их расшифровок представлены с сохранением регистра, орфографии и согласования слов в таблице 1.

Из данных таблицы 1 следует, что значительная доля извлекаемых полнотекстовых расшифровок аббревиатур не находилась в единственном или множественном числе именительного падежа (начальной форме), требовала коррекции орфографических ошибок, а также числа и падежа с использованием морфологического разбора. Важно отметить, что подобная коррекция слов и фраз является одной из наиболее сложных задач NLP. В данном исследовании предпринимались попытки приведения полнотекстовых расшифровок аббревиатур к начальной форме с использованием лемматизатора Natural Language Toolkit (NLTK) и морфологического анализатора Pymorphy2 [20]. Однако при использовании данных инструментов удавалось обеспечить приведение слов лишь к нормальной форме (отличающейся от начальной). Для существительных нормальной форме соответствует единственное число именительного падежа, для прилагательных — единственное число именительного падежа в мужском роде, для глаголов, причастий, деепричастий — глагол в неопределённой форме несовершенного вида. Примеры полнотекстовых расшифровок аббревиатур в исходной, нормальной и начальной формах представлены на рисунке 1.

Данные, представленные на рисунке 1, демонстрируют невозможность использования нормальных форм отдельных слов по причине искажения правильности написания полнотекстовых расшифровок терминов: в большинстве случаев адекватное согласование родов прилагательного и существительного в терминах отсутствовало.

В связи с этим была предпринята попытка приведения расшифровок к начальной форме при билингвальном переводе с использованием

**Таблица 1 — Примеры пар аббревиатур и дословно найденных для них полнотекстовых расшифровок из текстов аннотаций к научным статьям**

| № | Аббревиатура | Дословная полнотекстовая расшифровка из текста |
|---|---|---|
| 1 | ДН | «дыхательная недостаточность» |
| 2 | ИМТ | «индекс массы тела» |
| 3 | НД | «нормативной документации» |
| 4 | АГ | «артериальная гипертензия» |
| 5 | ПК | «прекалликреина» |
| 6 | rAAV | «Рекомбинантный аденоассоциированный вирус» |
| 7 | ЩЖ | «щетовидная железа» |
| 8 | ПСМТ | «позвоночно-спинномозговой травмой» |
| 9 | ИИ | «ишемического инсульта» |
| 10 | ГАМК | «гамма-аминомасляной кислоты» |
| 11 | БНЧС | «Боль в нижней части спины» |
| 12 | аксСпА | «Аксиального спондилоартрита» |

