МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ М. В. ЛОМОНОСОВА

На правах рукописи

САВКИН ИГОРЬ АЛЕКСЕЕВИЧ

ПЕРСОНАЛИЗИРОВАННАЯ МЕДИЦИНА: ПРОГНОЗИРОВАНИЕ НА ОСНОВЕ БАЙЕСОВСКИХ СЕТЕЙ

Специальность 03.01.02 – Биофизика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва – 2019

Работа выполнена на кафедре биофизики физического факультета Московского Государственного Университета имени М.В. Ломоносова и в лаборатории вычислительных систем и прикладных технологий программирования Научно-исследовательского вычислительного центра МГУ имени М.В. Ломоносова

Научные руководители	**Твердислов Всеволод Александрович** доктор физико-математических наук, профессор
	Сулимов Владимир Борисович доктор физико-математических наук, зав. лаб.
Официальные оппоненты	**Пантелеев Михаил Александрович** доктор физико-математических наук, профессор РАН, ВРИО директора Центра теоретических проблем физико-химической фармакологии РАН
	Бирюков Александр Сергеевич доктор физико-математических наук, профессор, зав. теорсектором, Федеральное государственное бюджетное учреждение науки научный центр волоконной оптики Российской академии наук (НЦВО РАН)
	Романов Алексей Николаевич, кандидат физико-математических наук старший научный сотрудник, Федеральный исследовательский центр химической физики им. Н.Н.Семенова РАН

Защита диссертации состоится 21 ноября 2019 года в 15.30 на заседании диссертационного совета МГУ.01.04 Московского государственного университета имени М.В.Ломоносова по адресу: 119991 Москва, Ленинские горы, д. 1, стр. 2, Физический факультет МГУ, ЦФА. E-mail: info@physics.msu.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М.В.Ломоносова (Ломоносовский пр-т, д. 27) и на сайте ИАС «ИСТИНА»: https://istina.msu.ru/dissertations/243130741/

Автореферат разослан «___» _____ 2019 г.

Ученый секретарь диссертационного совета МГУ.01.04,
кандидат технических наук Сидорова А.Э

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Поиск подходов к построению экспертной медицинской системы, позволяющей получать хорошее качество и высокую надежность предсказания (развития заболевания или диагноза) и в то же время не являющейся моделью “черного ящика” является важной задачей в современной медицине. Байесовские сети реализуют сравнительно сложный подход для применения в области прогнозирования, если нет достаточно глубокого понимания соответствующих алгоритмов. Кроме того, методы прогнозирования на основе технологии байесовских сетей сравнительно плохо освещены в научной литературе, хотя, именно использование байесовских позволяет дать структурированную и хорошо визуализированную топологию причинно-следственных связей, что особенно важно при выборе лечения конкретного пациента при данном заболевании. Именно эти причины делают данную работу особенно актуальной.

Степень разработанности избранной темы

В отечественной литературе практически нет публикаций, посвященных применению технологии байесовских сетей для задач персонализированной медицины, в которых строятся вероятностные прогностические модели на основе баз данных пациентов с теми или иными заболеваниями. В данной работе впервые байесовские сети применены для построения прогностических моделей для следующих медицинских проблем: предсказания состояния пациентов, перенесших острый коронарный синдром, для пациентов с раком молочной железы, прогнозирования специфического гуморального иммунного ответа на основании исходных параметров иммунного статуса детей, привитых против кори, краснухи и эпидемического паротита, полногеномного анализа генетических ассоциаций для постановки диагноза полигенной гиперхолестеринемии и для определения особенностей течения хронического гепатита С.

Цели исследования. Разработать математические методы, алгоритмы и программы, позволяющие применять технологию байесовских сетей для решения задач персонального прогнозирования состояния пациентов с различными заболеваниями.

Основные задачи. Основными задачами данного исследования являются:

- разработка математических методов, алгоритмов и их программная реализация для использования байесовских сетей при прогнозировании развития или исхода заболеваний на основе персональных данных пациентов;
- разработка программ для построения и работы с байесовскими сетями: программ обучения байесовских сетей на заданных базах данных пациентов и программы предсказания на обученных сетях состояния пациентов;
- разработка программ для оценки надежности предсказаний;
- проведение тестирования разработанных программ;
- построение вероятностных моделей на основе байесовских сетей для прогнозирования состояния пациентов на имеющихся базах данных пациентов, и выявление с их помощью критических прогностических параметров;
- применение разработанных алгоритмов и программ для прогнозирования состояния пациентов на имеющихся базах данных пациентов, в том числе для баз данных пациентов с диагнозом рака молочной железы и базы данных пациентов, содержащей однонуклеотидные полиморфизмы, для прогноза гипергликемии.

Объект и предмет исследований

Объектом данного исследования являются пациенты (больные или здоровые люди) с их клиническими, генетическими и другими параметрами вплоть до образа жизни с теми или другими заболеваниями или без них, а также состояние этих пациентов. Предметом исследований является прогноз

состояния пациентов. Исследование заключается в изучении прогноза состояния пациентов в зависимости от их параметров. Квинтэссенция объекта исследований является база данных, содержащая параметры пациентов и их состояние.

