УПРАВЛЕНИЕ В СОЦИАЛЬНЫХ И ЭКОНОМИЧЕСКИХ СИСТЕМАХ

SOCIAL AND ECONOMIC SYSTEMS MANAGEMENT

Научная статья
УДК 004.02:004.06
https://doi.org/10.24143/2072-9502-2022-2-87-96

Применение методов парного сравнения количественных и бинарных выборок в биомедицинских исследованиях с целью принятия управленческих решений

***Лев Игоревич Евельсон*[1 *]*, Эмилия Владимировна Гегерь*[2]*, Ирина Романовна Козлова*[3]**

[1]*Научно-инновационный центр информационных и дистанционных технологий, Брянск, Россия, levelmoscow@mail.ru* [*]

[2]*Брянский клинико-диагностический центр, Брянск, Россия*

[3]*Брянский государственный технический университет, Брянск, Россия*

Аннотация. Решение исследовательских задач в рамках создания единого цифрового контура в здравоохранении требует проведения исследований, реализуемых на основе деперсонализированных медицинских данных, накопленных в информационных системах лечебных учреждений. Описаны методы математической статистики, направленные на сравнение средних значений выборок двух видов: количественных и бинарных – с целью определения связи между показателями анализа крови и условиями труда. Выполнено сопоставление методов и результатов сравнения количественных и бинарных выборок. Показано, в каких случаях целесообразно использовать те или иные методы, когда есть возможность выбора между ними. Исследование проводилось с использованием медицинских данных, накопленных в медицинской информационной системе транзакционного типа. В процессе подготовки к исследованию данные подвергались деперсонализации, очистке от неизбежных шумов и дефектов. Бинаризация значений показателей производилась путем сопоставления с известными границами интервала медицинской нормы. Разработана методика приведения выборок к однородности одновременно по признакам пола и возраста пациентов. Выявлены показатели лабораторных исследований, которые имеют статистически значимую взаимосвязь с условиями труда в рассматриваемых 4 группах. Эти группы соответствовали следующим комплексам условий труда: воздействие промышленных электромагнитных излучений, воздействие на рабочем месте шума и вибраций, условия работы в региональных офисных службах. Предлагаемые методы и полученные результаты повысят точность выполняемых оценок риска профессиональной заболеваемости и станут основой для исследования механизма влияния производственных факторов, что будет способствовать улучшению условий труда и снижению негативного воздействия вредных производственных факторов на здоровье человека. Они также будут способствовать совершенствованию анализа данных, накопленных в медицинских информационных системах, и принятию управленческих решений в здравоохранении.

Ключевые слова: математическая статистика, анализ данных, бинарные выборки, медицинские информационные системы, анализ крови, пределы нормы

Для цитирования: *Евельсон Л. И., Гегерь Э. В., Козлова И. Р.* Применение методов парного сравнения количественных и бинарных выборок в биомедицинских исследованиях с целью принятия управленческих решений // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2022. № 2. С. 87–96. https://doi.org/10.24143/2072-9502-2022-2-87-96.

© Евельсон Л. И., Гегерь Э. В., Козлова И. Р., 2022

Original article

Applying methods of twin comparing quantitative and binary samples in biomedical information systems for decision making

Lev I. Evelson[1] *, Emiliya V. Geger[2], Irina R. Kozlova[3]

*[1]Innovation Scientific Centre of Information and Remote Technologies, LLC, Bryansk, Russia, levelmoscow@mail.ru **

[2]Bryansk Clinical and Diagnostic Center, Bryansk, Russia

[3]Bryansk State Technical University, Bryansk, Russia

Abstract. Solving research problems within the framework of creating a single digital circuit in healthcare requires a research conducted on the basis of depersonalized medical data stored in the information systems of medical institutions. There are described the methods of mathematical statistics aimed at comparing the average values of two types of samples: quantitative and binary in order to determine the relationship between blood test indicators and working conditions. Comparison of methods and results of comparison of quantitative and binary samples is made. The expediency of processing small structured samples taken out from the medical information system is substantiated. The study was conducted by using medical data stored in a transactional medical information system. During the preparation process, the data were depersonalized, cleaned from the inevitable noise and defects. Binarization of the values of the indicators was performed by comparing them with the known boundaries of the interval of the medical norm. A method was developed to bring the samples to uniformity simultaneously on the gender and age signs of the patients. There have been revealed the parameters of laboratory tests, which have a statistically significant relationship with working conditions identified for 4 groups under study. These groups were corresponding to the following work conditions complexes: influence of electromagnetic emanation, noise and vibrations, working conditions in regional office services. The proposed methods and received results will increase the accuracy of the performed risk assessments of occupational morbidity and become the base for studying the mechanism of the work conditions influencing the health. They will contribute to improvement of the analysis of the data collected in the medical information systems and management decision-making in healthcare.

Keywords: mathematical statistics, data analysis, binary sampling, medical information systems, blood test, norm limits

For citation: Evelson L. I., Geger E. V., Kozlova I. R. Applying methods of twin comparing quantitative and binary samples in biomedical information systems for decision making. *Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and Informatics.* 2022;2:87-96. (In Russ.) https://doi.org/10.24143/2073-5529-2022-2-87-96.

Введение

К настоящему времени в информационных системах (ИС) медицинских организаций накоплено уже много различных данных, связанных с медицинской помощью населению. В данный момент создается единая цифровая сеть, в цифровой контур вовлечено все больше лечебных учреждений, и цифровые технологии позволяют оказывать все более эффективные и персонализированные услуги.

