

Применение закона Бенфорда для оценки качества данных профилактического скрининга в центрах здоровья

Старунова О.А.¹, Руднев С.Г.^{2,1}, Иванова А.Е.¹, Семенова В.Г.¹, Стародубов В.И.¹

¹ *Центральный научно-исследовательский институт
организации и информатизации здравоохранения Минздрава России, Москва*

² *ИБМ им. Г.И. Марчука РАН, Москва*

ХIII конференция «Математические модели и численные методы в биологии и медицине»
2-3 ноября 2021 г., ИБМ РАН, Сеченовский университет, Москва

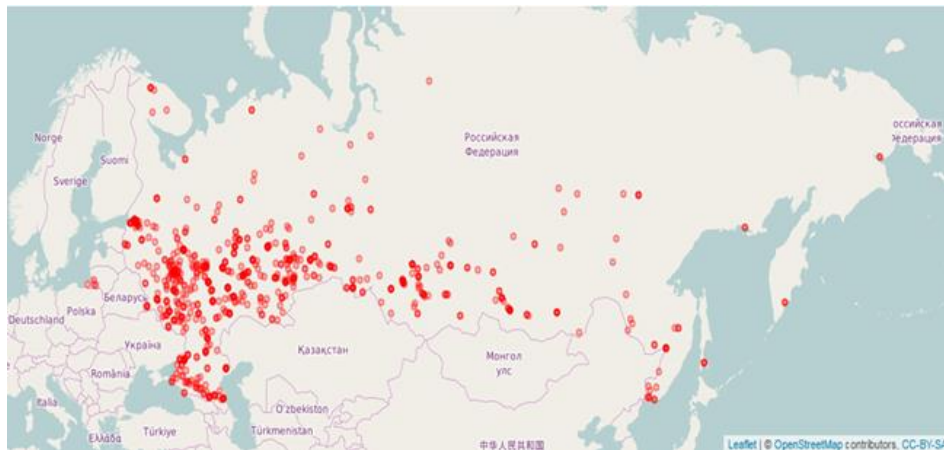
Структура доклада

1. Введение, цель исследования
2. Закон Бенфорда
3. Структура данных биоимпедансных измерений
4. Алгоритм экспертной оценки качества и результаты его применения
5. Применение закона Бенфорда, сравнение методов
6. Обсуждение и выводы

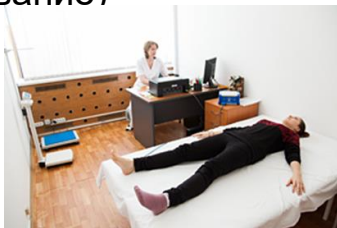
Национальная сеть центров здоровья

Методы диагностики, применяемые в центрах здоровья:

- Ангиологический скрининг (плече-лодыжечный индекс)
- Антропометрия (МТ, ДТ, ОТ, ОБ), калиперометрия, кистевая динамометрия
- Биоимпедансный анализ состава тела
- Биохимический анализ крови (холестерин, глюкоза, креатинин и др.)
- Кардиологический скрининг (дисперсионное картирование)
- Пульсоксиметрия
- САД, ДАД
- Спирометрия
- Тест на наличие курения
- и другие



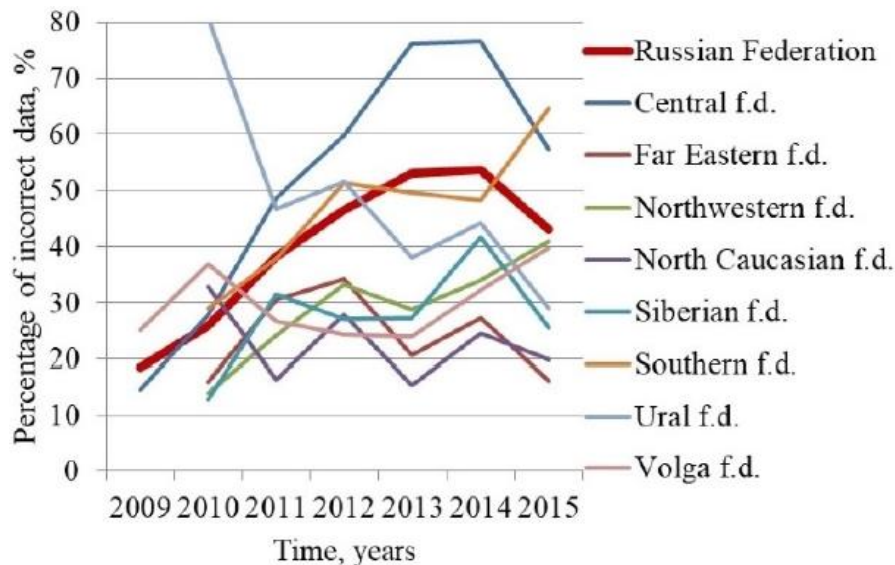
Географическое распределение центров здоровья



