БОЛЬШИЕ ДАННЫЕ В МЕДИЦИНЕ: БАЗА ДАННЫХ РЕНТГЕНОВСКИХ ИЗОБРАЖЕНИЙ ДЛЯ РЕШЕНИЯ ЗАДАЧ ДИАГНОСТИКИ, ЛЕЧЕНИЯ И ПРОВЕДЕНИЯ НАУЧНЫХ ИССЛЕДОВАНИЙ

В.А. Ковалев
Заведующий лабораторией анализа биомедицинских изображений ОИПИ НАН Беларуси, кандидат технических наук, Республика Беларусь

В.А. Лапицкий
Заместитель генерального директора по научной работе ОИПИ НАН Беларуси, Республика Беларусь

А.А. Дмитрук
Младший научный сотрудник лаборатории анализа биомедицинских изображений ОИПИ НАН Беларуси, Республика Беларусь

А.А. Калиновский
Научный сотрудник лаборатории анализа биомедицинских изображений ОИПИ НАН Беларуси, Республика Беларусь

Лаборатория анализа биомедицинских изображений, Объединенный институт проблем информатики Национальная академия наук Беларуси, alex.bsu@gmail.com

Целью настоящей работы является представление большой базы данных цифровых изображений грудной клетки, полученной в результате широкомасштабного скрининга населения на основе региональной телемедицинской системы. Приводятся основные количественные параметры, характеризующие базу данных и некоторые предварительные результаты анализа. По данным авторов, представляемая база изображений на сегодняшний день является крупнейшим цифровым архивом медицинских изображений указанного типа в мире. Представленные сведения могут быть полезны для научных работников и ИТ-инженеров, ведущих исследования в области обработки и анализа больших объемов биомедицинских данных, а так же разрабатывающих соответствующие методы, алгоритмы и программное обеспечение.

Введение

Большие коллекции медицинских изображений являются ценным источником информации, играющим важную роль в образовании, исследованиях по медицинской тематике и поддержке принятия решений в клинической практике. Известно, что одной из основных проблем при этом является постоянный рост размера цифровых архивов медицинских учреждений, который обусловлен большой доступностью соответствующего оборудования и медицинских сканирующих систем различных типов, способных производить цифровые снимки. К такому оборудованию относятся рентгеновские аппараты, томографы, ультразвуковые сканеры и микроскопы. Среднестатистическое отделение радиологии в настоящее время производит несколько терабайт данных в год [1]. Указанные причины побуждают исследователей к разработке инстру-

ментов для быстрого и эффективного доступа к данным в архивах медицинских изображений, прежде всего поиск похожих случаев из клинической практики.

Системы поиска медицинских изображений обычно позволяют осуществлять текстовый поиск по аннотациям или описаниям к изображениям, которые были заранее внесены вручную.

В последние десятилетия, в дополнение к традиционному текстовому поиску, возник поиск изображений по их содержанию (CBIR). Его отличительной особенностью является отсутствие необходимости в ручной аннотации изображений. Вместо этого, содержание изображений автоматически описывается с помощью визуальных признаков (например, характеристики цвета, формы или текстуры). Поиск похожих изображений осуществляется на основе сравнения визуальных признаков (дескрипторов) изображения-образца и остальных изображений коллекции с целью нахождения наиболее близких в пространстве признаков.

Другим важным направлением исследований является поиск новых связей между визуальным представлением диагноза и его текстовым описанием. Комбинация различных источников информации может быть применена для нахождения отличительных особенностей классов изображений или для визуализации скрытых структур, т.е. к открытию нового знания.

В Беларуси на базе корпоративной телекоммуникационной сети медицинских учреждений г. Минска функционирует уникальная для стран СНГ распределенная телемедицинская система реального времени по цифровой флюорографии [2], позволяющая повысить оперативность раннего выявления заболеваний легкого (прежде всего это туберкулез и рак).

В рамках системы, снимки с флюорографов городских поликлиник поступают в противотуберкулезные диспансеры, в которых врачи-рентгенологи выполняют анализ цифровых рентгеновских изображений и описывают результаты исследований в текстовом виде для поликлиник.