Новизна

Впервые разработаны и реализованы методы, алгоритмы и программы на основе технологии байесовских сетей для прогнозирования состояния пациентов по их персональным данным. Впервые предложен и реализован в виде суперкомпьютерной программы метод оптимизации байесовских сетей по числу узлов для повышения надежности предсказаний и выявления наиболее значимых прогностических параметров пациентов. Впервые разработанные методы, алгоритмы и программы применены для построения вероятностных моделей на основе байесовских сетей для прогнозирования состояния пациентов и выявления наиболее значимых прогностических параметров для следующих заболеваний: острого коронарного синдрома, рака молочной железы, прогнозирования специфического гуморального иммунного ответа на основании исходных параметров иммунного статуса детей, привитых против кори, краснухи и эпидемического паротита, полногеномного анализа генетических ассоциаций для постановки диагноза полигенной гиперхолестеринемии, для определения особенностей течения хронического гепатита С.

Теоретическая и практическая значимость работы

Теоретическая значимость работы заключается в разработке методологии постановки задачи персонализированного подхода к прогнозу состояния пациентов, в выработке методики прогноза, методики подготовки базы данных, программ построения и обучения байесовских сетей, программ построения ROC-кривых (Receiver Operating Characteristic) и вычисления площади под ними, представляющей собой важный критерий надежности прогноза. Важное теоретическое значение имеет разработанная методика и

программа оптимизации наивных байесовских сетей по числу узлов с целевой функцией, представляющей собой площадь под ROC-кривой. Такая оптимизация позволяет существенно повысить надежность прогноза и выявить наиболее важные прогностические параметры.

Практическая значимость работы заключена в полученных оптимальных байесовских сетях, которые можно использовать для прогноза состояния пациентов, перенесших острый коронарный синдром, пациентов с раком молочной железы, для прогнозирования специфического гуморального иммунного ответа на основании исходных параметров иммунного статуса детей, привитых против кори, краснухи и эпидемического паротита, для полногеномного анализа генетических ассоциаций для постановки диагноза полигенной гиперхолестеринемии и для определения особенностей течения хронического гепатита С. Разработанная методика гистограмм риска позволяет стратифицировать пациентов по их персональным данным по группам риска.

Методология диссертационного исследования

Проведенные исследования позволили выработать методологию подготовки баз данных пациентов и представление их в форме, которая удобна для использования для обучения и прогноза с помощью байесовских сетей. Разработать методологию постановки задачи прогноза и методику дискретизации непрерывных параметров пациентов.

Положения, выносимые на защиту

1. Разработка математических методов и алгоритмов для построения вероятностных прогностических моделей на основе байесовских сетей для персонального прогноза состояния пациентов на основе информации о них из соответствующих баз данных.

2. Реализация разработанных моделей и алгоритмов в виде комплекса программ, осуществляющих построение байесовских сетей, их обучение,

опрос и валидацию. Построение вероятностных прогностических моделей и выявление с их помощью критических прогностических параметров.

3. Применение разработанного комплекса программ для прогноза летального исхода как результата рака молочной железы для выявления нового прогностического фактора – экспрессию мРНК гена YB-1 в опухоли.

4. Разработка программ для построения вероятностных прогностических моделей на основе байесовских сетей с использованием однонуклеотидных полиморфизмов (снипов) для прогноза содержания липопротеинов низкой и высокой плотности.

Публикации. По теме диссертации опубликовано шесть работ в журналах, индексируемых в базах Web of Science, RSCI, Scopus и РИНЦ.

Степень достоверности и апробация результатов

Степень достоверности следует из корректности постановки задачи и подтверждается тем, что разработанная методика опирается на теорию вероятности и на использование технологии байесовских сетей, а также на применение апробированных численных методов. Апробация результатов и методов проведена на российских и международных конференциях (шесть докладов, а также два тезиса докладов, опубликованных в журналах).

Личный вклад

Автор диссертации принимал непосредственное участие в поиске и анализе литературных данных по методам и алгоритмам, в написании и отладке программ NetGen (для построения наивной байесовской сети), ANN (для обучения, опроса и валидация наивных байесовских сетей), SiLVIA (для оптимизации наивной байесовской сети по числу узлов), ANN-RARE (для обучения, опроса и валидации наивных байесовских сетей на базах данных содержащих редкие значения), GeNA (для построение базы данных и наивных байесовских сетей для данных, использованных при прогнозировании генетического влияния на гипергликемию), Pirson (для оценки по методу Пирсона для оценки связанности параметров). В обработке

и анализе полученных результатов, формулировании выводов, а также в подготовке публикаций и докладов на научных конференциях.

Структура и объем диссертации диссертационная работа состоит из введения и четырех глав. Объем диссертации составляет 92 страницы текста, включая 20 рисунков и 4 таблицы. Список литературы состоит из 108 наименований.

СОДЕРЖАНИЕ РАБОТЫ

Во ***Введении*** приводится описание основных моментов в развитии машинного обучения, которые привели к возможности использования данных методов в медицине.

Также описано текущее состояние в области персонализированной медицины и даны сведения об имеющихся тенденциях как в персонализированной медицине, так и в выборе различных математических подходов к решению задач в данной области и причины использования именно байесовских сетей в данной области.

