СПОСОБ КЛАСТЕРИЗАЦИИ ФАКТОРОВ РИСКА СВЯЗАННЫХ С АТЕРОСКЛЕРОЗОМ ЗАБОЛЕВАНИЙ

В.Г. Вилков, С.А. Шальнова, А.Д. Деев
ФГБУ "Государственный научно-исследовательский центр профилактической медицины", Москва

Резюме. *Разработан новый способ расчета риска фатальных событий в самоорганизующихся кластерах факторов риска. Проведен вторичный анализ данных проспективных исследований неорганизованных когорт России и США (the First National Health and Nutrition Examination Survey – NHANES I, NHANES I Epidemiologic Followup Study – NHEFS, the Second National Health and Nutrition Examination Survey – NHANES II, NHANES II Mortality Study – NH2MS), общее число наблюдений 26840 человек, длительность наблюдения до 28 лет.*
Ключевые слова*: Сердечно-сосудистые заболевания, суммарный сердечно-сосудистый риск, общий риск, фактор риска, моделирование, искусственные нейронные сети, Россия, США.*
Key words*: Cardiovascular disease, total risk, risk factor, model analysis, artificial neural networks, Russia, United States of America.*

Концепция факторов риска (ФР) в последние десятилетия является основополагающей для многих медицинских дисциплин и в значительной степени определяет профилактические, лечебные и диагностические стратегии в отношении ряда заболеваний, в частности связанных с атеросклерозом. У индивида нередко сочетаются несколько ФР, многие из которых взаимосвязаны. Следовательно, при изучении ФР необходимо использовать многомерные статистические методы.

Современные знания о ФР, включая способы прогноза неблагоприятных исходов [1, 2, 3, 4], получены, как правило, с использованием методов статистического анализа, подразумевающих линейность зависимостей и нормальный закон распределения величин показателей. Однако к настоящему времени накоплены данные, свидетельствующие о том, что для сложных медико-биологических систем характерны многообразие связей между элементами, нелинейность этих связей и ненормальные распределения величин показателей [5, 6]. В частности, среди 23 характеризующих ФР сердечно-сосудистых заболеваний (ССЗ) показателей не обнаружено ни одного, распределение величин которого строго подчинялось нормальному закону [7].

Мы предположили, что использование нелинейных статистических методов может способствовать обнаружению новых фактов и уточнению оценок риска.

Цель настоящего исследования – разработка нового способа оценки суммарного риска развития фатального события.

Методика

Использовали данные проспективных российских исследований, выполненных в Государственном научно-исследовательском центре профилактической медицины [8, 9, 4] и нескольких исследований неорганизованной популяции США: the First National Health and Nutrition Examination Survey (NHANES I), NHANES I Epidemiologic Followup Study (NHEFS), the Second National Health and Nutrition Examination Survey (NHANES II), NHANES II Mortality Study (NH2MS) [10, 11, 12, 13, 14, 15, 16, 17, 18]. В настоящей работе использованы данные обследования лиц в возрасте 35 лет и старше (только белых), всего 15355 мужчин (6935 смертей на 251 тысячу «человеко-лет» наблюдения) и 11485 женщин (2694 смерти на 174 тысячи «человеко-лет» наблюдения). Анализировали смертность от ССЗ и общую смертность раздельно в российской и американской когортах. Причины смерти кодировались по международной классификации болезней 8 (Россия) или 9 (США) пересмотра. Первичные обследования проводились в 1971-1982 гг., последняя информация о жизненном статусе обследуемых получена в 2002 г. (Россия) и 1992 г. (США).

После предварительного анализа принято решение использовать в настоящей работе следующие характеризующие ФР переменные: возраст, пол, статус курения, уровень образования, величины систолического артериального давления (САД) и частоты сердечных сокращений (ЧСС).

