УДК 681.518

А.И. Поворознюк, Н.А. Чикина, Ю.Л. Геворкян, И.В. Антонова

Национальный технический университет "ХПИ", Харьков

ПРОГНОЗ РАЗВИТИЯ ПРОФЕССИОНАЛЬНО ОБУСЛОВЛЕННЫХ ЗАБОЛЕВАНИЙ С ПОМОЩЬЮ ДИСКРИМИНАНТНОГО АНАЛИЗА

В статье предложен способ прогноза развития профессионально обусловленных аллергодерматозов у рабочих химико-фармацевтических предприятий на основе результатов дискриминантного анализа, учитывающий суммарное влияние внутренних факторов риска. Полученные результаты позволяют с достаточно высокой степенью точности различать состояния здоровья обследованных, что может быть использовано при построении экспертной системы прогноза риска развития профессионально обусловленных аллергодерматозов у рабочих химико-фармацевтических предприятий.

Ключевые слова: *прогноз развития заболевания, дискриминантный анализ, факторы риска, экспертная система.*

Введение

Постановка проблемы и анализ литературы. В условиях современного технологически развитого производства остается актуальной задача прогноза развития заболеваний, обусловленных действием профессиональных факторов риска. На предприятиях химико-фармацевтической промышленности к профессиональным рискам относится воздействие активных химических веществ, вызывающих различные заболевания аллергического генеза.

В ходе исследований, проводимых в рамках научно-исследовательских работ Харьковского НИИ Дерматологии и Венерологии совместно с НТУ "ХПИ" на предприятиях химико-фармацевтической промышленности, кроме профессиональных рисков развития аллергодерматозов у рабочих химико-фармацевтических предприятий авторами выявлены групповые и внутренние факторы риска [1 – 5]. Наличие таких факторов риска позволяет использовать их при решении задачи прогноза развития профессионально обусловленных аллергодерматозов.

При разработке способа прогноза возникла необходимость объединения выявленных внутренних факторов риска в один обобщенный внутренний фактор риска. Решение этой задачи позволит, с одной стороны, учесть суммарное воздействие внутренних факторов риска без потери информации о силе воздействия каждого из них, а с другой стороны, использовать его в качестве правила классификации обследованных в соответствии с оценкой риска развития профессионально обусловленных заболеваний.

Для построения такого обобщенного фактора риска был выбран дискриминантный анализ, который не только выявляет различия между группами объектов, но и помогает классифицировать объекты по принципу максимального сходства. Для того чтобы при решении задачи прогноза состояния обследуемого можно было учесть влияние выявленных внутренних факторов риска, необходимо построить решающие правила, относящие обследуемых к одному из возможных известных состояний (обучение с учителем). С помощью линейного дискриминантного анализа в соответствии с введенным множеством возможных состояний строятся разделяющие гиперплоскости, каждая из которых описывается уравнением канонической дискриминантной функции (ДФ) [6 – 8].

Целью статьи является разработка способа прогноза состояния обследуемых с помощью дискриминантных функций, построенных на основе значимых внутренних факторов риска развития профессионально обусловленных заболеваний у рабочих химико-фармацевтической промышленности.

Основная часть

Выбор классификационных признаков – одна из главных проблем методов классификации. Важность (информативность) признаков при классификации определяется целью исследования. В настоящих исследованиях целью классификации является синтез решающих правил для прогноза развития аллергодерматозов на основе выявленных в предыдущих исследованиях индивидуальных факторов риска [1, 2, 5]. Результатом дискриминантного анализа является построение ДФ

$$y = \sum_{i=1}^{n} b_i x_i + b_0 ,$$

где x_i $(i = \overline{1,n})$ – значения независимых переменных, соответствующих рассматриваемым случаям, b_i $(i = \overline{0,n})$ – коэффициенты (константы), которые оцениваются с помощью дискриминантного анализа.