Далее рассмотрена актуальность темы исследования, ее степень разработанности, объект, цели и задачи диссертационной работы, описан предмет исследований. Также показана научная новизна, теоретическая и практическая значимость работы, методология диссертационного исследования, положения, выносимые на защиту, степень достоверности и апробация результатов, личный вклад и практическая значимость выполненной работы.

В **первом пункте первой главы** проводится обзор литературы по применению различных методов в построении экспертных систем (ЭС) в биологии и медицине в исторической последовательности от методов, основанных на правилах до полноценных математических моделей. Основную часть в данном обзоре занимают системы, основанные на байесовских сетях (БС), построению которых и посвящена диссертация.

Также рассмотрены примеры использования других современных методов от простых регрессионных до использования групп сложных моделей для построения одной модели (ансамблирование). **Второй пункт первой главы** посвящен математике методов машинного обучения и начинается с типов основных задач и отличий между ними. Данный пункт дает представление о проблемах при использовании различных методов в медицине и в особенности проблемам моделей, представляющих из себя “черную коробку”. **Раздел один второго пункта** посвящен логистической регрессии: математическим основам при построении и обучения. Дальше рассматриваются наиболее популярный в настоящее время метод нейронных сетей. **Второй раздел второго пункта посвящен** сверточным нейронным сетям – алгоритму, который позволяет достаточно просто строить ЭС для графических данных из-за которого нейронные сети и получили огромную популярность в медицинских задачах, связанных с различными рода изображениями. Завершается этот пункт рассмотрением метода опорных векторов.

Глава 2 посвящена теории и методологии применения байесовских сетей. **Пункт один** посвящен основным терминам и уравнениям, необходимым для построения байесовских сетей для дальнейшего использования при построении программ и понимания полученных с их помощью моделей.

Рисунок 1. Пример простейшей байесовской сети

Основной теоремой для теории байесовских сетей является теорема Байеса, связывающая условные вероятности событий A и B:

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} = \frac{P(A,B)}{P(B)} \quad (1)$$

где $P(A|B)$ – условная вероятность события A при условии наблюдения события B, $P(B|A)$ – условная вероятность события B при условии

наблюдения события A, $P(A)$ и $P(B)$ – вероятности событий A и B, $P(A,B)$ – совместная вероятность наблюдения событий A и B.

А также важная для построения экспертной системы формула классификатора на основе байесовской сети:

$$P(V_t^k|V_1^m,\dots,V_n^m) = \frac{P(V_t^k|parents(V_t^k)\prod_{i=1;i\neq t}^{n} P(V_i^m|parents(V_i^m))}{P(V_1^m,\dots,V_{t-1}^m\dots,V_{t+1}^m,\dots,V_n^m)} \quad (2)$$

где $V_t^k -$ k-ое значение целевого узла V_t; $P(V_i^m|parents(V_i^m))$ – условная вероятность наблюдения значения V_i^m соответствующее e_i^m; $parents(V_i^m))$ – конфигурация родителей соответствующая e^m, а величина в знаменателе $P(V_1^m,\dots,V_{t-1}^m\dots,V_{t+1}^m,\dots,V_n^m)$ представляет собой совместную вероятность наблюдения величин $V_1^m,\dots,V_{t-1}^m\dots,V_{t+1}^m,\dots,V_n^m$. **Второй пункт** посвящен описанию работы с различными данными и в частности реализованных алгоритмах дискретизации непрерывных переменных для получения интервалов. Рассмотрены проблемы ограничения при работе с непрерывными данными без использования дискретизации. **Третий пункт второго раздела** посвящен ненаивным байесовским сетям. В данном пункте

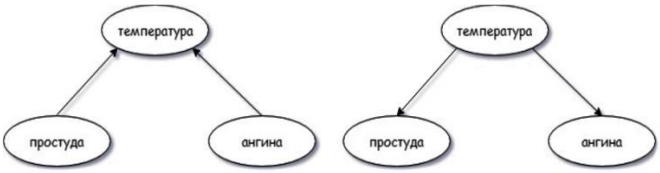

Рисунок 2. Ненаивная (слева) и наивная топология сети (справа).

приводятся определения и математические выкладки, показывающие возможность использования наивных байесовских сетей даже в случаях сложных задач. **Четвертый пункт** данного раздела посвящен алгоритмам построения ненаивной топологии. В **пятом пункте второго раздела** описан процесс обучения сети. В **первом и втором разделе** этого пункта показан прямой и итерационный (EM) алгоритмы обучения байесовских сетей. В **шестом пункте** описан процесс опроса сети. В **разделе один** данного пункта идет описание и алгоритм использования деревьев сочленений для ускорения

процесса опроса. В **седьмом пункте** описан наиболее популярный в настоящее время в медицине подход к оценке качества модели, основанный на использовании чувствительности и специфичности. В **восьмом пункте** описана проблема редких значений и методы ее решения.

Глава 3 посвящена использованным подходам в создании программ и построении моделей для анализа различных медицинских проблем. **Первый пункт** посвящен оптимизации байесовской сети по числу узлов. В нем рассматриваются причины необходимости оптимизации байесовской сети и сложность поиска оптимальной топологии. Во **втором пункте** описан использовавшийся алгоритм, построенный на основе жадного поиска. Данный алгоритм состоит из следующих шагов:

1. Рассчитываем начальное значение AUC для сети $G_{m,initial}$
2. Удаляем каждый узел по отдельности и рассчитываем величины AUC для всех $G_{m,i}$
3. Удаляем узел i, удаление которого ведет к наибольшей максимизации AUC
4. Возвращаемся к пункту 1.

