УДК 614.1: 311.3(075.32)
DOI: 10.35693/2500-1388-2023-8-3-189-197

Технологии расчета и визуализации статистики распространенности и заболеваемости на примере сведений о полипозном риносинусите в Самарской области

© С.А. Палевская, А.В. Гущин, М.К. Блашенцев

ФГБОУ ВО «Самарский государственный медицинский университет» Минздрава России (Самара, Россия)

Аннотация

Цель – определить в статистическом смысле особенности распределения хронических заболеваний в хронологии наблюдений; показать специфику приемов проверок гипотез в количественном и вероятностном прогнозе распространяемости полипозного риносинусита.

Материалы и методы. Использованы данные случаев амбулаторно-поликлинической помощи за 2017–2021 годы и количественные сведения о зарегистрированных основных или сопутствующих фактах заболеваний полипозным риносинуситом в медицинских организациях в 25 районах Самарской области.

Результаты. Синтез статистик исходных данных, которые составили объем числового расширения первичных показателей в следующем соотношении: категории 15,8%, счетные данные 26,3%; количественные значения 21,1%; 26,7% – это относительные данные заболеваемости и распространенности. Оставшийся объем – это описательная статистика и показатели в виде таблиц коэффициентов корреляций. Для синтезированных расширений данных произведены оценки распределений и проверки гипотез с использованием статистических критериев.

Выводы. Счет числа заболеваний хронического характера аппроксимируются плотностями атипичных распределений. Приблизительно 58% выборок по диагнозам не подтверждаются, подчиняясь закону распределения. В подобной ситуации при подготовке прогноза для перехода к временному ряду необходимо решать проблему получения последовательностей со стационарными характеристиками. При машинном обучении данные в расчетах предсказания должны пройти проверку на вероятностное подтверждение совпадения распределений параметров связанных выборок. Результаты прогноза следует воспринимать как вероятностный вывод на уровне неотвергаемой гипотезы.

Ключевые слова: описательная статистика, корреляции количественных признаков, статистические тесты, проверка статистических гипотез, регрессия, прогноз временных рядов, p-values.

Конфликт интересов: не заявлен.

Для цитирования:
Палевская С.А., Гущин А.В., Блашенцев М.К. **Технологии расчета и визуализации статистики распространенности и заболеваемости на примере сведений о полипозном риносинусите в Самарской области.** *Наука и инновации в медицине*. 2023;8(3):189-197.
doi: 10.35693/2500-1388-2023-8-3-189-197

Сведения об авторах
Палевская С.А. – д-р мед. наук, MBA, профессор MBA, директор ИПО.
ORCID: 0000-0001-9263-9407 E-mail: s.a.palevskaya@samsmu.ru
Гущин А.В. – канд. техн. наук, доцент кафедры менеджмента ИПО.
ORCID: 0000-0002-6128-2334
E-mail: a.v.guschin@samsmu.ru
Блашенцев М.К. – аспирант кафедры оториноларингологии им. академика РАН И.Б. Солдатова, ассистент ФАЦ.
ORCID: 0000-0002-9820-4292 E-mail: mblashentsev@gmail.com

Автор для переписки
Гущин Андрей Викторович
Адрес: Самарский государственный медицинский университет, ул. Гагарина, 18, г. Самара, Россия, 443001.
E-mail: a.v.guschin@samsmu.ru

Рукопись получена: 15.04.2023
Рецензия получена: 08.06.2023
Решение о публикации принято: 11.06.2023

Technologies for calculating and visualizing statistics on prevalence and incidence on the example of information about polypous rhinosinusitis in the Samara region

© Svetlana A. Palevskaya, Andrei V. Gushchin, Mikhail K. Blashentsev

Samara State Medical University (Samara, Russia)

Abstract

Aim – to statistically determine the distribution of chronic diseases in the chronology of observations; to show the specifics of methods for testing hypotheses in the quantitative and probabilistic prediction of the prevalence of polypous rhinosinusitis.

Material and methods. The outpatient data for the period of 2017–2021 and quantitative information about the cases with polypous rhinosinusitis as main or concomitant diagnosis registered by medical organizations of 25 districts of the Samara region were used in the study.

Results. The synthesis of the initial data statistics, which amounted to the volume of the numerical expansion of primary indicators in the following ratio: categories 15.8%, counting data 26.3%; quantitative values 21.1%; 26.7% – relative incidence and prevalence data. The rest of the data is the descriptive statistics and indicators in the form of tables of correlation coefficients. For extensions of the synthesized data, distributions were evaluated and hypotheses tested using statistical criteria.

Conclusion. The count of the number of chronic diseases is approximated by the density of atypical distributions. Approximately 58% of samples for diagnoses are not confirmed as obeying the law of distribution. In such a situation, when preparing a forecast for the transition to a time series, it is necessary to solve the problem of obtaining sequences with stationary characteristics. In machine learning, data in predictive calculations must be

checked for probabilistic confirmation of the coincidence of related sample parameter distributions. The results of the forecast should be taken as a probabilistic conclusion at the level of an unrejected hypothesis.

Keywords: descriptive statistics, quantitative correlations, statistical tests, statistical hypothesis testing, regression, time series forecasting, p-values.

Conflict of interest: nothing to disclose.

Citation

Palevskaya SA, Gushchin AV, Blashentsev MK. **Technologies for calculating and visualizing statistics on prevalence and incidence on the example of information about polypous rhinosinusitis in the Samara region.** *Science and Innovations in Medicine*. 2023;8(3):189-197. doi: 10.35693/2500-1388-2023-8-3-189-197

Information about authors

Svetlana A. Palevskaya – PhD, MBA, Director of the Institute of Postgraduate Education.
ORCID: 0000-0001-9263-9407 E-mail: s.a.palevskaya@samsmu.ru

Andrei V. Gushchin – PhD, Associate professor of the Department of Management, Institute of the Postgraduate Education.
ORCID: 0000-0002-6128-2334
E-mail: a.v.guschin@samsmu.ru

Mikhail K. Blashentsev – a postgraduate student of the Department of Otorhinolaryngology n.a. Academician of the Russian Academy of Sciences I.B. Soldatov, assistant of the FAC.
ORCID: 0000-0002-9820-4292 E-mail: mblashentsev@gmail.com

Corresponding Author

Andrei V. Gushchin
Address: Samara State Medical University, 18 Gagarina st., Samara, Russia, 443001.
E-mail: a.v.guschin@samsmu.ru

Received: 15.04.2023
Revision Received: 08.06.2023
Accepted: 11.06.2023

ВВЕДЕНИЕ

Развитие современного программного обеспечения для изучения и обобщения свойств статистик регрессионных зависимостей [1–3] открывает возможности обработки большого объема комбинаторно сочетающихся признаков количественных и категориальных данных. При этом возникает проблема смысловой нагрузки на эти сочетания, чтобы достижение целей анализа и обработки данных не было нарушено противоречивым набором результатов машинных действий. В статье показана расчетная траектория, выводящая на итоговые групповые операции подготовки временных рядов к прогнозу на основании сочетания категорий, счетчиков и вещественных данных. Обращается внимание на переход от категорий и счета к временному характеру соотношения данных при подготовке прогноза.

