УДК 303.732.4

СРАВНИТЕЛЬНАЯ ОЦЕНКА ЭФФЕКТИВНОСТИ МЕТОДОВ КЛАССИФИКАЦИОННОГО АНАЛИЗА В СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Г. Г. Рапаков, В. А. Горбунов

Воронежский государственный университет

Поступила в редакцию 14.10.2014 г.

Аннотация. Рассмотрены возможности статистического анализа результативности медицинской профилактики на основе анкетного опроса. Разработаны статистические модели принятия решений с целью стимулирования здоровьесберегающих активностей населения и повышения эффективности профилактики неинфекционных заболеваний. С помощью методов логистической регрессии и дискриминантного анализа созданы адекватные, статистически значимые модели и вычислены классификационные функции. Проведен сравнительный анализ использованных методов..

Ключевые слова: дискриминантный анализ, логистическая регрессия, анкетный опрос, медицинская профилактика, принятие решений.

Annotation. In the article are considered the opportunities of statistical analysis of health preventive efficiency on the base of questionnaires. The paper presents developed statistical models for making the decision in order to stimulation the medical activities and improve the effectiveness of prevention of non-communicable diseases. Using the methods of the logistic regression and discriminant analysis the adequate and statistically significant models has been created and the classifying functions are calculated. The comparative analysis of methods was conducted.

Keywords: discriminant analysis, logistic regression, sociological research, prophylaxis, decision making.

ВВЕДЕНИЕ

Заболеваемость артериальной гипертензией (АГ) и смертность от болезней системы кровообращения (БСК) представляет собой значимый социально-экономический фактор в Российской Федерации [1]. Распространенность АГ среди взрослого населения в развитых странах составляет ≈ 35–40 %. Два из семи ведущих факторов риска (ФР) сердечно-сосудистых заболеваний (ССЗ) – АГ и гиперхолестеринемия – обуславливают 58–59 % преждевременной смертности [2]. Так, в Вологодской области в 2010–2011 гг. первичная заболеваемость АГ на 100 тыс. населения составляет величину от 622 до 764 [3–5]. Для снижения смертности населения от ССЗ необходимо создание единой профилактической среды, которая предусматривает осознанную модификацию ФР пациентами, стимулирование активности граждан, направленной на сохранение здоровья, и проведение эффективной немедикаментозной терапии [6].

В целях повышения качества человеческого капитала и улучшения демографический ситуации на территориальном уровне актуальной является задача формирования здорового образа жизни (ЗОЖ) и создания регионального здоровьесберегающего образовательного пространства (РЗОП). Настоящая работа является продолжением исследований, посвященных применению методов анализа данных в оценке эффективности областных целевых программ (ОЦП) в здравоохранении региона [7–9]. Ее новизна состоит

в создании на основе методов логистической регрессии и дискриминантного анализа адекватной, статистически значимой модели принятия решений в целях стимулирования здоровьесберегающих активностей населения и повышения эффективности региональной профилактики неинфекционных заболеваний.

ПРЕДПОСЫЛКИ СОЗДАНИЯ И ОСНОВЫ МОДЕЛИ

Объектом исследования является система организации медицинской профилактики АГ на территориальном уровне. Цель исследования состоит в анализе данных здоровьесберегающих активностей населения в части модификации факторов риска артериальной гипертензии при внедрении областной целевой программы.

В соответствии с планом организационных мероприятий Департамента здравоохранения Вологодской области для оценки качества реализации ОЦП «Профилактика и лечение артериальной гипертензии среди населения Вологодской области» и обеспечения аналитической поддержки при принятии управленческих решений в задаче формирования РЗОП было выполнено мониторинговое медико-социологическое исследование. Опрос осуществлялся в образовательных учреждениях области в 2011 г. и был проведен Государственным научно-исследовательским центром профилактической медицины МЗ РФ (отдел разработки политики и стратегии профилактики неинфекционных заболеваний и укрепления здоровья населения) совместно с Вологодским областным центром медицинской профилактики [10].

