УДК 004.02:004.6

СОВЕРШЕНСТВОВАНИЕ МЕТОДОВ ОБРАБОТКИ ДАННЫХ В ИНФОРМАЦИОННЫХ СИСТЕМАХ ПОДДЕРЖКИ ПРИНЯТИЯ УПРАВЛЕНЧЕСКИХ РЕШЕНИЙ

[1]Гегерь Э.В., [2]Евельсон Л.И., [2]Федоренко С.И., [2]Козлова И.Р.

[1]ГАУЗ «Брянский клинико-диагностический центр», Брянск, e-mail: emiliya_geger@mail.ru;
[2]ФГБОУ ВО «Брянский государственный технический университет», Брянск

В статье рассматривается применение метода сравнения бинарных выборок к анализу данных в медицинских информационных системах. Показано, что предлагаемый метод позволяет выявлять статистическую взаимосвязь между вредными условиями труда и заболеваемостью, устанавливать заболевания, характерные для определенного комплекса вредных производственных факторов. Это дает возможность врачу получить новые знания, лучше определять тактику диагностики и лечения. Для медицинского менеджмента становится возможным на более высоком уровне планировать профилактические мероприятия. Описывается методология статистического исследования, направленного на выявление характерных заболеваний, присущих действию вредных производственных факторов: электрического и магнитного полей промышленной частоты, с целью принятия управленческих и врачебных решений. Обосновывается необходимость внедрения новых методов анализа медицинских данных, которые позволяют лучше понять закономерности и определить значимые различия в показателях заболеваемости между лицами, подвергающимися воздействию опасных производственных факторов, и лицами, в работе которых отсутствуют вредные производственные источники. Предлагаемые методы направлены на анализ данных, накопленных в медицинских информационных системах транзакционного типа. Сделаны выводы о целесообразности дальнейших исследований и анализа показателей здоровья в группах лиц, имеющих вредные производственные факторы и у лиц, в профессиональной деятельности которых согласно специальной оценке труда отсутствуют опасные источники.

Ключевые слова: медицинские информационные системы, оценка риска, анализ данных, бинарные выборки, вредные производственные факторы

IMPROVEMENT OF DATA PROCESSING METHODS IN INFORMATION SYSTEMS FOR SUPPORT OF ADMINISTRATION OF DECISION-MAKING DECISIONS

[1]Geger E.V., [2]Evelson L.I., [2]Fedorenko S.I., [2]Kozlova I.R.

[1]Bryansk Clinical and Diagnostic Center, Bryansk, e-mail: emiliya_geger@mail.ru;
[2]Bryansk State Technical University, Bryansk

The article considers the application of the binary sample comparison method to data analysis in medical information systems. It is shown that the proposed method allows to identify the statistical relationship between harmful working conditions and morbidity, to establish diseases common to a certain complex of harmful production factors. This allows the doctor to gain new knowledge, better determine the tactics of diagnosis and treatment. It becomes possible for medical management to plan preventive measures at a higher level. It is described the methodology of statistical research aimed at identifying the characteristic diseases inherent in the action of harmful production factors: electric and magnetic fields of industrial frequency, in order to make managerial and medical decisions. The article demonstrates the necessity of implementation of new methods for the analysis of medical data to better understand patterns and to identify significant differences in incidence of disease between persons exposed to hazardous industrial factors and those in which there are no harmful production sources. The proposed methods are aimed at analyzing the data accumulated in medical information systems of transactional type. Conclusions about expediency of further researches and the analysis of indicators of health in groups of the persons having harmful production factors and at persons in which professional activity according to a special assessment of work there are no dangerous sources are made.

Keywords: medical information systems, risk assessment, data analysis, binary data selections, harmful industrial factors

В программе «Цифровая экономика Российской Федерации», утвержденной Распоряжением Правительства РФ 28 июля 2017 г., которая ориентируется на Стратегию развития информационного общества в Российской Федерации на 2017–2030 гг., сфера здравоохранения выделена в приоритетное направление. Применение информационных технологий, используемых при организации и оказании медицинской помощи, является одним из векторов развития цифрового здравоохранения. Современные информационно-аналитические методы предоставляют в распоряжение лечащего врача разнообразные методики для принятия врачебных решений [1–3].