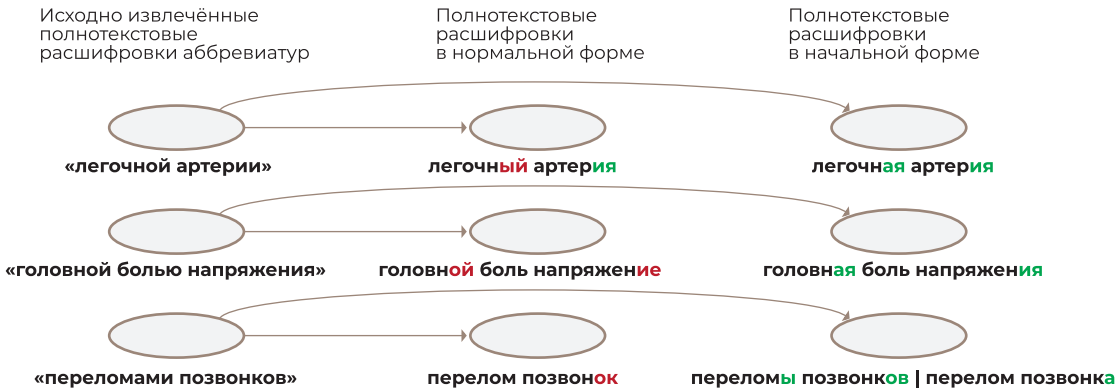


**Рисунок 1 — Примеры полнотекстовых расшифровок аббревиатур в исходной, нормальной и начальной формах.**

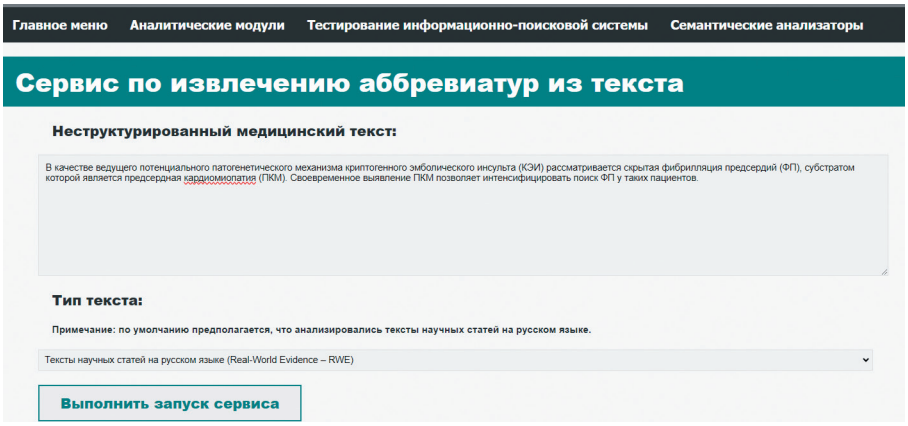


**Рисунок 2 — Фрагмент пользовательского решения для работы с сервисом по извлечению аббревиатур из текста.**

программного интерфейса приложения (API) Googletrans. Googletrans позволяет неограниченно осуществлять перевод текстов на разных языках и является единственным открытым ресурсом, свободно интегрируемым в приложения разработчиков. Благодаря открытому доступу Googletrans API позволил реализовать автоматический билингвальный перевод полнотекстовых формулировок на английский язык и обратно.

Полнотекстовые формулировки аббревиатур на английском языке сопоставлялись с терминами УНМН, разрабатываемой на базе ИЦТМ с 2022 г. При нахождении полного совпадения термина при регистронезависимом дословном поиске найденной аббревиатуре сопоставлялся номер концепта УНМН.

Разработанный алгоритм реализован в виде сервиса в составе аналитической системы ИЦТМ (рис. 2), выступающей в качестве платформы для создания баз знаний и СППКР на основе УНМН. Сервис предусматривает возможность внесения свободного неструктурированного текста с последующим его отнесением к данным реальной клинической практики или научнопрактическим материалам.

Предполагается экспертная проверка извлечённых аббревиатур и их полнотекстовых расшифровок с возможностью реализации трёх сценариев. Первый (основной) сценарий предполагает немедленное добавление аббревиатуры в состав концепта УНМН. Второй сценарий предусматривает возможность экспертной проверки и коррекции характеристик аббревиатуры

(номера сопоставленного концепта, непосредственно формулировки и её регистра) на уровне врачаэксперта с последующим добавлением в УНМН. В случае некорректной отработки алгоритма реализуется третий сценарий, согласно которому аббревиатура считается некорректной и ни при каких условиях не может быть добавлена в УНМН.

Для количественной оценки качества работы алгоритма определялись следующие статистические параметры: абсолютные (n) и относительные (%) доли найденных в текстах аббревиатур, аббревиатур с исходно правильной падежной формой, аббревиатур с верной падежной формой после коррекции с использованием билингвального перевода и аббревиатур с корректно подобранными для них концептами УНМН.

Валидированный алгоритм был использован для обработки корпуса из 168 тыс. текстов аннотаций к русскоязычным статьям. Независимо от результата сопоставления с концептами УНМН, найденные аббревиатуры и их полнотекстовые расшифровки сохранялись в базу данных.

## РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

При экспертной разметке исследуемого набора из 728 текстов аннотаций к научным статьям было выделено 305 аббревиатур. Некоторые аббревиатуры могли встречаться сразу в нескольких аннотациях. При автоматической обработке текста получено 285 аббревиатур, из них 266 были уникальными. Всего в одном случае вместо аббревиатуры было найдено обычное слово, в дальнейшем исключённое из

**Таблица 2 — Распределение аббревиатур по типам концептов УНМН**

| Семантический класс УНМН | Тематическая группа (tui) терминов УНМН, сопоставленных с аббревиатурами | Доля концептов – n (%) |
|---|---|---|
| Клинические расстройства | Всего, из них: | 81 (34,8%) |
| | T047 – Заболевания или синдромы | 56 (69,1%) |
| | T033 – Клинические находки | 12 (14,8%) |
| | T037 – Травмы и отравления | 7 (8,64%) |
| | Другие | 6 (7,41%) |
| Процедуры | Всего, из них: | 51 (21,9%) |
| | T059 – Лабораторные процедуры | 16 (31,3%) |
| | T060 – Диагностические процедуры | 12 (23,5%) |
| | T061 – Лечебно-профилактические мероприятия | 16 (31,3%) |
| | Другие | 7 (13,7%) |
| Химические вещества и лекарства | Всего, из них: | 34 (14,6%) |
| | T121 – Лекарственные вещества | 14 (41,2%) |
| | T116 – Аминокислоты, пептиды и белки | 6 (17,6%) |
| | T197 – Органические химические вещества | 4 (11,8%) |
| | Другие | 10 (29,4%) |
| Анатомия | Всего, из них: | 18 (7,73%) |
| | T023 – Части тела, органы или части органов | 10 (55,6%) |
| | T022 – Анатомические и функциональные системы | 3 (16,7%) |
| | T024 – Ткани | 3 (16,7%) |
| | Другие | 2 (11,1%) |
| Физиологические процессы, всего: | | 15 (6,44%) |
| Абстрактные понятия и категории, всего: | | 13 (5,58%) |
| Живые организмы, всего: | | 8 (3,43%) |
| Организации, всего: | | 6 (2,58%) |
| Гены, белки и аминокислоты, всего: | | 3 (1,29%) |
| Явления, процессы и их результаты, всего: | | 3 (1,29%) |
| Деятельность и поведение, всего: | | 1 (<1%) |
| Сумма по всем семантическим классам и группам УНМН: | | 225 (100%) |

**Примечание:** УНМН — унифицированная национальная медицинская номенклатура.

**Таблица 3 — Результаты количественной оценки качества работы алгоритма извлечения аббревиатур терминов УНМН**

| Краткое описание количественного критерия оценки качества | Значение – n (%) |
|---|---|
| Общая правильность распознавания аббревиатур | 261 (100) |
| Правильность падежных форм исходно извлечённых полнотекстовых формулировок | 28 (10,7) |
| Правильность падежных форм полнотекстовых формулировок, скорректированных при билингвальном переводе | 257 (98,1) |
| Соответствие исправленной формулировки дословно извлечённой | 229 (87,4) |
| Корректность подбора концептов УНМН для извлечённых аббревиатур | 261 (100) |

**Примечание:** УНМН – Унифицированная национальная медицинская номенклатура.

исследования. Мера оценки, эквивалентная специфичности, оказалась равной 99,6% (284 аббревиатуры из 285 действительно оказались аббревиатурами). Мера, эквивалентная чувствительности, была равна 93,1% (284 аббревиатуры найдено среди 305).

При использовании англоязычных полнотекстовых расшифровок аббревиатур с применением регистронезависимого поиска удалось сопоставить 130 (49,1%) из 265 найденных аббревиатур с 225 уникальными концептами УНМН из разных классов и групп (таблица 2).

Из таблицы 2 следует, что большинство сопоставленных аббревиатур относилось к тематическим группам из четырёх основных семантических классов УНМН: «Клинические расстройства» (~35%), «Процедуры» (~22%), «Химические вещества и лекарства» (~15%), «Анатомия» (~8%). Не сопоставлялись с УНМН аббревиатуры названий научнообразовательных учреждений, малоизвестных диагностических и терапевтических методов, узкоспециализированные понятия, а также экономические и географические термины, не относящиеся к медицине.