Привлечение программных средств подготовки данных не снижает актуальности экспертного вывода о последовательности действий в подготовке статистического анализа и визуализации результатов. Особенно это значимо при применении современных алгоритмических средств на основе машинного обучения. Важен не только результат обучения и прогноза, но обобщение действий как эмпирического опыта рассматриваемой темы.

№	Признаки (наименование поля)	Тип признака	Примечание
1	Отчетный год	Категория (число)	
2	Код района	Категория (число)	
3	Наименование района	Категория (строка)	
4	Население (в районе)	Счетчик	Счетчик – счетные целые данные
5	Диагноз	Категория (строка)	
6	Зарегистрировано заболеваний	Счетчик	
7	Мужчины	Счетчик	Итог равен полю №6
8	Женщины		
9	Дети	Счетчик	Итог равен полю №6
10	Подростки		
11	Взрослые		
12	Старше трудоспособного возраста (муж. > 60, жен. > 55)	Счетчик	Подмножество поля №11
13	Заболеваемость первичная на 100 000 населения	Количество	
14	Распространенность на 100 000 населения	Количество	

Таблица 1. *Типы данных полей регистрации подсчета заболеваний*

Table 1. *Data types and registration fields for disease count*

В основной части статьи демонстрируются приемы агрегирования и кроссирования исходных измерений. Применяется современный p-value метод оценки допуска ошибки отклонения по вероятностной мере, функционально связанной с критерием статистического теста [4, 5]. На всех этапах вероятностных выводов используются приемы сбалансированного визуального и численно-графического представления результата [6]. Этот важный прием технологии визуализации и расчета статистик демонстрируется для условий фактически полной неопределенности в общих категориях диагнозов «Другие полипы синуса» и «Полип носа неуточненный». Здесь тривиальный численный ряд геометрического распределения объективно дополнен визуальным подтверждением атипичности самого явления. Показана технология реформирования распределения редких событий в закон Пуассона, а также особенности выбора количественных данных с оценкой средних для осуществления целей прогноза. Все сделанные предположения о новом качестве синтезируемой информации подтверждаются критериями оценок распределений.

ЦЕЛЬ

Определить в статистическом смысле особенности распределения хронических заболеваний на основе сведений о полипозном риносинусите, показать основные технологические приемы выбора критериев проверки гипотез, подбора модели регрессии и осуществления прогноза для определенного типа заболеваний.

МАТЕРИАЛ И МЕТОДЫ ИССЛЕДОВАНИЯ

Для числового расширения представления категорий и признаков применялись методы кроссирования и агрегирования исходных данных. В качестве базовой информации использовались счетные и количественные данные регистрации первичной заболеваемости и распространения заболеваний в относительных показателях на 100 000 населения. Используемые в расчетах ключевые поля приведены к понятию «Признаки» и описаны **в таблице 1**, где сумма пунктов 7 и 8, пунктов 9, 10 и 11 равны пункту 6; п. 12 – группа старшего трудоспособного возраста численно входит в состав п. 11. Строки «объекты» составляют районы Самарской области с исключенными городами. Значения счетчиков и расчет количественных признаков локализуются в исходных данных по периодам одного

№	Выполненный запрос к исходным данным	Диаграмма
1	Число районов, где было диагностировано заболевание по годам	Есть
2	Распределение числа типов диагнозов по районам за период 2017–2021 гг.	Есть
3	Корреляции числа типов диагнозов по районам за 2017–2021 гг.	Есть
4	Общее распределение числа заболеваний по районам за период 2017–2021 гг.	Нет
5	Корреляции диагнозов по числу заболеваний по районам за 2017–2021 гг.	Есть
6	Описательная статистика числа диагнозов по районам за 2017–2021 гг.	Нет
7	Корреляции числа заболеваний по районам с разделением по полу	Нет
8	Описательная статистика числа заболеваний по районам с разделением по полу	Нет
9	Корреляции диагнозов общего числа заболеваний по годам за 2017–2021 гг.	Есть
10	Число заболеваний по годам с разделением по полу	Есть
11	Описательная статистика числа заболеваний по годам с разделением по полу	Нет
12	Описательная статистика первичной заболеваемости и распространенности по районам за период 2017–2021 гг.	Нет
13	Корреляция первичной заболеваемости и распространенности по районам	Нет
14	Описательная статистика первичной заболеваемости и распространенности диагноза J33.0 по районам за период 2017–2021 гг.	Нет

Таблица 2. *Расширение исходной информации отношениями по результатам запросов к исходным данным (см. таблицу 1)*

Table 2. *Extension of initial information by relations based on the results of queries to the source data (see table 1)*

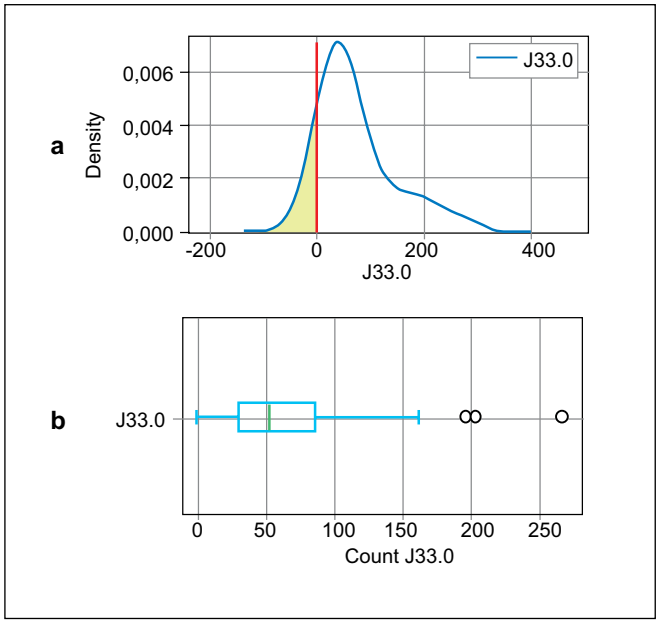

Рисунок 1. *Оценка плотности распределений заболеваний по диагнозу J33.0 по районам за период 2017–2021 гг.: a – ядерная оценка плотности положительных значений; b – межквартильный размах (коробка, 50% реализаций случайной величины с наибольшей плотностью), медиана (вертикальная линия, разделяющая коробку), разброс (ограниченные отрезки справа и слева от коробки) и отдельные выбросы (точки за линией отрезков), которые можно воспринимать как атипичные значения для данного характера распределения.*

Figure 1. *Estimation of the density of distribution of cases with J33.0 diagnosis by districts for the period 2017–2021: a – a kernel estimate of the density of positive values; b – interquartile range (box, 50% of the realizations of the random variable with the highest density), median (vertical line separating the box), spread (limited segments to the right and left of the box) and individual outliers (points behind the line segments), which can be perceived as atypical values for a given distribution.*

года. Периоды представляют временной диапазон наблюдений 2017–2021 гг.

