УДК 004.6
DOI: 10.46548/21vek-2020-0950-0029

МЕТОДИКА СРАВНЕНИЯ БИНАРНЫХ ВЫБОРОК ПРИ АНАЛИЗЕ МЕДИЦИНСКИХ ДАННЫХ ДЛЯ ПРИНЯТИЯ УПРАВЛЕНЧЕСКИХ РЕШЕНИЙ

Гегерь Эмилия Владимировна, доктор биологических наук,
профессор кафедры «Безопасность жизнедеятельности и химия»
Брянский государственный технический университет
(241035, Россия, г. Брянск, Бульвар 50-летия Октября, д. 7, e-mail: emiliya_geger@mail.ru)
Козлова Ирина Романовна, аспирант,
Брянский государственный технический университет
(241035, Россия, г. Брянск, Бульвар 50-летия Октября, д. 7, e-mail: kozlowa.iri2014@yandex.ru)
Юркова Ольга Николаевна, кандидат экономических наук,
доцент кафедры «Информационные технологии»
Брянский государственный инженерно-технологический университет
(241037, Россия, г. Брянск, проспект Станке Димитрова, 3, e-mail: yurkova_olga@mail.ru)
Евельсон Лев Игоревич, кандидат технических наук,
доцент, директор по научным исследованиям и инновациям
Научно-инновационный центр информационных и дистанционных технологий
(241007, Россия, г. Брянск, Россия, ул. Бежицкая, 1/4, e-mail: levelmoscow@mail.ru)

Аннотация. В статье описывается исследование зависимости заболеваемости, а также показателей лабораторных исследований от наличия вредных производственных факторов шума и вибрации. Развивается и используется подход, основанный на сравнении бинарных выборок. Анализ медицинских данных, накопленных в медицинской информационной системе транзакционного типа, дает возможность выявлять заболевания, характерные для определенных комплексов вредных факторов, связанных с производственной деятельностью, что позволит усовершенствовать диагностику и лечение, используя цифровые технологии, и поможет принятию правильных управленческих решений. Результаты выполнявшихся общих анализов крови и мочи приводились к бинарному виду путем их сопоставления с известным интервалом статистической нормы, а выставлявшиеся диагнозы рассматривались как изначально бинарные величины. Полученные в результате бинаризации выборки для двух групп, первая группа включает в себя лица, в производственной деятельности которых присутствуют вредные факторы, а вторая – тех, у которых эти факторы отсутствуют, сравнивались между собой. Разработана и применена методика корректировки выборок, позволяющая привести выборки, исходно неоднородные по признакам пола и возраста, к однородным одновременно по обоим признакам, что дает возможность корректно сравнивать показатели лабораторных исследований, а также диагнозы. Рекомендовано внедрение разработанного метода для анализа данных, содержащихся в медицинских информационных системах, применительно к различным профессиональным группам. Это позволит контролировать риски в системе управления охраной труда и принимать грамотные врачебные решения.

Ключевые слова: медицинские данные, бинарные выборки, анализ данных, вредные производственные факторы, шум, вибрация.

METHOD FOR COMPARING BINARY SAMPLES IN THE ANALYSIS OF MEDICAL DATA FOR MAKING MANAGERIAL DECISIONS

Geger Emiliya Vladimirovna, doctor of biological sciences,
professor of the Department "Life Safety and Chemistry"
Bryansk State Technical University
(241035, Russia, Bryansk, Blvd. 50 years of October, 7, e-mail: emiliya_geger@mail.ru)
Kozlova Irina Romanovna, postgraduate student,
Bryansk State Technical University
(241035, Russia, Bryansk, Blvd. 50 years of October, 7, e-mail: kozlowa.iri2014@yandex.ru)
Yurkova Olga Nikolaevna, candidate of economic sciences,
associate Professor of the Department «Information technologies»
Bryansk State Engineering and Technological University
(241037, Russia, Bryansk, Dimitrov prospect, 3, e-mail: yurkova_olga@mail.ru)
Evelson Lev Igorevich, candidate of technical sciences,
associate professor Director of research and innovation
«Innovation Scientific Centre of Information and Remote Technologies», Limited Liability Company,
(241007, Russia, Bryansk, Bezhitskaya St. 1/4, e-mail: levelmoscow@mail.ru)

