DOI:

А.А. Баранов[1], Л.С. Намазова-Баранова[1], И.В. Смирнов[2], Д.А. Девяткин[2], А.О. Шелманов[2], Е.А. Вишнёва[1], Е.В. Антонова[1], В.И. Смирнов[1]

[1] Научный центр здоровья детей, Москва, Российская Федерация
[2] Институт системного анализа Федерального исследовательского центра «Информатика и управление» Российской академии наук, Москва, Российская Федерация

Технологии комплексного интеллектуального анализа клинических данных

Обоснование. *Медицинские учреждения генерируют большой поток данных, содержащих важную информацию о пациентах. В структурированном виде, как правило, хранятся результаты анализов. Такие данные, как анамнезы, результаты осмотров, описания результатов обследований (УЗИ, ЭКГ, рентген) и другие, имеют неструктурированную форму (в виде текстов на естественном языке). Используя методы интеллектуальной обработки накопленных массивов данных, можно автоматизировать решение многих задач, возникающих в клинической практике, повысив таким образом качество медицинской помощи.* ***Цель исследования****: создание комплексной системы интеллектуальной обработки данных в многопрофильном педиатрическом центре.* ***Методы***. *Извлечение информации из клинических текстов на русском языке осуществляется на основе их полного лингвистического анализа. Из текстов извлекаются названия заболеваний, симптомов, областей тела, к которым относится заболевание, а также лекарственных препаратов. В тексте распознаются такие атрибуты заболеваний, как «отрицание» (указывает на то, что заболевание отсутствует), «не пациент» (указывает на то, что заболевание относится не к пациенту, а к его родственнику), «тяжесть заболевания», «течение заболевания». Для извлечения информации применяют медицинские тезаурусы, набор вручную составленных шаблонов, а также различные методы на основе машинного обучения. Полученные из текстов данные используются для решения задачи автоматической диагностики хронических заболеваний. Предложен метод на основе машинного обучения для классификации пациентов со схожими нозологиями, а также метод для определения наиболее информативных признаков.* ***Результаты***. *Экспериментальное исследование разработанных методов проводилось на обезличенных историях болезни пациентов педиатрического центра. Проведена оценка качества разработанных методов извлечения информации из клинических текстов на русском языке. Проведена экспериментальная оценка метода автоматической диагностики для пациентов с болезнями органов дыхания, с аллергическими, нефрологическими и ревматическими болезнями. Определены наиболее информативные признаки, а также подходящие методы машинного обучения для классификации пациентов по группам заболеваний. Получены шаблонные комбинации признаков заболеваний. Использование данных позволило повысить качество диагностики хронических заболеваний.* ***Заключение***. *Разработанные методы были реализованы в системе интеллектуальной обработки данных в многопрофильном педиатрическом центре. Проведенные исследования свидетельствуют о перспективности использования системы для повышения качества медицинской помощи пациентам детской возрастной категории.*

Ключевые слова*: клинические тексты, извлечение информации, машинное обучение, анализ медицинских данных, интеллектуальный анализ данных.*

*(****Для цитирования****:* Баранов А.А., Намазова-Баранова Л.С., Смирнов И.В., Девяткин Д.А., Шелманов А.О., Вишнёва Е.А., Антонова Е.В., Смирнов В.И. Технологии комплексного интеллектуального анализа клинических данных. *Вестник РАМН.* 2016;71(2):160–171. doi:)

Обоснование

Системы интеллектуального анализа медицинской информации применяются для поддержки принятия решений при диагностике заболеваний, выполнении лечебных мероприятий, в целях контроля действий медицинского персонала и для предупреждения о наступлении потенциально опасных изменений в состоянии здоровья пациентов [1]. Применению методов машинного обучения для анализа структурированных медицинских данных посвящено значительное число работ. Так, у N. Isa [2] представлена система, использующая методы интеллектуальной диагностики для прогнозирования рака груди. В работе A. Al-Hyari и соавт. [3] описана система, позволяющая выполнить диагностику стадии хронической почечной недостаточности. M. Fazel Zarandi и соавт. [4] решают задачу автоматизированной диагностики астмы. В качестве входных данных ученые используют показатели здоровья в количественной и качественной форме, антропометрические данные, генетические тесты, по которым строится набор нечетких правил, составляющих основу для последующей классификации пациентов. Аналогичная диагностика проводится и при помощи деревьев решений [5]. Метод логистической регрессии применяется для диагностики ревматоидного артрита [6]. Среди исследований, посвященных выделению значимых признаков заболеваний, следует отметить работы A. Wright и соавт. и S. Doddi и соавт. [7, 8]. Используя априорный алгоритм, авторы получили эмпирические оценки взаимосвязи различных медикаментов, результатов лабораторных тестов и заболеваний. Этот метод применялся также для интеллектуальной диагностики диабета. Ассоциативные правила, сформированные априорным алгоритмом, использовались затем для построения классификатора [9].

Помимо структурированных данных медицинские учреждения генерируют большой объем неструктурированных текстов, содержащих важную информацию о здоровье пациентов. К ним относятся анамнезы, результаты осмотров, описания результатов обследований, таких как ультразвуковые (УЗИ), электрокардиографические (ЭКГ), рентгенологические и др. Анализ клинических текстов на сегодняшний день является одним из быстроразвивающихся актуальных научных направлений, которое находится на стыке компьютерной лингвистики

и медицины [10, 11]. Методы извлечения информации из клинических текстов позволяют повысить эффективность анализа клинических данных и улучшить качество медицинского обслуживания пациентов.

На практике часто возникает потребность в универсальной системе, которая могла бы анализировать разнородные данные, как в структурированной форме, так и в неструктурированной, и позволяла бы автоматизировать широкий спектр мероприятий в рамках лечебного процесса. При создании такой системы необходима агрегация разнообразных методов анализа медицинских данных и текстов.

В статье представлена комплексная система интеллектуальной обработки данных в многопрофильном педиатрическом центре, которая решает следующие задачи:

- автоматическая диагностика хронических заболеваний у детей;
- выявление наиболее значимых для диагностики признаков заболеваний;
- выявление скрытых зависимостей в клинических данных.

Представлены также методы, примененные в системе для извлечения информации из клинических текстов и анализа медицинских данных. Проведено экспериментальное исследование разработанной системы на данных многопрофильного педиатрического центра.

Цель исследования: создание комплексной системы интеллектуальной обработки данных в многопрофильном педиатрическом центре.

Методы

Дизайн исследования

Проведено экспериментальное исследование с целью машинного извлечения информации из клинических текстов деперсонализованных историй болезни пациентов многопрофильного педиатрического центра.

Условия проведения

Исследование проведено на базе ФГБУ «Научный центр здоровья детей» Минздрава России (ФГБУ «НЦЗД» Минздрава России).

Продолжительность исследования

Исследование проведено в 2013–2015 гг.

Описание медицинского вмешательства

Медицинское вмешательство не проводилось.