Типы четырех диагнозов (пункт 5 таблицы 1) зашифрованы по МКБ-10 как J33.0, J33.1, J33.8 и J33.9, где J33.0 – полип полости носа, J33.1 – полипозная дегенерация синуса, J33.8 – другие полипы синуса, J33.9 – полип носа неуточненный. Категориальные признаки, счетчики и диагностированные заболевания составляют основную группу категориальных признаков (все пункты таблицы 1, кроме пунктов 13, 14). Корреляции, таблицы кроссирования, агрегирования по категориям и счетчикам – соответствующие таблицы, диаграммы – используются в исследовании, но не приводятся в данной статье. Описание этих отношений сформулированы **в таблице 2**.

К счетным конечным выборкам можно применять статистики, оценивающие тип и характеристики вероятностных процессов. В этих вопросах исследований используются современные методы реконструкции гладких вероятностных функций. Одно из возможных решений аппроксимации распределения счетчиков – использовать метод ядерной оценки плотности – непараметрический способ оценки плотности случайной величины на конечных выборках. Фактически осуществляется попытка аппроксимировать гистограмму непрерывной функцией p(x) при помощи ядерной оценки плотности

$$p(x) = \lim_{h \to 0} \frac{1}{2h} P[x - h, x + h],$$

где $p[a,b]$ – вероятностная мера на отрезке $[a,b]$, x – значение признака, h – шаг аппроксимации.

Сглаживающее представление способно дать приближение любого распределения и определить дальнейший характер исследования (**рисунок 1a**). В исследованиях для ядерной оценки плотности предусмотрен специальный количественный способ визуального отображения информации (**рисунок 1b**). Это диаграмма в следующем сочетании: коробки разброса по квартилям в 50% (IQR) относительно медианы, линии с границами разброса в 1.5 длины IQR от 1-й и 3-й квартили. Соответственно в характеристиках разброса будет использоваться межквартильный диапазон или интерквартильный размах. Интерквартильный размах определяет половину выборки, которая центрирована относительно медианы (вертикаль внутри коробки). Показатели, как сглаженное распределение и размах, в данной ситуации есть показатели изменчивости признака для асимметричных распределений. Кроме того, все наборы данных подсчета заболевших обладают аномальными значениями отдельных выбросов. Размах IQR мы рассматриваем в качестве аналога дисперсии, который является робастным (слабо чувствительным) к выбросам в данных.

Аппроксимацию распределения подсчета диагнозов по районам следует воспринимать в изображении на положительном интервале числа наблюдений. В рассуждениях и выводах использование метрических свойств интерквартильного размаха должно пониматься как альтернатива среднеквадратичного отклонения (с.к.о), которое применимо только для нормально распределенных данных, что неприемлемо буквально для рассматриваемых случаев. Интерквартильный размах будет служить непараметрической метрической оценкой, не требующей предположений

относительно распределения. Применяется для любых типов данных, что актуально с имеющимся типом часто не идентифицируемых распределений с атипичными выбросами вне диапазона приемлемых значений вероятности.

В исследованиях применяются классические методы проверки статистических гипотез на тип распределения. Например, для редких наблюдаемых событий – распределение Пуассона с предположением вероятностного предсказания числа заболеваний на период по конкретному диагнозу. Для рассматриваемых случаев произведено специальное кроссирование данных из общей статистики распространенности первичной заболеваемости по районам за весь период наблюдений (**пункт 12 таблицы 2**).

Организация исследований строится на последовательном переходе от категориальных к количественным статистикам. При такой организации таблично-графическое представление результатов анализа позволяет правильно и безальтернативно в плане оптимальности подготовить выборки достижения главной целевой составляющей – регрессии и прогноза основных зависимостей.

РЕЗУЛЬТАТЫ

В основании технологии визуализации данных, специфика которых представлена **таблицей 1**, находятся отношения категорий «Код», «Диагноз» к количественному счетному признаку «Зарегистрировано заболеваний» (далее «Заболевшие»). На этом этапе расширения исходных статистик ситуация с формализацией случайного поведения наблюдений достаточно проблематична. Например, корреляционные оценки взаимодействия категорий и счетчиков (**пункт 5 таблицы 2**), как и корреляции предыдущих отношений, показывают независимость процессов счета диагнозов по районам. Описательная статистика полученных отношений (**пункт 6 таблицы 2**) показывает, что средние числа диагнозов сильно смещены вправо. Это превышает половину вероятностной меры (медианы), приближаясь к квантилю 0.75. Фактически среднее и чуть выше среднего – это уже примерно 2/3 от реализации меры вероятности заболеть (**рисунок 1**). Диагнозы, отличные от J33.0, еще менее характерны в определении распределения. Имеют максимальную вероятность заболевания менее чем для одного человека в районных масштабах. Группа из одного человека и более заболевают с вероятностью, распадающейся от 25% меры вероятности и меньше. Причем распад внешне напоминает экспоненциальный, что требует дополнительной проверки. Диапазоны с малозначащей вероятностью заполняют аномальные выбросы вероятности заболевания больших групп, выходящих за пределы квартильного размаха, за которым

i	x_i	n_i	Px_i	n'_i	n'_i-n_i	$(n'_i$-$n_i)^2$	χ^2
0	0	15	0,3263	8,1570	6,8430	46,8267	5,7407
1	1	4	0,3654	9,1358	-5,1358	26,3768	2,8872
2	2	2	0,2046	5,1161	-3,1161	9,7099	1,8979
3	3	2	0,0764	1,9100	0,0900	0,0081	0,0042
4	4	1	0,0214	0,5348	0,4652	0,2164	0,4047
5	10	1	0,0048	0,1198	0,8802	0,7748	6,4674
							17,402

Таблица 3. *Расчетные значения критерия Пирсона χ^2 для оценок распределения Пуассона заболеваемости по районам с диагнозом J33.1*

Table 3. *Estimated values of Pearson's χ^2 criterion for estimates of the Poisson distribution of morbidity by districts with a J33.1 diagnosis*

сложно дать оценку вероятности выбросов и их подчинения некоторому закону распределения (**рисунок 1b**). Вероятность числа заболеваний по среднему также сильно снижена – менее 2/3 оставшейся меры заболевших малыми группами. Метрически малая часть подчинена расчетам распределения – фактически 90% диапазона наблюдений – это аномальные выбросы заболевания больших групп на отрезках с минимальной, фактически не аппроксимируемой вероятностью.

