На правах рукописи
УДК 519.7:616-053.2

Татаринцев Павел Борисович

Разработка систем диагностики, дифференциальной диагностики и прогнозирования заболеваний методами многомерного статистического анализа

05.13.01 – системный анализ, управление и обработка информации

Автореферат
диссертации на соискание ученой степени
кандидата технических наук.

Барнаул – 2006

Работа выполнена на кафедре «Дифференциальные уравнения» Алтайского государственного университета

Научные руководители: кандидат ф.-м. наук, доцент
Семенов Сергей Петрович

Официальные оппоненты: доктор т.н., профессор
Жмудяк Леонид Моисеевич,

Ведущая организация: Омский государственный университет

Защита диссертации состоится 6 июля 2006г., в 12^{00} часов, на заседании регионального диссертационного совета КМ 212.004.01. в Алтайском государственном техническом университете по адресу: 656038, г. Барнаул, пр. Ленина, 46.

С диссертацией можно ознакомиться в библиотеке Алтайского государственного технического университета им. И.И. Ползунова.

Автореферат разослан 6 июня 2006г.

Ученый секретарь регионального
диссертационного совета
к.э.н., доцент А.Г. Блем

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Процессы диагностики, дифференциальной диагностики и прогнозирования заболеваний имеют решающее значение в деятельности врача. Только своевременно установленный диагноз позволяет выбрать адекватный метод лечения. При этом на выбор стратегии лечения влияет также оценка риска неблагоприятного исхода заболевания для данного пациента.

Точность диагноза и быстрота, с которой его можно поставить, зависят от очень многих факторов: от состояния больного, от имеющихся данных о симптомах, признаках заболевания и результатах лабораторных анализов, от общего объема медицинской информации о наблюдении таких симптомов при самых различных заболеваниях и от квалификации самого врача. Исходя из этих знаний о процессе диагностики, можно определить условия, при которых диагноз может быть поставлен максимально быстро и точно.

Многочисленные публикации последних лет, приводят факты некорректного применения математических методов в медико-биологических исследованиях и свидетельствуют о том, что процедуры диагноза не согласуются с опытом врачей. Поэтому задачи разработки алгоритмов диагностики необходимо решать совместно медикам и математикам. Обычно совместное исследование медико-биологических данных начинается с формализации цели этого исследования. Причем, формализовать цель нужно так, чтобы она была понятна, как экспериментатору, который эти данные собирает, так и для тех, кто их будет анализировать с помощью математических методов. Опыт формализации целей медицинских исследований можно получить только при систематическом взаимодействии двух наук медицины и математики.

Для медицинской практики врача на первом плане всегда стоят три задачи: диагностика, дифференциальная диагностика и прогнозирование.

- Задача диагностики — правильно определить заболевание и правильно назначить лечение;
- Задача дифференциальной диагностики — правильно определить нозологическую форму заболевания;
- Задача прогнозирования — дать прогноз исхода лечения заболевания и предупредить осложнения.

Следует подчеркнуть, что до настоящего времени диагностика многих заболеваний осуществляется дорогостоящими методами, нередко обладающими низкой чувствительностью, требующими больших затрат времени и средств. При этом зачастую врачу приходится быстро принимать решения, не дожидаясь результатов анализов, опираясь лишь на данные клинических показателей и свой личный опыт. Такого рода заболевания являются *слабо диагностируемыми*. Поэтому являются актуальными задачи систематизации, создания банка данных клинических показателей и создания способов решения задач диагностики, дифференциальной диагностики и прогнозирования слабо диагностируемых заболеваний с помощью математических методов.

Анализ литературы показывает, что математический подход к задачам диагностики и прогнозирования позволяет отвлечься от обсуждения конкретных действий врача при том или ином заболевании и перейти к изучению вопросов алгоритмизации диагностического процесса, а также разработать принципы построения диагностических информационных систем и конкретных прикладных программ для реализации диагностических алгоритмов. Данный подход позволяет сократить затраты средств и времени на диагностику заболевания, увеличить точность диагностики, способствует оптимизации процесса принятия решений в медицине.

Поэтому проблема разработки формальных подходов к исследованию медико-биологических данных с учетом их специфики для решения задач диагностики и прогнозирования конкретных слабо диагностируемых заболеваний в настоящее время является актуальной.

Цель исследования. Разработка математических методов обработки слабоструктурированных многомерных данных с учетом особенностей медико-биологических закономерностей и их использование при разработке диагностических и прогностических алгоритмов и систем для конкретных слабо диагностируемых заболеваний.

В соответствии с поставленной целью были определены следующие задачи:

1. Разработка концептуальной модели процесса исследования медико-биологических данных для решения задач медицинской диагностики и прогнозирования слабо диагностируемых заболеваний.
2. Разработка логической схемы процесса исследования медико-биологической информации и построения систем диагностики и прогнозирования заболеваний.
3. Разработка критериев отбора информативных признаков из исходного массива данных.
4. Разработка алгоритма построения систем диагностики, дифференциальной диагностики и прогнозирования конкретных слабо диагностируемых заболеваний.
5. Апробация разработанного алгоритма построения систем диагностики и прогнозирования заболеваний.
6. Экспериментальное подтверждение работоспособности разработанных систем диагностики и прогнозирования заболеваний.

Объект исследования. Системы диагностики, дифференциальной диагностики и прогнозирования слабо диагностируемых заболеваний.

Предметом исследования являются математические методы исследования медико-биологических данных, построения систем диагностики, дифференциальной диагностики и прогнозирования слабо диагностируемых заболеваний с помощью статистических методов.

Методы исследования. При решении поставленных задач применялись методы системного анализа, математической статистики, многомерного статистического анализа, нейросетевого моделирования, компьютерные технологии.