Управление профессиональными рисками представляет собой комплекс организационно-технических мероприятий, который должен основываться на достоверных результатах периодических медицинских осмотров работников, занятых во вредных условиях труда, появляется необходимость использовать методы статистического ана-

лиза и аналитический инструмент, адекватные поставленной задаче [4, 5].

Процесс оценки производственного риска должен базироваться на результатах объективного анализа медицинских данных, накопленных в медицинских информационных OLTP (On-Line Transaction Processing) – системах, поэтому технологиям работы с данными должно уделяться особое внимание [6].

Для выявления причинно-следственных связей развития профессиональных заболеваний необходима разработка новых методических подходов к оценке уровня показателей здоровья работающего населения в динамике трудовой деятельности, обусловленных комплексом вредных факторов на рабочем месте, что позволит совершенствовать системы поддержки принятия решений.

Вышесказанное определило актуальность темы исследования.

Цель исследования: оценка рисков профессиональной заболеваемости с применением метода сравнения бинарных выборок на основе анализа данных медицинской информационной системы транзакционного типа.

Материалы и методы исследования

Как правило, данные в OLTP-системах собираются без их связи с дальнейшим анализом. Поэтому первоначально для достижения поставленных в исследовании целей необходимо решить задачи консолидации, очистки и предобработки данных, собранных в медицинской информационной OLTP-системе. Следующим этапом для анализа данных в данном исследовании являлось создание методики корректировки неоднородных выборок, являющейся необходимым условием достоверной обработки данных для принятия управленческих и врачебных решений.

Далее опишем группы обследуемых лиц и обоснуем формирование выборок для дальнейшего анализа.

Для решения поставленных задач исследования – выявления заболеваний, характерных для лиц, занимающихся профессиональной деятельностью, связанной с действием электрического и магнитного полей и излучений промышленной частоты на основе статистических и аналитических подходов, – нами были исследованы лица (112 чел.), постоянно работающие во вредных условиях труда (ЭМИ), и лица, которые по результатам специальной оценки труда не испытывают в своей производственной деятельности вышеуказанные вредные воздействия. В качестве таковых была взята группа лиц, являющихся офисными работниками (251 чел.).

На первом этапе исследования эта выборка принималась в качестве контрольной группы (КГ). Показатели здоровья работающих во вредных условиях труда анализировались на основании данных периодических медицинских осмотров. Ранее в статьях [4, 5] нами было описано сопоставление тех же групп по показателям лабораторных исследований, сделанное на основе статистического метода бинарных выборок. Этот метод предусматривал бинаризацию результатов лабораторных анализов по признаку соответствия принятой норме, принимающих только два возможных значения – «да» или «нет», т.е. «соответствует» или «не соответствует». В настоящей статье нами ставилась цель – на основе данных медосмотров по вышеуказанным двум группам и с помощью того же метода бинарных выборок изучить влияние электромагнитных полей промышленной частоты и излучений на заболеваемость. В данном случае бинаризация проводилась по признакам наличия/отсутствия соответствующих диагнозов, определяемых врачами в ходе медицинского обследования. Исследования проводились в Брянском клинико-диагностическом центре, результаты отражались в медицинской информационной системе. Перечень профессиональных заболеваний определялся согласно Приказу МЗСР РФ № 417н «Об утверждении перечня профессиональных заболеваний» и в соответствии с Приказом МЗСР РФ № 302 н «Об утверждении перечней вредных и (или) опасных производственных факторов и работ… и Порядка проведения обязательных предварительных и периодических медосмотров (обследований)» [7, 8]. Эксперимент проводился в соответствии с ФЗ РФ № 152 «О персональных данных» [9].

Математическая модель, описывающая собранные данные, строилась на статистической оценке значимости разницы между заболеваемостью в группе с наличием вредных производственных факторов и в группе с отсутствием таких факторов. Использовался подход, основанный на анализе бинарных выборок. В соответствии с этим подходом, в данной задаче сначала для обеих групп было определено множество D имевших место диагнозов, т.е. была сформирована совместная совокупность, в которую вошли все диагнозы, поставленные хотя бы одному из лиц, входящих в эти группы. Далее по каждому из диагнозов определялась их частота по данной группе.

Если в предыдущих работах специально проводилась бинаризация с использованием статистической нормы по лабораторным показателям, правомерность которой нуждается в дополнительном исследовании, то в настоящей работе метод бинарных выборок использован по «прямому» назначению, так как наличие/отсутствие у конкретного пациента рассматриваемого диагноза, естественно, описывается бинарной величиной.

