УДК 004.932.2

В.А. Ковалев[1], В.А. Левчук[1], И.В. Сафонов[1], О.В. Тарасов[2]

ПОИСК ВЗАИМОСВЯЗЕЙ МЕЖДУ ЛЕКАРСТВЕННОЙ УСТОЙЧИВОСТЬЮ И МОРФОЛОГИЕЙ ИЗОБРАЖЕНИЙ ПРИ ТУБЕРКУЛЕЗЕ

Приводятся результаты исследований, направленных на поиск взаимосвязей между лекарственной устойчивостью (ЛУ) туберкулеза легких и структурными признаками рентгеновских и компьютерно-томографических (КТ) изображений легкого. Представлена многоступенчатая процедура, включающая в себя вычисление признаков изображений, извлечение главных компонент, корреляцию главных компонент с клиническими данными пациентов, а также обратное отображение статистически значимых компонент на исходные признаки и далее – на исходные изображения с целью выделения соответствующих ключевых структурных образований. Проводится детальный статистический анализ взаимосвязей между степенью ЛУ и признаками изображений, включающий нахождение однофакторных корреляций, поиск многофакторных связей и проведение детальных кросс-проверок.

Введение

Спустя более чем 100 лет после открытия микобактерии туберкулеза легких эта болезнь продолжает оставаться угрозой здоровью и жизни людей во многих странах мира [1]. При наихудшем варианте развития заболевания болезнетворные микобактерии становятся невосприимчивыми к двум или более стандартным противотуберкулезным препаратам [1, 2]. Такие лекарственно-устойчивые формы туберкулеза требуют долговременного, тяжелого и затратного лечения, а в ряде случаев становятся неизлечимыми. Недавние исследования лекарственно-устойчивого туберкулеза указывают на чрезвычайную важность данной проблемы, что обусловлено постоянным ростом количества случаев устойчивости к одному и более лекарствам, вплоть до появления так называемой тотальной устойчивости, когда пациент не чувствителен ни к одному из известных лекарств [2]. Последнее особенно актуально в условиях Беларуси, где наблюдается значительное количество пациентов с тотальной устойчивостью к лекарствам [3].

Даная работа является частью проекта, одна из целей которого – создание открытых информационных ресурсов, касающихся проблемы лекарственной чувствительности туберкулеза, на национальном уровне [4]. Ключевой частью таких ресурсов служит база данных пациентов, содержащая клинические данные, рентгеновские и КТ-изображения грудной клетки пациентов, а также генетические данные бактерий *M. Tuberculosis*, взятых у пациентов с различными степенями лекарственной чувствительности (генетические данные будут добавлены в базу данных в 2013 г.).

Важной задачей проекта является исследование возможности предсказания степени лекарственной чувствительности по изображениям легких пациентов как одного из возможных дополнительных способов детектирования этой важнейшей характеристики заболевания. Если раннее предсказание ЛУ на основе рентгеновских и КТ-изображений окажется возможным (при условии приемлемого уровня вероятности правильного предсказания), то данный метод может быть применен на практике. Цель настоящей работы состоит в рассмотрении результатов предварительных исследований, посвященных поиску возможных корреляций между степенью ЛУ и текстурными (структурными) признаками изображений легких пациентов, больных туберкулезом. Указанная задача пока остается неизученной. Известны лишь попытки ее решения на основе традиционных рентгенологических (т. е. чисто экспертных) методов, предполагающих визуальный анализ и подсчет количества определенных образований в легких врачом-рентгенологом, что пока не привело к результатам, которые могут быть использованы в целях диагностики [5, 6]. Данная работа является первой попыткой поиска взаимосвязи между текстурными свойствами изображений и ЛУ туберкулеза с использованием методов компьютерного анализа изображений.

1. Материалы

Рентгеновские и КТ-изображения были использованы независимо для поиска возможных взаимосвязей между их структурой и степенью ЛУ пациентов. Все рентгеновские изображения были отсняты на аппарате KODAK Point-of-Care 260 и имели разрешение 2248×2248 пикселов. КТ-изображения были получены на томографе GE LightSpeed Pro 16 с расстоянием между слоями 2,5 мм. Общее количество аксиальных слоев изображений варьировало от 100 до 160 в зависимости от размера области интереса. На рентгеновских снимках области легких были выделены вручную, а на КТ-изображениях – полуавтоматически (предварительная автоматическая сегментация с последующей ручной корректировкой). Конечные результаты сегментации были проверены и подтверждены врачом-рентгенологом.