Сопоставленные с концептами УНМН аббревиатуры встретились в использованном наборе текстов аннотаций 261 раз. Каждый случай анализировался отдельно на предмет корректности извлечения аббревиатуры и качества обработки её полнотекстовой расшифровки. Сформировано пять признаков, для которых по каждому срабатыванию алгоритма экспертным способом определялись значения бинарных меток.

Первый признак определял общую правильность распознавания аббревиатуры. Наличие положительной метки этого признака означало, что аббревиатура и её расшифровка были агрегированы из текста полностью без лишних слов и символов. Второй признак характеризовал соответствие исходно извлечённой полнотекстовой расшифровки аббревиатуры начальной форме. Третий признак указывал на соответствие полнотекстовой формулировки начальной форме после билингвального перевода. Четвёртый признак характеризовал отсутствие искажения исходной расшифровки аббревиатуры после исправления (формулировка «фибрилляции предсердий» исправлялась на «фибрилляция предсердий», а не на «мерцательная аритмия»). Пятый признак определял корректность подбора концептов УНМН для извлечённых аббревиатур. Результат оценки вышеперечисленных признаков представлен в таблице 3.

Из таблицы 3 следует, что все извлечённые аббревиатуры были распознаны верно. Однако лишь малая часть (~10%) их полнотекстовых расшифровок исходно имела начальную форму. После попытки коррекции с использованием билингвального перевода к начальной форме приводилось до 98% формулировок. Тем не менее лишь 87% из них не теряли исходного варианта написания: все остальные могли заменяться на синонимичные понятия (например, «фибрилляция предсердий» исправлялась на термин «мерцательная аритмия»), что искажало полнотекстовую формулировку извлечённой аббревиатуры. Для всех извлеченных аббревиатур независимо от качества их исправления концепты УНМН подбирались верно.

При автоматической обработке корпуса текстов из 168 тыс. аннотаций к научным статьям на русском языке извлечено 16307 аббревиатур, из них 6617 было сопоставлено с концептами

УНМН. Полученные данные будут подвергнуты экспертной оценке и включению в ЕСМА.

## ОБСУЖДЕНИЕ

Большинство работ, посвящённых описанию алгоритмов автоматического анализа неструктурированной медицинской информации, связано с разработкой СППКР [21–24]. Значительная часть подобных СППКР базируется на использовании современных нейросетевых архитектур (BioBERT, Transformer, LSTM) и анализа больших данных [24, 25]. В некоторых ситуациях (например, при работе с редкими заболеваниями или при отсутствии источников больших данных) модели машинного обучения не позволяют достичь клинически значимого результата. В подобных случаях наибольшую эффективность демонстрируют алгоритмы интерпретации данных, построенные на использовании правил [26].

Наиболее часто в реальную практику внедряются гибридные СППКР, обеспечивающие достижение максимальной точности и предусматривающие одновременное использование результатов машинного обучения, экспертных правил и знаний [2]. Одним из способов реализации гибридных СППКР является онтологический подход, предполагающий предварительное формирование свода терминов для описания изучаемой области медицины [1, 2, 6].

В подавляющем большинстве исследований, связанных с разработкой СППКР на базе онтологического подхода, создаваемые словари понятий не имеют интеграции с крупными терминологическими сводами. В рамках системного решения данной проблемы ведётся создание УНМН — одной из крупнейших онтологических моделей на русском языке. Важным этапом разработки УНМН является поиск разнообразных формулировок клинических терминов (синонимов, сокращений и аббревиатур) с использованием автоматической обработки естественного языка (NLP — natural language processing) [23, 27–29]. Программные решения, применяемые в данной области, могут быть основаны на использовании как машинного обучения, так и семантических правил [12, 30–36].

