ИНФОРМАЦИОННЫЕ СИСТЕМЫ В БИОМЕДИЦИНСКИХ СИСТЕМАХ

INFORMATION SYSTEMS IN BIOMEDICAL SYSTEMS

УДК 004.89 DOI: 10.17587/it.25.97-106

В. В. Грибова, д-р техн. наук, зам. директора по научной работе, зав. лаб., gribova@iacp.dvo.ru,
Ф. М. Москаленко, канд. техн. наук, ст. науч. сотр., philipmm@iacp.dvo.ru,
Институт автоматики и процессов управления ДВО РАН, г. Владивосток,
К. И. Шахгельдян, д-р техн. наук,
директор института информационных технологий, carinash@vvsu.ru,
Д. В. Гмарь,
руководитель Центра информационно-технического обеспечения, dmitriy.gmar@vvsu.ru,
Владивостокский государственный университет экономики и сервиса, г. Владивосток,
Б. И. Гельцер, д-р мед. наук,
директор департамента клинической медицины Школы биомедицины, boris.geltser@vvsu.ru,
Дальневосточный федеральный университет, г. Владивосток

Концепция гетерогенного хранилища биомедицинской информации[1]

Предложена концепция хранилища биомедицинских данных и знаний. Архитектурной особенностью хранилища является интеграция первичных клинических и статистических данных, результатов их обработки и анализа, объединение последних с известными медицинскими знаниями по диагностике и лечению заболеваний и использование их для развития персонализированных трендов в клинической медицине. Хранилище имеет распределенную архитектуру и объединяет информационные и программные компоненты.

Ключевые слова: *гетерогенное хранилище, биомедицинская информация, интеллектуальные информационные системы, методы анализа данных, онтологии, базы знаний*

Введение

В последние годы существенно увеличиваются объемы информации в различных сегментах системы здравоохранения и клинической медицины в связи с накоплением данных из электронных историй болезни, результатов лабораторных и инструментальных исследований, мобильных устройств для мониторинга физиологических функций человека, экономических показателей деятельности медицинских организаций и др. [1]. Естественным следствием этих процессов является возможность дальнейшего использования этой информации в клинических и популяционных исследованиях, при построении интеллектуальных систем поддержки принятия решений для организаторов здравоохранения и клиницистов. Необходимость таких исследований обусловлена развитием персонифицированной медицины, где роль интеллектуальных средств поддержки лечебно-диагностической деятельности многократно возрастает. Вместе с тем существует ряд проблем, затрудняющих повторное использование биомедицинских данных в целях извлечения из них новых знаний и построения интеллектуальных систем для профилактики, диагностики и лечения заболеваний. Первая из них связана с хранением первичных данных в разрозненных источниках: медицинских и лабораторных информационных системах (МИС и ЛИС), медицинской статистической отчетности, отчетах Росстата, Роспотребнадзора и др. При этом отсутствуют единые стандарты хранения информации, а первичные данные обычно представлены в таблицах Excel, в файлах формата doc или pdf в свободной, часто меняющейся

[1] Исследование выполнено при частичной финансовой поддержке РФФИ в рамках научных проектов № 18-29-03131, 17-07-00956.

форме. Вторая проблема связана с тем, что вторичные данные, представленные в виде результатов обработки, также не имеют стандартов или общепринятых форм хранения. Третья проблема заключается в том, что биомедицинские данные, как правило, слабо формализованы и плохо структурированы. Кроме того, полученные в результате анализа и обработки первичных данных новые медицинские знания не используются интеллектуальными информационными системами, обеспечивающими поддержку принятия врачебных решений. Решение указанных проблем возможно только на основе создания новой технологии, включающей модели, методы и программные средства, обеспечивающие хранение и обработку гетерогенных первичных и вторичных биомедицинских данных. Использование такой технологии позволит, с одной стороны, проводить научные исследования и получать новые знания в области организации здравоохранения, профилактической и клинической медицины, а с другой — разрабатывать интеллектуальные информационные системы, повышающие эффективность диагностики, лечения и прогнозирования исходов заболеваний.

Целью исследования является разработка концепции гетерогенного хранилища (ГХ) биомедицинской информации.

1. Обзор существующих решений

Создание хранилищ данных широко практикуется уже много лет в различных областях знаний. Системы для хранения электронных медицинских записей стали активно развиваться только с начала 1990-х годов [2]. Для электронных историй болезни были предложены стандарты, например, такие как HL7, ГОСТ 52636—2006 [3, 4]. Вместе с тем создание хранилищ для здравоохранения долгие годы представляло серьезную проблему ввиду сложности и гетерогенности медицинских, в том числе клинических, данных [5]. Кроме того, возникла необходимость создания информационных систем, которые бы не ограничивались только ведением историй болезни и амбулаторных карт, но и обеспечивали поддержку принятия профессиональных врачебных решений, необходимых для эффективного лечения больных [2]. На необходимости интеграции, очистки и форматирования биомедицинских данных фиксировали внимание самые первые публикации по созданию гетерогенных хранилищ [6, 7]. Для интеграции клинических данных были разработаны несколько стандартов, относящихся не только к историям болезни, но и к медицинской терминологии (англоязычной): RxNorm, SNOMED-CT, ICD, LOINC, UMLS, DRG code [8, 9]. Понимание преимуществ совместного владения данными из различных организаций стало мотивом для создания федеральных информационных систем по отдельным заболеваниям [10, 11]. Некоторые базы клинических данных преобразовались в деперсонализированные хранилища, доступные для исследовательских и образовательных целей [12].