Научная новизна предлагаемой работы заключается в достижении следующих научных результатов:

1. Разработан комплексный метод исследования медицинских данных, состоящий из процедур статистической обработки данных с привлечением экспертной информации, учитывающий специфику слабо структурированных медико-биологических данных.
2. В рамках разработанного метода проведены исследования больших массивов медико-биологической информации и разработаны алгоритмы и компьютерные программы, предназначенные для использования в процессе диагностики следующих слабо-диагностируемых заболеваний: описторхоза, панкреатита, абдоминального сепсиса, а также для прогнозирования развития климактерического синдрома у женщин перименопаузальном периоде апробированные на практике.
3. Разработаны алгоритмы и программы, которые могут быть использованы на этапах предварительного анализа данных, позволяющие более точно оценивать следующие характеристики исследуемых нечисловых признаков: точечная и интервальная оценка вероятности в схеме Бернулли, вероятность ошибки I-го рода в точном критерии Фишера при анализе таблиц сопряженности 2x2.
4. Даны оценки диагностической значимости комплексов признаков в Алтайском крае на основе имеющихся статистических данных.

Теоретическая значимость результатов работы. Предложена концептуальная схема метода разработки систем диагностики, дифференциальной диагностики и прогнозирования слабо диагностируемых заболеваний.

Практическая значимость результатов работы. Выявлены дискриминантные переменные для использования в системах диагностики описторхоза, панкреатита, сепсиса в Алтайском крае. Выявлены факторы риска развития климактерического синдрома у женщин в перименопаузальном периоде. Разработаны эффективные алгоритмы диагностики и прогнозирования вышеуказанных заболеваний в Алтайском крае, которые внедрены в нескольких медицинских учреждениях края и используются в медицинской практике.

Основные положения, выносимые на защиту.

1. Концептуальная модель и логическая схема процесса исследования медико-биологических данных для решения задач медицинской диагностики и прогнозирования слабо диагностируемых заболеваний
2. Результаты сравнения работоспособности различных математических моделей при решении задач диагностики, дифференциальной диагностики и прогнозирования.
3. Алгоритмы построения и практические результаты применения комплексных систем диагностики, дифференциальной диагностики и прогнозирования конкретных слабо диагностируемых заболеваний.

Апробация результатов. Основные положения и отдельные результаты исследования докладывались и обсуждались на VI-й краевой конференции по математике (Барнаул, 2001), I Международной юбилейной конференции «Актуальные проблемы инфектологии и паразитологии» (Томск, 2001),V-й краевой

конференции по математике (Барнаул, 2002), научно-практической конференции хирургов Сибирского региона (Барнаул – Белокуриха, 2002), Всероссийской конференции «Современные технологии лабораторной диагностики нового столетия» (Москва, 2002), .

Публикации. По теме диссертации опубликовано 12 печатных работ, в том числе 3 статьи в периодических журналах, 7 тезисов докладов на конференциях. Получены 2 свидетельства о регистрации программных продуктов.

Структура и объем работы. Диссертация состоит из введения, двух глав, заключения, списка литературы из 80 источников, 15 приложений. Общий объем работы составляет 120 страниц, содержит 10 рисунков, 5 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность исследования, определены цели, задачи, объект, предмет, методы исследования. Раскрыты научная новизна, теоретическая и практическая ценность, сформулированы положения, выносимые на защиту.

В первой главе проведен анализ процесса медицинской диагностики. Рассмотрены проблемы анализа медико-биологических данных и построения систем медицинской диагностики с помощью различных методов. Изложены основные идеи подхода с использованием многомерного статистического анализа, в рамках которого проводится алгоритмизация процесса диагностики и построение поэтапной логической схемы указанного процесса, а также выделение необходимых для ее реализации математических моделей и методов.

Медицинская диагностика — это сложный многоэтапный процесс, возникающий в системе «больной-врач», упрощенная схема которой изображена на рис. 1.

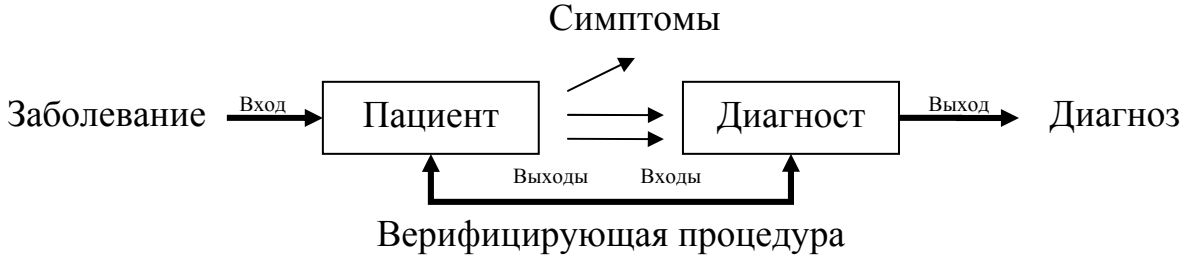

Рис. 1. Схема системы «больной врач»

Диагностика как наука, развивается по трем направлениям исследований данной системы:

1. Изучение методов наблюдения и исследования больного
2. Изучение диагностического значения симптомов болезней
3. Изучение особенностей мышления при распознавании заболевания

Наибольшее внимание в диссертации уделено второму направлению. Основной задачей здесь является разработка диагностических алгоритмов. В диссертации выяснено, какие диагностические алгоритмы используются в настоящее время. Основные из них: диагностические таблицы и диагностические де-

ревья. Данные алгоритмы обладают целым рядом недостатков по сравнению с алгоритмом, которым пользуются диагносты на практике. Его упрощенная схема приведена на рис. 2.

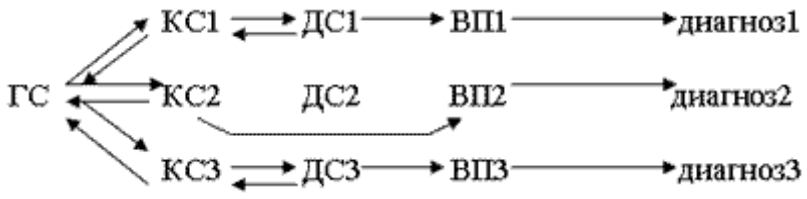

Рис. 2. Схема диагностического алгоритма, используемого диагностами на практике. Здесь ГС – главный симптом, КС – конкурирующий симптом, ДС – дополнительные симптомы, ВП – верифицирующая процедура. Обратные стрелки показывают, что симптомы могут меняться местами

В диссертационной работе в результате анализа литературы показано, что методы многомерного статистического анализа в диагностике заболеваний используются недостаточно широко. Данное положение обусловлено:

1. Трудностями сокращения факторного пространства.
2. Наличием в наборе признаков слабо информативных, а также существенно коррелирующих с другими признаками.
3. Математическими сложностями многомерного статистического анализа при числе признаков десятки и сотни.
4. Кроме того, в многомерном случае возникают сложности дискриминации диагностируемых состояний.