На первом этапе формирования тестовой выборки были отобраны 150 пациентов, по которым имелся полный набор клинических данных. Из указанной группы 24 пациента были исключены, поскольку они не подходили по дате проведения КТ-обследования. Далее из оставшихся 126 пациентов один был исключен из-за проблем с сегментацией легких на КТ-снимке. Еще 14 были исключены, поскольку их КТ-снимки были получены на томографах, отличных от GE LightSpeed Pro 16. В итоге еще четыре пациента были исключены из-за наличия у них ВИЧ-инфекции. В результате для проведения исследования были отобраны 107 пациентов, включая 64 мужчины и 43 женщины. Их средний возраст составлял 42,7 и 48,3 года соответственно. Возрастные различия были статистически незначимы ($p = 0{,}11$). Ключевой клинической характеристикой каждого пациента, представляющей наибольший интерес для данного исследования, являлась степень ЛУ. В рамках настоящего исследования параметр ЛУ был представлен бинарной величиной, причем значение 0 соответствовало лекарственно-чувствительной форме туберкулеза, а значение 1 – формам туберкулеза со всеми возможными степенями ЛУ. Кроме того, в статистическом анализе были задействованы такие характеристики пациента, как пол, возраст, рост, вес, признак повторного лечения и некоторые другие.

Все исходные данные, используемые в настоящей работе, за исключением рентгеновских снимков, были взяты с интернет-портала [4].

2. Алгоритмы и методы, использованные для поиска взаимосвязей

С целью выявления возможных взаимосвязей между характеристиками изображений и ЛУ, проверки найденных закономерностей, а также визуализации конкретных участков и структур на изображениях, которые имеют непосредственное отношение к найденным закономерностям, была реализована следующая многошаговая процедура:

1. Вычисление дескрипторов изображений. Для вычисления количественных характеристик изображений в данной работе был применен подход, основанный на использовании расширенных многомерных матриц совместной встречаемости. Такой подход обладает достаточной гибкостью и способен охватить широкий диапазон структурных свойств как двухмерных, так и трехмерных медицинских изображений [7–9]. Так, для описания структуры трехмерных КТ-изображений легкого были использованы шестимерные массивы частот совместной встречаемости, обозначаемые *IIGGAD* [7], в которых вычисляются частоты встречаемости пар вокселов с различными уровнями интенсивности (*I*), величинами градиентов (*G*), величинами углов между направлениями векторов градиентов (*A*) и расстояниями между вокселами (*D*). Следуя терминологии, сложившейся в области текстурного анализа изображений, указанные массивы в данной работе называются шестимерными матрицами совместной встречаемости. Для проверки устойчивости полученных результатов к изменению типа дескриптора также вычислялись редуцированные, упрощенные версии матриц, обозначаемые *IID*, учитывающие только значения интенсивности вокселов (*I*) на определенных расстояниях между ними (*D*). Описание структуры рентгеновских изображений производилось при помощи измененной версии четырехмерных матриц совместной встречаемости, обозначаемых *IIID* [8] и вычисляющих частоту совместной встречаемости троек пикселов с определенными интенсивностями (*I*), все попарные расстояния между которыми равны *D*.

2. Формирование таблицы данных. На данном этапе для каждого конкретного пациента элементы матрицы совместной встречаемости выписывались в отдельную строку таблицы данных. Из этих строк была сформирована стандартная таблица данных вида «объект – характеристики», в которой каждая ячейка матрицы совместной встречаемости соответствовала отдельной характеристике, т. е. столбцу таблицы данных по пациентам. Получившаяся в результате таблица содержала 107 строк (количество пациентов), в то время как количество столбцов (количество признаков) зависело от типа использованной матрицы встречаемости и изменялось от нескольких сотен до нескольких десятков тысяч в случае сложного типа дескриптора. Таблицы данных, представленные в таком виде, крайне неудобны (а в случае очень большого количества столбцов и непригодны) для проведения статистического анализа. Это вызвано тем, что количество признаков в них значительно превышает количество объектов и, следовательно, существует большой шанс обнаружения псевдозакономерностей. Кроме того, многие признаки (элементы матрицы совместной встречаемости) обычно сильно коррелированы между собой.

3. Уменьшение количества характеристик. Для уменьшения размерности пространства признаков был применен метод главных компонент (Principal Component Analysis, PCA). Поскольку элементы матриц встречаемости сильно коррелированы, метод главных компонент, будучи применен к входным данным, обычно дает на выходе 5–15 некоррелированных главных компонент (Principal Component, PC) [9]. В этом случае результирующая таблица данных, в которой количество столбцов на порядок меньше количества строк, а столбцы не коррелированы, может быть легко подвергнута статистическому анализу, а полученные результаты будут формально корректны.

4. Статистический анализ. Для выявления возможных взаимосвязей между клиническими данными и характеристиками изображений был применен корреляционный анализ, включавший в себя вычисление стандартных однофакторных корреляций Пирсона между клиническими показателями и полученными в результате работы PCA главными компонентами. Из числа всех PC были отобраны те, которые показали статистически значимые корреляции с ЛУ. Для анализа надежности и правомерности найденных взаимосвязей был применен более глубокий статистический анализ, включающий нахождение взаимных корреляций медицинских данных, вычисление взаимных корреляций PC, которые были получены для разных типов дескрипторов, и их корреляций с медицинскими показателями, а также многофакторный корреляционный анализ.

