Математическая биология и биоинформатика
2022. Т. 17. № 2. С. 230–249. doi: 10.17537/2022.17.230

==========ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ ==========
==========ТЕХНОЛОГИИ В БИОЛОГИИ И МЕДИЦИНЕ==========

Применение закона Бенфорда для оценки качества данных профилактического скрининга

Старунова О.А.[1], Руднев С.Г.[2], Иванова А.Е.[3], Семёнова В.Г.[4], Стародубов В.И.[5]

Центральный научно-исследовательский институт организации и информатизации здравоохранения, Москва, Россия

Аннотация. Эмпирический закон Бенфорда, описывающий вероятность появления определённых первых значащих цифр во многих распределениях, взятых из реальной жизни, используется для выявления аномалий в различного рода данных. Целью исследования является апробация закона Бенфорда для анализа качества массовых данных профилактического скрининга на примере данных биоимпедансных измерений в центрах здоровья Москвы. Как было установлено ранее, особенностью таких данных является их сильное зашумление искусственно сгенерированными и поддельными данными. Сформированная база данных биоимпедансометрии центров здоровья Москвы за 2010–2019 гг. содержала 1361019 записей результатов измерений в возрастном диапазоне обследованных от 5 до 96 лет. Применение алгоритма экспертной оценки качества данных, использованного в качестве эталона для анализа эффективности Бенфорд-анализа, выявило высокий процент некорректных данных (66.5 %) с преобладанием сфальсифицированных данных. Для характеристики степени соответствия данных закону Бенфорда для каждого центра здоровья рассчитывали средние абсолютные отклонения частот встречаемости первой и первых двух значащих цифр от должных значений и статистики χ^2 для десятых степеней стандартизованных значений активного, реактивного сопротивлений импеданса и индекса активного сопротивления. Установлена значимая корреляция между отклонением данных от закона Бенфорда и процентом некорректных данных согласно алгоритму экспертной оценки качества ($\rho_{max} = 0.66$ и 0.62 для среднего абсолютного отклонения и величины χ^2, соответственно, на основе параметра активного сопротивления импеданса и первой значащей цифры). Получено, что отклонение данных от закона Бенфорда является достаточным условием их компрометированности. Для центров здоровья, где основную часть некорректных данных составляли многократные измерения одного человека под видом разных, данные хорошо соответствовали закону Бенфорда. Если же в структуре некорректных данных преобладали измерения калибровочного блока, программные эмуляты измерений и выбросы, то использование закона Бенфорда позволяло эффективно ранжировать центры здоровья по уровню компрометированности данных.

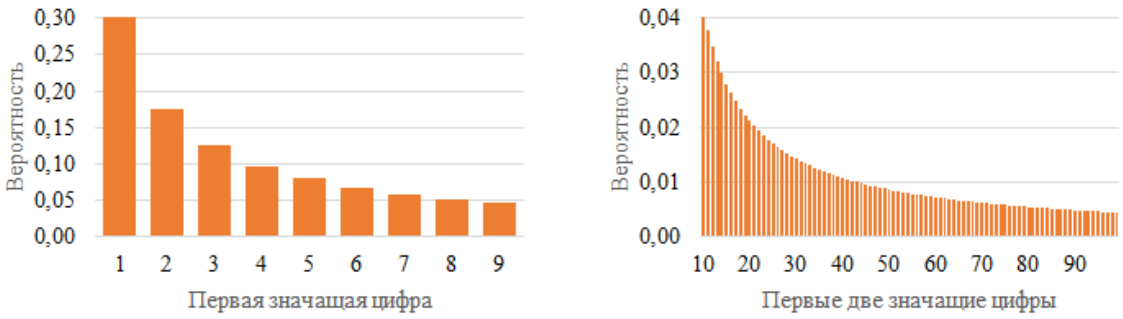
Ключевые слова: *центры здоровья, профилактический скрининг, большие данные, биоимпедансометрия, качество данных, алгоритм экспертной оценки качества, закон Бенфорда.*

[1]o.a.starunova@gmail.com
[2]rdnv2019@yandex.ru
[3]ivanova-home@yandex.ru
[4]vika-home@yandex.ru
[5]starodubov@mednet.ru

ВВЕДЕНИЕ

Назовём множество положительных чисел *удовлетворяющим закону Бенфорда* [1], если вероятность того, что случайно выбранное число из данного множества начинается с цифры $d\left(d \in \{1,...,9\}\right)$, равна $P(d) = \log_{10}(d+1) - \log_{10} d = \log_{10}\left(1+\frac{1}{d}\right)$ (рис. 1,а).

Закон Бенфорда хорошо описывает вероятность появления первых цифр в числовых множествах природных, биомедицинских, социологических и иных данных. Примерами являются распределения численностей людей в населённых пунктах [2], результаты выборов [3], суммарная длительность нот в классических музыкальных произведениях [4] и данные заболеваемости и смертности от новой коронавирусной инфекции [5]. Из приведённого определения следует, что числовое множество удовлетворяет закону Бенфорда, если логарифмы первых цифр его элементов имеют равномерное случайное распределение. Вообще говоря, выборочные данные соответствуют закону Бенфорда тем лучше, чем больше их вариация (на несколько порядков). В свою очередь, данные, вариация которых невелика (величины одного порядка – например, росто-весовые данные), в исходном виде закону Бенфорда заведомо не удовлетворяют.



а) б)

Рис. 1. Распределение Бенфорда для первой (а) и первых двух значащих цифр (б).

Считается, что впервые на неравномерность появления первых цифр в числовых данных, взятых из реальной жизни, обратил внимание американский астроном Ньюком в 1881 году [6]. Ввиду этого закон Бенфорда нередко именуют законом Ньюкома-Бенфорда. При работе со сборниками таблиц десятичных логарифмов Ньюком заметил повышенную истёртость тех страниц, где данные начинались с единицы, и относительную нетронутость страниц, где данные начинались с цифры девять. При этом с увеличением номера первой значащей цифры истёртость страниц снижалась. Позднее это свойство было отмечено и в работах других авторов (см., например, [7, 8]), а математическая формулировка данной зависимости предложена в 1938 году в статье Бенфорда [1].

В упомянутой выше работе Ньюком также отметил неравномерность появления первых двух значащих цифр в данных [6]. В статье Хилла [9] была приведена общая формулировка закона Бенфорда для первых k значащих цифр: вероятность того, что мантисса случайного выбранного числа начинается с цифр $\underline{d_1 d_2 ... d_k}$, составляет

$$P\left(D_1 = d_1, ..., D_k = d_k\right) = \log_{10}\left(1 + \frac{1}{\sum_{i=1}^{k} d_i \cdot 10^{k-i}}\right), d_1 \in \{1,...,9\}, d_{i\neq 1} \in \{0,...,9\},$$

где D_i – i-я значащая цифра числа, $i \in \{1, \ldots, k\}$. Там же была установлена эквивалентность приведённой формулировки статистическим свойствам выборки.

Математическая биология и биоинформатика. 2022. Т. 17. № 2. doi: 10.17537/2022.17.230

Распределение вероятностей значений первых двух значащих цифр для данных, удовлетворяющих общей формулировке закона Бенфорда, показано на рисунке 1,б.

В работе Нигрини [10] описаны способы проверки числовых множеств на соответствие закону Бенфорда. Для этого рассматривают частотные распределения первой, второй или первых двух значащих цифр одновременно, а также первых трёх или последних двух значащих цифр. Кроме того, применяются разности соседних значений упорядоченных данных (тест второго порядка). Иногда также рассматривают суммы чисел, начинающихся с одних и тех же значащих цифр (суммирующий тест), при этом соответствующая случайная величина должна иметь равномерное распределение. Гипотезу о соответствии данных закону Бенфорда проверяют при помощи различных статистических критериев, включая критерии согласия Пирсона и Колмогорова – Смирнова.

На платформе http://www.benfordonline.net/ [11, 12] представлено свыше 1500 оригинальных исследований, посвящённых закону Бенфорда. Более 85 % из них опубликованы за последние 20 лет, когда в связи с резким увеличением объёмов информации, производимой в различных сферах деятельности, выросла необходимость обработки больших и сверхбольших массивов данных с сопутствующей проверкой их на предмет достоверности.

Основной причиной смертности населения в большинстве стран мира, включая Россию, являются хронические неинфекционные заболевания, при этом значительные усилия направлены на меры по их профилактике [13, 14]. Хронические неинфекционные заболевания связаны с нарушениями нутритивного статуса и изменениями массы тела, традиционно оцениваемыми на основе расчёта индекса массы тела (ИМТ) [15]. Высокие значения ИМТ, избыточный вес и ожирение ассоциированы с повышенной восприимчивостью и тяжестью течения коронавирусной инфекции COVID-19 [16]. Уточнённую характеристику нутритивного статуса даёт оценка состава тела, под которым понимается представление массы тела в виде суммы нескольких компонентов, имеющих физиологическое и патофизиологическое значение [17]. Наиболее распространённым методом оценки состава тела является биоимпедансный анализ, основанный на измерении электрического сопротивления (импеданса) тела в ответ на приложенное внешнее электрическое поле низкой интенсивности с использованием специального оборудования – биоимпедансного анализатора состава тела. С использованием биоимпедансного анализа можно получить информацию о состоянии гидратации тела человека, а также оценить жировую, мышечную, активную клеточную массу, уровень основного обмена и другие характеристики [18].