Abstract. The article describes the study of the dependence of morbidity, as well as indicators of laboratory research

on the presence of harmful industrial factors of noise and vibration. An approach based on comparing binary samples is being developed and used. The analysis of medical data accumulated in a transactional medical information system makes it possible to identify diseases that are characteristic of certain complexes of harmful factors associated with industrial activities, which will improve diagnostics and treatment using digital technologies, and help to make the right management decisions. The results of the performed clinical blood analysis and urine tests were brought to a binary form by comparing them with a known interval of statistical norm, and the diagnoses presented were considered as initially binary values. The samples obtained as a result of binarization for two groups, the first group includes people who have harmful factors in their production activities, and the second group includes those who do not have these factors, were compared with each other. A method of sample adjustment has been developed and applied that allows to bring samples that are initially heterogeneous in terms of gender and age to homogeneous simultaneously in terms of both characteristics, which makes it possible to compare correctly laboratory research indicators, as well as diagnoses. It is recommended to implement the developed method for analyzing data contained in medical information systems in relation to various professional groups. It will allow to control risks in the occupational safety management system and make competent medical decisions.

Keywords: medical data, binary samples, data analysis, harmful occupational factors, noise, vibration.

Введение. В системе здравоохранения быстрыми темпами идет информатизация, что приводит к накоплению больших объемов данных о лечебно-диагностическом процессе медицинских учреждений, являющихся источником информации для подготовки управленческих решений [1, с. 82; 2, p. 309].

Сохранение здоровья работающего населения является приоритетным направлением государственной политики в области охраны труда и профилактики профессиональной заболеваемости. Оценка уровня вредного воздействия на работников в процессе их трудовой деятельности отдельных факторов трудового процесса и выработка механизмов управления ими с целью снижения до уровней приемлемых рисков позволяет сохранять профессиональное здоровье работающих и ведет к сбережению трудовых ресурсов [3, с. 80; 4, с. 155; 5, с. 15; 6, с. 78; 7, с. 3-5].

Анализ клинических данных о пациенте на основе обработки различных массивов медицинских данных с целью принятия обоснованных врачебных решений занимают в настоящее время особое место в информационных технологиях [8, с.122; 9].

Комплексная статистическая обработка результатов исследований представляет собой сложную задачу и является грамотным и достоверным инструментом для интерпретации данных и принятия правильных управленческих решений [2, p. 310-312; 10, с. 12; 11, с.82].

Статистическое описание данных медицинских исследований и оценка значимости различий величин, отражающих действенность проводимых профилактических, диагностических и лечебных процедур, являются основой доказательной медицины. Для анализа медицинской информации в настоящее время используются разнообразные статистические методы. В современной литературе они достаточно широко освещены [12, с. 8-9; 13, с.126-127; 14, с. 178; 15; 16, с. 24-27].

Анализ медицинских данных, как правило, основывается на строгом учете статистических закономерностей. Для обработки медицинской информации используют различные методы математической статистики, выбор одного из которых в каждом конкретном случае основывается на характере распределения анализируемых данных [17, с. 50].

В нашем исследовании для анализа данных в медицинских информационных системах, в частности, выявления характерных заболеваний, присущих действию вредных производственных факторов шума и вибрации, предложено и изучается применение метода сравнения бинарных выборок.

Преимущества данного метода в отличие от параметрических методов заключается в том, что он не требует выполнения серьезных допущений о виде закона распределения. По сравнению с непараметрическими методами его преимущество заключается в том, что он менее чувствителен к объему выборок и значительно проще в реализации.

