

**О.Э. КАРПОВ,**

д.м.н., член-корр. РАН, профессор, Федеральное государственное бюджетное учреждение «Национальный медико-хирургический Центр имени Н.И. Пирогова» Министерства здравоохранения Российской Федерации, Москва, Россия, e-mail: karpov_oe@mail.ru

С.А. СУББОТИН,

Федеральное государственное бюджетное учреждение «Национальный медико-хирургический Центр имени Н.И. Пирогова» Министерства здравоохранения Российской Федерации, Москва, Россия, e-mail: subbotinsa@pirogov-center.ru

Д.В. ШИШКАНОВ,

к.ф.-м.н., Федеральное государственное бюджетное учреждение «Национальный медико-хирургический Центр имени Н.И. Пирогова» Министерства здравоохранения Российской Федерации, Москва, Россия, e-mail: shishkanovdv@pirogov-center.ru

ИСПОЛЬЗОВАНИЕ МЕДИЦИНСКИХ ДАННЫХ ДЛЯ СОЗДАНИЯ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ

УДК 614:004.62

Карпов О.Э., Субботин С.А., Шишканов Д.В. Использование медицинских данных для создания систем поддержки принятия врачебных решений (Федеральное государственное бюджетное учреждение «Национальный медико-хирургический Центр имени Н.И. Пирогова» Министерства здравоохранения Российской Федерации, г. Москва, Россия)

Аннотация. Медицинские данные – неотъемлемая часть рабочих процессов в деятельности медицинских организаций. Ежедневно все большее количество данных обрабатывается в цифровом формате. Что с ними делать, и какими они должны быть для использования системами поддержки принятия врачебных решений? В статье приведены важные критерии, определяющие качество данных.

Ключевые слова: медицинская информационная система, медицинские данные, системы поддержки принятия врачебных решений.

UDC 614:004.62

Karpov O.E., Subbotin S.A., Shishkanov D.V. Medical data usage to create medical decision support systems (Federal state budgetary institution "National Medical and Surgical Center named after N.I. Pirogov" of the Ministry of Healthcare of the Russian Federation)

Abstract. Medical data is an integral part of the business-processes in the activities of medical organizations. Every day more and more data is processed digitally. What should we do with them and what should they be like to use medical decision support systems? The article presents important criteria that determine the quality of data.

Keywords: medical information system, medical data, medical decisions support systems.

Национальная программа «Цифровая экономика Российской Федерации» [1] призвана обеспечить цифровую трансформацию российской экономики и социальной сферы, в том числе здравоохранение. Она включает в себя самостоятельный федеральный проект «Цифровые технологии», основной задачей которого является создание «сквозных» технологий преимущественно на основе отечественных разработок. В их число входят «Большие данные» и «Нейротехнологии и искусственный интеллект». Ключевыми методами для создания прикладных систем на основе таких технологий на сегодняшний день являются нейронные сети и машинное обучение. Именно на них базируется абсолютное большинство инновационных решений, предлагаемых медицинским учреждениям как отечественными стартапами и научными коллективами, так и ведущими мировыми технологическими компаниями.

История применения систем на основе искусственного интеллекта в здравоохранении насчитывает десятки лет, но только современные подходы, заключающиеся в том, что «машину обучают с помощью большого количества данных», обеспечили прорыв [2]. Создание прикладных решений на базе технологий машинного обучения, в первую очередь интеллектуальных систем поддержки принятия врачебных решений (СППВР), из научного исследования с непредсказуемым результатом стало опытно-конструкторской работой. Если так, то, казалось бы, все просто, ведь каждое медицинское учреждение в процессе деятельности накапливает громадное количество данных – как по количеству медицинских записей, изображений, вспомогательной информации, так и по их объему. Так, ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России (Пироговский Центр) является крупным многопрофильным учреждением, и только за последний год в базах данных накопилось более полумиллиона протоколов амбулаторных приемов, десятки тысяч протоколов оперативных вмешательств, медицинских изображений, относящихся к разным модальностям и инструментальной диагностики, и лабораторных исследований. И все эти данные взаимосвязаны, структурированы в виде, удобном для медицинских работников (но зачастую не для машинного обучения), проверены экспертами. Аналогичная ситуация в большинстве крупных медицинских центров, а региональные системы здравоохранения ведут активную работу по созданию централизованных архивов медицинских изображений, объединяющих данные десятков и даже сотен учреждений.