Результаты проведенной экспертизы представлены в виде наборов заполненных анкет школьников и их ближайшего окружения – родителей и учителей. Рандомизированная выборка, используемая в исследовании, создавалась методом случайного отбора и представлена 274 анкетами родителей школьников. Анкета содержит 134 показателя и представляет собой разновидность социологического опроса с номинальными признаками. Ошибка выборки не превышает 6 % ($\alpha = 0,95$). Статистический анализ выполнен с использованием вычислительных модулей программного комплекса IBM SPSS [11–14].

Поскольку мониторинговое медико-социологическое исследование представляет собой опрос с номинальными признаками, для выявления наличия и направления связи между многовариантными переменными используются таблицы перекрестных распределений (сопряженности, кросстабуляции).

Кросстабуляционный анализ является эффективным, прежде всего, для номинальных и порядковых переменных. В случае использования переменных с интервальной шкалой применяют корреляционный анализ [15, 16]. Визуальный анализ перекрестных распределений, построенных в ходе исследования, показывает, что лишь 2,2 % анкетируемых считают повышенное кровяное давление значимым ФР, отрицательно влияющим на здоровье. Это свидетельствует о низком текущем уровне медицинской активности населения в части факторов риска АГ.

Для коррекции здоровьесберегающих активностей населения в ходе реализации ОЦП и повышения эффективности профилактики неинфекционных заболеваний (НИЗ) необходимо уметь определять вероятность попадания респондента в одну из двух целевых групп: «Повышенное АД–значимый ФР = да» и «Повышенное АД–значимый ФР = нет» на основе медико-демографических характеристик опрашиваемых. Для этого могут быть использованы методы классификационного анализа. Если целевые группы известны заранее, в статистическом анализе данных применяют логистическую регрессию и дискриминантный анализ. Выбор возможной статистической процедуры – бинарной, мультиномиальной или порядковой регрессии определяется типами зависимой переменной и множества независимых переменных q_i, принимающих участие в расчете.

РЕАЛИЗАЦИЯ ЛОГИСТИЧЕСКОЙ РЕГРЕССИОННОЙ И ДИСКРИМИНАНТНОЙ МОДЕЛИ

При проведении исследования в качестве групп были рассмотрены варианты ответа для показателя переменной опроса «Повышенное АД–значимый ФР», которая кодирует целевые уровни. Вероятностная модель позволяет выявить значимые критерии сегментирования и определить, насколько сильно выражено их влияние на зависимую переменную. В случае двух групп используется бинарная логистическая регрессия. В результате предварительной обработки наблюдений в анализ было включено 78,1 % данных.

При задании параметров построения модели был использован пошаговый метод постепенного включения независимых переменных в регрессионный анализ, выполненный за 6 шагов. Объединенные тесты для коэффициентов модели позволяют оценить ее качество на основании статистической значимости в последней строке табл. 1. Модель обладает высокой значимостью (Знч. < 0,001) и является практически пригодной для выявления значимых критериев сегментирования.

Сводка для построенной модели демонстрирует высокую долю совокупной дисперсии на основе величины R^2 Нэйджелкерка (табл. 2).

Результаты, приведенные в таблице классификации, дают возможность сопоставить фактически наблюдаемые показатели принадлежности к каждой группе со значениями, предсказанными на основе логистической регрессионной модели (табл. 3). Общий процент для модели на шаге 6 свидетельствует о том, что она позволяет корректно классифицировать 99,5 % респондентов. Качество сегментации для группы «Повышенное АД–значимый ФР = да» равно 83,3 %, для «Повышенное АД–значимый ФР = нет» – 100 %.