Национальная сеть центров здоровья

Проблема качества данных

- Сильная зашумленность искусственно сгенерированными и поддельными данными
- Необходимость создания автоматизированных программных комплексов для обработки данных профилактического скрининга (решение - **HCViewer**)
- Сравнительно редкая доступность алгоритмов экспертной оценки качества



Динамика процентной доли некорректных данных биоимпедансных измерений в центрах здоровья России, 2010-2015 гг. (n=2.294.843)
Starunova et al., 2017

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

Закон Бенфорда

Назовем множество положительных чисел *удовлетворяющим закону Бенфорда*, если вероятность того, что случайно выбранное число из данного множества начинается с цифры d , равна $\lg(1+1/d)$.



a



б

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

Общая формулировка закона Бенфорда для первых k значащих цифр:

Вероятность того, что мантисса случайного выбранного числа в числовых данных начинается с цифр $d_1, d_2, \dots, d_k$, равна

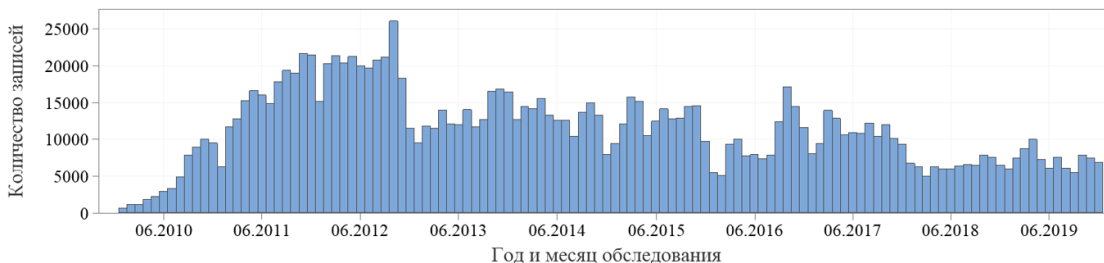
$$P(D_1 = d_1, \dots, D_k = d_k) = \log_{10} \left(1 + \frac{1}{\sum_{i=1}^k d_i \cdot 10^{k-i}} \right), d_1 \in \{1, \dots, 9\}, d_{i \neq 1} \in \{0, \dots, 9\},$$

База данных биоимпедансных измерений

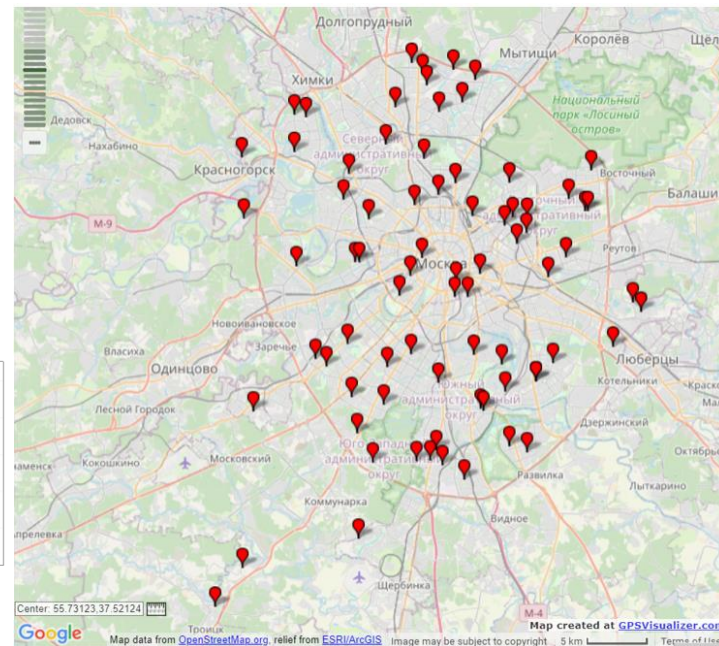


Анализатор состава тела ABC-01 «Медасс»

- Данные биоимпедансных измерений в центрах здоровья Москвы за 2010-2019 гг. из 79 центров здоровья ($n=1.362.333$)
- Измерения анализаторами состава тела ABC-01 «Медасс»
- Объединение данных из нескольких источников

