СОЗДАНИЕ УНИФИЦИРОВАННОЙ НАЦИОНАЛЬНОЙ МЕДИЦИНСКОЙ НОМЕНКЛАТУРЫ

Раузина С.Е., доцент каф. медицинкой кибернетики и информатики им. С.А. Гаспаряна, зав. лаб. семантического анализа медицинской информации ИЦТМ РНИМУ им. Н.И. Пирогова

Международный конгресс ИТМ,
Москва, 12-13 октября 2023

СТРАТЕГИЧЕСКИЙ ПРОЕКТ РНИМУ ИМЕНИ Н.И. ПИРОГОВА «ИНСТИТУТ ЦИФРОВОЙ ТРАНСФОРМАЦИИ МЕДИЦИНЫ»

Программа академического лидерства «Приоритет 2030»

Стратегический проект РНИМУ имени Н.И. Пирогова «Институт цифровой трансформации медицины»

Задачи на разработку:

- *Информационно-поисковая система для предоставления врачу семантически связанной информации*
- *Сервисы поддержки принятия клинических решений*
- *Среда для обучения работе в среде МИС МО студентов «врачебных» факультетов и последипломного образования*
- *Семантический тренажер для подготовки врачей-аналитиков*

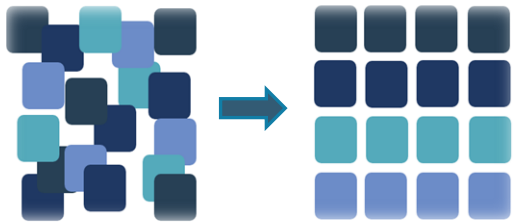
Национальная медицинская терминологическая база для унификации работ по созданию поисковых и интеллектуальных систем, а также обработки неструктурированных медицинских текстов

ПРОБЛЕМЫ НЕСТРУКТУРИРОВАННОЙ КЛИНИЧЕСКОЙ ИНФОРМАЦИИ

Отсутствие общепринятой формализации при описании клинической картины пациента, единой медицинской терминологии приводит к:

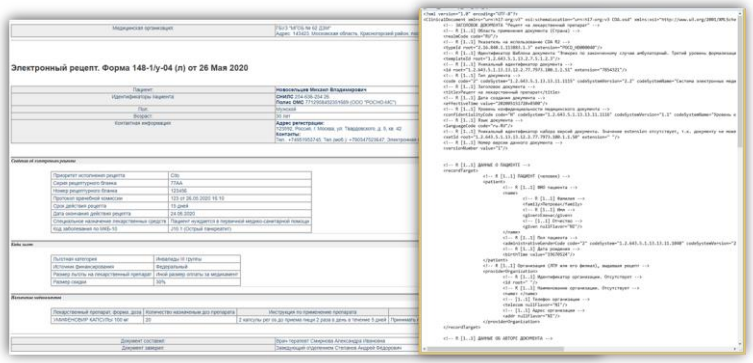
- Искажению интерпретации информации
- Проблемам анализа на основе данных
- Проблемам создания dataset для проведения научных исследований
- Возможности построения одноплатформенных систем ИИ и информационо-поисковых алгоритмов

Автоматизация обработки медицинской информации требует унификации решений при создании алгоритмов ее поиска и консультативной поддержки врача

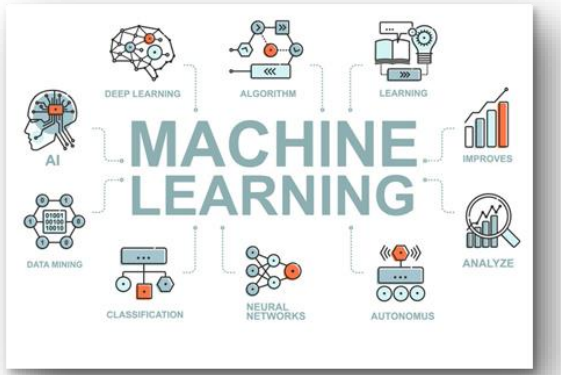


ПУТИ УНИФИКАЦИИ РЕШЕНИЙ

1. Разработка структурированных электронных медицинских документов **(СЭМД)**



2. Развитие **алгоритмов обработки** неструктурированных медицинских текстов



Разработка унифицированной медицинской терминологии

Инженеры по знаниям

ФУНДАМЕНТ ДЛЯ ПОСТРОЕНИЯ УНИФИЦИРОВАННОЙ НАЦИОНАЛЬНОЙ МЕДИЦИНСКОЙ НОМЕНКЛАТУРЫ (УНМН)