Значимым элементом системы профилактического скрининга в России является национальная сеть центров здоровья [19]. Данная сеть была создана в конце 2009 года с целью обследования здорового населения трудоспособного возраста для выявления факторов риска неинфекционных заболеваний и насчитывает порядка 800 стационарных и мобильных центров при поликлиниках, больницах и других медицинских учреждениях. На базе центра здоровья, согласно штатным расписаниям, может проводиться комплексное обследование до 20–30 и более пациентов в день с использованием ряда инструментальных методов, таких как измерение роста и веса, измерение артериального давления, кистевая динамометрия, пульсоксиметрия, кардиоскрининг и биоимпедансный анализ состава тела [20]. Таким образом, национальная сеть центров здоровья в постоянном режиме генерирует значительный объём данных инструментальных измерений. Часть из них хранится в локальной информационной системе центра здоровья и/или в базе данных устройства, на котором производилось измерение. До 2014 года часть указанных данных централизованно выгружалась на федеральный уровень с использованием возможностей многоуровневой медицинской информационной системы «Программный комплекс центров здоровья», разработанной компанией СофТраст (г. Белгород) [21], а после 2014 года из всего объёма

генерируемых данных систематически, но в ручном режиме, аккумулировались только данные биоимпедансных измерений [21–23]. Массовость таких данных и невозможность оперативной ручной проверки их качества привели к необходимости автоматизации обработки данных и созданию специализированного программного комплекса HCViewer [22, 24].

Расходы на здравоохранение в большинстве стран мира представляют собой ведущую статью бюджетных трат и потому являются привлекательной сферой для мошенничества [25]. Общемировые потери от мошенничества в сфере здравоохранения оцениваются величиной порядка 6 % от глобальных расходов на здравоохранение [26]. В нашей стране эта тема сравнительно мало изучена, а надзорная деятельность в здравоохранении, в отличие от банковской сферы и области компьютерной безопасности, сегодня ведётся без использования возможностей технологий больших данных [27].

Анализ первичных данных биоимпедансных измерений в центрах здоровья России за 2010–2015 гг. с использованием программного комплекса HCViewer (общий размер базы данных составил 2.4 млн. записей результатов измерений, выполненных в 335 центрах здоровья 62 субъектов Российской Федерации) выявил наличие их сильного зашумления искусственно сгенерированными и поддельными данными: суммарный процент некорректных данных, согласно полученной оценке, составил 44.8 % [22]. При этом центры здоровья отличались высоким уровнем неоднородности качества данных с выполнением закона Парето: около 80 % всего объёма некорректных данных были сгенерированы в 20 % центров здоровья [22]. Использованный в указанной работе алгоритм оценки качества данных биоимпедансных измерений в центрах здоровья был основан на знании особенностей методики измерений, биоимпедансного оборудования и программного обеспечения центров здоровья. Однако на практике алгоритмы экспертной оценки качества данных бывают доступны далеко не всегда, поэтому желательно иметь возможность проверки качества данных без участия экспертов. Одна из таких возможностей заключается в использовании закона Бенфорда.

Цель исследования – апробация закона Бенфорда для анализа качества массовых данных профилактического скрининга на примере данных биоимпедансных измерений в центрах здоровья Москвы.

МАТЕРИАЛЫ И МЕТОДЫ

База данных биоимпедансных измерений

Массив данных биоимпедансных измерений в центрах здоровья Москвы за 2010–2019 гг. был сформирован в результате аккумулирования данных из различных источников. Данные за 2018–2019 гг. были получены по письму ЦНИИОИЗ Минздрава России №7-5/1498 от 17.12.2019 г. и №7-5/1020 от 31.08.2020 г., а за 2015–2017 гг. – по письму ЦНИИОИЗ №7-5/1067 от 27.11.2017 г. (ранее эти данные не обрабатывались). Данные за предшествующие годы были получены по письму Минздрава России №14-1/10/2-3200 от 24.10.2012 г. [23], из федерального информационного ресурса центров здоровья [21] и по письму ЦНИИОИЗ №7-5/434 от 02.07.2015 г. [22].

По сведениям официальной статистики, на сегодняшний день в Москве действует 62 центра здоровья (https://data.mos.ru/opendata/tsentry-zdorovya/data/table). Их количество и принадлежность к тому или иному лечебно-профилактическому учреждению, а также адреса за прошедшие 10 лет изменялись. В итоге, данные за разные годы были получены из 79 центров здоровья Москвы (рис. 2). Все они соответствовали измерениям анализаторами состава тела АВС-01 «Медасс» (ООО НТЦ Медасс, г. Москва).

Данные биоимпедансных измерений, полученные в разные годы из различных источников, имели непустые пересечения. В связи с этим при объединении данных для каждой записи биоимпедансного обследования пациента был сформирован

квазиуникальный код на основе названия центра здоровья, даты рождения, даты измерения и пола пациента, и удалены повторы. Кроме того, были удалены записи повторных измерений в течение одного визита за исключением последнего по времени измерения (так как повторные измерения пациента обычно проводятся при наличии сомнений в качестве результатов обследования). Также были удалены записи, для которых отсутствовала информация о росте, весе, дате рождения, дате и времени измерения, поле пациента или значениях активного и реактивного сопротивлений импеданса. В процессе объединения данные были преобразованы из форматов разработчика fmd и fmd2 в единый промежуточный формат xlsx, а в окончательном виде сохранены в форматах mwx и csv. Сформированный начальный массив данных биоимпедансных измерений в центрах здоровья Москвы за 2010–2019 гг. содержал 1.362.333 записи.

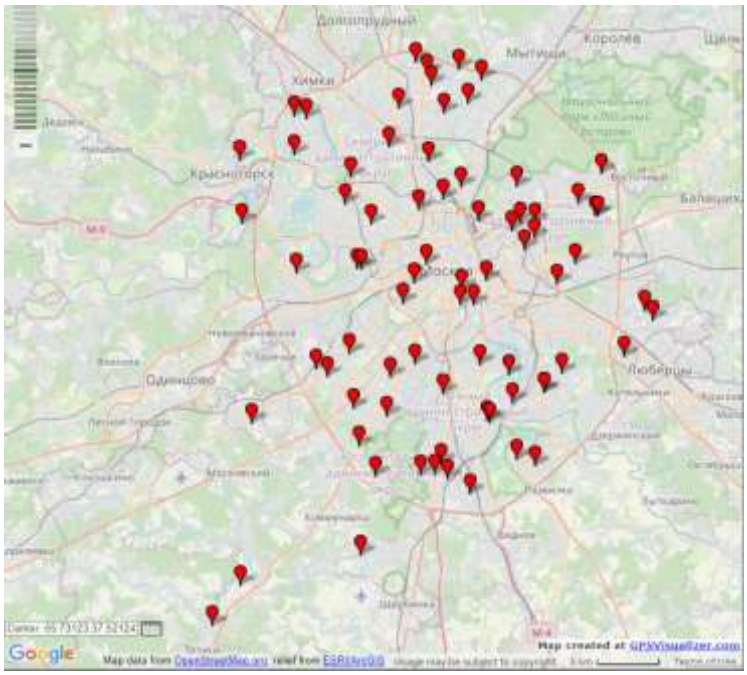


Рис. 2. Географическое положение центров здоровья Москвы, предоставивших данные биоимпедансных измерений.

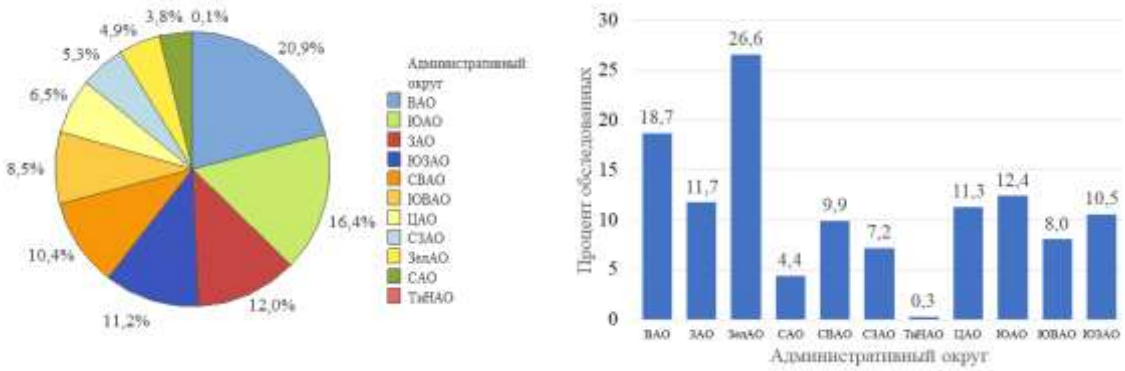


Рис. 3. Диаграмма распределения общего количества записей в базе данных (n = 1.362.333) по административным округам Москвы (слева) и процентное отношение количества записей к численности постоянного населения административного округа (справа). Обозначения административных округов Москвы: ВАО – Восточный; ЗАО – Западный; ЗелАО – Зеленоградский; САО – Северный; СВАО – Северо-Восточный; СЗАО – Северо-Западный; ТиНАО – Троицкий и Новомосковский; ЦАО – Центральный; ЮАО – Южный; ЮВАО – Юго-Восточный; ЮЗАО – Юго-Западный.

