УДК: 004.6
DOI: 10.30987/ 2658-6436-2020-3-21-27

Э.В. Гегерь, И.Р. Козлова

МОДЕЛИРОВАНИЕ И УПРАВЛЕНИЕ МЕДИЦИНСКИМИ ДАННЫМИ

В статье описывается статистический метод анализа медицинских данных, основанный на сравнении бинарных выборок. Обработка данных, которые накапливаются в медицинских информационных системах транзакционного типа, на основе анализа бинарных выборок, позволяет определять те показатели лабораторных исследований и диагнозы, которые характерны для вредных производственных факторов. Это будет способствовать развитию цифровых технологий в здравоохранении, которые позволят совершенствовать как диагностику, так и методы лечения, а также будет содействовать принятию компетентных управленческих решений.

Результаты исследований приводились к бинарному виду путем их сопоставления с интервалом статистической нормы. 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.

Исходная группа оказалась неоднородной по отношению к другой группе в связи с чем было принято решение провести дальнейшее исследование, основанное на разработке и апробации методики корректировки выборок с целью достижения однородности при максимальном сохранении используемых для анализа медицинских данных.

Ключевые слова: *медицинские данные, бинарные выборки, анализ данных.*

E.V. Geger, I.R. Kozlova

MODELING AND MANAGING MEDICAL DATA

The article describes a statistical method for analyzing medical data based on the comparison of binary samples. Processing data that is accumulated in transactional medical information systems, based on the analysis of binary samples, allows you to determine the indicators of laboratory research and diagnoses that are characteristic of harmful production factors. This will contribute to the development of digital technologies in healthcare, which will improve both diagnostics and treatment methods, as well as facilitate the adoption of competent management decisions.

The research results were converted to binary form by comparing them with the statistical norm interval. Diagnoses were considered initially as a binary variable. The samples obtained as a result of binarization for two groups, the first group includes people whose production activities contain harmful factors, and the second – those who do not have these factors, were compared with each other.

The initial group turned out to be heterogeneous in relation to the other group, so it was decided to conduct a further study based on the development and testing of methods for adjusting samples in order to achieve uniformity while maximizing the preservation of medical data used for analysis.

Keywords: *medical data, binary samples, data analysis.*

Введение

Моделирование – важный инструмент планирования, прогнозирования и управления в современной медицине.

Сбор и анализ исходных данных является стратегической функцией разработки

моделей. Расширение возможностей сбора и анализа медицинских данных является крайне актуальным.

Информационная сфера здравоохранения – одна из самых быстрорастущих среди исследованных данных [1].

В медицинских информационных системах накоплены большие объемы информации о лечебно-диагностическом процессе. Особенность этих данных заключается в том, что они идут непрерывным потоком и постоянно накапливаются. Это огромная часть медико-биологических данных, которыми необходимо эффективно управлять и использовать для анализа и получения персональных превентивных рекомендаций [2].

В сложившейся ситуации чрезвычайно сложно в огромном потоке информации выделить ведущие факторы этиологии, патогенеза и клинических симптомов.

Отличительной чертой медико-биологической информации является то, что она обычно представлена в слабоструктурированном или неструктурированном формате [3].

Разнообразие задач, решаемых при изучении анализируемых медицинских данных, особенностях их получения и обработки, диктует необходимость совершенствования подходов к формированию систем сбора и обработки данных медицинских информационных системах транзакционного типа [4].

Анализ предметной области показывает необходимость развития технологий сбора и преобразования медицинских данных из небольших выборок, статистический анализ этих данных позволит провести прогнозную аналитику, увеличить производительность медицинских информационных систем и принять грамотные управленческие решения.

Методы исследования

Обработка данных из небольших слабоструктурированных выборок была сделана нами на основании **анализа** результатов периодических медицинских осмотров в соответствии с Федеральным законом № 152 «О персональных данных» [5].

Построение математической модели, которая описывает собранные медицинские данные, основывалось на статистической оценке значимости разницы между показателями лабораторных исследований и заболеваемостью в группе с наличием вредных производственных факторов и в группе с отсутствием таких факторов [6, 7, 8].

Для построения модели было предложено использовать подход, основанный на анализе бинарных выборок [6, 7, 8].

Преимущества данного метода заключается в том, что в отличие от параметрических методов он не требует выполнения серьезных допущений о виде закона распределения. По сравнению с непараметрическими методами он менее чувствителен к объему выборок и значительно проще в реализации [6].