Систематизированный словарь терминов

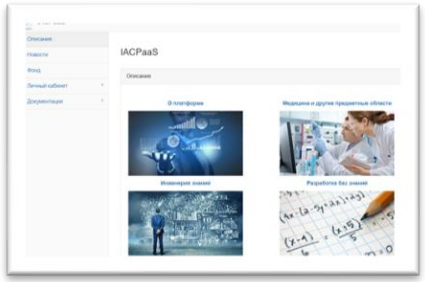
Связанное представление терминов/концептов

Широкий спектр областей медицины

Отечественный опыт

UMKB - Объединенная база медицинских знаний

Терминология, созданная на платформе IACPaaS



Система унифицированного медицинского языка, Unified Medical Language System

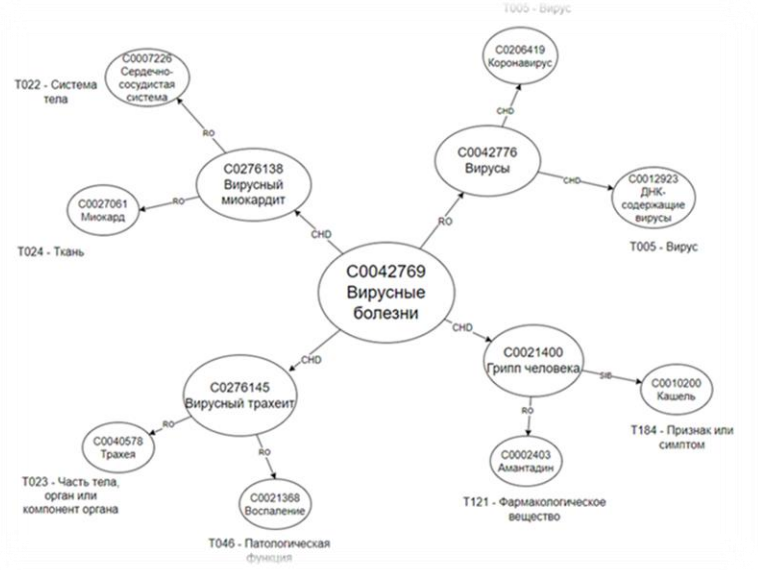
UMLS

4,6 млн Понятий

11,2 млн Синонимов

98 млн Связей

76 актуальн. Справочников



MeSH

SNOMED CT

LOINC® From Regenstrief

MEDICOMP SYSTEMS

ICD-10-CM

RxNorm

ОЦЕНКА ПОТЕНЦИАЛЬНОЙ ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ UMLS

Верификация: анализ клинических рекомендаций в части описания клинической картины, данных анамнеза и факторов риска развития заболеваний по **22** группам нозологий.

Всего выделено **579** медицинских терминов

- Аналогичное или полностью схожее наименование – **74**% терминов
- Частичное совпадение – **16**% терминов
- Не найдено - **10**% терминов

№	Наименование КР	ID	Год принят.
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

Причиной неполного совпадения или ненайденных терминов чаще всего является их сложное с медицинской точки зрения звучание, вероятно, иное в англоязычной среде, требующее привлечения профильных специалистов