Однако активно применяемых в российской практике СППВР на базе технологий искусственного интеллекта крайне мало. Более того, на дату подготовки статьи имелась единственная публичная проведенная в России количественная оценка применимости СППВР для автоматизированного выявления заболеваний [3]. Авторы пришли к выводу, что система применима «только для массовых периодических осмотров в популяциях с низкой претестовой вероятностью наличия патологии», однако «технология может быть рекомендована для полуавтоматизированного формирования групп риска по туберкулезу легких для последующей верификации результатов врачом-рентгенологом». Таким образом, на современном этапе поддержка искусственного интеллекта наиболее полезна в массовых рутинных медицинских процессах, прежде

всего при скрининге, и без врачей-экспертов их создание и эксплуатация невозможны.

Рассмотрим причины, препятствующие созданию СППВР на основе данных медицинского учреждения и способы их преодоления. При создании СППВР медицинская организация может выступать в различных ролях, что накладывает специфические требования к данным:

- в качестве заказчика разработки медицинское учреждение должно обеспечить пригодные для машинного обучения данные в необходимом объеме; при этом очень сложно оценить применимость создаваемого решения в других организациях;
- в качестве эксперта или организатора клинических испытаний медицинское учреждение должно количественно оценить качество СППВР на своих данных;
- как пользователь учреждение должно включить новации в регулярные процессы оказания медицинской помощи и доказательно повысить ее качество и (или) объемы, снизить затраты. Необходимо учитывать порождаемые СППВР риски, основным из которых является работа на «новых» данных, а они всегда будут не полностью соответствовать обучающей выборке, даже если СППВР разрабатывался внутри организации, например, в эксплуатацию ввели новое диагностическое оборудование, изменилась структура входного потока пациентов.

Во всех случаях медицинской организации критически важно не потерять контроль над данными и обеспечить возможность работы с множеством СППВР от разных поставщиков, что накладывает повышенные требования к интероперабельности используемых автоматизированных систем. Отметим, что во всех случаях требуется, чтобы данные не просто «были», а соответствовали требованиям, которые могут отличаться для создания, апробации и эксплуатации разных СППВР. В соответствии с ГОСТ Р ИСО 9001–2015 «Системы менеджмента качества. Основные положения и словарь» степень соответствия совокупности присущих характеристик объекта требованиям называется качеством. Так что же такое данные и их качество, сколько их нужно для создания СППВР и как их обеспечить?

Ключевые понятия в серии ГОСТ Р ИСО 8000 «Качество данных» и серии ISO/IEC25000 «System and Software Quality Requirements and Evaluation (SQuaRE)» – «Программная инженерия. Требования и оценка качества программного продукта».



Первая часть ИСО 8000 [6] определяет принципы, лежащие в основе серии:

1) качество данных затрагивает данные, имеющие определенное назначение, т.е. участвующие в принятии какого-либо решения;

2) качество данных затрагивает нужные и подходящие данные, уместные в подходящем месте в подходящее время;

3) качество данных отвечает требованиям потребителя;

4) качество данных предотвращает повторение дефектов данных и сокращает избыточные расходы.

Вторая часть ИСО 8000 «Словарь» [4] вводит следующие определения:

- данные (data): символическое представление чего-либо, частично зависящего в своем значении от метаданных;

- управление качеством данных (data quality management): согласованная деятельность по контролю и управлению организацией, имеющей отношение к качеству данных.

В стандарте ИСО 8000 «Основные данные. Обмен данными характеристик. Словарь» [5] приводятся другие, более простые для практического применения термины и определения из других стандартов этой серии:

- данные: интерпретируемое представление информации в официальной форме, удобной для передачи, интерпретации и обработки;

- информация: значимые данные. Также – знания или сведения, относящиеся к таким объектам, как факты, события, предметы, процессы или идеи,

включая концепции, которые в соответствующих контекстах имеют конкретное значение.

Стандарт ИСО 8000 «Основные данные. Обмен данными характеристик. Обзор» [15] вводит терминологию данных и требования к описанию данных (см. рис. 1).