Нестандартизированные коэффициенты регрессии, представляющие собой множители при переменных в уравнении функции z, позволяют с вероятностью p оценить принадлежность респондента к группе классификации зависимой переменной:

$$z = -57,246 - 3,719 \cdot q_{31б\_1} - 3,832 \cdot q_{31б\_2} - \\ -20,341 \cdot q_{31б\_2} + 48,770 \cdot q_{3902} + 32,592 \cdot q_{3904} - \\ -97,334 \cdot q_{3907} + 30,841 \cdot q_{4003} + 64,279 \cdot q_{4007},$$

$$p = 1/(1+e^z).$$

Модель при исключении члена из уравнения функции z:

Таблица 1

Объединенные тесты для коэффициентов модели

		χ^2	Степень свободы	Значимость
Шаг 6	Ступенька	5,158	1	0,023
	Блок	50,222	8	0,000
	Модель	50,222	8	0,000

Таблица 2

Сводка для модели

Ступенька	Правдоподобие	R^2 Кокса и Снелла	R^2 Нэйджелкерка
1	46,535	0,038	0,166
2	38,406	0,073	0,325
3	26,948	0,122	0,539
4	18,714	0,155	0,686
5	9,657	0,190	0,842
6	4,499	0,209	0,927

Таблица 3

Таблица классификации

	Наблюденные		Предсказано		
			Повышенное АД		Процент корректно классифицированных респондентов
			да	нет	
Шаг 6	Повышенное АД	да	5	1	83,3
		нет	0	208	100,0
	Общий процент корректно классифицированных респондентов				99,5

Таблица 4

Оценка вероятности принадлежности к группе «Повышенное АД–значимый ФР = да» для логистической регрессионной модели

ФР-повышенное АД	В кризисе ребенок опасается негативной реакции учителей, q31б	Коэффициенты q31б	Курение: убеждение и внушение, q3902	Курение: просмотр ТВ, q3904	Курение: наказание и угроза, q3907	Алкоголь: личный полож. пример, q4003	Алкоголь: наказание и угроза, q4007	*Z*	*p*
да	ученики не хотят обращаться к учителям	–20,34	да	нет	нет	нет	да	–32,34	1,00
да	ученики не хотят обращаться к учителям	–20,34	да	нет	нет	нет	да	–32,34	1,00
да	ученики опасаются обращаться к учителям	–3,83	да	да	нет	да	нет	–14,98	1,00
да	ученики не хотят обращаться к учителям	–20,34	да	нет	нет	да	нет	1,10	0,250
да	ученик полностью доверяет учителям	–3,72	да	нет	нет	нет	да	–15,72	1,000
да	ученики не хотят обращаться к учителям	–20,34	да	да	нет	да	нет	–31,49	1,000

$$z = -7,254 \cdot q_{31б} - 7,623 \cdot q_{3902} - 7,039 \cdot q_{3904} -$$
$$-13,301 \cdot q_{3907} - 4,828 \cdot q_{4003} - 13,685 \cdot q_{4007}.$$

График функции $p(z)$ представлен на рис. 1.

Анализ данных, представленных в табл. 4, демонстрирует существующую отчужденность между семьей и школой. Меры по ограждению от наркотиков не вошли в число значимых критериев сегментирования, что является тревожным симптомом и свидетельствует о низкой медико-социальной активности населения. При корректировке поведенческих факторов риска по отношению к курению во всех случаях были использованы убеждение и внушение, а также отказ от

наказания и угроз. Эффективность влияния телевидения на здоровьесберегающее поведение незначительна: 66 % респондентов не участвуют в организации просмотра телепередач. Модификация поведенческих факторов риска в части алкоголя демонстрирует инверсное поведение родителей: анкетируемые, использующие для ограничения личный положительный пример, не нуждаются в том, чтобы прибегать к наказанию и угрозам.

Так как в дискриминантном анализе участвуют одна категориальная зависимая переменная и несколько независимых переменных любого типа, при выборе зависимой переменной рекомендуется ограничивать количество категорий. Их рост ведет к снижению точности и надежности статистической модели. В результате предварительной обработки наблюдений в анализ было включено 89,1 % случаев. Диапазон изменения границ зависимой переменной выбран от 1 до 2. Для ввода независимых переменных в модель в исследовании был применен пошаговый метод с использованием вероятности F (включение – 0,05; исключение – 0,1).

