Supplementary materials: "RuCCoN: Clinical Concept Normalization in Russian"

Anonymous ACL submission

A Appendix: Инструкция для аннотаторов. Задача нормализации медицинских концепций

Annotation guidelines are in Russian; will be translated upon acceptance of the paper.

A.1. Обзор задачи

Задача состоит в том, чтобы просмотреть записи электронных медицинских карт, в которых выделены упоминания о медицинских концепциях, и сопоставить каждое из выделенных упоминаний концепций с уникальным идентификатором концепта (CUI) из предоставленной онтологии клинических терминов. Цель нормализации сущностей - присвоить один идентификатор различным синонимам одной и той же медицинской концепции. Например, строки «ишемический инфаркт сердца» и «инфаркт миокарда» написаны разными словами, но относятся к одной и той же концепции с идентификатором C0027051.

A.2. Данные и ресурсы

Данные. Документы, которые вы будете аннотировать, являются обезличенными записями в электронных медицинских картах. В записях заранее были выделены фрагменты текста, соответствующие медицинским концептам. Кроме, для облегчения и ускорения процесса разметки, каждая выделенная медицинская концепция была сопоставлена с CUI в автоматическом режиме.

Словари Каждая фраза, обозначенная в тексте как упоминание медицинской концепции, должна быть связана с CUI из Unified Medical Language System (UMLS), которая представляет собой единый международный словарь медицинских понятий. UMLS объединяет различные медицинские и клинические словари. Словари, относящиеся к этой задаче аннотации, включают:

- MedDRA (MDRRUS) — русская версия медицинского словаря для регуляторной деятельности, включающая большое количество медицинских концептов.
- MeSH Russian (MSHRUS) — русская версия словаря медицинских предметных заголовков, всеобъемлющего контролируемого словаря, созданного для индексирования журнальных статей и книг по наукам о жизни.

Дополнительные ресурсы. Вы можете использовать следующие дополнительные ресурсы, чтобы правильно определить наиболее подходящий CUI:

- «UMLS Metathesaurus Browser» — это англоязычный веб-сервис для поиска и определения оптимального CUI, доступный по ссылке: `https://uts.nlm.nih.gov/uts/umls/home`. Несмотря на то, что это англоязычный сервис, в нем доступен поиск русскоязычных концептов (рисунок 1). Для доступа к данному ресурсу вам нужно будет пройти регистрацию по ссылке: `https://uts.nlm.nih.gov/uts/login`
- Google — Вы можете использовать Google, если какая либо концепция вам незнакома, или если вам встретилась неизвестная ранее аббревиатура или сокращение. Рисунок 1. Страница с выводом результата поиска медицинского концепта на русском языке в сервисе «UMLS Metathesaurus Browser».