Для оценки риска влияния факторов производственной среды на здоровье работников нами были сформированы две группы:

К I группе были отнесены лица, трудовая деятельность которых связана с воздействием вредных производственных факторов.

II группу составили лица в профессиональной деятельности, которых отсутствовал вредный производственный фактор.

Рассматривались бинарные данные, которые являются результатами измерений противоположного признака и принимают два возможных значения – «0» и «1» [6, 7, 8].

В процессе исследования ставилась задача определения 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).

Выборка определяется 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* – биномиальная случайная величина *B(n,p)* с параметрами *n* – объем выборки и *p* – вероятность определенного ответа (например, «да») [6].

Такая случайная величина может быть представлена в виде:

$$m = X_1 + X_2 + ... + X_i, \quad (1)$$

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 $P(X_i = 0) = 1 - p$ [10, 11].

В данной задаче применение метода бинарных выборок базируется на сравнении значений индикаторных показателей с общепризнанной нормой, что дает возможность косвенно использовать результаты проводимых статистических исследований, которые позволили установить границы интервала нормы [7, 8, 9].

Метод, основанный на сопоставлении исследуемых групп по показателям **лабораторных исследований,** предусматривал бинаризацию результатов лабораторных анализов – общего анализа крови (ОАК) и общего анализа мочи (ОАМ) по признаку соотношения с принятой нормой, принимающей только два возможных значения – «да» или «нет», т.е. «соответствует» или «не соответствует» [6, 7, 8].

Если значение какого-либо показателя выходит за пределы нормы, то соответствующей бинарной величине присваивается значение «1», в противном случае – значение «0».

Такой же метод предлагается использовать и для сравнения выборок по диагнозам.

В этом случае не нужно делать предварительную бинаризацию, поскольку бинарными данными здесь являются непосредственно факты наличия или отсутствия данного диагноза у конкретного лица.

Предобработка здесь сводится к подсчету, сколько раз встречается конкретный диагноз в данной группе обследовавшихся лиц.

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

Как критерий однородности по признаку «пол» использовалась величина Q**,** определяемая по формуле критерия сравнения частот бинарных выборок (2) [10], а по количественному признаку «возраст» использовался критерий Крамера - Уэлча (3) [11].

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

где звездочками обозначены выборочные частоты бинарных выборок, являющиеся оценками соответствующих вероятностей:

$p_i^* = m_i / n_i$,

где n_1 – объем выборки I;

n_2 – объем выборки II;

m_1 – количество значений, выходящих за пределы нормы в выборке I;

m_2 – количество значений, выходящих за пределы нормы в выборке II.

Применялся критерий Крамера – Уэлча t_k (3). В данном случае критерий используется традиционным в статистике образом как критерий значимости разницы средних значений двух количественных выборок [6, 7]:

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

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.

Были получены результаты сравнения бинарных выборок по лабораторным показателям (по числу выходов этих показателей за пределы нормы) и по выставлявшимся диагнозам, сопоставимые с результатами наших предыдущих исследований [6].

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.

Результаты и их обсуждение

Данное исследование посвящено решению важной задачи, заключающейся в разработке методов и алгоритмов получения и обработки информации для оценки рисков и принятия решений в сфере профилактики профзаболеваний.

В обработке данных использовались средства электронных таблиц MS Excel 2007 с применением встроенных функций. Она включала предварительную обработку и анализ с использованием формул (1), (2), (3). Подобным образом была проведена обработка данных, полученных в ранее проведенных нами исследованиях [9].

Анализировались **показатели лабораторных исследований ОАК и ОАМ и первичная заболеваемость работников промышленной отрасли** и контрольной группы по данным периодических медицинских осмотров, проводилась оценка влияния вредных производственных факторов на здоровье работающих.

Сопоставлялись выборки, относящиеся к исходным группам без объединения и корректировки выборок. Сопоставление выборок осуществлялось на основе методики сравнения бинарных выборок.

Методика этих расчетов опиралась на методы сравнения бинарных выборок по критерию Q (по признаку пола) и сравнения средних значений количественных выборок по критерию Крамера – Уэлча (по признаку возраста).

Показано, что для ситуации, когда частоты в сравниваемых бинарных выборках не слишком малы, эти два критерия дают мало отличающиеся результаты и приводят к одинаковым выводам.

Но когда сравниваемые частоты малы, устойчивость статистических выводов ухудшается, что имеет причиной низкую точность асимптотической аппроксимации биномиального распределения стандартным нормальным в случаях с малыми частотами.