Методы регистрации исходов

Интеграция методов интеллектуальной обработки данных и текстов в единой комплексной системе

Комплексная система интеллектуальной обработки данных в многопрофильном педиатрическом центре разработана в соответствии с предложенной ранее сервисориентированной распределенной архитектурой [12], согласно которой каждый компонент системы представляет собой отдельный сервис, а взаимодействие между ними осуществляется посредством унифицированного протокола доступа к объектам (Simple

A.A. Baranov[1], L.S. Namazova-Baranova[1], I.V. Smirnov[2], D.A. Devyatkin[2], A.O. Shelmanov[2], E.A. Vishneva[1], E.V. Antonova[1], V.I. Smirnov[1]

[1] Scientific Center of Children's Health, Moscow, Russian Federation

[2] Institute for Systems Analysis, Federal Research Center «Computer Science and Control» of Russian Academy of Sciences, Moscow, Russian Federation

Technologies for Complex Intelligent Clinical Data Analysis

The paper presents the system for intelligent analysis of clinical information. Authors describe methods implemented in the system for clinical information retrieval, intelligent diagnostics of chronic diseases, patient's features importance and for detection of hidden dependencies between features. Results of the experimental evaluation of these methods are also presented. ***Background****: Healthcare facilities generate a large flow of both structured and unstructured data which contain important information about patients. Test results are usually retained as structured data but some data is retained in the form of natural language texts (medical history, the results of physical examination, and the results of other examinations, such as ultrasound, ECG or X-ray studies). Many tasks arising in clinical practice can be automated applying methods for intelligent analysis of accumulated structured array and unstructured data that leads to improvement of the healthcare quality.* ***Aims****: the creation of the complex system for intelligent data analysis in the multi-disciplinary pediatric center.* ***Materials and methods****: Authors propose methods for information extraction from clinical texts in Russian. The methods are carried out on the basis of deep linguistic analysis. They retrieve terms of diseases, symptoms, areas of the body and drugs. The methods can recognize additional attributes such as «negation» (indicates that the disease is absent), «no patient» (indicates that the disease refers to the patient's family member, but not to the patient), «severity of illness», «disease course», «body region to which the disease refers». Authors use a set of hand-drawn templates and various techniques based on machine learning to retrieve information using a medical thesaurus. The extracted information is used to solve the problem of automatic diagnosis of chronic diseases. A machine learning method for classification of patients with similar nosology and the method for determining the most informative patients' features are also proposed.* ***Results****: Authors have processed anonymized health records from the pediatric center to estimate the proposed methods. The results show the applicability of the information extracted from the texts for solving practical problems. The records of patients with allergic, glomerular and rheumatic diseases were used for experimental assessment of the method of automatic diagnostic. Authors have also determined the most appropriate machine learning methods for classification of patients for each group of diseases, as well as the most informative disease signs. It has been found that using additional information extracted from clinical texts, together with structured data helps to improve the quality of diagnosis of chronic diseases. Authors have also obtained pattern combinations of signs of diseases.* ***Conclusions****: The proposed methods have been implemented in the intelligent data processing system for a multidisciplinary pediatric center. The experimental results show the availability of the system to improve the quality of pediatric healthcare.*

Key words*: data mining in healthcare, natural language processing of clinical texts, hospital information system, information extraction.*

(For citation*:* Baranov AA, Namazova-Baranova LS, Smirnov IV, Devyatkin DA, Shelmanov AO, Vishneva EA, Antonova EV, Smirnov VI. Technologies for Complex Intelligent Clinical Data Analysis. *Annals of the Russian Academy of Medical Sciences*. 2016;71(2):160–171. doi:)

Object Access Protocol, SOAP) [13]. Основными компонентами системы являются лингвистический процессор Exactus [14], подсистема извлечения информации из клинических текстов, а также подсистема анализа структурированных данных, в которую входят различные классификаторы, основанные как на системах правил, так и на машинном обучении. Для решения различных задач обработки данных в системе определены наборы правил, определяющие порядок обработки информации, и компоненты, которые будут задействованы при этом.

Рассмотрим процесс обработки информации при решении задачи диагностики (рис.). Полуструктурированные данные экспортируются из медицинской информационной системы (МИС) в формате XML. Эти документы содержат как структурированные строковые поля и числовые данные, так и крупные неструктурированные текстовые блоки: анамнезы, результаты осмотров и др. XML-парсер (расширяемый язык разметки документов) анализирует поступающие документы, выделяет текстовые и структурированные данные, перенаправляет тексты на лингвистический анализ. В результате лингвистического анализа строится реляционно-ситуационная структура текста [15], содержащая информацию о морфологических характеристиках слов, синтаксической и ролевой структуре предложений, а также семантических отношениях. Построенная структура используется для извлечения информации из текста — медицинских терминов, их атрибутов, числовых данных и др. Извлеченная информация передается подсистеме анализа структурированных данных, в которой из нее формируются признаки для работы классификаторов на основе правил или моделей машинного обучения.

Методы извлечения информации из клинических текстов на русском языке

Подробное описание методов извлечения информации из клинических текстов на русском языке представлено в исследовании A. Shelmanov и соавт. [16]. В настоящей работе рассмотрим основные принципы, заложенные в них.

Цепочка извлечения информации из клинических текстов состоит из 7 компонентов: первый осуществляет извлечение и нормализацию результатов различных клинических анализов, которые обычно представляют собой строковые или числовые значения; второй распознает в тексте медицинские термины по заданным кодификаторам (заболевания, препараты, медицинские процедуры и др.); затем подключаются к работе компоненты, распознающие конструкции, указывающие на отсутствие заболеваний, на отношение заболевания не к пациенту, определяющие тяжесть и течение заболевания, сопоставляющие заболевание и связанную с ним область тела.

Извлечение и нормализация результатов клинических анализов осуществляется с помощью системы правил и регулярных выражений.

В основе метода распознавания медицинских терминов по заданным кодификаторам лежит подход, реализованный в системе MetaMap [17]. Для терминов из заданного кодификатора строится обратный поисковый индекс по словам. В тексте отыскиваются упоминания слов из поискового индекса (стоп-слова не учитываются). Затем на основе единичных слов порождаются варианты терминов, представляющие собой более сложные конструкции в тексте. При порождении вариантов используются синтаксические связи, линейный контекст слева и справа от слова и эвристические (от др.-греч. Εὑρίσκω —