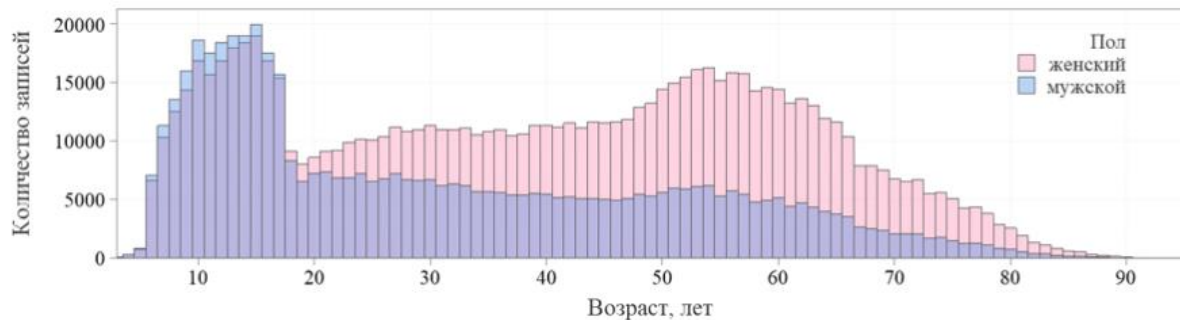


Гистограмма распределения общего количества записей в базе данных ($n=1.362.333$) по году и месяцу обследования

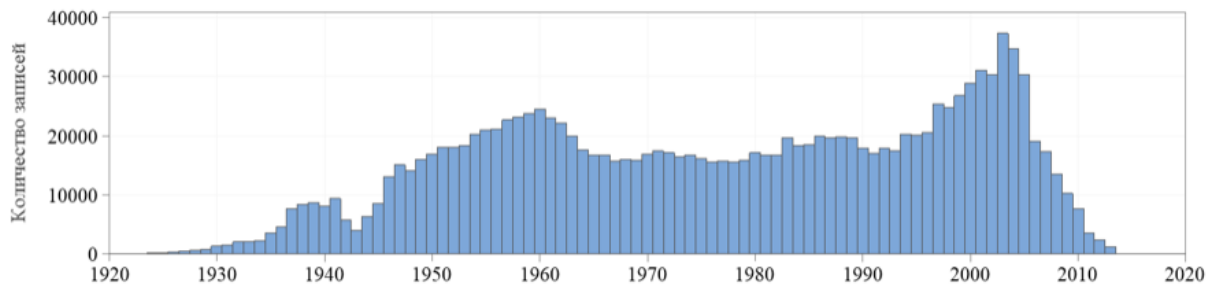


Географическое положение центров здоровья Москвы, представивших данные биоимпедансных измерений

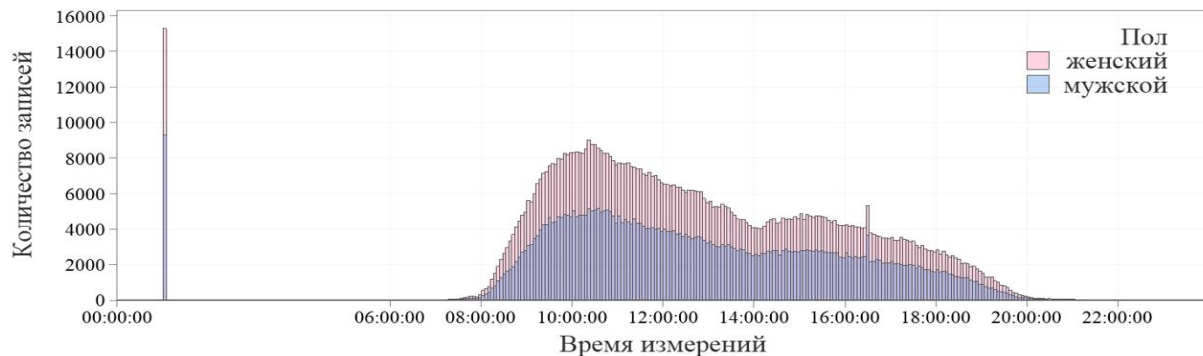
Характеристика данных



Половозрастная структура обследованных



Распределение данных по году рождения обследованных



Распределение данных по времени суток обследования

Применение алгоритма экспертной оценки качества

Структура некорректных данных

1. Выбросы

Рост, см	Вес, кг	ИМТ, кг/м ²	ЖМ, кг	%ЖМ	R50, Ом	X50, Ом	Фазовый угол, град.
130-210	35-150	12-55	0,5-75	0-55	250-1000	20-150	3,0-10,2