Предлагаемый нами подход состоит в использовании прямого расчета рисков в однородных в отношении ФР (с учетом их взаимосвязей) частях когорты. Это достигается посредством выделения в многомерном пространстве *«самоорганизующихся кластеров факторов риска» (СОК ФР)**, объединяющих лиц с близкими величинами показателей, характеризующих учтенные ФР. Для кластеризации с неуправляемым обучением (классификацией без учителя) применяли искусственные нейронные сети (ИНС), а именно самоорганизующиеся карты Кохонена [19, 20, 21, 22, 23, 24, 25, 26, 27, 28]. Среду нейрокомпьютера эмулировали с использованием программы-имитатора [22].

Для сравнения с результатами традиционного подхода были рассчитаны риски фатальных событий для тех же когорт с использованием модели пропорциональных интенсивностей Кокса [29, 30, 31].

Результаты исследования и их обсуждения

Мы использовали прямой расчет

**В качестве англоязычного эквивалента нашего термина предлагаем "Self-Organizing Clusters of Risk Factors (SOC RF)"*

рисков в однородных многомерных СОК ФР вместо моделирования с неизбежно присущими последнему ограничениями, упрощениями и искажениями [32]. Для каждого СОК ФР по реальным данным рассчитывается абсолютный риск фатального события, величина которого в дальнейшем соотносится с соответствующими ячейками таблицы риска типа представленной на рис. 1.

Применение ИНС обеспечивает следующие принципиальные преимущества: ИНС представляют собой исключительно мощный метод моделирования, позволяющий воспроизводить зависимости практически любой сложности; они нелинейны по своей природе, тогда как до сих пор рассматриваемая задача решалась средствами линейного моделирования; по сравнению с традиционными методами ИНС лучше работают при большом числе переменных; они обучаются на примерах, не требуя априорных знаний о характере связей между входными данными и результатом, в нашем случае – о механизме влияния ФР на выживаемость.

Однако применение ИНС сопряжено с рядом трудностей. В отличие от методов традиционной статистики при использовании ИНС для конкретной практической задачи может быть построено много работоспособных и удовлетворяющих формальным критериям точности сетей, и, соответственно, найдено не одно, а несколько решений, поэтому возникает проблема выбора лучшей сети. Применительно к кластеризации и сетям Кохонена выбор усложняется отсутствием формальных критериев оценки успешности решения поставленной задачи данной конкретной сетью. Имеется также ряд технических вопросов (размер сети, ее конфигурация, выбор параметров пре- и постпроцессирования и др.), не имеющих на данный момент общепринятых способов решения, однако сильно влияющих на конечный результат.

Попытки найти способ выбора единственной лучшей сети Кохонена не дали удовлетворившего нас результата. Полезной находкой, использованной в дальнейшем, оказался критерий оценки качества сети, основанный на стабильности результатов осуществляемой с ее помощью кластеризации в нескольких случайных подвыборках. Работа сети признавалась удовлетворительной, если величины характеризующих ФР и исходы показателей в данном выделенном с ее использованием кластере значимо не различались в нескольких случайных подвыборках. Однако описанная техника не позволяет выбрать единственную сеть.

Проблема была решена следующим образом. Не пытаясь найти единственную наилучшую сеть, для

каждой задачи обучали большое количество подходящих по размерам сетей Кохонена. Из их числа отбирали несколько перспективных сетей, учитывая известные эмпирические правила построения ИНС [21] и описанный выше критерий, основанный на стабильности результатов кластеризации. Дальнейшие действия производили в 2 этапа.

На первом этапе работали с каждой отобранной сетью по отдельности. Вначале определяли номера кластеров, соответствующих ячейкам таблицы риска, образец которой представлен на рис. 1. Для этого через сеть прогоняли набор условных наблюдений, в которых величины учитываемых в таблице ФР соответствуют «центрам» ячеек таблицы. Например, для представленной на рис. 1 таблицы риска через сеть необходимо прогнать 96 условных записей. В первой записи значения ФР: пол – женский, САД 120 мм рт.ст, ЧСС 70 уд./мин, возраст 40 лет, статус курения – некурящие, образовательный уровень – среднее или ниже; во второй записи: пол – женский, САД 120 мм рт.ст, ЧСС 90 уд./мин, возраст 40 лет, статус курения – некурящие, образовательный уровень – среднее или ниже; и т.д. В результате каждую условную запись сеть отнесет к одному из кластеров и всем ячейкам таблицы риска можно будет присвоить номера соответствующих кластеров.