Реализуем на данных, описанных выше, попытку проверки гипотез предполагаемых типов распределения. Рассмотрим два основных отношения данных диагнозов: счетные заболеваний (**пункт 4 таблицы 2**) и количественные заболеваемости и распространенности по районам (**пункт 12 таблицы 2**). Группировка диагнозов в общем числе заболевших предполагает гипотезу вероятности события определенного числа заболеваний. Для этого нужно вероятностное подчинение закону Пуассона. Визуально изображение ядерной аппроксимации (**рисунок 1a**) воспринимается как распределение Пуассона. Но проверки по критерию согласия Пирсона χ^2 отвергли данную нулевую гипотезу для всех диагнозов. Покажем это на примере J33.1. Требуется при уровне значимости α=0.05 проверить гипотезу о том, что случайная величина J33.1, обозначенная как X, не отклоняется от распределения по закону Пуассона. Счетчик частоты эмпирических значений n_i и число $N=\sum n_i$ заболевших, где i=1.5, определяют теоретическую частоту $n_i' = Px_i N$ заболевших при известном теоретическом распределении Px, единственный параметр $\bar{\lambda}$ которого оценивается при знании эмпирических частот и счетчиков заболеваемости $\bar{\lambda} = \frac{1}{N}\sum_i x_i n_i$. Составляем расчетную **таблицу 3**, где также отражаем составляющие выражения расчета наблюдаемое значение критерия

$$\chi^2 = \sum_i \frac{(n_i' - n_i)^2}{n_i'} .$$

При расчете степени свободы S учитываем свертку, которая производится по строкам, где $n_i \leq \max(x_i)$, тогда $S=r-1=1$, т.к. число строк после свертки определим как r=2. Соответственно табличное значение критерия χ_t^2 (0.05;1)=3.8 меньше расчетного χ^2=17,402. Гипотеза H_0 распределения Пуассона отвергается. Малое значение степеней свободы – основная причина опровержения гипотезы для всех распределений заболевших, а также заболеваемости и распространенности (**пункт 12 таблицы 2**) по районам за период 2017–2022 гг.

Табличное представление категорий «Код» и «Диагноз» также информирует о невозможности неотклонения гипотез о популярных распределениях, поскольку числовое поле таблицы фрагментировано нулевыми значениями, что говорит в пользу свертки числа степеней свободы (числа районов) с завышением требований к граничным значениям критериев. Естественно, что без средств программной автоматизации перебор поиска аппроксимации функции вероятности или характера плотности распределения физически невозможен. Поэтому актуальными для категорий и счетчиков в плане информативности остаются только табличные данные. Для дальнейшего движения к цели требуется перестройка и новое кроссирование выбранных категорий и счетчиков. Для этого задействуем хронологию событий, но, исходя

	Мужчины				Женщины			
	J33.0	J33.1	J33.8	J33.9	J33.0	J33.1	J33.8	J33.9
J33.0	1,0000	0,2913	0,9148	0,0674	1,0000	-0,8109	0,4314	-0,7200
J33.1	0,2913	1,0000	0,1750	0,8376	-0,8109	1,0000	-0,2287	0,5437
J33.8	0,9148	0,1750	1,0000	0,1392	0,4314	-0,2287	1,0000	-0,0716
J33.9	0,0674	0,8376	0,1392	1,0000	-0,7200	0,5437	-0,0716	1,0000

Таблица 4. *Корреляции числа заболеваний по годам с разделением по полу*
Table 4. *Correlations of the number of cases by years with a division by sex*

из предоставленных данных, это будет достаточно общее представление за годовой период. Тем не менее произведем визуализацию, которую дополним проверкой гипотез на тип распределения, на связи и идентичность распределений, используемых как входные данные прогноза.

Сформируем данные статистик типа распределения числа заболеваний по годам. Исследуем перегруппировку данных по категориям «Год», «Пол», «Диагноз» со счетчиком количества заболевших. Сначала агрегируем подсчет диагнозов по годам (без учета пола), придаем характер временного ряда. В хронологии сразу проявляются корреляционные свойства J33.1 и J33.9 (**пункт 9 таблицы 2**). Экспертам следует обратить на это внимание, так как положительная корреляция изменений данных этих диагнозов будет присутствовать в графических отображениях для этих отношений. Разделение заболеваний по полу в годовой хронологии представлено количественно (**пункт 10 таблицы 2**), и корреляционные зависимости приведены **в таблице 4**. Графические сравнения показаны **на рисунке 2**. Хронология и разделение по полу приводит к обнаружению линейных зависимостей диагнозов для числа заболеваний. Положительная корреляция J33.1 и J33.9 присутствует в мужской подгруппе, где фиксируется сильное взаимодействие J33.0 и J33.8 (**таблица 4**). В женской группе значимые зависимости отрицательны: J33.0 и J33.1, J33.0 и J33.9 (**таблица 4**). Минимальной метрической разницей между подгруппами обладает диагноз J33.0 с положительной и самой сильной корреляцией (**рисунок 2a**) к J33.8. Максимальное расстояние по заболеванию у J33.1 с отрицательной корреляцией (**рисунок 2b**) в женской подгруппе.

Переходим к анализу статистик, которые составляют основную цель построения технологий исследований – поиск зависимостей для определения регрессии и прогноза. Главный объект данных – общая статистика заболеваемости и распространенности по районам. Это вещественные и относительные количественные признаки. Основные распределения количественных признаков – это «Заболеваемость» (**пункт 13 таблицы 1**) и «Распространенность» (**пункт 14 таблицы 1**) по районам (по годам). Источники данных – отношения **пунктов 12–14 таблицы 2**. Общая оценка ядерной плотности показывает сильный распад вероятности при росте подсчета заболеваемости и распространенности по районам (**рисунок 3**). Имеем характер той же нетипичной плотности, полученный для категорий без хронологии (**рисунок 1**).