Данный подход позволяет выделять наиболее значимые узлы байесовской сети. В **третьем пункте** описана процедура использования ROC-кривых (Receiver Operating Characteristic curve) и вычисления значения AUC (Area

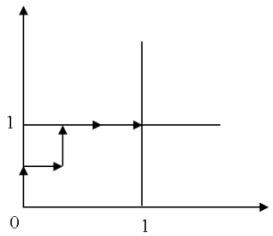

Рисунок 3. Пример ROC-кривой

Under the Curve) для использования в качестве характеристика качества модели. **Четвертый пункт третьей главы** посвящен гистограммам риска –

предложенному способу упрощения работы с результатами прогнозов систем, основанных на байесовских сетях. Данный подход позволяет визуализировать качество прогнозирования для групп с различными

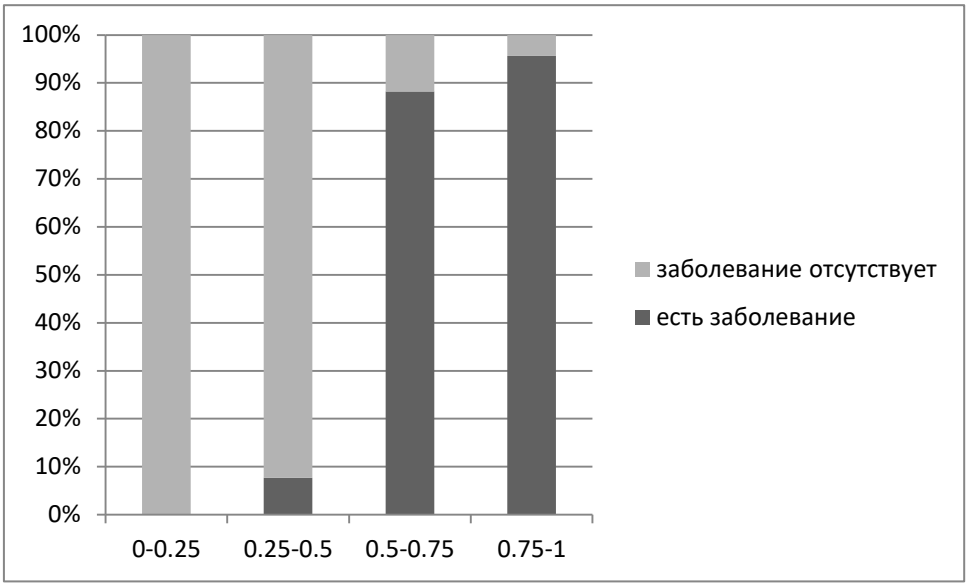

Рисунок 4. Пример гистограммы риска. По горизонтали отложена условная вероятность наблюдения заболевания при имеющихся у пациентов наборе значений остальных узлов.

значениями условных вероятностей заболевания при условии имеющихся знаний о пациенте. В **пятом пункте** перечислены основные программы, написанные для решения задач в данной работе. В частности, для подготовки базы данных из набора генетических данных была создана программа GeNa. Для анализа связанности переменных с помощью критерия хи-квадрат в ходе работы была написана программа Pirson, позволяющая оценивать связанность параметров с целевой переменной исходя из рассчитанных значений хи-квадратов. Для построения файлов сетей была написана программа NetGen. Для обучения и опроса была написана программа ANN, для обучения и опроса в случае редких значений – программа ANN-RARE, для проведения оптимизации наивной БС по числу узлов с помощью метода жадного поиска разработана программа SiLVIA. Описаны основные форматы входных и выходных данных.

Четвертая глава посвящена построению моделей на основе байесовских сетей с использованием написанных программ: коронарный синдром [1], прогнозирование иммунного статуса после прививок [2], рак молочной железы [3], оценка влияния генетических факторов на сердечные нарушения [4], гепатит С [5].

В **пункте один четвертой главы** рассматривалась задача прогнозирования неблагоприятного исхода для двух интервалов времени пол года и через полтора года для пациентов с диагнозом острый коронарный синдром была применена описанный в третьей главе технологию оптимизации байесовских сетей[1] для базы данных пациентов, в которой содержалось 1193 пациента, поступившие в стационар в связи с развитием острого коронарного синдрома (ОКС). В этой базе данных содержалось более 400 параметров (переменных), описывающих статус пациента, демографические параметры, рутинные биомаркеры, данные электрокардиограмм и эхокардиограмм, медицинская история и история семьи, генетические параметры и состояние пациента (исход) через определенные периоды времени после попадания пациента в стационар и получения стандартного лечения. Начальные байесовские сети строились исходя из одинакового набора данных по 204 параметрам с отличием в целевом узле – исход через полгода или исход через полтора года. Для этих начальных сетей надежность предсказания (AUC) оказалась низкой: для прогноза через полгода AUC=0.61, а для прогноза через полтора года AUC=0.65. После оптимизации сети для прогноза на полгода величина AUC выросла до 0.80, и оптимизированная байесовская сеть содержала 48+1 узлов: 48 узлов, соответствующих параметрам пациентов, и один корневой узел, соответствующий исходу через полгода. Для прогноза на полтора года AUC вырос до 0.76, а оптимальная сеть содержала 54+1 узлов. При этом только 19 узлов в каждой из этих двух сетей соответствовали одним и тем же параметрам пациентов. Это говорит о том, что существенно различные

факторы влияют на прогрессирование ИБС через полгода и через полтора года.