Начиная с конца 90-х годов появляются научные работы по применению методов Data mining в целях повышения эффективности управления здравоохранением и отдельными медицинскими организациями. Основными направлениями здесь являлись: оптимизация ресурсов, страховая медицина, оптимальные объемы коечного фонда, маршрутизация пациентов к специалистам различных профилей и др. [13, 14]. Наряду с решением экономических и управленческих задач тогда же появились исследования, в которых использовались факторный анализ, деревья принятия решений, векторные машины, байесовские сети, анализ Кокса и др., которые способствовали получению новых знаний, повышающих качество клинических решений [15—17]. В конце 90-х стало понятно, что хранилища позволяют не только повысить эффективность и снизить расходы на здравоохранение, но и улучшить качество медицинской помощи за счет извлечения знаний из данных и повторного использования баз знаний [18].

С начала 2000-х годов в тренде доказательной медицины активно развиваются системы поддержки принятия врачебных решений. Уже тогда отмечалась важность интеграции в них историй болезни и известных ранее знаний с новыми научными результатами, которые могли бы расширить такой репозиторий знаний [19]. В последнее время увеличилось число работ по оценке эффективности применения этих систем в клинической практике [20].

Анализ научной литературы показал, что существующие системы хранения клинических данных не обеспечивают решение вышеописанных проблем ГХ биомедицинской информации, что подчеркивает необходимость реализации новых подходов для их решения.

2. Компоненты гетерогенного хранилища

Общая классификация

ГХ состоит из **информационных и программных компонентов** (рис. 1). **Информационными компонентами** ГХ являются:

- **онтологии** — системы формализации с помощью концептуальных схем, в соответствии с которыми формируются другие типы информационных и программных компонентов;
- **данные —** совместно используемый набор формализованных или неформализованных (включая слабоформализованные) данных разных типов. Формализованные данные — логически связанные входные или выходные данные, организованные в соответствии с поддерживаемой моделью. Неформализованными (или слабоформализованными) данными являются текстовые документы, графики, диаграммы, графические изображения и др., которые являются либо результатами работы программных систем и сервисов, требующими обработки, либо файлами с документацией;
- **знания** — формально представленные зависимости, причинно-следственные связи между данными, предназначенные для решения задач в практической медицине и образовании.

Программными компонентами являются: исходные коды программ, программные агенты, инструментальные оболочки, прикладные программные системы и сервисы. Они делятся на:

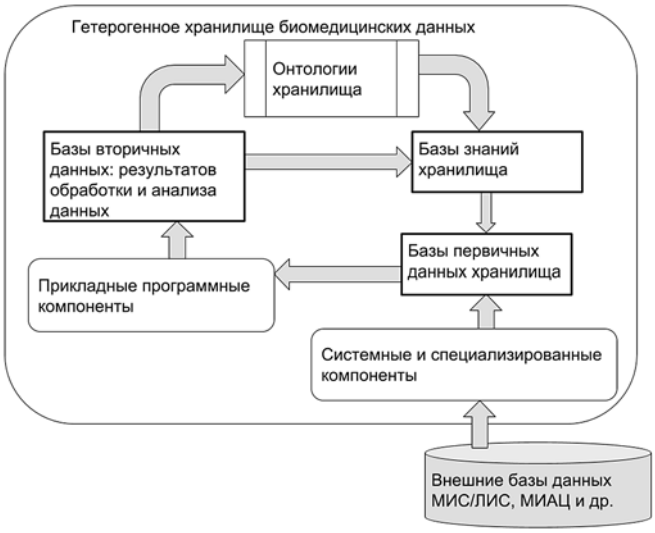

Рис. 1. Концептуальная схема ГХ

- **системные программные компоненты,** которые предназначены для организации эффективного функционирования ГХ, а их примерами являются сервис авторизации и управления правами пользователей, генератор агентов, генератор редакторов информационных ресурсов, сервис импорта и экспорта данных в ГХ и др.;
- **специализированные программные компоненты,** которые предлагают пользователю ряд сервисных функций, необходимых, прежде всего, для подготовки данных к обработке. Примерами таких компонентов являются сервисы преобразования информации из одного формата в другой, сервис формализации текста из неформализованного вида и др. К специализированным программным компонентам также относятся программные оболочки, предназначенные для создания на их основе программных систем и сервисов, ориентированных на заданный в оболочке класс задач, например, специализированная оболочка для создания интеллектуальных систем по диагностике заболеваний;
- **прикладные программные компоненты,** которые предназначены для решения научно-исследовательских и практически-значимых задач в медицине, включая обучающие системы. К прикладным программным компонентам относятся информационно-аналитические системы, системы мониторинга и прогноза, поддержки принятия врачебных решений, система поддержки проведения научных исследований, компьютерные тренажеры для отработки моторных навыков и знаний студентов.

Онтологии

Для описания используемых в хранилище структурированных информационных компонентов (данных и знаний), программных компонентов (сервисов, агентов, шаблонов сообщений), а также структуры хранилища и связей между его логическими элементами используются *онтологии*. В настоящее время использование онтологического подхода положено в основу создания многих современных компьютерных систем и тематических порталов знаний для упрощения навигации, формирования и сопровождения сложно-структурированной информации различных уровней абстракции в привычной для ее носителей си-