Интервалы допустимых значений параметров биоимпедансометрии для взрослых людей, использованные для обнаружения «выбросов»

1. Поддельные данные

- Измерение шаблона
- Программная эмуляция данных
- Нереалистично короткий интервал времени между измерениями
- Многократные измерения одного человека под видом разных

Критерии выявления поддельных данных

Поддельными считали те записи в базе данных, для которых измерение человека либо не проводилось вовсе, либо они соответствовали многократному измерению одного человека под видом разных. Указанные типы поддельных данных определялись следующим образом:

1) **Измерение шаблона** (эквивалентной электрической схемы биообъекта) на частоте 50 кГц, как правило, дает значения R50 и X50 в интервале от 381 до 391 Ом и от 38 до 48 Ом, соответственно.

2) **Программная эмуляция (имитация) измерения**, обычно применяемая в центрах здоровья для настройки программного обеспечения анализатора, определялась при значениях R50 = 444, 444.4, 555.5 и 556 Ом.

3) **Нереалистично короткий интервал времени между измерениями**. Интервал времени между двумя соседними измерениями разных людей в центре здоровья составляет менее 90 секунд.

4) **Измерение одного человека под видом разных**. Записи двух соседних обследований в базе данных отличаются менее чем на 1% по величине R50 и, одновременно, менее чем на 7% по величине X50.

Качество данных биоимпедансных измерений в центрах здоровья 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	178577	62,6	197108	69,1
ЗАО	163766	26567	16,2	38324	23,4	1562	1,0	90874	55,5	115046	70,3	124704	76,2
ЗелАО	66575	7358	11,1	33307	50,0	512	0,8	32830	49,3	37741	56,7	39967	60,0
САО	52154	10760	20,6	8563	16,4	22075	42,3	22241	42,6	38252	73,3	40911	78,4
СВАО	141682	26217	18,5	10225	7,2	329	0,2	62902	44,4	70688	49,9	83257	58,8
СЗАО	72795	45824	63,0	16744	23,0	2293	3,2	40494	55,6	54067	74,3	56816	78,1
ТпНАО	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	46896	53,1	64212	72,7
ЮАО	222629	13013	5,9	204	0,1	103	0,1	120498	54,1	118865	53,4	143314	64,4
ЮВАО	114375	35554	31,1	39568	35,0	17968	15,7	32728	28,6	74329	65,0	80285	70,2
ЮЗАО	152523	12087	7,9	21932	14,4	103	0,1	51683	33,9	51809	34,0	73818	48,4
Итого	1361019	242438	17,8	248016	18,2	45990	3,4	601693	44,2	786831	57,8	905005	66,5

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

Зависимость процентной доли некорректных данных от количества обследований в центре здоровья



Точки на рисунке соответствуют центрам здоровья, а цвет маркера – административному округу Москвы. Размеры точек пропорциональны доле подделок.

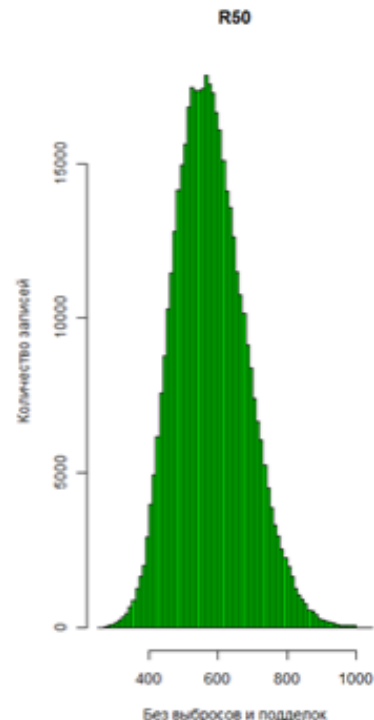
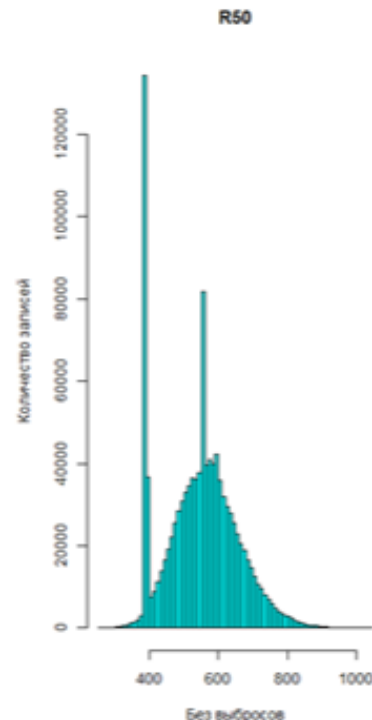
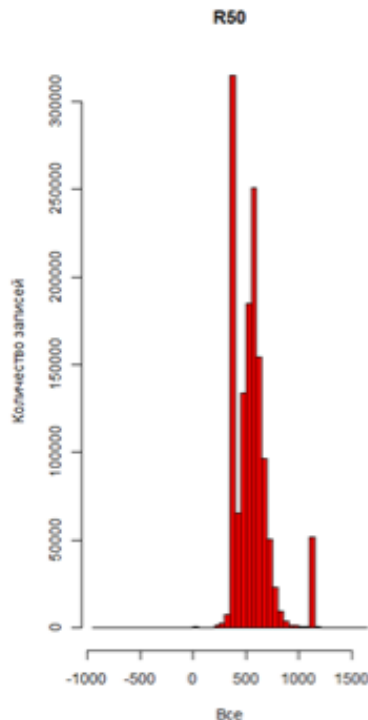
Динамика процентной доли некорректных данных биоимпедансных измерений в центрах здоровья Москвы (на основе алгоритма экспертной оценки качества)