5. Визуализация релевантных участков изображений. Использованный в работе метод поиска взаимосвязей позволяет «проецировать» выбранные PC, коррелирующие с ЛУ, обратно на элементы дескриптора и далее – на исходные изображения с целью выделения соответствующих ключевых структурных образований [8]. Такая техника позволяет визуализировать участки изображений, которые являются причиной найденных корреляций. Это несколько облегчает задачу интерпретации выявленных закономерностей. Алгоритм визуализации таких релевантных участков включает в себя шаги, перечисленные ниже:

Шаг 1. Для каждого элемента дескриптора (матрицы встречаемости) вычисляется коэффициент корреляции его значения с отобранной для визуализации PC.

Шаг 2. Для выбранного изображения создаются два массива прямоугольной формы, по размерам равные исходному изображению, с нулевыми начальными значениями. Один массив предназначен для хранения информации о положительных корреляциях, другой – для отрицательных.

Шаг 3. При помощи алгоритма, аналогичного алгоритму подсчета матриц встречаемости, перебираются все пары (тройки) пикселов исходного изображения. Естественно, что каждая пара (тройка) пикселов соответствует определенному элементу дескриптора. В ячейках описанных выше массивов, которые покоординатно соответствуют пикселам рассматриваемого изображения, производится суммирование квадратов коэффициентов корреляций соответствующего элемента дескриптора с отобранной PC. Суммирование производится отдельно для случаев положительной и отрицательной корреляций.

Шаг 4. Полученные массивы, соответствующие положительным и отрицательным корреляциям, преобразовываются в бинарные маски с использованием подходящего порога. В данной работе значения порогов выбирались вручную.

Шаг 5. Участки изображений, соответствующие найденным корреляциям с выбранной РС, визуализируются (подсвечиваются) согласно полученным бинарным маскам.

Приведенный алгоритм дает возможность наблюдать области изображений, ответственные за положительные и отрицательные корреляции с выбранной характеристикой.

3. Статистический анализ

В данном разделе приводятся результаты процедуры поиска и детального статистического анализа взаимосвязей между ЛУ и признаками изображений.

1. Получение главных компонент и корреляционный анализ данных. Для описания текстурных свойств КТ-изображений были использованы дескрипторы типа *IIGGAD*, различающие 12 градаций по интенсивностям (*I*), 4 градации по величинам градиента (*G*), 8 градаций по углам между градиентами (*A*) и рассматривающие пары точек на расстояниях $D = \{1, 2, 3, 4\}$. Суммарное количество элементов в дескрипторе такого типа составляло 73 728, большинство из которых, как и следовало ожидать, были нулевыми. При помощи метода главных компонент были получены и отобраны пять некоррелированных РС (далее CT PC), удовлетворивших критерию отбора Кайзера [10]. В случае рентгеновских изображений использовались дескрипторы типа *IIID*, различающие восемь градаций по интенсивностям (*I*), вычисляющие встречаемость троек пикселов на взаимных расстояниях $D = \{1, 3, 5\}$ с общим количеством элементов 1 536. В результате работы PCA были получены шесть некоррелированных главных компонент (X-ray PC).

Результаты анализа корреляционных связей межу полученными характеристиками изображений обоих типов (CT PC и X-ray PC) и имеющимися в базе клиническими данными приведены в табл. 1. В таблице отображены признаки изображений, значимо коррелированные с медицинскими показателями. В ячейках представлены значения коэффициентов корреляции r и соответствующих им показателей статистической значимости *p-value*. Полужирным шрифтом выделены значения, соответствующие статистически значимым корреляциям с $p < 0,01$. Как видно из таблицы, значимые корреляции с ЛУ имеют место для третьей главной компоненты CT PC3 ($r = 0,34$, $p = 0,00038$) в случае КТ-изображений и шестой компоненты X-ray PC6 ($r = 0,31$, $p = 0,0010$) для рентгеновских снимков легкого. Признаки CT PC3 и X-ray PC6 также показали следующие статистически значимые корреляции с медицинскими данными: CT PC3 – вес пациента ($r = 0,29$, $p = 0,0025$), CT PC3 – объем пораженной части легкого ($r = 0,34$, $p = 0,00032$) и X-ray PC6 – признак повторного лечения ($r = –0,29$, $p = 0,0024$). Анализ взаимных корреляций медицинских данных показал, что среди клинических характеристик, отобранных для анализа, два показателя значимо коррелируют с показателем ЛУ: признак повторного лечения ($r = –0,31$, $p = 0,0013$) и объем пораженной части легкого ($r = –0,25$, $p = 0,011$). Проверка возможных взаимосвязей между выбранными признаками рентгеновских (X-ray PC6) и КТ-изображений (CT PC3) показала, что эти характеристики взаимно не коррелированы ($r = –0,0017$, $p = 0,986$). Данный факт, вероятно, можно объяснить существенно разными принципами формирования рентгеновских и КТ-изображений.