Проверялась гипотеза об однородности двух выборок бинарных (двоичных) данных. Такая выборка характеризуется объемом n и частотой $p^* = m/n$ (где m – число людей, которым поставлен рассматриваемый диагноз), по которой оценивается вероятность p.

«В статистической модели предполагалось, что m является биномиальной случайной величиной $B(n, p)$, т.е. случайной величиной с параметрами n – объем выборки и p – вероятность наличия соответствующего диагноза. Такая случайная величина может быть представлена в виде

$$m = X_1 + X_2 + ... + X_n, \qquad (1)$$

где 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].

Сначала была проведена консолидация данных на основе медицинской информационной системы. Осуществлялась выгрузка обезличенных данных в формате MS Excel. Далее выполнялась очистка и предобработка данных. В ходе очистки выявлялись и устранялись дефекты и шумы, 4 записи в группе ЭМИ были исключены из дальнейшего анализа вследствие неустранимых дефектов данных.

Объем первой исходной выборки уменьшился с первоначальной 112 человек до 108 человек. Далее следовал этап анализа, который был выполнен после предобработки. Имеются две выборки по каждому рассматриваемому диагнозу. В соответствии с первоначальным планом исследования первая выборка относится к лицам, профессиональная деятельность которых связана с вредными производственными факторами ЭМИ, а вторая считается контрольной.

Объемы бинарных выборок обозначим как n_1 и n_2. Пусть в первой выборке число лиц с рассматриваемым диагнозом равно m_1, а во второй – m_2. В рамках вероятностной модели предположим, что m_1 и m_2 – биномиальные случайные величины, соответствующая вероятность значения «1», т.е. наличия этого диагноза для конкретного лица, входящего в первую выборку, равна p_1, а во вторую выборку – p_2. Проверяем нулевую гипотезу об однородности выборок: H_0: $p_1 = p_2$ при альтернативной гипотезе H_0: $p_1 \neq p_2$.

«В основу построения критерия проверки нулевой гипотезы положим статистическую функцию от выборочных частот p_i^* и объемов выборок n_i, обозначим ее q (p_1, p_2, n_1, n_2); обозначим Q предельное значение функции q, которое соответствует выходу за границы доверительного интервала разности вероятностей p_1, p_2. Критерий однородности выборок (проверки нулевой гипотезы) имеет вид

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

где знаком «*» обозначены выборочные частоты, являющиеся оценками соответствующих вероятностей: $p_i^* = m_i/n_i$). Модуль величины Q следует сравнивать с граничным значением критерия проверки однородности, которое, как показано в [10], можно определить на основании условий наследования сходимости из соотношения

$$K_{(\alpha)} = \frac{\Phi^{-1}(1+\alpha)}{2}, \qquad (3)$$

где Ф – функция стандартного нормального распределения, «–1» означает, что речь идет об обратной функции, α – уровень статистической значимости. Для уровня значимости α = 0,05, имеем К = 1,96 (таблицы значений функций Ф и Ф-1 [10]). Если Q по модулю меньше 1,96, то разница между выборками признается статистически незначимой и принимается нулевая гипотеза об однородности выборок. Если Q по модулю больше 1,96, то принимается альтернативная гипотеза о неоднородности. Необходимо обратить внимание на знак Q, по нему можно судить о том, какая из сравниваемых частот (вероятностей) выше.

Результаты исследования и их обсуждение

Для анализа заболеваемости была использована «Международная статистическая классификация болезней и проблем, связанных со здоровьем» десятого пересмотра (МКБ-10) [11].

В табл. 1 частично представлены результаты обработки диагнозов по методике бинарных выборок. В ней собраны все

диагнозы, по которым разница между двумя исследовавшимися выборками оказалась статистически значима (т.е. Q по модулю больше 1,96). Если $Q < 0$, то больше частота в группе КГ, а если $Q > 0$, то, соответственно, в группе ЭМИ. Также в таблице присутствуют для примера некоторые диагнозы, разница по которым оказалась незначимой. Диагноз «Здоров» позволяет сравнить уровень общей заболеваемости в обеих группах. По диагнозу «Здоров» получили значение $Q = 3{,}61$ (Q положительное, т.е. частота этого диагноза в группе ЭМИ значимо превышает частоту в группе КГ).

Таким образом, заболеваемость в группе КГ оказалась значимо выше, чем в группе ЭМИ.