Каждая запись биоимпедансного обследования в базе данных содержала наименование и адрес центра здоровья, где проводилось измерение, а также название административного округа Москвы. Информация о пациенте также включала сведения о поле, дате рождения, дате и времени измерения, росте и весе, окружности талии и бёдер, активном и реактивном сопротивлениях импеданса на частоте 50 кГц, фазовом угле импеданса и параметрах состава тела.

Порядка 80 % записей в базе данных относились к шести административным округам: Восточному, Южному, Западному, Юго-Западному, Северо-Восточному и Юго-Восточному (рис. 3 слева), а по соотношению количества записей к численности постоянного населения округа лидировал Зеленоградский административный округ (рис. 3 справа). Ввиду малочисленности данных Троицкий и Новомосковский округа были объединены. Отношение общего количества записей в базе данных к численности постоянного населения Москвы на 1 января 2020 года составило 10.7 %. Таким образом, общее количество записей в базе данных примерно соответствовало обследованию 1 % населения Москвы в год.

Далее рассматривались только записи результатов измерений в возрастном диапазоне обследованных от 5 до 96 лет. Количество таких записей в базе данных составило 1.361.019.

Алгоритм экспертной оценки качества данных

Разметка данных, основанная на алгоритме экспертной оценки качества, проводилась с использованием программного комплекса HCViewer [22, 24] в два этапа.

Таблица 1. Интервалы допустимых значений параметров биоимпедансометрии для взрослых людей, использованные для обнаружения «выбросов» [22]

Рост, см	Вес, кг	ИМТ, кг/м^2	ЖМ, кг	% ЖМ	*R*, Ом	*X*, Ом	Фазовый угол, град.
130–210	35–150	12–55	0.5–75	0–55	250–1000	20–150	3.0–10.2

Примечания: ИМТ – индекс массы тела, ЖМ – жировая масса, % ЖМ – процентное содержание жира в массе тела, *R* и *X* – активное и реактивное сопротивления импеданса при измерении на частоте 50 кГц.

На первом этапе отмечались «выбросы» (outliers) – ошибки измерений и/или ввода данных на основе интервалов допустимых значений роста, веса, ИМТ, жировой массы (ЖМ), процентного содержания жира в массе тела (% ЖМ), активного (*R*), реактивного (*X*) сопротивлений и фазового угла импеданса, определяемого как $\mathrm{arctg}\left(X / R\right)\times 180° / \pi.$ Отметим, что субстратом активного (резистивного) сопротивления *R* в биологическом объекте являются вне- и внутриклеточные жидкости, обладающие ионным механизмом проводимости, а реактивного сопротивления *X* (диэлектрический компонент импеданса) – клеточные мембраны [17,18]. Поэтому величина фазового угла импеданса обычно интерпретируется как характеристика целостности клеточных мембран; пониженные значения фазового угла нередко встречаются в клинической медицине, где могут служить индикатором хронических заболеваний, а повышенные значения характерны для физически развитых людей и спортсменов высокой квалификации [18].) Если значения указанных параметров выходили за границы интервалов допустимых значений, то соответствующие записи в базе данных отмечались как содержащие выбросы. Для взрослых людей (возраст от 18 лет и старше) использовались постоянные (табл. 1), а для детей и подростков – зависящие от возраста интервалы.

На втором этапе выявлялись сфальсифицированные данные. В базе данных отмечали записи, для которых соответствующие измерения были классифицированы как

поддельные: либо измерение человека не проводилось вовсе, либо проводились многократные измерения одного человека под видом разных. В первом случае данные представляли собой результат измерения шаблона (эквивалентной электрической схемы биообъекта, используемой для проверки корректности измерений анализатором АВС-01 «Медасс») или программную эмуляцию измерения.

Указанные типы поддельных данных определяли следующим образом [22]:

1) Измерение шаблона (эквивалентной электрической схемы биообъекта) на частоте 50 кГц, как правило, даёт значения R и X в интервале от 381 до 391 Ом и от 38 до 48 Ом, соответственно. При одновременном нахождении значений R и X в указанных интервалах рассматриваемую запись считали измерением шаблона.

2) Программная эмуляция (имитация) измерения, обычно применяемая в центрах здоровья для настройки программного обеспечения анализатора, определялась при значениях R = 444, 444.4, 555.5 и 556 Ом (в нашем случае, в отличие от [22], дополнительно использовались значения 444 и 556 Ом ввиду особенностей округления данных для двух центров здоровья Москвы).

3) Если интервал времени между двумя соседними измерениями разных людей в центре здоровья составлял менее 90 секунд, то оба измерения считались поддельными, так как стандартная процедура обследования пациента (с измерением роста и веса, окружности талии и бёдер, внесением соответствующей информации в программное обеспечение анализатора, укладыванием пациента на медицинскую кушетку, фиксацией электродов и измерением импеданса) не может занимать столь короткое время.

4) Если записи двух соседних обследований в базе данных не классифицировались как измерение шаблона или программная эмуляция измерения и отличались менее чем на 1 % по величине R и, одновременно, менее чем на 7 % по величине X, то их считали измерениями одного человека под видом разных. Если на гистограмме распределения значений R в центре здоровья высота какого-либо столбца более чем в 1.5 раза превышала среднее арифметическое высоты четырёх соседних с ним столбцов (взятых по два слева и справа), а количество записей для данного интервала значений R превышало 50, то такие записи также считались измерениями одного человека под видом разных (таким способом выявляли измерения одного человека под видом разных, выполненные не подряд).

Можно отметить, что в указанной реализации первый, второй и четвёртый типы сфальсифицированных данных независимы. При этом третий тип сфальсифицированных данных (измерения с чрезмерно коротким интервалом времени между ними) является зависимым и может характеризовать манеру получения некорректных данных других типов, а «выбросы» могут также, формально, являться сфальсифицированными данными.

Реализация закона Бенфорда

Для проверки степени соответствия данных закону Бенфорда использовался пакет benford.analysis на языке R [28]. Данные были сгруппированы в соответствии с центром здоровья, где проводилось измерение. При этом были исключены из рассмотрения 2 из 79 центров здоровья с малым (1 и 59) количеством измерений. Для остальных 77 центров здоровья количество записей варьировало от 970 до 51214.

Соответствие закону Бенфорда проверяли с использованием величин активного (R), реактивного сопротивления импеданса (X) и индекса активного сопротивления $IR = H^2 / R$, где H – рост человека. При измерениях по традиционной схеме «запястье-голеностоп» указанные величины варьируют в пределах одного порядка, от 250 до 1000–1100 Ом для R, от 20 до 150 Ом для X, и от 10 до 100 см^2/Ом для IR (см. рис. 5), поэтому в исходном виде закону Бенфорда они не удовлетворяют. Однако известно, что естественные распределения величин R и X близки к нормальному (см., например, [29]),

Математическая биология и биоинформатика. 2022. Т. 17. № 2. doi: 10.17537/2022.17.230

а десятые степени стандартных нормально-, логнормально- и экспоненциально распределённых случайных величин хорошо соответствуют закону Бенфорда [30]. В связи с этим вместо исходных распределений величин *R*, *X* и *IR* рассматривались распределения десятых степеней их стандартизованных значений.

Таблица 2. Классификация значений *MAD*, применяемая для оценки степени соответствия данных закону Бенфорда на основе первой и первых двух значащих цифр [10, 28]

Количество первых значащих цифр	Соответствие закону Бенфорда			
	Тесное	Приемлемое	Минимальное	Несоответствие
1	0.000–0.006	0.006–0.012	0.012–0.015	> 0.015
2	0.000–0.0012	0.0012–0.0018	0.0018–0.0022	> 0.0022

Для характеристики отклонения данных от закона Бенфорда для каждого центра здоровья, а также первой и первой двух значащих цифр десятых степеней стандартизованных (с новым средним 0 и дисперсией 1) значений *R*, *X* и *IR*, рассчитывали среднее абсолютное отклонение (*MAD*, mean absolute deviation) наблюдаемых частот от теоретических [10]. При использовании первой значащей цифры данная величина имеет вид $MAD = \sum_{i=1}^{9} \frac{|p_i - x_i|}{9}$, а для первых двух значащих цифр $MAD = \sum_{i=10}^{99} \frac{|p_i - x_i|}{90}$, где p_i – теоретическое значение вероятности появления числа i в качестве первой или, соответственно, первых двух значащих цифр случайной величины, удовлетворяющей закону Бенфорда (см. общую формулировку закона Бенфорда в разделе «Введение»), а x_i – фактическая частота для тестируемой выборки. При этом использовали эмпирическую классификацию значений *MAD*, приведённую в таблице 2. Отметим, что величина *MAD* не зависит от размера выборки.