Вместе с тем анализ литературы по многомерному статистическому анализу показывает, что данный метод с успехом применяется в других сферах. Разработка и его применение особенно актуальны в настоящее время в связи с использованием современных диагностических систем (томографов, морфоденситометров, аппаратов ИВЛ и др.) с огромным количеством признаков.

Для преодоления этих трудностей предлагается метод диагностики, дифференциальной диагностики и прогнозирования заболеваний с привлечением экспертной информации не только на стадии многомерного статистического анализа, но и процессе настройки диагностических процедур. Предлагаемая информационная технология включает несколько этапов:

1. Разведочный анализ
2. Многомерный статистический анализ
3. Разработка алгоритма диагностики

Во второй главе излагаются основные этапы процесса исследования медико-биологических данных, конечной целью которого является разработка системы диагностики или прогнозирования. При этом выделяются методические задачи, которые возникают при реализации разработанной концептуальной схемы, и указываются математические инструменты (модели и методы), необходимые для решения этих задач.

Как правило, результаты медицинских исследований представляют собой таблицу, в которой столбцы соответствуют некоторым измеренным показателям, а строки соответствуют наблюдаемым пациентам. Далее столбцы таблицы

будем называть **переменными**, а строки — **случаями**. Количество переменных и случаев может составлять от нескольких десятков до нескольких сотен. В таблице могут присутствовать переменные различной природы: количественные и качественные. Первоначально нужно решить несколько задач, чтобы подготовить таблицу к дальнейшему анализу. В зависимости от природы исследуемых переменных выбираются различные подходы и статистические критерии, иначе результаты могут получиться лишенными смысла или возникнут трудности в их интерпретации. Поэтому в первую очередь производится **классификация** переменных по типам.

Для дальнейшей обработки статистическими методами очень важно, чтобы данные были корректными и не содержали выбросов. Задача второго шага — это **цензурирование** переменных и исправление ошибок, сделанных при занесении данных в компьютер. После выполнения процедуры цензурирования нужно выяснить информационную ценность каждой переменной и **исключить** те переменные, которые либо не несут в себе никакой информации (являются константами), либо дублируют ее (линейно зависимые переменные).

Затем требуется более компактно представить имеющиеся данные, чтобы иметь возможность **описать** и понять их структуру.

Общая схема действий такова:

1. Определение типа каждой переменной.
2. Устранение ошибок и выбросов.
3. Исключение малоинформативных переменных.
4. Описание данных.
5. Проверка различных гипотез о структуре зависимостей.

Решение вышеперечисленных задач производится методами разведочного анализа данных.

При решении задач диагностики и прогнозирования наиболее подходящей моделью для описания данных является дискриминантная модель, когда точки данных относятся к нескольким группам.

В рамках выбранной модели структуры данных для описания структуры зависимостей между переменными может использоваться одна из следующих моделей:

1. Модель независимых переменных.
2. Модель линейно зависимых переменных.
3. Древообразная модель.
4. Факторная модель.
5. Кластерная модель.
6. Иерархическая модель.

Трудоемкость процесса сбора медико-биологических данных или высокая стоимость чаще всего являются непреодолимыми препятствиями для сбора необходимого объема данных за короткий промежуток времени. В таких условиях при описании зависимостей между переменными указанные модели не могут применяться. В качестве решения данной проблемы предлагается максимально избавится от избыточной информации в исходных данных и тем самым уменьшить количество переменных. Процедура выявления и исключения малоин-

формативных признаков в матрицах данных больших размеров должна проводиться в два шага. На первом шаге вычисляются выборочные дисперсии признаков и исключаются признаки с нулевой дисперсией, т.е. признаки являющиеся константами. Низкая информационная ценность таких признаков в матрицах с большим числом признаков очевидна. На втором шаге определяются линейно зависимые признаки путем многократного применения метода наименьших квадратов. Необходимость применения данного метода обусловлена спецификой медико-биологических данных: переменные могут описывать одну и ту же характеристику объекта, но в разных единицах измерения, либо какие-то из признаков являются интегральными для остальных, например, общая площадь эритроцита является суммой площадей выпуклой, вогнутой и переходной его частей.

В процессе применения данного метода получается набор множественных коэффициентов детерминации R^2 для всевозможных сочетаний предикторов и зависимых переменных. Для получения конечного результата необходимо сравнить все полученные коэффициенты между собой и выделить переменные, сильно линейно зависимые от остальных (R^2 превышает некоторое пороговое значение, например, 0.95), Такие переменные подлежит исключению из модели, так как являются источником дублирования информации.

На этапе отсеивания малоинформативных признаков может использоваться экспертная информация о структуре матрицы данных. Использование такой информации способствует уменьшению количества выполняемых расчетов.

Важной частью диагностического исследования является описательная часть. Она позволяет сопоставлять результаты данного исследования с другими аналогичными работами. Кроме этого информация, содержащаяся в описательной части исследования позволяет оценить значимость данной работы и ее актуальность, а также область и рамки применимости сделанных выводов, обобщений и полученных результатов. С ее помощью можно решить, соответствует ли эта генеральная совокупность той, в которой диагност практикует и, соответственно, может ли он применить результаты исследования в своей практике. Математическая часть на этапе описания данных состоит в вычислении описательных статистик. Описательные статистики являются приближенными характеристиками генеральной совокупности.