Далее, для выявления наиболее важных прогностических факторов была проведена следующая процедура. Удаление узлов-параметров из БС было продолжено до тех пор, пока величина AUC не стала достаточно сильно уменьшаться: этот процесс останавливался, если удаление любого узла приводило к уменьшению величины AUC на значение больше порогового 0.003. Такие минимальные сети содержали 17 и 16 узлов-параметров для прогноза исхода через полгода и через полтора года соответственно. Для обеих сетей оказались важными: полиморфизм гена TNF-альфа, наличие родственников со случаями инфаркта и использование спиронолактона в течение последних десяти дней перед госпитализацией. В данной работе впервые показана важность TNF-альфа гена для предсказания состояния пациентов после перенесенного ОКС.

В **пункте два** описывается построение модели для прогнозирования специфического гуморального иммунного ответа детей, привитых против краснухи, кори и паротита. Целью работы [2] является прогнозирование специфического иммунного ответа через месяц и через год на вакцину Приорикс (Бельгия). Исследовались 40 детей (21 мальчиков и 19 девочек), проходивших плановую вакцинацию. Ранее эти дети не болели корью, краснухой и эпидемиологическим паротитом и не имели антител к нему.

В качестве параметров были выбраны различные параметры, связанные с иммунным ответом: уровни лейкоцитов и лимфоцитов, иммуноглобулинов, относительные и абсолютные значения субпопуляций лимфоцитов и часто используемых цитокинов-маркеров. Всего было использовано 19 параметров. Первоначальные AUC были равны: 0.8, 0.74 и 0.17 соответственно.

В результате оптимизации были построены модели для прогнозирования иммунного ответа. Для моделей ответа на вирус краснухи (AUC=0.95) и кори

(AUC=0.79) важными оказались по семь частично пересекающихся параметра. А для эпидемиологического паротита (AUC=0.66) – четыре.

Как видим, и в данном случае оптимизация наивной байесовской сети по числу узлов привела к выявлению сравнительно небольшого числа прогностических параметров и существенно увеличило надежность предсказания – это особенно ярко видно на примере ответа на вирус краснухи и кори, где по семи прогностическим параметрам можно делать прогноз с надежностью (с величиной AUC) 0.95 и 0.79 соответственно.

В **пункте три**, посвященному задаче прогнозирования рака молочной железы, были использованы данные о 323 пациентах с подтвержденным диагнозом РМЖ I-IV стадий в возрасте 30-85 лет. Были построены начальные модели на основе наивных байесовских сетей с 32+1 узлами (32 узла-параметра и один узел – конечная точка, соответствующая целевой переменной, определяющей исход) для предсказания прогрессирования опухоли и для предсказания смерти больных. В эти 32 параметра вошли основные клинические и молекулярно-биологические данные, информация о видах лечения. В результате начальные байесовские сети дали значения AUC равные 0.626 и 0.621 для прогрессирования опухоли и для предсказания смерти пациентов. Так что начальные сети со всеми 32 параметрами имеют низкую надежность предсказания. После оптимизации сетей по числу узлов были получены модели с AUC 0.833 и 0.907 для предсказания прогрессирования опухоли и для предсказания смерти пациентов соответственно, содержащие по семь значимых узлов.

Для предсказания прогрессирования опухоли важными оказались в основном классические параметры: гормональный статус опухоли (рецепторы прогестерона), экспрессия рецепторов HER-2/neu и поражение лимфатических узлов (стадия по N). Из вариантов лечения значимыми для прогноза прогрессирования РМЖ стали предоперационная лучевая терапия и неоадъювантная химиотерапия. Предоперационная лучевая терапия сейчас

давно не применяется как неэффективная, и такая терапия была только у 8% пациенток, которые получали лечение в 90-х годах прошлого века.

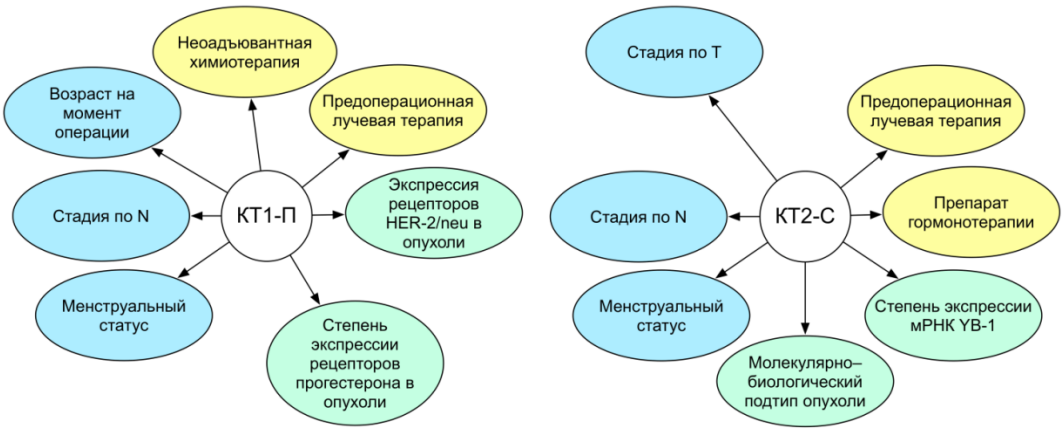

Рисунок 7. Оптимальные байесовские сети для предсказания прогрессирования опухоли (левая сеть) и для предсказания смерти больных (правая сеть).