Помимо величины *MAD*, для каждого центра здоровья на основе первой и первых двух значащих цифр десятых степеней стандартизованных значений *R*, *X* и *IR* рассчитывали статистику $\chi^2 = \sum_{i=1}^{K} \frac{n(p_i - x_i)^2}{p_i}$, где n – размер выборки, а K – общее количество элементов суммы (9 и 90 соответственно). Для оценки значимости отклонений от закона Бенфорда использовались критические значения χ^2, равные 15.51 для распределений первой значащей цифры и 112.02 для распределений первых двух значащих цифр, при уровне значимости $p = 0.05$ и соответствующем количестве степеней свободы (8 и 89).

Эффективность применения закона Бенфорда оценивали путём расчёта ранговых корреляций Спирмена величин *MAD* и $\frac{\chi^2}{n}$, соответственно, с оценками процентного содержания некорректных данных в центрах здоровья на основе алгоритма экспертной оценки качества с учётом структуры некорректных данных.

РЕЗУЛЬТАТЫ

Применение алгоритма экспертной оценки качества

По результатам разметки данных в соответствии с алгоритмом экспертной оценки качества, 66.5 % данных биоимпедансных измерений в центрах здоровья Москвы за 2010–2019 гг. были некорректными (табл. 3).

Таблица 3. Результаты разметки данных биоимпедансометрии в центрах здоровья Москвы за 2010–2019 гг. (n = 1.361.019) программой HCViewer на основе алгоритма экспертной оценки качества. Данные для административных округов Москвы

Адм. округ Москвы	Общее кол-во записей	Выбросы		Сфальсифицированные данные								Всего некорректных данных	
				1		2		3		4			
		Абс.	%	Абс.	%	Абс.	%	Абс.	%	Абс.	%	Абс.	%
ВАО	285159	59119	20.7	79096	27.7	561	0.2	95168	33.4	100183	35.1	197108	69.1
ЗАО	163766	26567	16.2	38324	23.4	1562	1.0	90874	55.5	75203	45.9	124704	76.2
ЗелАО	66575	7358	11.1	33307	50.0	512	0.8	32830	49.3	4240	6.4	39967	60.0
САО	52154	10760	20.6	8563	16.4	22075	42.3	22241	42.6	7820	15.0	40911	78.4
СВАО	141682	26217	18.5	10225	7.2	329	0.2	62902	44.4	60125	42.4	83257	58.8
СЗАО	72795	45824	63.0	16744	23.0	2293	3.2	40494	55.6	35234	48.4	56816	78.1
ТиНАО	1047	9	0.9	1	0.1	1	0.1	588	56.2	561	53.6	613	58.6
ЦАО	88314	5930	6.7	52	0.1	483	0.6	51687	58.5	46420	52.6	64212	72.7
ЮАО	222629	13013	5.9	204	0.1	103	0.1	120498	54.1	117833	52.9	143314	64.4
ЮВАО	114375	35554	31.1	39568	35.0	17968	15.7	32728	28.6	17296	15.1	80285	70.2
ЮЗАО	152523	12087	7.9	21932	14.4	103	0.1	51683	33.9	29164	19.1	73818	48.4
Итого	1361019	242438	17.8	248016	18.2	45990	3.4	601693	44.2	494079	36.3	905005	66.5

Примечания: 1 – измерение шаблона (эквивалентной электрической схемы биообъекта); 2 – программная эмуляция измерения; 3 – измерения, выполненные подряд с коротким интервалом времени (менее чем через 90 секунд); 4 – многократное измерение одного человека под видом разных.

Обозначения административных округов Москвы: ВАО – Восточный; ЗАО – Западный; ЗелАО – Зеленоградский; САО – Северный; СВАО – Северо-Восточный; СЗАО – Северо-Западный; ТиНАО – Троицкий и Новомосковский; ЦАО – Центральный; ЮАО – Южный; ЮВАО – Юго-Восточный; ЮЗАО – Юго-Западный.

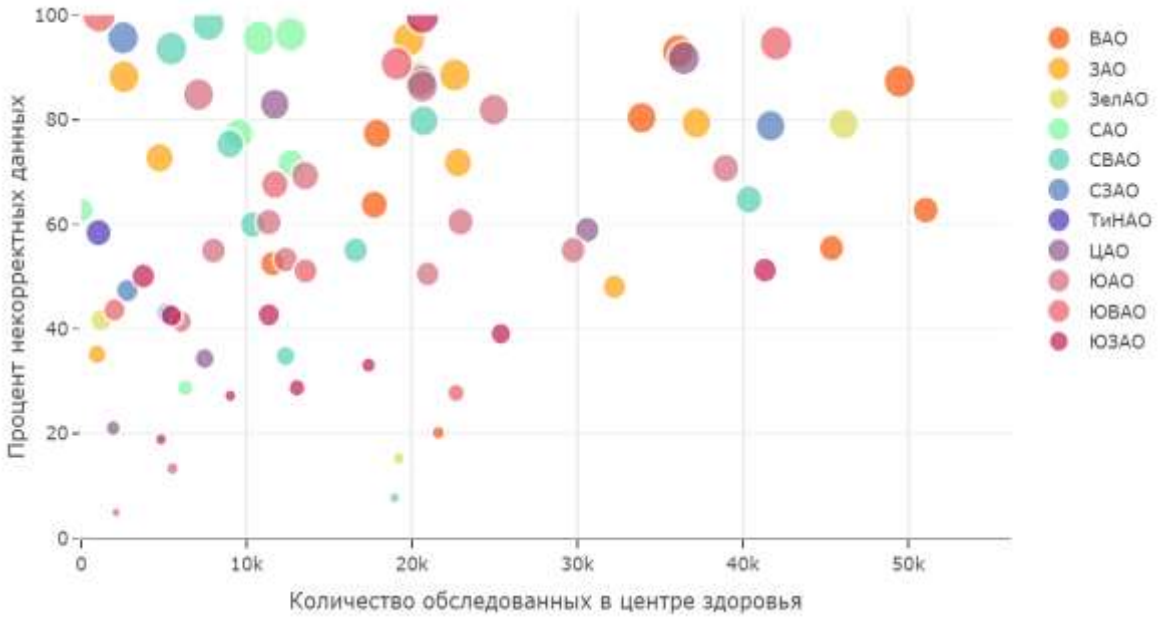


Рис. 4. Зависимость качества данных биоимпедансных измерений в центрах здоровья Москвы за 2010–2019 гг. (n = 1.361.019), определяемого программой HCViewer на основе алгоритма экспертной оценки качества, от количества обследованных в центре здоровья. Точка на рисунке соответствует центру здоровья, а цвет маркера – административному округу Москвы. Размер точек пропорционален доле подделок. Обозначения административных округов Москвы: ВАО – Восточный; ЗАО – Западный; ЗелАО – Зеленоградский; САО – Северный; СВАО – Северо-Восточный; СЗАО – Северо-Западный; ТиНАО – Троицкий и Новомосковский; ЦАО – Центральный; ЮАО – Южный; ЮВАО – Юго-Восточный; ЮЗАО – Юго-Западный.

На уровне административных округов Москвы, наибольшим процентом некорректных данных отличались Северный и Северо-Западный округа (78.4 % и 78.1 % соответственно), а наименьшим – Юго-Западный округ (48.4 %). Из таблицы 3 видно, что преимущественным источником таких данных по Москве в целом были многократные измерения одного человека под видом разных (36.3 %). Такие измерения чаще выполнялись сериями с коротким промежутком времени между ними (в 70.2 % случаев, данные не показаны). В одном административном округе Москвы (СЗАО) первым по значимости источником некорректных данных (63.0 %) были выбросы, в двух других (ЗелАО, ЮВАО) – измерения шаблона (50.0 % и 35.0 %, соответственно), и ещё в одном округе (САО) – программная эмуляция измерения (42.3 %).

Центры здоровья Москвы отличались высоким уровнем неоднородности качества данных биоимпедансных измерений (рис. 4). Приемлемым качеством отличались лишь некоторые центры здоровья с относительно небольшим количеством измерений – менее 20–22 тысяч за 10 лет, что соответствует средней частоте обследований в центре здоровья не более 7–8 человек в день. Лишь в пяти центрах здоровья Москвы доля некорректных данных биоимпедансометрии составила менее 20 %, и в 16 – менее 40 % от общего количества измерений.