Затем через сеть прогоняли весь набор реальных данных и для каждого кластера рассчитывали риск фатального события, этот риск соотносили с ячейками таблицы риска, относящимися к данному кластеру. В итоге каждой ячейке таблицы будет соответствовать рассчитанный с использованием данной сети риск, выраженный как отношение числа умерших в течение указанного периода времени (числитель) к числу подверженных риску (знаменатель).

Описанную процедуру повторяли с каждой из отобранных сетей для сердечно-сосудистой и общей смертности и периодов наблюдения 10 и 20 лет.

На втором этапе производили усреднение результатов для каждой ячейки таблицы риска. Для этого суммировали для данной ячейки таблицы величины числителей и знаменателей, рассчитанные для всех использованных на первом этапе сетей, после чего посредством операции деления суммы величин числителей на сумму величин знаменателей вычисляли риск.

Описанный подход позволил обойти принципиальную трудность, связанную с выбором единственной лучшей сети, а также избежать чрезмерных колебаний величин риска в близких ячейках таблицы. Имеются и другие плюсы, связанные со спецификой нейросетевых методов, в

частности становится возможным сочетать сети с разным числом нейронов и др. характеристиками, каждая из которых привносит в конечный результат свои преимущества. Увеличивая количество используемых на первом этапе сетей, можно добиться требуемого сглаживания величин риска при перемещении по таблице. В данной работе при построении каждой таблицы использовали от 16 до 20 сетей.

Описанным выше способом были построены таблицы рисков обусловленных ССЗ и всеми причинами фатальных событий в течение 10 и 20 ближайших лет для когорт России и США, всего 16 таблиц (рис. 1 – 16).

Недостатком нашего способа является снижение надежности расчета риска для ячеек таблиц, которым соответствует малое число реальных наблюдений в когорте. Однако эта проблема в неявном виде присутствует и при использовании основанных на моделировании способах-прототипах. Малое число лиц с данным сочетанием значений ФР в выборке при любом способе оценки риска обусловит ненадежный и нестабильный результат. В зависимости от техники расчета риска это может проявляться явно (при нашем способе) или маскироваться за счет усреднения, интер- и экстраполяции (при регрессионных методах). В настоящей работе использовали следующее решение – в ячейках таблиц с числом наблюдений менее 10 рассчитанные посредством кластерной техники величины риска были заменены на полученные с использованием соответствующей модели регрессии (см. ниже). Величины риска в таких ячейках цветных таблиц (рис. 1 – 16) маркированы синим цветом.

Известные способы прогнозирования риска основаны на многомерных регрессионных моделях [1, 2, 3, 4, и др.]. Преимуществами такого подхода являются наличие хорошо разработанных статистических методов, относительная простота математического аппарата, наглядность, непротиворечивость и удобство интерпретации результатов. Поскольку модель базируется на всей совокупности наблюдений, число которых в популяционных исследованиях весьма велико, обеспечивается высокая статистическая достоверность. Однако это достигается за счет минимизации ошибок для наблюдений с наиболее часто встречающимися сочетаниями величин ФР, тогда как для относительно редко встречающихся наблюдений ошибки прогноза могут быть большими. Недостатки данного подхода обусловлены налагаемыми линейным моделированием ограничениями, влекущими, в частности, утрату нелинейных составляющих связей между переменными.