В создании единого цифрового контура в здравоохранении важную роль играет аналитика, основанная на первичной информации [1]. Медицинские информационные системы (МИС) оперируют большими объемами детализированной информации о здоровье пациента с помощью технологии оперативной обработки транзакций – OLTP (Online Transaction Processing – обработка транзакций в реальном времени) [2, 3]. Они предназначены для «цифровизации» непосредственно текущих информационных процессов учреждения. Данные, хранящиеся в МИС, как правило, для исследовательских задач не используются [4–6]. Однако они могут быть консолидированы, обезличены (деперсонализированы), очищены от неизбежных шумов и дефектов и выгружены в аналитические системы либо в электронные таблицы MS Excel для дальнейших исследований. Конкретные задачи, которые при этом ставятся, могут быть весьма разнообразны. При выборе методов их решения следует рационально подходить к учету особенностей методов, характеристик имеющихся доступных наборов данных для конкретной задачи и вычислительных ресурсов. Характерный объем анализируемых выборок медицинских данных для многих задач составляет порядка сотен или тысяч записей, поэтому целесообразно ориентироваться на методы, предназначенные для работы с относительно небольшими структурированными выборками. Подходы Big Data, часто используемые в мировой практике [7, 8], предназначены для работы с большими наборами данных, формирующихся из разнообразных по структуре и формату источников медицинской информации, представляющих собой неструктурированный набор файлов, таблиц, рисунков,

графиков, их описаний, зачастую противоречивых выводов и суждений. Для малых и средних выборок технологии Big Data неэффективны. При применении классических методов математической статистики возникает ряд типичных проблем, которые «в чистой математике» считаются как бы заранее кем-то решенными, однако на практике их приходится решать, и от этого существенно зависит достоверность результатов и выводов.

В статье на важном для охраны труда практическом примере продемонстрированы некоторые типичные проблемы, показаны возможные пути и разработанные методики их решения с помощью нетрадиционного применения хорошо известных математических методов. Рассматривается проблема оценки статистической значимости зависимости между лабораторными показателями анализа крови и условиями труда пациента. Такая задача является частью общей актуальной проблемы оценки риска профессиональной заболеваемости. Ее решение играет существенную практическую роль в планировании мероприятий по охране труда, а также способствует развитию цифровых технологий в медицине для принятия управленческих решений на основе точных, своевременных и полных данных и адекватных аналитических инструментов.

Цель работы – выявление важных особенностей и закономерностей практического применения различных известных методов математической статистики, направленных на исследование зависимости показателей крови от производственных факторов.

Материалы и методы исследования

Исследования проводились в клинико-диагностической лаборатории Брянского клинико-диагностического центра, результаты были отражены в медицинской информационной системе транзакционного типа. В качестве первичного источника данных использовались результаты общего анализа крови (ОАК) и биохимические показатели крови у лиц, работа которых связана с вредными условиями труда – с воздействием электромагнитных излучений промышленной частоты (I группа, или ЭМИ, 108 чел.), с воздействием шума и вибраций (II группа, или ШиВ, 149 чел.). Также использовались результаты медицинских осмотров работников офисных служб (III группа, или ТАМ, 251 чел. и группа IV, или АДМ, 147 чел.). Биомедицинские исследования выполнены в строгом соответствии с законодательством Российской Федерации, ведомственными приказами и инструкциями [9]. В работе использовались методы математической статистики, направленные на сравнение средних значений выборок двух видов: количественных и бинарных. Количественные включали непосредственно числовые действительные значения показателей анализа крови. Бинарные (да/нет, 1/0 и т. д.) получаются с помощью операции сопоставления этих числовых значений с известным интервалом нормы (попадает/не попадает). Такой метод замены изначально количественных данных на бинарные для подобных задач описан в работах [10–12]. Метод сравнения бинарных выборок для общего случая основанный на распределении Бернулли и теореме Муавра – Лапласа, подробно описан, например, в [13]. Основная конечная формула для критерия значимости разницы:

$$Q = \left(p_1^* - p_2^*\right) \Big/ \sqrt{\frac{p_1^*(1 - p_1^*)}{n_1} + \frac{p_2^*(1 - p_2^*)}{n_2}}, \quad (1)$$

где p_1^* и p_2^* – частоты появления бинарного значения «1» в первой и второй сравниваемых выборках, соответственно; n_1 и n_2 – объемы выборок.

Для количественных выборок сравнение средних значений, а точнее оценка значимости разницы между ними при разных неизвестных дисперсиях, называется задачей Беренса – Фишера. Она не имеет точного теоретического решения, а для приближенного решения в данной работе использовался критерий Крамера – Уэлча [14], в котором фигурируют выборочные оценки дисперсий. Формула для расчетного значения критерия:

$$K = \frac{1}{s}(\bar{x} - \bar{y}), \quad (2)$$

где

$$s = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; \; s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n}(x_i - \bar{x})^2;$$

$$s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{m}(y_i - \bar{y})^2,$$

где s – несмещенная оценка дисперсии разности выборочных средних рассматриваемых выборок; s_1^2 – несмещенная оценка дисперсии выборки 1; s_2^2 – несмещенная оценка дисперсии выборки 2; $\bar{x}$ – выборочное среднее арифметическое значение элементов выборки 1; $\bar{y}$ – выборочное среднее арифметическое значение элементов выборки 2.

До проведения анализа данных после консолидации была произведена очистка от дефектов и шумов, деперсонализация данных. В выборки включались значения различных показателей крови, а также бинарные значения пола (мужской/женский) и количественные значения возраста. Пол и возраст являются важными признаками, которые могут существенно влиять на заболеваемость и показатели крови. В ходе анализа поочередно лица каждой группы, подвергшиеся воздействию вредных производственных факторов (ЭМИ и ШиВ) либо являющиеся сотрудниками одной организации (ТАМ, АДМ), сравнивалась с объединенной группой, в которую входили лица остальных групп. Объединенные группы были следующие: «Все остальные, кроме группы ЭМИ» ($ВО_Э$),

«Все остальные, кроме группы ШиВ» ($ВО_Ш$), «Все остальные, кроме группы ТАМ» ($ВО_Т$) и «Все остальные, кроме группы АДМ» ($ВО_А$). Прежде всего, пары групп ЭМИ и $ВО_Э$, ШиВ и $ВО_Ш$, ТАМ и $ВО_Т$, АДМ и $ВО_А$ были проверены на однородность по признакам пола путем сравнения бинарных выборок по формуле (1) и возраста с помощью критерия Крамера–Уэлча по формуле (2). Все пары друг относительно друга оказались неоднородными по обоим этим признакам, поэтому далее была произведена корректировка выборок с целью добиться однородности. Разработанная методика корректировки основывалась на принципах рандомизации и эвристических способах. В соответствии с первым принципом корректировка осуществлялась таким образом, чтобы порядок записей, проверяемых на предмет удовлетворения критерию удаления, был случайным. В соответствии со вторым принципом корректировка сразу же прекращалась, как только оба (по полу и по возрасту) расчетные значения критериев однородности становились меньше или равны критическим значениям, причем алгоритм корректировки был сформирован так, чтобы число удаляемых записей было минимальным.