Таблица 4. Центры здоровья Москвы с наименьшим процентом некорректных данных биоимпедансных измерений

№ п/п	Лечебно-профилактическое учреждение центра здоровья	Общее количество записей	Процент некорректных данных*
1	ГП №19 ДЗМ ЮВАО	2106	4.9
2	ГП №107 ДЗМ СВАО	18943	7.7
3	ГП №52 ДЗМ ЮАО	5517	13.4
4	ДГП №105 ДЗМ ЗелАО	19206	15.2
5	КДП №121 ДЗМ филиал №6 ЮЗАО	4830	18.9
6	ГП №66 ДЗМ ВАО	21590	20.2
7	ДГП №38 ДЗМ ЦАО	1933	21.1
8	ДКЦ №1 ДЗМ филиал №4 ЮЗАО	9015	27.2
9	ДЦ №3 ДЗМ филиал №2 ЮВАО	22662	27.8
10	ГП №134 ДЗМ филиал №3 ЮЗАО	13043	28.7
		Среднее значение	18.5

* На основе алгоритма экспертной оценки качества.

Обозначения: ГП – городская поликлиника; ДГП – детская городская поликлиника; ДКЦ – диагностический клинический центр; ДЦ – диагностический центр; КДП – консультативно-диагностическая поликлиника; ДЗМ – Департамент здравоохранения г. Москвы.

В таблице 4 указаны 10 центров здоровья Москвы с наименьшим процентом некорректных данных согласно алгоритму экспертной оценки качества.

После удаления «выбросов» и сфальсифицированных данных в соответствии с алгоритмом экспертной оценки качества распределения значений активного (R) и реактивного (X) сопротивлений стали близки к нормальному (рис. 5а,б), а распределение значений индекса активного сопротивления (IR) приобрело несколько более сложный вид (рис. 5,в).

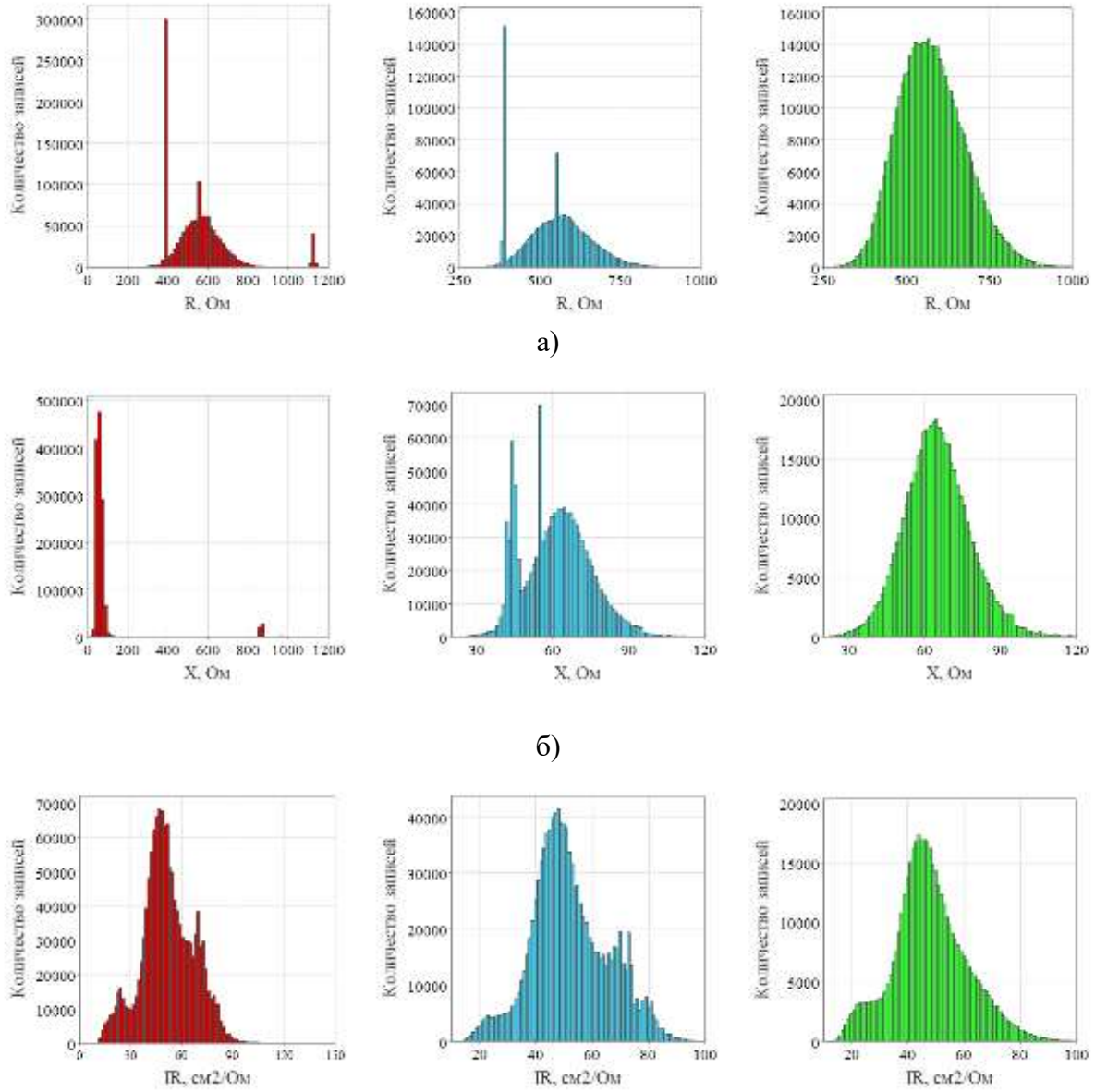


Рис. 5. Распределения значений активного сопротивления *R* (а), реактивного сопротивления *X* (б) и индекса активного сопротивления *IR* (в) на различных этапах фильтрации данных биоимпедансометрии в центрах здоровья Москвы за 2010-2019 гг. согласно алгоритму экспертной оценки качества: слева – начальное распределение, в центре – после удаления «выбросов», справа – после удаления «выбросов» и сфальсифицированных данных.

Бенфорд-анализ данных

Наблюдалась выраженная неоднородность центров здоровья Москвы по степени соответствия данных биоимпедансных измерений закону Бенфорда на основе значений *MAD* (табл. 5а). При использовании статистики χ^2 и общепринятом уровне значимости данные большинства центров здоровья закону Бенфорда не соответствовали, что объясняется зависимостью статистики χ^2 от объёма выборки (табл. 5б).

Ранговые корреляции распределений значений *MAD* для различных параметров и количества первых значащих цифр были достаточно высоки, от 0.76 до 0.95 (табл. 6), что свидетельствует о хорошей взаимной согласованности критериев. Ранговые корреляции распределений значений χ^2 соответствовали тому же диапазону, от 0.76 до 0.94. При этом «перекрёстные» ранговые корреляции величин *MAD* и χ^2 были в пределах 0.59–0.93 (данные не показаны).

Таблица 5а. Распределение центров здоровья Москвы по степени соответствия данных закону Бенфорда на основе пороговых значений *MAD*

Количество первых значащих цифр	Параметр*	Соответствие закону Бенфорда			
		Тесное	Приемлемое	Минимальное	Несоответствие
1	*R*	13	14	2	48
1	*X*	23	17	2	35
1	*IR*	42	18	7	10
2	*R*	5	11	10	51
2	*X*	17	13	8	39
2	*IR*	36	6	5	30

* Использованы десятые степени стандартизованных величин.

Таблица 5б. Распределение центров здоровья Москвы по степени соответствия данных закону Бенфорда на основе пороговых значений χ^2

Количество первых значащих цифр	Параметр*	Соответствие закону Бенфорда	
		$p < 0.05$	$p > 0.05$
1	*R*	2	75
1	*X*	10	67
1	*IR*	27	50
2	*R*	1	76
2	*X*	2	75
2	*IR*	22	55

* Использованы десятые степени стандартизованных величин.

Таблица 6. Корреляции Спирмена распределений значений *MAD*, характеризующих отклонение от закона Бенфорда, для центров здоровья Москвы

Параметр*, количество первых значащих цифр	*R*, 1	*X*, 1	*IR*, 1	*R*, 2	*X*, 2
X, 1	0.89	1			
IR, 1	0.81	0.76	1		
R, 2	0.95	0.88	0.81	1	
X, 2	0.86	0.93	0.76	0.93	1
IR, 2	0.87	0.95	0.86	0.92	0.91

* Использованы десятые степени стандартизованных величин.

Тесно соответствовали закону Бенфорда на основе значений *MAD* для всех рассмотренных в таблице 5а критериев одновременно лишь данные 4 из 77 центров здоровья. Они перечислены в таблице 7, где также приведены соответствующие количества записей результатов измерений и проценты некорректных данных согласно алгоритму экспертной оценки качества. Указанные центры здоровья отличались более низким средним значением процента некорректных данных (36.4 %) в сравнении со средним значением для центров здоровья Москвы (61.5 %). При этом процент некорректных данных для одного из четырёх центров здоровья (60.5 %) соответствовал общему среднему. Отметим, что на основе значений χ^2 согласно всем рассмотренным в таблице 5б критериям одновременно закону Бенфорда не соответствовал ни один из 77 центров здоровья.