A.3. Описание задачи

Для каждого выделенного упоминания концепции в тексте (т.е. для каждого фрагмен-

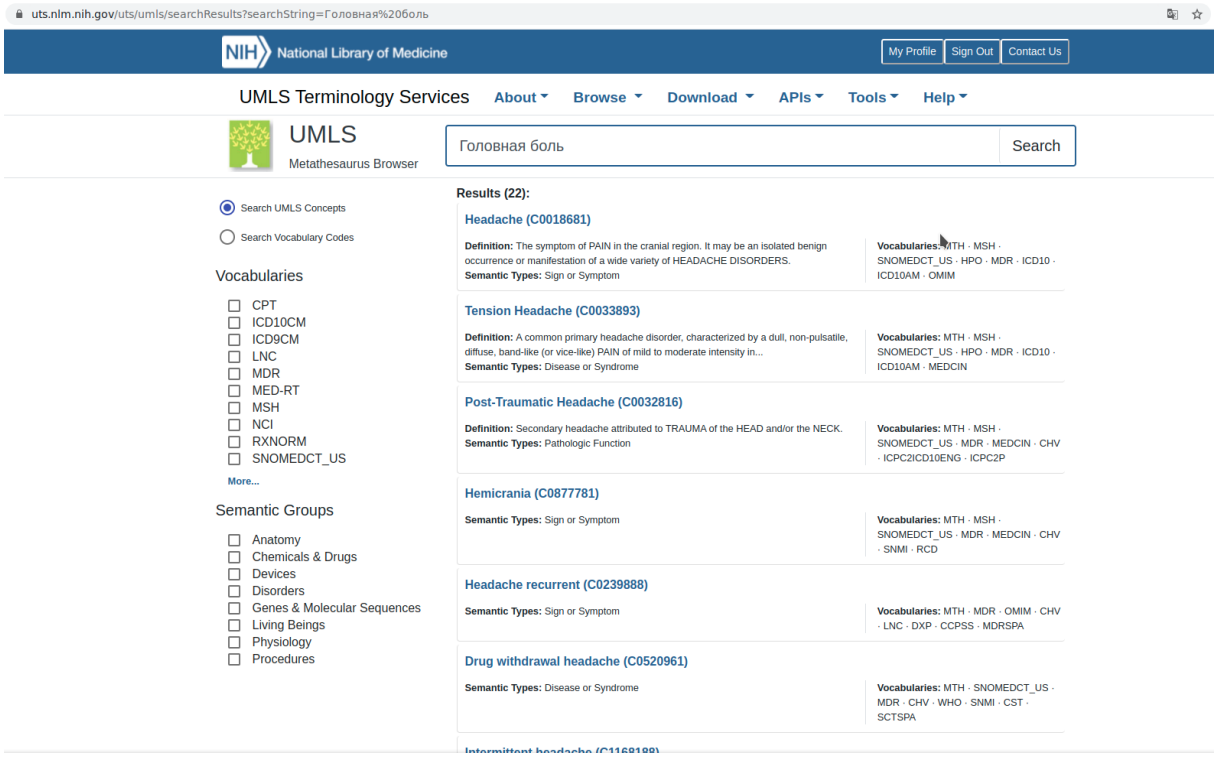

Рис. 1: Поиск русскоязычных концептов в «UMLS Metathesaurus Browser».

та текста, соответствующего клинической концепции) вам необходимо будет идентифицировать CUI.

Пример: «Пациент поступил с пониженным артериальным давлением». Выделенный заранее фрагмент текста «пониженным артериальным давлением» следует сопоставить с CUI C0020649 (Гипотония).

Для каждого выделенного упоминания должна быть только одна основная аннотация. Однако во многих случаях выделенные концепции могут соответствовать нескольким CUI, в данных случаях вам следует руководствоваться следующими правилами:

1. Следует выбирать такие концепты, которые максимально соответствуют семантическому типу выделенного фрагменту. Например, если идентифицированное упоминание является лабораторным тестом, предпочтительнее использовать CUI с семантическим типом «Лабораторная процедура» (семантические типы будут указаны при поиске концепции в сервисе «UMLS Metathesaurus Browser»). В приведенном ниже примере CUI, который следует выбрать, отмечен звездочкой.

   Пример: 2019-10-07 05:30 Анализ крови WBC - 6.5 RBC - 3,23 * Hgb - 9,5 * Hct - 27,6 * MCV - 86 MCH - 29,4 MCHC - 34,4 RDW - 14,0 Plt - 356 #

   - WBC
     - C0023516 (лейкоциты) - семантический тип: клетка
     - C0023508 (процедура подсчета лейкоцитов) * - семантический тип: лабораторная процедура
   - RBC
     - C0014792 (эритроциты)
     - C0014772 (измерение количества эритроцитов) *
   - Hgb
     - C0019046 (гемоглобин)
     - C0474563 (Анализ гемоглобина плазмы)
     - C0474536 (Процедура определения гемоглобина)
     - C0369320 (свободный гемоглобин (процедура))
     - C0587341 (оценка уровня гемоглобина) *
     - C2711614 (измерение общей концентрации гемоглобина)
   - Hct