Стандарт «Модель качества данных ИСО 25000 [9] определил набор характеристик качества данных (таблица 1) которые делятся на:

- системно независимые, внутренне присущие качества данных, которые определяют степень, в которой имеют естественную способность удовлетворить заявленные и предполагаемые потребности при использовании в определённых условиях;

- системно зависимые, которые определяют степень, в которой качество данных достигается и сохраняется только в компьютерной системе при использовании в определённых условиях.

Анализ характеристик качества данных объясняет, почему не вся информация в базах данных медицинских учреждений может использоваться для создания СППВР. Для «первичных» данных, непосредственно хранимых в медицинских информационных системах (МИС), крайне сложно оценить степень согласованности или точности. Даже в более зрелой, с точки зрения цифровой трансформации, бизнес-среде «...только 35% компаний управляют данными централизованно, а более половины (57%) выявляют ошибки в данных постфактум... И хотя большинство организаций используют какое-либо технологическое решение для контроля, подготовки и очистки, почти треть компаний (29%) до сих пор

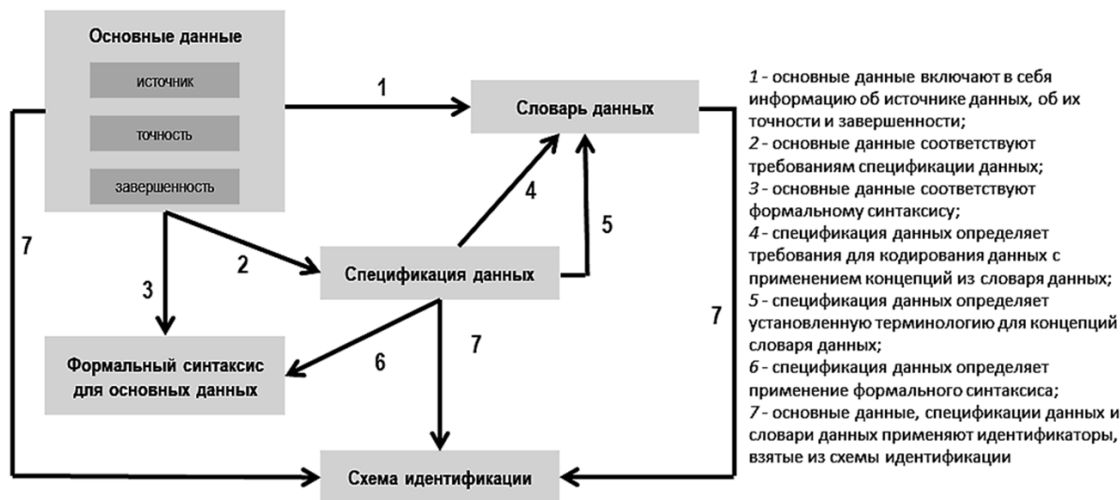


Рис. 1. Построение данных по ИСО 8000



Таблица 1

Характеристики качества данных [9]

№	Характеристика качества данных		Системно	
			независимое	зависимое
1	Accuracy	Точность	+	-
2	Completeness	Полнота	+	-
3	Consistency	Согласованность	+	-
4	Credibility	Достоверность	+	-
5	Correctness	Правильность	+	-
6	Accessibility	Простота доступа	+	+
7	Compliance	Соответствие	+	+
8	Confidentiality	Конфиденциальность	+	+
9	Efficiency	Эффективность	+	+
10	Precision	Точность	+	+
11	Traceability	Контролируемость	+	+
12	Understandability	Понятность	+	+
13	Availability	Доступность	-	+
14	Portability	Мобильность	-	+
15	Recoverability	Восстанавливаемость	-	+

проверяют и очищают свои данные «вручную». 77% СЮ [старшее должностное лицо в сфере информатизации] справедливо рассматривают данные в качестве стратегического актива, который не до конца используется в организации» [11].

Ситуация осложняется тем, что в большинстве случаев для целей машинного обучения требуется добавлять данные, которые отсутствуют в МИС. Необходимая работа по добавлению необходимых атрибутов, описанию и интерпретации первичной информации (разметка) является трудоемкой, и не только сама зависит от качества первичных данных, но и сильно влияет на качество данных для СППВР. В частности, оценка выявленных ошибок и пропусков в данных медицинской информационной системы Пироговского Центра [14] относится именно к первичным системно-зависимым данным и не может быть использована для оценки качества данных с точки зрения пригодности для создания СППВР. Опытные врачи-эксперты могут при выполнении разметки это качество повысить, а неопытные – потерять. Практика Пироговского Центра показывает, что для получения значимых результатов должны рассматриваться целостные наборы взаимосвязанных первичных данных, что еще более повышает требования к их качеству, но обеспечивает наиболее важные характеристики для поддержки принятия решений – согласованность,

достоверность и точность. Такое обогащение дает возможность врачу проследить причины рекомендуемого СППВР решения.