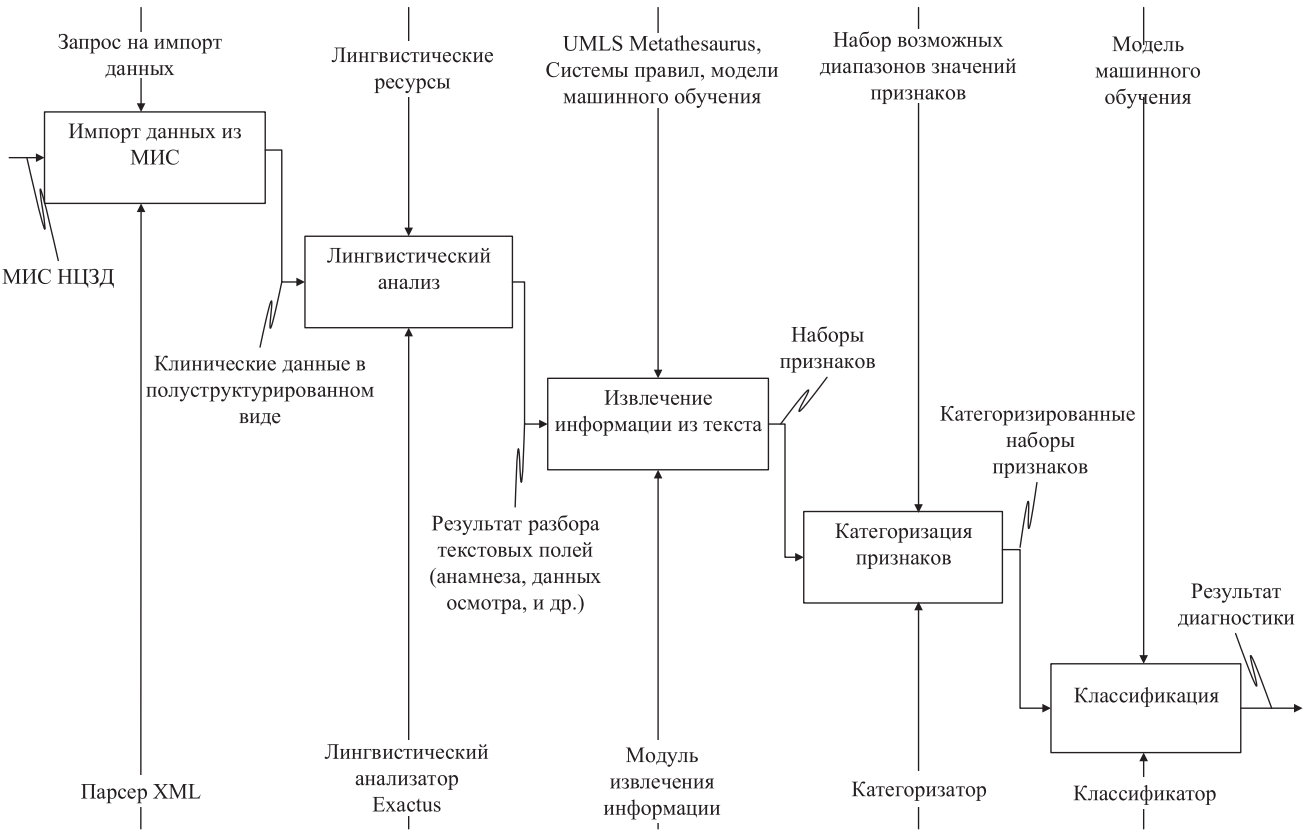

Рис. Процесс комплексной интеллектуальной обработки данных в системе
Примечание. МИС — медицинская информационная система.

отыскиваю, открываю) правила. Полученные варианты сопоставляются с терминами из кодификатора и ранжируются. Оценивается, насколько термин из кодификатора близок к сгенерированному варианту. На оценку влияют пересечение слов сгенерированного варианта и термина кодификатора; синтаксическая связность слов в варианте; способ, с помощью которого был сгенерирован вариант и другие эвристики. Варианты с низкими оценками отсекаются по порогу. Оставшиеся варианты, найденные в тексте, обычно имеют от одного до трех соответствующих им кодов из кодификатора. По кодам определяются типы найденных медицинских терминов (заболевание, симптом, лекарственный препарат и т.п.). Разработанный метод позволяет находить в тексте не только точные термины из кодификаторов, но и различные варианты их написания. В качестве кодификаторов использовались Unified Medical Language System (UMLS) Metathesaurus [18] (на русский язык переведен MeSH [19]), а также государственный реестр лекарственных средств (предварительно преобразован так, чтобы препаратам с одинаковыми действующими веществами был сопоставлен единый код) [20].

В методах распознавания конструкций, указывающих на отсутствие заболевания и на то, что заболевание не относится к пациенту, используются эвристики и лексико-синтаксические шаблоны. Ко всем упоминаниям заболеваний, найденным с помощью метода распознавания в тексте медицинских терминов по кодификаторам, применяется набор правил и шаблонов. Отыскиваются устойчивые словосочетания, указывающие на отрицания, упоминания о родственниках и др. (как в примере: «*Наследственность – у матери лекарственная аллергия на пенициллины, у старшего брата пищевая аллергия*»), которые находятся в контексте упоминания заболевания или же синтаксически связаны с ним.

Методы определения атрибутов, тяжесть заболевания и течение заболевания, а также метод установления связей между заболеваниями и областями тела основаны на машинном обучении. Алгоритм определения тяжести и течения заболевания одинаковый. Для заданной аннотации заболевания в предложении к каждому слову классификатор проставляет метку, указывающую на принадлежность к атрибуту заболевания. После того как все слова в предложении просмотрены, рядом стоящие слова с одинаковыми метками группируются в единый атрибут. После выделения атрибутов из текста другой набор классификаторов осуществляет их нормализацию, т.е. определяет значение атрибута. Для классификации используются лексические, морфологические и синтаксические признаки, извлеченные из контекста слов. Алгоритм установления связей между заболеваниями и областями тела заключается в том, что каждая пара аннотаций «заболевание–область тела», найденная в предложении, оценивается бинарным классификатором, который определяет, связаны они или нет. В качестве признаков для классификации используются дальность в токенах (обособленная последовательность символов) между аннотацией заболевания и областью тела, количество аннотаций заболеваний между ними, наличие синтаксического подчинения единому родителю, его часть речи, наличие синтаксической связи между словами аннотаций.

Методы интеллектуального анализа медицинских данных

Для решения задачи диагностики хронических заболеваний предлагается использовать методы машинного обучения на основе деревьев решений. Рассмотрим примененный метод диагностики подробнее.

Пусть $D = \{d_1, d_2,..., d_m\}$, где $m \in N^+$ — множество всех возможных диагнозов, определяемых системой. Ранее извлеченные численные показатели здоровья пациентов переводятся в категориальную форму. Для этого используется функция категоризации показателей:

$$Cat(i, val, d_{pr}, age, sex) : I \times R \times D \times R \times Sx \to C^l,\ l \in N^+$$

где $Sx = \{m,f\}$ — множество возможных значений признака «пол пациента», $C = \{low, normal, high\}$ — множество возможных категориальных значений показателя здоровья. Функции S и Cat задаются в реляционном виде экспертами в предметной области.

Для того чтобы классификатор мог работать с категориальными признаками, производится их предварительная бинаризация:

$$X = \bigcup_{\forall i \in sI} i \times C,$$

где X – пространство бинарных признаков, sI – пространство категориальных признаков.

Рассмотрим способ построения классификатора на основе деревьев решений CART [21]. Обозначим полученные ранее наборы признаков объектов как $x_i \in R^n$, где $i=1,2,...,l$, и вектор их меток, соответствующих различным заболеваниям из D, как $y \in d^n$. Дерево решений рекурсивно разбивает пространство объектов таким образом, что похожие объекты группируются вместе.

Пусть Q — подмножество объектов, разделяемых некоторой вершиной m дерева. Зададим разделитель $\theta = (j, t_m)$, состоящий из признака j и порога t_m, дробящий исходное множество объектов, соответствующих вершине на подмножества $Q_{left}\theta$ и $Q_{right}\theta$:

$$Q_{left}(\theta) = \{(x,y) \mid x_j \le t_m\}$$

$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta).$$

Оценка разделителя $G(Q,\theta)$, заданного для вершины m, вычисляется с использованием функции неоднородности по Джини ($Gin(\cdot)$):

$$G(Q,\theta) = \frac{n_{left}}{N_m} Gin(Q_{left}(\theta)) + \frac{n_{right}}{N_m} Gin(Q_{right}(\theta)),$$

где N_m — размер множества объектов, разделяемых вершиной m, n_{left}, n_{right} — размер подмножеств объектов, соответствующих левому и правому поддеревьям вершины m. При построении дерева выберем для вершины m разделитель с минимальной оценкой:

$$\theta^* = arg\ min_\theta G(Q, \theta).$$

Затем рекурсивно выберем разделители для подмножеств $Q_{left}(\theta^*)$ и $Q_{right}(\theta^*)$. Процесс продолжается до тех пор, пока не будет достигнута требуемая высота дерева.

Для каждого предварительного диагноза на основе тестового набора данных строится отдельный классификатор. В качестве положительных примеров для обучения используются результаты обследований пациентов с целевым диагнозом, а в качестве отрицательных — результаты обследований пациентов со схожей симптоматикой.

Для повышения качества работы методов диагностики используются композиции классификаторов на основе деревьев решений.