Последовательное исключение некорректных данных на основе алгоритма экспертной оценки качества

Распределение значений R50:

- слева – начальное;
- в центре – после удаления «выбросов»;
- справа – после удаления «выбросов» и поддельных данных.



Применение закона Бенфорда

Характеристика отклонения от закона Бенфорда - средняя абсолютная ошибка (CO):

$$CO = \sum_{i=10}^{99} |b_i - x_i|/90,$$

где b_i – теоретическое значение вероятности появления числа i в качестве первых двух значащих цифр случайной величины, удовлетворяющей закону Бенфорда, а x_i – фактическая частота для тестируемой выборки.

Классификация значений CO (Cinelli, 2018)

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

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>

Распределение центров здоровья Москвы по качеству данных биоимпедансных измерений в соответствии со значением СО

Количество первых значащих цифр	Параметр*	Соответствие закону Бенфорда			
		Тесное	Приемлемое	Минимальное	Несоответствие
1	R50	13	14	2	48
1	X50	23	17	2	35
1	DT ² /R50	42	18	7	10
2	R50	8	13	6	50
2	X50	17	13	8	39
2	DT ² /R50	36	6	5	30

Корреляции Спирмена распределений значений СО, характеризующих отклонение от закона Бенфорда, для центров здоровья Москвы

Параметр*, количество первых значащих цифр	R50, 1	X50, 1	DT ² /R50, 1	R50, 2	X50, 2
X50, 1	0,89	1			
DT ² /R50, 1	0,83	0,77	1		
R50, 2	0,95	0,88	0,83	1	
X50, 2	0,83	0,91	0,75	0,90	1
DT ² /R50, 2	0,86	0,85	0,87	0,90	0,88

* Использованы десятичные степени указанных величин.

Зависимость процентной доли некорректных данных от величины средней абсолютной ошибки (СО)