В табл. 5 представлены независимые переменные, которые оказались включенными в результирующую дискриминантную модель на последнем шаге анализа.

Качество приближения дискриминантной модели можно оценить на основе статистической значимости. Ее величина (Знч. < 0,001) указывает на существенные различия между средними значениями дискриминантных функций в двух группах зависимой переменной. Нормированные коэффициенты канонической дискриминантной функции позволяют сравнивать относительный вклад каждой независимой переменной в различие двух исследуемых групп (табл. 6).

Коэффициенты канонической дискриминантной функции позволяют построить уравнение ее функции z и с вероятностью p оценить принадлежность респондента к группе классификации. График функции $p(z)$ представлен на рис. 2.

$$z = -4,406 + 0,628 \cdot q_{3902} + 1,938 \cdot q_{3904} -$$

$$-5,342 \cdot q_{3907} + 5,262 \cdot q_{4007},$$

$$p = 1/(1+e^{z}).$$

Точность вероятностной модели обеспечивает правильную классификацию 94,3 % исходных сгруппированных наблюдений (табл. 7). Результаты классификации, приведенные в таблице, дают возможность сопоставить фактически наблюдаемые показатели принадлежности к каждой группе со значениями, предсказанными на основе дискри-

Таблица 5

Независимые переменные, включенные в дискриминантную модель

Независимые переменные	Толерантность	Значение F исключения	λ Уилкса
Курение – наказание, угроза наказания	0,503	0,000	0,945
Алкоголь – наказание, угроза наказания	0,456	0,000	0,890
Организация просмотра телепередач	0,869	0,001	0,817
Убеждение, внушение	0,996	0,023	0,796

Таблица 6

Нормированные коэффициенты

Независимые переменные	Стандартизированные коэффициенты
Убеждение, внушение	0,311
Организация просмотра телепередач	0,489
Алкоголь – наказание, угроза наказания	–1,112
Курение – наказание, угроза наказания	1,256

Таблица 7

Результаты классификации

Повышенное АД	Предсказанная принадлежность к группе		Итого
	да	нет	
Да (частота)	5,0	1,0	6,0
Нет (частота)	13,0	225,0	238,0
Да (%)	83,3	16,7	100,0
Нет (%)	5,5	94,5	100,0

Таблица 8

Оценка вероятности принадлежности к группе «Повышенное АД–значимый ФР = да» для дискриминантной модели

ФР-повышенное АД	Курение: убеждение и внушение, q3902	Курение: просмотр ТВ, q3904	Курение: наказание и угроза, q3907	Алкоголь: наказание и угроза, q4007	*Z*	*p*
да	да	нет	нет	да	–5,321	0,995
да	да	нет	нет	да	–5,321	0,995
да	да	да	нет	нет	–1,998	0,881
да	да	нет	нет	нет	–0,059	0,515
да	да	нет	нет	да	–5,321	0,995
да	да	да	нет	нет	–1,998	0,881