Первым рассмотренным способом построения композиции стал случайный лес (Random Forest) [22]. Он представляет собой лес деревьев решений:

$$H=\{h(x,\Theta_k),\ k=1, ..,n\},$$

где $n \in N^+$, $h(x, \theta)$ — решающее дерево, $\{\Theta_k\}$ — векторы параметров, и все деревья в нем участвуют в голосовании за диагноз, соответствующий вектору признаков заболе-

вания *x*. Все деревья в лесу обучаются на подмножествах данных, случайным образом сгенерированных на основе исходной обучающей выборки, причем для обучения каждого дерева отбирается случайное подмножество признаков заболеваний. Для выбора разделяющих признаков, как и в деревьях CART, используется функция неоднородности по Джини.

Вторым рассмотренным способом построения композиций стал градиентный бустинг (от англ. Boosting — улучшение; процедура последовательного построения композиции алгоритмов машинного обучения) на деревьях решений [23]. В этом методе составляется линейная свертка классификаторов:

$$H_m(x) = \sum_{m=1}^{M} b_m h(x, \theta_m),$$

где $b_m \in R$ и m — длина композиции.

Эта композиция должна быть оптимальной в смысле минимизации функционала ошибки Q:

$$Q = \sum_{i=1}^{l} L(y_i, H_m(x_i)) \to \min.$$

Будем последовательно строить такую композицию при помощи рекурсивного алгоритма, каждый раз добавляя в сумму классификатор, оптимизирующий Q:

$$H_m(x) = H_{m-1}(x) + b_m h(x, \theta_m).$$

Обозначим ∇Q_i i-компонент градиента функции Q. Тогда параметры и вес m-слагаемого могут быть получены методом градиентного спуска:

$$b_m = \arg\min_{b \in R} \sum_{i=1}^{l} L(y_i, H_{m-1}(x_i) - b \nabla Q_i),$$

$$\theta_m = \arg\min_{\theta \in R^n} \sum_{i=1}^{l} L(y_i, b_m h(x_i, \theta)).$$

Процесс обучения прекращается, когда величина функционала ошибки становится меньше предварительно заданного параметра $L \in R$, либо когда длина свертки превысит заранее определенную величину $M \in N^+$.

Для оценки качества диагностики использовалась F_1-мера. В качестве метода вычисления этой оценки применялась десятикратная перекрестная проверка, при помощи которой для каждого метода диагностики и для каждого заболевания была получена оценка F_1, а также величина ее среднеквадратичного отклонения.

Для определения относительной важности признаков, связанных с заболеваниями, использовалась мера значимости по Джини [24]. По результатам всех проходов перекрестной проверки по этой мере вычислялась суммарная значимость признаков, и отбирались наиболее значимые признаки.

Для решения задачи выявления скрытых зависимостей между признаками заболеваний, лекарственными средствами и диагнозами использовался метод формирования наборов ассоциативных правил [25]. С помощью априорного алгоритма [26] составлялись шаблонные комбинации признаков различных заболеваний.

Рассмотрим *процесс формирования ассоциативных правил* более подробно.

Пусть $I = \{i_1, i_2, ..., i_n\}$ — набор булевых признаков, $D = \{t_1, t_2, ..., t_m\}$ — набор транзакций. Каждая транзакция в D состоит из элементов множества I. Пусть ассоциативное правило является импликацией вида $X \Rightarrow Y$, где $X, Y \subseteq I$ и $X \cap Y = \varnothing$. Каждое такое правило включает подмножества X и Y, где X называется антецедентом (от лат. *Antecedens* — предшествующее), а Y — следствием. Величина поддержки подмножества X относительно $T \subset D$ определяется как доля транзакций в T, которые содержат X. Величина доверия к ассоциативному правилу $X \Rightarrow Y$ относительно множества T — мера того, как часто элементы из Y появляются среди транзакций в T, содержащих X. Тогда задача построения набора ассоциативных правил состоит в выявлении подмножеств с высоким уровнем поддержки и формировании в дальнейшем набора ассоциативных правил с высоким доверием. Для уменьшения вычислительной сложности построения ассоциативных правил используется априорный алгоритм. В нем для выявления наборов признаков с высоким уровнем поддержки используется поиск в ширину. При построении ассоциативных правил применяется антимонотонное свойство поддержки:

$$\forall A, B: (A \subseteq B) \Rightarrow s(A) \geqslant s(B).$$

Это свойство состоит в том, что мера поддержки множества никогда не превышает поддержку его подмножеств. Отсюда следует, что любое подмножество признаков будет часто встречающимся тогда и только тогда, когда все его подмножества будут встречаться также часто. Это означает, что формирование набора правил можно выполнять рекурсивно: сначала генерируются правила на основе подмножеств длины k-1, затем составляются правила, в которые входят подмножества признаков длиной k.

Благодаря низкой вычислительной сложности такой подход к формированию наборов ассоциативных правил применим для анализа больших наборов признаков пациентов.

Этическая экспертиза

Исследование проводилось в соответствии с Федеральным законом № 152-ФЗ от 27.07.2006 (ред. от 21.07.2014) «О персональных данных».

Статистический анализ

Принципы расчета размера выборки

Использование выборок малого размера для настройки методов интеллектуального анализа данных или текстов может приводить к проблеме переобучения. Теоретическая оценка размера выборки, достаточного для корректной настройки методов машинного обучения, может быть определена для используемого класса алгоритмов на основе размерности Вапника–Червоненкиса [27]. Однако такие оценки сильно завышены и плохо применимы на практике [28], поэтому для настройки методов интеллектуального анализа данных или текстов используют наибольшие доступные выборки, а оценка их применимости для обучения производится эмпирически.

Методы статистического анализа данных

Для получения оценки обобщающей способности методов машинного обучения обычно используют процедуру многократной перекрестной проверки. В рамках этой процедуры анализируемая выборка произвольно разбивается на n блоков одинакового размера. Затем выполняется n тестовых итераций: в ходе каждой из них один из блоков используется для контроля (определения оценки качества работы метода), а остальные применяются для обучения (настройки) исследуемого метода. По итогам выполнения этой процедуры вычисляется среднее арифметическое полученной эмпирической оценки качества обучения метода. В качестве эмпирической оценки обобщающей способности предложенных методов использовалась F_1-мера.

Результаты

Основные результаты исследования

Результаты экспериментальных исследований методов извлечения информации из клинических текстов

Размеченный корпус клинических текстов. Для применения методов машинного обучения в задачах извлечения информации из клинических текстов и для проведения экспериментальных исследований этих методов был создан размеченный корпус клинических текстов на русском языке. В состав корпуса вошли более 120 деперсонализованных историй болезни пациентов педиатрического центра с аллергическими, ревматическими и нефрологическими заболеваниями, а также болезнями органов дыхания. Обезличенные карты включали в себя эпикризы, рекомендации и отчеты, фиксирующие результаты различных медицинских обследований (УЗИ, ЭКГ, рентгенографию и др.). Совместно с экспертами педиатрического центра была определена важная для врачей информация, требующая извлечения и автоматической обработки, а также выработаны соглашения и инструкции по разметке сущностей, атрибутов и связей. При составлении инструкций по разметке учитывался опыт зарубежных семинаров, таких как CLEF eHealth [29], для которых создавались схожие ресурсы. Специалисты в области медицины разметили в корпусе более 18 000 сущностей, а также более 12 000 атрибутов и связей.