Основываясь лишь на данных однофакторного корреляционного анализа, нельзя сделать достоверный вывод о наличии на рассматриваемых изображениях признаков ЛУ. В частности, найденные взаимосвязи могли быть обусловлены:

– наличием некоторого причинного связующего фактора, коррелированного с ЛУ и проявляющегося на изображениях (так называемая «проблема третьей переменой» [11]); в этом случае зависимость между ЛУ и характеристиками изображений лишь косвенная;

– случайностью, связанной с организацией данных; в этом случае найденные зависимости не будут выявлены при проведении такого же анализа на подвыборке из всех имеющихся данных;

– случайностью, связанной со спецификой способа количественного описания изображений; в этом случае полученный результат будет сильно чувствителен к изменению типа используемого дескриптора.

Таблица 1

Корреляция характеристик изображений и клинических данных

Клинический показатель	CT PC1	CT PC2	CT PC3	CT PC5	X-ray PC2	X-ray PC4	X-ray PC5	X-ray PC6
Возраст	***r* = 0,33** ***p* = 0,0004**	*r* = –0,14 *p* = 0,15	*r* = 0,062 *p* = 0,53	*r* = 0,086 *p* = 0,38	*r* = 0,18 *p* = 0,063	*r* = 0,080 *p* = 0,41	*r* = 0,051 *p* = 0,60	*r* = –0,10 *p* = 0,30
Пол	***r* = 0,35** ***p* = 0,0002**	*r*=0,18 *p*=0,064	*r*=0,11 *p*=0,25	*r*=–0,24 *p*=0,014	***r*=0,24** ***p*=0,012**	*r*=–0,13 *p*=0,19	*r*=0,15 *p*=0,12	*r*=–0,092 *p*=0,35
Рост	***r* = –0,47** ***p* < 0,0001**	*r*=–0,18 *p*=0,059	*r*=–0,11 *p*=0,25	*r*=0,19 *p*=0,047	*r*=–0,16 *p*=0,10	*r*=0,13 *p*=0,17	*r*=–0,22 *p*=0,023	*r*=–0,013 *p*=0,89
Вес	*r* = –0,13 *p* = 0,18	***r*=–0,45** ***p*<0,0001**	***r*=–0,29** ***p*=0,0020**	***r*=0,31** ***p*=0,0012**	*r*=0,18 *p*=0,070	***r*=0,37** ***p*=0,0001**	***r*=–0,30** ***p*=0,002**	*r*=–0,015 *p*=0,88
Повторное лечение	*r* = –0,06 *p* = 0,53	*r*=0,046 *p*=0,64	*r*=0,14 *p*=0,15	*r*=0,001 *p*=0,92	*r*=–0,15 *p*=0,12	*r*=–0,005 *p*=0,96	*r*=–0,064 *p*=0,52	***r*=–0,29** ***p*=0,0024**
Наличие симптомов	*r* = 0,13 *p* = 0,18	*r*=0,091 *p*=0,36	*r*=0,24 *p*=0,013	*r*=0,007 *p*=0,95	*r*=–0,15 *p*=0,12	*r*=0,006 *p*=0,95	*r*=0,041 *p*=0,68	*r*=–0,11 *p*=0,24
Объем легкого	***r* = –0,60** ***p* < 0,0001**	***r*=–0,38** ***p*=0,0001**	*r*=–0,16 *p*=0,11	***r*=0,41** ***p*<0,0001**	*r*=–0,23 *p*=0,017	*r*=0,13 *p*=0,17	*r*=–0,14 *p*=0,14	*r*=–0,056 *p*=0,57
Объем очагов поражения	*r* = 0,11 *p* = 0,27	*r*=–0,016 *p*=0,87	***r*=0,34** ***p*=0,0003**	*r*=–0,015 *p*=0,88	*r*=0,014 *p*=0,89	*r*=–0,09 *p*=0,38	*r*=–0,075 *p*=0,44	*r*=0,038 *p*=0,69
Лекарственная устойчивость	*r* = –0,15 *p* = 0,11	*r*=–0,13 *p*=0,18	***r*=0,34** ***p*=0,0004**	*r*=0,006 *p*=0,95	*r*=–0,094 *p*=0,34	*r*=–0,07 *p*=0,49	*r*=–0,036 *p*=0,71	***r*=–0,31** **p=0,0010**
Уменьшение объема легкого	*r* = 0,041 *p* = 0,68	*r*=–0,0055 *p*=0,96	*r*=0,24 *p*=0,015	*r*=–0,026 *p*=0,79	*r*=–0,090 *p*=0,36	*r*=0,086 *p*=0,38	*r*=–0,081 *p*=0,41	*r*=–0,14 *p*=0,15
Распространенность процесса	*r* = 0,095 *p* = 0,33	*r*=0,048 *p*=0,62	*r*=0,15 *p*=0,12	*r*=–0,18 *p*=0,061	*r*=0,23 *p*=0,019	*r*=0,002 *p*=0,98	*r*=–0,15 *p*=0,12	*r*=–0,11 *p*=0,25
Наличие полостей распада	*r* = –0,071 *p* = 0,47	*r*=0,14 *p*=0,15	*r*=0,029 *p*=0,77	*r*=–0,15 *p*=0,13	*r*=–0,096 *p*=0,33	*r*=–0,076 *p*=0,438	*r*=–0,077 *p*=0,43	*r*=0,043 *p*=0,66
Количество полостей распада	*r* = –0,053 *p* = 0,59	*r*=0,18 *p*=0,066	*r*=–0,063 *p*=0,52	*r*=–0,21 *p*=0,032	*r*=–0,11 *p*=0,25	*r*=–0,028 *p*=0,78	*r*=–0,015 *p*=0,88	*r*=0,004 *p*=0,97
Синдром диссеминации	*r* = 0,039 *p* = 0,70	*r*=–0,015 *p*=0,88	*r*=0,15 *p*=0,12	*r*=0,085 *p*=0,38	*r*=0,080 *p*=0,42	*r*=–0,11 *p*=0,26	*r*=0,034 *p*=0,73	*r*=0,13 *p*=0,19