Например:
аускультация легких при эмфиземе

RUS: коробочный звук

ENG: Hyperresonance

Геморрагия при панкреатите

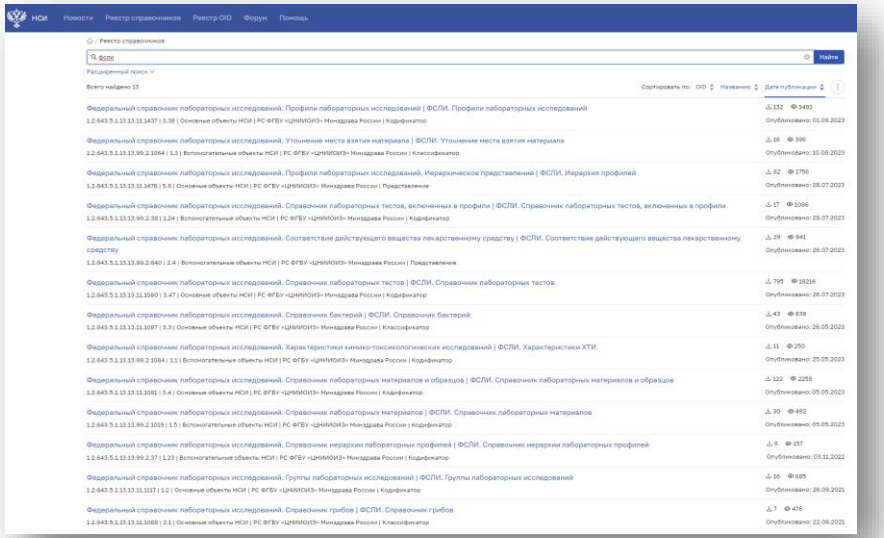
Симптом Тужилина

СОЗДАНИЕ УНИФИЦИРОВАННОЙ НАЦИОНАЛЬНОЙ МЕДИЦИНСКОЙ НОМЕНКЛАТУРЫ (УНМН)

1 версия

- **Переводы**
  - ✓ Исходно в UMLS представлено **3**% русскоязычных термина
  - ✓ Нейросетевые переводы в медицине дают большой % ошибок
  - ✓ Переведено **190** тыс. терминов из семантических групп *Симптомы, Клинические находки, Диагностика*

- **Сопоставление с утвержденными национальными справочниками (ФР НСИ МЗ РФ)** SNOMED_CT, LOINC, RadLex
  - ✓ МКБ-10 (100%)
  - ✓ ФСЛИ (32%)
  - ✓ ФСИДИ (75%)
  - ✓ Анатомические локализации (90%)
  - ✓ Выявленные патологии (100%)



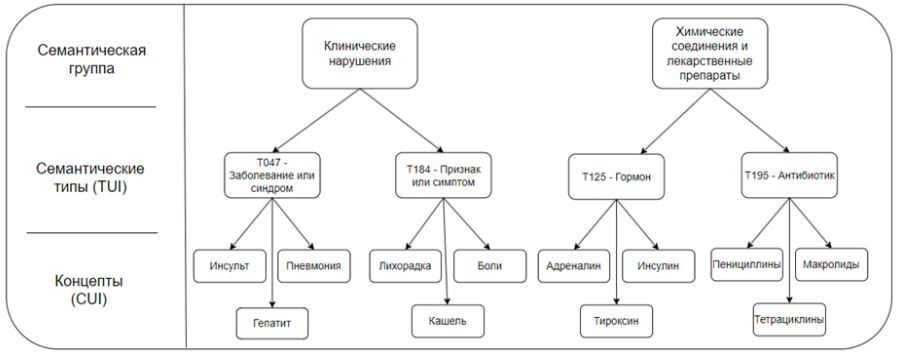
Всего экспертно верифицированных переводов - > **215** тыс.
исходно представлено на русском языке в UMLS - **304** тыс.

По оценке с помощью разработанной метрики: в UMLS присутствует ≈ **1,5** млн уникальных клинически значимых терминов

МЕТОДЫ АВТОМАТИЗИРОВАННОЙ РАБОТЫ С ТЕРМИНОЛОГИЧЕСКОЙ БАЗОЙ: ИСХОДНЫЕ ЗАКОНОМЕРНОСТИ

Использование закономерностей в семантической организации терминов и связей, исходно представленных в UMLS

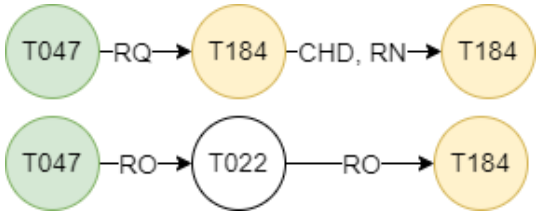
- ✓ Выбор клинически значимых справочников
- ✓ Выявление релевантных групп терминов и типов связей
- ✓ Использование иерархических зависимостей