В число факторов прогноза смерти больных вошли и стандартные прогностические факторы такие, как категория T, категория N, и молекулярно-биологический подтип опухоли. Но кроме них вошел и новый прогностический маркер (YB-1).

В **пункте четыре** описана задача оценки генетического влияния на гипергликемию. Было рассмотрено два фактора риска: LDL-C и HGL-C, ассоциированных с липидным метаболизмом. Переменная LDL-C принимала значение 0, если концентрация LDL-C была меньше, чем 4.9 ммоль/Л, и она принимала значение 1, если концентрация LDL-C была больше, чем 4.9 ммоль/Л. HDL-C принимала значение 0, если концентрация HDL-C была больше, чем 1.2 ммоль/Л, и она принимала значение 1, если концентрация HDL-C была меньше, чем 1.2 ммоль/Л.

Применить в данном случае метод оптимизации величины AUC при варьировании (уменьшении) числа узлов, как это делалось в предыдущих разделах, не представляется возможным ввиду очень большого объема информации. Поэтому сначала для отбора наиболее важных узлов

использовался либо полногеномный поиск ассоциаций (GWAS) [6] с использованием уровня значимости p<0.0001, либо критерий согласия Пирсона χ^2. Дальше использовалась оптимизация байесовской сети по числу узлов.

Основным выводом проделанной работы являлось то, что начальное применение методов отбора узлов-параметров для байесовских сетей методами GWAS или Pirson и дальнейшая оптимизация байесовских сетей по числу узлов с использованием AUC нельзя использовать для выявления критических прогностических генетических параметров, так как эти параметры существенно зависят от группы пациентов к которой применяются эти методы.

В **пункте пять** описано построение модели для хронического гепатита С и цирроза печени. В нашей работе [5] использовались данные 253 больных хроническим гепатитом С (ХГС) и циррозом печени (ЦП) с набором основных клинических и биохимических параметров, характеризующих факторы вируса и хозяина, позволяющих оценить клиническое течение заболевания и исходы, а также точечных мутаций генов из пяти групп [5]: генов участвующих в воспалительных реакциях и противовирусном иммунитете, активаторов локального печеночного фиброза, генов гемохроматоза, генов тромбоцитарных рецепторов, а также генов белков свертывающей системы и эндотелиальной дисфункции.

Модели строились для прогнозирования четырех состояний – четырех конечных точек (КТ): развитие цирроза печени (КТ1), скорость фиброза (КТ2), наличие портальной гипертензии (КТ3) и наличие криоглобулинов (КТ4). Начальные наивные байесовские сети несколько различались для этих исходов как по числу узлов-листьев, так и по самим параметрам пациентов, соответствующих этим узлам, и величины AUC для этих начальных сетей не были впечатляюще большими: для КТ1 сеть была 40+1 и AUC=0.86, для КТ2 сеть была 42+1 и AUC=0.60, для КТ3 сеть была 39+1 и AUC=0.48, для КТ4

сеть была 45+1 и AUC=0.64. Учитывая эти результаты, была проведена оптимизация этих сетей по числу узлов. В результате получились оптимизированные сети со значительно большими значениями AUC: для КТ1 была оптимальная сеть 14+1 и AUC=0.90, для КТ2 – 9+1 и AUC=0.76, для КТ3 – 12+1 и AUC=0.81, для КТ4 – 9+1 и AUC=0.76.

Анализ прогностических факторов, соответствующих узлам-листья оптимизированных сетей, показал, что влияние некоторых из них на конечные точки, особенно клинико-демографических параметров, известно, а для связи других, фактически новых прогностических параметров, особенно полиморфизмов, с конечными точками либо можно выдвинуть разумные соображения, подтверждаемые, хотя и косвенными, опубликованными экспериментальными данными, либо ничего рационального сказать невозможно, и тут требуются дополнительные исследования.

Для других трех конечных точек (КТ2, КТ3 и КТ4) были обнаружены ассоциации с другим или некоторыми теми же мутациями, однако при проверке надежности прогноза оптимизированный байесовских сетей на контрольной группе (10% пациентов от всей базы) выяснилось, что для этих трех конечных точек надежность прогноза падает так сильно, что верить ему нельзя: величина площади под ROC-кривой равна 0.57, 0.21 и 0.44 для КТ2, КТ3 и КТ4 соответственно. Только для прогноза цирроза печени надежность прогноза с помощью оптимизированной байесовской сети осталась достаточно высокой – AUC=0.80. Столь сильное падение AUC на контрольной группе по сравнению с величинами AUC, полученными для группы отбора (90% пациентов от полной базы данных), по-видимому, связано со слишком малым общим числом пациентов в контрольной группе.