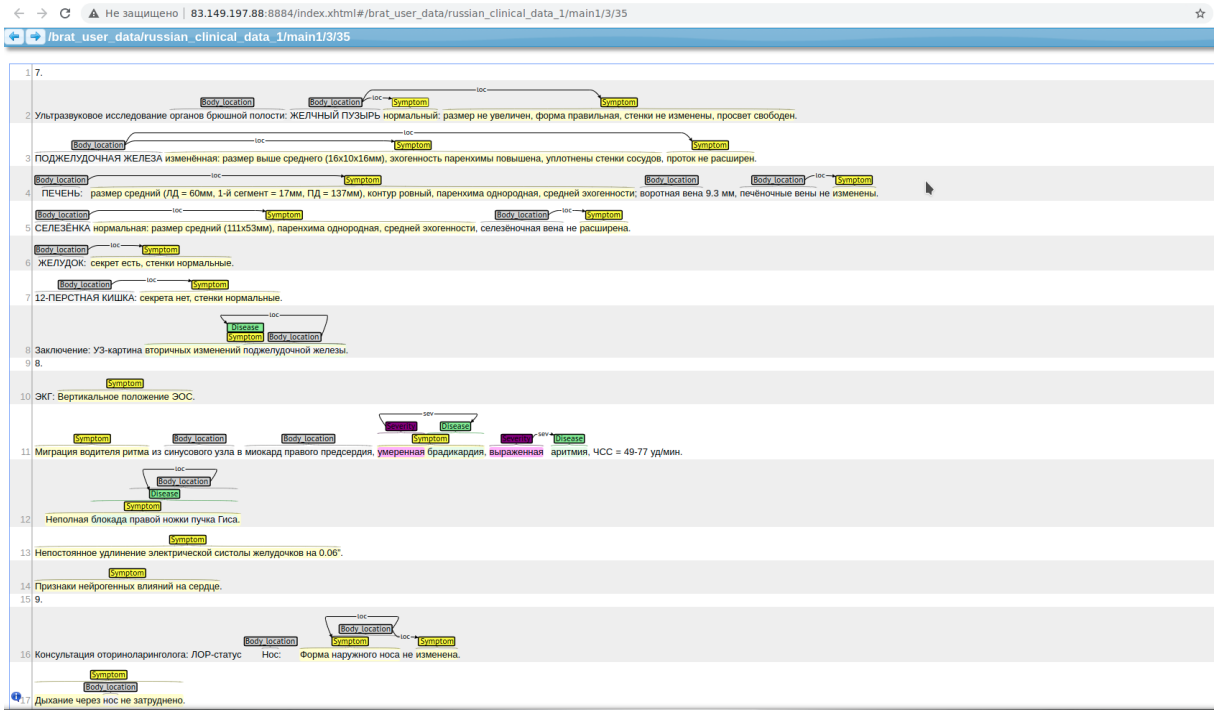

Рис. 2: Пример записи электронной медицинской карты.

  - C0018935 (процедура гематокрита) *
- MCV
  - C0863148 (анализ среднего корпускулярного объема) *
- MCH
  - C0369183 (тест на средний корпускулярный гемоглобин эритроцитов) *
- MCHC
  - C0474535 (определение средней концентрации корпускулярного гемоглобина) *
- RDW
  - C0427460 (определение ширины распределения эритроцитов) *
- Plt
  - C0005821 (Тромбоциты)
  - C0032181 (измерение количества тромбоцитов) *

2. Если выделенный фрагмент является общим понятием, нормализуйте его как общее понятие.