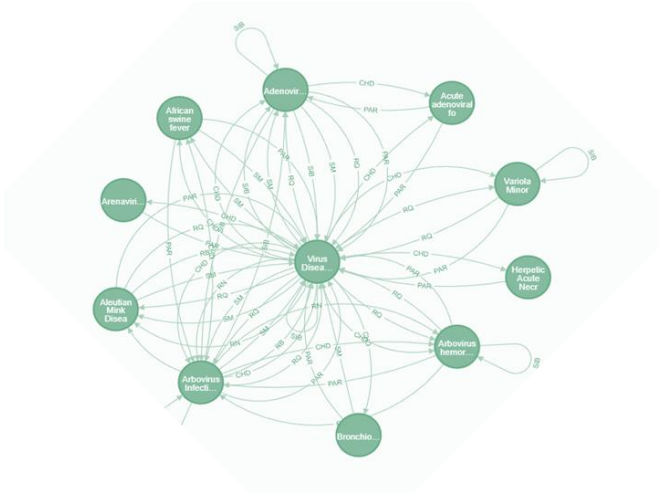
127 семантических типов терминов, образующих **15** групп: (7 социально-экономических и 8 биомедицинских)

Связи объединены в **11** групп и свыше **980** подтипов

- ✓ Паттерны для выделения симптомов
- ✓ Паттерны для инструментальной и лабораторной диагностики
- × Паттерны для выбора методов лечения



менее 70% релевантной информации



МЕТОДЫ АНАЛИЗА СЕМАНТИЧЕСКИХ СЕТЕЙ

✓ **Связность концептов:**

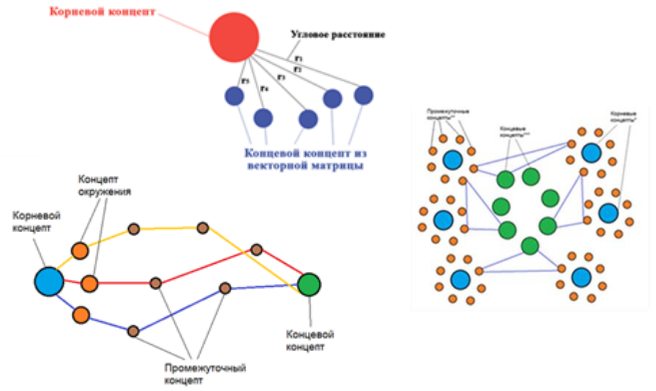
- прямые связи
- кратчайшие пути
- графовое расстояние
- геометрические фигуры

Взвешенный коэффициент кластеризации*:

обеспечивает учет количественной меры связности концептов и степень ее неоднородности

✓ **Метрики оценки клинической значимости и специфичности терминов**

инструменты ранжирования узлов графа



$$ВКК_n = C_n \cdot R_n^{-1,29}$$

где Cn – число геометрических контуров или незамкнутых путей, в образовании которых участвует узел n; Rn – число прямых связей между узлом n и любыми другими вершинами графа

Клиническая значимость термина: на основе анализа атрибутов метатезауруса (тематическая группа, актуальность, источник-справочник и др.)

Неспецифичные (обобщающие) термины: «симптом», «признак», «заболевание», «пациент»

**Астанин П. А., Раузина С. Е., Зарубина Т. В. Построение этиопатогенетического образа концептов метатезауруса UMLS с использованием графовых метрик // Программные системы: теория и приложения. 2023. Т. 14. № 3. С. 59–94. (Рус., англ.). https://psta.psiras.ru/2023/3_59-94*

ДОБАВЛЕНИЕ ПРЯМЫХ СВЯЗЕЙ В УНМН

✓ **27 млн аннотаций к медицинским англоязычным статьям PubMed**

Семантический анализатор **SemRep** (набор лингвистических правил для анализа англоязычных текстов) ≈ 4 млн уникальных связей для УНМН

Лаборатория вычислительной физиологии Массачусетского технологического института

PhysioNet

✓ **Данные реальной клинической практики: dataset MIMIC-IV (330 тыс. ЭМК)**

Агрегатор концептов UMLS **MetaMap** (автоматическое извлечение концептов из англоязычных текстов) ≈ 16 млн уникальных связей для УНМН между симптомами и заболеваниями

ПОСТРОЕНИЕ БАЗЫ ЗНАНИЙ НА ОСНОВЕ УНМН

- Закономерности, исходно представленные в UMLS
- Методы анализа семантических сетей
- Метрика автоматизированной оценки клинической значимости концептов
- Метрика оценки относительной специфичности терминов
- Получение дополнительных прямых связей

с **70**% до **90**% релевантной информации:

искомый диагноз встречается среди трёх первых заболеваний в ранжированном перечне