Снимки, а также результаты работы врачей-рентгенологов сохраняются в базах данных диспансеров [2]. Поскольку процедура обязательного рентгеновского обследования в стране является ежегодной, то ожидается, что через несколько лет количество единиц хранения может достигнуть числа, сопоставимого со всем населением Беларуси. Типичные примеры снимков грудной клетки, хранящихся в базе данных, представлены на рисунке 1.

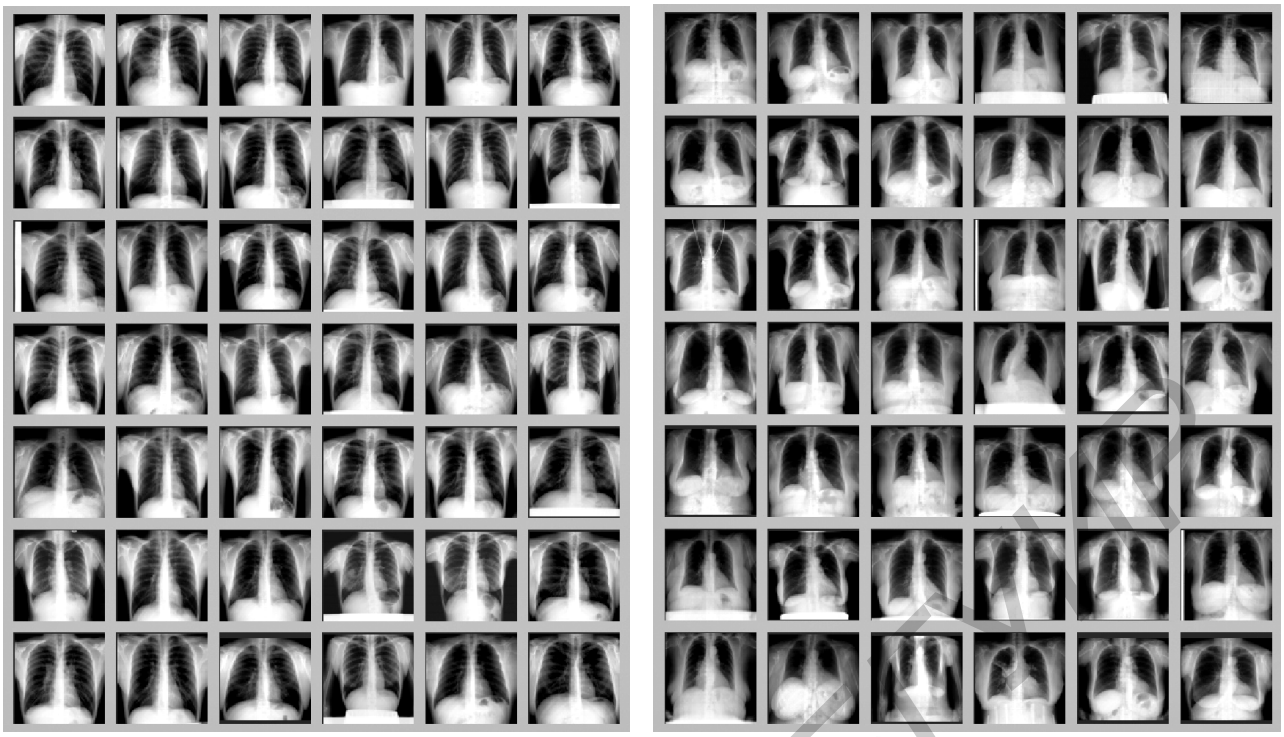

Рис.1. Типичные примеры цифровых изображений грудной клетки здоровых мужчин (21 год, левая панель) и женщин (70 лет, правая панель), хранящихся в описываемой базе данных

Таким образом, целью настоящей работы является описание существующего архива анонимизированных рентгеновских изображений и его предварительный анализ.

Тестовая база данных

Изображения в базе хранятся в формате DICOM, который является общепризнанным стандартом хранения медицинских изображений. База отличается большой разнородностью, поскольку содержит изображения с различных флюорографических аппаратов, которые могут отличаться разрешением, структурой полей DICOM-заголовка и т.д. Для удобства работы с изображениями, а также их унификации, все файлы были конвертированы в формат PNG (без потерь, 16 бит на пиксел). В результате была сформирована база, общая информация о которой представлена в таблице 1.