Задачей дискриминантного анализа является определение таких коэффициентов b_i, чтобы по значениям ДФ можно было с минимальной ошибкой провести разделение всей совокупности на классы.

Если значения x_i нормированы, то чем больше значения коэффициентов b_i, тем больше вклад соответствующей переменной в дискриминацию совокупности. Если классов более двух, то строится несколько ДФ.

Пусть для описания множества Ω, состоящего из m объектов $\{A_1, A_2, ..., A_i, ..., A_m\}$, используется n признаков (выявленных факторов риска): $\{X_1, X_2, ..., X_j, ..., X_n\}$.

Каждому объекту A_i $(i = \overline{1,m})$ соответствует некоторое значение целевого признака X_0. Признаки X_j $(j = \overline{0,n})$ измерены в порядковой или номинальной шкале. Целевой признак X_0 имеет k_0 градаций, т.е. принимает k_0 различных значений: $X_0^{(1)}, X_0^{(2)}, ..., X_0^{(l_0)}, ..., X_0^{(k_0)}$. Тогда задачей классификации множества объектов Ω является отнесение каждого объекта $A_i \in \Omega$ $(i = \overline{1,m})$, задаваемого в виде совокупности его характеристик (признаков) $\{X_{1i}, X_{2i}, ..., X_{ji}, ..., X_{ni}\}$, к одному из k_0 взаимоисключающих классов (состояний) Ω_p $(p = \overline{1,k_0})$ множества Ω, т.е. таких Ω_p, что $\Omega_p \bigcap \Omega_q = \varnothing$, $\bigcup_p \Omega_p = \Omega$. Наличие соответствующих правил классификации D означает, что существует однозначное отображение конечной совокупности наблюдений Ω на множество $\{\Omega_p\} : \Omega \xrightarrow{D} \{\Omega_p\}$.

В настоящем исследовании база данных обследованных рабочих и служащих предприятия (365 человек) была разделена на три класса ($k_0 = 3$) по состоянию здоровья: Ω_1 – ГР1 (группа риска 1 – больные аллергодерматозами или имеющие аллергические заболевания в анамнезе); Ω_2 – ГР2,3 (группа риска 2, 3 – обследованные, имеющие различные соматические заболевания и заболевания кожи неаллергического генеза), Ω_3 – ПЗ (практически здоровые). Такое разбиение множества Ω соответствует значениям целевого признака X_0: $X_0^{(1)}$ = "ГР1", $X_0^{(2)}$ = "ГР2,3", $X_0^{(3)}$ = "ПЗ", т.е. значения целевой переменной отражают риск перехода в соответствующее состояние.

Дискриминантный анализ проводился средствами статистического пакета SPSS [9]. Поскольку при проведении дискриминантного анализа в пакете SPSS входные переменные должны быть дихотомическими, то соответствующие признаки были переведены из номинальной в дихотомическую шкалу ("1" – признак присутствует, "0" – признак отсутствует). Классификация проводилась по значениям целевого признака X_0.

В рамках пакета SPSS процедура дискриминантного анализа может проводиться двумя методами: 1) метод одновременного учета переменных, 2) пошаговый метод. В результате применения первого метода при наличии большого количества переменных некоторые из них могут оказаться лишними, т.е. не вносящими никакого вклада в анализ, а лишь усложняющими его, некоторые могут нести одинаковую информацию, будучи при этом хорошими дискриминаторами. Пошаговый метод является одним из способов исключения лишних переменных и состоит в использовании процедуры последовательного отбора наиболее полезных дискриминантных переменных, хотя полученное множество дискриминантных переменных может и не быть наилучшей их комбинацией. Для выбора наиболее оптимального по критерию качества классификации вида ДФ дискриминантный анализ проводился обоими методами.

