Известия Гомельского государственного университета
имени Ф. Скорины, № 3 (96), 2016

ИНФОРМАТИКА

УДК 681.32.001

Автоматизация оценки риска смертности человека с использованием дискриминантного анализа

Н.Б. ОСИПЕНКО[1], А.Н. ОСИПЕНКО[2]

Описывается один из вариантов трехэтапного скрининга здоровья населения. Предлагается способ использования паспортных данных для предварительной экспресс-диагностики групп риска обследуемого человека. Для решения задачи распознавания двух групп риска смертности использованы три метода: дискриминантного анализа, голосования и коллектива решающих правил. На примере пробного исследования показывается практическая целесообразность применения этих данных на первом этапе скрининга.
Ключевые слова: паспортные данные, распознавание образов, дискриминантный анализ данных, риск смертности.

One of options of three-stage screening of health of the population is described. A method of using the passport data for preliminary express diagnostics of risk groups of the examined person is offered. For the solution of a problem of recognition of two risk groups of mortality three methods are used: discriminant analysis, vote and collective of decisive rules. On the example of trial research practical expediency of application of these data at the first stage of screening is shown.
Keywords: passport data, recognition of images, discriminant analysis of data, risk of mortality.

Введение. Задача выявления причин смертности и, в частности, особенностей основных групп риска по статистическим данным является одной из ведущих в сфере санитарно-гигиенических исследований. Особенность современного момента состоит в том, что появляются технические возможности для массового скрининга здоровья населения [1], [2]. Разрабатываются и внедряются различные концепции такого скрининга по поводу ранней диагностики и профилактики онкозаболеваний, туберкулеза, сердечнососудистых и других заболеваний. В связи с этим возникает проблема согласования и синхронизации всех этих обследований в рамках единой концепции общереспубликанского скрининга здоровья населения.

В настоящей работе предлагается для этих целей разработать методологию и программно-технологическое обеспечение предварительной экспресс-диагностики на основе паспортных сведений, данных о группе крови, антропометрии, анамнезе и иной стандартной информации, имеющейся в поликлинических базах данных. В перспективе к этой информации могут быть добавлены данные биометрии (отпечатки пальцев, фотографии сетчатки глаза), фрагменты почерка. В еще более далекой перспективе – данные недорогого экспресс-анализа ДНК. В результате такой диагностики для каждого человека будут сформированы оценки вероятностей принадлежности к основным группам риска.

На втором этапе скрининга должно осуществляться подробное анкетирование граждан по поводу соответствующих целевых проблем в наиболее вероятных для них группах риска.

Наконец, только на третьем этапе предполагается перейти к более дорогостоящему детальному клиническому и амбулаторному обследованию (если в этом появляется необходимость).

Обоснование целесообразности использования паспортных данных в статистическом анализе причин смертности населения. Наиболее спорным в предлагаемой трехэтапной схеме массового скрининга является первый этап. Основной довод – высокие ошибки диагностики. Как показывает проведенный нами пробный анализ статистических связей паспортных данных и основных причин смерти [2], такие связи имеются, и они вполне пригодны для предварительного распределения жителей по группам риска. Что касается ошибок этой диагностики, то, во-первых, окончательное решение об отнесении человека к группе риска (или к нескольким группам риска) на завершающей фазе первого этапа должен принимать участковый терапевт. Ре-

зультаты компьютерной обработки по этому человеку носят статус дополнительной ориентирующей информации. Они позволят специалисту быстрее сфокусироваться на потенциально более слабых сторонах здоровья обследуемого. Особенно это касается молодежи и лиц среднего возраста, практически никогда самостоятельно не обращавшихся за медицинской помощью.

Во-вторых, ошибки первого этапа скрининга (отнесение не к своей группе риска) могут быть выявлены на втором этапе. В этой ситуации обследуемому придется провести дополнительное анкетирование по иной группе риска, подобранной для него врачом с учетом компьютерной обработки его первой анкеты. В любом случае, наличие первого этапа скрининга позволит уйти от обременительного анкетирования сразу по всем группам риска. В оптимальном варианте их должно быть не меньше десятка.