Таким образом, оптимизированная нами наивная байесовская сеть может быть использована для прогнозирования цирроза печени (КТ1) со сравнительно высокой надежностью прогноза AUC=0.80.

В **заключении** приведено возможное применение выработанных методик и полученных результатов.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ:

1. Разработаны математические методы и алгоритмы для построения вероятностных прогностических моделей на основе байесовских сетей для персонального прогноза состояния пациентов на основе информации о них из соответствующих баз данных.
2. Разработанные модели и алгоритмы реализованы в виде комплекса программ, осуществляющих построение байесовских сетей, их обучение, опрос и валидацию. С использованием разработанных программ на исследуемых базах данных осуществлено построение вероятностных прогностических моделей и выявление с их помощью критических прогностических параметров.
3. Применение разработанного комплекса программ для прогноза летального исхода как результата рака молочной железы выявило новый прогностический фактор – экспрессию мРНК гена YB-1 в опухоли.
4. Разработаны программы для построения вероятностных прогностических моделей на основе байесовских сетей с использованием однонуклеотидным полиморфизмов (снипов) для прогноза содержания липопротеинов низкой и высокой плотности. Показано, что методика начального отбора снипов с помощью одного из двух методов: полногеномного поиска ассоциаций – GWAS, или с помощью метода Пирсона, существенно зависит от группы пациентов, к которой она применяется.
5. Разработанная методика построения вероятностных прогностических моделей на основе байесовских сетей показала свою работоспособность и эффективность для персонализированного прогноза состояния пациентов. Эта методика может быть применена для разработки вероятностных прогностических моделей для широкого круга персонализированных

прогнозов состояния пациентов и особенностей течения заболеваний.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ в журналах, индексируемых в базах данных Web of Science, RSCI, Scopus и РИНЦ:

1. Генс Г.П., Сулимов А.В., Моисеева Н.И., Овсий О.Г., Вельшер Л.З., Рыбалкина Е.Ю., Селезнева И.И., Савкин И.А., Сулимов В.Б., Поиск подходов к прогнозированию исходов рака молочной железы с помощью байесовских сетей // «Онкология. Журнал им. П.А.Герцена». 2014. Номер: 5, С. 37-46.

   Импакт-фактор журнала по РИНЦ 0,170
2. E.D.Maslennikov, A.V.Sulimov, I.A.Savkin, M.A.Evdokimova, D.A.Zateyshchikov, V.V.Nosikov, V.B.Sulimov, An intuitive risk factors search algorithm: usage of Bayesian network technique in personalized medicine // Journal of Applied Statistics. 2015. Vol.42, Issue 1. P.71-87.

   Импакт-фактор по SCOPUS (SJR) 0.505
3. Топтыгина А. П., Азиатцева В. В., Савкин И А., Кислицин А. А., Семикина Е. Л., Гребенников Д. С., Алешкин А.В., Сулимов А. В., Сулимов В. Б., Бочаров Г. А. Прогнозирование специфического гуморального иммунного ответа на основании исходных параметров иммунного статуса детей, привитых против кори, краснухи и эпидемического паротита, IMMUNOLOGIYA, 2015, Т.36. № 1 (январь-февраль). С.22-30.

   Импакт-фактор журнала по SCOPUS (SJR) 0,128
4. Г. П. Генс, А. В. Сулимов, Н.И. Моисеева, Е.В. Каткова, Л.З. Вельшер, Л.И. Коробкова, И.А. Савкин, В.Б. Сулимов, Применение генных сигнатур и медицинских экспертных систем для прогнозирования клинических исходов рака молочной железы // Вестник РОНЦ им. Н. Н. Блохина РАМН. 2015-2016. Т. 26-27. № 4-1. С. 47 – 52.

   Импакт-фактор журнала по РИНЦ 0,733

5. A.V. Sulimov, A.N. Meshkov, I.A. Savkin, E.V. Katkova, D.C. Kutov, Z.B. Hasanova, N.V. Konovalova, V.V. Kukharchuk, V.B. Sulimov, Genome-wide analysis of genetic associations for prediction of polygenic hypercholesterolemia with bayesian networks // Journal of Computational and Engineering Mathematics (на английском языке), 2015. Vol. 2. N 4. pp.11-26.

   Импакт-фактор журнала по РИНЦ 0,298

6. Л.М. Самоходская, Е.Е. Старостина, А.В. Сулимов, Т.Н. Краснова, Т.П.Розина, В.Г.Авдеев, И.А. Савкин, В.Б. Сулимов, Н.А. Мухин, В.А. Ткачук, В.А. Садовничий, Прогнозирование особенностей течения хронического гепатита С с использованием байесовских сетей, Терапевтический архив, 2019, №2, 32-39.

   Импакт-фактор журнала по РИНЦ 1,004; по SCOPUS (SJR 2018) 0,146.

ТЕЗИСЫ ДОКЛАДОВ НА МЕЖДУНАРОДНЫХ И ВСЕРОССИЙСКИХ НАУЧНЫХ КОНФЕРЕНЦИЯХ,

опубликованные в журналах:

1) Генс Г.П., Мохнюк К.С., Шиндяпин В.В., Савкин И.А., Сулимов А.В., Сулимов В.Б. Прогнозирование исходов рака толстой кишки с помощью Байесовских сетей // Тезисы первого международного форума онкологии и радиологии, 23-28 сентября 2018 года, Москва, Исследования и практика в медицине, 2018, Том 5, № 2S, С. 60

   https://doi.org/10.17709/2409-2231-2018-5-S2

   Импакт-фактор журнала по РИНЦ 0,587.