Таблица 7. Центры здоровья Москвы с наименьшим отклонением данных от закона Бенфорда на основе значений *MAD* («тесное соответствие» согласно табл. 2 для каждого из рассмотренных в табл. 5а критериев)

№ п/п	Лечебно-профилактическое учреждение центра здоровья	Общее количество записей	Процент некорректных данных*
1	ВФД №17 ДЗМ СВАО	12352	34.8
2	ГП №107 ДЗМ СВАО	18943	7.7
3	ГП №170 ДЗМ филиал №1 ЮАО	22935	60.5
4	ДГП №103 ДЗМ ЮЗАО	11345	42.7
		Среднее значение	36.4

* На основе алгоритма экспертной оценки качества.

Обозначения: ВФД – врачебно-физкультурный диспансер; ГП – городская поликлиника; ДГП – детская городская поликлиника; ДЗМ – Департамент здравоохранения г. Москвы.

Таблица 8. Центры здоровья Москвы с наибольшим отклонением данных от закона Бенфорда на основе значений *MAD* («несоответствие» согласно табл. 2 для каждого из рассмотренных в табл. 5а критериев)

№ п/п	Лечебно-профилактическое учреждение центра здоровья	Общее количество записей	Процент некорректных данных*
1	ГП №11 ДЗМ филиал №1 ЮЗАО	20625	99.8
2	ГП №115 ДЗМ СЗАО	2530	96.7
3	ГП №209 ДЗМ филиал №40 ЗАО	4740	72.7
4	ГП №23 ДЗМ ЮВАО	1079	100
5	ГП №6 ДЗМ филиал №1 САО	10760	95.7
6	ГП №6 ДЗМ филиал №3 САО	12656	96.3
7	ГП №64 ДЗМ ВАО	49460	87.3
8	ГП №69 ДЗМ ВАО	45363	55.5
9	ДГП №148 ДЗМ ЮВАО	19068	90.7
10	ДГП №39 ДЗМ САО	9518	77.3
		Среднее значение	87.2

* На основе алгоритма экспертной оценки качества.

Обозначения: ГП – городская поликлиника; ДГП – детская городская поликлиника; ДЗМ – Департамент здравоохранения г. Москвы.

Не соответствовали закону Бенфорда согласно каждому из рассмотренных критериев на основе значений *MAD* данные 10 центров здоровья (табл. 8). Указанная группа центров здоровья отличалась повышенным средним значением процента некорректных данных согласно экспертному алгоритму оценки качества (87.2 %) в сравнении со средним значением для центров здоровья Москвы (61.5 %). При этом, как и в случае табл. 7, процент некорректных данных для одного из десяти центров здоровья (55.5 %) примерно соответствовал общему среднему. На основе значений χ^2 не соответствовали закону Бенфорда согласно каждому из рассмотренных в таблице 5б критериев одновременно данные 44 центров здоровья.

Выраженные различия средних значений процента некорректных данных для подгрупп центров здоровья с наименьшим и наибольшим отклонением данных от закона Бенфорда на основе значений *MAD* в табл. 7 и 8 и значительный уровень взаимной коррелированности величин *MAD* и χ^2 позволяют предположить возможность использования закона Бенфорда для ранжирования центров здоровья по качеству данных. Более подробный анализ приведён в следующем разделе.

Анализ эффективности закона Бенфорда

Наибольшая корреляция величин *MAD* и $\frac{\chi^2}{n}$ с процентом некорректных данных биоимпедансных измерений в центрах здоровья Москвы, согласно алгоритму экспертной оценки качества, наблюдалась для распределения первой значащей цифры десятой степени стандартизованных значений *R* и составила 0.66 и 0.62 соответственно (табл. 9). Анализ соответствующих диаграмм рассеяния выявил наличие выраженного роста минимальных значений процентной доли некорректных данных: почти линейного с увеличением *MAD* и близкого к логарифмическому с увеличением $\frac{\chi^2}{n}$ (рис. 6). Остальные диаграммы рассеяния имели аналогичный вид (данные не показаны).

Таблица 9. Корреляции Спирмена распределений значений *MAD* и $\frac{\chi^2}{n}$, соответственно, с процентом некорректных данных согласно алгоритму экспертной оценки качества для центров здоровья Москвы

Критерий	Параметр*, количество первых значащих цифр					
	***R*, 1**	***X*, 1**	***IR*, 1**	***R*, 2**	***X*, 2**	***IR*, 2**
MAD	0.66	0.59	0.50	0.58	0.50	0.46
$\frac{\chi^2}{n}$	0.62	0.58	0.51	0.56	0.49	0.45

* Использованы десятые степени стандартизованных величин.

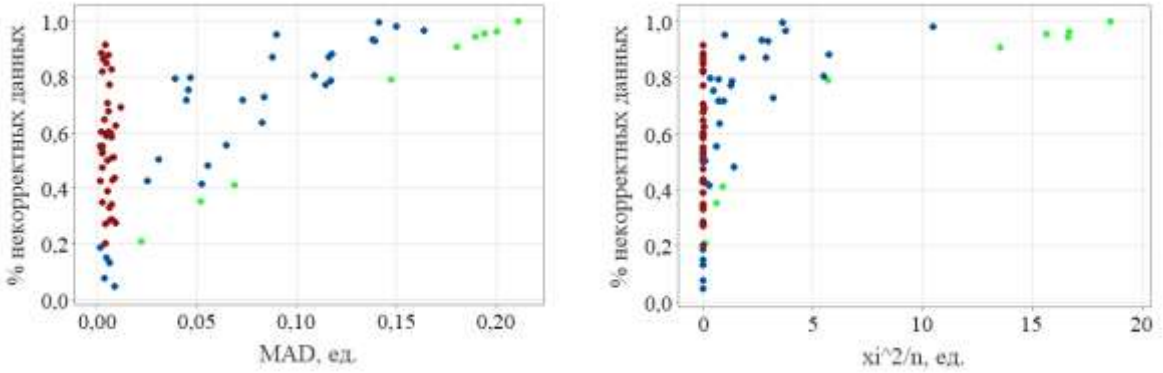


Рис. 6. Диаграмма рассеяния процента некорректных данных биоимпедансометрии в центрах здоровья Москвы в зависимости от величин отклонения от закона Бенфорда *MAD* (слева) и $\frac{\chi^2}{n}$ (справа), рассчитанных на основе распределения первой значащей цифры десятой степени стандартизованных значений *R*. Точки соответствуют центрам здоровья. Цветом выделены отдельные подгруппы центров здоровья (см. ниже)

Полученный результат означает, что Бенфорд-анализ позволяет выявлять некоторые центры здоровья с заведомо высоким процентом некорректных данных биоимпедансных измерений на основе пороговых значений *MAD* и $\frac{\chi^2}{n}$ и, таким образом, даёт возможность строить нижние оценки уровня компрометированности данных. Например, на рисунке 6 видно, что при использовании распределения первой значащей цифры величины R^{10} все центры здоровья при значениях *MAD* и $\frac{\chi^2}{n}$ выше 0.07 и 0.9 соответственно содержали не менее 40 % некорректных данных, а при значениях выше 0.15 и 5.5 – не менее 80 %

некорректных данных, что имеет диагностические значение. При этом каждый из двух указанных наборов критериев для *MAD* и $\frac{\chi^2}{n}$ даёт примерно равное количество центров здоровья с соответствующей долей некорректных данных, что говорит об их хорошей согласованности. Вместе с тем, как следует из рисунка 6, высокий процент некорректных данных наблюдался и для многих центров здоровья с минимальными значениями *MAD* и $\frac{\chi^2}{n}$.

Для выяснения возможной связи вида диаграмм рассеяния на рис. 6 со структурой некорректных данных рассматривались две подгруппы центров здоровья. Первая из них, отмеченная красным цветом, включала те центры здоровья, которые характеризовались тесным или приемлемым соответствием данных закону Бенфорда для первой значащей цифры (значения *MAD* не выше 0.012, см. табл. 2) при доле некорректных данных от 20 % и выше (см. рис. 6 слева). Вторая подгруппа, отмеченная зелёным цветом, включала те центры здоровья, находящиеся на наклонной правой границе облака рассеяния на рис. 6 слева, для которых значения *MAD* не соответствовали закону Бенфорда (значения *MAD* выше 0.015, см. табл. 2). Отметим, что вторая подгруппа центров здоровья характеризовалась пропорциональностью доли некорректных данных и величины *MAD*. Количество центров здоровья в указанных подгруппах составило 39 и 9 соответственно. Можно отметить хорошую согласованность диаграмм рассеяния на рис. 6 слева и справа для величин *MAD* и $\frac{\chi^2}{n}$, включая сходное относительное положение упомянутых подгрупп центров здоровья (они выделены на рис. 6 справа тем же цветом.)