Описание исходных данных для статистического анализа. Для решения задачи оценки риска смертности от некоторых заболеваний (в нашем случае два вида: с быстрой или медленной потерей трудоспособности) поиск исходной информации осуществлялся в Интернете, и поэтому многие данные, например, показатели медицинских обследований, которые обычно применяют в задачах такого рода, использованы быть не могли в связи с проблемами их доступности. Но схема данного исследования может быть использована в качестве прототипа решения задачи оценка риска смертности.

В качестве исходного материала для пробного исследования возможностей использования паспортных сведений на первом этапе скрининга послужила выборка из 106 знаменитых людей. Данные подготавливались в системе Excel. По каждому человеку из Интернета брались данные в виде строки: имя; фамилия; страна, в которой жил человек; тип менталитета; день, месяц и год рождения; день, месяц и год смерти; количество жен (мужей); количество детей (включая приемных); основной род занятий по жизни (профессия); общая причина смерти (заболевание сердечнососудистой системы – инфаркт или инсульт – 1, онкозаболевание – 2, другие болезни – 3, несчастные случаи – 4, преднамеренное убийство – 5, иное – 6); детализация подсистем организма в структуре общей причины смерти:

(1) причина смерти (сердечнососудистые заболевания): сердце (инфаркт и др), мозг (инсульт и др.);

(2) причина смерти (онкозаболевания) система: нервная (мозг и т. д.), эндокринная (железы: щитовидная, поджелудочная и т. д.), дыхательная, пищеварительная, мочевая (почки и т. д.), половая, опорнодвигательная (кости), кроветворная и иммунная, органов чувств (глаза, кожа, уши, гортань, нос), иное, включая: лимфа, селезенка, мягкие ткани, ткани сосудов;

(3) причина смерти (хронические болезни): детализация подсистем организма та же, что в причине смерти (2) (онкозаболевания);

(4) причина смерти (несчастные случаи): авария, смертельные травмы, самоубийство, отравление.

Для обработки данных использовался пакет «Statistica», в частности его средства подготовки новых переменных путём того или иного преобразования исходных признаков, а также программы анализа вариантов, построения гистограмм, матриц корреляций признаков, классификации данных и дискриминантного анализа. Кроме того, в рамках этого пакета была написана программа обучения распознаванию групп риска.

Входной информацией в методе дискриминантного анализа системы Statistica являются: количество супругов, количество детей, значения элементов психоматрицы человека, рассчитанной по алгоритму квадрата Пифагора (КП1–КП9), тип заболевания. Исходными данными для задачи распознавания групп риска явились полученные компоненты десятимерного вектора психоматрицы человека по квадрату Пифагора и другие признаки, описывающие человека, такие, как пол, возраст, количество детей и др.

Перевод даты рождения, имени и фамилии в качественные признаки. Перевод даты рождения в набор числовых качественных признаков осуществлен с помощью общеизвестного алгоритма Пифагора и описан в работе [3]. В итоге из даты рождения получаются следующие признаки: число (цифра) даты (получено путем поэтапного сложения всех ее цифр пока в сумме не получится одна цифра); количества встречаемости цифр 0, 1, … ,9 в рабочих числах алгоритма Пифагора [3]; образованные из предыдущих десяти номинальных признаков бинарные признаки, показывающие наличие или отсутствие, например, ноль двоек, одной двойки и т. д. в квадрате Пифагора.