Определенные трудности возникают при использовании в диагностических процедурах качественных признаков. Так, описывая качественные признаки, исследователь оценивает точность, с которой доли $\tilde{p}_i$, вычисленные по выборке, соответствуют вероятностям p_i во всей генеральной совокупности. В настоящее время наиболее популярны три метода оценки точности долей. Это метод кратчайших доверительных интервалов, метод построения фидуциальных интервалов и метод, опирающийся на центральную предельную теорему (ЦПТ). В любом из методов для оценки точности доли представляет интерес лишь

суммарное число успехов $S = \sum_{i=1}^{n} y_i$, достигнутое в серии из n испытаний Бернулли, независимо от порядка их следования.

Методом Монте-Карло выявлены свойства трех вышеперечисленных методов оценки доверительных интервалов, используемых в известных диагностических процедурах. Генерировались случайные числа с биномиальным распределением с параметрами $p = 0.3$ и $n = 20$. Каждое сгенерированное число использовалось для получения 95% доверительной интервальной оценки для p. Затем подсчитывалось количество попаданий p в полученные интервалы. Пример расчетов приведен в таблице 1. Метод построения кратчайших доверительных интервалов основан на свойствах биномиального распределения. Интервальные оценки для p, полученные данным методом чаще всего оказываются более широкими и не соответствуют требуемой доверительной вероятности — 95%, а, следовательно, являются менее точными. В основе второго метода оценки лежит фидуциальное распределение Фишера. Фидуциальное распределение в данном случае представляет собой бета–распределение. Полученные данным методом интервальные оценки соответствуют требуемой доверительной вероятности — 95%. Кроме того, данный метод приводит к наиболее узким интервальным оценкам. Третий метод из исследуемых является наиболее простым, но в 20% случаев метод не способен оценить доверительный интервал для p.

С точки зрения вычислительной трудности, все три метода легко реализуются с помощью компьютера. Поэтому для оценки точности может быть использован любой из них. Наиболее точным на малых выборках является второй метод, основанный на фидуциальном распределении. На больших выборках ($n > 1000$) удобно использовать третий метод.

Таблица 1. Качество интервальных оценок

Метод	*Количество испытаний*	*Количество попаданий в интервал*	*Частота*	*Особенности интервала*
Кратчайших доверительных интервалов	1000	980	98,0%	Симметричный, доверительная вероятность завышена
Фидуциальных интервалов	1000	953	95,3%	Ассиметричный, доверительная вероятность в норме
ЦПТ	775	759	97,9%	Симметричный, доверительная вероятность завышена

Наиболее часто в статистических исследованиях медико-биологических данных требуется обосновать, что наблюдаемые различия между несколькими выборками являются неслучайными и, следовательно, отражают действительные различия между генеральными совокупностями. В зависимости от типа задачи, поставленной в исследовании — диагностика, дифференциальная диагностика, прогноз или лечение, возникают различные виды сравниваемых генеральных совокупностей.

Для проверки нулевой гипотезы применяют разнообразные статистические критерии, которые можно разделить на три группы: непараметрические, порядковые и параметрические.

В задачах диагностики, дифференциальной диагностики и прогнозирования исхода признак, определяющий принадлежность случая к какой-либо группе больных относится к данным номинального типа. Поэтому задача выявления различий между двумя группами по некоторому номинальному признаку сводится к исследованию зависимости между двумя номинальными переменными.

Точное значение вероятности ошибки первого рода можно вычислить комбинаторными методами. На этом принципе построен точный критерий Фишера, когда $n = m = 2$.

Предлагаемый в диссертационной работе алгоритм собственной разработки позволяет рассчитывать P для N порядка 10^6 и может использоваться в прикладных программах вместо приближенного критерия χ^2 при анализе таблиц сопряженности 2x2.

В диссертации рассмотрена математическая постановка задачи диагностики. Предположим, что рассматривается ограниченная группа g различных заболеваний $E = (G_1, G_2\ldots, G_g)$, и что каждый больной страдает только одним из них (т.е. события $G_1, G_2\ldots, G_g$ несовместны). Допустим также, что имеется список из n признаков, симптомов или результатов лабораторных анализов $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Для простоты будем считать, что все симптомы дискретны, т.е. каждый из них относится к одному из двух или нескольких конкретных классов.

Для того чтобы немедленно начать лечение, и, возможно, назначить специальные дополнительные проверки, необходимо поставить предварительный диагноз. Выражаясь математическим языком, врачу нужно знать вероятность каждого заболевания при данном симптомокомплексе, т.е. $P(G_k \mid \mathbf{x})$. Поэтому здесь целесообразно применить формулу Байеса. Но метод Байеса в строгом смысле оправдан, только если альтернативные гипотезы (в данном случае заболевания) имеют априорные вероятности. Однако всегда имеется возможность выбрать модель, в которой априорные вероятности существуют и задаются соответствующими правилами даже при полном отсутствии информации. Например, всегда можно предположить, что все заболевания равновероятны.

В диагностических процедурах наличие большого числа признаков создает дополнительные трудности. Так, если дискриминантный признак один, то его значения можно разбить на g непересекающихся интервалов $I_1, I_2, \ldots, I_g$ и каждому интервалу присвоить метку с названием соответствующего диагноза $G_{d1}, G_{d2}\ldots, G_{dg}$. Тогда диагностическое правило такое: «Если значение дискриминантного признака принадлежит интервалу I_k, диагностируется заболевание G_{dk}». Аналогичное правило можно построить, если дискриминантных признаков два или три. Но, если, например, имеется десять дискриминантных признаков для пары взаимоисключающих диагнозов, диагностическое правило должно описать около 1000 различных сочетаний интервалов. Разработка и дальнейшее применение подобного метода диагностики не представляется оптимальным вариантом решения задачи диагностики. Можно также предположить, что, при применении методов многомерного статистического анализа на прак-

тике при большом количестве признаков проявятся проблемы, предусмотреть которые невозможно без использования экспертной информации от специалистов-диагностов. Поэтому в диссертационной работе предложено использовать методы дискриминантного анализа — канонический анализ и элементарные классифицирующие функции с использованием экспертных оценок, для получения нескольких интегральных признаков, с помощью которых можно построить систему диагностики. При этом подчеркивается, что подключение врача к процессу математического моделирования позволяет избежать проблем, связанных с интерпретацией результатов диагностических процедур и их применимостью, пониманием диагностической сущности выделенных главных компонент и т.д. Таким образом, сотрудничество с врачом превращает решение задачи построения диагностической системы в динамический, контролируемый процесс, при этом результаты промежуточных этапов полностью устраивают обе участвующие в этом процессе стороны.