Таблица 1. Размер базы данных

Общий объем изображений	6,4 ТБ
Число пациентов	1090330
Число флюорографических снимков	1923802

На рисунке 2 показано распределение снимков по их разрешению в пикселах (показаны разрешения изображений, число которых в базе более 10000).

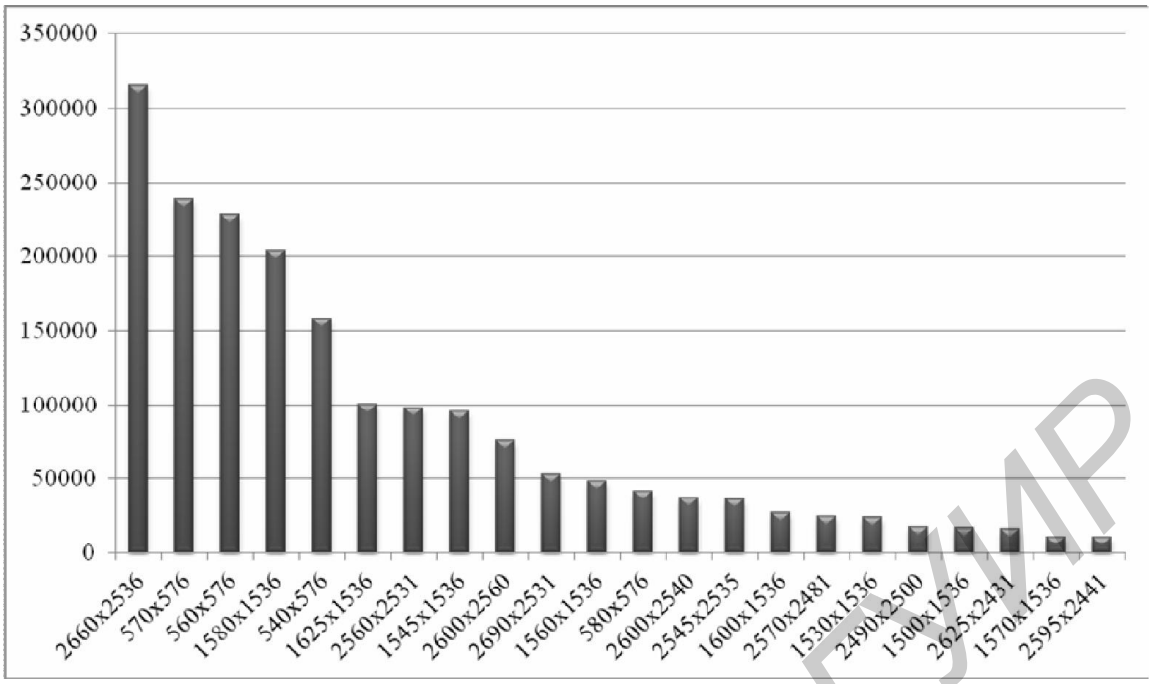

Рис.2. Наиболее часто встречающееся разрешение у снимков из базы данных

Как видно из рисунка, в базе содержатся снимки как высокого, так и низкого разрешения. Возможно, снимки низкого разрешения были получены с морально устаревших цифровых аппаратов, поскольку дата их получения начинается с 2001 года.

Подавляющее большинство пациентов (около 700 тыс.) обследовалось 1 раз, хотя в базе имеются пациенты с числом обследований до 15 раз (рисунок 3).