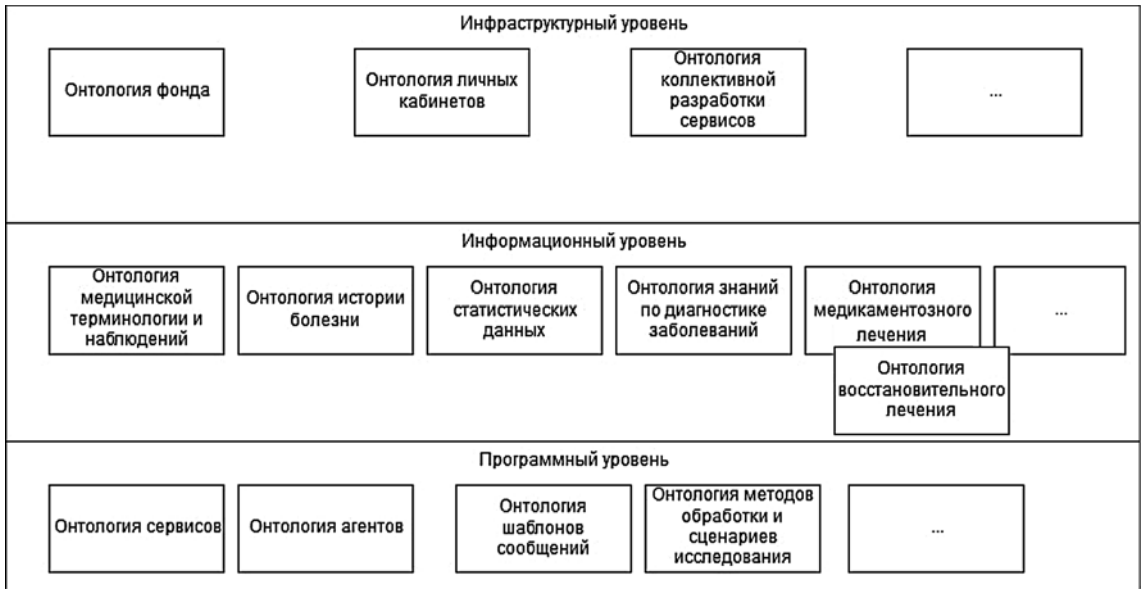

Рис. 2. Онтологии хранилища биомедицинских данных

стеме понятий — без специального обучения последних или участия в этом процессе специалистов-посредников.

Онтологии ГХ условно можно разделить на три логических уровня (рис. 2). Первый из них является *инфраструктурным*. Данный уровень включает онтологии, разработанные для организации и поддержки инфраструктуры ГХ. К ним можно отнести онтологию фонда программных и информационных ресурсов, личных кабинетов пользователей, коллективной разработки и др. [21].

Второй логический уровень — *информационный*, содержит онтологии, поддерживающие формирование информационных компонентов ГХ — баз данных и баз знаний, предназначенных для решения научных и практических задач в медицине, а также обучения студентов и повышения квалификации врачей. Среди таких онтологий, прежде всего, необходимо выделить: онтологию медицинской терминологии и наблюдений, которая задает унифицированную структуру описания всех медицинских терминов и наблюдений, онтологии истории болезни, статистических данных, лекарственных препаратов, научных исследований и другие онтологии данных, которые к настоящему времени уже созданы и находятся в ГХ, а также те, которые будут создаваться пользователями ГХ при постановке новых научно-практических задач. К данному уровню также относятся онтологии знаний, которые, прежде всего, необходимы для создания интеллектуальных систем поддержки принятия врачебных решений: онтология диагностики острых и хронических заболеваний [22, 23], онтология назначения лечения — медикаментозного и восстановительного [24, 25]. Здесь же располагаются онтологии, описывающие экспериментальные медицинские знания, полученные в результате проведения научных исследований на основе данных ГХ.

Третий логический уровень — *программный*. Данный уровень содержит множество онтологий, описывающих структуры программных сервисов и их компонентов: онтологии сервисов, агентов, шаблонов сообщений, онтология методов обработки, сценариев научных исследований и др.

Данные

Данные ГХ делятся на *первичные*, полученные из различных внешних источников или сгенерированные средствами программных компонентов ГХ в целях их дальнейшей обработки, и *вторичные*, являющиеся результатом обработки, например, проведенного научного исследования в соответствии с некоторым сценарием, также представляющим собой вторичные данные.

Первичные данные могут быть строго-, слабо- и не формализованными. Все данные, введенные из внешних источников, требуют предварительной обработки для формализации, верификации и преобразования в соответствии со структурой терминов. ГХ поддерживает следующие типы данных:

1) истории болезни, внесенные в ГХ через интерфейс программных компонентов, извлеченные из МИС, либо представленные в формате текстовых файлов, подготовленных врачами в медицинских организациях, где отсутствуют МИС;

2) формализованные данные медицинской организации, характеризующие ее процессы и эффективность работы;

3) данные медицинской статистики, описывающие региональные и федеральные системы здравоохранения;

4) данные экологического, социально-экономического и других видов мониторинга.

Реализация алгоритмов диагностики заболеваний и поиска аналогичных случаев, определение и оценка корректности процедур лечения, проведение исследований методами машинного обучения в целях извлечения новых медицинских знаний возможны только на основе строго формализованных данных. Такие данные вводятся в ГХ через предоставляемый им программный и пользовательский интерфейс. Для повышения качества исследований и принятия решений необходимо проводить процедуры верификации строго формализованных данных, которые обеспечивают взаимосвязь историй болезни одного и того же пациента в течение жизни вплоть до смерти.

Различные МИС содержат в большинстве случаев слабоформализованные данные, хотя уровень формализации существенно варьируется: от 10 до 90 % объема информации истории болезни. Интерес для исследователей представляют и неформализованные истории болезни, например, формата doc, которых еще достаточно много в медицинских организациях России. Эти два типа данных должны быть подвергнуты специализированным процедурам формализации. Результаты лабораторных и инструментальных исследований фактически являются частью историй болезни и представляют собой строго формализованные лабораторные данные, изображения или различные виды сигналов. Информация о процессах медицинской организации обычно может быть получена из МИС и может использоваться для оценки эффективности управления и деятельности таких организаций.

Данные медицинской статистики описывают региональную или федеральную системы здравоохранения и извлекаются из систем медицинской статистики в медицинских информационно-аналитических центрах региона. Несмотря на то что данные представлены в табличной форме, они требуют процедур предварительной обработки. Сложность вызывает тот факт, что формы статистической отчетности меняются ежегодно как по объектам учета, так и по их атрибутике и типам. Кроме того, возникают проблемы, связанные с институциональными изменениями системы регионального здравоохранения, которые необходимо учитывать при анализе временных рядов.

Данные экологического мониторинга (вода/воздух/почва) могут быть получены из систем Центров гигиены и эпидемиологии, но они также требуют предварительной обработки. Данные социально-экономического мониторинга могут быть получены из открытых источников в виде таблиц, они характеризуют социально-экономический статус различных регионов и муниципальных образований. Для извлечения этих данных требуются специализированные программные компоненты предварительной обработки.