Результаты экспериментальных исследований метода извлечения заболеваний и лекарственных препаратов

Для экспериментальных исследований метода извлечения заболеваний из разработанного корпуса был выделен подкорпус, состоящий из 30 случайно выбранных историй болезней. Остальная часть корпуса использовалась для настройки параметров метода. Рассчитывались точность, полнота и F_1-мера (подробное описание различных метрик оценки качества методов машинного обучения приведено в [30]). При этом пересечение аннотации, полученной автоматически с помощью разработанной системы, с аннотацией «золотого стандарта» считалось правильным ответом системы. Разработанный метод сравнивался с двумя другими, более простыми «базовыми» методами. Первый «базовый» метод отмечает в тексте любые слова, которые встречаются в терминах из тезауруса, относящихся к заболеваниям. Этот алгоритм потенциально обладает максимальной полнотой, но низкой точностью. Второй «базовый» метод помечает словосочетание как термин, только если все слова, входящие в словосочетание, присутствуют в термине из тезауруса, представленном в виде «мешка слов». Этот алгоритм потенциально обладает низкой полнотой, но максимальной точностью. Результаты оценки методов представлены в табл. 1.

Для экспериментальной оценки метода извлечения лекарственных препаратов из клинических текстов использовалась та же методика, что и для оценки метода извлечения заболеваний. Точность метода составила 84,3%, полнота — 74,6%, F_1-мера — 79,2%.

Результаты экспериментальных исследований методов распознавания конструкций, указывающих на отсутствие заболевания и на то, что заболевание не относится к пациенту

Экспериментальная оценка методов распознавания конструкций, указывающих на отсутствие заболевания (атрибут «отрицание») и на то, что заболевание не относится к пациенту (атрибут «не пациент»), проводилась на всем размеченном корпусе, поскольку количество подобных аннотаций в нем было невелико. Так как оба метода основаны на правилах, это допущение не снижало объективности оценки методов. Стоит также отметить, что в экспериментах, помимо атрибутов заболеваний, учитывались атрибуты аннотаций симптомов. Результаты экспериментов представлены в табл. 2.

Результаты экспериментальных исследований методов извлечения и нормализации атрибутов «тяжесть» и «течение» заболевания, а также метода, устанавливающего связи между областями тела и заболеваниями

В экспериментальных исследованиях методов извлечения и нормализации атрибутов «тяжесть заболевания» и «течение заболевания», а также метода, устанавливающего связи между областями тела и заболеваниями, использовалась пятикратная перекрестная проверка на размеченном корпусе. Отдельно оценивалось качество извлечения атрибутов и их нормализации. Чтобы исключить ошибки модуля определения заболеваний, мы оценивали только атрибуты, которые относятся к заболеваниям, выделенным нашей системой. В задаче извлечения атрибутов «тяжесть заболевания» и «течение заболевания» рассчитывались нестрогие оценки точности, полноты и F_1-меры. При нестрогой оценке пересечение аннотации, полученной автоматически с помощью системы, с аннотацией «золотого стандарта» считалось правильным ответом системы.

В задаче нормализации атрибутов оценивалась только аккуратность (Accuracy), поскольку это задача классификации без «пустого» класса.

В задаче установления связей между областями тела и заболеваниями учитывались только связи между заболеваниями и областями тела, которые были определены на предыдущих этапах анализа текста.

Для каждого метода и для каждой задачи проводились эксперименты с четырьмя классификаторами на основе машинного обучения: линейный метод опорных векторов, метод опорных векторов с радиальным яром, случайный лес, AdaBoost (сокращение от Adaptive Boosting — алгоритм усиления классификаторов путем объединения их в комитет) [31].

В табл. 3 представлены результаты экспериментальной оценки методов извлечения атрибутов «тяжесть заболевания» и «течение заболевания», а также метода установления связей между областями тела и заболеваниями,

Таблица 1. Результаты экспериментальной оценки методов извлечения заболеваний

Метод	Полнота, %	Точность, %	F_1-мера, %
Разработанный метод	72,8	95,1	82,4
Базовый 1	84,9	9,3	16,7
Базовый 2	69,8	99,2	81,9

Таблица 2. Результаты экспериментальной оценки методов распознавания конструкций, указывающих на отсутствие заболевания и на то, что заболевание не относится к пациенту

Метод	Полнота, %	Точность, %	F_1-мера, %
Определение атрибута «отрицание»	98,7	95,3	97,0
Определение атрибута «не пациент»	90,9	96,8	93,8

Таблица 3. Результаты экспериментальной оценки методов извлечения атрибутов «тяжесть заболевания» и «течение заболевания», а также метода установления связей между областями тела и заболеваниями

Задача	Классификатор	Полнота, %	Точность, %	F_1-мера, %
Извлечение атрибута «тяжесть заболевания»	Случайный лес	93,6	82,6	87,5
Извлечение атрибута «течение заболевания»	Линейный SVM	92,3	99,2	95,7
Установление связей между областями тела и заболеваниями	RBF SVM	91,4	76,6	83,3

Таблица 4. Результаты экспериментальной оценки методов нормализации атрибутов «тяжесть заболевания» и «течение заболевания»

Задача	Классификатор	Аккуратность, %
Нормализация атрибута «тяжесть заболевания»	AdaBoost	89,8
Нормализация атрибута «течение заболевания»	Случайный лес	92,7

Таблица 5. Состав тестового набора данных

Заболевание	Похожие заболевания
Бронхиальная астма	Бронхит Аллергический ринит Муковисцидоз
IgA-нефропатия	Широкий спектр гломерулярных заболеваний
Юношеский артрит	Спондилит

а в табл. 4 — результаты оценки качества нормализации атрибутов «тяжесть заболевания» и «течение заболевания». Для каждой задачи указан результат наилучшего классификатора.

Результаты экспериментальных исследований методов анализа медицинских данных

Описание тестовых данных. Для исследования методов интеллектуального анализа медицинских данных использовался набор данных, состоящий из историй болезни пациентов НЦЗД с основными клиническими диагнозами: бронхиальная астма, юношеский артрит и спондилит, IgA-нефропатия, а также эпикризов пациентов с похожей симптоматикой (табл. 5).

Всего в тестовом наборе данных представлено более 1000 историй болезни. Они содержат данные осмотра, анамнез и результаты проведения дополнительных исследований (томографии, рентгенографии, кожные пробы) в текстовой форме (заключения) и результаты анализов (анализы мочи, крови, посевы микрофлоры) в полуструктурированном виде. Для формирования представления данных в структурированном виде использовался метод анализа медицинских текстов.

Таблица 6. Результаты экспериментов по диагностике хронических заболеваний (полный набор признаков)

Нозология	Категоризация численных признаков	Метод	F_1	σ
Бронхиальная астма	Нет	Деревья решений	0,82	0,23
		Случайный лес	0,95	0,07
		Градиентный бустинг на деревьях решений	0,78	0,23
	Да	Деревья решений	0,86	0,19
		Случайный лес	0,98	0,04
		Градиентный бустинг на деревьях решений	0,81	0,20
IgA-нефропатия	Нет	Деревья решений	0,87	0,21
		Случайный лес	0,74	0,24
		Градиентный бустинг на деревьях решений	0,88	0,19
	Да	Деревья решений	0,92	0,15
		Случайный лес	0,74	0,24
		Градиентный бустинг на деревьях решений	0,90	0,16
Юношеский артрит	Нет	Деревья решений	0,91	0,07
		Случайный лес	0,94	0,05
		Градиентный бустинг на деревьях решений	0,96	0,03
	Да	Деревья решений	0,90	0,06
		Случайный лес	0,95	0,05
		Градиентный бустинг на деревьях решений	0,97	0,03

Таблица 7. Результаты экспериментов по диагностике хронических заболеваний (только численные признаки)