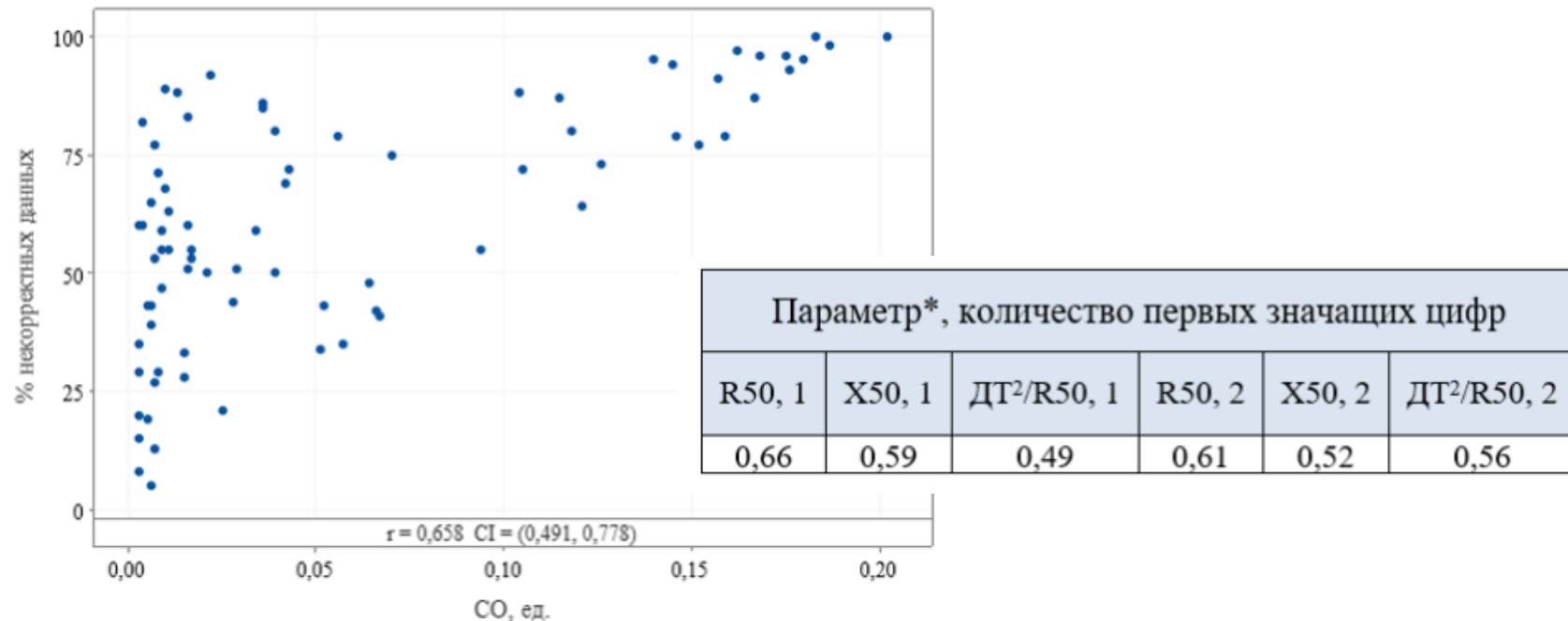
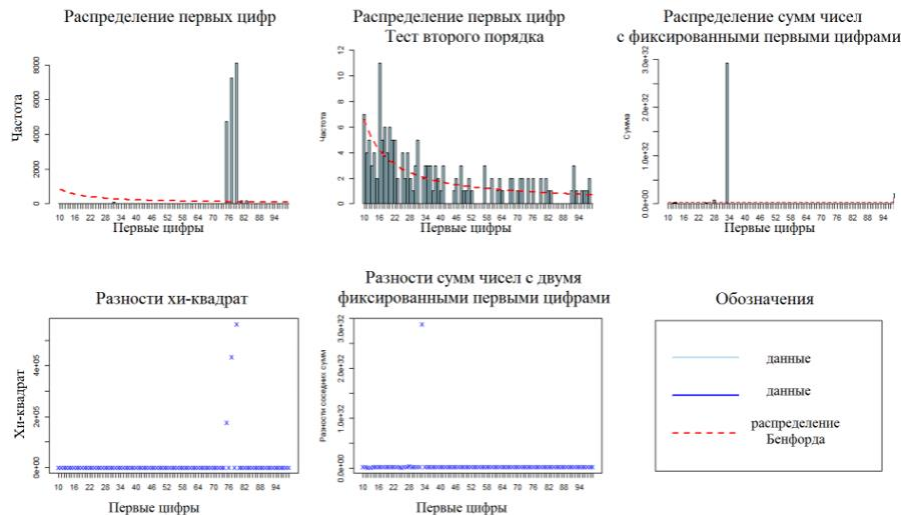
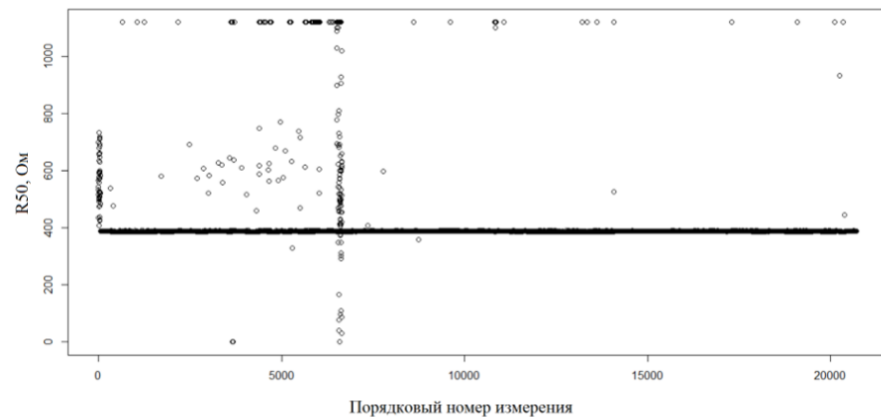


Диаграмма рассеяния процентной доли некорректных данных биоимпедансометрии в центрах здоровья Москвы в зависимости от величины СО, рассчитанной на основе распределения первых значащих цифр величины R50¹⁰. Точки соответствуют центрам здоровья. Приведены корреляция Спирмена и 95%-ный доверительный интервал.