К первичным данным относится и *справочник медицинской терминологии и наблюдений*. Этот базовый ресурс ГХ содержит формальное описание медицинских терминов, относящихся к определенным группам: морфология, физиология, патология, фармакология и др.; наблюдения, содержащие группы признаков (жалобы, данные объективного исследования, данные лабораторных и инструментальных исследований), события (лечебные, диагностические мероприятия, осложнения и др.), факторы (профессиональные вредности, вредные привычки и др.), справочник заболеваний МКБ (международной классификации болезней). В настоящее время справочник медицинской терминологии и наблюдений включает около 24 000 различных понятий и активно развивается всеми пользователями ГХ.

Вторичные данные представляют собой, во-первых, результаты обработки первичных данных методами статистического анализа, машинного обучения и искусственного интеллекта. Во-вторых, к вторичным данным относятся описания проведенных научных исследований, включая дизайн проекта, условия выборки первичных данных, последовательность применения методов и др. Научные исследования проводятся в соответствии с некоторым сценарием, в который включаются условия выборки первичных данных, методы обработки и входные параметры, результаты обработки, их интерпретация, скрипты обра-

ботки (например, на языках R, Python и др.). Важной целью сохранения в ГХ информации об исследовании является воспроизводимость (повторяемость) научных результатов, сравнение их между собой при изменении условий выборки, расширении объема первичных данных, условий применения методов обработки и др. К вторичным данным относят также все случаи применения интеллектуальных сервисов профилактики, диагностики и лечения заболеваний с протоколированием предложенной врачебной помощи и результатами применения. Эти данные необходимы для постоянного совершенствования полученных знаний, разрабатываемых алгоритмов и интеллектуальных систем.

Базы знаний

К настоящему времени в медицине накоплены уникальные знания по диагностике, лечению и профилактике заболеваний, которые представлены в медицинской литературе, научных отчетах и многочисленных статьях. Однако такие знания в неформализованном виде используются только специалистами в своей узкой области. Для повышения эффективности врачебных решений необходимо расширение применения знаний, а для этого они должны быть доступны для обработки программными системами, которые помогут врачам принять обоснованные клинические решения. С этой целью знания формализуют и формируют базы знаний, структура которых представлена в онтологиях. Основными базами знания ГХ являются: базы знаний по диагностике острых, хронических заболеваний и синдромов, базы знаний по медикаментозному и восстановительному лечению, база знаний по фармакотерапии, базы экспериментальных знаний, полученных в результате научных исследований с данными ГХ и др.

Базы знаний по диагностике острых, хронических заболеваний и синдромов. Эти базы знаний включают: комплекс диагностических признаков и вариативность их динамики; специфичность симптомов заболеваний; средства уточнения диагнозов с учетом этиологии, патогенеза, варианта течения; определение необходимых условий для развития заболеваний и др. Базы знаний групп заболеваний, доступные в ГХ в настоящее время, представлены на рис. 3 (см. вторую сторону обложки), а на рис. 4 (см. вторую сторону обложки) представлен фрагмент базы знаний по диагностике геморрагической лихорадки с почечным синдромом (ГЛПС).

База знаний по назначению лечения (медикаментозного и восстановительного). В ГХ содержатся базы знаний по назначению медикаментозного и восстановительного лечения, включающие описание группы заболеваний с общими принципами лечения, как правило, симптоматического и/или патогенетического, в том числе модель, вид, цель и схему терапии. Каждый раздел базы знаний описывается набором условий, представляющих собой клиническое наблюдение, относящееся или к персональным данным пациента, или к клинической картине заболевания, результатам лабораторных и инструментальных исследований. Базы знаний по восстановительному лечению описывают возможные варианты использования методов восстановительного лечения заболеваний, учитывающие назначенную медикаментозную терапию, персональные особенности пациента, индивидуальность клинической картины заболевания.

База знаний фармакотерапии. База фармакотерапии включает описание действующих веществ с их фармакологическими группами, свойствами, заболеваниями. Ее особенностью и принципиальным отличием от аналогов является формальное описание таких значимых для систем поддержки принятия решений атрибутов, как противопоказания, побочные действия, передозировка, взаимодействие с другими лекарственными средствами. Данные атрибуты описываются как множество значений признаков, событий и факторов, связанных логическими условиями, каждый из которых имеет структуру, определенную в базе терминологии и наблюдений.

База экспериментальных медицинских знаний. База представляет собой новые медицинские знания, полученные в результате обработки и анализа первичных клинических и статистических данных ГХ методами искусственного интеллекта. Знания имеют различные формы, например, линейной, логистической, мультиномиальной, порядковой регрессий, искусственной нейронной сети и др. Эти модели вместе со своими весовыми коэффициентами, предикторами, свободными членами и архитектурой описывают физиологические процессы, определяют прогноз течения и степень тяжести заболеваний, вероятность осложнений, длительность жизни после опе-

рации и др. [26, 27]. Полученные знания должны заново верифицироваться при расширении базы первичных данных. Они используются наравне с ранее введенными в ГХ знаниями из внешних источников: стандартами лечения, правилами диагностики заболевания, назначения лечения и др. Компоненты, которые реализуют интеллектуальные сервисы по профилактике, диагностике и лечению, обеспечивая поддержку врачебных решений, используют эти знания и фиксируют все случаи их применения. Эта информация в дальнейшем учитывается для уточнения медицинских знаний.

Программные компоненты

Программные компоненты ГХ — это, прежде всего, **прикладные программные сервисы интеллектуальной обработки данных и знаний** для решения научных задач, а также задач практической медицины и образования. Для их создания используются фреймворки и программные средства общего назначения, например, платформа IACPaaS [28], предназначенная для создания интеллектуальных сервисов в произвольных предметных областях, платформы с библиотеками, поддерживающие обработку больших массивов данных, например, R-studio (с языком R), Anaconda (с языком Python), а также специализированные оболочки и системы, ориентированные на создание интеллектуальных систем и сервисов, решающих задачи определенных классов, например, диагностики процессов (заболеваний).