Нозология	Категоризация численных признаков	Метод	F_1	σ
Бронхиальная астма	Нет	Деревья решений	0,69	0,22
		Случайный лес	0,82	0,22
		Градиентный бустинг на деревьях решений	0,70	0,20
	Да	Деревья решений	0,61	0,25
		Случайный лес	0,73	0,19
		Градиентный бустинг на деревьях решений	0,65	0,24
IgA-нефропатия	Нет	Деревья решений	0,63	0,24
		Случайный лес	0,75	0,25
		Градиентный бустинг на деревьях решений	0,69	0,24
	Да	Деревья решений	0,64	0,24
		Случайный лес	0,75	0,24
		Градиентный бустинг на деревьях решений	0,69	0,24
Юношеский артрит	Нет	Деревья решений	0,79	0,16
		Случайный лес	0,87	0,10
		Градиентный бустинг на деревьях решений	0,83	0,14
	Да	Деревья решений	0,76	0,14
		Случайный лес	0,89	0,10
		Градиентный бустинг на деревьях решений	0,84	0,12

Описание результатов. В ходе первого эксперимента выполнялось исследование методов диагностики заболеваний с/без использования признаков, выделенных из текстовых полей (данные осмотра, анамнез, результаты обследований). Для каждого заболевания выявлялось качество его дифференциации от похожих заболеваний. Результаты представлены в табл. 6 и 7. Выявлено, что использование текстовых признаков позволяет значительно улучшить качество диагностики. Отметим также, что использование категоризации признаков позволяет повысить качество диагностики и уменьшить дисперсию ошибки. Наиболее высокого качества классификации на имеющемся тестовом наборе удалось добиться с помощью случайного леса.

В ходе второго эксперимента выделялись наиболее значимые признаки, связанные с диагностируемыми заболеваниями. В качестве исходных данных использовались данные осмотра, результаты анализов и информация из анамнеза, в том числе применяемые ранее препараты. Результаты этого эксперимента представлены в табл. 8.

В результате проведения третьего эксперимента получены шаблонные комбинации признаков заболеваний:

1. Набор ассоциативных правил, выражающих шаблонные комбинации признаков для болезней верхних дыхательных путей:
   (‘ГИПЕРСЕНСИБИЛИЗАЦИЯ’, ‘АСТМА БРОНХИАЛЬНАЯ’, ‘РИНИТ’) => (‘Монтелукаст’)
   (‘ГИПЕРСЕНСИБИЛИЗАЦИЯ’, ‘АСТМА БРОНХИАЛЬНАЯ’) => (‘Будесонид’)
   (‘ГИПЕРСЕНСИБИЛИЗАЦИЯ’, ‘РИНИТ’) => (‘АСТМА БРОНХИАЛЬНАЯ’)
   (‘КАШЕЛЬ’, ‘АСТМА БРОНХИАЛЬНАЯ’) => (‘РИНИТ’)
   (‘АСТМА БРОНХИАЛЬНАЯ’) => (‘Флутиказон’)
   (‘IGE повышен’, ‘РИНИТ’) => (‘КАШЕЛЬ’)
   (‘IGE повышен’) => (‘Уровень базофилов повышен’)
   (‘IGE повышен’, ‘ГИПЕРСЕНСИБИЛИЗАЦИЯ ‘) => (‘РИНИТ’)
2. Набор ассоциативных правил, выражающих шаблонные комбинации признаков для нефрологии:
   (‘Белок в моче повышен’, ‘Холестерин повышен’, ‘СОЭ повышена’) => (‘Преднизолон’)
   (‘Холестерин повышен’, ‘Эритроциты в моче повышены’) => (‘Белок в моче повышен’)
   (‘IgG понижен’, ‘Холестерин повышен’) => (‘Белок в моче повышен’)
   (‘Холестерин повышен’, ‘СОЭ повышена’, ‘Альбумин понижен’) => (‘Белок в моче повышен’)
3. Набор ассоциативных правил, выражающих шаблонные комбинации признаков для ревматологии:
   (‘ЭКЗАНТЕМА’) => (‘Метотрексат’, ‘Метилпреднизолон’)
   (‘ЭКЗАНТЕМА’) => (‘ Метилпреднизолон’, ‘Тоцилизумаб’)
   (‘ГОЛЕНЬ’, ‘ЭКЗАНТЕМА’) => (‘Тоцилизумаб’)
   (‘Метотрексат’, ‘ЭКЗАНТЕМА’) => (‘Метилпреднизолон’)
   (‘Холестерин повышен’) => (‘Эритроциты повышены’)
   (‘С-реактивный белок повышен’) => (‘Лейкоциты повышены’)
   (‘СОЭ повышена’) => (‘Лейкоциты повышены’, ‘Холестерин повышен’)
   (‘Эритроциты в моче повышены’) => (‘Количество патологических цилиндров в моче повышено’)
   (‘Эритроциты в моче повышены’) => (‘Лейкоциты повышены’)
   (‘Количество патологических цилиндров в моче повышено’) => (‘Круглый эпителий в моче повышен’).

Таким образом, работоспособность предложенной системы была подтверждена экспериментально.

Таблица 8. Наиболее значимые признаки, связанные с заболеваниями

Нозология	Тип признаков	Наиболее значимые признаки
Юношеский артрит	Текстовые признаки	Локализация: ягодичная область Симптом: лейкоцитоз Препарат: сульфасалазин Препарат: этанерцепт
	Численные признаки	Удельное количество лейкоцитов в крови Удельное количество тромбоцитов в крови Уровень креатинина в крови Уровень гемоглобина в крови СОЭ
IgA-нефропатия	Текстовые признаки	Симптом: васкулит Симптом: повышен уровень IgA Симптом: сердечный тон Симптом: гиперемия Локализация: спина
	Численные признаки	Уровень креатинина в крови Удельное количество цилиндров в моче Удельный вес мочи Уровень холестерина в крови Наличие слизи в моче
Бронхиальная астма	Текстовые признаки	Препарат: флутиказон Препарат: сальбутамол Симптом: сенсибилизация подтверждена кожными пробами Препарат: будесонид Симптом: бронхитические изменения в легких
	Численные признаки	Уровень эозинофилов в крови Уровень лимфоцитов в крови Уровень IgE в крови СОЭ Уровень базофилов в крови

Примечание. СОЭ — скорость оседания эритроцитов.

Обсуждение

Резюме основного результата исследования

В результате исследований была экспериментально подтверждена работоспособность предложенных методов извлечения информации из клинических текстов. В том числе проведена апробация методов извлечения заболеваний и лекарственных препаратов, распознавания конструкций, указывающих на отсутствие заболевания и на то, что заболевание не относится к пациенту, методов извлечения и нормализации атрибутов «тяжесть» и «течение» заболевания, а также метода, устанавливающего связи между областями тела и заболеваниями.

Произведена также апробация методов анализа медицинских данных: диагностики хронических заболеваний; определения значимых признаков, связанных с заболеваниями; выявления шаблонных комбинаций признаков заболеваний. Показано, что использование дополнительных признаков, полученных с помощью методов извлечения информации из клинических текстов, позволяет улучшить качество диагностики. Экспериментально подтверждена работоспособность подхода к интеграции различных методов анализа данных и текстов в комплексной системе. Таким образом, работоспособность предложенной системы была подтверждена экспериментально.

Обсуждение основного результата исследования

Анализ результатов экспериментальных исследований методов извлечения информации из клинических текстов

Анализ результатов экспериментальных исследований методов извлечения заболеваний и лекарственных препаратов. По результатам проведенных экспериментов, разработанный метод извлечения заболеваний обладает наибольшим качеством (наибольшей F_1-мерой — 82,4%). Как и ожидалось, он обладает большей точностью, чем первый «базовый» метод, и большей полнотой, чем второй. Стоит отметить, что преимущество разработанного метода над вторым «базовым» по F_1-мере будет более значительным в задаче нормализации медицинских терминов и при строгом сопоставлении результатов автоматической разметки с «золотым стандартом». Второй «базовый» метод часто находит общие однословные термины, представленные в кодификаторе, но не отыскивает сложных словосочетаний.