Например, при анализе применимости технологий искусственного интеллекта в маммографии вместе с проектом «Третье мнение» в качестве решаемой задачи была выбрана не оценка вероятности отдельных патологий, а автоматизированное определение категории по шкале BI-RADS. Для этого:

- врачами-экспертами были определены классификаторы, подготовлены справочники и домены данных предметной области, необходимые для разметки маммограмм;
- разработчиками создано приложение для формирования базы данных обезличенных размеченных медицинских изображений, предназначенных для обучения нейронной сети; его интерфейс приведен на *рис. 2*;
- совместно проведено его тестирование на реальных данных Пироговского Центра.

Важно, что качество как степень соответствия характеристик полученного набора данных отвечает требованиям по созданию СППВР, а база данных размеченных медицинских изображений будет регистрироваться как результат интеллектуальной деятельности. Отдельно следует подчеркнуть, что при этом используются только общепринятые

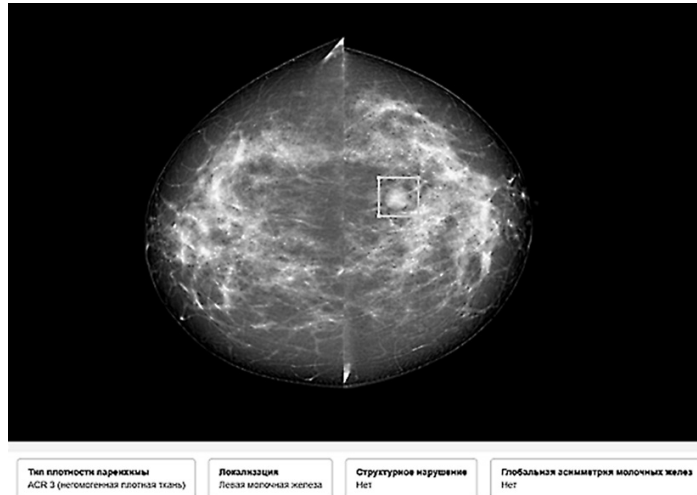


Рис. 2. Интерфейс решения для описания и интерпретации маммограмм врачами-экспертами

семантические активы, что согласно [13] является единственным способом построения экосистемы цифрового здравоохранения, в которой СППВР могут применяться широким кругом медицинских учреждений с минимальными затратами.

Еще более показателен проект, в котором технологическим партнером является компания «Новая

медицина» – искусственный интеллект применяется именно для проверки первичных данных. В качестве результата ожидается автоматизированный анализ качества медицинской документации. В специальном интерфейсе (см. рис. 3) врачи-эксперты оценивают обезличенный протокол приема гинеколога по перечню установленных службой контроля качества