Для перевода имени и фамилии в признаки для распознавания групп риска нумерология предлагает множество различных соответствий букв и цифр. В настоящей работе использована числовая азбука [4], с помощью которой можно вместо буквы ставить числа: А-1, Б-2, В-3, Г-4, Д-5, Е-6, Ж-7, З-8, И-9, К-10, Л-20, М-30, Н-40, О-50, П-60, Р-70, С-80, Т-90, У-100, Ф-200, Х-300, Ц-400, Ч-500, Ш-600, Щ-700, Ю-800, Я-900, Э-1000. В работе [4] приведена таблица содержательных значений этих числовых кодов. Например, для известного баснописца Ивана Крылова кодирование имеет следующий вид: И–9, В–3, А–1, Н–40. В сумме 53; К–10, Р–70, Ы–0, Л–20, О–50, В–3. В сумме 153. Складываем обе полученные суммы (53+153=206). Полученное число ищем в приведенной в [4] таблице. Но так как точно числа 206 в ней нет, то его следует разбить на 200 + 6. Для числа 200 находим: хладнокровие, слабохарактерность, для 6–труд, свободолюбие, успех. Как видно из биографии знаменитого баснописца, все эти качества были налицо. Теперь поступим с числом 206 иначе. Сложим его цифры до двузначного или однозначного вида. И снова обратимся к таблице содержательной интерпретации чисел: 8 – величие, кротость, справедливость.

При наличии обучающей выборки в несколько сотен тысяч человек можно было бы воспользоваться полностью описанной здесь кодировкой. В нашем же пробном примере применялась только однозначная кодировка. При этом были образованы группы бинарных признаков для имени, для фамилии и для имени вместе с фамилией. Соответствующие числа (цифры) преобразовывались в последовательность из девяти бинарных признаков. Например, у семерки на седьмом месте в такой последовательности стоит единица, а на других местах – нули.

Алгоритм распознавания групп риска методом голосования. В связи с небольшим объемом пробной выборки вместо четырех общих групп риска были выделены две. В первую группу вошли люди, умершие от сердечнососудистых заболеваний и от несчастных случаев. Для нее характерна быстрая потеря жизнеспособности организма. Во вторую группу вошли те, у кого такая потеря шла постепенно – это умершие от онкозаболеваний и других хронических болезней. Также из-за недостаточного объёма исходной выборки не удалось осуществить полноценное выделение ее однородных подвыборок, в частности использовать для этого признаки пола, менталитета и продолжительности жизни (ранние смерти и смерти в зрелом возрасте). Сам алгоритм распознавания групп риска разбит на два этапа.

Этап 1. Обучение распознаванию групп риска.

1.1. Построение матрицы взаимных корреляций по всем признакам, включая целевой признак принадлежности к первой или второй обобщенным группам риска.

1.2. Выбор признаков, имеющих более или менее значимую связь с целевым свойством (в качестве критерия выбора информативного признака в нашем случае бралось условие для коэффициента корреляции: $|r| \geq 0{,}1$). Положительная корреляция говорит о том, что признак в большей степени «голосует» за вторую группу риска и, наоборот, отрицательная корреляция говорит о предпочтении первой группы риска.

1.3. Формирование дискриминантной функции.

Пусть $P_1 = \{p_1^1, p_2^1, ..., p_{n_1}^1\}$ – множество признаков, голосующих за первую группу риска, а $P_2 = \{p_1^2, p_2^2, ..., p_{n_2}^2\}$ – множество признаков, голосующих за вторую группу риска. Вероятность B_1 отнесения объекта *x* к классу 1 вычисляется по формуле:

$$B_1 = (\sum_{i=1}^{n_1} x_i^1 / n_1) / ((\sum_{i=1}^{n_1} x_i^1) / n_1 + (\sum_{j=1}^{n_2} x_j^2) / n_2),$$

где $x_i^1, (i = 1,..., n_1)$ – значения признаков из множества P_1 для объекта *x*; $x_j^2, (i = 1,..., n_2)$ – значения признаков из множества P_2 для объекта *x*.

Соответственно, $B_2 = 1 - B_1$ – вероятность отнесения объекта *x* к классу 2.

Этап 2. Экзамен алгоритма распознавания групп риска.