Оценка полноты метода извлечения лекарственных препаратов оказалась немного ниже ожидаемой, поскольку разметчики при создании корпуса отмечали не только зарегистрированные в России лекарства, но и некоторые терапевтические косметические средства (например, специальные местные средства по уходу при кожных проявлениях аллергии). Было также отмечено, что некоторые ошибки, вызывающие снижение полноты, связаны с сокращениями наименований препаратов в клинических записях и использованием обобщенных названий: например, «пенициллин», тогда как в тезаурусе имеется лишь официальное наименование «бензатина бензилпенициллин».

Анализ результатов экспериментальных исследований методов распознавания конструкций, указывающих на отсутствие заболевания и на то, что заболевание не относится к пациенту. Полученные оценки качества методов показывают, что применение довольно простых правил и ша-

блонов позволяет добиваться высокого качества решения задач выделения атрибутов «отрицание» и «не пациент».

Анализ результатов экспериментальных исследований методов извлечения и нормализации атрибутов «тяжесть» и «течение» заболевания, а также метода, устанавливающего связи между областями тела и заболеваниями. Результаты экспериментов свидетельствуют о том, что разработанные методы извлечения и нормализации атрибутов «тяжесть» и «течение» заболевания, а также метод, устанавливающий связи между областями тела и заболеваниями, обладают высокой F_1-мерой и позволяют решать указанные задачи с приемлемым качеством. Тем не менее остается значительное пространство для улучшения этих методов, что может быть достигнуто с помощью применения новых методов машинного обучения и разработки более сложного признакового пространства.

Анализ результатов экспериментальных исследований методов интеллектуального анализа медицинских данных

Анализ результатов экспериментальных исследований метода диагностики хронических заболеваний. В результате проведенных экспериментов было выявлено, что использование методов градиентного бустинга и случайного леса позволяет добиться наибольшей точности и полноты диагностики хронических заболеваний по сравнению с другими методами (деревья решений). В ходе экспериментов метод диагностики на основе градиентного бустинга показал большую устойчивость, что выразилось в меньшем значении среднеквадратичного отклонения F_1-меры. Отметим, что совместное использование методов анализа структурированных и текстовых данных в рамках единой процедуры, а также выполнение предварительной категоризации числовых показателей здоровья пациентов позволило значительно повысить качество диагностики заболеваний.

Анализ результатов экспериментальных исследований метода определения значимых признаков, связанных с заболеваниями. В ходе эксперимента выполнялось обучение классификатора на основе случайного леса, и вычислялась значимость признаков заболеваний по Джини. Определялась значимость не только числовых показателей здоровья, но и текстовых признаков, таких как симптомы, их локализация, используемые препараты. Экспертный анализ выявленных значимых признаков подтвердил их связь с соответствующими хроническими заболеваниями.

Анализ результатов экспериментальных исследований метода выявления шаблонных комбинаций признаков заболеваний. В результате эксперимента выявлено наличие устойчивых неявных связей между различными показателями здоровья пациентов, а также применяемыми лекарственными препаратами и диагнозами. Произведена эвристическая оценка достоверности полученных связей. Отмечена перспектива использования метода в целях автоматизации процесса диагностики.

Заключение

В работе выполнено экспериментальное исследование методов извлечения информации из клинических текстов, методов интеллектуальной диагностики хронических заболеваний у детей, метода определения относительной значимости признаков заболевания, а также метода построения ассоциативных правил. Случайный лес показал лучшее качество диагностики по сравнению с деревьями решений CART и градиентным бустингом на имеющемся тестовом наборе данных.

К основным научно-техническим проблемам, решенным при создании системы, относятся организация первичной обработки разнородных данных о пациентах, неравномерное распределение заболеваний в обучающей выборке и интеграция различных методов интеллектуального анализа данных и текстов в рамках единой системы. Первая проблема была разрешена с помощью методов извлечения информации из текстов.

Для решения второй проблемы потребовалось применение композиции алгоритмов классификации, таких как бэггинг (случайный лес) и градиентный бустинг на деревьях решений. Использование композиций классификаторов позволяло нивелировать перекосы обучающего набора данных.

Для решения третьей проблемы была разработана специфическая архитектура системы на основе спецификации OGSA [32], которая позволила произвольным образом комбинировать разнородные методы обработки данных и текстов в рамках выполнения лечебно-диагностических процедур.

Отметим, что совместное использование методов анализа структурированных и текстовых данных в рамках единой процедуры позволило значительно повысить качество диагностики хронических заболеваний системой. Использование предварительной категоризации числовых показателей здоровья пациентов также позволило добиться устойчивого прироста качества диагностики. Выявление наиболее значимых для диагностики признаков заболеваний и скрытых зависимостей в клинических данных, по результатам экспериментального исследования разработанной системы в многопрофильном педиатрическом центре, свидетельствует о перспективах ее использования в целях автоматизации процесса диагностики. Более того, созданная комплексная универсальная система способна анализировать разнородные данные как структурированного, так и неструктурированного характера и позволяет автоматизировать не только диагностический этап медицинской помощи, но и широкий спектр мероприятий в рамках лечебного процесса при ряде хронических нозологических форм в педиатрической практике.

Использованные методы извлечения информации из клинических текстов и анализа медицинских данных, а также их агрегации позволили ускорить и усовершенствовать процесс дифференциально-диагностического поиска при таких тяжелых, хронических, инвалидизирующих заболеваниях, как бронхиальная астма, юношеский артрит и спондилит, IgA-нефропатия. Для пациента педиатрического профиля несвоевременность установки диагноза и промедление в назначении патогенетической терапии являются основной составляющей прогноза течения и исхода болезни, определяют экономическое бремя данных нозологий для здравоохранения.

Дальнейшее совершенствование системы, назначение и внедрение подсистем поддержки принятия решений и анализа результатов, а также дополнение новыми нозологическими формами необходимо для обеспечения процесса качественного оказания медицинской помощи детям. В условиях внедрения современных возможностей информационно-коммуникационных технологий интеграция разработанной модели в клиническую практику обеспечит своевременность, качество и в то же

время позволит осуществлять аудит и контроль оказания медицинской помощи пациентам детской возрастной категории.

Источник финансирования

Исследование поддержано грантом РФФИ № 13-04-12062 «офи_м» «Исследование методов и средств интеллектуального анализа данных для построения комплексной системы интеллектуальной обработки данных (на примере многопрофильного педиатрического центра)».

