# Benford's Law and Health Statistics: Application to Russian Preventive Screening Data

Olga Starunova[1], Sergey Rudnev[2,1]

[1]Federal Research Institute for Health Organization and Informatics, Moscow, Russia

[2]Institute of Numerical Mathematics, Moscow, Russia

**BIOMAT 2017 Oct. 29 – Nov. 03 INM RAS Moscow, Russia**

# Outline

- Background, data source
- Evidence for the presence of incorrect data
- Empirical selection criteria
- Classification and statistics of the incorrect data
- Motivation of using Benford’s law
- Software for data filtering and analysis
- Benford analysis: some results
- Summary and prospects

# Healthcare fraud

- Fraud = criminal deception, the use of false representations to gain an unjust advantage (*Concise Oxford Dictionary*)
- Healthcare system: major expenditure in most countries
- Healthcare expenditure, % of GDP (as of 2014):
  - Brazil: 8.3%
  - Russia: 7.1%
  - India: 4.7%
  - China: 5.5%
  - USA: 17.1%
- Estimated global loss due to healthcare fraud: 6% of the global health care spending (GHCAN, 2017)
- Requires big data analytics
- In Russia: largely unrecognized problem



Bolton R J, Hand D J (2002) Statistical fraud detection. Stat. Sci. 17(3):235-255.

WHO GHE database (2017) Health expenditure, total (% of GDP). Available at: https://data.worldbank.org/indicator/SH.XPD.TOTL.ZS

Global Health Care Antifraud Network (2017) The health care fraud challenge: worldwide, health care fraud is a lucrative, though illicit, line of work. Available at: http://www.ghcan.org

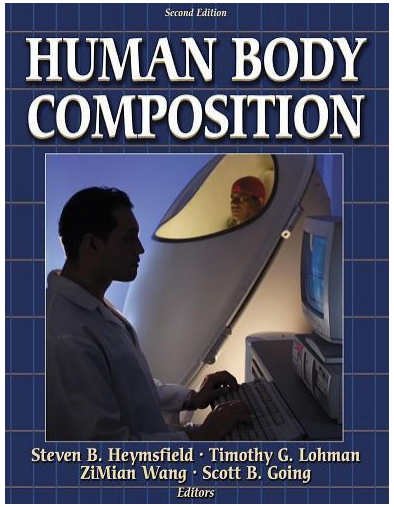
# Human body composition research

| | | | |
|---|---|---|---|
| N, Ca, P, K, Na, Cl | Lipid | Adipocytes | Adipose Tissue |
| H | Water | Cells | Skeletal Muscle |
| C | | | |
| O | Proteins | Extracellular Fluid | Visceral Organs & Residual |
| | Glycogen | | |
| | Minerals | Extracellular Solids | Skeleton |
| *Atomic* | *Molecular* | *Cellular* | *Tissue-System* |

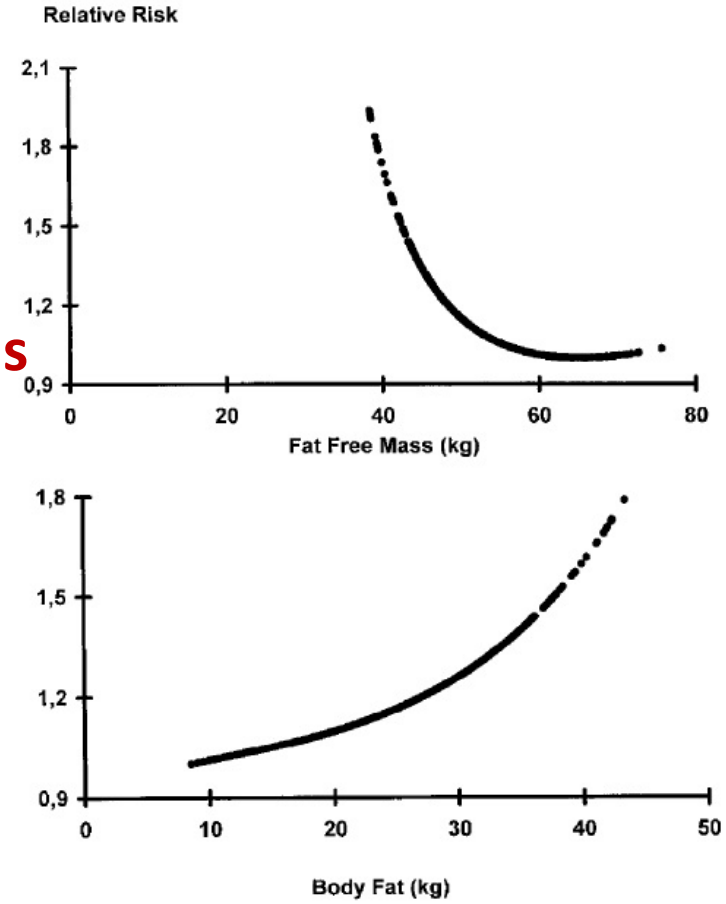
Some body composition models at various structural levels of the organism (Heymsfield, Wang, et al., 1997)