Выполнявшийся анализ данных был условно разделен на следующие этапы:

1. Консолидация обезличенных данных, выбираемых из транзакционной МИС в соответствии с поставленной задачей анализа.

2. Очистка данных от дефектов и шумов [15].

3. Для проведения каждого расчета, касающегося очередной группы, – операция слияния всех остальных групп в группу ВО (таким образом получались группы $ВО_Э$, $ВО_Ш$, $ВО_Т$ и $ВО_А$).

4. Корректировка групп ВО с целью получения выборок, однородных с изучаемой в данном расчете, одновременно по признакам пола и возраста. Для проверки однородности применялись критерии K и Q (формулы (1) и (2)).

5. Сравнение средних значений показателей крови с определением статистической значимости разницы по критерию Крамера–Уэлча, которое делалось поочередно для каждой из 4 исходных групп, сравниваемых с соответствующей объединенной и скорректированной группой ВО.

6. Бинаризация показателей крови путем сопоставления их значений с интервалом нормы.

7. Сравнение средних частот с определением статистической значимости разницы по критерию Q сравнения бинарных выборок для каждой из 4 исходных групп, сравниваемых с соответствующей объединенной и скорректированной группой ВО.

Результаты исследования и их анализ

По результатам расчетов были сформированы таблицы. В табл. 1 приведены данные, связанные с проверкой однородности групп по важным признакам пола и возраста.

Таблица 1

Table 1

Результаты корректировки выборок

Results of sample adjustment

Группа		Объем исходной выборки	Средний возраст исходной выборки	Половой состав исходной выборки (м/ж)	$K_{исх}$	$Q_{исх}$	Объем скорректированной выборки	Средний возраст скорректированной выборки	Половой состав скорректированной выборки (м/ж)	$K_{кор}$	$Q_{кор}$
ЭМИ/$ВО_Э$	ЭМИ	108	42,67	106/2	1,07	10,31	106	42,47	106/0	1,94	0
	$ВО_Э$	547	43,94	409/138			407	44,87	407/0		
ШиВ/$ВО_Ш$	ШиВ	149	51,76	139/10	11,12	6,72	149	51,76	139/10	1,89	1,83
	$ВО_Ш$	506	41,37	376/130			136	49,54	118/18		
ТАМ/$ВО_Т$	ТАМ	251	41,08	223/28	–5,88	5,55	251	41,08	223/28	–1,86	1,93
	$ВО_Т$	404	45,38	292/112			277	42,65	230/47		
АДМ/$ВО_А$	АДМ	147	40,91	47/100	–3,89	–14,93	147	40,91	47/100	–1,87	–1,9
	$ВО_А$	508	44,55	468/40			73	43,45	33/40		

Показан половой состав и средний возраст в исходных и скорректированных выборках, а также представлены полученные расчетные значения критериев сравнения средних значений возраста по критерию K и бинарных выборок по признаку пола по критерию Q – до корректировки и после.

Как видно из столбцов $K_{исх}$ и $Q_{исх}$, исходные группы ЭМИ, ШиВ, ТАМ, АДМ оказались неоднородны с соответствующими группами ВО по одному из двух признаков, причем группы ШиВ, ТАМ и АДМ неоднородны сразу по обоим признакам, а группа ЭМИ – только по полу. В связи с этим производилась корректировка, результаты которой отражены в правой части табл. 1. Отметим: фактически корректировались только объединенные группы ВО. Группы ШиВ, ТАМ и АДМ не изменялись, а из исходной группы ЭМИ, в которой из 108 человек было только 2 женщины, были сразу

удалены 2 соответствующие им записи. Из соответствующей группы ВОэ также сразу были удалены все записи, относящиеся к женщинам, вследствие чего при применении далее общего алгоритма корректировки однородность проверялась только по критерию *K*, а критерий *Q* при сравнении ЭМИ и $ВО_Э$ равнялся 0. Изменение числа записей при корректировке группы ВО оказалось очень большим (соответственно почти в 7 и 4 раза) для групп АДМ и ШиВ, в отличие от групп ЭМИ и ТАМ. Это обусловлено тем (как видно из данных табл. 1), что в исходных группах половой состав групп АДМ и возрастной состав группы ШиВ отличаются от остальных групп. В то же время целесообразно провести дополнительное исследование и оптимизацию предложенного алгоритма корректировки.

В табл. 2–5 приводятся результаты сравнения средних количественных значений лабораторных показателей по критерию Крамера – Уэлча для исходных (до корректировки) и скорректированных выборок; приводятся результаты сравнения частот выхода за пределы нормы количественных значений лабораторных показателей для исходных (до корректировки) и скорректированных выборок, т. е. результаты сравнения бинарных выборок по критерию *Q*, при этом расчетные значения по критериям Крамера – Уэлча и *Q* приведены с учетом знака: наличие перед числом знака «–» означает, что среднее значение для рассматриваемой группы оказалось меньше, чем для совокупности остальных, а отсутствие знака (что подразумевает знак «+») говорит о том, что значение было больше.