Пример Бенфорд-анализа данных (1)



а)

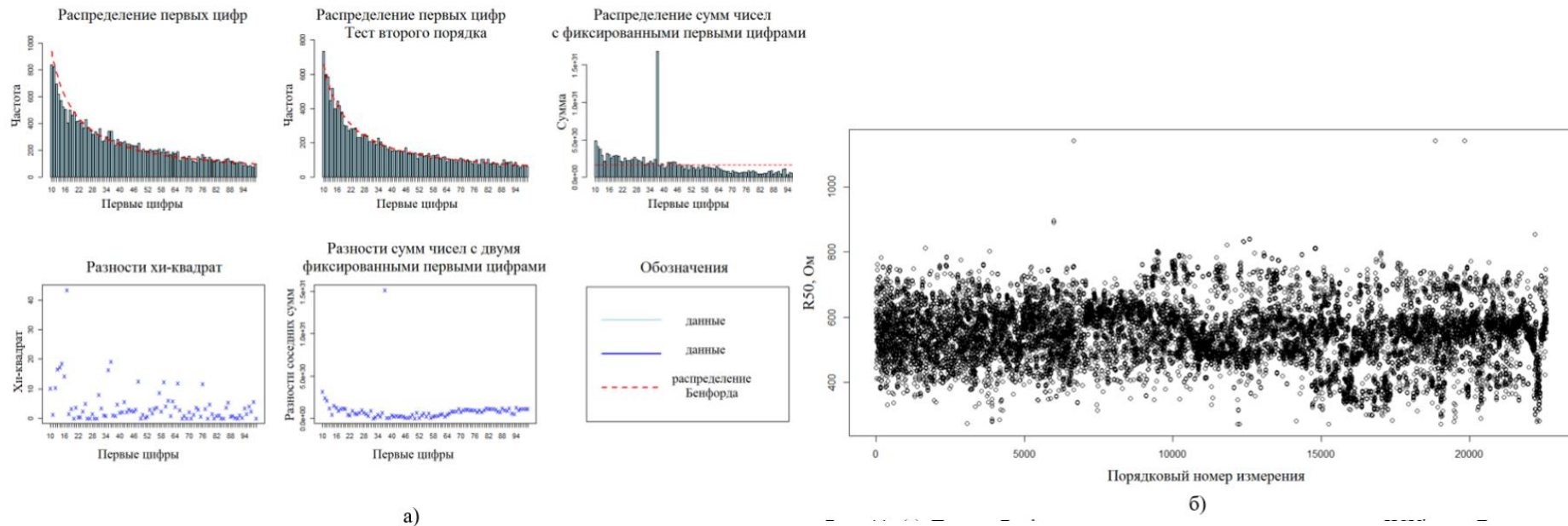


б)

(а) Пример Бенфорд-анализа с использованием программы HCViewer. Данные для рассматриваемого центра здоровья Москвы (ГП №11 ДЗМ филиал №1 ЮЗАО) не соответствуют закону Бенфорда ($CO=0,020$).

(б) Распределение значений параметра R_{50} в том же центре здоровья (измерения упорядочены по времени). На основании экспертного алгоритма оценки качества 98,3% данных классифицированы как некорректные.