   - дальнейшая терапия ⇒ C0087111 (Лечебная процедура)
   - болезнь ⇒ C0012634 (Болезнь)
   - процедура ⇒ C0184661 (Интервенционная процедура)
   - Инфекция ⇒ C0009450 (Инфекционные заболевания)

3. Если выделенный фрагмент текста состоит из нескольких отдельных концепций, вы можете аннотировать его несколькими концепциями и записать их CUI в поле комментариев. Например, «чрескожная транслюминальная коронарная ангиопластика проксимального отдела левой передней нисходящей ветви» может быть аннотирована двумя CUI: C2936173 «Чрескожная транслюминальная коронарная ангиопластика» и C0226033 «Структуры проксимальной части передней нисходящей ветви левой коронарной артерии». Запишите CUI в порядке появления в выделенном фрагменте текста, разделяя пробелом.

4. Если упоминание концепции включает такие модификаторы, как «легкий», «тяжелый», «известный», «положительный» и т.д., модификатор следует включить при поиске соответствующего CUI. Однако часто бывает, что в словаре будет только более общая концепция, которая не включает вышеуказанный модификатор. В этом

случае выберите оптимальный CUI, игнорируя модификатор, а идентификатор модификатора запишите в поле комментариев. Существуют модификаторы, которые отсутствуют в словаре (например точные размеры органов), такие модификаторы следует игнорировать. Однако модификаторы, которые неотделимы по смыслу от основной концепции следует обязательно учитывать при выборе оптимального CUI (например, «Острая ишемия миокарда»).

- Пример: Теплая конечность
  I. C0424742 (Теплые конечности)
- Пример: повышенный уровень холестерина.
  I. C0020443 (Гиперхолестеринемия)
- Пример: положительный семейный анамнез
  I. C0241889 (Семейный анамнез)
  II. C1446409 (положительный)
- Другие возможные модификаторы:
  I. C1302234 (со смертельным исходом)
  II. C0205081 (Умеренный (модификатор серьезности))
  III. C1299392 (от легкой до умеренной)
  IV. C1299393 (от умеренного до тяжелого)

Обратите внимание, если модификатор не относится напрямую к выделенной концепции, вы не должны его учитывать.

5. Если выделенное упоминание концепции записано во множественном числе и словаре существует соответствующая концепция во множественном числе, следует выбрать идентивикатор CUI данной концепции. В противном случае следует искать концепции в единственном числе.

6. Когда не удается найти подходящую концепцию, выделенному участку текста назначается метка «без CUI».

7. Иногда в словаре встречаются концепции, на первый взгляд абсолютно идентичные. В таких случаях необходимо воспользоваться сервисом «UMLS Metathesaurus Browser», в котором есть дополнительная информация о концепциях словаря, которая поможет выбрать наиболее верный идентификатор.

A.4. Инструмент аннотации

Процесс аннотации проводится с использованием специализированного веб-сервиса brat (`https://brat.nlplab.org/`). Вам будет предоставлена гугл-таблица, в которой содержаться ссылки на записи электронных медицинских карт. Каждая ссылка веден на отдельную запись. На рисунке 2 представлен пример записи электронной медицинской карты в сервисе для аннотации brat.

Каждый выделенный фрагмент текста — это медицинская концепция, которую нужно связать с соответствующим идентификатором CUI. Для того, чтобы вызвать меню выбора идентификатора вам необходимо сделать двойной щелчок на метке с типом сущности, расположенном над выбранным фрагментом текста (рисунок 3).

В открывшемся окне сопоставьте концепт, указанный в строке «Ref» в поле «Normalization», с выделенным фрагментом текста, указанным в поле «Text». Если концепт указан верно, нажмите кнопку «OK» и перейдите к следующему выделенному фрагменту текста. Если концепт указан неверно, дважды щелкните по строке «Ref» в поле «Normalization», откроется окно поиска концептов (рисунок 4).