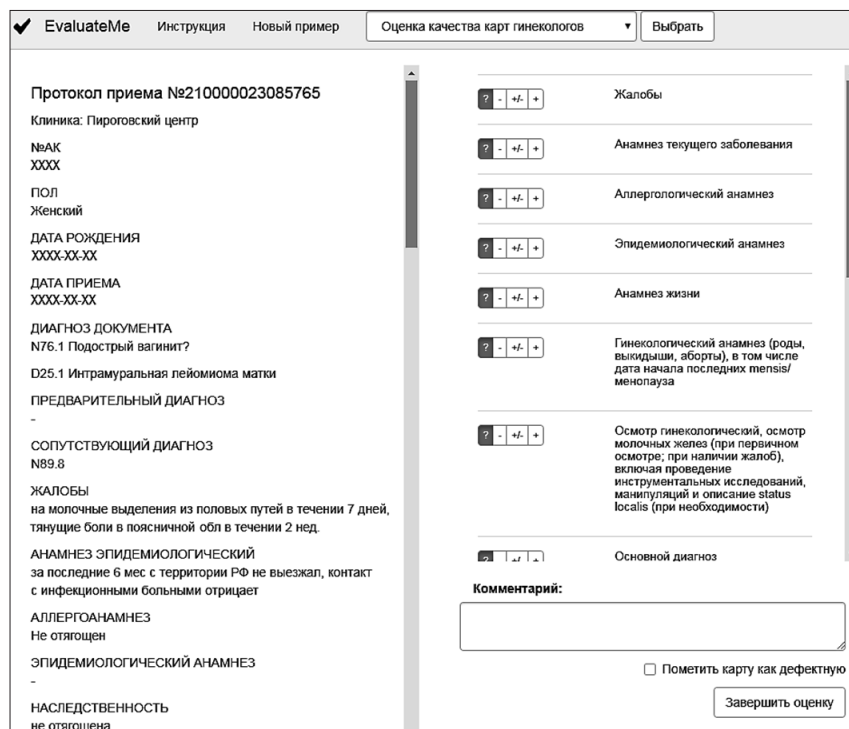


Рис. 3. Интерфейс решения для экспертизы врачами-экспертами протоколов приема



критериев, выделяя причины, по которым он может быть дефектован.

Размеченные протоколы будут использоваться для автоматического выделения подмножества документов, которые предназначены для повторной ручной проверки. Ожидается, что после накопления обучения на нескольких тысячах размеченных протоколов использование СППВР позволит усовершенствовать контрольно-экспертную работу в амбулаторных подразделениях Пироговского Центра.

Где разработчики СППВР могут получить качественные данные вне работы с конкретными медицинскими учреждениями? Ключевым источником являются свободно распространяемые наборы размеченных данных. Крупнейший мировой репозиторий реальных и модельных задач машинного обучения – UCI Machine Learning repository – ведет свою историю с 1987 года! В разделе «Наука о жизни» сейчас содержится 107 наборов, два из которых входят в число наиболее популярных (таблица 2).

Имеются и другие популярные репозитории, например:

- библиотека DICOM-изображений (<http://www.osirix-viewer.com/resources/dicom-image-library/>);
- открытая онлайн-база данных медицинских изображений, учебных кейсов и клинических материалов, интегрированных изображений и текстовых метаданных (<https://medpix.nlm.nih.gov/>);
- набор баз данных по маммограммам <http://www.mammoimage.org/databases/>, который представляет собой реальные снимки груди с известными типами заболеваний;
- коллекция PhysioBank (<http://www.physionet.org/cgi-bin/atm/ATM>), которая включает в себя наборы ЭКГ, ЭЭГ и других биомедицинских цифровых результатов от здоровых людей и пациентов с различными состояниями;
- наборы данных сайта <https://www.kaggle.com>, на котором постоянно проводится масса конкурсов, в том числе и связанных со здравоохране-

нием (примеры см. – <https://www.kaggle.com/datasets?sortBy=hottest&group=public&page=1&pageSize=20&size=all&filetype=all&license=all&tags=4202>).

Большинство этих и аналогичных наборов используется студентами, научными коллективами и стартапами для обучения и создания прототипов СППВР. Во многом именно с ограниченностью качественных наборов данных связано то, что многие из них делают сходные решения, в частности, автоматизированный анализ рентгеновских снимков грудной клетки стал популярным именно после одного из конкурсов Kaggle.

При этом не только качество, но и размеры наборов имеют первостепенное значение. При работе внутри организации объемы данных нужного качества постоянно растут, но для выхода на рынок разработчики обычно прибегают либо к приобретению готовых данных (заказывают разметку), либо вступают в коллаборацию с медицинскими учреждениями. Так, корпорации IBM пришлось купить компанию Merge Healthcare за 1 миллиард долларов США, чтобы получить 30 млрд. медицинских снимков [2], а патентная заявка Google, предлагающая СППВР, которая будет собирать и компилировать данные ЭМК, чтобы предупредить поставщиков о предстоящих клинических событиях, основана на том, что «...Google, Калифорнийский университет в Сан-Франциско, медицинская школа Стэнфордского университета и Чикагский университет медицинских исследований использовали 46,864,534,945 элементов медицинских карт, полученных из историй болезни 216,221 взрослых пациентов, госпитализированных в течение не менее 24 часов...» [7].