Системная программная компонента импорта и экспорта данных в/из ГХ. Импорт клинических данных выполняется в соответствии с онтологиями истории болезни с использованием базы данных медицинской терминологии и наблюдений. Импортировать можно только строго формализованные данные.

Специализированная программная компонента предварительной обработки клинических данных. Предварительная обработка клинических данных включает: структурирование историй болезни и формализацию ее разделов. Структурирование историй болезни выполняется в соответствии с онтологией истории болезни, когда из неструктурированных файлов, обычно формата doc, выделяются разделы (жалобы, дневники, выписка, эпикриз и др.) и фиксируется текст внутри каждого раздела. Формализация истории болезни подразумевает превращение текста внутри разделов в числовые, строковые и категориальные переменные. По определенным правилам, соответствующим разделам историй болезни, формируются грамматики, которые с помощью средств обработки текста, например Томита-парсер, обеспечивают формализацию текста. В качестве словарей используются база данных медицинских терминов и наблюдений и база знаний диагностики и лечения.

Специализированная программная компонента предварительной обработки статистических данных. Компонента обеспечивает решение проблем данных, связанных с ежегодным изменением форм статистической отчетности, изменением объектов учета, их атрибутов, и позволяет формировать временные ряды, которые пригодны для соответствующего анализа [29]. Также эта компонента позволяет приводить к табличному виду статистические данные, собранные из открытых источников (РОССТАТ, НИИ Информатизации здравоохранения и др.).

Специализированная оболочка для создания систем поддержки принятия решений для практической медицины [30, 31]. Она позволяет создавать системы поддержки принятия клинических решений по диагностике и лечению заболеваний в различных разделах медицины. Она использует, помимо информационных компонентов ГХ, решатели задач по диагностике и назначению лечения. Расширение областей применения оболочки связано с формированием новых разделов медицины в базах данных и знаний ГХ. Оболочка предоставляется пользователям как облачный сервис.

Заключение

Быстрый рост объемов биомедицинской информации позволяет надеяться на развитие тренда персонализированной медицины. Одним из серьезных препятствий к исследованиям в этом направлении является отсутствие технологии хранения, обработки, анализа и верификации такой информации. В работе предложена концепция хранилища биомедицинских данных и знаний, объединяющего данные из разных источников, разной природы. Архитектурной особенностью хранилища является интеграция первичных клинических и статистических данных, результатов их обработки и анализа, объединение последних с известными медицинскими знаниями по ди-

агностике и лечению заболеваний и использование их для развития персонализированных трендов в клинической медицине. ГХ имеет компонентную распределенную архитектуру и объединяет информационные и программные компоненты. К первым относятся: данные, онтологии и базы знаний, вторые делятся на: системные, специализированные и прикладные.

К настоящему времени разработаны и используются многие программные и информационные компоненты ГХ, продолжаются работы по созданию новых компонентов ГХ и усовершенствованию имеющихся.

Список литературы

1. **Andreu-Perez J., Poon C. C. Y., Merrifield R. D., Wong S. T., Yang G. Z.** Big Data for Health // IEEE Journal of biomedical and health informatics. 2015. Vol. 19, N. 4. P. 1193—1208.

2. **Evans R. S., Lloyd J. F., Pierce L. A.** Clinical Use of an Enterprise Data Warehouse // AMIA Annu Symp. Proc. 2012. P. 189—198.

3. **HL7 Standards** URL: http://www.hl7.org/implement/standards/index.cfm?ref=nav (дата обращения: 10.11.2018).

4. **Национальный** стандарт РФ ГОСТ Р 52636—2006 "Электронная история болезни. Общие положения". URL: http://dokipedia.ru/document/5324420 (дата обращения: 12.11.2018).

5. **Chute C. G., Beck S. A., Fisk T. B., Mohr D. N.** The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data // Journal of the American Medical Informatics Association. 2010. Vol. 17, N. 2. P. 131—135.

6. **Verma R., Harper J.** Life cycle of a data warehousing project in healthcare // Journal of healthcare information management. 2001. Vol. 15, N. 2. P. 107—117.

7. **Kerkri E. M., Quantin C., Allaert F. A., Cottin Y., Charve P., Jouanot F., Yetongnon K.** An approach for integrating heterogeneous information sources in a medical data warehouse // Journal of Medical Systems. 2001. Vol. 25, N. 3. P. 167—176.

8. **Kamal J., Liu J., Ostrander M., Santangelo J., Dyta R., Rogers P., Mekhjian H. S.** Information warehouse — a comprehensive informatics platform for business, clinical, and research applications // AMIA Annual Symposium Proceedings. 2010. N. 13. P. 452—456.

9. **Grant A., Moshyk A., Diab H., Carona P., Fabien de Lorenzi, Bissona G., Menarda L., Lefebvrea R., Gauthier P., Grondinb R., Desautels M.** Integrating feedback from a clinical data warehouse into practice organisation // International Journal of Medical Informatics. 2006. Vol. 75, N. 3—4. P. 232—239.

10. **Cao X., Wong S. T., Kent S., Tjandra D. C., Lowenstein D. H.** A web-based federated neuroinformatics model for surgical planning and clinical research applications in epilepsy // Neuroinformatics. 2004. Vol. 2, N. 1. P. 101—117.

11. **Livne O. E., Schultz N. D., Narus S. P.** Federated Querying Architecture with Clinical & Translational Health IT Application // Journal of medical systems. 2011. Vol. 35, N. 5. P. 1211—1224.

12. **Erdal B. S., Liu J., Ding J., Chen J., Marsh C. B., Kamal J., Clymer B. D.** A Database De-identification Framework to Enable Direct Queries on Medical Data for Secondary Use // Methods of information in medicine. 2012. Vol. 51, N. 3. P. 229—241.