Одним из прототипов предлага-

Рис. 1-16

емого нами способа является российская шкала 10-летнего риска смерти от ССЗ на базе регрессионной модели SCORE, построенная с использованием той же когорты [4]. Поскольку в нашей работе набор ФР отличается наличием ЧСС и отсутствием концентрации общего холестерина в крови, для сравнения нового и традиционного подходов с применением последнего были построены регрессионные модели и на их основе рассчитаны таблицы 10- и 20-летних сердечно-сосудистых и общих рисков для когорт России и США, которые далее считали ближайшим прототипом (рис. 17 – 32). Как и следовало ожидать, по данным регрессионных моделей связи риска с возрастом, полом, статусом курения, уровнем образования и величинами САД существенно не отличались от известных [2, 4]. Связь ЧСС с суммарным риском по данным регрессионных моделей однозначная – при тахикардии риск выше, на иллюстрациях (рис. 17 – 32) в подгруппах с очень низким риском этот эффект может маскироваться вследствие округления величин риска до целых значений.

Разработанный нами способ (рис. 1 – 16) позволил выявить ряд особенностей влияния ФР на суммарный риск в сравнении с базирующимся на регрессионной технике прототипом (рис. 17 – 32).

Курение. По данным основанных на регрессионных моделях таблиц (рис. 17 – 32) у курящих по сравнению с некурящими риск в 1,5-2 раза выше вне зависимости от популяции, пола, возраста, образовательного уровня, величин САД и ЧСС (за исключением подгрупп с очень низким, близким к нулю суммарным риском). Такое единообразие может быть обусловлено техническими особенностями линейного моделирования. Более корректным будет утверждение, что для большинства включенных в когорту лиц курение при 20-летнем наблюдении в среднем в 1,5-2 раза повышает сердечно-сосудистый и общий риски. Этого в принципе достаточно для общего вывода о важной и безусловно негативной роли данного ФР, и соответствующих практических рекомендаций, ориентированных на популяцию в целом.

Наш способ позволил дать более дифференцированную оценку влияния курения на суммарный риск (рис. 1 – 16). В частности, выявлены категории лиц, у которых значимость курения особенно велика – риск у курящих более чем в 2 раза (до 9 раз) выше в сравнении с некурящими. Это женщины без высшего образования с нормальной ЧСС 40-50 лет с нормальным или пограничным САД (Россия) или 50 лет вне зависимости от уровня САД (США); в российской когорте лица 40 лет с высшим обра-

Рис. 17-32

зованием, тахикардией, нормальным или пограничным САД (у женщин данный эффект проявляется только при 20-летнем наблюдении).

Систолическое АД оказывает сильное и однозначное влияние на смертность по данным обоих способов – по мере повышения САД риск увеличивается. Зависимость вычисленных по регрессионным моделям рисков от САД близка к линейной (рис. 17 – 32), что согласуется и с данными литературы [4, 2]. По результатам нашего способа (рис. 1 – 16) данная зависимость наблюдается как для общей, так и сердечно-сосудистой смертности, у женщин и у мужчин, однако использование нелинейных методов позволило выявить ряд существенных особенностей – в ряде случаев обращает внимание более выраженный в сравнении с регрессионными моделями прирост риска при САД 180 мм рт.ст. (иногда 160-180 мм рт.ст.) по отношению к предыдущим градациям САД. Например, при изменении САД от 160 до 180 мм рт.ст. 10-летний сердечно-сосудистый риск в российской когорте у некурящих женщин со средним или ниже образованием и нормальной ЧСС увеличился с 9 до 20% в возрасте 60 лет и с 3 до 7% в возрасте 50 лет (рис. 1 – 16). По данным регрессионного метода (рис. 17 – 32) в соответствующих категориях лиц прирост риска не превышает 2%, сходные результаты демонстрирует и таблица SCORE [4]. Выраженный прирост риска при повышении САД до 160-180 мм рт.ст. обнаруживается преимущественно у некурящих с нормальной ЧСС при умеренных величинах рисков (рис. 1 – 16). Можно предположить, что при сочетанном воздействии высоких уровней других ФР данный эффект становится менее заметным.

Частота сердечных сокращений. Рассчитанный по регрессионным моделям риск фатальных событий во всех случаях выше у лиц с тахикардией (рис. 17 – 32).