В открывшемся окне нажмите на кнопку «Search UMLS», система выполнит поиск в словаре схожих по тексту концептов и выведет их списком (рисунок 5)

Выберите подходящий концепт из списка и нажмите кнопку «OK» (или дважды щелкните по нужному концепту). Система сохранит ваш выбор и вернет в предыдущее окно, там так же необходимо нажать на кнопку «OK», система запомнит ваш выбор, можно переходить к аннотации следующего выбранного участка текста. Если в списке найденных системой концептов не было найдено подходящего, можно попробовать изменить поисковую фразу в поле «Query», по которой производится поиск, и выполнить поиск вновь. В большинстве случаев правильный подбор поисковой фразы позволяет найти наиболее подходящую концепцию в словаре. Если даже после изменения поиско-

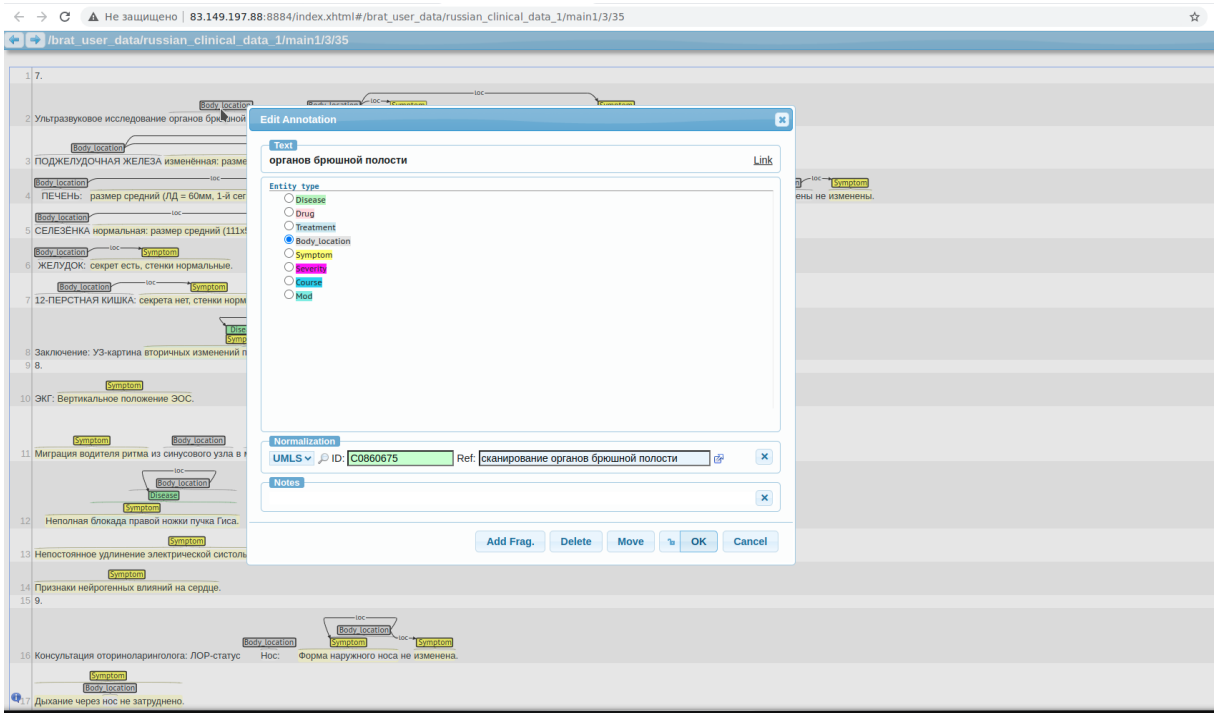

Рис. 3

вой фразы не удается найти подходящий концепт, вернитесь в предыдущее меню, нажав на кнопку «cancel» и удалите идентификатор, находящийся в строке ID в поле «Normalization», в открывшемся окне. Удаление идентификатора очистит строку «Ref», это будет служить индикатором того, что к выделенному фрагменту текста не удалось подобрать подходящий концепт.

Некоторые выделенные фрагменты могут быть автоматически аннотированы несколькими концептами одновременно. Вам нужно оставить только тот концепт, семантический тип которого лучше подходит. Те концепты, которые не подходят, нужно удалить, щелкнув по ним и удалив идентификатор, находящийся в строке ID в поле «Normalization», в открывшемся окне.