Материалы и результаты исследования. В соответствии с поставленной целью выявления взаимосвязи лабораторных показателей общего анализа крови (ОАК) и общего анализа мочи (ОАМ), а также выставляемых диагнозов согласно международной классификации болезней (МКБ-10) [18], связанных с наличием производственных вредностей, для работников, имеющих вредные производственные факторы – шум и вибрацию (группа I), из медицинской информационной системы (МИС) по результатам периодических медицинских осмотров отбирались данные, относящиеся к этой группе.

Вторую группу лиц, работа которых не связана с вредными производственными факторами шума и вибрации, планировалось взять в качестве контрольной группы условно здоровых людей (группа II).

Рассматривались бинарные данные, которые являются результатами измерений альтернативного признака. Бинарные случайные величины принимают только два возможных значения – «0» и «1».

В процессе исследования решался вопрос о значимости различия средних частот двух выборок бинарных (двоичных) данных, т.е. данных, которые могут быть представлены закодированным ответом на вопрос, на который можно ответить «да» или «нет» («да», выходит за границы нормы, или «нет», не выходит).

Такая выборка характеризуется объемом n и

частотой $p=m/n$, с которой в рассматриваемой выборке встречается ответ «да» m и по которой оценивается соответствующая вероятность p. В вероятностной модели предполагается, что m является биномиальной случайной величиной $B(n,p)$, т.е. случайной величиной с параметрами n – объем выборки и p – вероятность определенного ответа (например, «да»). Такая случайная величина может быть представлена в виде: $m=X_1+X_2+...+X_i$. (1)

где m – число ответов «да»; X_i – это независимые одинаково распределенные случайные величины, которые могут принимать одно из двух значений (1 или 0), причем, если $P(X_i=1)=p$, то $P(X_i=0)=1-p$ [19, 20].

В нашей задаче применение метода бинарных выборок основано на сравнении значений индикаторных показателей с известной нормой, что дает возможность неявно использовать результаты ранее проводившихся статистических исследований, в результате которых были установлены границы интервала нормы.

Метод, основанный на сопоставлении исследуемых групп по показателям лабораторных исследований, предусматривал бинаризацию результатов лабораторных анализов ОАК и ОАМ по признаку соответствия принятой норме, принимающих только два возможных значения – «да» или «нет», т.е. «соответствует» или «не соответствует». Если значение какого-либо показателя выходит за пределы нормы, то соответствующей бинарной величине присваивается значение «1», в противном случае – значение «0».

На предварительном этапе осуществлялась консолидация данных на основе медицинской информационной системы.

Рассматриваемая исходная группа оказалась неоднородной по отношению к другой группе. Было принято решение провести исследование, разработать и апробировать методику корректировки выборок с целью достижения однородности при максимальном сохранении данных, используемых для анализа.

Графически методика представлена на рисунке 1.