Таблица 1

Диагнозы, встречающиеся в группах ЭМИ и КГ

№ п/п	Диагноз по МКБ	Критерий однородности Q
1	Здоров	3,86
2	G90.9	2,02
3	E78.0	7,35
4	H52.4	3,10
5	H35.4	–2,01
6	I11.9	–0,60
7	H25.0	–3,92
8	H27.8	–2,25
9	M51.1	–2,01
10	M19.0	–2,25
11	E06.3	–3,23
12	E80.4	–2,85
13	E04.1	–3,55
14	E04.2	–4,66
15	E01.0	–2,87
16	R94.3	–2,25

Электромагнитные излучения вряд ли положительно влияют на здоровье человека, логично предположить, что для КГ характерно влияние определенной неблагоприятной производственной среды, не учитываемой при аттестации рабочих мест, проводимой при специальной оценке условий труда (повышенная зрительная нагрузка, напряжение мышц спины и т.д.).

Сравнение бинарных выборок по другим диагнозам позволяет выделить те диагнозы, которые имеют значимо более высокую частоту в группе ЭМИ. Как видно из табл. 1, для рассматриваемых групп оказались характерны следующие диагнозы:

Для группы ЭМИ: E78.0 – Чистая гиперхолестеринемия; G90.9 – Расстройство вегетативной [автономной] нервной системы неуточненное; H52.4– Пресбиопия.

Для группы КГ: H35.4 – Периферические ретинальные дегенерации; H25.0 – Начальная старческая катаракта; H27.8 – Другие уточненные болезни хрусталика; M51.1 Поражения межпозвоночных дисков поясничного и других отделов с радикулопатией; M19.0 – Первичный артроз других суставов; E01.0 – Диффузный (эндемический) зоб, связанный с йодной недостаточностью; E06.3 – Аутоиммунный тиреоидит; E80.4 – Синдром Жильберта; E04.1 (E04.2) – Нетоксический одноузловой (многоузловой) зоб; R94.3 – Отклонения от нормы, выявленные при проведении функциональных исследований сердечно-сосудистой системы.

Были проверены гипотезы об однородности рассматриваемых выборок по признаку пола, который важен для выявления заболеваемости. Анализ данных показал: в группе ЭМИ большинство составляют мужчины, в то время как в группе КГ, несмотря на то, что мужчин больше, чем женщин, эта разница существенно меньше. При объеме выборки ЭМИ 108 человек, 106 из них составляют мужчины. В выборке КГ (251 человек), число мужчин – 223.

Поскольку признак «пол» с достаточной для целей настоящего исследования точностью можно считать бинарным, то можно применить описанный выше метод бинарных выборок по той же схеме и к этому признаку. Проведенный расчет показал значение $Q = 3{,}92$. Поскольку модуль Q больше критического значения 1,96, можно сделать вывод о неоднородности рассматриваемых выборок по признаку пола. В связи с этим далее был сделан расчет, в котором сопоставлялись две выборки, полученные исключением из исходных выборок всех записей, относящихся к женщинам. Логическим обоснованием такого приема может служить то, что число женщин в группе ЭМИ – 2, при этом общий объем обеих выборок и после исключения «женских» записей остается достаточно большим для того, чтобы можно было корректно проверить нулевую гипотезу однородности, используя метод бинарных выборок. Таким образом, описанный выше расчет был полностью повторен для скорректированных выборок, в которых были оставлены только «мужские» записи. В табл. 2 представлены полученные результаты по диагнозам.

Структура табл. 2 полностью аналогична структуре табл. 1.

Таблица 2

Диагнозы, встречающиеся в группах ЭМИ и КГ после корректировки выборок по признаку «пол»

№ п/п	Диагноз по МКБ	Критерий однородности Q
1	Здоров	3,73
2	G90.9	1,57
3	E78.0	7,29
4	H52.4	2,97
5	H35.4	–2,01
6	I11.9	–0,65
7	H25.0	–3,40
8	H27.8	–2,25
9	M51.1	–2,02
10	M19.0	–2,26
11	E06.3	–3,23
12	E80.4	–2,85
13	E04.1	–3,56
14	E04.2	–4,42
15	E01.0	–2,88
16	R94.3	–2,25

Сопоставление значений табл. 1 и 2 показывает, что корректировка выборок по признаку «пол» по большинству диагнозов не привела к изменению вывода, знаки неравенства $Q > 1{,}96$ или $Q > 1{,}96$ остались прежними. Однако по диагнозу G90.9 «Расстройство вегетативной [автономной] нервной системы неуточнённое» результат изменился: вместо вывода о значимо большей заболеваемости по этому диагнозу в группе ЭМИ сделан вывод о незначимой разнице между ЭМИ и КГ по этому диагнозу.