Путем программного перебора аппроксимаций рассчитаем оценку принадлежности к распределениям рассматриваемых признаков по районам за период 2017–2021 гг. [5, 6]. При этом используется наблюдаемый *p*-value уровень значимости (*p*-значение). Рассчитывается *p*-value вероятность отклонения наблюдаемого распределения от предполагаемого по тесту Колмогорова – Смирнова. Здесь *p*-value функционально зависит от критерия согласия Колмогорова. Мера *p*-value должна превышать или быть равной мере критической области уровня значимости α=0.05. В результате, как и для категорий-счетчиков, имеем дело с менее популярными типами распределений или гипотеза закона плотности не принимается вообще. Итоги расчетов обозначены **в таблице 5**.

Для придания свойств «нормальности» предпримем попытку агрегирования диагнозов по периодам наблюдения с осреднением относительных признаков «Заболеваемость» и «Распространенность». Результаты размещены **в таблице 6**.

Выбираем объект прогнозирования – J33.0. Экспериментально было определено, что диагноз J33.0 (**рисунок 1**) имеет минимальное значение всплесков вероятности за границей наблюдаемых значений. Сделаем проверку обобщенного по среднему J33.0 на нормальное (или близко к нормальному) распределение. Другими словами, проведем оценку распределения среднего

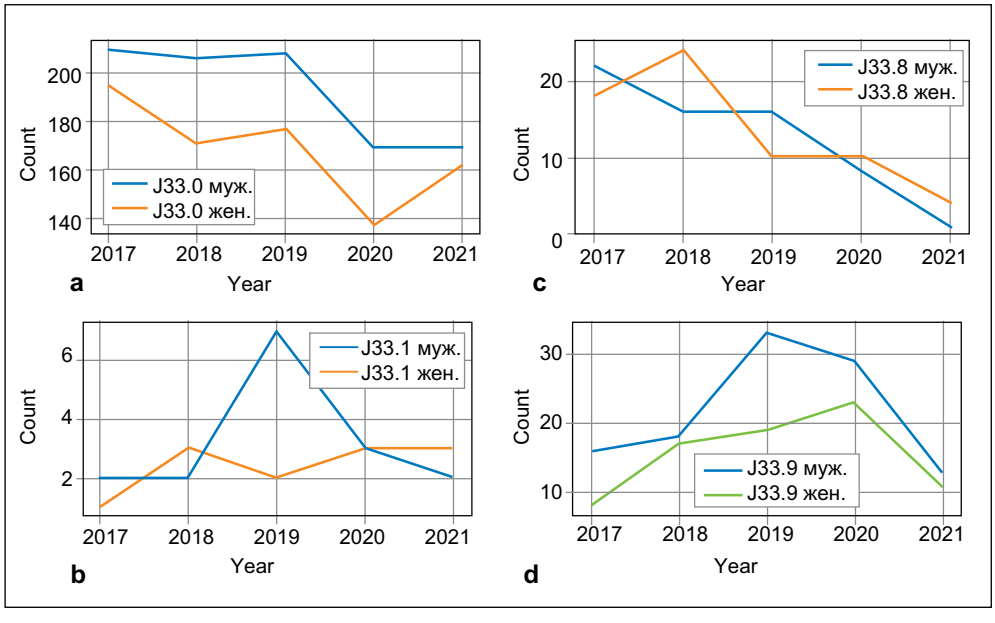

Рисунок 2. *Линейные графики парного сравнения количества заболеваний мужчин и женщин по годам, ‖·‖ – относительное расстояние между парами нормированных линий: a – диагноз J33.0, ‖·‖=0.00036; b – диагноз J33.1, ‖·‖=0.1118; c – диагноз J33.8, ‖·‖=0.0104; d – диагноз J33.9, ‖·‖=0.0090.*

Figure 2. *Line graphs of pairwise comparison of the number of diseases in men and women by years, ‖·‖ – relative distance between pairs of normalized lines: a – diagnosis J33.0, ‖·‖=0.00036; b – diagnosis J33.1, ‖·‖=0.1118; c – diagnosis J33.8, ‖·‖=0.0104; d – diagnosis J33.9, ‖·‖=0.0090.*

J33.0, полученного при агрегировании вещественных относительных признаков (**таблица 6, значения выделены фоном**). Получены положительные результаты проверки по критерию Пирсона (χ^2-квадрат, обозначено Statistics) на соответствие эмпирического распределения нормальному, при α=0.05, за период 2017–2021 гг.:

Заболеваемость: Statistics=2.625, p-value=0.269;

Распространенность: Statistics=1.736, p-value =0.420,

где по условию p-value≥α, гипотезы нормальности обоих распределений не отклоняются.

Остается проверить гипотезу сравнения выборочных средних зависимых выборок для J33.0 по годам по модифицированному критерию Стьюдента (парный студенческий Т-тест). Конкретно проверяем гипотезу H_0 по парному T-тесту с критерием T_d, что выборки J33.0 по годам из каждой совокупности «Заболеваемость» и «Распространенность» зависимы, то есть распределения из параметров тождественны. Если p-value наблюдаемого критерия T_d больше или равно уровню значимости α, то гипотеза схожести параметров распределения зависимых выборок не отвергается. Зависимости соседних периодов должны обеспечить значимость параметров регрессии и достоверность прогноза. Результаты проверок приведены **в таблице 7**.

После положительных проверок исходных данных прогноза агрегируемые в среднем по годам диагнозы заболеваемости и распространенности составят временной ряд (**таблица 6**). Для аппроксимации статической зависимости времени и количества требуется выбор и оценивание параметров модели регрессии по каждому из диагнозов на четырех временных отрезках [7–13]. Для построения прогноза на заданное число шагов требуется динамическая модель фазового движения по траектории прогноза. Но нехватка физических характеристик процесса приводит к неопределению типа уравнения движения. Тогда источник наблюдаемого движения принимаем за «черный ящик», формирующий временной ряд путем генерации последовательно во времени случайных величин. В этой ситуации воспользуемся