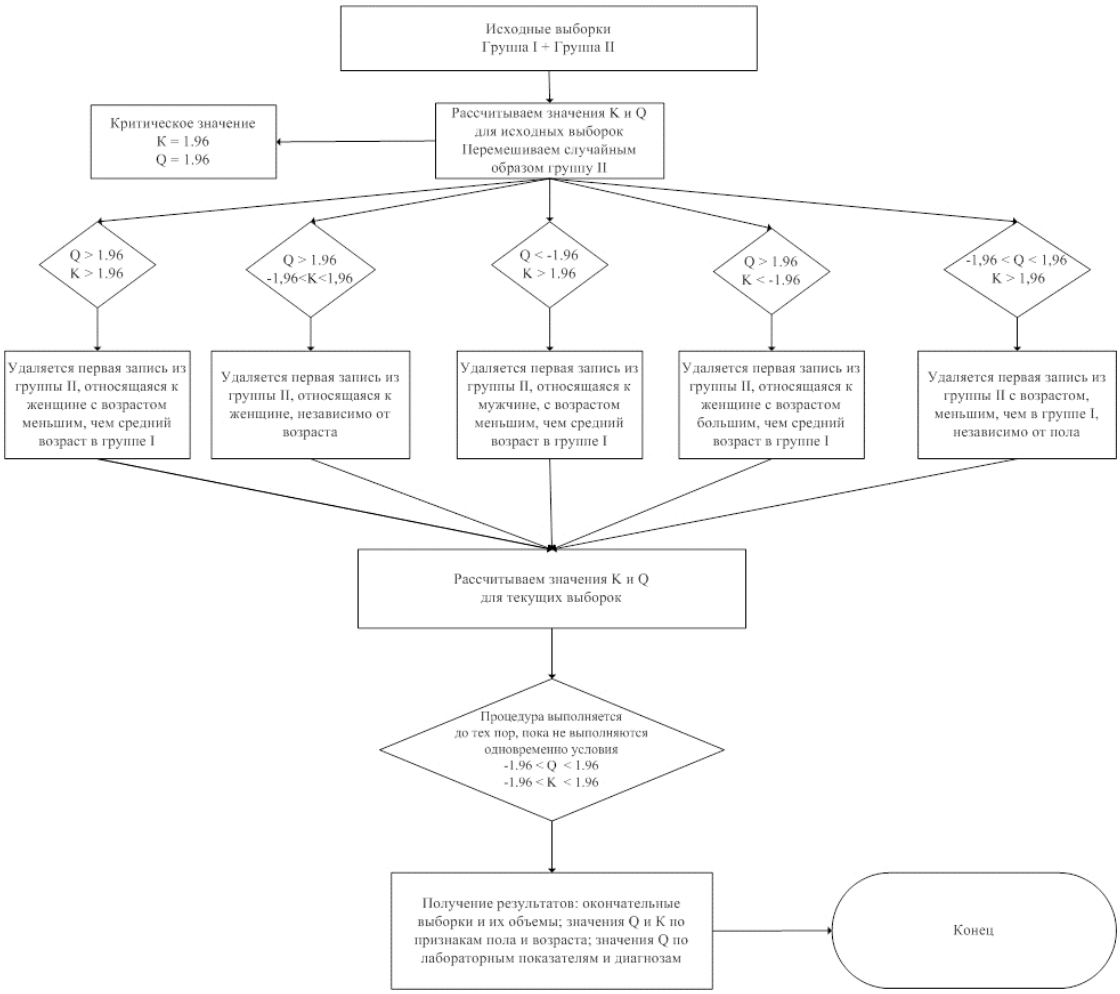

Рисунок 1 – Алгоритм методики корректировки выборок при сравнении групп I и II (Q – критерий по полу; K – критерий по возрасту)

С учетом анализа задачу нашего исследования можно сформулировать следующим образом: разработка и применение методики корректировки выборок, позволяющей привести выборки, исходно неоднородные по признакам пола и возраста, к однородным одновременно по обоим признакам, что

дает возможность корректно сравнивать лабораторные показатели исследований, а также диагнозы (рис. 1).

Предварительный расчет показал, что две группы неоднородны по полу и возрасту. Методика также построена на принципах рандомизации и экономии информации. Рандомизация означает, что строки сначала случайным образом перемешиваются, а уже затем последовательно удаляются по одной, вплоть до выполнения критерия однородности. После этого производится новый расчет по лабораторным показателям. В качестве критерия однородности по признаку пола использовалась величина Q, определяемая по формуле критерия сравнения частот бинарных выборок (2) [19], а по количественному признаку возраста использовался критерий Крамера – Уэлча (3) [20].

$$Q = \frac{p_1^* - p_2^*}{\sqrt{\frac{p_1^*(1-p_1^*)}{n_1} + \frac{p_2^*(1-p_2^*)}{n_2}}} \quad (2)$$