В работе также отмечено, что наибольший интерес для практического здравоохранения представляют системы диагностики и дифференциальной диагностики заболеваний основанные на комплексе самых разнообразных данных, таких как анамнез, клинический осмотр, результаты лабораторных тестов и сложных функциональных методов. Смесь таких разнородных данных слабо подается классификации методами дискриминантного анализа с его линейными функциями классификации и ограничениями на параметры распределения признаков. Анализ литературы и проведенная серия экспериментов показывают, что выходом из этой ситуации является применение искусственных нейронных сетей. Но в отличие от дискриминантного анализа, набор параметров нейросетевой модели не может быть обоснован статистически. Поэтому экспертная информация от специалиста-диагноста в этом случае оказывается наиболее ценной. В противном случае модель может оказаться не согласованной с опытом врача.

В третьей главе диссертационной работы приведены результаты четырех исследований различных, конкретных заболеваний. В ходе этих исследований решались задачи диагностики, дифференциальной диагностики и прогнозирования.

Первые два заболевания — это сепсис и панкреонекроз. Существующие методы обследования больных данными заболеваниями обладают недостаточной прогностической информативностью, что обосновывает необходимость разработки комплексных клинических и лабораторных методов, а также надежных скрининг–систем как для ранней диагностики данных заболеваний, прогнозирования их осложнений и исхода, так и для оценки эффективности лечения. Наиболее активно разрабатываемые в настоящее время методики лабораторной диагностики сложны, трудоемки и требуют высоких затрат. Это делает необходимым поиск таких объектов исследования, которые были бы легко доступны и их изменения коррелировали бы с самими патологическими процессами.

Были исследованы клетки периферической крови. Эритроциты и лимфоциты. Изменения морфологии эритроцитов и лимфоцитов при различных воз-

действиях тесно связаны с функциональной активностью этих клеток, что может использоваться в диагностике различных заболеваний.

У пациентов изучались морфоденситометрические показатели данных клеток при помощи метода компьютерной телевизионной морфоденситометрии (КМДМ).

В основу исследования сепсиса положены данные о 104 пациентах, которые находились на лечении в хирургических отделениях Городской больницы №1 г. Барнаула с июля 1999 года до мая 2002 года. Группу контроля составили 20 клинически здоровых лиц.

Исходная таблица данных содержит 101 переменную + данные повторных измерений на 3 и 10 день течения заболевания. Т.е. в общей сложности 303 переменные.

В ходе исследования решались следующие задачи:

1. **Дифференциальная диагностика.** Определить основное заболевание, вызвавшее сепсис до проведения операции.
2. **Диагностика.** Определить наличие сепсиса и тяжесть его течения.
3. **Прогноз.** Определить вероятность благоприятного (выздоровление) или неблагоприятного (летальный) исхода заболевания.

Решение каждой поставленной задачи проводилось в два этапа. На первом этапе проводился разведочный анализ. В качестве базовой модели данных была выбрана дискриминантная модель с независимыми переменными. Оценивалась статистическая значимость различий средних в группах. В случае если гипотеза о принадлежности значений показателя КМДМ нормальному распределению подтверждается, применялся *t*–критерий Стьюдента или критерий Ньюмена-Кейлса. При отсутствии оснований считать гипотезу о принадлежности нормальному распределению состоятельной, применялся порядковый критерий Крушкаль-Уоллиса. Для удаления избыточной информации производилась процедура отсеивания переменных методом множественной регрессии. Кроме этого привлекалась экспертная информация о взаимосвязи геометрических характеристик эритроцитов и лимфоцитов, что позволило сократить признаковое пространство примерно в два раза.

На втором этапе проводился дискриминантный анализ, с целью получения интегрального количественного выражения множественных различий. Выбор дискриминантного анализа, обусловлен в данном случае простотой вычислений и интерпретации результатов.

В результате получены новые данные об изменении строения и оптических характеристик форменных элементов периферической крови (эритроциты и лимфоциты) у больных с абдоминальным сепсисом и их взаимосвязи с тяжестью состояния пациента. Впервые определены морфоденситометрические критерии изменений эритроцитов и лимфоцитов крови у пациентов с абдоминальным сепсисом, позволившие провести дифференциальную диагностику наиболее частых острых хирургических заболеваний, осложняющихся сепсисом. Впервые определены морфоденситометрические критерии изменений эритроцитов и лимфоцитов крови для раннего прогнозирования течения, исхода абдоминального сепсиса.

Впервые разработанные дискриминантные функции показателей компьютерной телевизионной морфоденситометрии эритроцитов и лимфоцитов периферической крови у больных с абдоминальным сепсисом позволяют проводить раннюю диагностику сепсиса при острых хирургических заболеваниях органов брюшной полости и дифференцировку первичной патологии. По данным изменений показателей компьютерной морфоденситометрии эритроцитов и лимфоцитов крови возможно раннее прогнозирование исходов абдоминального сепсиса и оценка эффективности проводимого лечения.

В основу исследования панкреонекроза положены данные о 80 больных различными его формами, которые находились на лечении в хирургических отделениях Городской больницы №1 г. Барнаула с 1999 года по 2001 годы. В группе контроля исследовано 20 человек без признаков каких-либо заболеваний. В итоге получена таблица, содержащая 170 переменных.

В ходе исследования решались следующие задачи:

1. **Диагностика.** Выявление изменений показателей компьютерной морфоденситометрии эритроцитов и лимфоцитов периферической крови у больных панкреонекрозом.
2. **Дифференциальная диагностика.** Определение изменений показателей компьютерной морфоденситометрии эритроцитов и лимфоцитов периферической крови у больных панкреонекрозом в зависимости от формы панкреонекроза.
3. **Прогноз.** Выявление возможностей прогноза течения панкреонекроза по данным компьютерной морфоденситометрии эритроцитов и лимфоцитов периферической крови.