**Measurement methods:**

- Anthropometry
- **Bioimpedance analysis**
- HW
- ADP
- DEXA
- CT
- MRI
- NAA
- NIR
- …



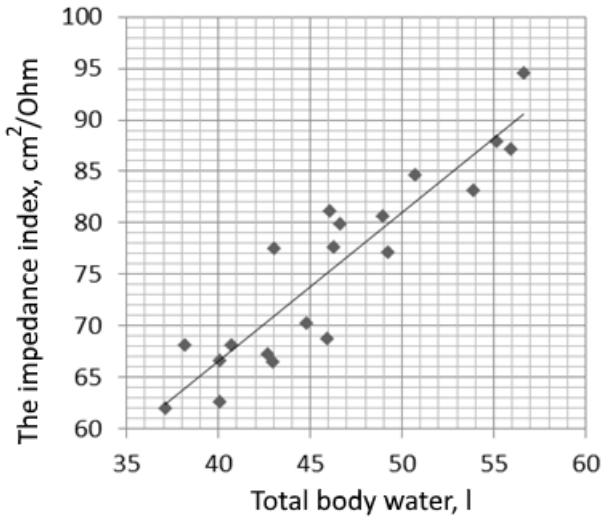

HLWG'2005



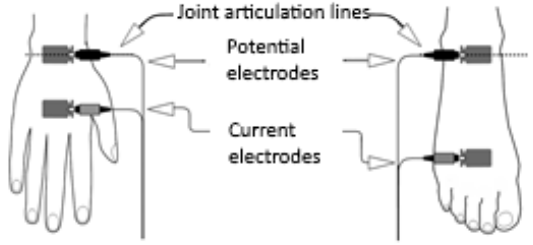

Relative risk of death: the dependence on lean (fat-free) and fat mass (Heitmann et al., 2000)

# Bioelectrical impedance analysis (BIA)

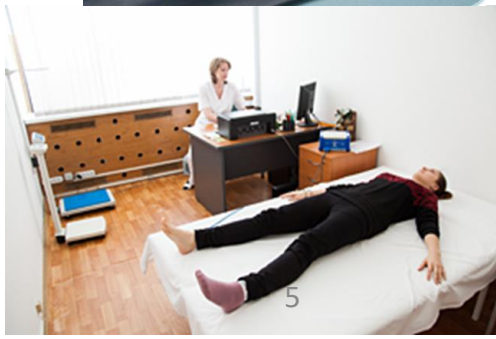
- BIA deals with passive electric properties of biological tissues which characterize their ability to oppose (impede) electric current flow (Grimnes, Martinsen, 2014)
- Portative, non-invasive, fast, relatively inexpensive
- Most commonly used method of body composition assessment in population and clinical studies
- Recommended by the European society for parenteral and enteral nutrition (ESPEN guidelines, 2004)
- Population-specific body composition formulae
- Lack of cross-calibration studies between different types of BIA instruments
- Availability of the population-based reference data for various countries: USA (Chumlea et al., 2002), Germany (Bosy-Westphal et al., 2006), UK (McCarthy et al., 2006; Franssen et al., 2014), China (Du et al., 2014), Russia (Rudnev et al., 2014) and other.