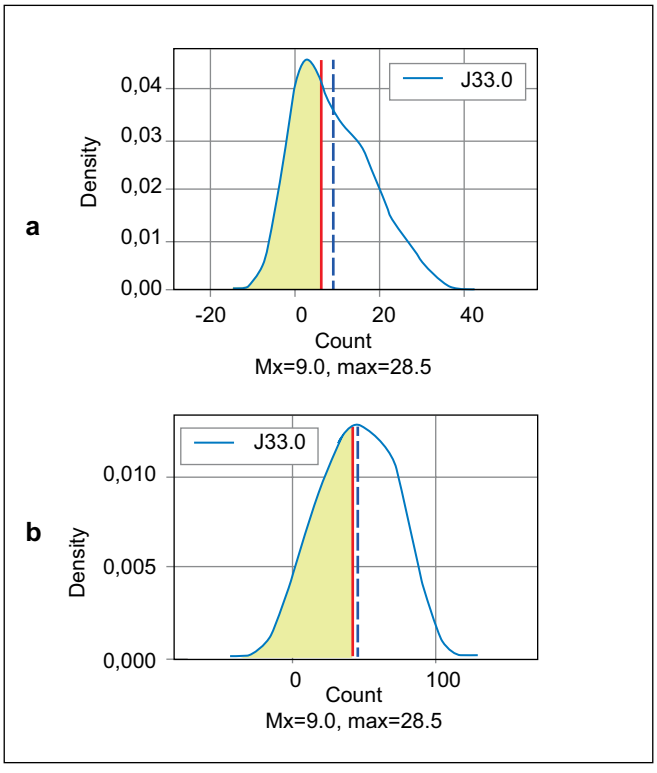

Рисунок 3. *Оценка плотности распределений J33.0 по районам за период 2017–2021 гг. (визуальная оценка для знака аргумента > 0); вертикальная сплошная линия – медиана распределения; вертикальная пунктирная линия – оценка среднего: Mx – среднее, max – максимальное учитываемое количество при аппроксимации плотности: a – заболеваемость; b – распространенность.*

Figure 3. *Estimation of distribution density for J33.0 by districts for the period 2017–2021 (visual evaluation for argument sign > 0); the vertical solid line is the distribution median; vertical dotted line – average estimate: Mx – average, max – the maximum number taken into account when approximating the density: a – incidence; b – prevalence.*

техникой машинного обучения при аппроксимации формулы движения [14–16]. Для формирования достаточной по объему тестовой выборки (≈20% от обучающей) временной ряд J33.0 (**таблица 6**) был интерполирован. Далее использовалась модель машинного обучения с автоматическим обучением. Программный код построен с использованием библиотеки машинного обучения PyCaret (Python). Для достижения оптимальной точности прогноза применялась модель на обучающих данных Random Forest Regressor. Выбираем шаги прогноза в интервале $I_p \in [\frac{1}{6}; \frac{1}{4}]$ от периода года.

	p-value вероятность не отклонения гипотезы распределения, параметры распределения		
Диагноз	Заболели	Заболеваемость	Распространенность
J33.0	Обратное Гаусса (Вальда) p=0.815 (11.66, 87.68)	Обобщенное экстремальных значений (genextreme) p=0.75 (-0.4, 4.07, 4.96)	Вейбулла, непрерывное, ограниченное снизу (weibull_min) p=0.956 (3.38, -26.45, 80.31)
J33.1	Нормальное p=0.017 (1.12, 2.14) Отвергаем H_0	Логистическое p=0.000 (–, –) Отвергаем H_0	Логистическое p=0.015 (1.49, 1.79) Отвергаем H_0
J33.8	Косинусоидальное p=0.008 (2.16, 2.09) Отвергаем H_0	Логистическое p=0.000 (–, –) Отвергаем H_0	Нормальное p=0.009 (5.92, 12.5) Отвергаем H_0
J33.9	Альфа p=0.14 (3.78, -0.57, 0.96)	Логистическое p=0.003 (0.60, 8.89) Отвергаем H_0	Альфа p=0.139 (2.37, -1.33, 2.17)

Таблица 5. *Вероятностная оценка аппроксимации распределений по данным значения меры p-value теста Колмогорова – Смирнова*

Table 5. *Probabilistic estimation of approximation of distributions based on the data of the value of the p-value measure of the Kolmogorov–Smirnov test*

	Заболеваемость				Распространенность			
Год	J33.0	J33.1	J33.8	J33.9	J33.0	J33.1	J33.8	J33.9
2017	14,93	0,00	3,53	0,51	60,42	5,47	19,14	8,16
2018	10,63	0,28	0,00	2,21	46,62	4,74	20,15	15,12
2019	8,71	0,23	0,95	1,28	49,16	4,83	17,53	19,63
2020	6,90	0,28	4,28	2,70	40,73	3,61	12,76	21,37
2021	6,79	0,41	0,54	0,67	47,95	5,52	2,81	7,56

Таблица 6. *Временной ряд подготовки прогноза – агрегированные по среднему относительные показатели диагноза; выделенные столбцы – кандидаты на прогноз*

Table 6. *Forecast preparation time series – aggregated mean relative diagnosis data; selected columns – candidates for the forecast*

Выборки (пара)	Заболеваемость	Распространенность	H_0
(2017; 2018)	Statistics T_d =1.075, p=0.294	Statistics T_d =1.810, p=0.083	Не отклонена
(2018; 2019)	Statistics T_d =0.606, p=0.550	Statistics T_d =-0.472, p=0.642	Не отклонена
(2019; 2020)	Statistics T_d =0.818, p=0.422	Statistics T_d =1.329, p=0.197	Не отклонена
(2020; 2021)	Statistics T_d =0.208, p=0.837	Statistics T_d =-0.384, p=0.704	Не отклонена

Таблица 7. *Проверка H_0: распределения параметров равны для зависимых парных выборок (J33.0)*

Table 7. *Checking the hypothesis H_0: parameter distributions are equal for dependent paired samples (J33.0)*

Временные ряды, решение регрессии и прогноза графически отображены **на рисунках 4, 5**.

Поведение детерминированной модели регрессии в обоих случаях прогнозирования также интерпретируется как вероятностный вывод на уровне гипотез (без факта утверждения). Для распространенности по районам не отвергаем гипотезу незначительного положительного дрейфа заболевания (**рисунок 4**). Для первичной заболеваемости не отвергаем гипотезу устойчивого тренда количественного снижения показателей (**рисунок 5**).

ОБСУЖДЕНИЕ

Вероятностные отклонения категориальных и счетных статистик общего периода наблюдений по районам формируют аномальные выбросы. Выбросы смещают оценки кроссируемых данных. Формируются распределения, у которых оценки средних находятся в районе процентиля 0.75% реализованной меры вероятности. Следовательно, в большинстве случаев попытки принятия гипотезы распределения окажутся неуспешными. Вопросы регрессии и прогноза будут неразрешимы. Актуальными для эксперта остаются численные таблицы и диаграммы. В качестве исключения из общего следует обратить внимание на диагноз J33.0, проявляющий себя, согласно предварительно проведенным оценкам, как самый адекватный процесс для использования вероятностных выводов и проверки гипотез при его исследовании (**рисунок 1**).