13. **Silver M., Sakata T., Su H. C., Herman C., Dolins S. B., O'Shea M. J.** Case study: how to apply data mining techniques in a healthcare data warehouse // Journal of healthcare information management. 2001. Vol. 15, N. 2. P. 155—164.

14. **Chen Y., Matsumura Y., Nakagawa K., Ji S., Nakano H., Teratani T., Zhang Q., Mineno T., Takeda H.** Analysis of yearly variations in drug expenditure for one patient using data warehouse in a hospital // Journal of medical systems. 2007. Vol. 31, N. 1. P. 17—24.

15. **Zhou X., Chen S., Liu B., Zhang R., Wang Y., Li P., Guo Y., Zhang H., Gao Z., Yan X.** Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support // Artificial Intelligence in Medicine. 2010. V. 48, N. 2—3. P. 139—152.

16. **Cho I. S., Haug P. J.** The contribution of nursing data to the development of a predictive model for the detection of acute pancreatitis // Stud Health Technol Inform. 2006. N. 122. P. 139—142.

17. **Botsis T., Anagnostou V. K., Hartvigsen G., Hripcsak G., Weng C.** Developing a multivariable prognostic model for pancreatic endocrine tumors using the clinical data warehouse resources of a single institution // Applied Clinical Informatics. 2010. Vol. 1, N. 1. P. 38—49.

18. **Hollis J.** Deploying an HMO's data warehouse // Health management technology. 1998. V. 19, N. 8. P. 46—48.

19. **Sim I., Gorman P., Greenes R. A., Haynes R. B., Kaplan B., Lehmann H., Tang P. C.** Clinical decision support systems for the practice of evidence-based medicine // Journal of the American Medical Informatics Association. 2001. Vol. 8, N. 6. P. 527—534.

20. **Moja L., Friz H. P., Capobussi M., Kwag K., Banzi R., Ruggiero F., González-Lorenzo M., Liberati E. G., Mangia M., Nyberg P., Kunnamo I., Cimminiello C., Vighi G., Grimshaw J., Bonovas S.** Implementing an evidence-based computerized decision support system to improve patient care in a general hospital: the CODES study protocol for a randomized controlled trial // Implementation Science. 2016. N. 11:89. P. 1—10.

21. **Грибова В. В., Клещев А. С., Москаленко Ф. М., Тимченко В. А., Федорищев Л. А., Шалфеева Е. А.** Управляемая графовыми грамматиками разработка оболочек интеллектуальных сервисов на облачной платформе IACPaaS // Программная инженерия. 2017. Т. 8, № 10. С. 435—447.

22. **Gribova V., Okun D., Petryaeva M., Shalfeeva E., Tarasov A.** Ontology for Differential Diagnosis of Acute and Chronic Diseases // Communications in Computer and Information Science. 2018. Vol. 934. P. 152—163.

23. **Грибова В. В., Петряева М. В., Окунь Д. Б., Шалфеева Е. А.** Онтология медицинской диагностики для интеллектуальных систем поддержки принятия решений // Онтология проектирования. 2018. Т. 8, № 1 (27). С. 58—73.

24. **Грибова В. В., Окунь Д. Б.** Онтологии для формирования баз знаний и реализации лечебных мероприятий в медицинских интеллектуальных системах // Информатика и системы управления. 2018. № 3 (57). С. 71—80.

25. **Грибова В. В., Окунь Д. Б.** Онтология база знаний восстановительного лечения // Системный анализ в медицине (САМ 2018): Материалы XII международной научной конференции под общ. ред. В. П. Колосова. (Благовещенск, 2018 г.). 2018. С. 47—50.

26. **Гельцер Б. И., Скляр Л. Ф., Елисеева В. С., Шахгельдян К. И., Маркелова Е. В., Емцева Е. Д., Бениоваи С. Н.** Оценка иммунологических показателей при лекарственной устойчивости ВИЧ на фоне ВААРТ // ВИЧ-инфекция и иммуносупрессии. 2018. Т. 10, № 1. С. 54—62.

27. **Гельцер Б. И., Шахгельдян Б. И., Курпатов И. А., Котельников В. Н.** Результаты моделирования должных величин силы дыхательных мышц на основе методов искусственного интеллекта // Российский физиологи-

ческий журнал им. И. М. Сеченова. 2018. Т. 104, № 9. С. 1065—1074.

28. **Грибова В. В., Клещев А. С., Москаленко Ф. М., Тимченко В. А., Федорищев Л. А., Шалфеева Е. А.** Облачная платформа IACPaaS для разработки оболочек интеллектуальных сервисов: состояние и перспективы развития // Программные продукты и системы. 2018. Т. 31, № 3. С. 527—536.

29. **Шахгельдян К. И., Гельцер Б. И., Гмарь Д. В., Кривелевич Е. Б., Теук К. А., Транковская Л. В.** Проблемы анализа данных медицинской статистики // Проблемы социальной гигиены, здравоохранения и истории медицины. 2018. Т. 26, № 3. С. 132—136.

30. **Gribova V. V., Petryaeva M. V., Okun D. B., Tarasov A. V.** Software Toolkit for Creating Intelligent Systems in Practical and Educational Medicine // 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). (Vladivostok, 18—25 Aug. 2018). IEEE Xplore, 2018. P. 1—5.

31. **Грибова В. В., Москаленко Ф. М., Окунь Д. Б., Петряева М. В.** Облачная среда для поддержки клинической медицины и образования // Врач и информационные технологии. 2016. № 1. С. 60—66.