Когда вы закончите аннотирование записи электронной медицинской карты вернитесь в гугл-таблицу, в которой содержаться ссылки на соответствующие записи, в поле «status» поставьте цифру 1 и переходите по следующей ссылке.

A.5. Пример аннотации

На электрокардиограмме зафиксирован нормальный синусовый ритм, увеличение левого предсердия, отклонение оси влево, нельзя исключить старый инфаркт миокарда в передне-перегородочной области. В AVL — заметные зубцы Q, соответствующие боковому инфаркту миокарда. Неспецифические изменения сегмента ST-T.

Аннотация:

- электрокардиограмме ⇒ C1623258 (Электрокардиография)
- увеличение левого предсердия ⇒ C0238705 (гипертрофия левого предсердия)
- отклонение оси влево ⇒ C0232297 (отклонение оси влево)
- старый инфаркт миокарда в передне-перегородочной области ⇒ C0027051 (Инфаркт миокарда), C0580836 (Старый), C0225904 (Структура миокарда передне-перегородочной области)
- AVL - заметные зубцы Q ⇒ C1287077 (зубец Q - обнаружение), C0205402 (заметный), C0449216 (aVL - участок отведения)
- боковому инфаркту миокарда ⇒ C0027051 (Инфаркт миокарда), C0225823 (Структура бокового миокарда)

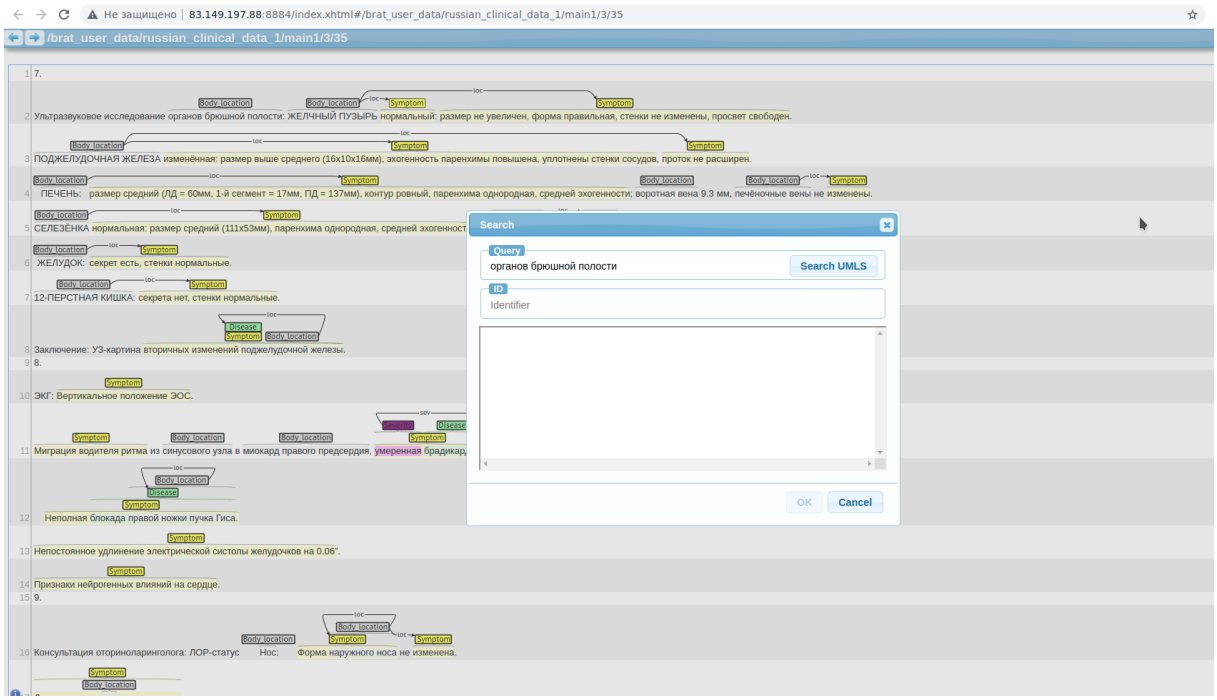

Рис. 4

- Неспецифические изменения сегмента ST-T ⇒ C1997940 (Неспецифические отклонения ST-T на электрокардиограмме)