Результаты исследований показывают следующее: статистически значимой оказалось различие между двумя выборками по лейкоцитам и эритроцитам в общем анализе крови, лейкоцитам в общем анализе мочи и содержанию глюкозы. Следует отметить, что отдельные показатели имеют значимую разницу как у лиц I группы, так и у лиц II группы. Данные анализа подтверждают полученные нами ранее результаты [7, 8, 9].

Однако в результате проведенного исследования оказалось, что рассмотренные группы значимо отличаются между собой по признакам пола и возраста.

Соответствующие значения критериев Q и Крамера - Уэлча К оказались по модулю больше критического значения 1,96 (α=0,05).

Результаты проведенных исследований выявили необходимость в проведении дополнительного исследования по скорректированным выборкам лабораторных показателей и диагнозов для обоих критериев.

Выводы исследования

1. Выявлены лабораторные показатели ОАК и ОАМ, для которых выходы за пределы нормы встречаются значимо как в первой, так и во второй группах.

2. Выявлены диагнозы, которые значимо чаще встречаются в группе I: H35.0 (Периферические ретинальные дегенерации); H52.0 (Гиперметропия); E78 (Чистая гиперхолестеринемия); J44.9 (Хроническая обструктивная легочная болезнь неуточненная); R73.0 (Отклонения результатов нормы теста на толерантность к глюкозе); R72 (Аномалия лейкоцитов, не классифицированная в других рубриках).

3. 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.

4. Определена необходимость разработки методики корректировки исходных выборок для приведения их к однородности по признакам пола и возраста.

5. В связи с тем, что рассматривавшиеся группы оказались неоднородными между собой по признакам пола и возраста, а в контрольной группе выход за пределы нормы показателей лабораторных исследований и заболеваемость оказалась выше, чем в группе лиц, работа которых связана с наличием производственных вредностей, появилась необходимость в проведении дополнительного этапа исследования, добавив новую выборку с лицами, в работе которых отсутствует вредный производственный фактор. Каждая из исходных групп должна быть сопоставлена со всеми остальными группами. Такую совмещенную новую группу следует принимать в качестве контрольной именно для конкретной одной рассматриваемой исходной группы.

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

7. Показана актуальность повышения эффективности управленческих решений с целью повышения уровня производственной безопасности на основе статистического анализа заболеваемости работников, взаимосвязанной с условиями труда.

Список литературы:

1. 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// Министерство цифрового развития, связи и массовых коммуникаций Российской Федерации [Электронный ресурс]. – URL:

References:

1. Programma «Cifrovaya ekonomika Rossijskoj Federacii», utverzhdennaya protokolom zasedaniya prezidiuma Soveta pri Prezidente Rossijskoj Federacii po strategicheskomu razvitiyu i nacionalnym proektam ot 4 iyunya 2019 g. № 7// Ministerstvo cifrovogo razvitiya, svyazi i massovyh kommunikacij Rossijskoj Federacii [Elektronnyj resurs]. – URL: https://digital.gov.ru/ru/activity/directions/858/ (data

https://digital.gov.ru/ru/activity/directions/858/ (дата обращения: 10.06.2020)

2. Гегерь, Э.В. Совершенствование методов обработки данных в информационных системах поддержки принятия управленческих решений / Э.В. Гегерь, Л.И. Евельсон, С.И. Федоренко, И.Р. Козлова // Современные наукоемкие технологии. Серия Информатика, вычислительная техника и управление. 05.13.10 – Управление в социальных и экономических системах (технические науки). – 2019. – № 12 (часть 2). – С. 276-281.

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

4. Каширин, И.Ю. Интерактивная аналитическая обработка данных в современных OLAP-системах/ И.Ю. Каширин, С.Ю. Семченков. // Бизнес-информатика. – 2009. – №2 (8). – С. 12-19.

5. О персональных данных: Федеральный закон от 27.07.2006 № 152-ФЗ (ред. от 31.12.2017) // Консультант Плюс [сайт]. – URL: http://www.consultant.ru/document/cons_doc_LAW_61801/ (дата обращения: 05.06.2020).

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

7. Гегерь, Э.В. Разработка метода оценки риска профессиональной заболеваемости, основанного на статистике нечисловых данных /Э.В. Гегерь, С.И. Федоренко, Л.И. Евельсон // Перспективы науки. – 2017. –№11 (98). – С. 7-13.

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

9. Гегерь, Э.В. Разработка метода статистической оценки риска профессиональной заболеваемости, основанного на анализе бинарных выборок / Э.В. Гегерь, С.И. Федоренко, И.Р. Козлова.// Наука и бизнес: пути развития. - 2018. - №3 (81). - С. 97-101.