В настоящем исследовании предпринята попытка создания автоматизированного алгоритма извлечения аббревиатур терминов УНМН из текста научных статей. Данный алгоритм был построен на основе правил, представленных в виде регулярных выражений и позволивших добиться извлечения аббревиатур и их полнотекстовых расшифровок с ~93% чувствительностью и ~99% специфичностью. С использованием билингвального перевода (на английский язык и обратно) удалось скорректировать орфографические ошибки в полнотекстовых расшифровках и привести их к начальной форме для ~87% аббревиатур. Половина (~49%) аббревиатур, среди которых подавляющая часть относилась к различным областям клинической медицины, автоматически сопоставлялась с одним или несколькими концептами УНМН. Оставшаяся часть аббревиатур требует отдельного анализа и сопоставления с концептами УНМН экспертным способом.

Обработка текстов аннотаций к научным статьям (168 тыс.) на русском языке позволила сформировать основу для создания Единого ЕСМА. В настоящее время ЕСМА включает свыше 6,6 тыс. уникальных записей, сопоставленных с УНМН. В дальнейшем планируется расширение ЕСМА по мере накопления и автоматической обработки неструктурированных данных из различных областей медицины.

К перспективам настоящего исследования следует отнести извлечение медицинских аббревиатур и их расшифровок при полнотекстовом анализе научных статей с предварительной доработкой инструмента приведения фраз и слов к начальной форме. Также актуальными задачами являются разработка алгоритма автоматического извлечения синонимов клинических терминов из неструктурированного текста и создание способов решения проблемы лексической неоднозначности.

## ЗАКЛЮЧЕНИЕ

Повышение качества автоматической обработки неструктурированного текста возможно при использовании справочников специализированных аббревиатур. Для разработки подобных ресурсов могут применяться тексты научных статей.

В рамках настоящего исследования был разработан и валидирован алгоритм извлечения

аббревиатур из текста аннотаций к научным статьям. Автоматическая обработка 167 тыс. текстов позволила сформировать основу для создания крупнейшего справочника аббревиатур терминов УНМН на русском языке.

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

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

## ЛИТЕРАТУРА/REFERENCES

1. Осмоловский И.С., Зарубина Т.В. Разработка и апробация прототипа экспертной системы для диагностики подагры // Социальные аспекты здоровья населения. — 2023. — Т.69. — №4. — С.1-24. [Osmolovsky IS, Zarubina TV. Developing and testing a prototype expert system for gout diagnosis. Social Aspects of Population Health. 2023; 69(4): 1-24. (In Russ.)] doi: 10.21045/2071-5021-2023-69-4-15.
2. Зарубина Т.В., Кобринский Б.А., Белоносов С.С. и др. Медицинская информатика: учебник. 2-е издание, переработанное и дополненное // Москва: ГЭОТАР-Медиа, 2022. — 464 с. [Zarubina TV, Kobrinskii BA, Lipkin YuG. Medical Informatics: Textbook. M.: GEOTAR-Media, 2022. 464 p. (In Russ.)] doi: 10.33029/9704-6273-7-TMI-2022-1-464.
3. Киселев К.В., Потехина А.В., Осяева М.К. и др. Разработка номенклатуры понятий для системы поддержки принятия врачебных решений в области диагностики стенокардии I-IV функциональных классов // Евразийский кардиологический журнал. — 2018. — №3. — С.14-25. [Kiselev KV, Potekhina AV, Osyaeva MK. Development of concepts nomenclature for clinical decision support system in diagnostics of angina pectoris. Eurasian heart journal. 2018; 3: 14-25. (In Russ.)]
4. Нугуманова А.Б., Байбурин Е.М., Мансурова М.Е., Барахнин В.Б. Автоматическое извлечение решеток понятий из медицинских текстов на основе комбинации анализа формальных понятий и технологий бутстраппинга // Вестник Новосибирского государственного университета. Серия: Информационные технологии. — 2018. — Т.16. — №4. — С.140-152. [Nugumanova AB, Bayburin EM, Mansurova ME, Barakhnin VB. Automatic extraction of formal lattices from medical texts based on the combination of the formal concept analysis and bootstrapping technologies. Vestnik NSU. Series: Information Technologies. 2018; 16(4): 140-152. (In Russ.)] doi: 10.25205/1818-7900-2018-16-4-140-152.
5. Сбоев А.Г., Селиванов А.А., Рыбка Р.Б. и др. 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 ehkstraktsii svyazannykh imenovannykh sushchnostei na primere biomeditsinskikh tekstovykh dannykh. Vestnik Voennogo innovatsionnogo tekhnopolisa «Ehra». 2022; 3(1): 57-67. (In Russ.)] doi: 10.56304/S2782375X22010193.
6. 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. 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.
7. Шахмаметова Г.Р., Худоба Е.В. Разработка метода структурирования данных и знаний клинических рекомендаций // Информационные технологии интеллектуальной поддержки принятия решений (ITIDS'2019): Труды VII Всероссийской научной конференции (с приглашением зарубежных ученых). — 2019. — Т.2. — С.237-240. [Shakhmametova GR, Khudoba EV. Razrabotka metoda strukturirovaniya dannykh i znanii klinicheskikh rekomendatsii. Informatsionnye tekhnologii intellektual'noi podderzhki prinyatiya reshenii (ITIDS'2019): Trudy VII Vserossiiskoi nauchnoi konferentsii (s priglasheniem zarubezhnykh uchenykh). 2019; 2: 237-240. (In Russ.)]
8. 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. — С.30-42. [Astanin PA, Ronzhin LV, Rauzina SE. Algorithm for UMLS metathesaurus concepts specificity estimation using example of analysis of the semantic model describing axial spondyloarthritis differential diagnostics. Medical doctor and information technologies. 2023; 3: 30-42. (In Russ.)] doi: 10.25881/18110193_2023_3_30.