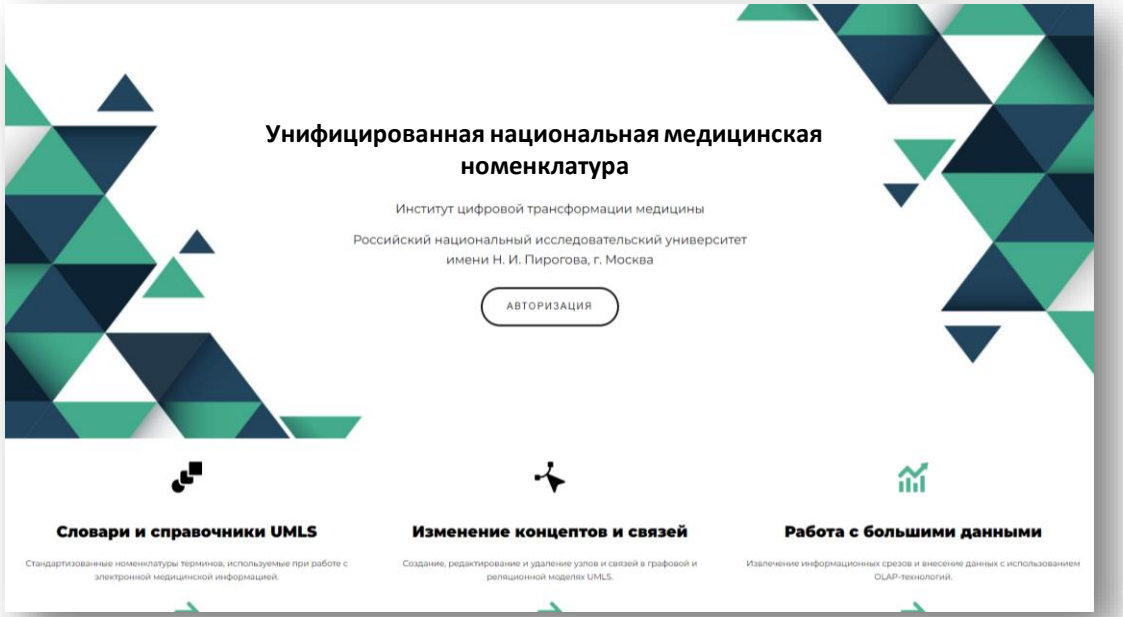
Добавление полученных метрик к УНМН в качестве атрибутов позволило получить онтологическую модель для создания компонентов информационно-поисковых систем и СППКР

ПРОГРАММНЫЕ РЕШЕНИЯ

Аналитическая система для проведения исследовательских работ и обучения

3 основных рабочих блока:

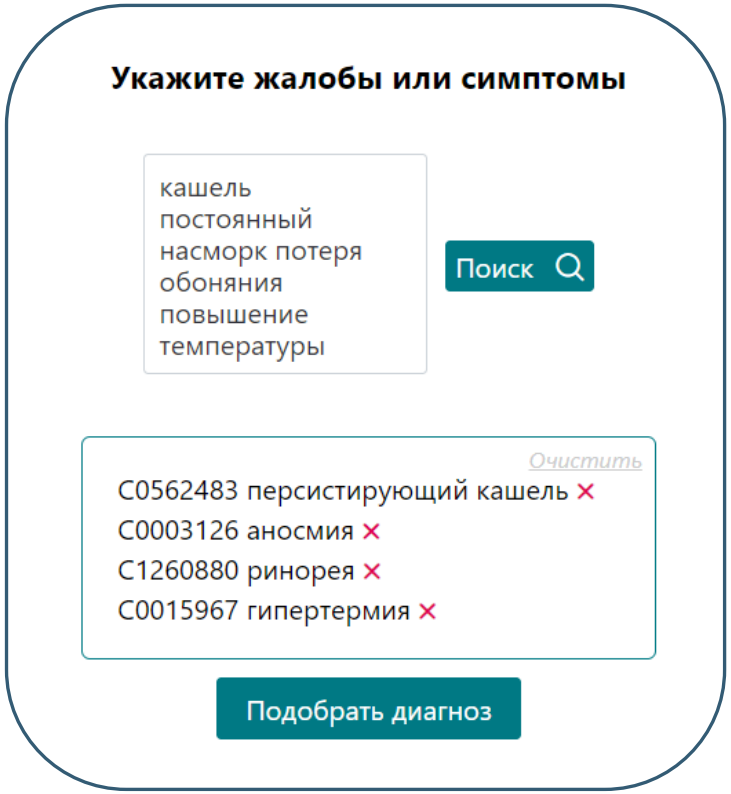
- ✓ Нормативно-справочная информация – международные и национальные словари;
- ✓ Редактор базы знаний – инструмент для наполнения УНМН экспертными знаниями;
- ✓ Интеллектуальное ядро – комплекс аналитических решений, предназначенных для извлечения OLAP-срезов из УНМН.