2. Анализ взаимосвязей. Множественная регрессия. Для более детального выявления причинных взаимосвязей был проведен многофакторный анализ. Для этого использовалась модель множественной регрессии вида $U \sim CT + Xray + Y$, в которой в качестве зависимой переменной U выступала ЛУ, в качестве независимых переменных CT и $Xray$ – коррелированные с ней признаки изображений (CT PC3 и X-ray PC6), а в качестве дополнительной независимой переменной Y по очереди выступал каждый из оставшихся клинических признаков. Для каждого прогона модели вычислялись значения p-value, показывающие статистическую значимость частных корреляций зависимой переменной с каждой из независимых. Из табл. 2 видно, что для каждого из клинических признаков Y уровень статистической значимости частных корреляций признаков КТ-изображений с U был не хуже $p = 0,0029$. Это достигалось при подстановке вместо Y объема очагов поражения. Для рентгеновских изображений аналогичные частные корреляции были не хуже $p = 0,0042$ при использовании в качестве дополнительного клинического параметра Y признака повторного лечения. В то же время среди медицинских данных наиболее значимую частную корреляцию с ЛУ U показал признак повторного лечения $(p = 0,032)$.

Таблица 2

Результаты многофакторного корреляционного анализа.

Переменная *Y*	CT *p*-value	Xray *p*-value	Y *p*-value
Возраст	0,00013	0,00024	0,13
Пол	0,00026	0,00053	0,68
Рост	0,00015	0,00047	0,38
Вес	0,000050	0,00053	0,090
Повторное лечение	0,00046	0,0042	0,032
Наличие симптомов	0,00060	0,00062	0,42
Объем легкого	0,000090	0,00052	0,14
Объем очагов поражения	0,0029	0,00027	0,071
Уменьшение объема легкого	0,00012	0,00032	0,34
Распространенность процесса	0,00026	0,00051	0,88
Наличие полостей распада	0,00022	0,00036	0,29
Количество полостей распада	0,00015	0,00040	0,27
Синдром диссеминации	0,00053	0,00018	0,070

3. Кросс-проверка. Для проверки устойчивости полученных взаимосвязей между признаками изображений и ЛУ была применена процедура кросс-проверки [12], также известная под названием Cross-Validation или скользящий контроль. Процедура кросс-проверки применялась в двух вариантах: контроль по *k* блокам (*k*-fold cross-validation) с количеством блоков данных *k*, равным 10, и контроль по циклическому исключению одного объекта (leave-one-out cross-validation). В обоих типах кросс-проверки производится разбиение всего массива данных на обучающую и контрольную выборки, причем каждый объект входит в контрольную выборку строго один раз. В случае контроля по *k* блокам размер контрольной выборки приблизительно равен $1/k$ от размера всей выборки. В случае циклического исключения одного объекта контрольная выборка состоит лишь из одного объекта, а именно того, который был исключен из обучающей выборки.

Суть процедуры кросс-валидации заключается в проверке повторяемости найденных закономерностей на части исследуемой выборки изображений (пациентов). Указанная процедура в данном случае состоит из следующих шагов:

Шаг 1. К обучающей выборке применялся метод главных компонент. Запоминались получившиеся в результате работы метода средние значения элементов дескриптора (PC0) и факторные координаты переменных (матрица нагрузок в PCA, матрица поворота, loadings).