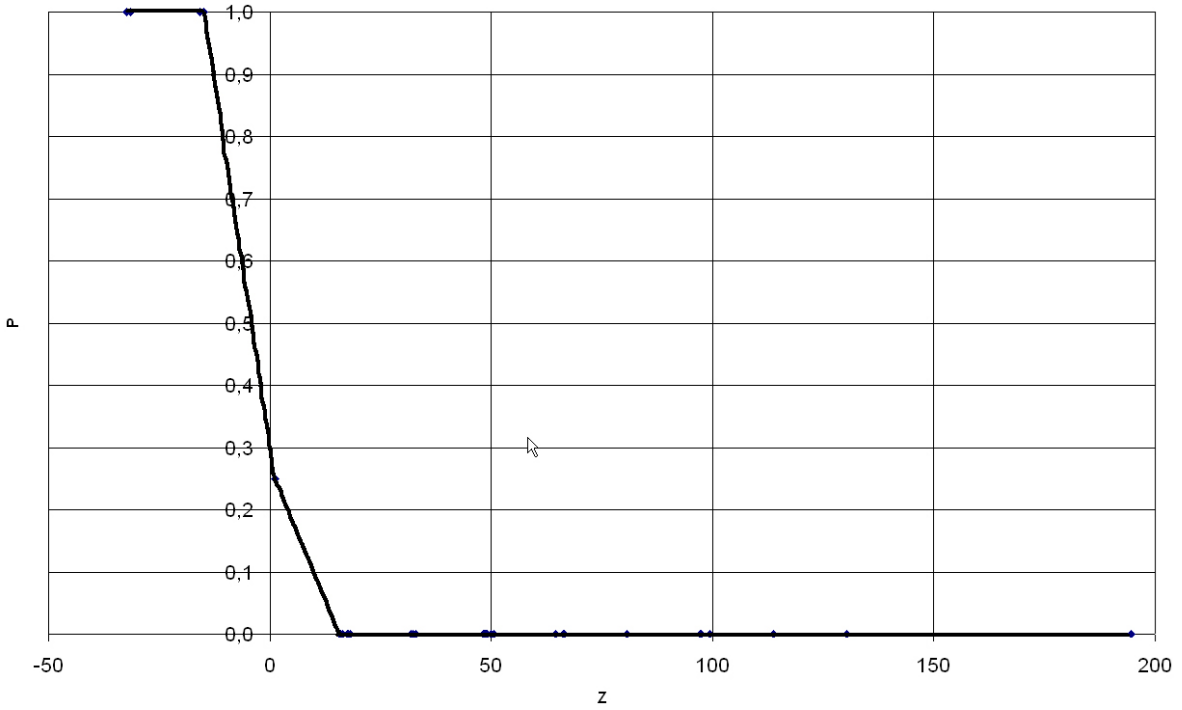

Рис. 1. Классификация на основе логистической регрессионной модели. Оценка вероятности p принадлежности респондента к группе классификации переменной z для логистической регрессионной модели