Таблица 2

Table 2

Результаты расчетов по критериям Крамера–Уэлча и *Q* для группы ЭМИ

Results of calculations by using the Cramer-Welch and *Q* criteria for EMR group

Группа	Показатели ОАК									Биохимия	
	Гемоглобин	Лейкоциты	Лимфоциты	Моноциты	Эритроциты	Тромбоциты	Эозинофилы	Гематокрит	СОЭ	Общий холестерин	Глюкоза
Результаты расчета по критерию Крамера–Уэлча											
ЭМИ–$ВО_{исх}$	–1,55	–0,64	3,1	6,47	–0,89	2,03	8,43	–1,13	–0,04	3,98	0,96
ЭМИ–$ВО_{кор}$	–2,01	–0,57	3,01	6,02	–3,33	1,87	8,52	–2,57	0,02	3,14	0,74
Результаты расчета по критерию *Q*											
ЭМИ–$ВО_{исх}$	–1,55	–0,639	3,1	6,467	–0,886	2,03	8,43	–1,13	–0,04	3,97	0,96
ЭМИ–$ВО_{кор}$	–2,01	–0,57	3,01	6,02	–3,32	1,87	8,52	–2,57	0,02	3,14	0,74
Сводные результаты для группы ЭМИ											
ЭМИ–$ВО_{исх}$	+ 0	0 0	+ +	– +	0 0	0 +	– +	– 0	0 0	+ +	+ 0
ЭМИ-$ВО_{кор}$	0 –	0 0	0 +	– +	– –	0 0	– +	– –	0 0	0 +	+ 0

Таблица 3

Table 3

Результаты расчетов по критериям Крамера–Уэлча и *Q* для группы ШиВ

Results of calculations by using the Cramer-Welch and *Q* criteria for NV group

Группа	Показатели ОАК									Биохимия	
	Гемоглобин	Лейкоциты	Лимфоциты	Моноциты	Эритроциты	Тромбоциты	Эозинофилы	Гематокрит	СОЭ	Общий холестерин	Глюкоза
Результаты расчета по критерию Крамера–Уэлча											
ШиВ–$ВО_{исх}$	3,05	3,89	–3,02	0,65	0,64	–1,37	3,41	4,200	–3,43	0,83	0,7
ШиВ–$ВО_{кор}$	–0,09	2,33	–2,28	1,72	–1,15	0,87	3,33	1,8	–2,62	–0,93	–1,63
Результаты расчета по критерию *Q*											
ШиВ–$ВО_{исх}$	0,239	1,25	–0,73	–1,57	–0,24	–0,71	–2,13	2,68	–1,87	–0,17	1,83
ШиВ–$ВО_{кор}$	–0,17	0,391	–0,29	–2,11	–1,37	–0,65	–2,96	1,71	–1,81	–1,03	0,77
Сводные результаты для группы ШиВ											
ШиВ-$ВО_{исх}$	+ 0	+ 0	– 0	0 0	0 0	0 0	+ –	+ +	– 0	0 0	0 0
ШиВ–$ВО_{кор}$	0 0	+ 0	– 0	0 –	0 0	0 0	+ –	0 0	– 0	0 0	0 0

Таблица 4

Table 4

Результаты расчетов по критериям Крамера–Уэлча и *Q* для группы ТАМ

Results of calculations using the Cramer-Welch and *Q* criteria for TAM group

Группа	Показатели ОАК									Биохимия	
	Гемоглобин	Лейкоциты	Лимфоциты	Моноциты	Эритроциты	Тромбоциты	Эозинофилы	Гематокрит	СОЭ	Общий холестерин	Глюкоза
Результаты расчета по критерию Крамера–Уэлча											
ТАМ–$ВО_{исх}$	6,34	–0,98	1,4	3,81	6,74	0,41	2,5	6,03	0,41	1,82	–2,46
ТАМ–$ВО_{кор}$	3,744	–1,64	1,28	3,73	4,00	0,68	2,71	4,327	1,81	1,82	–1,9
Результаты расчета по критерию *Q*											
ТАМ–$ВО_{исх}$	2,886	–1,41	-0,53	–2,96	6,74	0,066	–5,25	1,54	0,9	1,03	–1,77
ТАМ–$ВО_{кор}$	2,479	–1,44	–1,67	–3,37	4,51	0,07	–5,11	0,24	1,2	1,25	–1,65
Сводные результаты для группы ТАМ											
ТАМ–$ВО_{исх}$	+ +	0 0	0 0	+ –	+ +	0 0	+ –	+ 0	0 0	0 0	– 0
ТАМ–$ВО_{кор}$	+ +	0 0	0 0	+ –	+ +	0 0	+ –	+ 0	0 0	0 0	0 0

Таблица 5

Table 5

Результаты расчетов по критериям Крамера–Уэлча и *Q* для группы АДМ

Results of calculations using the Cramer-Welch and *Q* criteria for ADM group

Группа	Показатели ОАК									Биохимия	
	Гемоглобин	Лейкоциты	Лимфоциты	Моноциты	Эритроциты	Тромбоциты	Эозинофилы	Гематокрит	СОЭ	Общий холестерин	Глюкоза
Результаты расчета по критерию Крамера–Уэлча											
АДМ–$ВО_{исх}$	–12,5	–3,31	–1,19	2,08	–8,33	1,23	–0,91	–5,53	2,9	–6,65	–0,57
АДМ–$ВО_{кор}$	–3,33	–1,5	1,8	1,69	–2,41	0,11	–0,77	–0,76	–0,94	–2,63	0,99
Результаты расчета по критерию *Q*											
АДМ–$ВО_{исх}$	–2,35	0,66	–1,79	–1,98	–6,86	–1,8	0,95	–3,98	0,66	–4,11	–1,29
АДМ–$ВО_{кор}$	–0,83	0,35	–0,82	0,65	–1,71	–2,16	2,7	–0,25	–0,27	–3,39	0,224
Сводные результаты для группы АДМ											
АДМ–$ВО_{исх}$	– –	– 0	0 0	+ –	– –	0 0	0 0	– –	+ 0	– –	0 0
АДМ–$ВО_{кор}$	– 0	0 0	0 0	0 0	– 0	0 –	0 +	0 0	0 0	– –	0 0

Критические значения по обоим критериям (Крамера–Уэлча и *Q*) принимались во всех случаях равными 1,96, что соответствует уровню значимости $\alpha = 0,05$ [14, 15]. Знаки «+» и «–» в сводных результатах расчетных значений критериев означают знак разности между средними в случае, если разница оказалась статистически значимой, а знак «0» говорит о ее незначимости.