Вначале были получены ДФ, разделяющие все три возможных состояния, соответствующих определенным выше классам. Однако качество дискриминации этой функции составило в среднем 48%, что является недостаточным для ее дальнейшего использования. Поэтому в дальнейшем все объекты были разделены на два класса. В первый класс вошли объекты со значением целевого признака $X_0^{(1)}$ ("ГР1"), а во второй были объединены объекты со значением целевого признака $X_0^{(2)}, X_0^{(3)}$ ("ГР2,3" и "ПЗ"), т.е. решалась задача построения правила классификации, по которому можно было бы выделить обследуемых с максимально высоким риском развития профессионально обусловленных аллергодерматозов.

В качестве дискриминантных переменных (из общего числа 7 признаков и 37 их градаций) в эту ДФ вошли следующие внутренние факторы риска: X_1 – "Пол" ("0" – мужской, а значение "1" – женский); $X_2^{(2)}$ – одна из градаций признака "Группа крови" со значением "Группа крови 2 (А)"; $X_3^{(3)}$ – одна из градаций признака "Конституционные особенности" со значением "Гиперстеник"; $X_5^{(4)}$, $X_5^{(8)}$, $X_5^{(10)}$ – градации признака "Дата рождения" с 12-месячной цикличностью (периоды с 22.06 по 22.07, с 24.10 по 22.11, а также с 22.12 по 20.01); $X_6^{(1)}$, $X_6^{(11)}$ – градации признака "Год рождения" с 12-летней цикличностью (к которым относятся 1934, 1946, 1958, 1970 годы, а также 1936, 1948, 1960, 1972 годы рождения); X_7 – "Резус-фактор".

Качество дискриминации в среднем составило 65,2% (табл. 1). Каноническая ДФ при таком разделении выглядит следующим образом:

$$y_1 = 1,223X_5^{(4)} - 0,973X_5^{(8)} + 0,949X_5^{(10)} - $$
$$-0,761X_6^{(1)} + 1,658X_6^{(11)} + 1,086X_1 + 0,715X_3^{(3)} +$$
$$+0,513X_7 - 0,097X_2^{(2)} - 3,046.$$

Здесь дискриминантные переменные расположены в порядке убывания значений (по абсолютной величине) их внутригрупповой корреляции со стандартизованной канонической функцией.

Таблица 1

Фактическая и предсказанная после проведения дискриминантного анализа принадлежность к классам (ДФ y_1)

		Предсказанная принадлежность к классам		Всего
		"ГР1"	"ГР2,3+ПЗ"	
Количество обследованных	"ГР1"	89	36	125
	"ГР2,3+ПЗ"	91	149	240
%	"ГР1"	71,2	28,8	100,0
	"ГР2,3+ПЗ"	37,9	62,1	100,0

Далее была построена ДФ, классифицирующая объединенные на предыдущем шаге классы объектов (группы риска "ГР2,3" и "ПЗ").

В качестве дискриминантных переменных в нее вошли следующие внутренние факторы риска: X_1 – "Пол"; $X_2^{(1)}$ – одна из градаций признака "Группа крови" со значением "Группа крови 1 (0)"; $X_4^{(2)}$ – одна из градаций признака "Цвет волос" со значением "Шатен"; $X_5^{(5)}$, $X_5^{(6)}$, $X_5^{(7)}$, $X_5^{(10)}$ – градации признака "Дата рождения" с 12-месячной цикличностью (периоды с 23.07 по 23.08, с 24.08 по 23.09, с 24.09 по 23.10, а также с 22.12 по 20.01); $X_6^{(1)}$, $X_6^{(5)}$, $X_6^{(11)}$ – градации фактора "Год рождения" с 12-летней цикличностью (к которым относятся 1934, 1946, 1958, 1970 годы, 1940, 1952, 1964, 1976 годы, а также 1936, 1948, 1960, 1972 годы рождения), X_7 – "Резус-фактор". Качество дискриминации в среднем составило 63,8% (табл. 2).