Конфликт интересов

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

ЛИТЕРАТУРА

1. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Biomedical informatics. Springer; 2014. p. 643–674. Doi:10.1007/978-1-4471-4474-8_22.
2. Isa NAM. Towards intelligent diagnostic system employing integration of mathematical and engineering model. In: Proceedings of International Conference on Mathematics, Engineering and Industrial Applications. AIP Publishing; 2015. p. 030002–1–030002–13. Doi:10.1063/1.4915633.
3. Abeer YA, Ahmad MA, Majid AA. Clinical decision support system for diagnosis and management of chronic renal failure. In: Proceedings of Applied Electrical Engineering and Computing Technologies. IEEE; 2013. p. 1–6. Doi:10.1109/aeect.2013.6716440.
4. Zarandi MHF, Zolnoori M, Moin M, Heidarnejad H. A fuzzy rule-based expert system for diagnosing asthma. Transaction E: Industrial Engineering. 2010;17(2):129–142.
5. Prosperi MC, Marinho S, Simpson A, Custovic A, Buchan IE. Predicting phenotypes of asthma and eczema with machine learning. BMC medical genomics. 2014;7(1). Doi:10.1186/1755-8794-7-s1-s7.
6. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. Journal of the American Medical Informatics Association. 2012;19(e1):e162–e169. Doi:10.1136/amiajnl-2011-000583.
7. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. Journal of biomedical informatics. 2010;43(6):891–901. Doi:10.1016/j.jbi.2010.09.009.
8. Doddi S, Marathe A, Ravi SS, T DC. Discovery of association rules in medical data. Informatics for Health and Social Care. 2001;26(1):25–33. Doi:10.1080/14639230117529.
9. Stilou S, Bamidis P, Maglaveras N, Pappas C. Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. Studies in health technology and informatics. 2001;(2):1399–1403.
10. Dligach D, Bethard S, Becker L, Miller TA, Savova GK. Discovering body site and severity modifiers in clinical texts. Journal of the American Medical Informatics Association (JAMIA). 2014;p. 448–454. Doi:10.1136/amiajnl-2013-001766.
11. Chikka VR, Mariyasagayam N, Niwa Y, Karlapalem K. Information Extraction from Clinical Documents: Towards Disease/Disorder Template Filling. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer; 2015. p. 389–401. Doi:10.1007/978-3-319-24027-5_41.
12. Баранов АА, Намазова-Баранова ЛС, Смирнов ИВ, Девяткин ДА, Шелманов АО, Вишнева ЕА, et al. Методы и средства комплексного интеллектуального анализа медицинских данных. Труды ИСА РАН. 2015;65(2):81–93.
13. Gudgin M, Hadley M, Mendelsohn N, Moreau JJ, Nielsen HF, Karmarkar A, et al. Soap version 1.2 part 1: Messaging framework. W3C Working Draft, DevelopMentor, Sun, IBM, Canon, Microsoft. 2002.
14. Shelmanov AO, Smirnov IV. Methods for semantic role labeling of Russian texts. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» (2014). 13; 2014. p. 607–620.
15. Osipov G, Smirnov I, Tikhomirov I, Shelmanov A. Relational-situational method for intelligent search and analysis of scientific publications. In: Proceedings of the Integrating IR Technologies for Professional Search Workshop; 2013. p. 57–64. Doi:10.3103/s0147688210060080.
16. Shelmanov AO, Smirnov IV, Vishneva EA. Information Extraction from Clinical Texts in Russian. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» (2015). 13; 2015. p. 560–572.
17. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010;17(3):229–236. Doi:10.1136/jamia.2009.002733.
18. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bulletin of the Medical Library Association. 1993;81(2).
19. 2014AA UMLS MeSH Russian Source Information URL: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHRUS/; 2015.
20. Государственный реестр лекарственных средств (ГРЛС) URL: http://grls.rosminzdrav.ru/Default.aspx; 2015.
21. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984. Doi:10.2307/2530946.
22. Breiman L. Random forests. Machine learning. 2001;45(1):5–32. Doi:10.1023/A:1010933404324.
23. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001;p. 1189–1232. Doi:10.1214/aos/1013203451.
24. Breiman L. Technical note: Some properties of splitting criteria. Machine Learning. 1996;24(1):41–47. Doi:10.1007/bf00117831.
25. Agrawal R, Inski T, Swami A. Mining association rules between sets of items in large databases. În: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. vol. 22. ACM; 1993. p. 207–216. Doi:10.1145/170036.170072.
26. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of 20th International Conference on Very Large Data Bases. vol. 1215; 1994. p. 487–499.
27. Vapnik V. The nature of statistical learning theory. Springer Science & Business Media; 1998.
28. Воронцов КВ. Комбинаторный подход к оценке качества обучаемых алгоритмов. Математические вопросы кибернетики. 2004;13:5–36.
29. Kelly L, Goeuriot L, Suominen H, Schreck T, Leroy G, Mowery DL, et al. Overview of the SHARE/CLEF eHealth evaluation lab 2014. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction. Springer; 2014. p. 172–191. Doi:10.1007/978-3-319-11382-1_17.
30. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2011;2(1):37–63.

31. Freund Y, Schapire R, Abe N. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence. 1999;14(771-780):1612.
32. Corcho O, Alper P, Kotsiopoulos I, Missier P, Bechhofer S, Goble C. An overview of S-OGSA: A reference semantic grid architecture. Web Semantics: Science, Services and Agents on the World Wide Web. 2006;4(2):102–115. Doi:10.1016/j.websem.2006.03.001.

КОНТАКТНАЯ ИНФОРМАЦИЯ

Баранов Александр Александрович, доктор медицинских наук, профессор, академик РАН, директор Федерального государственного бюджетного учреждения «Научный центр здоровья детей» Министерства здравоохранения Российской Федерации
Адрес: 119991, Москва, Ломоносовский проспект, д. 2, стр. 1, **тел**.: +7 (499) 134-30-83, **e-mail**: baranov@nczd.ru

Намазова-Баранова Лейла Сеймуровна, доктор медицинских наук, профессор, член-корреспондент РАН, заместитель директора по научной работе, директор НИИ педиатрии ФГБУ «НЦЗД» Минздрава России
Адрес: 119991, Москва, Ломоносовский проспект, д. 2, стр. 1, **тел**.: +7 (499) 967-14-14, **e-mail**: namazova@nczd.ru

Смирнов Иван Валентинович, кандидат физико-математических наук, доцент, заведующий лабораторией «Компьютерная лингвистика и интеллектуальный анализ информации» Института системного анализа Федерального исследовательского центра «Информатика и управление» Российской академии наук (ИСА ФИЦ ИУ РАН)
Адрес: 117312, Москва, проспект 60-летия Октября, д. 9, **тел**.: +7 (499) 135-90-20, **e-mail**: ivs@isa.ru

Девяткин Дмитрий Алексеевич, младший научный сотрудник лаборатории «Интеллектуальные технологии и системы» ИСА ФИЦ ИУ РАН
Адрес: 117312, Москва, проспект 60-летия Октября, д. 9, **тел**.: +7 (499) 135-90-20, **e-mail**: devyatkin@isa.ru

Шелманов Артём Олегович, кандидат технических наук, младший научный сотрудник лаборатории «Компьютерная лингвистика и интеллектуальный анализ информации» ИСА ФИЦ ИУ РАН
Адрес: 117312, Москва, проспект 60-летия Октября, д. 9, **тел**.: +7 (499) 135-90-20, **e-mail**: shelmanov@isa.ru

Вишнёва Елена Александровна, кандидат медицинских наук, заведующая отделом стандартизации и клинической фармакологии ФГБУ «НЦЗД» Минздрава России
Адрес: 119991, Москва, Ломоносовский проспект, д. 2, стр. 1, **тел**.: +7 (495) 967-14-65, **e-mail**: vishneva@nczd.ru

Антонова Елена Вадимовна, доктор медицинских наук, заведующая отделом прогнозирования и планирования научных исследований ФГБУ «НЦЗД» Минздрава России
Адрес: 119991, Москва, Ломоносовский проспект, д. 2, стр. 1, **тел**.: +7 (495) 967-15-66, **e-mail**: antonova@nczd.ru

Смирнов Владимир Иванович, кандидат экономических наук, заместитель директора по информационным технологиям ФГБУ «НЦЗД» Минздрава России
Адрес: 119991, Москва, Ломоносовский проспект, д. 2, стр. 1, **тел**.: +7 (495) 967-15-66, **e-mail**: support@nczd.ru