Решение каждой поставленной задачи проводилось в два этапа. На первом этапе проводился разведочный анализ. В качестве базовой модели данных была выбрана дискриминантная модель с независимыми переменными. Оценивалась статистическая значимость различий средних в группах. В случае если гипотеза о принадлежности значений показателя КМДМ нормальному распределению подтверждается, применялся *t*–критерий Стьюдента или критерий Ньюмена-Кейлса. При отсутствии оснований считать гипотезу о принадлежности нормальному распределению состоятельной, применялся непараметрический критерий Крушкаль-Уоллиса. Для удаления избыточной информации производилась процедура отсева переменных методом множественной регрессии.

На втором этапе проводился дискриминантный анализ, с целью получения единого количественного выражения множественных различий. Выбор дискриминантного анализа, обусловлен в данном случае простотой вычислений и интерпретации результатов.

Получены новые данные об изменениях эритроцитов и лимфоцитов крови по данным компьютерной морфоденситометрии при панкреонекрозе и их взаимосвязи с развитием гнойно-септических осложнений, в динамике течения заболевания. Впервые определены компьютерные морфоденситометрические критерии изменений эритроцитов и лимфоцитов крови у больных различными формами панкреонекроза, а также изменения эритроцитов и лимфоцитов крови в динамике течения воспалительного процесса при панкреонекрозе. Впервые

обнаружены изменения морфоденситометрических показателей эритроцитов и лимфоцитов крови, которые могут быть использованы для прогноза течения панкреонекроза.

Впервые разработанные дискриминантные функции показателей компьютерной морфоденситометрии эритроцитов и лимфоцитов периферической крови у больных панкреонекрозом, имеющие высокую вероятность, способствуют дифференциальной диагностике различных форм панкреонекроза между собой. По данным изменений показателей компьютерной морфоденситометрии эритроцитов и лимфоцитов крови у больных панкреонекрозом в динамике возможно прогнозирование течения заболевания.

Результатом следующего исследования является нейросетевая система диагностики описторхоза. В основу исследования положены данные лабораторных анализов, отражающих клиническое состояние пациента. Были взяты результаты лабораторных и функциональных исследований: значения уровней иммуноглобулинов типа A, G, M, а также наличие специфического иммуноглобулина в крови, изменения в печени по данным УЗИ, наличие таких проявлений описторхоза как сыпь, бронхоспазмы, боли в печени, артральгии, диспепсия, субфебрилитет. Всего одиннадцать параметров.

Исходный набор данных состоит из 158 примеров. Объемы контрольной (практически здоровых лиц) и основной (лиц с заболеванием) групп примерно равны.

Задача исследования состояла в построении системы диагностики данного заболевания. В результате предварительного анализа было принято решение разработать соответствующую систему диагностики (СД) с помощью модели искусственной нейронной сети.

Процесс создания СД проводился в рамках разработанной схемы и включал: разведочный анализ данных, обучение нейросетей с различной структурой, разработку алгоритма, разработку компьютерной программы, испытание системы на примерах, не входящих в обучающую выборку, «доучивание» системы на этих примерах.

Хотя данные даже приближенно не удовлетворяют модели Фишера, с помощью формального дискриминантного анализа было найдено линейное правило классификации по всем имеющимся переменным. Правило состоит из двух функций классификации (по одной для каждой группы таблица 2). Классификация производится подстановкой наблюдаемых значений переменных в каждую из функций классификации. Задача унификации переменных решена следующим образом. Качественные признаки перед подстановкой в функции кодируются числами: 1 — признак присутствует, минус 1 — признак отсутствует. Использование данных кодов не препятствует формальной процедуре применения дискриминантного анализа, т.к. для дихотомических признаков числовые коды их значений не влияют на результат классификации. Затем все переменные перед подстановкой в функции нормируются относительно среднего и среднеквадратичного отклонения обучающей выборки. Т.е. от значений переменной отнимается общее среднее, и затем полученная разность делится на

среднеквадратичное отклонение. Объект относится к тому классу, чья функция классификации имеет большее значение.

Таблица 2. Коэффициенты функций классификации для применения в диагностике описторхоза

Переменная	Описторхоза нет	Описторхоз есть	λ Уилкса (0,1608)	*p*
Ig G	–0,0010	0,0211	0,1691	0,0069
Ig A	0,0943	0,0736	0,1663	0,0270
Ig M	0,1651	0,1131	0,1841	0,0000
Специфический белок	***–2,6431***	***–2,8257***	***0,1608***	***0,8058***
УЗИ	–2,4490	–0,7902	0,1682	0,0106
Сыпь	–3,5170	–1,1946	0,1771	0,0002
Бронхоспазмы	–13,5914	–8,2430	0,1879	0,0000
Боли в печени	–6,1677	0,6214	0,3034	0,0000
Артралгии	***–11,3687***	***–11,3567***	***0,1607***	***0,9926***
Диспепсия	–1,2326	0,3017	0,1673	0,0161
Субфебрилитет	–6,7435	–3,8075	0,1748	0,0005
Константа	–43,1556	–23,9255		

Показателем значимости переменной для дискриминации является значение статистики λ Уилкса. В заголовке таблицы 2 приведено значение λ Уилкса для модели в целом (0,1608). В строках таблицы указаны значения данной статистики при временном исключении соответствующей переменной из модели. λ Уилкса в данном случае, возможно, приближенно подчиняется распределению $F(11, 144)$. Достигнутый уровень значимости приводится в последнем столбце таблицы. Две переменные, выделенные полужирным курсивным шрифтом, являются незначимыми для модели. Этот факт подтверждается после сравнения двух таблиц классификации — до исключения незначимых переменных и после (таблицы 3, 4).

Таблица 3. Таблица классификации при использовании всех переменных

	Истинная классификация	
Классификация теста	Заболевание есть	Заболевания нет
Заболевание есть	69	0
Заболевания нет	7	80
Итог	76	80

Оценка чувствительности теста 90%, специфичности 99%, *PPV* 99%, *NPV* 91%.