V. V. Gribova, D. Sc. (Technical Sciences), Deputy Director,
Head of intelligent systems lab, gribova@iacp.dvo.ru,
Ph. M. Moskalenko, Ph. D. (Technical Sciences),
Senior Researcher in intelligent systems lab, philipmm@iacp.dvo.ru,
Institute of Automation and Control Processes FEB RAS, Vladivostok,
C. I. Shahgeldyan, D. Sc. (Technical Sciences),
Head of information technologies institute, carinash@vvsu.ru,
D. V. Gmar', Head of information technology center, dmitriy.gmar@vvsu.ru,
Vladivostok State University of Economics and Service, Vladivostok,
B. I. Geltser, D. Sc. (Medical Sciences),
Head of clinical medicine department of Biomedicine school, boris.geltser@vvsu.ru,
Far-Eastern State University, Vladivostok

A Concept for a Heterogeneous Biomedical Information Warehouse

One of the major obstacles of expanding research in the field of personified medicine is the lack of technology for storing, processing, analyzing and verifying of biomedical information. The paper proposed the concept of biomedical data and knowledge warehouse, which combines information from different sources of various nature. The architectural feature is the integration of primary clinical and statistical data, results of their processing and analysis, combination of the latter with known medical knowledge on diagnostics and disease treatment and their use in clinical medicine. Heterogeneous warehouse has a component distributed architecture and combines information (ontologies, data and knowledge bases) and software (system, specialized and applied) units. Ontologies are divided into three layers: infrastructure (meant for infrastructure support of the warehouse), information (support formation of data and knowledge components) and software (these ontologies set structures for software components). Data is divided into primary (received form outer sources — case records, formalized data of the medical organizations, medical statistics, monitoring logs) and secondary (acquired while data analysis by statistical methods, machine learning and artificial intelligence). Knowledge bases describe diagnostics, cures and pharmacotherapy, treatment assignment. Several information resources and software components are described (medical terminology and observations glossary, export/import tool, clinical data preprocessing tool, statistics data preprocessing tool, specialized shell for decision support system development). To date, many software and information components of the warehouse have been developed and are being used, work continues on the creation of new components and the improvement of existing ones.

Keywords: *heterogeneous repository of biomedical information, intelligent information systems, data analysis methods, ontologies, knowledge bases*

DOI: 10.17587/it.25.97-106

References

1. **Andreu-Perez J., Poon C. C. Y., Merrifield R. D., Wong S. T., Yang G. Z.** Big Data for Health, *IEEE Journal of biomedical and health informatics,* 2015, vol. 19, no. 4, pp. 1193—1208.

2. **Evans R. S., Lloyd J. F., Pierce L. A.** Clinical Use of an Enterprise Data Warehouse, *AMIA Annu Symp Proc.*, 2012, pp. 189—198.

3. **HL7** Standards, available at: http://dokipedia.ru/document/5324420 (date of access: 10.11.2018) (in Russian).

4. **Nacional'nyj** standart RF GOST R 52636—2006 "Jelektronnaja istorija bolezni. Obshhie polozhenija" (Digital case record. General information), available at: http://dokipedia.ru/document/5324420 (date of access: 12.11.2018) (in Russian).

5. **Chute C. G., Beck S. A., Fisk T. B., Mohr D. N.** The Enterprise Data Trust at Mayo Clinic: a semantically integrated

warehouse of biomedical data, *Journal of the American Medical Informatics Association*, 2010, vol. 17, no. 2, pp. 131—135.

6. **Verma R., Harper J.** Life cycle of a data warehousing project in healthcare, *Journal of healthcare information management*, 2001, vol. 15, no. 2, pp. 107—117.

7. **Kerkri E. M., Quantin C., Allaert F. A., Cottin Y., Charve P., Jouanot F., Yetongnon K.** An approach for integrating heterogeneous information sources in a medical data warehouse, *Journal of Medical Systems*, 2001, vol. 25, no. 3, pp. 167—176.

8. **Kamal J., Liu J., Ostrander M., Santangelo J., Dyta R., Rogers P., Mekhjian H. S.** Information warehouse — a comprehensive informatics platform for business, clinical, and research applications, *AMIA Annual Symposium Proceedings*, 2010, no. 13, pp. 452—456.

9. **Grant A., Moshyk A., Diab H., Carona P., Fabien de Lorenzi, Bissona G., Menarda L., Lefebvrea R., Gauthier P., Grondinb R., Desautels M.** Integrating feedback from a clinical data warehouse into practice organization, *International Journal of Medical Informatics*, 2006, vol. 75, no. 3—4, pp. 232—239.

10. **Cao X., Wong S. T., Kent S., Tjandra D. C., Lowenstein D. H.** A web-based federated neuroinformatics model for surgical planning and clinical research applications in epilepsy, *Neuroinformatics*, 2004, vol. 2, no. 1, pp. 101—117.

11. **Livne O. E., Schultz N. D., Narus S. P.** Federated Querying Architecture with Clinical & Translational Health IT Application, *Journal of medical systems*, 2011, vol. 35, no. 5, pp. 1211—1224.

12. **Erdal B. S., Liu J., Ding J., Chen J., Marsh C. B., Kamal J., Clymer B. D.** A Database De-identification Framework to Enable Direct Queries on Medical Data for Secondary Use, *Methods of information in medicine*, 2012, vol. 51, no. 3, pp. 229—241.

13. **Silver M., Sakata T., Su H. C., Herman C., Dolins S. B., O'Shea M. J.** Case study: how to apply data mining techniques in a healthcare data warehouse, *Journal of healthcare information management,* 2001, vol. 15, no. 2, pp. 155—164.

14. **Chen Y., Matsumura Y., Nakagawa K., Ji S., Nakano H., Teratani T., Zhang Q., Mineno T., Takeda H.** Analysis of yearly variations in drug expenditure for one patient using data warehouse in a hospital, *Journal of medical systems*, 2007, vol. 31, no. 1, pp. 17—24.

15. **Zhou X., Chen S., Liu B., Zhang R., Wang Y., Li P., Guo Y., Zhang H., Gao Z., Yan X.** Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support, *Artificial Intelligence in Medicine*, 2010, vol. 48, no. 2—3, pp. 139—152.