Таблица 2

Фактическая и предсказанная после проведения дискриминантного анализа принадлежность к классам (ДФ y_2)

		Предсказанная принадлежность к классам		Всего
		"ГР2,3"	"ПЗ"	
Количество обследованных	"ГР2,3"	88	38	126
	"ПЗ"	49	65	114
%	"ГР2,3"	69,8	30,2	100,0
	"ПЗ"	43,0	57,0	100,0

Каноническая ДФ при таком разделении объектов выглядит следующим образом:

$$y_2 = 1,724X_5^{(6)} + 1,071X_5^{(7)} + 1,239X_5^{(10)} +$$
$$+1,612X_6^{(1)} + 0,856X_6^{(5)} - 0,855X_1 - 1,533X_6^{(11)} -$$
$$-0,556X_4^{(2)} + 0,585X_2^{(1)} + 1,337,$$

где дискриминантные переменные также расположены в порядке убывания их внутригрупповой корреляции со стандартизованной канонической функцией.

С помощью стандартного для дискриминантного анализа критерия λ-Уилкса производился тест на то, значимо ли в классах различаются центры (табл. 3). Поскольку данная мера является "обратной", величины более близкие к нулю говорят о высоком различении классов объектов (т.е. центры классов хорошо разделены и сильно отличаются друг от друга по отношению к степени разброса внутри групп). В нашем случае значения критерия λ-Уилкса для функций y_1 и y_2 не позволяют сказать об их высокой способности различать классы, что объясняется сутью дискриминантных переменных – они не являются симптомами заболеваний, а лишь факторами риска, действие которых увеличивает риск развития заболевания, но при этом они не могут являться причиной заболевания. Используя значения критерия λ-Уилкса, можно получить тест на проверку значимости различий центров классов на основе распределения χ^2. Значения p (величина ошибки I рода) указывают на значимые различия.

Проверка качества классификации по полученным ДФ проводилась по всем объектам исходной выборки. Способ прогноза развития профессионально обусловленных аллергодерматозов заключается в том, что при идентификации состояния объекта A_i $(i = \overline{1,m})$ на основе построенных решающих правил D объект A_i в зависимости от совокупности его внутренних факторов риска $\{X_{1i}, X_{2i}, ..., X_{ji}, ..., X_{ni}\}$ попадает в один из трèх возможных классов Ω_p – "ГР1", "ГР2,3", "ПЗ", различающихся уровнем риска развития профессиональных заболеваний. Таким образом, принадлежность обследуемого к одному из классов интерпретируется как прогноз возможного состояния в будущем.

Таблица 3

Проверка значимости различия центров классов

Тест ДФ	Значения критерия λ -Уилкса	Значения критерия χ^2	Степени свободы	Критические значения критерия χ^2	Значимость, p
y_1	0,905	35,744	9	27,877	0,0001
y_2	0,930	16,926	9	16,919	0,05

Выводы

Построенные в результате дискриминантного анализа ДФ y_1 и y_2 содержат 14 из 37 выявленных внутренних факторов риска. Таким образом, с помощью дискриминантного анализа было получено информативное подмножество факторов риска развития профессионально обусловленных аллергодерматозов, необходимое для разделения всей совокупности на три состояния, соответствующие различным уровням риска развития профзаболевания. Результаты классификации, составившие соответственно 65,2 и 63,8% правильных ответов, дают возможность применять полученные ДФ для прогноза развития профессионально обусловленных заболеваний. Полученные результаты предполагается использовать в дальнейших исследованиях при построении экспертной системы оценки риска развития профессионально обусловленных аллергодерматозов у рабочих химико-фармацевтических предприятий.

Список литературы

1. Чикина Н.А. Изучение влияния внутренних факторов риска на развитие аллергодерматозов у рабочих химико-фармацевтических предприятий / Н.А. Чикина, И.В. Антонова // Тр. Меж. науч.-техн. конф. "MicroCAD – 2007". – Х.: НТУ "ХПИ", 2007. – Ч.8. – С. 195-200.