Шаг 2. Вычислялись корреляции полученных главных компонент с ЛУ. Запоминался номер главной компоненты с максимальным модулем коэффициента корреляции.

Шаг 3. Главные компоненты для контрольной выборки были получены не с помощью нового вызова метода PCA, а с помощью матрицы нагрузок и средних значений элементов дескриптора PC0, вычисленных в работе PCA на обучающей подвыборке. Для этого из данных контрольной выборки вычитались средние значения элементов дескриптора PC0, после чего полученная матрица чисел умножалась на матрицу нагрузок (факторные координаты переменных).

Шаг 4. Из вычисленных главных компонент контрольной выборки была отобрана PC с тем же номером, что и у главной компоненты обучающей выборки, максимально коррелированной с ЛУ.

Шаг 5. Для каждого объекта (пациента) значение отобранной PC запоминалось.

Шаг 6. После того как значения отобранной PC были подобным образом найдены для всех пациентов (перебор всех контрольных выборок завершен), вычислялся коэффициент корреляции этих значений с ЛУ.

Поскольку в случае контроля по *k* блокам разбиение на блоки происходило случайно, для получения стабильных результатов вся процедура кросс-проверки повторялась 100 раз и выбиралось медианное значение получившихся коэффициентов корреляции *r* и соответствующее ему значение доверительной вероятности *p*. Результаты кросс-проверок для рентгеновских и КТ-изображений (табл. 3) показывают, что связи признаков КТ-изображений с ЛУ сохраняют прежний уровень статистической значимости при обоих типах кросс-проверки. Корреляции

признаков рентгеновских изображений с ЛУ проявляют меньшую устойчивость к вариации обучающей выборки по сравнению с КТ-изображениями. Это можно объяснить большим вкладом шумовых факторов в последнюю (шестую по счету) отобранную главную компоненту признаков рентгеновских изображений.

Таблица 3

Результаты кросс-проверок

Тип изображений	Контроль по *k* блокам	Контроль по отдельным объектам
КТ	r = 0,32 p = 0,00039	r = 0,32 p = 0,00038
Рентген	r = 0,21 p = 0,013	r = 0,26 p = 0,0034

4. Зависимость от типа дескриптора изображения. Для проверки устойчивости полученных зависимостей к изменениям способа количественного описания изображений базовая процедура поиска взаимосвязей, включающая вычисление дескрипторов, PCA и однофакторный корреляционный анализ, была повторена несколько раз с использованием дескрипторов, различающихся типом и/или параметром бининга (количеством градаций используемых параметров пары/тройки пикселов). В случае КТ-изображений у дескрипторов типа *IIGGAD* варьировались следующие параметры: количество градаций интенсивности (*I*), количество градаций величины градиента (*G*), количество градаций величины угла между градиентами (*A*), а также рассматриваемые расстояния между вокселами (*D*). В случае когда количество градаций величин градиентов (*G*) и углов (*A*) равно единице, дескриптор типа *IIGGAD* превращается в его упрощенную версию, обозначаемую *IID*, в которой градиенты яркости изображения не учитываются. Для рентгеновских изображений варьировались следующие параметры дескриптора *IIID*: количество градаций интенсивности (*I*) и набор взаимных расстояний между пикселами (*D*).

Таблица 4

Корреляция характеристик КТ-изображений с ЛУ для разных типов дескрипторов

Параметры дескриптора *IIGGAD*				Номер отобранной PC	Коэффициент корреляции PC с ЛУ	*p*-value
I	*G*	*A*	*D*			
12	4	8	{1, 2, 3, 4}	3	0,34	0,00038
12	4	4	{1, 2, 3, 4, 5, 6}	4	0,35	0,00022
12	4	8	{1, 2}	3	0,32	0,00071
12	4	8	{1}	3	0,33	0,00053
8	4	8	{1, 2, 3, 4}	3	0,27	0,0052
8	4	4	{1, 2, 3, 4}	3	0,27	0,0055
12	1	1	{1, 2, 3, 4}	3	0,27	0,0055
8	1	1	{1, 2, 3, 4}	3	0,31	0,0011