Сравним с объемами оказания медицинской помощи в России. Извлечения из данных Росстата по отчету за 2017 год [10] приведены в таблице 3. Объемы сопоставимы с крупнейшими рассмотренными выше наборами данных. Ежегодно российское здравоохранение формирует много больше миллиарда протоколов приемов, десятки миллионов протоколов оперативных вмешательств,

Таблица 2

Характеристики наборов данных репозитория UCI Machine Learning repository

Наименование		Дата предоставления	Количество скачиваний на 01/03/2019	Ссылка на набор
Breast Cancer Wisconsin (Diagnostic) Data Set	Висконсинский диагностический набор данных по раку груди	01/11/1995	857 300 5 по популярности	http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
Heart Disease Data Set	Набор данных по заболеваниям сердца	01/07/1998	834 281 7 по популярности	http://archive.ics.uci.edu/ml/datasets/Heart+Disease



что сопровождается значительными объемами лабораторных и инструментальных диагностических исследований.

Однако в России, по контрасту с распространением доступных медицинских данных в мире, имеется единственный зарегистрированный в установленном порядке «Репозиторий открытых данных для машинного обучения и «искусственного интеллекта» [8]. Функциональные возможности этой базы данных обеспечивают использование формализованных знаний врачей-экспертов в виде деперсонализированных компьютерных томограмм с отмеченными патологическими изменениями в легочной ткани работы и обучение врачей-диагностов, валидацию и исследование диагностической ценности алгоритмов анализа биомедицинских данных. При этом ценность данных уже осознается, так только

в 2018 году на покупку данных о перемещении граждан власти столицы потратили 101,8 млн. руб. Всего с 2015 года на эти цели из городского бюджета потрачено 516 млн. руб. (<https://www.rbc.ru/politics/04/03/2019/5c7cd5fe9a794760d9cfb900>).

Более того, специализированное решение компании Google для поиска наборов данных <https://toolbox.google.com/datasetsearch> не находит ни одного набора по ключевым словам «медицинская карта» или «ЭМК» (электронная медицинская карта), в то время как по «EHR» (electronic health record) результатов много и, что очень важно, постоянно появляются новые. Так, 22 февраля 2019 года сделан доступным типовой набор ЭМК Министерства по делам ветеранов США <https://catalog.data.gov/dataset/va-personal-health-record-non-identifiable-data>.

Таблица 3

Количественные характеристики системы здравоохранения Российской Федерации

№	Показатель	Год	2005	2010	2013	2014	2015	2016
1	Число посещений врачей в амбулаторном звене, млн.*		1264,0	1332,6	1357,8	1323,4	1288,8	1237,0
2	Численность госпитализированных пациентов на 100 человек населения		22,4	22,2	21,1	21,4	20,8	20,6
3	Среднее число дней пребывания пациента на койке		13,8	12,6	12,1	11,8	11,5	11,1
4	Число проведенных операций в стационаре, тыс.		8735	9277	9503	9740	9882	9974
5	Число проведенных операций в амбулаторно-поликлинических организациях		6062	5822	5757	5709	5661	5590

* не включая количество посещений среднего медицинского персонала (143,6 млн. в 2016 году)



Рис. 4. Структура управления качеством данных [12]

Чтобы добиться требуемого качества данных, стандарт «Основные данные. Структура управления качеством» ИСО 8000 рекомендует следующую структуру управления качеством основных данных [12] (рис. 4). Для медицинской организации роли в этой структуре распределены между поставщиками МИС (только в части системно зависимых характеристик) и своими сотрудниками. Желательно, чтобы они представляли не ИТ-подразделения, а функции контроля и обеспечения качества информации должны быть включены в их должностные обязанности. Эти функции – от обеспечения форматно-логического контроля форм ввода и подготовки операторов до создания самостоятельных аналитических подсистем – обеспечивают качество первичных данных для целей организации. Как указано выше, для целей создания СППВР даже в рамках отдельного

медицинского учреждения требуются дополнительные усилия врачей-экспертов и оценка качества обогащенных данных.

Чтобы создаваемые наборы данных были пригодны для широкого использования, требуется иметь общепризнанные правила их разметки, способы проверки качества и прозрачные условия доступа к этой информации. Очевидно, что создание таких наборов российских данных требует методической поддержки на уровне регулятора отрасли и привлечения к разметке наиболее профессиональных экспертов. Такие шаги смогут привлечь к разработке СППВР новые талантливые команды математиков и программистов, повысят культуру работы с информационными системами и данными внутри медицинского сообщества, как следствие – приблизят цифровую трансформацию здравоохранения.