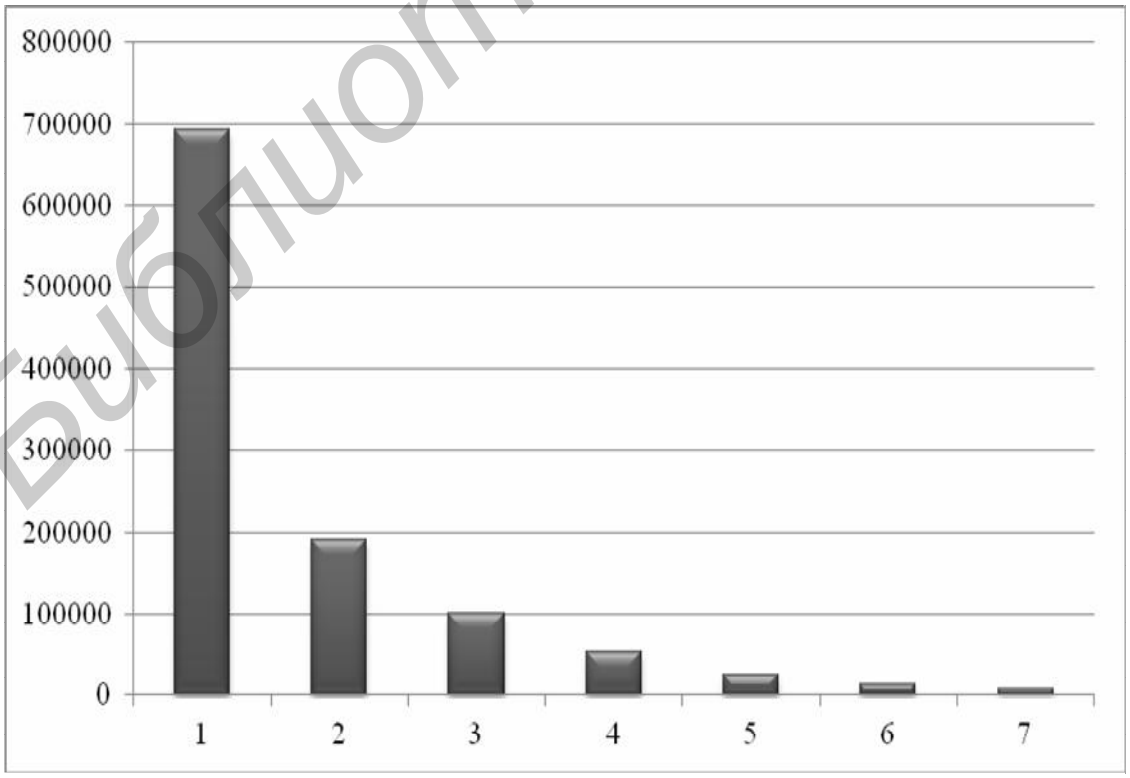

Рис. 3. Число обследований, приходившихся на одного пациента

Зная дату обследования и год рождения, а также пол пациентов, было построено соответствующее распределение (рисунок 4). Из рисунка видно, что начинаются флюорографические исследования примерно с 17-летнего возраста (более 42 тыс. исследований). Максимумы приходятся на возраст 18 лет (около 76 тыс. исследований) и 51 год (около 38 тыс. исследований). Небольшой всплеск также приходится на возраст 70 лет (более 15 тыс. исследований).