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

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

obrashcheniya: 10.06.2020)

2. Geger, E.V. Sovershenstvovanie metodov obrabotki dannyh v informacionnyh sistemah podderzhki prinyatiya upravlencheskih reshenij / E.V. Geger, L.I. Evelson, S.I. Fedorenko, I.R. Kozlova // Sovremennye naukoemkie tekhnologii. Seriya Informatika, vychislitelnaya tekhnika i upravlenie. 05.13.10 – Upravlenie v socialnyh i ekonomicheskih sistemah (tekhnicheskie nauki). – 2019. – № 12 (chast 2). – S. 276-281.

3. Baranov, A.A. Metody i sredstva kompleksnogo intellektualnogo analiza medicinskih dannyh / A.A. Baranov, L.S. Namazova-Baranova, I.V. Smirnova. i dr //Trudy ISA RAN. – 2015.– Tom 65. 2. – S. 81-93.

4. Kashirin, I.YU. Interaktivnaya analiticheskaya obrabotka dannyh v sovremennyh OLAP-sistemah / I.YU. Kashirin, S.YU. Semchenkov. // Biznes-informatika. – 2009. – №2 (8). – S. 12-19.

5. O personalnyh dannyh: Federalnyj zakon ot 27.07.2006 № 152-FZ (red. ot 31.12.2017) // Konsultant Plyus [sajt]. – URL: http://www.consultant.ru/document/cons_doc_LAW_61801/ (data obrashcheniya: 05.06.2020).

6. Geger, E.V. Metodika sravneniya binarnyh vyborok pri analize medicinskih dannyh dlya prinyatiya upravlencheskih reshenij / E.V. Geger, I.R. Kozlova, O.N. YUrkova, L.I. Evelson. // XXI vek: itogi proshlogo i problemy nastoyashchego plyus. Informatika, vychislitelnaya tekhnika, upravlenie. – 2020. – №2 (50), T.9. – S. 164-170.

7. Geger, E.V. Razrabotka metoda ocenki riska professionalnoj zabolevaemosti, osnovannogo na statistike nechislovyh dannyh / E.V. Geger, S.I. Fedorenko, L.I. Evelson // Perspektivy nauki. – 2017. – №11 (98). – S. 7-13.

8. Geger, E.V. Razrabotka metoda ocenki professionalnyh zabolevanij dlya sozdaniya informacionnoj sistemy proizvodstvennoj bezopasnosti / E.V Geger, S.I. Fedorenko, L.I. Evelson, I.R. Kozlova // Vestnik NC BZHD. – 2019. – №1 (39). – S. 79-87.

9. Geger, E.V. Razrabotka metoda statisticheskoj ocenki riska professionalnoj zabolevaemosti, osnovannogo na analize binarnyh vyborok / E.V. Geger, S.I. Fedorenko, I.R. Kozlova. // Nauka i biznes: puti razvitiya. - 2018. - №3 (81). - S. 97-101.

10. Kobzar, A.I. Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnyh rabotnikov. – M.: Fizmatlit, 2006. – 816 s.

11. Orlov, A.I. Prikladnaya statistika / A.I. Orlov. – M.: Izdatelstvo «Ekzamen», 2006. – 671 s.

Статья поступила в редколлегию 20.07.2020.
Рецензент:
канд. техн. наук, доц.,
Брянский государственный технический университет
Подвесовский А.Г.
Статья принята к публикации 25.07.2020.

Сведения об авторах:

Гегерь Эмилия Владимировна
доктор биологических наук, доцент, заведующая кабинетом статистики, Брянский клинико-диагностический центр
241050, г. Брянск, Россия, ул. Бежицкая, 2
E-Mail*:* emiliya_geger@mail.ru

Козлова Ирина Романовна
преподаватель, кафедра "Информационные технологии", Брянский государственный инженерно-технологический университет
241037, Брянск, Россия,
проспект Станке Димитрова, 3
E-Mail*:* kozlowa.iri2014@yandex.ru

Information about authors:

Geger Emiliya Vladimirovna
Doctor of biological sciences, Associate Professor, Head of Department of Statistic, Bryansk clinical diagnostic center
241050, Bryansk, Beziskaya st., 2
E-mail: emiliya_geger@mail.ru

Kozlova Irina Romanovna
lecturer, Department "Information technologies", Bryansk State Engineering and Technological University
241037, Bryansk, Russia, Stanke Dimitrova prospect, 3,
E-mail: kozlowa.iri2014@yandex.ru