9. 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.1-28. [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): 1-28. (In Russ.)] doi: 10.21045/2071-5021-2023-69-3-14.
10. Gusev A, Korsakov I, Novitsky R, et al. Feature extraction method from electronic health records in Russia. Proceedings of the 26th FRUCT Conference. 2020: 497–500. doi: 10.5281/zenodo.4007408.
11. 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.
12. Cossin S, Margaux J, Larrouture I, et al. Semi-Automatic Extraction of Abbreviations and their Senses from Electronic Health Records. 2021: 1-12.
13. Ежков А.А. Анализ исследований в области обработки неструктурированных текстов в медицине // Наука и Просвещение: сборник статей II Международной научно-практической конференции «Научное обозрение». — 2022. — С.23-26. [Ezhkov AA. Analiz issledovanii v oblasti obrabotki nestrukturirovannykh tekstov v meditsine. Nauka i Prosveshchenie: sbornik statei II Mezhdunarodnoi nauchno-prakticheskoi konferentsii «Nauchnoe obozrenie». 2022: 23-26. (In Russ.)]
14. Шрайберг Я.Л., Дмитриева Е.Ю., Смирнова О.В. и др. Разработка системы взаимосвязанных классификаций: сопоставление Государственного рубрикатора научно-технической информации и Универсальной десятичной классификации // Научные и технические библиотеки. — 2023. — №11. — С.36-65. [Shraiberg YaL, Dmitrieva EYu, Smirnova OV. Developing the system of interconnected classifications: Comparing the State Rubricator of Sci-tech Information and Universal Decimal Classification. Scientific and Technical Libraries. 2023; 11: 36-65. (In Russ.)] doi: 10.33186/1027-3689-2023-11-36-65.
15. Пикалёв Я.С. Разработка системы нормализации текстовых корпусов // Проблемы искусственного интеллекта. — 2022. — №25(2). — С.64-78. [Pikalev YaS. Razrabotka sistemy normalizatsii tekstovykh korpusov. Problemy iskusstvennogo intellekta. 2022; 25(2): 64-78. (In Russ.)]
16. Астапов Р.Л., Мухмадеева Р.М. Автоматизированная предобработка текста для определения эмоциональной окраски текста // Актуальные научные исследования в современном мире. — 2021. — №5-2(73). — С.19-23.
17. Логунова Т.В., Щербакова Л.В., Васюков В.М., Шимкун В.В. Анализ алгоритмов классификации текстов // Universum: технические науки. — 2023. — №2-2(107). — С.4-20. [Astapov RL, Mukhmadeeva RM. Avtomatizirovannaya predobrabotka teksta dlya opredeleniya ehmotsional'noi okraski teksta. iScience. 2021; 5-2(73): 19-23. (In Russ.)] doi: 10.32743/UniTech.2023.107.2.15064.
18. Груздев Д.Ю., Макаренко А.С., Коджебаш Д.О. Принципы создания аннотации корпуса текстов // Вестник МИТУ — МАРХИ. — 2023. — №1. — С.88-97. [Gruzdev DYu, Makarenko AS, Kodzhebash DO. Corpus annotation development principles. Vestnik MITU — MARHI. 2023; 1: 88-97. (In Russ.)] doi: 10.52470/2619046X_2023_1_88.
19. Пашук А.В., Гуринович А.Б., Волорова Н.А., Кузнецов А.П. Анализ методов разрешения лексической многозначности в области биомедицины // Доклады БГУИР. — 2019 — №5(123). — С.60-65. [Pashuk AV, Gurinovich AB, Volorova NA, Kuznetsov AP. Analysis of the methods of word sense disambiguation in the biomedical domain. Doklady BGUIR. 2019; 5(123): 60-65. (In Russ.)] doi: 10.35596/1729-7648-2019-123-5-60-65.
20. Валиев А.И., Лысенкова С.А. Применение методов машинного обучения для автоматизации процесса анализа содержания текста // Вестник кибернетики. — 2021. — №44(4). — 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; 44(4): 12-15. (In Russ.)] doi: 10.34822/1999-7604-2021-4-12-15.