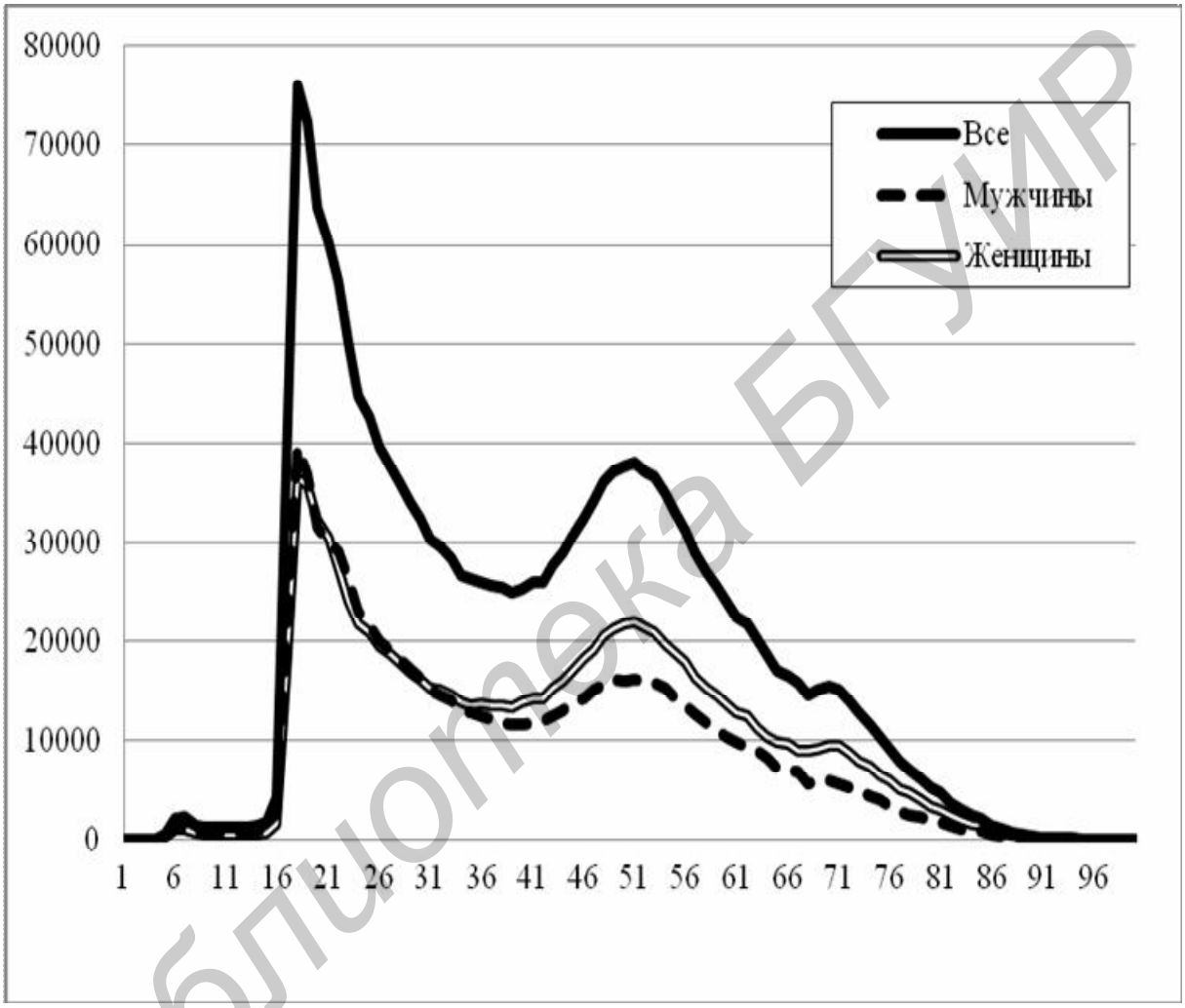

Рис. 4. Распределение числа исследований по возрасту

Подавляющее большинство изображений содержат текстовые описания, произведенные врачами-рентгенологами, также у примерно 24 тыс. изображений имеется экспертное заключение. Большинство описаний изображений соответствуют норме. Текстовые данные, для их последующего анализа, требуют дополнительной обработки, поскольку неструктурированны, содержат много специфических терминов и сокращений, ошибки в тексте и т.д.

Таким образом, в статье показан пример реального архива изображений медицинских учреждений Беларуси, который может использоваться при проведении научных исследований и разработок в области Big Data. Так, например, в

работе [3] представлены результаты разведочного анализа (Data Mining) с целью поиска закономерностей изменения размеров и формы легких в процессе естественного старения, а также с целью обнаружения статистически-значимых различий формы легкого у мужчин и женщин. В указанном исследовании было использовано более 100 000 рентгеновских изображений, представляющих испытуемых в возрасте от 21 до 70 лет, которые были отобраны из представленной базы данных цифровых рентгеновских снимков.

Литература

1. Partik, B. Digital (R)Evolution in Radiology / B. Partik, C. Schaefer-Prokop // Digital radiology in chest imaging / Ed. W. Hruby. – Springer Vienna, 2001. – P. 189-203.

2. Применение цифровых сканирующих аппаратов и передовые телемедицинские инновационные технологии в диагностике заболеваний легких / В.В. Анищенко [и др.]. – ОИПИ НАН Беларуси, 2010. – 136 с.

3. Kovalev V., Prus A., Vankevich P. Mining lung shape from X-ray images. In: *International Conf. on Machine Learning and Data Mining (MLDM-2009),* Leipzig, Germany, P.Perner (Ed.), LNAI, vol. 5632, Springer Verlag, 2009, pp. 554–568.