На рис. 1 изображена гистограмма по количественному показателю «Эозинофилы», построенная для скорректированной группы ЭМИ (после удаления двух записей, относящихся к женщинам), а на рис. 2 – такая же гистограмма для группы $ВО_{Э(кор)}$ (указаны средние арифметические значения и назначенные интервалы нормы).

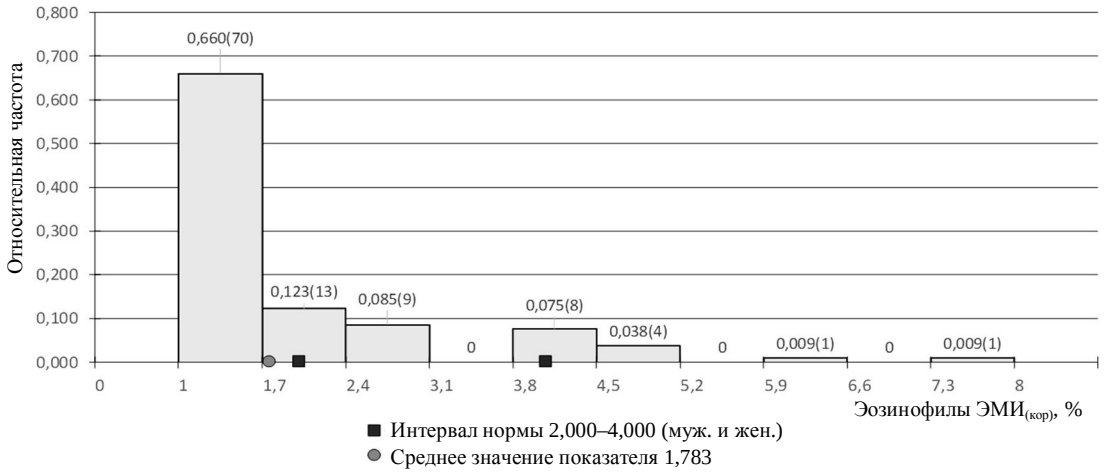

Рис. 1. Гистограмма по количественному показателю «Эозинофилы» для группы $ЭМИ_{(кор)}$

Fig. 1. Histogram for the quantitative indicator “Eosinophils” for $EMR_{(kor)}$ group

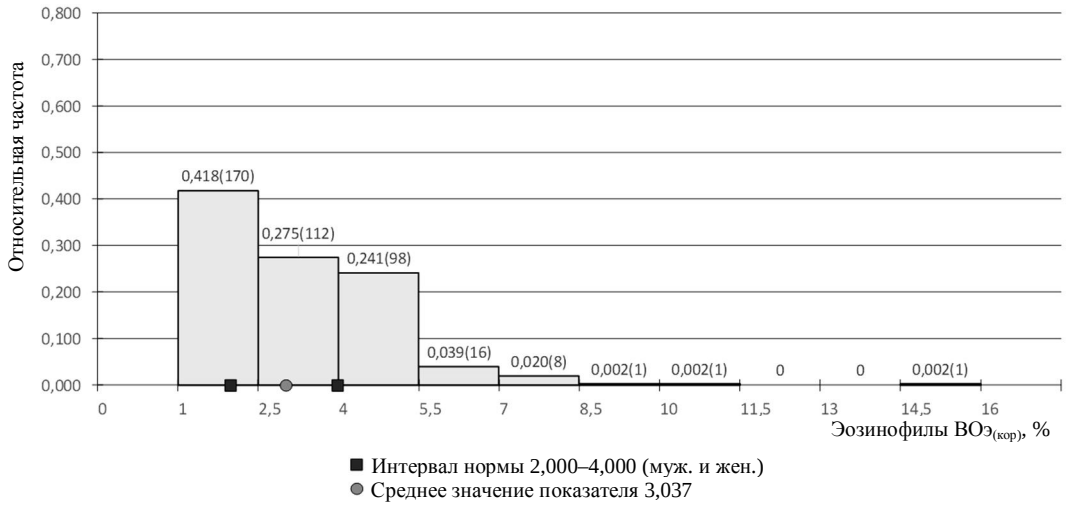

Рис. 2. Гистограмма по количественному показателю «Эозинофилы» для группы $ВО_{Э(кор)}$

Fig. 2. Histogram for the quantitative indicator “Eosinophils” for $BO_{(кор)}$ group

Ситуация по эозинофилам в группах $ЭМИ_{(кор)}$ и $ВО_{Э(кор)}$ является примером того, что результаты сравнения (определения знака и значимости разницы) количественных и бинарных выборок показателей крови могут не только не совпадать, но даже быть противоположными. В этом примере прослеживается зависимость показателя «Эозинофилы» от воздействия ЭМИ. Судя по рис. 1 и 2, вид распределения – как в ЭМИ, так и в ВОЭ – далек от нормального, т. е. важная предпосылка применения традиционных параметрических методов здесь не выполняется. Значимость зависимости лабораторных показателей от условий труда, по нашему мнению, выражается наличием второго знака «+» во второй строке в секциях «Сводные результаты…» табл. 2–5, который отражает сравнение бинарных выборок по критерию *Q* для скорректированных выборок. Это связано с тем, что сравнение неоднородных по полу и возрасту исходных выборок нелегитимно, сравнение количественных выборок по критерию *K* не имеет прямой связи с заболеваемостью. В то же время наличие второго знака «–» вряд ли говорит о том, что производственные факторы положительно влияют на показатели крови.