Таблица 10. Структура некорректных данных для центров здоровья, выделенных красным (подгруппа 1) и зелёным (подгруппа 2) цветом на рис. 6 (см. описание в тексте)

Подгруппа центров здоровья	**Выбросы**	**Сфальсифицированные данные***				**Общий процент некорректных данных**	**Среднее значение *MAD***
		1	**2**	**3**	**4**		
1 (n = 39)	6.7	0.2	0.4	40.3	43.4	56.1	0.015
2 (n = 9)	20.1	41.6	18.1	36.2	10.7	72.7	0.130

* Типы сфальсифицированных данных: 1 – измерения калибровочного блока; 2 – программная эмуляция измерений; 3 – чрезмерно короткий интервал времени между измерениями; 4 – многократное измерение одного человека под видом разных.

Для первой из рассматриваемых подгрупп центров здоровья (n = 39) было характерно формирование массива некорректных данных за счёт множественных измерений одного человека под видом разных людей (сфальсифицированные данные 4 типа) при практически полном отсутствии сфальсифицированных данных 1 и 2 типов (измерения калибровочного блока и программная эмуляция измерений, соответственно), табл. 10. Доля выбросов была невелика (6.7 % от общего количества записей), а высокий процент сфальсифицированных данных 3 типа (40.3 %), как уже отмечалось в разделе «Материал и методы», в данном случае объясняет манеру получения данных 4 типа. Указанная подгруппа центров здоровья характеризовалась низкими значениями *MAD* и, таким образом, закон Бенфорда оказался в данном случае неэффективен.

Во второй подгруппе центров здоровья (n = 9) преимущественную часть некорректных данных составляли измерения калибровочного блока (41.6 %), выбросы (20.1 %) и программная эмуляция измерений (18.1 %), а доля сфальсифицированных данных 4 типа, в отличие от первой подгруппы, была сравнительно низкой (10.7 %).

Указанная подгруппа центров здоровья характеризовалась высокими значениями *MAD* (см. табл. 10), и, следовательно, применение закона Бенфорда оказалось в данном случае эффективным.

ОБСУЖДЕНИЕ

Ввиду массовости, широкого географического охвата и применения инструментальных методов диагностики данные профилактического скрининга в российских центрах здоровья представляют значительный интерес для анализа здоровья населения [19, 21, 23]. Особенностями таких данных являются отсутствие возможности их оперативной ручной обработки и сильная зашумлённость искусственно сгенерированными и сфальсифицированными данными [22, 23, 31], что затрудняет возможность их последующего использования в медико-биологических исследованиях. Поэтому предобработка данных профилактического скрининга должна включать применение эффективных методов и алгоритмов поиска и удаления (фильтрации) некорректных данных. Эта функция была реализована нами в специализированном программном комплексе HCViewer [22, 24] на основе рассмотренного выше алгоритма экспертной оценки качества и применялась при обработке данных биоимпедансных измерений в центрах здоровья России за 2010–2012 и 2010–2015 годы [22, 23]. Были получены оценки диагностической эффективности указанного алгоритма: его чувствительность на тестовых наборах данных составила 95.2–98.9 %, а специфичность – 94.5–99.2 % [22].

Применение алгоритма экспертной оценки качества данных к обновлённому массиву данных биоимпедансных измерений в центрах здоровья Москвы за 2010–2019 годы выявило высокий процент некорректных данных (66.5 %), что в 1,5 раза превышало общероссийский уровень в данных за 2010–2015 годы (44.8 %) [22] и на порядок превосходит среднемировую оценку потерь от мошенничества в сфере здравоохранения от общих расходов на здравоохранение (6 %) [26], что свидетельствует о недостаточности принимаемых мер контроля. В сравнении с общероссийскими данными центров здоровья за 2010–2015 годы некорректные данные были распределены по центрам здоровья Москвы более равномерно, с нарушением закона Парето («закон 20/80»), так что 80 % некорректных данных были сгенерированы в 43 % центров здоровья Москвы.

В отличие от алгоритма экспертной оценки качества, Бенфорд-анализ представляет собой групповой метод анализа данных. Его преимущество состоит в относительной простоте реализации и возможности применения к любым наборам числовых данных независимо от наличия или отсутствия критериев экспертной оценки качества. Элементом оценки здесь являются целые совокупности данных, а не индивидуальные записи измерений. В качестве структурной единицы Бенфорд-анализа в настоящей работе рассматривался центр здоровья. Каждому центру здоровья ставилось в соответствие некоторое число (значение *MAD* или $\frac{\chi^2}{n}$, характеризующее отклонение от закона Бенфорда), что служило основой для их ранжирования. (По аналогии с этим уместно рассмотреть задачу динамической оценки качества данных биоимпедансометрии в каждом отдельно взятом центре здоровья на основе регулярно проводимого Бенфорд-анализа.)

Сравнение результатов применения алгоритма экспертной оценки качества данных биоимпедансных измерений и пакета benford.analysis [28], встроенного в программный комплекс HCViewer [24], показало возможность использования Бенфорд-анализа для надёжного выявления некоторых типов некорректных данных, таких как программная эмуляция измерений, измерения шаблона и выбросы. Этот результат не зависел от выбора параметра биоимпедансометрии (десятые степени стандартизованных величин

R, *X* или *IR*) и количества рассматриваемых первых значащих цифр (одна или две). При этом несколько выше была значимая корреляция между процентом некорректных данных согласно алгоритму экспертной оценки качества и величиной *MAD*, оцениваемой на основе параметра *R* и одной первой значащей цифры ($\rho = 0.66$ при общем диапазоне значений 0.46–0.66, табл. 9). Аналогичный результат был получен и для величины $\frac{\chi^2}{n}$ ($\rho = 0.62$ при общем диапазоне значений 0.45–0.62, табл. 9). При относительно малой процентной доле некорректных данных любого типа Бенфорд-анализ не давал непропорционально высоких значений *MAD* и $\frac{\chi^2}{n}$ (рис. 6). Это указывает на возможность использования отклонений от закона Бенфорда в качестве достаточного (но не необходимого) условия компрометированности данных, и, кроме того, на адекватность и относительную полноту используемых критериев экспертной оценки качества данных. Для центров здоровья, в которых основную часть некорректных данных составляли многократные измерения одного человека под видом разных, данные хорошо соответствовали закону Бенфорда и, таким образом, в этом случае Бенфорд-анализ был неэффективен. В остальных случаях использование закона Бенфорда позволяло эффективно ранжировать центры здоровья по уровню компрометированности данных.

Одна из возможностей использования Бенфорд-анализа заключается в применении гибридных алгоритмов фильтрации данных биоимпедансных измерений. Например, сначала на его основе можно исключить центры здоровья с заведомо высоким уровнем измерений калибровочного блока, программных эмулятов измерений и выбросов, а к остальным данным применить алгоритм экспертной оценки качества. Перспективным направлением дальнейшего развития работы является применение методов машинного обучения, таких как нейронные сети и искусственные иммунные системы, для выявления аномалий в данных профилактического скрининга с последующей разработкой и внедрением системы динамического мониторинга качества данных.

Полученный в результате фильтрации массив данных биоимпедансных измерений центров здоровья Москвы за 2010–2019 годы, содержащий 456.014 записей, будет использован для характеристики физического развития населения Москвы и оценки рисков хронических заболеваний.

Проведённый ретроспективный анализ затронул только данные биоимпедансных измерений. Полученные данные были неполны, так как необходимая информация в соответствии с запросами предоставлялась не всегда и, кроме того, некоторые центры здоровья сообщали на этапах сбора данных о полной или частичной утрате информации ввиду поломки оборудования. Первичные данные измерений в центрах здоровья Москвы с использованием других инструментальных методов в настоящее время централизованно не собираются. Для обеспечения возможности оперативного сбора и обработки данных, динамического контроля качества данных профилактического скрининга представляет интерес подключение центров здоровья Москвы к Единой медицинской информационно-аналитической системе (ЕМИАС).

ВЫВОДЫ

Анализ данных биоимпедансных измерений в центрах здоровья Москвы за 2010–2019 годы выявил высокий процент некорректных данных согласно алгоритму экспертной оценки качества (66.5 %), что свидетельствует о низкой эффективности применяемых мер контроля. В структуре некорректных данных биоимпедансометрии преобладали сфальсифицированные данные.

Установлена значимая корреляция между отклонением данных от закона Бенфорда и процентом некорректных данных согласно алгоритму экспертной оценки качества.

Отклонение данных биоимпедансных измерений в центрах здоровья от закона Бенфорда является достаточным условием их компрометированности.

Работа выполнена в ФГБУ «ЦНИИОИЗ» Минздрава России при поддержке Российского научного фонда (грант № 20-15-00386, рук. В.И. Стародубов).