Предлагаемый нами способ (рис. 1 – 16) в большинстве случаев демонстрирует сходный результат, однако в 1/4 случаев связи между ЧСС и суммарным риском не столь однотипны – наблюдаются небольшие (максимум ±5%) разнонаправленные различия суммарного риска в подгруппах с нормальной ЧСС и тахикардией. Это характерно для лиц с низким суммарным риском (не более 2-3%); нормальным САД; а также некурящих женщин с образованием выше среднего (Россия). Вышеизложенное в целом согласуется с данными ряда эпидемиологических исследований, в которых выявлены ассоциации смертности с ЧСС, наблюдавшиеся, однако, не во всех группах обследованных лиц [33, 34].

Оценка сочетаний ФР представ-

ляет особый интерес. В известных способах, в частности основанных на системе SCORE, по существу используется суммирование эффектов ФР: по мере повышения АД риск увеличивается по линейному закону, курение удваивает риск и т.п. [4]. Использованные нами нейросетевые методы нелинейны и в принципе не имеют ограничений, свойственных упомянутым выше прототипам. Это позволило выявить неочевидные ранее особенности.

• Зависимость суммарного абсолютного риска от таких ФР, как пол, возраст, статус курения, уровень образования, величины САД и ЧСС не является столь однозначной, как считалось ранее, в частности исходя из системы SCORE.

• Как при низком (до 2-3%), так и при очень высоком суммарном риске он относительно слабо связан с уровнями отдельных ФР.

• При средних для данного вида смертности и срока наблюдения уровнях суммарного риска особенности влияния на смертность отдельных ФР проявляются сильнее. В частности, увеличение САД более 170 (в части случаев более 150) мм рт.ст. влечет более резкий прирост риска в сравнении с предыдущими градациями САД.

• В когорте США в сравнении с российской смертность в целом ниже, различия наиболее выражены у мужчин в возрасте до 55 лет. У женщин и мужчин старше 55 лет различия смертности в указанных когортах выражены меньше.

Построенные разработанным нами способом таблицы риска демонстрируют согласующиеся друг с другом результаты в российской и американской когортах (рис. 1 – 16). Это свидетельствует о стабильности и воспроизводимости результатов предлагаемой статистической техники и косвенно подтверждает корректность примененного нами подхода.

Заключение

Разработанный нами способ оценки риска в многомерных самоорганизующихся кластерах ФР (СОК ФР) и традиционный подход, основанный на моделировании дожития, имеют свои плюсы и минусы. При разработке профилактических стратегий, ориентированных на популяцию в целом, представляются более удобными основанные на использовании регрессии системы типа SCORE, при этом свойственные им эффекты усреднения не являются помехой, а простота и однозначность интерпретаций влияния отдельных ФР представляют собой важное преимущество. Наш способ более перспективен в случаях, когда требуется дать дифференцированные рекомендации, учитывающие особенности конкретной подгруппы.

В частности, на его основе можно сформулировать некоторые дополнения к общепринятым [4, 2] рекомендациям:

У лиц с высоким суммарным риском фатальных событий целесообразно стремиться в первую очередь скорректировать наиболее значимые для прогноза ФР, а именно повышение САД более 170 (иногда более 150) мм рт.ст. и, в части подгрупп, курение.

После достижения среднего для данных возраста и пола уровня суммарного риска для максимальной эффективности профилактических мероприятий последние могут быть дифференцированы в зависимости от особенностей подгруппы с тем, чтобы сконцентрировать усилия на коррекции наиболее значимых для данного контингента ФР. В одних случаях более перспективно дальнейшее снижение уровня АД, в других – нормализация повышенной ЧСС, либо сочетание указанных мероприятий.

Лица, для которых характерна сильная связь суммарного риска с ЧСС, составляют значительную часть популяции. Для этого контингента снижение ЧСС может оказаться одним из приоритетных направлений профилактики. При наличии артериальной гипертонии перспективны антигипертензивные лекарственные средства, обладающие отрицательным хронотропным эффектом. У нормотоников можно рекомендовать немедикаментозные способы, в частности повышение уровня физической активности.