Таблица 5

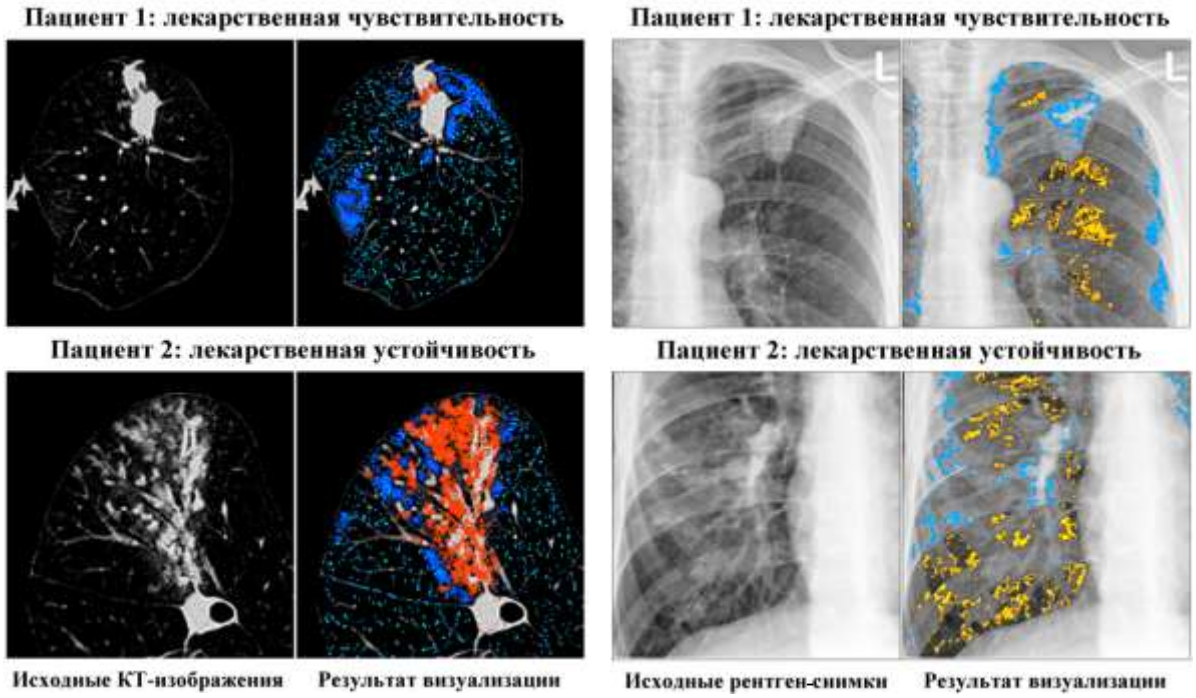
Корреляции характеристик рентгеновских изображений с ЛУ для разных типов дескрипторов

Параметры дескриптора *IIID*		Номер отобранной PC	Коэффициент корреляции PC с ЛУ	*p*-value
I	*D*			
8	{1, 3, 5}	6	0,31	0,00099
8	{2, 6, 10}	6	0,30	0,0017
8	{4, 12, 20}	6	0,32	0,00089
16	{1, 3, 5}	6	0,26	0,0071
16	{2, 6, 10}	6	0,28	0,0042
16	{4, 12, 20}	6	0,28	0,0033

В табл. 4 и 5 приведены результаты проведения базовой процедуры поиска закономерностей с использованием различных типов дескрипторов. Из таблиц видно, что результаты поиска закономерностей слабо зависят от типа дескрипторов. Это выражается в почти везде одинаковом порядковом номере главной компоненты, коррелированной с ЛУ. Также видно, что статистическая значимость найденных корреляций во всех случаях не хуже $p = 0,0055$ для компьютерной томографии и $p = 0,0071$ для рентгена.

4. Визуализация релевантных участков изображений

На рисунке показаны результаты подсветки участков рентгеновских и КТ-изображений, соответствующих полученным ключевым признакам изображений CT PC3 и X-ray PC6, которые коррелированы с ЛУ. Процедура подсветки была осуществлена при помощи алгоритма, описанного выше. Оттенками красного на изображениях выделены участки, соответствующие положительной корреляции с выбранным признаком и, следовательно, с ЛУ, а оттенками синего – соответствующие отрицательным корреляциям. Для визуализации были отобраны два пациента с характерными значениями полученных характеристик изображений: для пациента 1 с лекарственно чувствительным туберкулезом значения признаков изображений имели значения CT PC3 = –0,0065 и X-ray PC6 = 0,000073, для пациента 2 с множественной лекарственной устойчивостью эти значения были CT PC3 = 0,0071 и X-ray PC6 = –0,0034 соответственно. Значения элементов дескрипторов, соответствующих подсвеченным областям, имеют наибольшие статистические различия для лекарственно-чувствительных и лекарственно-устойчивых пациентов, соответствующие $p = 0,000062$ для КТ-изображений и $p = 0,099$ для рентгеновских. Для более полного понимания полученных результатов требуется глубокая медико-биологическая интерпретация полученных зависимостей, однако обсуждение таких вопросов выходит за рамки данной работы.



а) *б)*

Визуализация участков изображений, соответствующих найденным корреляциям с ЛУ: *а)* КТ-изображения; *б)* рентгеновские изображения. Оттенками красного подсвечены области, соответствующие положительным корреляциям с ЛУ, оттенками синего – отрицательным

Заключение

Результаты исследований, представленные в данной статье, указывают на наличие статистически значимых взаимосвязей между вычисленными количественными признаками рентгеновских и КТ-изображений и степенью ЛУ пациентов, больных туберкулезом легких. На основании имевшихся в распоряжении медицинских данных не удалось выявить какого-либо дополнительного фактора, тривиальным образом объясняющего полученные корреляции и указывающего на их косвенность (транзитивность).