Обычно экзамен проводится для объектов, не участвовавших при построении дискриминантной функции. Принятие решения об отнесении объекта к группе риска имеет вид:

$$R = \begin{cases} 1, \text{если } B_1 - B_2 \geq \alpha; \\ 2, \text{если } B_2 - B_1 \geq \alpha; \\ 0, \text{если } |B_1 - B_2| < \alpha, \end{cases}$$

где $0 \le \alpha \le 0.5$ (в нашем примере α =0.2).

Если R = 0, то программа не может различить группу риска для объекта *x*.

Результаты распознавания групп риска. Результатом дискриминантного анализа системы Statistica являются функции классификации, построенные одним из её методов (стандартным). С помощью полученных классификационных функций можно вычислить значения удаленности от центра группы для произвольного человека и отнести его к группе с медленной (Slow) или быстрой (Rapid) потерей трудоспособности:

Slow = 13,6КП1 + 15,1КП2 + 17,4КП3 + 13,9КП4 + 14,8КП5 + 16,0КП6 + 13,3КП7 + 14,9КП8 + + 13,3КП9 + 1,6КС + 0,05КД-94,632,

Rapid = 14,0КП1 + 15,4КП2 + 17,8КП3 + 14,4КП4 + 14,8КП5 + 16,3КП6 + 14,2КП7 + 14,8КП8 + + 13,1КП9 + 1,4КС + 0,3КД-98,0136,

где КП1–КП9 – элементы психоматрицы квадрата Пифагора; КС – количество супругов; КД – количество детей.

Результатом обучения метода голосования, используя паспортные данные для распознавания двух групп риска заболеваний с медленной (Slow) или быстрой (Rapid) потерей трудоспособности: по обучающей выборке объема 106, явились два множества информативных признаков, приведенные в таблице 1. Для первого класса (быстрая потеря трудоспособности организма) и для второго (постепенная потеря трудоспособности организма) множества информативных признаков не пересекаются.

Таблица 1 – Результаты обучения метода голосования

Класс	КП1	КП2	КП3	КП4	КП5	КП6	КП7	КП8	КП9	Дети	СЦД	СЦИ	СЦФ	СЦИФ
1	3	0-1		0			≥ 2	≥ 2	≥ 2	≤ 1	9	5,6,7,8	3,5,9	1,4,5
2		≥ 2		≥ 1			≤ 1	≤ 1	1	≥ 2	4,5,7	1,3,9	1,4	2

Как видим, первую группу от второй отличают высокие потенциалы воли (КП1(3)), творчества (КП7($\ge$ 2)), внушаемости (КП8($\ge$ 2)), памяти (КП9($\ge$ 2)) и низкие потенциалы энергии (КП2(0-1)) и здоровья (КП4(0)). Для первой группы риска характерны также недостаток детей (Дети($\le$ 1)), суммарная цифра даты рождения 9, суммарная цифра имени 5, 6, 7 или 8, суммарная цифра фамилии 5 или 9, суммарная цифра имени и фамилии 1, 4 или 5. Отметим, что для второй группы риска характерны немалое количество детей в семье (Дети($\ge$ 2)) и отсутствие эмпатии (КП8($\le$ 1)). Все это способствует дисгармоничной семейной жизни и стрессам. По мнению многих специалистов, продолжительные стрессы и депрессивные состояния являются самым главным фактором онкозаболеваний.

В целом интерпретация этих признаков не противоречит имеющимся представлениям о различии первой и второй групп риска.

Итоговые интегральные результаты сравнения оценки смертности методами голосования и дискриминантного анализа данных системой Statistica приведены в таблице 2. Оценка качества дискриминантной функции проводилась по исходной выборке объема 106. При этом ошибка первого рода оценки смертности методами голосования и дискриминантного анализа данных системой Statistica (отнесение объекта первого класса ко второму) составила 0,16 и 0.04, ошибка второго рода – 0,01 и 0,16, доля отказов – 0,41 и 0,27, вероятность правильного распознавания – 0.42 и 0,53.