На рис. 3, 4 представлены гистограммы количественного показателя «Моноциты» для групп $ТАМ_{(кор)}$ и $ВО_{Т(кор)}$ соответственно.

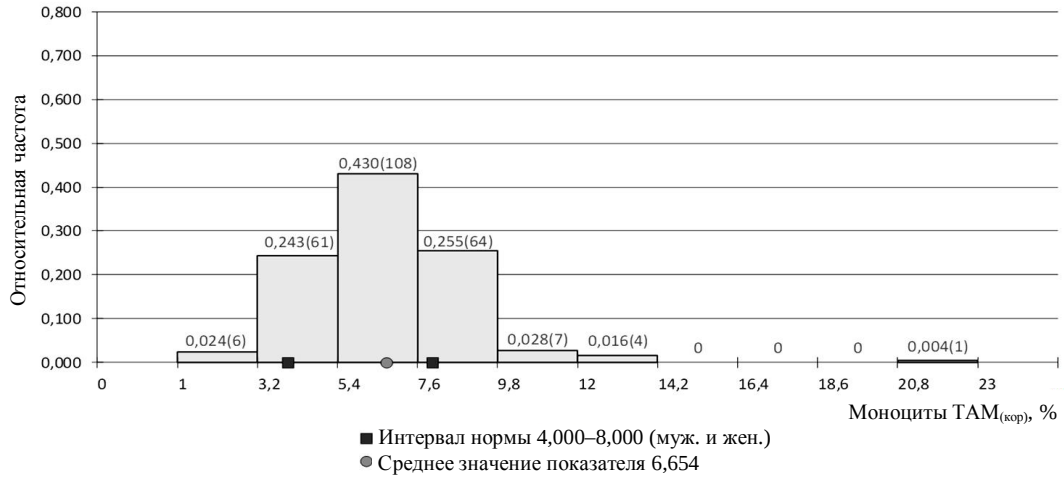

Рис. 3. Гистограмма по количественному показателю «Моноциты» для группы $ТАМ_{(кор)}$

Fig. 3. Histogram for the quantitative indicator “Monocytes” for ТАМ (кор) group

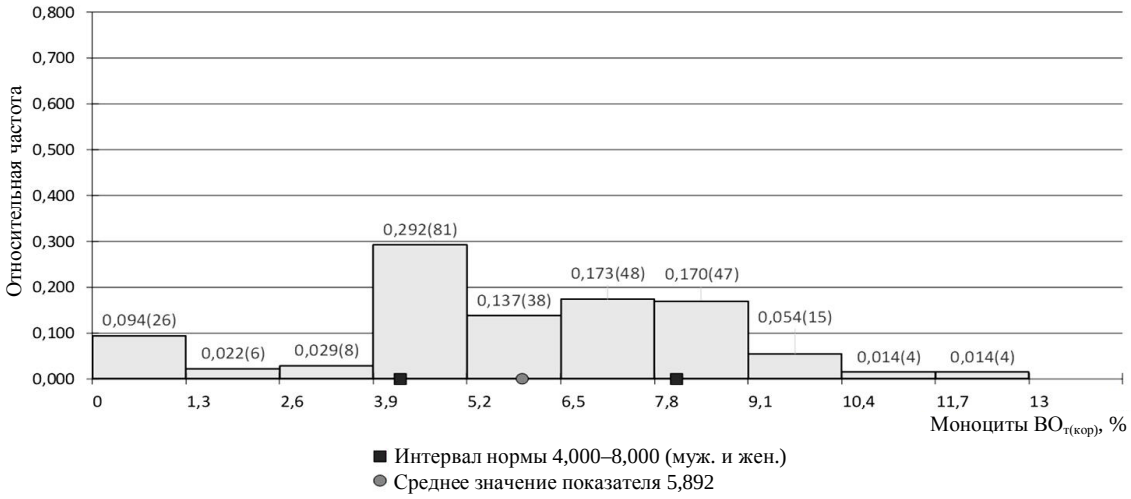

Рис. 4. Гистограмма по количественному показателю «Моноциты» для группы $ВО_{Т(кор)}$

Fig. 4. Histogram for the quantitative indicator “Monocytes” for $ВО_{Т(кор)}$ group

В результате анализа данных табл. 1–5 установлено, что в группе ЭМИ значимо больше выходов за пределы нормы обнаружено по лимфоцитам, моноцитам, эозинофилам и общему холестерину; в группе ШиВ таких показателей крови не нашлось; в группе ТАМ значимо больше выходов за пределы нормы обнаружено по гемоглобину; в группе АДМ – по эозинофилам. Сопоставление результатов по исходным и скорректированным выборкам при расчетах по *K* и *Q* (см. табл. 2–5) показывает, что по обоим критериям статистический вывод нигде не получался противоположным (не было случаев, когда по исходным выборкам получалось бы, что средние значения или частоты выхода за пределы нормы в рассматриваемой группе больше, а по скорректированным – наоборот, меньше). В то же время он не всегда получался одинаковым: по многим показателям в разных группах разница оказывалась значимой для исходных выборок и незначимой для скорректированных и наоборот. Это можно рассматривать как подтверждение необходимости приведения сравниваемых выборок к однородности по полу и возрасту. Сопоставление результатов по *K* и *Q* по скорректированным выборкам (первый и второй знаки во второй сводной строке в табл. 2–5) показывает следующее. По группе ЭМИ знак «+» не совпал нигде, знак «–» совпал для 2-х показателей, знак «0» – для 3-х показателей. По группе ШиВ знаки «+» и «–» не совпали нигде, знак «0» совпал для 6 показателей. По группе ТАМ знак «+» совпал для 2 показателей, знак «–» не совпал нигде, знак «0» совпал для 6 показателей. По группе АДМ знак «+» не совпал нигде, знак «–» совпал для 1 показателя, знак «0» – для 6 показателей. Таким образом, выводы по критериям *K* и *Q* намного чаще совпадали в случае незначимости разницы