Найденные зависимости показали свою устойчивость при проведении кросс-проверок, где данные, необходимые для вычисления нужных признаков изображений, брались из подвыборки всех данных.

Полученные результаты требуют дальнейшей медико-биологической интерпретации, а также исследования возможности распознавания (предсказания) ЛУ туберкулеза каждого конкретного пациента на основании признаков изображений и базовых данных о пациенте, получение которых не требует проведения сложных биомедицинских анализов. Задача распознавания может решаться путем применения известных классификаторов к данным, представленным в настоящей работе.

Работа выполнена при частичной финансовой поддержке Национального института аллергических и инфекционных заболеваний США в рамках CRDF проектов BOB1-31055-MK-11 и BOB1-31120-MK-13

Список литературы

1. Ferguson, L.A. Multidrug-resistant and extensively drug-resistant tuberculosis : The new face of an old disease / L.A. Ferguson, J. Rhoads // Journal of American Academy Nurse Practitioners. – 2009. – Vol. 21, № 11. – P. 603–609.

2. Chiang, C.Y. Drug-resistant tuberculosis: Past, present, future / C.Y. Chiang, R. Centis, G.B. Migliori // Respirology. – 2010. – Vol. 15, № 3. – P. 413–432.

3. Multidrug-resistant tuberculosis in Belarus: the size of the problem and associated risk factors / A. Skrahina [et al.] // Bulletin of the World Health Organization. – 2013. – Vol. 91. – P. 36–45.

4. Belarus Tuberculosis Portal [Электронный ресурс]. – Режим доступа : http://tuberculosis.by. – Дата доступа : 20.02.2013.

5. Radiological Findings of Extensively Drug-Resistant Pulmonary Tuberculosis in Non-AIDS Adults : Comparisons with Findings of Multidrug-Resistant and Drug-Sensitive Tuberculosis / J. Cha [et al.] // Korean Journal of Radiology. – 2009. – Vol. 10. – P. 207–216.

6. Computed Tomography Features of Extensively Drug-Resistant Pulmonary Tuberculosis in Non-HIV-Infected Patients / E.S. Lee [et al.] // Journal of Computer Assisted Tomography. – 2010. – Vol. 34. – P. 559–563.

7. Three-dimensional texture analysis of MRI brain datasets / V.A. Kovalev [et al.] // IEEE Transactions on Medical Imaging. – 2001. – Vol. 20, № 5. – P. 424–433.

8. A method for identification and visualization of histological image structures relevant to the cancer patient conditions / V.A. Kovalev [et al.] // Proc. of the 27-th Int. congress on Computer Analysis of Images and Patterns (CAIP-2011). – Spain, 2011. – Vol. 6854, № 1. – P. 460–468.

9. Kovalev, V.A. Gender and age effects in structural brain asymmetry as measured by MRI texture analysis / V.A. Kovalev, F. Kruggel, D.Y. von Cramon // NeuroImage. – 2003. – Vol. 19. – P. 896–905.

10. Kaiser, H.F. The application of electronic computers to factor analysis / H.F. Kaiser // Educational and Psychological Measurement. – 1960. – Vol. 20. – P. 141–151.

11. Sapsford, R. Data Collection and Analysis / R. Sapsford, V. Jupp. – London : Sage, 2006. – 332 p.

12. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection / R. Kohavi // Proc. of the Fourteenth Int. Joint Conf. on Artificial Intelligence (IJCAI'95). – USA, 1995. – Vol. 2. – P. 1137–1143.

Поступила 29.03.2013

[1]*Объединенный институт проблем информатики НАН Беларуси, Минск, Сурганова, 6 e-mail: vassili.kovalev@gmail.com*

[2]*Республиканский научно-практический центр пульмонологии и фтизиатрии, Минск, Долгиновский тракт, 157*

V.A. Kovalev, V.A. Liauchuk, I.U. Safonau, A.U. Tarasau

EXAMINATION OF POSSIBLE LINKS BETWEEN DRUG RESISTANCE AND MORPHOLOGY OF LUNG IMAGES OF TUBERCULOSIS PATIENTS

The purpose of this paper is to present the results of an exploratory study of possible correlations between the drug resistance and the structural features of CT and X-ray images of lungtuberculosis patients. A multi-step procedure is suggested which includes calculation of textural image features, extracting their principal components, correlating them to patients' clinical data and mapping the significant principal components back to image descriptor elements and then to the corresponding image structures they found to be linked with. The results of a detailed statistical analysis of the revealed links between the drug resistance and the image features are presented. The analysis includes finding one-factor correlations, performing multivariate regression analysis and cross-validation.