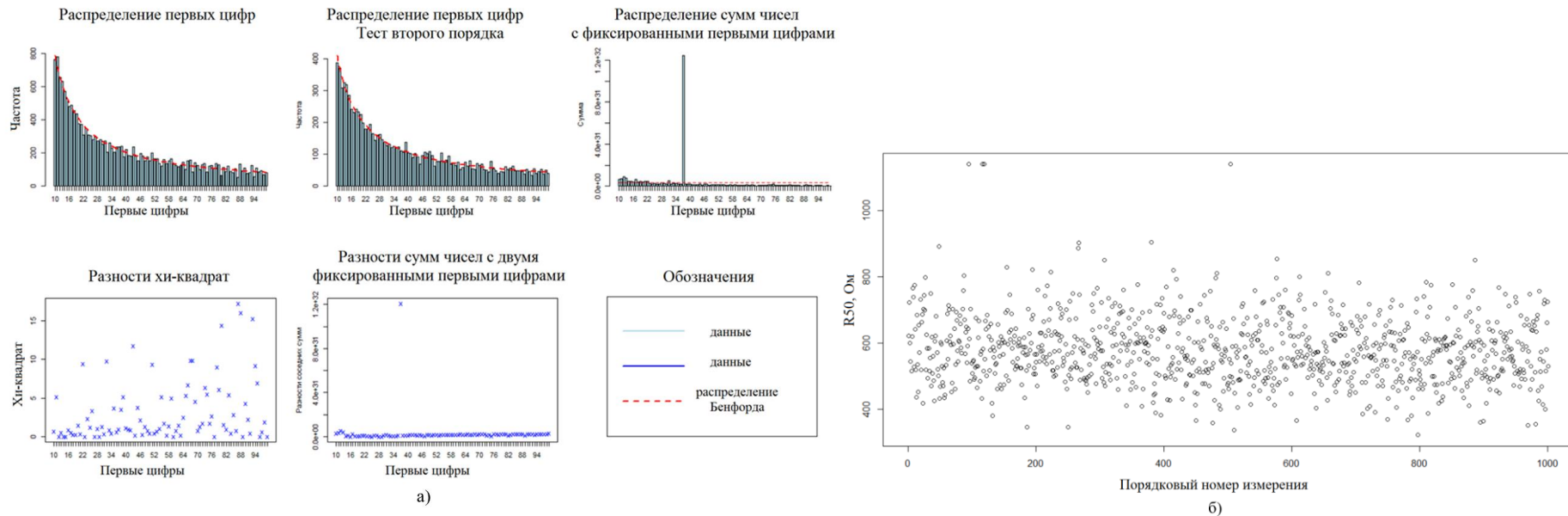
Пример Бенфорд-анализа данных (2)



(а) Пример Бенфорд-анализа с использованием программы HCViewer. Данные для рассматриваемого центра здоровья Москвы (ГП №8 ДЗМ филиал №3 ЗАО) тесно соответствуют закону Бенфорда ($CO=0,0012$).

(б) Распределение значений параметра R50 в том же центре здоровья (измерения упорядочены по времени). На основании экспертного алгоритма оценки качества 88,5% данных классифицированы как некорректные.

Пример Бенфорд-анализа данных (3)



(а) Пример Бенфорд-анализа с использованием программы HCVIEWER. Данные для рассматриваемого центра здоровья Москвы (ГП №107 ДЗМ СВАО) тесно соответствуют закону Бенфорда ($CO=0,001$). (б) Распределение значений параметра R50 в том же центре здоровья (измерения упорядочены по времени). На основании экспертного алгоритма оценки качества только 7,7% данных классифицированы как некорректные.

Выводы

1. Анализ данных биоимпедансных измерений в центрах здоровья Москвы за 2010-2019 годы выявил высокий процент некорректных данных согласно экспертному алгоритму оценки качества (66,5%), что свидетельствует о низкой эффективности применяемых мер контроля. В структуре некорректных данных биоимпедансометрии преобладали сфальсифицированные данные.
2. Установлена значимая корреляция между отклонением данных от закона Бенфорда и процентом некорректных данных согласно экспертному алгоритму оценки качества. Отклонение данных биоимпедансных измерений в центрах здоровья от закона Бенфорда является достаточным условием их компрометированности.

Публикации по теме исследования

1. Старунова О.А., Руднев С.Г., Иванова А.Е., Семенова В.Г., Стародубов В.И. (2021). Применение закона Бенфорда для анализа качества данных профилактического скрининга // Математическая биология и биоинформатика (представлено к публ.)
2. Старунова О.А., Руднев С.Г., Стародубов В.И. (2020). HCViewer: программа для автоматизированного анализа качества, фильтрации и обработки массовых данных профилактического скрининга в центрах здоровья. Свидетельство о государственной регистрации программы для ЭВМ №2020665580, дата регистрации 27.11.2020 г.
3. Starunova O.A., Rudnev S.G., Starodubov V.I. (2017). HCViewer: software and technology for quality control and processing raw mass data of preventive screening // Russ. J. Numer. Anal. Math. Model. V.32, N5. P.315-326.
4. Стародубов В.И., Руднев С.Г., Николаев Д.В., Коростылев К.А. (2015). О качестве данных профилактического скрининга в центрах здоровья и способе повышения эффективности бюджетных расходов // Аналитический вестник Совета Федерации ФС РФ. №44(597). С.43-49.

Спасибо за внимание!

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