ЛИТЕРАТУРА



1. Паспорт национального проекта «Цифровая экономика Российской Федерации» // <http://static.government.ru/media/files/urKHm0gTPPnzJlaKw3M5cNLo6gczMkPF.pdf> (Дата обращения: 27.02.2019).
2. Гусев А.В., Добридюк С.Л. Искусственный интеллект в медицине и здравоохранении / Информационное общество, 2017. – № 4–5. – С. 78–93.
3. Морозов С.П., Владзимирский А.В., Ледихова Н.В., Соколова И.А., Кульберг Н.С., Гомболевский В.А. Оценка диагностической точности системы скрининга туберкулеза легких на основе искусственного интеллекта // Туберкулез и болезни лёгких. – 2018. – Т. 96, № 8. – С. 42–49. DOI: 10.21292/2075-1230-2018-96-8-42-49.
4. ГОСТ Р ИСО 8000-2-2014 «Качество данных. Часть 2. Словарь».
5. ГОСТ Р ИСО 8000-102-2011 «Качество данных. Часть 102. Основные данные. Обмен данными характеристик. Словарь».
6. ГОСТ Р 56214-2014/ISO/TS8000-1:2011 «Качество Данных. Часть 1. Обзор».
7. Dave Muoio Google patent application offers new details on company's predictive EHR aggregation system // https://www.mobihealthnews.com/content/google-patent-application-offers-new-details-companys-predictive-ehr-aggregation-system?mkt_tok=eyJpIjoiWVRJd09USmtaRGRrTW1SaClslnciOiJadVhiQzNtSEdhMk1OSlg0KzRhZXBObHZKZyZwSmRjYaklrMVp4WGozUGNqQd0Rjc3BuTjZlMVZENHZcLzN5RHdwchJlTlVWVhOGlyUnJqUVVhac0dMXc80VUNyRmtRYUZIUU9BcjRGeDh1dW1Q0N6YTJBelhiNVRqUUDqQlhtcE8rbkNpln0%3D (дата обращения 27.02.2019).
8. Тегированные результаты компьютерных томографий легких: а.с. 2018620500 Рос. Федерация / Морозов С.П., Кульберг Н.С., Гомболевский В.А. с соавт.; заявитель и правообладатель: ГБУЗ «НПЦМР ДЗМ». – № 2018620148; заявл. 06.02.2018; опубл. 28.03.2018, Бюл. № 4. – 1 с.
9. ISO/IEC25012 – Data Quality model: defines a general data quality model for data retained in a structured format within a computer system. It focuses on the quality of the data as part of a computer system and defines quality characteristics for target data used by humans and systems // <https://www.iso25000.com/index.php/en/iso-25000-standards/iso-25012?limit=5&start=15> (Дата обращения: 06.03.2019).
10. Федеральная служба государственной статистики «Здравоохранение в России» 2017 // http://www.gks.ru/free_doc/doc_2017/zdrav17.pdf (Дата обращения: 04.03.2019).
11. Морозова О.А. Управление данными как активом предприятия: качество данных и бизнес-результат // <https://fd.ru/articles/158513-upravlenie-dannymi-kak-aktivom-predpriyatiya-rek-> (Дата обращения: 05.03.2019).
12. ГОСТ Р 56215-2014/ISO/TS8000-150:2011 «Качество данных. Часть 150. Основные данные. Структура управления качеством».
13. Карпов О.Э., Акаткин Ю.М., Конявский В.А., Шишканов Д.В., Ясиновская Е.Д. Цифровое здравоохранение в цифровом обществе. Экосистема и кластер. М.: ДПК Пресс, 2017. – 220 с.
14. Карпов О.Э., Субботин С.А., Замятин М.Н., Шишканов Д.В., Асташев П.Е., Прохорова Е.С. Имитационное моделирование деятельности современного многопрофильного медицинского учреждения // Вестник Российского экономического университета имени Г.В. Плеханова. 2018; (6): 57–66.
15. ГОСТ Р 54524-2011/ISO/TS8000-100:2009 «Качество данных. Часть 100. Основные данные. Обмен данными характеристик. Обзор».