ПРОГРАММНЫЕ РЕШЕНИЯ

Прототип информационно-поисковой системы (ИПС) для автоматического определения наиболее вероятного класса заболеваний при составлении дифференциально-диагностического ряда произвольного клинического образа и выбора методов диагностики для их уточнения

- ✓ Выделение медицинских понятий из свободно вводимых текстов
- ✓ Поиск отрицаний



- ✓ Вывод ранжированного списка заболеваний на основе введенных симптомов

ОБРАБОТКА НЕСТРУКТУРИРОВАННОГО ТЕКСТА

Шаг 1: Очистка(регулярные выражения), коррекция ошибок (алгоритм Левенштайна)

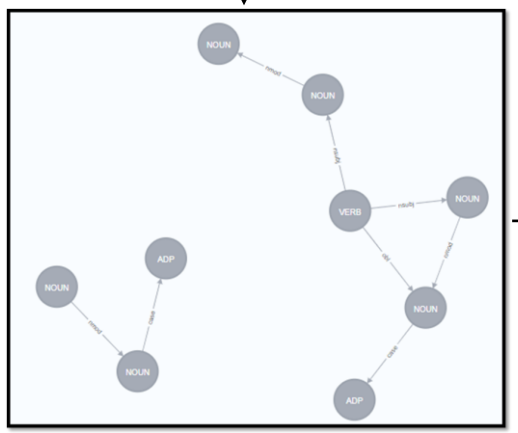
Шаг 2: Лемматизация – приведение к нормальной форме слова (pymorphy2, stanza*)

Шаг 3: Построение дерева синтаксических связей предложения (stanza depparser)*

Боль, одыжка, повышенная температура

боль одышка повышенная температура

боль одышка повышенный температура



- Боль
- Одышка
- Гипертермия

Поиск построчных совпадений с нормализованной леммой из метатезауруса

Проверка графовых представлений концептов на предмет вхождения в синтаксическое дерево предложения (neo4j)

*нейронная сеть stanza BERT

НАПРАВЛЕНИЯ РАЗВИТИЯ

Разработка новых национальных словарей, например, группы справочников, имеющих отношение к симптомам и клиническим находкам

Использование УНМН для формализации и унификации документов ЭМК, а также при создании СЭМД (единое кодирование информации в ЭМК)

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, а также **системы для разработки СППКР**, в которых будут реализованы автоматизированные рабочие места эксперта и инженера по знаниям

ПУБЛИКАЦИИ

1. 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 – 14.
2. 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: 10.3233/SHTI230458
3. Свидетельство о регистрации программы для ЭВМ № 2022684715/ 16.12.2022 Бюл. №12. Астанин П.А., Ронжин Л.В., Раузина С.Е. Модуль поиска релевантного окружения концептов семантической сети UMLS.
4. Свидетельство о регистрации программы для ЭВМ № 2022684714/ 16.12.2022 Бюл. №12. Астанин П.А., Ронжин Л.В., Королева Ю.И. Модуль поиска кратчайшего пути между концептами семантической сети UMLS.
5. Астанин П.А. Применение автоматизированного анализа семантической сети UMLS для решения задачи поиска релевантных знаний о ревматических заболеваниях. Сборник материалов Международной научно-практической конференции «Математическое моделирование систем и процессов», Псков. – 2022. С.6-12
6. Астанин П. А., Раузина С. Е., Зарубина Т. В. Построение этиопатогенетического образа концептов метатезауруса UMLS с использованием графовых метрик// Программные системы: теория и приложения. 2023. Т. 14. № 3. С. 59–94. (Рус., англ.). https://psta.psiras.ru/2023/3_59-94

БЛАГОДАРЮ ЗА ВНИМАНИЕ

rauzina_se@rsmu.ru