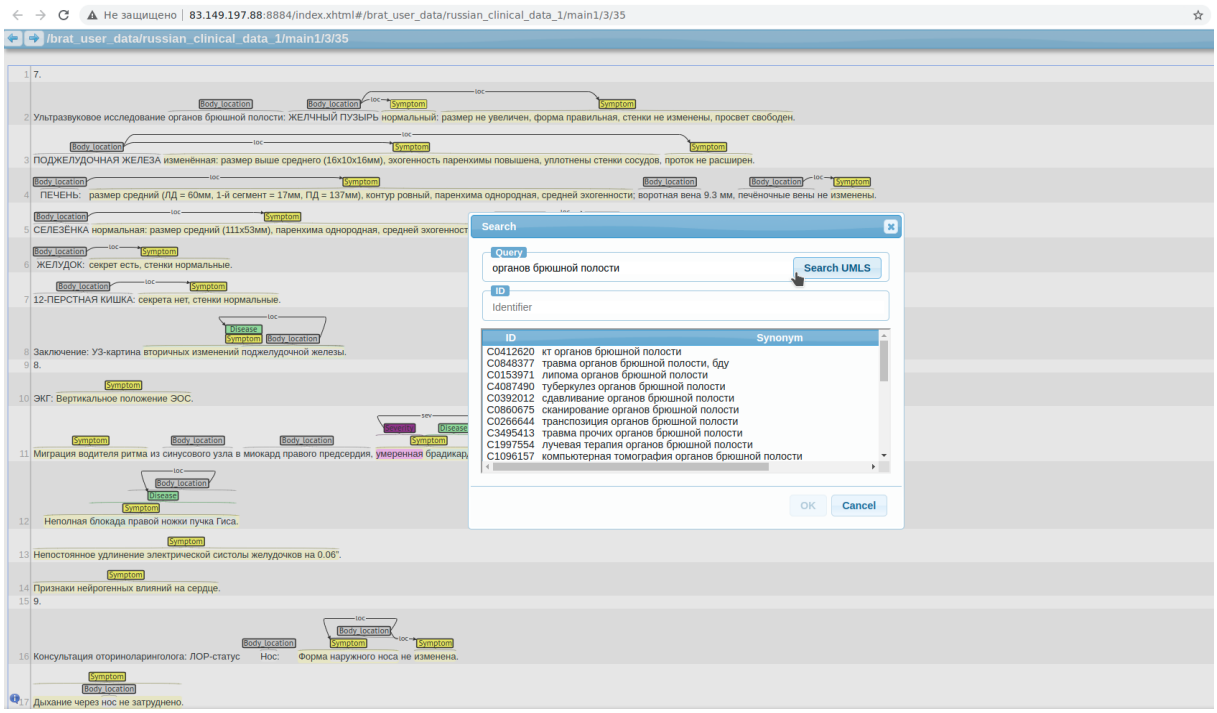
Не защищено | 83.149.197.88:8884/index.xhtml#/brat_user_data/russian_clinical_data_1/main1/3/35
/brat_user_data/russian_clinical_data_1/main1/3/35
Search
Query
органов брюшной полости
Search UMLS
ID
Identifier
ID
Synonym
C0412620 кт органов брюшной полости
C0848377 травма органов брюшной полости, бду
C0153971 липома органов брюшной полости
C4087490 туберкулез органов брюшной полости
C0392012 сдавливание органов брюшной полости
C0860675 сканирование органов брюшной полости
C0266644 транспозиция органов брюшной полости
C3495413 травма прочих органов брюшной полости
C1997554 лучевая терапия органов брюшной полости
C1096157 компьютерная томография органов брюшной полости
OK
Cancel

Рис. 5