**21.** Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020; 36(4): 1234-1240. doi: 10.1093/bioinformatics/btz682.

**22.** Zhang Y, Tiryaki F, Jiang M, et al. Parsing clinical text using the state-of-the-art deep learning based parsers: a systematic comparison. BMC Med Inform Decis Mak. 2019; 19(3): 77. doi: 10.1186/s12911-019-0783-2.

**23.** Ленивцева Ю.Д., Копаница Г.Д. Автоматическое определение типа аллергии из неструктурированных медицинских текстов на русском языке // Научно-технический вестник информационных технологий, механики и оптики. — 2021. — Т.21. — №3. — С.433-436. [Lenivtceva IuD, Kopanitsa GD. Automatic allergy classification based on Russian unstructured medical texts. Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2021; 21(3): 433-436. (In Russ.)] doi: 10.17586/2226-1494-2021-21-3-433-436.

**24.** Хоружая А.Н., Козлов Д.В., Арзамасов К.М., Кремнева Е.И. Анализ текстов описаний КТ-исследований головного мозга с признаками внутричерепных кровоизлияний с помощью алгоритма дерева решений // Соврем. технол. мед. — 2022. — Т. 14. — №6. — С. 34-41. [Khoruzhaya AN, Kozlov DV, Arzamasov KM, Kremneva EI. Text Analysis of Radiology Reports with Signs of Intracranial Hemorrhage on Brain CT Scans Using the Decision Tree Algorithm. Sovremennye tehnologii v medicine. 2022; 14(6): 34-41. (In Russ.)] doi: 10.17691/stm2022.14.6.04.

**25.** Кротова О.С., Москалев И.В., Хворова Л.А., Назаркина О.М. Реализация эффективных моделей классификации медицинских данных методами интеллектуального анализа текстовой информации // Известия Алтайского государственного университета. — 2020. — №111(1). — С.99-104. [Krotova OS, Moskalev IV, Khvorova LA, Nazarkina OM. Implementation of effective models for classifying medical data using text mining. Izvestiya of Altai State University. 2020; 111(1): 99-104. (In Russ.)] doi: 10.14258/izvasu(2020)1-16.