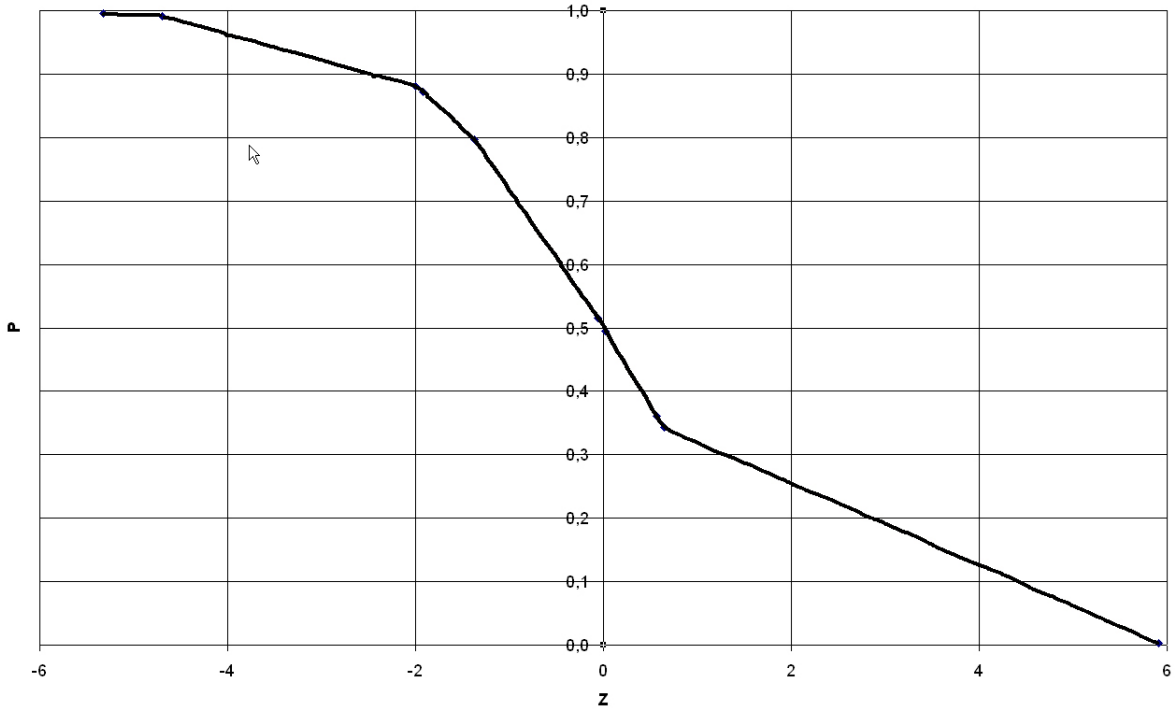

Рис. 2. Классификация на основе модели дискриминантного анализа. Оценка вероятности p принадлежности респондента к группе классификации переменной z для модели дискриминантного анализа

минантного анализа. Корректная сегментации принадлежности к группе «Повышенное АД–значимый ФР = да» обеспечена для 83,3 % респондентов, в отношении «Повышенное АД–значимый ФР = нет» – в 94,5 % случаев.

Результаты сегментации респондентов на основе их медико-демографических характеристик для дискриминантного анализа уступают аналогичному показателю логистической регрессионной модели. Снижение общего качества на 5 % вызвано уменьшением точности распознавания для объектов второй группы, которая является наиболее многочисленной. Точность классификации целевой группы «Повышенное АД–значимый ФР = да» не пострадала и обеспечивается за счет меньшего на треть количества независимых переменных в уравнении дискриминантной модели. Анализ графика, приведенного на рис. 2, и данных табл. 8 показывает, что выигрыш, связанный с уменьшением количества критериев сегментирования, приводит к снижению разрешающей способности классификатора. При этом диапазон значений аргумента сжимается более чем в 20 раз и возрастают трудности интерпретации результатов вероятностного прогноза при отнесении каждого конкретного респондента в выборке к определенной целевой группе.

ЗАКЛЮЧЕНИЕ

Проведенные исследования показали, что вероятностные модели классификационного анализа выявили значимые (Знч. < 0,001) критерии сегментирования респондентов по целевым группам при оценке медицинской активности населения в части факторов риска АГ. Высокая доля совокупной дисперсии (R^2 Нэйджелкерка = 0,927), описываемая моделью, делает ее практически пригодной. На основе медико-демографических характеристик опрашиваемых корректно классифицировано не менее 83,3 % респондентов. Нормированные коэффициенты позволяют сравнивать относительный вклад каждой независимой переменной. При переходе от логистической регрессионной модели к дискриминантному анализу с приемлемым для практики снижением точности классификации количество критериев сегментирования уменьшено на треть. Результаты мультиномиальной логистической регрессии

позволяют оценить представление целевых групп респондентов в категориях независимых переменных и количественно оценить влияние отношений, складывающихся в образовательных учреждениях, на реализацию программы ЗОЖ. Анализ данных мониторингового медико-социологического исследования позволил оценить здоровьесберегающие активности населения в части модификации факторов риска артериальной гипертензии. Результаты моделирования используются для поддержки принятия управленческих решений в сфере профилактики заболеваемости в ходе реализации ОЦП «Профилактика и лечение артериальной гипертензии среди населения Вологодской области» и оптимизации региональной профилактики НИЗ в условиях сокращения финансирования здравоохранения.

СПИСОК ЛИТЕРАТУРЫ

1. Социально-экономический ущерб от острого коронарного синдрома в Российской Федерации / А. В. Концевая, А. М. Калинина, И. Е. Колтунов, Р. Г. Оганов // Рациональная фармакотерапия в кардиологии. – 2011. – Т. 2, № 7. – С. 158–166.

2. Кардиоваскулярная профилактика – важнейшее условие снижения смертности населения: методические указания / Г. Т. Банщиков, Е. А. Барачевская, М. Н. Зайцева, Р. А. Касимов, Г. Г. Рапаков. – Вологда : ВоГУ, 2014. – 73 с.