где звездочками обозначены выборочные частоты бинарных выборок, являющиеся оценками соответствующих вероятностей: $p^*_i = m_i/n_i$, n_1 – объем выборки I; n_2 – объем выборки II; m_i – количество значений, выходящих за пределы нормы в выборке I; m_2– количество значений, выходящих за пределы нормы в выборке II. В методике корректировки выборок по возрасту использовался критерий Крамера – Уэлча t_k (3). В данном случае критерий используется традиционным для статистических методов образом как критерий значимости разницы средних значений двух количественных выборок [19]:

$$t_k = \frac{1}{s}(\bar{x} - \bar{y}) \quad (3)$$

где

$$s = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (4)$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (5)$$

$$s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{m}(y_i - \bar{y})^2 \quad (6)$$

где $\bar{x}$ – выборочное среднее арифметическое значение возраста выборки I; $\bar{y}$ – выборочное среднее арифметическое значение возраста выборки II; n_1 – количество значений в выборке I; n_2 – количество значений в выборке II; s_1^2 – несмещенная (исправленная) оценка дисперсии выборки I; s_2^2 – несмещенная (исправленная) оценка дисперсии выборки II; s – несмещенная (исправленная) оценка дисперсии разности выборочных средних рассматриваемых выборок.

В таблице 1 представлены исходные (до удаления строк) объемы выборок I и II и результаты сравнения этих двух выборок по критериям пола и возраста, произведенного по формулам (2) и (3). Приведены результаты корректировки: объемы скорректированных выборок и расчетные значения критериев однородности по полу и возрасту.

Таблица 1 – Сравнение исходных выборок (до и после корректировки)

Наименование выборки	Объем выборки	Кол-во мужчин в выборке	Кол-во женщин в выборке	Средний возраст	Расчетное значение Q*	Расчетное значение К*	Критическое значение критериев
Исходные объемы выборок (до удаления строк)							
Группа I	149	139	10	51,7	6,72	11,1	1,96
Группа II	506	376	130	41,4			
	Результаты корректировки выборок						
Группа I	149	139	10	51,7	1,896	1,896	1,96
Группа II	136	118	18	49,5			

*Q – критерий по полу; К – критерий по возрасту

Как видно из таблицы 1 сравниваемые исходные выборки оказались неоднородными как по полу, так и по возрасту.

Далее были получены результаты сравнения бинарных выборок по лабораторным показателям (по числу выходов этих показателей за пределы нормы) и по выставлявшимся диагнозам.

Процедура расчета, в соответствии с разработанной методикой корректировки выборок, была построена следующим образом:

1. Перемешиваем случайным образом выборку II и затем удаляем поочередно по одной записи, удовлетворяющей критерию удаления, но критерий проверяем сразу по двум признакам: полу и возрасту. В качестве критерия по возрасту примем Критерий Крамера – Уэлча, формула (3), критическое значение равно 1,96. В качестве критерия по полу берем формулу для вычисления Q (2), и то же критическое значение 1,96. До корректировки было Q>1,96 и расчетное значение по Крамеру – Уэлчу тоже K>1,96. Следующая запись в группе II удаляется, если она относится к женщине с возрастом меньше, чем средний возраст в группе I. Процедура такого удаления осуществляется до тех пор, пока хотя бы одно из значений Q и K не станет меньше 1,96.

2. Если возникнет ситуация Q<1,96; K>1,96, то далее удаляются только записи в группе II со средним возрастом меньше, чем средний возраст в группе

I, независимо от пола. Если возникнет ситуация, наоборот, Q>1,96; K<1,96, то далее удаляются только женщины в группе II, независимо от возраста. При этом оба признака продолжают контролироваться.

3. Если возникнет ситуация Q<-1,96, K>1,96, то далее удаляем мужчин со средним возрастом меньше, чем средний возраст в группе I. Если возникнет ситуация Q>1,96; K<-1,96, то далее удаляем женщин со средним возрастом больше, чем средний возраст в группе I.