Тесты подтвердили характер распределений по типу данных «Категория-счетчик» как отличный от популярных распределений в медицинской статистке. Ни одно распределение диагнозов по годам не обладает характеристикой монотонности и однородностью соотношения пиковых значений. Это предварительно говорит о слабых корреляциях между процессами как по категориям, так и по счетным данным. Большая часть распределений не идентифицируются как популярные функции плотности, а определяются как общие α или β распределения физических процессов, например, по типу медленно дрейфующего броуновского движения (**таблица 5**). Здесь положительным является возможность отследить изменение медианы в выборках и принять ее как оценку ситуации прогресса (регресса) первичных заболеваний за период. Тем не менее распределения с дрейфом характеристик, включая дисперсионные, не гарантируют адекватности регрессионных моделей и прогноза на их основании. В этом случае следует сконцентрироваться на получении для эксперта табличных данных, графических аппроксимаций и дискретных вероятностей.

Разделение заболевших в районах за период 2017–2021 гг. по полу кардинальных изменений в свойства распределений не вносит. Известно, что сумма похожих выборок сохраняет свойства распределения в целом. Действительно, корреляционные зависимости диагнозов при разделении числа диагностированных по полу значимых линейных зависимостей подсчета заболевших не обнаруживают (**пункт 7 таблицы 2**). В описательной статистике (**пункт 8 таблицы 2**) отслеживаются соотношения средних групп и пониженных вероятностей их наблюдения в районе квантиля 0,75, тем самым не меняя свойств по отношению к ранее сделанным выборкам. В результате можно сделать заключение, что комбинация категорий районов, диагнозов и счетного количества заболевших результата в получении особо характерных статистик не приносят.

Ситуация меняется, если категории и счетчики связать с количественными данными, зависящими от некоторой хронологии. Тогда в выборках начинают проявляться характерные статистические свойства. Например, зависимости между группами, разделенные по полу (**таблица 4, рисунок 2**). В мужской и женской группе различен состав коррелирующих пар диагнозов заболеваний по годам (**таблица 4**). Особой экспертной оценки требует факт разнополярной значимой корреляции мужской и женской группами (**рисунки 2b, 2c**). Линейная зависимость по диагнозам в подгруппах изображена **на рисунке 2**. Сильная линейная зависимость у J33.0 (**рисунок 2a**); отрицательная зависимость J33.1 и J33.8 (**рисунки 2c, 2b**); J33.9 – смешанная зависимость, которая положительна только на участке 2020–2021 гг.

Учет периодов измерения и количественных признаков отображает некоторые характерные зависимости (**таблица 5**), но попытки определить требуемые для прогноза распределения закончились неотрицанием гипотез неподходящих типов распределений или неопределением каких-либо законов вообще (**таблица 5, диагнозы J33.1, J33.8, J33.9**). Фактически следует отказаться от мало практикуемых для проводимых исследований формул распределений. Они более пригодны для

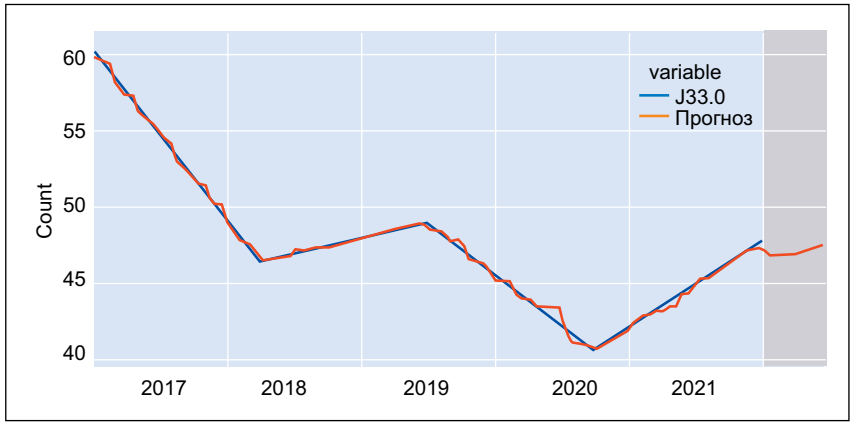

Рисунок 4. *Регрессия и прогноз распространенности заболевания, диагноз J33.0.*

Figure 4. *Regression and prediction of disease prevalence, diagnosis J33.0.*

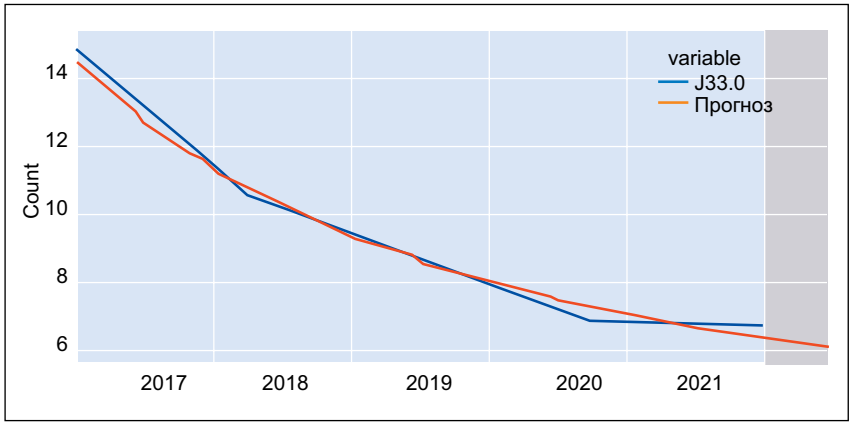

Рисунок 5. *Регрессия и прогноз первичной распространенности заболевания, диагноз J33.0.*

Figure 5. *Regression and prognosis of the primary prevalence of the disease, diagnosis J33.0.*

физических моделей. Оценки параметров этих функций не популярны в статистике практических подсчетов. Интерпретация этих функций сложна, кроме того, эти распределения чувствительны к объему выборки. Тогда на основании вероятностей неотклонения (**таблица 5**) для анализа на данном этапе следует применять только эмпирические расчеты дискретных функций вероятности при неизвестном законе плотности.

Все сказанное выше следует понимать как сложную ситуацию, возникшую в ходе анализа данных. При использовании только оригинальных показателей фактически отсутствует возможность принятия гипотез и вероятностных выводов, типичных для медицинской статистики. Ситуацию можно исправить применением компьютерных технологий группировки и агрегирования исходных данных. Современные программные средства позволяют для каждого подобного случая аппроксимировать плотность или построить ломаную вероятности дискретного распределения. Этого достаточно для выдвижения гипотез о средних и дисперсиях рассматриваемых процессов. В итоге акцент интерпретации данных переносится на таблично-графические показатели и опыт эксперта.

Тенденции корреляционной связи наблюдаются при разделении группы по полу и установке хронологии получения данных (**рисунок 2, таблица 4**). Временные ряды для непараметрической регрессии можно формировать только при наличии относительных данных, формирующих оценки среднего за период по районам (**таблица 6**). И только при неотвержении гипотезы схожести распределений (**таблица 7**) становится возможным достижение поставленной цели – сделать переход к машинной модели обучения для получения прогнозируемых траекторий.