3. Эпидемиологическая ситуация в отношении основных факторов риска и суммарного сердечнососудистого риска среди населения г. Вологды 35–64-летнего возраста / А. И. Попугаев, А. М. Калинина, Р. А. Касимов и [др.] // Кардиоваскулярная терапия и профилактика. – 2008. – Т. 7, № 8. – Ч. 2. – С. 12–19.

4. Статистический ежегодник Вологодской области: стат. сб./ Росстат, Территориальный орган Федеральной службы государственной статистики по Вологодской области (Вологдастат). – Вологда, 2011. – 374 с.

5. Демографический ежегодник Вологодской области: стат. сб./ Росстат, Территориальный орган Федеральной службы государственной статистики по Вологодской области. – Вологда, 2011. – 88 с.

6. Реализация программы «Профилактика и лечение артериальной гипертонии в Российской Федерации» на региональном уровне (опыт г. Вологды) / Г. Т. Банщиков, А. А. Колинько, Р. А. Касимов и [др.] // Кардиоваскулярная терапия и профилактика. – 2008. – Т. 3, № 3. – Ч.2. – С. 43–46.

7. Профилактика и лечение артериальной гипертонии среди населения Вологодской области на 2009–2011 годы [Электронный ресурс]: ведомственная целевая программа: постановление Правительства Вологодской области от 28 июня 2010 г. № 739 // КонсультантПлюс: справ.-правовая система / Компания«КонсультантПлюс».

8. *Рапаков Г. Г.* Организация системы раннего выявления больных артериальной гипертензией и доступность антигипертензивных средств в Вологодской области: опыт использования кластерного анализа / Г. Г. Рапаков, Г. Т. Банщиков //Архивъ внутренней медицины. – 2013. – №4. – С. 16 – 23.

9. *Рапаков Г. Г.* Интеллектуальный анализ данных в здравоохранении региона (на материалах Вологодской области): монография / Г. Г. Рапаков, Г.Т. Банщиков. – Вологда: ВоГУ, 2014. – 79 с.

10. Колинько А. А. Программа «Здоровые города, районы и поселки» в субъекте Российской Федерации: структура, этапы реализации / А. А. Колинько, Р.А. Касимов // Профилактическая медицина. – 2012. – Т. 15, № 5. – С. 16–20.

11. *Бююль А.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П. Цефель. – СПб. : ДиаСофтЮП, 2002. – 608 с.

12. *Field A.* Discovering Statistics Using SPSS/ A. Field. – Sage Publica-tions, 2005. – 779 p.

13. *Hair J. F.* Marketing research/ J. F. Hair, R. P. Bush, D. J. Ortinau. – McGraw-Hill/Irwin, 2003. – 720 p.

14. *Einspruch E. L.* An Introductory Guide to SPSS for Windows/ E. L. Einspruch – Sage Publications, 2005. – 145 p.

15. Айвазян С. А. Прикладная статистика: Исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М. : Финансы и статистика, 1985. – 487 с.

16. Гланц С. Медико-биологическая статистика / С. Гланц. – М. : Практика, 1999. – 459 с.

Рапаков Георгий Германович – кандидат технических наук, доцент, доцент кафедры информационных систем и технологий, Вологодский государственный университет.
Тел.: +7(921) 231-31-12; +7(8172) 72-95-71
E-mail: grapakov@yandex.ru

Горбунов Вячеслав Алексеевич – зав. кафедрой информационных систем и технологий, доктор физико-математических наук, профессор, Вологодский государственный университет.
Тел.: +7(921) 234-50-65; +7(8172) 72-95-71
E-mail: gorbunov1945@inbox.ru

Rapakov G. G. – PhD in Technical Science, Associate Professor, Information Systems and Technologies Department, Vologda State University.

Gorbunov V. A. – Doctor of Physical–Mathematical Science, Professor, Information Systems and Technologies Department, Head of Department. Vologda State University.