между выборками. Проиллюстрированные на рис. 1–4 примеры соответствуют случаям противоположных знаков в сводных строках табл. 2–5. Применение обоих критериев правомерно, при этом критерии дополняют друг друга. Критерий *K* позволяет выявить значимость влияния производственных факторов на среднее значение показателя крови, а критерий *Q* – оценить влияние производственных факторов на частоту выхода за пределы нормы. Для комплексного исследования, направленного на изучение биологического механизма влияния производственных факторов на показатели крови, целесообразно применять оба эти критерия. Учитывая, что критерий *K* относится к параметрическим методам, представляется целесообразным исследовать, насколько законы распределения действительно близки к нормальному и, возможно, применить непараметрические методы сравнения распределений количественного показателя. Применение критерия *Q* требует бинаризации, результат которой зависит от принятого интервала нормы, поэтому для метода бинарных выборок целесообразно исследовать влияние границ нормы на получаемые в результате по критерию *Q* статистические выводы.

Заключение

В работе предложен метод анализа медицинских данных, накапливаемых в транзакционных информационных системах медицинских учреждений. Метод направлен на выявление зависимости показателей крови и заболеваемости от производственных факторов. Он основан на известной формуле определения статистической значимости разности частот сравниваемых бинарных выборок. Продемонстрировано новое применение ключевых математических формул на медицинских данных. Для количественных показателей крови, определяемых лабораторно, предлагается использовать алгоритм бинаризации, использующий сопоставление значения показателя с заранее известными границами интервала нормы. В результате исследования были выявлены показатели крови, для которых число выходов за пределы нормы значимо больше в рассматриваемой группе, чем в совокупности остальных. Показано, что для оценки значимости зависимости используемых для диагностики лабораторных показателей крови от условий труда метод бинарных выборок является более информативным с точки зрения оценки профессиональной заболеваемости, в то время как методы сравнения средних значений двух количественных выборок более информативны для изучения биологического механизма этой зависимости.

Нами ведется разработка соответствующей программной оболочки, основу которой составит данный метод, и технологий наполнения контента, что позволит более эффективно управлять медицинскими данными с целью поддержки принятия врачебных решений.

Список источников

1. *Программа* «Цифровая экономика РФ» (утв. 04.06.2019 г., протокол № 7). URL: https://digital.gov.ru/ru/activity/directions/858/ (дата обращения: 10.03.2021).

2. *Стефанова Н. А., Андронова И. В.* Проблемы цифровизации сферы здравоохранения: российский и зарубежный опыт // Вестн. Самар. ун-та: экономика и управление. 2018. Т. 9. № 3. С. 31–35.

3. *Белышев Д. В.* Анализ методов хранения данных в современных медицинских информационных системах // Программные системы: теория и приложения. 2016. № 2 (29). С. 85–103.

4. *Новокрещенов В. С., Киселев С. Н.* Современные методы хранения данных в медицинских информационных системах // Соврем. науч. исслед. и инновации. 2017. № 4. URL: http://web.snauka.ru/issues/2017/04/81796 (дата обращения: 25.03.2021).

5. *Баранов А. А., Намазова-Баранова Л. С., Смирнов И. В., Девяткин Д. А., Шелманов А. О., Вишнёва Е. А., Антонова Е. В., Смирнов В. И.* Методы и средства комплексного интеллектуального анализа медицинских данных // Тр. ИСА РАН. 2016. Т. 65. № 2. С. 81–93.

6. *Карпов О. Э., Субботин С. А., Шишканов Д. В.* Использование медицинских данных для создания систем поддержки принятия решений // Врач и информ. технологии. 2019. № 2. С. 11–18.

7. *Belle A., Thiagarajan R., Soroushmehr S. M., Navidi F., Beard D. A., Najarian K.* Big Data Analytics in Healthcare // BioMed research international. 2015. V. 2015. P. 1–16.

8. *Yanase J., Triantaphyllou E.* The seven key challenges for the future of computer-aided diagnosis in medicine // Int. J. Med. Inform. 2019. V. 129. P. 413–422.

9. *О персональных* данных: Федеральный закон от 27 июля 2006 г. № 152-ФЗ (ред. от 24 апреля 2020 г.). URL: http://base.garant.ru/5635295/ (дата обращения: 30.01.2021).

10. *Geger E. V., Podvesovskii A. G., Kuzmin S. A., Tolstenok V. P.* Methods for the Intelligent Analysis of Biomedical Data // CEUR Workshop Proceedings of the 29th International Conference on Computer Graphics and Vision (GraphiCon 2019). 2019. V. 2485. P. 308–311.

11. *Гегерь Э. В., Козлова И. Р., Юркова О. Н., Евельсон Л. И.* Методика сравнения бинарных выборок при анализе медицинских данных для принятия управленческих решений // XXI век: итоги прошлого и проблемы настоящего плюс. Информатика, вычислительная техника, управление. 2020. Т. 9. № 2 (50). С. 164–170.

12. *Гегерь Э. В., Евельсон Л. И., Федоренко С. И., Козлова И. Р.* Совершенствование методов обработки данных в информационных системах поддержки принятия управленческих решений // Соврем. наукоемкие технологии. 2019. № 12, ч. 2. С. 276–281.

13. *Орлов А. И.* Прикладная статистика. М.: Экзамен, 2006. 671 с.

14. *Кобзарь А. И.* Прикладная математическая статистика. Для инженеров и научных работников. М.: Физматлит, 2006. 816 с.

15. *Mirkes E. M., Coats T. J., Levesley J., Gorban A. N.* Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes // Computers in Biology and Medicine. 2016. V. 75. P. 203–2016.