ВЫВОДЫ

Обобщенные распределения всех диагнозов заболевших по районам не имеют устойчивого или прогрессирующего характера. Это внешне напоминает экстремально-пиковые распределения с закономерностью последующего быстрого экспоненциального распада вероятности. В этом случае информативность категорий для эксперта будут обеспечивать в основном табличные данные. Вывод о каких-либо зависимостях внутри перечня диагнозов будет затруднен ввиду малого набора типов диагнозов, равного четырем. Можно сделать вывод, что визуализация категорий в данном случае отражает характерное состояние отдельных элементов (**рисунок 1b**), но предварительно не выявляет зависимости для более детальных исследований. Требуется анализ численных данных.

Размещение категорий, счетчиков и количества во временной ряд делает модель генерации случайных величин (счета заболеваний) более адекватной по статистическим свойствам. Определяются корреляционные зависимости разделенных по полу подгрупп (**таблица 4**). При этом диагнозы могут иметь противоположные знаки зависимостей. Нормирование векторов временных линий (**рисунок 2**) позволяет определить условную разницу заболеваемости по диагнозам мужчин и женщин за весь период. Но тесты не определили необходимых для прогноза распределений (**таблица 5**). Как и в предыдущем случае, проверки выявляют зависимости, подобные обратному гауссиану, описывающему распределение времени броуновского движения с положительным дрейфом. Наличие дрейфа по отношению к диагнозу можно характеризовать положительной медианой. Если дрейфа нет, то медиана равна нулю. На этом этапе значимыми для эксперта по-прежнему остаются табличные данные и дискретные оценки элементов вероятности с неизвестным законом распределения.

На примере сведений о полипозном риносинусите мы сталкиваемся со сложной в плане анализа ситуацией распределения редких событий. Ситуация осложняется тем, что попытка не опровергнуть наблюдаемую реализацию событий с малой вероятностью приводит к неуспеху – гипотеза распределения Пуассона не подтверждена (**таблица 3**). Для экспертной оценки актуальными остаются только табличные данные и графический вывод современных программных средств отображения параметров статистик. При этих условиях классическая регрессия не применима для групп (**рисунок 2, таблица 4**). Агрегирование данных счетного порядка приводит их к суммированию, то есть не изменяет распределение, отличное от нормального. Этим не обеспечивается предположение о постоянстве дисперсий с возможностью решения моделей линейной или множественной регрессии.

Решение проблемы обеспечивает наличие относительных данных вещественных признаков. Соответственно со стороны решения проблем обеспечения прогноза предварительно нелинейную регрессию можно получить по временному ряду, только если присутствуют относительные количественные признаки. Их осреднение при агрегировании и кроссировании нормализуют оценки за период. После прохождения парного студенческого Т-теста можно приступать к выбору модели машинного обучения для целей прогноза.

Конфликт интересов: *все авторы заявляют об отсутствии конфликта интересов, требующего раскрытия в данной статье.*

ЛИТЕРАТУРА / REFERENCES

1. You J, Tulpan D, Malpass MC, et al. Using machine learning regression models to predict the pellet quality of pelleted feeds. *Animal Feed Science and Technology.* 2022;293:115443. doi: 10.1016/j.anifeedsci.2022.115443

2. Hu Y, Xia X, Fang J, et al. A Multivariate Regression Load Forecasting Algorithm Based on Variable Accuracy Feedback. *Energy Procedia.* 2018;152:1152-1157. doi: 10.1016/j.egypro.2018.09.147

3. Kumari Kh, Yadav S. Linear regression analysis study. *Curriculum in cardiology – statistics*. 2018;4(1):33-36. doi: 10.4103/jpcs.jpcs_8_18

4. Almalik O. Combining dependent p-values resulting from multiple effect size homogeneity tests in meta-analysis for binary outcomes. *Journal of Medical Statistics and Informatics*. 2021;1. doi: 10.7243/2053-7662-10-1

5. Hart J. Comparison of p-value results between one versus two sample t testing: A case study. *Journal of Medical Statistics and Informatics.* 2021;10. doi: 10.7243/2053-7662-9-1

6. Iftikhar S. Modification in inter-rater agreement statistics-a new approach. *Journal of Medical Statistics and Informatics.* 2020;8(1):2. doi: 10.7243/2053-7662-8-2

7. Basu A. Does a country's scientific 'productivity' depend critically on the number of country journals indexed? *Scientometrics*. 2010;3. doi: 10.1007/sllI92-010-0186-8

8. Almalki A. Regression Analysis for COVID-19 Infections and Deaths Based on Food Access and Health Issues. *Healthcare*. 2022;10(2):324. doi: 10.3390/healthcare10020324

9. Silhavy R, Silhavy P, Prokopova Z. Analysis and selection of a regression model for the Use Case Points method using a stepwise approach. *Journal of Systems and Software.* 2017;125:1-14. doi: 10.1016/j.jss.2016.11.029

10. Trubiani C, Ghabi A, Egyed A. Exploiting traceability uncertainty between software architectural models and extra-functional results. *Journal of Systems and Software*. 2017;125:15-34. doi: 10.1016/j.jss.2016.11.032

11. García-Floriano A, López-Martín C, Yáñez-Márquez C, et al. Support vector regression for predicting software enhancement effort. *Information and Software Technology.* 2018;97:99-109. doi: 10.1016/j.infsof.2018.01.003

12. Bach P, Wallisch Ch, Klein N, et al. Systematic review of education and practical guidance on regression modeling for medical researchers who lack a strong statistical background: Study protocol. *National Library of Medicine,* 2020;21;15(12). doi: 10.1371/journal.pone.0241427

13. Rambotti S, Breiger RL. Extreme and Inconsistent: A Case-Oriented Regression Analysis of Health, Inequality, and Poverty. *Sage*. 2020;18. doi: 10.1177/2378023120906064

14. Crabtree BF, Ray SC, Schmidt PM, et al. The individual over time: Time series applications in health care research. *Journal of Clinical Epidemiology.* 1990;43(3):241-260. doi: 10.1016/0895-4356(90)90005-A

15. Festag S, Denzler J, Spreckelsen C. Generative adversarial networks for biomedical time series forecasting and imputation. *Journal of Biomedical Informatics.* 2022;129:104058. doi: 10.1016/j.jbi.2022.104058

16. Morid MA, Sheng Ol, et al. Time Series Prediction Using Deep Learning Methods in Healthcare. *Transactions on Management Information Systems.* 2023;14(1):1-29. doi: 10.1145/3531326