Таблица 2 – Результаты сравнения оценки смертности тремя методами

	Вероятность ошибки 1 рода	Вероятность ошибки 2 рода	Вероятность отказа	Вероятность правильного распознавания
Метод голосования	0,16	0,01	0,41	0,42
Метод дискриминантного анализа в системе Statistica	0,04	0,16	0,27	0,53
Метод коллектива решающих правил	0,1	0,11	0,12	0,67

Анализ полученных результатов показал, что точность оценки смертности, выдаваемой двумя методами голосования и дискриминантного анализа в системе Statistica, можно улучшить. Для этой цели был построен коллектив решающих правил на базе этих двух методов. Пусть R_{gol}, R_{discr} и R_{col} – решения по отнесению человека к группам с медленной (Slow) или быстрой (Rapid) потерей трудоспособности, полученные методами голосования, дискриминантного анализа системы Statistica и коллектива решающих правил. По полученным результатам двумя методами R_{gol}, R_{discr} – голосования и дискриминантного анализа системы Statistica построим коллектив решающих правил.

$$Rcol = \begin{cases} Rgol, \text{если } Rgol = 2 \ и \ Pgol(2) - Pgol(1) > Pdiscr(1) - Pdiscr(2), \\ Rdiscr, \text{если } Rdiscr = 1 \ и \ Pdiscr(1) - Pdiscr(2) > Pgol(2) - Pgol(1) \end{cases},$$

где $Pgol(1), Pgol(2)$ – вероятности отнесения к классам 1 (Slow) и 2 (Rapid) по методу голосования; а $Pdiscr(1), Pdiscr(2)$ – вероятности отнесения к классам 1 (Slow) и 2 (Rapid) по методу дискриминантного анализа системы Statistica.

Как видно из таблицы 2, использование коллектива решающих правил позволило получить более точный результат прогноза.

Заключение. Несмотря на небольшой объем выборки в проведенном пробном исследовании, можно утверждать, что паспортные данные вполне пригодны для включения их в список признаков при экспресс-диагностике группы риска на первом этапе скрининга здоровья населения.

В работе использованы три метода (дискриминантный анализ, реализованный в системе Statistica, голосование и коллектив решающих правил) для решения задачи распознавания двух групп риска смертности (с медленной или быстрой потерей трудоспособности), в основном от сосудистых или онкологических заболеваний для выборки знаменитых людей с известной общей причиной смерти.

Исследование представляет собой логически завершенный макет для разработки полноценного инструментария автоматизации сравнения методов оценки риска смертности. Доля ошибок первого и второго рода (на уровне 10 %) итогового метода распознавания для обучающей выборки с использованием в основном только паспортных данных, говорит о перспективности и целесообразности продолжения подобных работ, но уже с привлечением данных медицинской диагностики.

Литература

1. Большакова, Г.И. Построение модели факторов здоровья сельского населения по данным скринингового обследования / Г.И. Большакова [и др.] // Известия Гомельского гос. университета им. Ф. Скорины. – 2006. – № 4 (37). – С. 113–115.

2. Осипенко, Н.Б. Пример «выращивания» регрессионной модели социального явления на базе критерия правдоподобности ее интерпретации / Н.Б. Осипенко, А.Н. Осипенко, К.А. Осипенко // Проблемы физики, математики и техники. – 2013. – № 4 (17). – С. 85–88.

3. Осипенко, К.А. Метод регрессионного моделирования продолжительности жизни по дате рождения / К.А. Осипенко, Н.Б. Осипенко // Творчество молодых 2012: сборник научных работ студентов и аспирантов УО «ГГУ им. Ф. Скорины» : в 2 ч. / Гомельский гос. ун-т им. Ф. Скорины ; отв. ред. О.М. Демиденко. – Гомель, 2012. – Ч. 1. – С. 194–197.

4. Хигир, Б.Ю. Число имени / Б.Ю. Хигир. – СПб. : Астрель, 2008. – 42 с.

[1]Гомельский государственный
университет им. Ф. Скорины

[2]Гомельский государственный технический
университет им. П.О. Сухого

Поступила в редакцию 25.04.2016