References

1. *Programma «Tsifrovaia ekonomika RF» (utv. 4.06.2019 g., protokol № 7)* [Program Digital Economy of the Russian Federation (approved on June 4, 2019, protocol No. 7)]. Available at: https://digital.gov.ru/ru/activity/directions/858/ (accessed: 10.03.2021).

2. Stefanova N. A., Andronova I. V. Problemy tsifrovizatsii sfery zdravookhraneniia: rossiiskii i zarubezhnyi opyt [Problems of digitalization of healthcare sector: Russian and foreign experience]. *Vestnik Samarskogo universiteta: ekonomika i upravlenie*, 2018, vol. 9, no. 3, pp. 31-35.

3. Belyshev D. V. Analiz metodov khraneniia dannykh v sovremennykh meditsinskikh informatsionnykh sistemakh [Analysis of data storage methods in modern medical information systems]. *Programmnye sistemy: teoriia i prilozheniia*, 2016, no. 2 (29), pp. 85-103.

4. Novokreshchenov V. S., Kiselev S. N. Sovremennye metody khraneniia dannykh v meditsinskikh informatsionnykh sistemakh [Modern methods of data storage in medical information systems]. *Sovremennye nauchnye issledovaniia i innovatsii*, 2017, no. 4. Available at: http://web.snauka.ru/issues/2017/04/81796 (accessed: 25.03.2021).

5. Baranov A. A., Namazova-Baranova L. S., Smirnov I. V., Deviatkin D. A., Shelmanov A. O., Vishneva E. A., Antonova E. V., Smirnov V. I. Metody i sredstva kompleksnogo intellektual'nogo analiza meditsinskikh dannykh [Methods and means of complex intellectual analysis of medical data]. *Trudy ISA RAN*, 2016, vol. 65, no. 2, pp. 81-93.

6. Karpov O. E., Subbotin S. A., Shishkanov D. V. Ispol'zovanie meditsinskikh dannykh dlia sozdaniia sistem podderzhki priniatiia reshenii [Using medical data to create decision support systems]. *Vrach i informatsionnye tekhnologii*, 2019, no. 2, pp. 11-18.

7. Belle A., Thiagarajan R., Soroushmehr S. M., Navidi F., Beard D. A., Najarian K. Big Data Analytics in Healthcare. *BioMed research international*, 2015, vol. 2015, pp. 1-16.

8. Yanase J., Triantaphyllou E. The seven key challenges for the future of computer-aided diagnosis in medicine. *Int. J. Med. Inform.*, 2019, vol. 129, pp. 413-422.

9. *O personal'nykh dannykh Federal'nyi zakon ot 27 iiulia 2006 g. № 152-FZ (24 aprelia red. ot 2020 g.)* [On personal data Federal Law of July 27, 2006 No. 152-FZ (April 24 edition of 2020)]. Available at: http://base.garant.ru/5635295/ (accessed: 30.01.2021).

10. Geger E. V., Podvesovskii A. G., Kuzmin S. A., Tolstenok V. P. Methods for the Intelligent Analysis of Biomedical Data. *CEUR Workshop Proceedings of the 29th International Conference on Computer Graphics and Vision (GraphiCon 2019)*, 2019, vol. 2485, pp. 308-311.

11. Geger' E. V., Kozlova I. R., Iurkova O. N., Evel'son L. I. Metodika sravneniia binarnykh vyborok pri analize meditsinskikh dannykh dlia priniatiia upravlencheskikh reshenii [Methods for comparing binary samples in analysis of medical data for making managerial decisions]. *XXI vek: itogi proshlogo i problemy nastoiashchego plius. Informatika, vychislitel'naia tekhnika, upravlenie*, 2020, vol. 9, no. 2 (50), pp. 164-170.

12. Geger' E. V., Evel'son L. I., Fedorenko S. I., Kozlova I. R. Sovershenstvovanie metodov obrabotki dannykh v informatsionnykh sistemakh podderzhki priniatiia upravlencheskikh reshenii [Improving data processing methods in information systems for supporting managerial decision-making]. *Sovremennye naukoemkie tekhnologii*, 2019, no. 12, part 2, pp. 276-281.

13. Orlov A. I. *Prikladnaia statistika* [Applied statistics]. Moscow, Ekzamen Publ., 2006. 671 p.

14. Kobzar' A. I. *Prikladnaia matematicheskaia statistika. Dlia inzhenerov i nauchnykh rabotnikov* [Applied Mathematical Statistics. For engineers and scientists]. Moscow, Fizmatlit Publ., 2006. 816 p.

15. Mirkes E. M., Coats T. J., Levesley J., Gorban A. N. Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. *Computers in Biology and Medicine*, 2016, vol. 75, pp. 203-2016.

Статья поступила в редакцию 26.01.2022; одобрена после рецензирования 25.02.2022; принята к публикации 01.04.2022
The article is submitted 26.01.2022; approved after reviewing 25.02.2022; accepted for publication 01.04.2022

Информация об авторах / Information about the authors

Лев Игоревич Евельсон – кандидат технических наук, доцент; директор по научным исследованиям и инновациям; Научно-инновационный центр информационных и дистанционных технологий; levelmoscow@mail.ru

Эмилия Владимировна Гегерь – доктор биологических наук, доцент; заведующий кабинетом статистики; Брянский клинико-диагностический центр; emiliya_geger@mail.ru

Ирина Романовна Козлова – аспирант кафедры компьютерных технологий и систем; Брянский государственный технический университет; kozlowa.iri2014@yandex.ru

Lev I. Evelson – Candidate of Technical Sciences, Assistant Professor; Director of the Research and Innovation Centre; "Innovation Scientific Centre of Information and Distance Technologies", LLC; levelmoscow@mail.ru

Emiliya V. Geger – Doctor of Biology, Assistant Professor; Head of the Statistics Department; Bryansk Clinical Diagnostic Center; emiliya_geger@mail.ru

Irina R. Kozlova – Postgraduate Student of the Department of Computer Technologies and Systems; Bryansk State Technical University; kozlowa.iri2014@yandex.ru