**26.** Ткаченко С.А., Коломыцева Е.П. Разработка подходов по выявлению именованных сущностей в биомедицинских текстах с использованием методов нечеткой логики // Вектор развития современной науки: Сборник статей VII Международной научно-практической конференции. — 2020. — С.34-41. [Tkachenko SA, Kolomytseva EP. Razrabotka podkhodov po vyyavleniyu imenovannykh sushchnostei v biomeditsinskikh tekstakh s ispol'zovaniem metodov nechetkoi logiki. Vektor razvitiya sovremennoi nauki: Sbornik statei VII Mezhdunarodnoi nauchno-prakticheskoi konferentsii. 2020: 34-41. (In Russ.)]

**27.** 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 RKH, Yusupova NI, Smetanina ON. 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.

**28.** Клышинский Э.С., Грибова В.В., Шахгельдян К.И. и др. Алгоритм автоматического выделения жалоб пациентов из историй болезни // Новые информационные технологии в автоматизированных системах. — 2019. — №22. — С.204-209. [Klyshinskii EhS, Gribova VV, Shakhgel'dyan KI. Algoritm avtomaticheskogo vydeleniya zhalob patsientov iz istorii bolezni. Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh. 2019; 22: 204-209. (In Russ.)]

**29.** Легашев Л.В., Шухман А.Е., Болодурина И.П. и др. Обработка русскоязычных неструктурированных медицинских текстов и вероятностное прогнозирование групп заболеваний // Врач и информационные технологии. — 2022. — №4. — С.52-63. [Legashev LV, Shukhman AE, Bolodurina IP. Russian unstructured clinical texts processing and probabilistic classification of disease groups. Medical doctor and information technologies. 2022; 4: 52-63. (In Russ.)] doi: 10.25881/18110193_2022_4_52.

**30.** Сердюк Ю.П., Власова Н.А., Момот С.Р. Система извлечения упоминаний симптомов из текстов на естественном языке с помощью нейронных сетей // Программные системы: теория и приложения. — 2023. — Т.14. — №56(1). — С.95-123. [Serdyuk YuP, Vlasova NA, Momot SR. A system for extracting symptom mentions from texts by means of neural networks. Program Systems: Theory and Applications. 2023; 14(56(1)): 95-123. (In Russ.)] doi: 10.25209/2079-3316-2023-14-1-95-123.

**31.** Москалев И.В., Кротова О.С., Хворова Л.А. Автоматизация процесса извлечения структурированных данных из неструктурированных медицинских выписок с применением технологий интеллектуального анализа текстов // Высокопроизводительные вычислительные системы и технологии. — 2020. — Т.4. — №1. — С.163-167. [Moskalev IV, Krotova OS, Khvorova LA. Avtomatizatsiya protsessa izvlecheniya strukturirovannykh dannykh iz nestrukturirovannykh meditsinskikh vypisok s primeneniem tekhnologii intellektual'nogo analiza tekstov. High-performance computing systems and technolohies. 2020; 4(1): 163-167. (In Russ.)]

**32.** Du X, Zhu R, Li Y, Anjum A. Language model-based automatic prefix abbreviation expansion method for biomedical big data analysis. Future Gener Comput Syst. 2019; 98: 238-251. doi: 10.1016/j.future.2019.01.016.

**33.** Chang JT, Schütze H, Altman RB. Creating an online dictionary of abbreviations from MEDLINE. J Am Med Inform Assoc. 2002; 9(6): 612-620. doi: 10.1197/jamia.m1139.

**34.** Qiao J, Jinling L, Xinghua L. Deep contextualized biomedical abbreviation expansion. Proceedings of the 18th BioNLP Workshop and Shared Task in Florence, Italy. 2019: 88-96. doi: 10.18653/v1/W19-5010.

**35.** Juyong K, Gong L, Khim J, et al. Improved clinical abbreviation expansion via non-sense-based approaches. Proceedings of Machine Learning Research. 2020; 136: 161-178.

**36.** Skreta M, Arbabi A, Wang J, et al. Automatically disambiguating medical acronyms with ontology-aware deep learning. Nat Commun. 2021; 12(1): 5319. doi: 10.1038/s41467-021-25578-4.