СПИСОК ЛИТЕРАТУРЫ

1. Benford F. The law of anomalous numbers. *Proc. Am. Phil. Soc*. 1938. V. 78. № 4. P. 551–572.
2. Durtschi C., Hillison W., Pacini C. The effective use of Benford's law to assist in detecting fraud in accounting data. *J. Forensic Accounting*. 2004. V. 5. № 1. P. 17–33.
3. Mebane W.R. Jr. Election forensics: vote counts and Benford's law. URL: https://www-personal.umich.edu/~wmebane/pm06.pdf (дата обращения: 02.11.2022).
4. Khosravani A., Rasinariu C. Emergence of Benford's law in music. *ArXiv*: 1805.06506 [physics.soc-ph]. 2018. URL: https://arxiv.org/abs/1805.06506 (дата обращения: 11.10.2022).
5. Coeurjolly J.-F. Digit analysis for Covid-19 reported data. *ArXiv*: 2005.05009 [stat.AP]. 2020. URL: https://arxiv.org/pdf/2005.05009.pdf (дата обращения: 11.10.2022).
6. Newcomb S. Note on the frequency of use of different digits in natural numbers. *Am. J. Math*. 1881. V. 4. № 1. P. 39–40.
7. Franel J. A propos des tables de logarithmes. *Vjschr. Naturf. Ges. Zurich*. 1917. V. 62. № 1–2. P. 286–295.
8. Boring E.G. The logic of normal law of error in mental measurement. *Am. J. Psychology*. 1920. V. 31. № 1. P. 1–30.
9. Hill T.P. A statistical derivation of the significant-digit law. *Statist. Sci*. 1995. V. 10. № 4. P. 354–363. doi: 10.1214/ss/1177009869
10. Nigrini M.J. *Benford's law: application for forensic accounting, auditing and fraud detection*. Wiley and Sons: New Jersey, 2012. 352 p.
11. Berger A., Hill T.P., Rogers E. *Benford online bibliography*. *2009-2022*. URL: http://www.benfordonline.net (дата обращения: 11.10.2022).
12. Berger A., Hill T.P. A basic theory of Benford's law. *Probab. Surveys*. 2011. V. 8. P. 1–126. doi: 10.1214/11-PS175
13. Чучалин А.Г. Профилактика и контроль хронических неинфекционных заболеваний. *Пульмонология*. 2009. № 1. С. 5–10.
14. Кобякова О.С., Куликов Е.С., Малых Р.Д., Черногорюк Г.Э., Деев И.А., Старовойтова Е.А., Кириллова Н.А., Загромова Т.А., Балаганская М.А. Стратегии профилактики хронических неинфекционных заболеваний: современный взгляд на проблему. *Кардиоваскулярная терапия и профилактика*. 2020. Т. 18. № 4. С. 92–98. doi: 10.15829/1728-8800-2019-4-92-98
15. NCD Risk Factor Collaboration. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. *Lancet*. 2017. V. 390. № 10113. P. 2627–2642. doi: 10.1016/S0140-6736(17)32129-3
16. Silverio R., Goncalves D.C., Andrade M.F., Seelaender M. Coronavirus disease 2019 (COVID-19) and nutritional status: the missing link? *Adv. Nutr*. 2021. V. 12. № 3. P. 682–692. doi: 10.1093/advances/nmaa125
17. Heymsfield S.B., Lohman T.G., Wang Z., Going S.B. (eds.) *Human body composition*. 2nd ed. Champaign, IL: Human Kinetics, 2005. 533 p.
18. Николаев Д.В., Смирнов А.В., Бобринская И.Г., Руднев С.Г. *Биоимпедансный анализ состава тела человека*. М.: Наука, 2009. 392 с.

19. Погосова Н.В., Вергазова Э.К., Аушева А.К., Суворов С.С., Бойцов С.А. Центры здоровья: достигнутые результаты и перспективы. *Профилактическая медицина*. 2014. Т. 17. № 4. С. 16–24.
20. Кривонос О.В., Бойцов С.А., Погосова Н.В., Юферева Ю.М., Янушевич О.О., Кузьмина Э.М., Нероев В.В., Тутельян В.А., Батурин А.К., Погожева А.В., Брюн Е.А. *Оказание медицинской помощи взрослому населению в центрах здоровья: методические рекомендации*. Москва, 2012. 110 с.
21. Стародубов В.И., Руднев С.Г., Николаев Д.В., Коростылёв К.А. Федеральный информационный ресурс центров здоровья: современное состояние и перспективы развития. *Социальные аспекты здоровья населения*. 2015. 45 (5). URL: http://vestnik.mednet.ru/content/view/706/30/lang,ru/ (дата обращения: 11.10.2022).
22. Starunova O.A., Rudnev S.G., Starodubov V.I. HCViewer: software and technology for quality control and processing raw mass data of preventive screening. *Russ. J. Numer. Anal. Math. Model*. 2017. V. 32. № 5. P. 315–326. doi: 10.1515/rnam-2017-0030
23. Руднев С.Г., Соболева Н.П., Стерликов С.А., Николаев Д.В., Старунова О.А., Черных С.П., Ерюкова Т.А., Колесников В.А., Мельниченко О.А., Пономарёва Е.Г. *Биоимпедансное исследование состава тела населения России*. М.: РИО ЦНИИОИЗ, 2014. 493 с.
24. Старунова О.А., Руднев С.Г., Стародубов В.И. *HCViewer: программа для автоматизированного анализа качества, фильтрации и обработки массовых данных профилактического скрининга в центрах здоровья*. Свидетельство о гос. регистрации программы для ЭВМ № 2020665580 от 27.11.2020 г.
25. Mikkers M., Sauter W., Vincke P., Boertjens J. *Healthcare fraud, corruption and waste in Europe: national and academic perspectives*. The Hague: Eleven International Publishing, 2017. 336 p.
26. Global Health Care Anti-Fraud Network (2022). URL: http://www.ghcan.org (дата обращения: 11.10.2022).
27. Федеральная служба по надзору в сфере здравоохранения. Система оценки результативности и эффективности контрольно-надзорной деятельности. URL: https://roszdravnadzor.gov.ru/reform/effectiveness (дата обращения: 11.10.2022).
28. Cinelli C. Package «benford.analysis». Benford analysis for data validation and forensic analytics. Version 0.1.5. December 21, 2018. URL: https://cran.r-project.org/web/packages/benford.analysis/benford.analysis.pdf (дата обращения: 11.10.2022).
29. UK Biobank. Data-Field 23106. Impedance of whole body. Data. 2022. URL: https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=23106 (дата обращения: 11.10.2022).
30. Morrow J. *Benford's law, families of distributions and a test basis.* Centre for Economic Performance, London School of Economics and Political Science, 2014. 29 p. URL: http://cep.lse.ac.uk/pubs/download/dp1291.pdf (дата обращения: 11.10.2022).
31. Стародубов В.И., Руднев С.Г., Николаев Д.В., Коростылёв К.А. О качестве данных профилактического скрининга в центрах здоровья и способе повышения эффективности бюджетных расходов. *Аналитический вестник Совета Федерации ФС РФ*. 2015. Т. 44. № 597. С. 43–49.

Рукопись поступила в редакцию 31.10.2021, переработанный вариант поступил 19.10.2022.
Дата опубликования 05.11.2022.

Application of Benford's Law for Quality Assessment of Preventive Screening Data

Starunova O.A., Rudnev S.G., Ivanova A.E., Semenova V.G., Starodubov V.I.

Russian Research Institute of Health, Moscow, Russia

Abstract. An empirical Benford's law which describes the probability of the appearance of certain first significant digits in many distributions taken from real life, is used to identify anomalies in various kinds of data. Our aim was to test Benford's law to assess the quality of mass preventive screening data on the example of bioelectrical impedance analysis (BIA) data from Moscow health centers. As was shown earlier, such a data is characterized by a high level of contamination by artificially generated and falsified data. A generated 2010–2019 database of BIA measurements contained 1361019 measurement records in the age range of the examined persons from 5 to 96 years. Application of the expert quality assessment algorithm, which was used as a reference for evaluation of the effectiveness of Benford analysis, revealed a high percentage of incorrect data (66.5 %) which was dominated by falsified data. To characterize the degree of the data compliance with Benford's law, the mean absolute deviations of the frequency distributions of the first and first two significant digits deviations from the proper values and chi-squared statistics for the tenth powers of the standardized resistance, reactance, and resistance index values were assessed for each health center. A significant correlation was observed between the data deviation from Benford's law and the percentage of incorrect data as provided by the expert quality assessment algorithm (ρ_{max} = 0.66 and 0.62 for the mean absolute deviations and χ^2 statistics, respectively, based on the resistance value and the first significant digit). It is suggested that deviation of the BIA data from Benford's law serves as a sufficient, but not a necessary, condition for their contamination. For those health centers, in which most of the incorrect data were represented by multiple measurements of the same person under the guise of different ones, the data were in good agreement with Benford's law. If the structure of incorrect data was dominated by measurements of the calibration block, software emulations of BIA measurements and outliers, then the use of Benford's law made it possible to effectively rank health centers by the level of data authenticity.

Key words: *health centers, preventive screening, big data, bioelectrical impedance analysis, data quality, expert quality assessment algorithm, Benford's law.*