Таблица 4. Таблица классификации после удаления незначимых переменных

	Истинная классификация	
Классификация теста	Заболевание есть	Заболевания нет
Заболевание есть	69	0
Заболевания нет	7	80
Итог	76	80

Таблицы идентичны — исключение указанных переменных из модели не ухудшает качество диагностики. Кроме того, полученный результат согласует-

ся с мнением эксперта о том, что исключенные переменные не являются значимыми для диагностики данного заболевания.

Полученный математический метод диагностики хорошо определяет факт отсутствия заболевания (специфичность близка к 100%). Тест дал ошибочное заключение в 7 случаях, когда заболевание в действительности есть (чувствительность теста примерно 90%). Для повышения чувствительности было решено использовать нейросетевую технологию.

Экспериментальным путем была получена оптимальная структура сети — 3 нейрона на входном слое, 2 нейрона на промежуточном, 1 нейрон на выходе. В качестве функции активации для всех нейронов взята гиперболическая сигмоида. Интерпретация значений сигмоиды на выходе (на выходном нейроне) производится по принципу: значение сигмоиды больше нуля — тест положительный, значение меньше нуля — тест отрицательный. При обучении сети множество обучающих примеров было разбито на две части — обучающее (116 случаев) и контрольное множество (40 случаев). После обучения сети на контрольном множестве сеть допустила всего одну ошибку классификации. График динамики среднеквадратичной ошибки при обучении приведен на рис. 3. Две линии на рисунке, обозначающие динамику ошибки сети в процессе обучения, убывают одновременно. Это означает, что сеть правильно сформировала правило разделения объектов на классы. Этим подтверждается работоспособность построенной модели. Для увеличения точности классификации сеть была адаптирована к объединенному (обучающему и контрольному) множеству. Итоговая нейронная сеть обеспечивает уровень чувствительности и специфичности близкий к 100%.

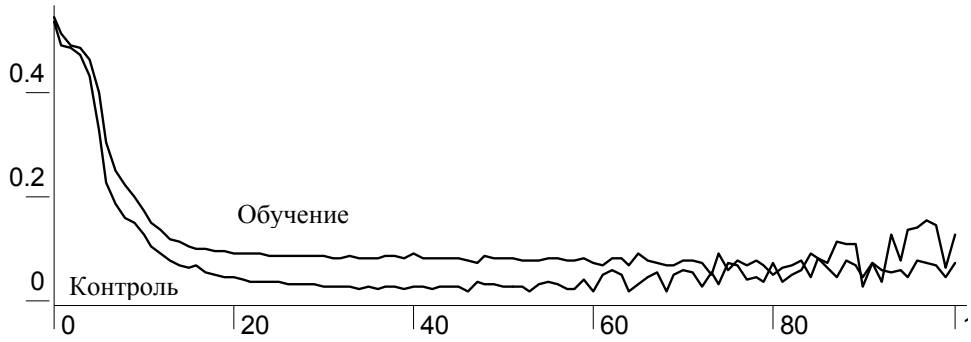

Рис. 3. Линия «Обучение» обозначает ошибку сети на множестве обучающих примеров, линия «Контроль» — на контрольном

Созданная на основе разработанного алгоритма компьютерная программа диагностики описторхоза, передана для использования в Барнаульский центр по диагностике, профилактике и лечению описторхоза и других гельминтозов. Программа зарегистрирована российским агентством по патентам и товарным знакам.

Основными результатами исследования следует считать: обоснование статистической значимости предложенных признаков для диагностики описторхоза, создание и обучение нейронной сети, разработка алгоритма и прикладной программы диагностики описторхоза.

Результатом следующего исследования является система «Прогнозирования риска развития климактерического синдрома у женщин в перименопаузальном периоде». В основу исследования положены данные анкетного опроса группы женщин проживающих в Алтайском крае. Объем группы опрошенных составляет 1642 случая. Анкетные данные представляют собой набор 29 дискретных переменных, из которых 7 являются порядковыми, 19 переменных являются номинальными и две переменные являются классификационными. Классификационные переменные это *КС* — закодированная двумя числами «0 – КС не возникает», «1 – КС возникает» и *КСС* — закодированная четырьмя числами «0 – КС не возникает», «1 – КС слабой», «2 – КС средней» и «3 – КС тяжелой степени».

Исследование проводилось в два этапа. Сначала был проведен разведочный анализ данных. С помощью методов непараметрической статистики: критерия Манна-Уитни и Хи-квадрат, с привлечением экспертной информации были выявлены признаки, которые различаются в группах.

На втором этапе производилось нейросетевое моделирование имеющихся данных. Исходный набор наблюдений методом случайного отбора был разделен на три части — обучающее множество, контрольное множество, тестовое множество в пропорции 4:1:1. Обучающее множество использовалось для обучения нейронных сетей. Контрольное множество использовалось для выбора наилучшей нейронной сети среди всех обученных. Тестовое множество использовалось для оценки чувствительности и специфичности модели.

Для прогнозирования риска развития КС использована модель персептрона с одним входным и одним выходным слоем (без промежуточных слоев). Всего в модели участвует три персептрона. В качестве функции активации выбрана экспоненциальная сигмоида для всех нейронов. В ходе многочисленных экспериментов были определены оптимальные по соотношению количество-качество наборы нейронов в трех сетях.

Для предъявления введенных данных нейронной сети, они подвергаются предварительной обработке. Числовые данные нормируются относительно среднего и стандартного отклонения (их оценки вычислены на основе обучающих данных и заложены в программу); нечисловые и логические переменные кодируются числами.

После предварительной обработки, данные последовательно предъявляются трем нейронным сетям.

Первая нейронная сеть состоит из 50 нейронов на первом слое и одного нейрона на выходном слое. Данная нейронная сеть была обучена решать задачу определения риска развития, в общем, без учета степени тяжести. Оценка чувствительности и специфичности теста составляет 96%. Если сеть дает положительный результат, данные предъявляются второй нейронной сети.

Вторая нейронная сеть состоит 13 нейронов на входном слое и одного нейрона на выходном. Задача данной сети — выявить риск развития КС слабой степени. Оценка чувствительности и специфичности для данной сети с учетом ошибок, допускаемых первым персептроном, составила 93%. Если сеть дает отрицательный результат, данные предъявляются третьей нейронной сети.