16. **Cho I. S., Haug P. J.** The contribution of nursing data to the development of a predictive model for the detection of acute pancreatitis*, Stud Health Technol Inform.*, 2006, no. 122, pp. 139—142.

17. **Botsis T., Anagnostou V. K., Hartvigsen G., Hripcsak G., Weng C.** Developing a multivariable prognostic model for pancreatic endocrine tumors using the clinical data warehouse resources of a single institution, *Applied Clinical Informatics,* 2010, vol. 1, no. 1, pp. 38—49.

18. **Hollis J.** Deploying an HMO's data warehouse, *Health management technology,* 1998, vol. 19, no. 8, pp. 46—48.

19. **Sim I., Gorman P., Greenes R. A., Haynes R. B., Kaplan B., Lehmann H., Tang P. C.** Clinical decision support systems for the practice of evidence-based medicine, *Journal of the American Medical Informatics Association*, 2001, vol. 8, no. 6, pp. 527—534.

20. **Moja L., Friz H. P., Capobussi M., Kwag K., Banzi R., Ruggiero F., González-Lorenzo M., Liberati E. G., Mangia M., Nyberg P., Kunnamo I., Cimminiello C., Vighi G., Grimshaw J., Bonovas S.** Implementing an evidence-based computerized decision support system to improve patient care in a general hospital: the CODES study protocol for a randomized controlled trial, *Implementation Science*, 2016, no. 11:89, pp. 1—10.

21. **Gribova V. V., Kleshhev A. S., Moskalenko F. M., Timchenko V. A., Fedorishhev L. A., Shalfeeva E. A.** *Upravljaemaja grafovymi grammatikami razrabotka obolochek intellektual'nyh servisov na oblachnoj platforme IACPaaS* (Graph grammar-driven development of intelligent service shells on the IACPaaS cloud platform), *Software Engineering,* 2017, vol. 8, no.10, pp. 435—447 (in Russian).

22. **Gribova V., Okun D., Petryaeva M., Shalfeeva E., Tarasov A.** Ontology for Differential Diagnosis of Acute and Chronic Diseases, *Communications in Computer and Information Science*, 2018, vol. 934, pp. 152—163.

23. **Gribova V. V., Petrjaeva M. V., Okun' D. B., Shalfeeva E. A.** *Ontologija medicinskoj diagnostiki dlja intellektual'nyh sistem podderzhki prinjatija reshenij* (The ontology of medical diagnostics for intelligent decision support systems), *Ontology of Designing*, 2018, vol. 8, no. 1 (27), pp. 58—73 (in Russian).

24. **Gribova V. V., Okun' D. B.** *Ontologii dlja formirovanija baz znanij i realizacii lechebnyh meroprijatij v medicinskih intellektual'nyh sistemah* (Ontologies for the formation of knowledge bases and the implementation of therapeutic measures in medical intelligent systems), *Information Science And Control Systems*, 2018, no. 3 (57), pp. 71—80 (in Russian).

25. **Gribova V. V., Okun' D. B.** *Ontologija baza znanij vosstanovitel'nogo lechenija* (Ontology of rehabilitation treatment knowledge base), *Sistemnyj analiz v medicine (SAM 2018): Materialy XII mezhdunarodnoj nauchnoj konferencii pod obshh. red. V. P. Kolosova. (Blagoveshhensk, 2018 g.)*, 2018, pp. 47—50 (in Russian).

26. **Gel'cer B. I., Skljar L. F., Eliseeva V. S., Shahgel'djan K. I., Markelova E. V., Emceva E. D., Beniovai S. N.** *Ocenka immunologicheskih pokazatelej pri lekarstvennoj ustojchivosti vich na fone VAART* (Evaluation of immunological parameters for drug resistance of HIV on the background of HAART), *HIV Infection and Immunosuppressive Disorders*, 2018, vol. 10, no. 1, pp. 54—62 (in Russian).

27. **Gel'cer B. I., Shahgel'djan B. I., Kurpatov I. A., Kotel'nikov V. N.** *Rezul'taty modelirovanija dolzhnyh velichin sily dyhatel'nyh myshc na osnove metodov iskusstvennogo intellekta* (The results of modeling the proper values of the strength of the respiratory muscles based on the methods of artificial intelligence), *Rossijskij fiziologicheskij zhurnal im. I. M. Sechenova*, 2018, vol. 104, no. 9, pp. 1065—1074 (in Russian).

28. **Gribova V. V., Kleshhev A. S., Moskalenko F. M., Timchenko V. A., Fedorishhev L. A., Shalfeeva E. A.** *Oblachnaja platforma IACPaaS dlja razrabotki obolochek intellektual'nyh servisov: sostojanie i perspektivy razvitija* (IACPaaS cloud platform for intelligent service shell development: state and development prospects), *Programmnye produkty i sistemy* (Software & Systems), 2018, vol. 31, no. 3, pp. 527—536 (in Russian).

29. **Shahgel'djan K. I., Gel'cer B. I., Gmar' D. V., Krivelevich E. B., Teuk K. A., Trankovskaja L. V.** *Problemy analiza dannyh medicinskoj statistiki* (Problems of medical statistics data analyzing), *Problems of Social Hygiene, Public Health and History of Medicine*, 2018, vol. 26, no. 3, pp. 132—136 (in Russian).

30. **Gribova V. V., Petryaeva M. V., Okun D. B., Tarasov A. V.** Software Toolkit for Creating Intelligent Systems in Practical and Educational Medicine, *2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC).* (Vladivostok, 18—25 Aug. 2018), IEEE Xplore, 2018, pp. 1—5.

31. **Gribova V. V., Moskalenko F. M., Okun' D. B., Petrjaeva M. V.** *Oblachnaja sreda dlja podderzhki klinicheskoj mediciny i obrazovanija* (Cloud environment for the support of clinical medicine and education), *Information technologies for the Physician*, 2016, no. 1, pp. 60—66 (in Russian).