2) Gelena Guens, Alexey Sulimov, Igor Savkin, Natalya Moiseeva, Vladimir Sulimov Application of Bayesian networks for prognosis of breast cancer patient’s outcomes // Journal of Cancer Science and Therapy 2018, Volume 10, P. 50 DOI: 10.4172/1948-5956-C1-124;

   Импакт-фактор журнала по РИНЦ 0,202.

СПИСОК КОНФЕРЕНЦИЙ, НА КОТОРЫХ БЫЛИ ДОЛОЖЕНЫ РЕЗУЛЬТАТЫ ДАННЫХ ИССЛЕДОВАНИЙ:

1) 2018 Прогнозирование исходов рака толстой кишки с помощью Байесовских сетей (Стендовый) Авторы: Генс Г.П., Мохнюк К.С., Шиндяпин В.В., Савкин И.А., Сулимов А.В., Сулимов В.Б., Первый Международный Форум онкологии и радиологии. Москва, 23-28 сентября 2018 г., Москва, Россия, 23-28 сентября 2018.
2) 2018 Application of Bayesian networks for prognosis of breast cancer patient′s outcomes (Устный), Авторы: Gelena Guens, Alexey Sulimov, Igor Savkin, Natalya Moiseeva, Vladimir Sulimov, 2nd Global Summit on Oncology & Cancer, Сингапур, 12-14 марта 2018
3) 2017 Оценка особенностей течения онкологических заболеваний на основе вероятностных моделей с использованием Байесовских сетей (Устный) Авторы: Генс Г.П., Сулимов А.В., Савкин И.А., Сулимов В.Б. Форум университетской науки 2017, МГМСУ имени Евдокимова, Россия, 16 мая 2017
4) 2015 Выявление генетических факторов, влияющих на развитие атеросклероза (Устный) Авторы: Сулимов А.В., Савкин И.А., Кутов Д.К., Каткова Е.В., Конференция молодых ученых, приуроченная к 60-летию НИВЦ МГУ, конференц-зал НОЦ "Суперкомпьютерные технологии" МГУ , Россия, 8 октября 2015
5) 2015 Применение технологии байесовских сетей для целей персонифицированной медицины (Устный), Авторы: Сулимов А.В., Савкин И.А., Сулимов В.Б., XXII Российский национальный конгресс "Человек и лекарство", Москва, Россия, 6-10 апреля 2015
6) 2014 Применение байесовских сетей для выявления генетические маркеров ассоциированных с факторами риска развития атеросклероза (Устный) Авторы: Савкин И.А., Мешков А.Н., Бойцов С.А., Сулимов

А.В., Сулимов В.Б., IV международная научно-практическая конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине» г. Казань, 29 октября – 1 ноября 2014, г. Казань, 2014

СПИСОК ЦИТИРУЕМОЙ ЛИТЕРАТУРЫ

1 Maslennikov E. D., Sulimov A. V., Savkin I. A., Evdokimova M. A., Zateyshchikov D. A., Nosikov V. V., Sulimov V. B. An intuitive risk factors search algorithm: usage of the Bayesian network technique in personalized medicine // Journal of Applied Statistics. — 2015. — Т. 42, № 1. — С. 71-87.

2 Топтыгина А. П., Азиатцева В., Савкин И., Кислицин А., Семикина Е., Гребенников Д., Алешкин А., Сулимов А., Сулимов В., Бочаров Г. Прогнозирование специфического гуморального иммунного ответа на основании исходных параметров иммунного статуса детей, привитых против кори, краснухи и эпидемического паротита // Иммунология. — 2015. — Т. 36, № 1.

3 Генс Г., et al.,. Применение генных сигнатур и медицинских экспертных систем для прогнозирования клинических исходов рака молочной железы // Вестник РОНЦ им. Н. Н. Блохина РАМН. — 2015-2016. — Т. 26-27, № 4-1. — С. 47 – 52.

4 Sulimov A. V., Meshkov A. N., Savkin I. A., Katkova E. V., Kutov D., Hasanova Z. B., Konovalova N. V. e., Kukharchuk V. V., Sulimov V. B. Genome-wide analysis of genetic associations for prediction of polygenic hypercholesterolemia with Bayesian networks // Journal of Computational and Engineering Mathematics. — 2016. — Т. 2, № 4. — С. 11-26.

5 Samokhodskaya L., Starostina E., Sulimov A., Krasnova T., Rosina T., Avdeev V., Savkin I., Sulimov V., Mukhin N., Tkachuk V. Prediction of

features of the course of chronic hepatitis C using Bayesian networks // Terapevticheskii arkhiv. — 2019. — T. 91, № 2. — C. 32-39.

6 Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A., Bender D., Maller J., Sklar P., De Bakker P. I., Daly M. J. PLINK: a tool set for whole-genome association and population-based linkage analyses // The American journal of human genetics. — 2007. — T. 81, № 3. — C. 559-575.