ЛИТЕРАТУРА

1. *Kannel W.B., McGee D., Gordon T.* // Am. J. Cardiol. - 1976. - V. 38. - N 1. - P. 46-51.
2. Европейские рекомендации по профилактике сердечно-сосудистых заболеваний в клинической практике. Кардиоваскулярная терапия и профилактика. - 2004. - Т. 3. - № 4. - С. 99-112.
3. *Elousua R., Marrugat J.* // Кардиоваскулярная терапия и профилактика. - 2002. - Т. 1. - № 1. - С. 84-85.
4. *Шальнова С.А., Оганов Р.Г., Деев А.Д.* // Кардиоваскулярная терапия и профилактика. - 2004. - Т. 3. - № 4. - С. 4-11.
5. *Hall J.C.* // Br. J. Surg. - 1982. - V. 69. - N 1. - P. 55-56.
6. *Максимов Г.К., Синицын А.Н.* Статистическое моделирование многомерных систем в медицине. - Л.: Медицина, 1983.
7. *Вилков В.Г.* Сердечно-сосудистый риск и дисфункция щитовидной железы (по данным популяционных исследований). - М.: Издатель "Гайнуллин", 2004.
8. *Константинов В.В., Жуковский Г.С., Тимофеева Т.Н.* // Кардиология. - 1996. - № 1. - С. 37-41.
9. *Шальнова С.А., Деев А.Д., Шестов Д.Б.* // Кардиология. - 1997. - № 9. - С. 49-54.
10. *Cox C.S., Mussolino M.E., Rothwell S.T., et al.* // Vital Health Stat. - 1997. - V. 1. - N 35.
11. http://www.cdc.gov/nchs/about/major/nhefs/nhefspuf.htm
12. *Miller H.W.* // Vital Health Stat. - 1973. - V. 1. - N 10a,b.
13. *Engel A., Murphy R.S., Maurer K., et al.* // Vital Health Stat. - 1978. - V. 1. - N 14.
14. http://www.cdc.gov/nchs/about/major/nhanes/nhanesi.htm
15. *Loria C.M., Sempos C.T., Vuong C.* // Vital Health Stat. - 1999. - V. 1. - N 38.
16. http://www.cdc.gov/nchs/r&d/nchs_datalinkage/nhanesii_data_linkage_activities.htm
17. National Center for Health Statistics. Plan and operation of the Second National Health and Nutrition Examination Survey, 1976-80. - Vital Health Stat. - 1981. - V. 15.

18. http://www.cdc.gov/nchs/about/major/nhanes/nhanesii.htm
19. *Kohonen T.* Self-Organizing Maps. - Heidelberg: Springer-Verlag, 1995.
20. *Mueller J.-A., Lemke F.* Self-Organising Data Mining. - Hamburg: Libri, 2000.
21. Нейронные сети: STATISTICA Neural Networks. - М.: Горячая линия - Телеком, 2001.
22. *Круглов В.В., Борисов В.В.* Искуственные нейронные сети: Теория и практика. - М.: Горячая линия - Телеком, 2001.
23. *Медведев В.С., Потемкин В.Г.* Нейронные сети: MATLAB 6. - М.: ДИАЛОГ-МИФИ, 2002.
24. http://www.cis.hut.fi/research/som-bibl/
25.http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml
26. http://www.cis.hut.fi/research/som_pak/
27.http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml
28. http://www.orc.ru/%7estasson/n3.zip
29. *Cox D.R.* // J Royal Statistical Society. - 1972. - V. 34. - P. 187-220.
30. *Кокс Д.Р., Оукс Д.* Анализ данных типа времени жизни. - М.: Финансы и статистика, 1988.
31. *Боровиков В.* STATISTICA: искусство анализа данных на компьютере. Для профессионалов. - СПб.: Питер, 2001.
32. http://www.megaputer.ru/download/book.zip
33. *Маколкин В.И., Зябрев Ф.Н.* // Кардиоваскулярная терапия и профилактика. - 2006. - Т. 5. - № 6. - С. 5-9.
34. *Шальнова С.А., Деев А.Д., Оганов Р.Г., и др.* // Кардиология. - 2005. - № 10. - С. 45-50.