Третья нейронная сеть состоит 12 нейронов на входном слое и одного нейрона на выходном. Данная сеть выявляет риск развития КС средней или тяжелой степени. Оценка для чувствительности и специфичности, с учетом ошибок первых двух моделей, составила 90%.

Сигнал выходного нейрона лежит в диапазоне (0, 1), при обучении минимизировалась среднеквадратичная ошибка $E(W) = \frac{1}{2}\sum_{i=1}^{N}(Net(X_i) - Y_i)^2$, где $Y_i = \{0 \text{ или } 1\}$ — требуемый результат, $Net(X_i)$ — ответ сети. После того, как сеть обучена, возникает задача интерпретации ответа нейронной сети, который никогда не равен строго 0 или 1. Обычно границей раздела классов является число 0,5. То есть, если ответ сети меньше 0,5, объект принадлежит первому классу, если больше, объект принадлежит ко второму классу. В данном случае обучающая выборка состоит из групп разной численности. Поэтому, с целью выравнивания чувствительности и специфичности теста, граница раздела смещена в пользу группы большего объема.

В результате исследования была разработан алгоритм и компьютерная программа прогнозирования риска развития КС.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В ходе проведенного исследования решены поставленные задачи и достигнуты следующие результаты:

1. Разработан комплексный метод исследования медицинских данных, состоящий из процедур статистической обработки данных и нейросетевого моделирования с привлечением экспертной информации, учитывающий специфику слабо структурированных данных об исследуемых заболеваниях.
2. Разработаны алгоритмы и компьютерные программы для диагностики описторхоза, диагностики и дифференциальной диагностики панкреатита и абдоминального сепсиса, а также для прогнозирования развития климактерического синдрома у женщин перименопаузальном периоде.
3. Даны оценки прогностической значимости выявленных комплексов диагностических факторов для диагностики вышеуказанных заболеваний в Алтайском крае.
4. Предложен алгоритм исключения неинформативных переменных, т.е. не несущих существенной информации с использованием метода множественной регрессии;
5. Предложен точный способ построения доверительных интервалов для долей, основанный на свойствах бета распределения, улучшающий точность оценки в используемых процедурах диагностики

6. Предложен усовершенствованный численный алгоритм вычисления вероятности для точного критерия Фишера при анализе таблиц сопряженности 2x2.

Результаты исследований переданы для использования в диагностические центры г. Барнаула.

ПУБЛИКАЦИИ ПО ТЕМЕ ИССЛЕДОВАНИЯ

1. Никулина М.А., Шевченко В.В., Лычев В.Г., Бабушкин И.Е., Татаринцев П.Б.. Психосоматические аспекты вирусных гепатитов. // Дальневосточный журнал инфекционной патологии №7, 2005. Хабаровск.
2. Petrova D.V., Shoikhet Ya.N., Tatarincev P.B. Mathematical methods of diagnostics in estabilishing the microbial cause of initial non-effective treated pneumonia. Pneumonia. 14-th National Congress on Lung Diseases. Abstract Book, Moscow, June 22-26, 2004.
3. Dina V. Petrova, Yakov N. Shoikhet, Pavel B. Tatarintsev. Rational antimicrobial therapy of slowly resolving or nonresolving pneumonia. European Respiratory Journal. Abstracts, 14th ERS Annual Congress, Glasgow, UK, September 4-8, 2004.
4. Беднаржевская Т.В., Шойхет Я.Н., Гранитова Л.В., Макарова И.Н., Татаринцев П.Б.. Выраженность вторичной легочной гипертензии у больных бронхиальной астмой. // Сибирский медицинский журнал. №3, 2004.
5. Шойхет Я.Н., Татаринцев П.Б., Толстокоров И.Г.. Острый билиарный панкреонекроз: возможности диагностики. Актуальные вопросы абдоминальной и сосудистой хирургии. Тр. научно-практической конференции хирургов Сибирского региона. Барнаул – Белокуриха, 2002.
6. Татаринцев П.Б., Карбышева Н.В., Семенов С.П.. Применение нейронных сетей в диагностике описторхоза // Актуальные проблемы инфектологии и паразитологии: Матер. 1-й международной юбилейной конф. – Томск, 2001. – с. 52.
7. Татаринцев П.Б., Карбышева Н.В., Семенов С.П., Нейросетевые методы диагностики описторхоза // Материалы 4-й краевой конференции по математике. – Барнаул, 2001. – с. 49.
8. Татаринцев П.Б.. Модель нейронной сети для диагностики заболеваний // Материалы 5-й краевой конференции по математике. – Барнаул, 2002. с. 72-74.
9. Татаринцев П.Б.. Диагностика заболеваний методами нейросетевого моделирования // Известия АГУ: Специальный выпуск, посвященный

пятилетию краевой конференции по математике. – Барнаул, 2002. с. 112-114.

10. Татаринцев П.Б., Карбышева Н.В., Семенов С.П.. Компьютерная технология в диагностике описторхоза // Современные технологии лабораторной диагностики нового столетия: Труды всероссийской конференции. – Москва, 2002. с. 105-106.
11. Татаринцев П.Б., Карбышева Н.В., Семенов С.П.. Тест на описторхоз (Opisthorchis). Рег. номер 2002610696 // Программы для ЭВМ. Базы данных. Топологии интегральных схем: Официальный бюллетень Российского агентства по патентам и товарным знакам. – 2002. – №3. с. 136.
12. Татаринцев П.Б., Кобозева Л.Н., Карбышева Н.В., Индивидуальное прогнозирование развития климактерического синдрома (Климакс). Рег. номер 2003611604 // Программы для ЭВМ. Базы данных. Топологии интегральных схем: Официальный бюллетень Российского агентства по патентам и товарным знакам. – 2003.

Подписано в печать 3.06.2006 г. Формат 60x84/16.

Бумага для множительных аппаратов. Печать офсетная.

Объем 1 п.л. Тираж 100 экз.

Лаборатория множительной техники экономического факультета АГУ.