4. Процедура выполняется до тех пор, пока оба расчетных значения критериев не станут одновременно меньше по модулю, чем 1,96. При этом процедура заканчивается, и результаты по лабораторным показателям и диагнозам (а также конечное значение числа записей в выборке II, n_2) принимаются за окончательные результаты расчета.

Для достижения однородности по обоим признакам (полу и возрасту) пришлось очень существенно сократить выборку II, так как иначе не удавалось достичь выполнения критерия однородности.

Результаты расчета по скорректированным выборкам лабораторных показателей для обоих критериев представлены в таблице 2.

Таблица 2 – Результаты сравнения частот выхода за пределы нормы по скорректированным выборкам лабораторных показателей

Наименование показателей	ШИВ		Все остальные		Расчетное значение Q после корректировки	Расчетное значение Q до корректировки
	p1		p2			
Гемоглобин	62	0,416	54	0,397	0,327	0,415
Лейкоциты (ОАК)	16	0,107	9	0,066	1,244	1,489
Тромбоциты	12	0,081	6	0,044	1,282	1,512
Лимфоциты	32	0,215	31	0,228	-0,268	-0,343
Моноциты	28	0,188	36	0,265	-1,550	-2,046
Эритроциты (ОАК)	74	0,497	34	0,50	4,461	5,450
Ретикулоциты	1	0,007	20	0,147	-4,513	-8,204
Эозинофилы	47	0,315	64	0,471	-2,709	-3,521
Гематокрит	53	0,356	45	0,331	0,441	0,558
СОЭ	21	0,141	27	0,199	-1,293	-1,715
Лейкоциты (ОАМ)	26	0,174	12	0,088	2,185	2,571
Эритроциты (ОАМ)	24	0,161	17	0,125	0,872	1,076
Общий холестерин	90	0,604	75	0,551	0,898	1,148
Глюкоза	15	0,101	9	0,066	1,058	1,277

Заключение. Таким образом, значимо больше отклонения от нормы в группе I имеют место только по эритроцитам в ОАК и лейкоцитам в ОАМ.

1. Разработана и применена методика корректировки выборок I и II, позволяющая привести выборки, исходно неоднородные по признакам пола и возраста, к однородным одновременно по обоим признакам, что дает возможность корректно сравнивать показатели лабораторных исследований и диагнозы.

2. Найдены лабораторные показатели ОАК и ОАМ, для которых выходы за пределы нормы встречаются значимо чаще в группе I: эритроциты в крови и лейкоциты в моче.

3. Найдены диагнозы, которые значимо чаще встречаются в группе I: *H*35.0 (Периферические ретинальные дегенерации); *H*52.0 (Гиперметропия); *E*78 (Чистая гиперхолестеринемия); *J*44.9 (Хроническая обструктивная легочная болезнь неуточненная); *R*73.0 (Отклонения результатов нормы теста на толерантность к глюкозе); *R*72 (Аномалия лейкоцитов, не классифицированная в других рубриках.

4. Для всех тех показателей ОАК и диагнозов, по которым получено значимо большее превышение в группе I, этот результат получен как до корректировки выборок, так и после нее. Это подчеркивает статистическую устойчивость полученных результатов и говорит о том, что в данной конкретной задаче неоднородность по полу и возрасту слабо влияет на окончательные выводы.

5. Выявлены лабораторные показатели, отклонения которых от нормы наблюдаются для исходной группы значимо чаще, чем в другой группе, что позволит разработать управленческие решения в проведении профилактических мероприятий.

6. Целесообразно внедрить разработанный метод для анализа данных, содержащихся в информационных системах медицинских организаций, применительно к различным профессиональным группам.

СПИСОК ЛИТЕРАТУРЫ:

1. Баранов А.А., Намазова-Баранова Л.С., Смирнова И.В. и др. Методы и средства комплексного интеллектуального анализа медицинских данных. Труды ИСА РАН. Том 65. 2. 2015. – С.81-93.