2. Чикина Н.А. Идентификация состояния здоровья на основе анализа типов реакции адаптации у рабочих предприятий химико-фармацевтической промышленности / Н.А. Чикина, И.В. Антонова // Вестник НТУ "ХПИ". Тематический выпуск: Информатика и моделирование. – Х.: НТУ "ХПИ". – 2008. – № 24. – С. 178-184.

3. Солошенко Э.Н. Математические модели лекарственной болезни: диагностика, прогнозирование, профилактика / Э.Н. Солошенко, Н.А. Чикина, А.М. Мануйлова // Кибернетика и вычисл. техника. – К.: Наукова думка, 1993. – Вып. 98. – С. 51-54.

4. Солошенко Э.Н. Математическое прогнозирование заболеваемости и осложнений лекарственной терапии / Э.Н. Солошенко, Н.А. Чикина, Л.П. Доильницина // Сб. статей "Проблемы восстановления и сохранения систем биосферы". – СПб.: Гидрометеостандарт, 1992. – Т.2. – С. 117-120.

5. Чикина Н.А. Математические модели адаптации к вредным условиям труда на основе метода корреляционной адаптометрии / / Н.А. Чикина, И.В. Антонова // Вестник НТУ "ХПИ". Тематический выпуск: Информатика и моделирование. – Х.: НТУ "ХПИ". – 2008. – № 49. – С. 184-189.

6. Ту Дж. Принципы распознавания образов / Дж. Ту, Р. Гонсалес. – М.: Мир, 1978. – 413 с.

7. Вапник В.Н. Теория распознавания образов / В.Н. Вапник, А.Я. Червоненкис. – М.: Наука, 1974. – 416 с.

8. Ким Дж.-О. Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка. – М.: Финансы и статистика, 1989. – 215 с.

9. Бююль А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П. Цèфель. – СПб.: ООО "ДиаСофтЮП", 2005. – 608 с.

Поступила в редколлегию 21.12.2009

Рецензент: д-р техн. наук, проф. В.Д. Дмитриенко, Национальный технический университет "ХПИ", Харьков.

ПРОГНОЗ РОЗВИТКУ ПРОФЕСІЙНО ОБУМОВЛЕНИХ ЗАХВОРЮВАНЬ ЗА ДОПОМОГОЮ ДИСКРИМІНАНТНОГО АНАЛІЗУ

А.І. Поворознюк, Н.О. Чікіна, Ю.Л. Геворкян, І.В. Антонова

У статті запропонований спосіб прогнозу розвитку професійно обумовлених алергодерматозів у робітників хіміко-фармацевтичних підприємств на основі результатів дискримінантного аналізу, що враховує сумарний вплив внутрішніх факторів ризику. Отримані результати дозволяють з досить високим ступенем точності розрізняти стани здоров'я обстежених, що може бути використане при побудові експертної системи прогнозу ризику розвитку професійно обумовлених алергодерматозів у робітників хіміко-фармацевтичних підприємств.

***Ключові слова**: прогноз розвитку захворювання, дискримінантний аналіз, фактори ризику, експертна система.*

PROGNOSIS OF DEVELOPMENT OF THE PROFESSIONALLY CONDITIONED DISEASES ON BASIS OF DISKRIMINANT ANALYSIS

A.I. Povoroznyuk, N.A. Chikina, Yu.L. Gevorkyan, I.V. Antonova

In the article the method of prognosis of development of professionally conditioned allergodermathoses for workers of chemical and pharmaceutical industry enterprises is offered. This method is based on results of diskriminant analysis and takes into account the summary influencing of internal factors of risk. The got results allow with enough high degree of exactnesses to distinguish the health state of inspected workers. This fact can be used for the construction of expert system of prognosis of risk of development of professionally conditioned allergodermathoses for chemical and pharmaceutical enterprises workers.

***Keywords**: prognosis of disease development, diskriminant analysis, risk factors, expert system.*