Выводы

1. Предлагаемый метод, основанный на анализе бинарных выборок, может продуктивно использоваться как составная часть информационной системы оценки риска профессиональных заболеваний. Применение данного метода показано на реальных данных, сбор, консолидация и обработка которых осуществлялись с помощью медицинской автоматизированной информационной OLTP-системы. Он позволит определять заболевания, которые присущи различным видам профессиональной занятости с последующим анализом скрытых закономерностей.

2. Группа работников КГ, которую первоначально планировалось использовать в качестве контрольной группы условно здоровых людей, не может быть выбрана в качестве таковой, поскольку заболеваемость в этой группе оказалась значимо выше, чем в группе ЭМИ. Целесообразно проведение дополнительного исследования, направленного на уточнение вредных факторов, присущих производственной деятельности работников группы КГ, а также их влиянию на заболеваемость.

3. Группы ЭМИ и КГ оказались неоднородными по признаку «пол». Было сделано дополнительное исследование, направленное на изучение влияния признака «пол» на результаты исследования, позволившего сделать вывод о целесообразности структурирования информации, приведению данных к однородности, что даст возможность сделать эти данные более пригодными для цифровой трансформации.

4. Необходимо провести дальнейшее исследование закономерностей влияния неоднородности сравниваемых выборок медицинских данных по характерным признакам на результаты сравнения, включив предлагаемый метод в аналитический модуль для медицинских информационных систем, определить эффективность его использования, что будет способствовать повышению качества и уровня управленческих решений.

Заключение

Таким образом, метод сравнения бинарных выборок может быть успешно применен для анализа медицинских данных, ранее накопленных в медицинских информационных системах транзакционного типа, и позволит выявлять заболевания, характерные для определенных комплексов вредных факторов, связанных с производственной деятельностью. Это будет способствовать совершенствованию диагностики и лечения с применением цифровых технологий, поможет принятию правильных управленческих решений в сфере медицинской профилактики профзаболеваний.

Список литературы

1. Катаев В.А., Зарипова Г.Р., Богданова Ю.А. Модели СППР в хирургической практике. Современные подходы к решению проблемы // Медицина. 2016. Т. 4. № 4 (16). С. 68–74.

2. Миронов П.И., Медведев О.И., Ишмухаметов И.Х., Булатов Р.Д. Прогнозирование течения и исходов тяжелого острого панкреатита // Фундаментальные исследования. 2011. № 10–2. С. 319–323.

3. Наркевич А.Н., Виноградов К.А., Катаева А.В., Пичугина Ю.А., Афанасьева Н.А. Средства интеллектуальной поддержки принятия решений в диагностике и лечении наркозависимых // Врач и информационные технологии. 2018. № 4. С. 20–26.

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

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

6. 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.

7. Приказ МЗСР РФ № 417н от 27.04.2012 г. «Об утверждении перечня профессиональных заболеваний». [Электронный ресурс]. URL: http://base.garant.ru/70177874/ (дата обращения: 23.11.2019).

8. Приказ МЗСР РФ от 12.04.2011 № 302 н (ред. 06.02.2018) «Об утверждении перечней вредных и (или) опасных производственных факторов и работ, при выполнении которых проводятся обязательные предварительные и периодические медосмотры (обследования), и Порядка проведения обязательных предварительных и периодических медосмотров (обследований)». [Электронный ресурс]. URL: http://base.garant.ru/12191202/ (дата обращения: 23.11.2019).

9. Федеральный закон от 27.07.2006 № 152-ФЗ (ред. 29.07.2017) «О персональных данных». [Электронный ресурс]. URL: http://base.garant.ru/5635295/ (дата обращения: 30.11.2019).

10. Орлов А.И. Прикладная статистика. М.: Изд-во «Экзамен», 2006. 671 с.

11. Международная классификация болезней 10-го пересмотра (МКБ-10) [Электронный ресурс]. URL: https://mkb-10.com (дата обращения: 20.11.2019).