Correlation of the impedance index Ht2/R with total body water in 20 healthy subjects, r=0.92 (Hoffer et al., 1969)


The conventional tetrapolar scheme of bioimpedance measurements



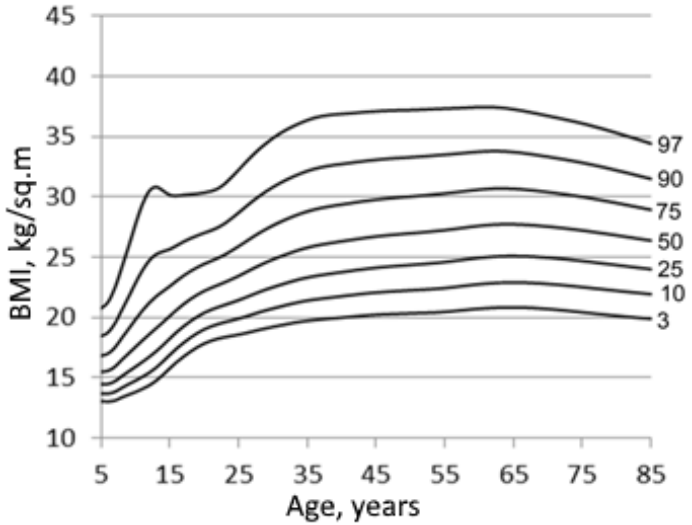
# National network of Health Centers (HCs): 1st round of data collection

BIA analyzer ABC-01 ‘Medas’ (Russia)

- Oct 2012: only BIA data (2010-2012)
- Direct submission of data from the HCs
- 52 out of 83 federal subjects
- 220 out of 800 Health Centers
- Cross-sectional data
- Initial number of BIA records: n=1,023,780
- Selected records: n=819,808

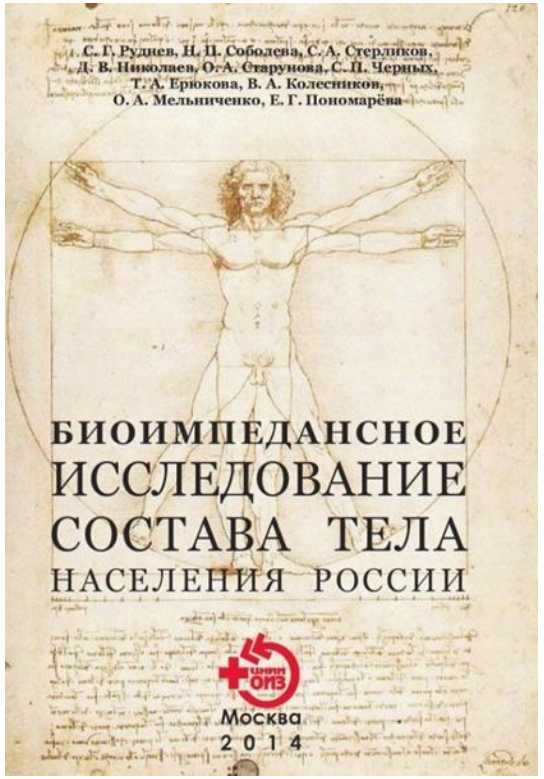
Automated anthropometric measuring system КМД 03 ТП-2 http://www.tves.com.ru

# Bioimpedance analysis: the Russian reference data




**Example:** smoothed centile reference curves for BMI in males: HCs' data

- Reference centile tables for BIA data
- R-macros in MS Excel (O. Starunova); GAMLSS package in R (Stasinopoulos, Rigby, 2007); BCT/BCPE distributions
- Parameters: R50, Xc50, Ht, Wt, BMI, LBM, FM, BCM, PA, etc (32 parameters in total)
- Inter-regional and international comparisons, health risks assessment
- Reference data for standardization




2014 book

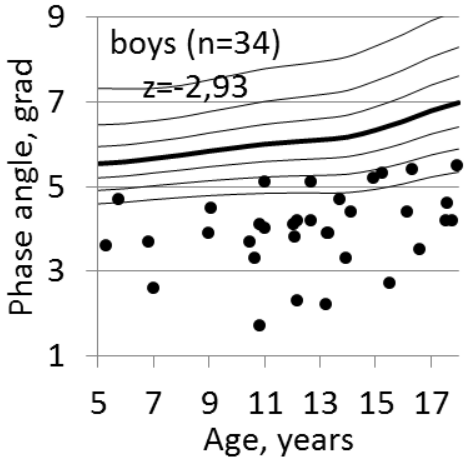
| Age, years | n | M | S | L | T | Percentile | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 3th | 10th | 25th | 50th | 75th | 90th | 97th |
| 5 | 1023 | 15.50 | 0.111 | -2.39 | 149.8 | 13.1 | 13.7 | 14.5 | 15.5 | 16.8 | 18.5 | 20.8 |
| 6 | 4624 | 15.58 | 0.117 | -2.32 | 141.0 | 13.0 | 13.7 | 14.5 | 15.6 | 17.0 | 18.8 | 21.3 |
| 7 | 9077 | 15.91 | 0.127 | -2.21 | 128.1 | 13.1 | 13.8 | 14.7 | 15.9 | 17.5 | 19.5 | 22.4 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |



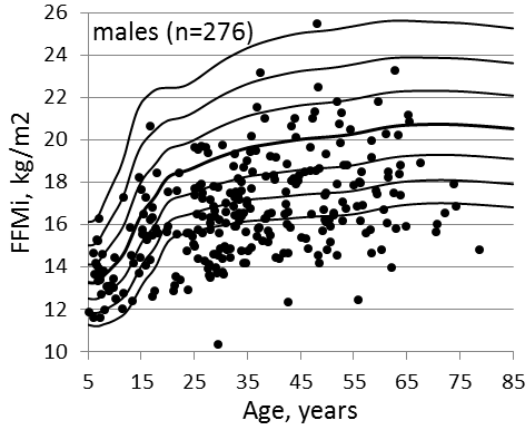
Rudnev S.G., Nikolaev D.V. et al. (2014) Bioimpedance study of body composition in the Russian population. Moscow: FRIHOI. 493 p

# Usefulness and validity of the selected BIA data

- Approved in clinical studies
  - Childhood cancer
  - Lung tuberculosis

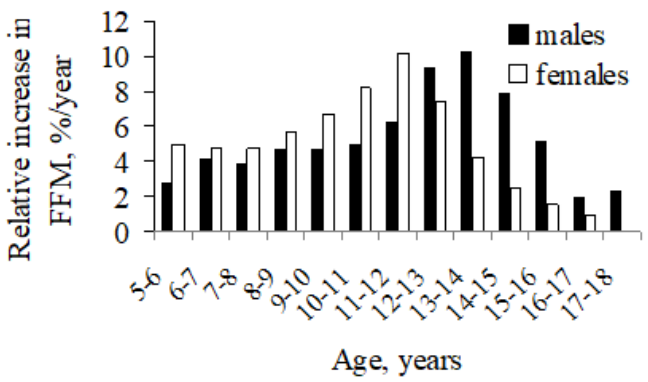


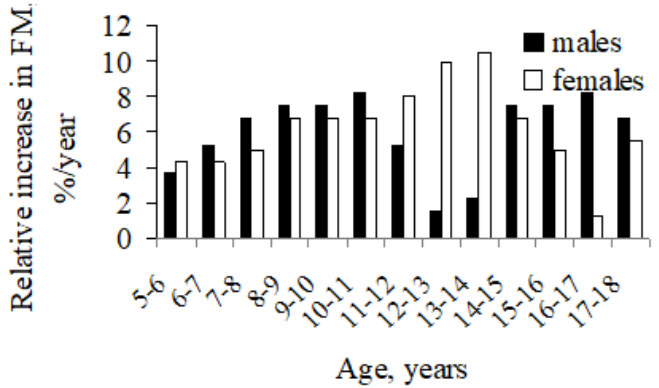

Phase angle in childhood cancer: chemotherapy / HSCT (Konovalova et al., 2014)




Fat-free mass index in lung tuberculosis patients (Rudnev et al., 2015)

- Show basic regularities of sexual dimorphism of the anthropometric and body composition parameters during growth in children