2. E. Geger, A. Podvesovskii, S. Kuzmin, V. Tolstenok. 2019. Methods for the Intelligent Analysis of Biomedical Data. GraphiCon 2019.Computer Graphics and Vision Proceedings of

the 29th International Conference on Computer Graphics and Vision (Sep. 2019), Vol. 2485. 308-311. DOI: 10.30987/graphicon-2019-2-308-311.

3. Гегерь Э.В., Федоренко С.И., Евельсон Л.И., Козлова И.Р. Разработка метода оценки профессиональных заболеваний для создания информационной системы производственной безопасности // Вестник НЦ БЖД. 2019. №1 (39). – С. 79-87.

4. Гегерь Э.В. Цифровое здравоохранение: перспективы развития // «Цифровой регион: опыт, компетенции, проекты» Матер. II международной науч.-практ. конф. Брянск. 2019. – С. 153-157.

5. Измеров Н.Ф., Актуализация вопросов профессиональной заболеваемости // Здравоохранение Российской Федерации. 2013. № 2. – С. 14-17.

6. Исмаилова Л.Н. Эффективное управление производственными рисками // Экономика и бизнес: теория и практика. 2016. №5. – С. 77-79.

7. Костенко Н.А. Условия труда и профессиональная заболеваемость как основа управления рисками для здоровья работников: автореф. дис. … канд. мед. наук. М., 2015. – 21 с.

8. Цыганкова И.А. Метод интеллектуальной обработки медико-биологических данных [Текст] / И.А. Цыганкова // Программные продукты и системы. 2009. № 3. – С. 120-123.

9. Bruce McCormick (2014) Update in Anaesthesia. World Federation of Societies of Anaesthesiologists. 466 p.

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

11. Куракова Н.А. Информатизации здравоохранения как инструмент создания «саморегулируемой системы организации медицинской помощи» // Врач и информационные технологии. 2009. №2. – 82 С.

12. Гусев А.В. Рынок медицинских информационных систем: обзор, изменения, тренды // Врач и информационные технологии. 2012. №3. – С. 6-15.

13. Ильин В.П. Корреляционный анализ количественных данных в медико-биологических // Бюлл. ВСНЦ СО РАМН. 2013. № 4. – С. 125–130.

14. Макарова Н.В. Статистический анализ медико-биологических данных с использованием пакетов статистических программ Statistica, SPSS, NCSS, SYSTAT : методическое пособие / Н.В. Макарова ; Всерос. центр экстрен. и радиац. медицины им. А.М. Никифорова МЧС России – СПб.: Политехника-сервис, 2012. – 178 с.

15. Программа «Цифровая экономика Российской Федерации», утвержденная протоколом заседания президиума Совета при Президенте Российской Федерации по стратегическому развитию и национальным проектам от 4 июня 2019 г. № 7 [Электронный ресурс] // КонсультантПлюс. URL: https://digital.gov.ru/ru/activity/directions/858/ (дата обращения: 10.03.2020).

16. Сташевский П.С. Поддержка принятия решений в здравоохранении с использованием показателя популяционного риска заболеваемости: автореф. дис.... канд. тех. наук. Новосибирск, 2014. – 20 с.

17. Свальковский А.В., Захаров С.Д. Аналитическая обработка баз данных внедренных информационных систем// Врач и информационные технологии. 2016. №5. – С. 49- 55.

18. Международная классификация болезней 10-го пересмотра (МКБ-10) [Электронный ресурс]. Режим доступа: https://mkb-10.com (дата обращения 20.07.2019).

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

20. Орлов А.И. Прикладная статистика / А.И. Орлов. – М.: Издательство «Экзамен», 2006. – 671 с.

Статья поступила в редакцию 10.05.2020
Статья принята к публикации 10.06.2020