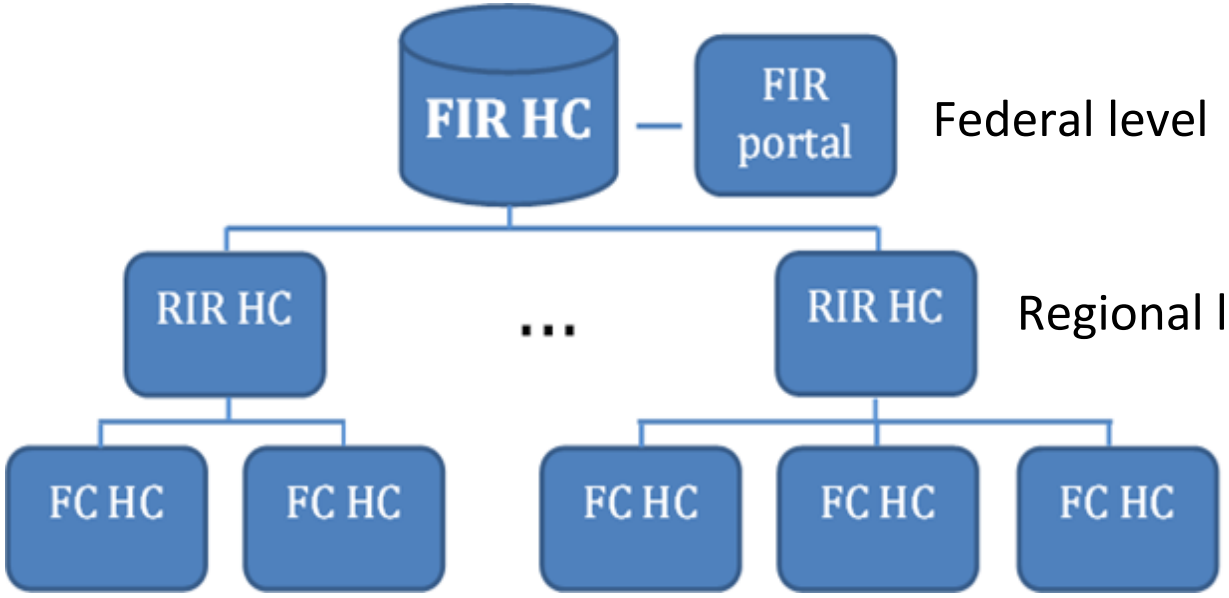




Relative growth rate of lean and fat mass in boys and girls: Health Centers' data (Starodubov et al., 2017)

# Next rounds of data collection (2014, 2015)

Federal database of the HCs (as of May, 2015)

- 3.72 million visitors
- 4.78 million meas. records

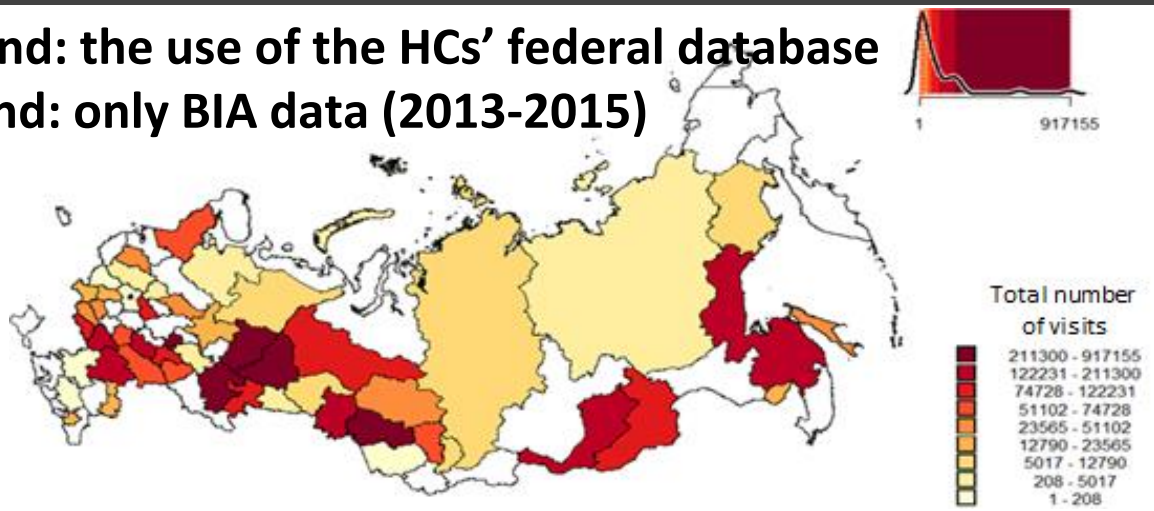



Automated data collection system of the HCs

*Abbreviations*: FIR HC - federal information resource of the Health Centers (HCs); RIR HC - regional information resource of the HCs; FC HC - functional component 'Health Center'

**2nd round: the use of the HCs' federal database**

**3rd round: only BIA data (2013-2015)**




**Measurement data**:

- Angiological screening (ankle-brachial index)
- Anthropometry, skinfolds, carpal dynamometry
- Bioelectrical impedance analysis
- Blood chemistry (cholesterol, glucose, etc.)
- Blood pressure (systolic/diastolic)
- Cardiological screening (dispersion mapping)
- Pulse oximetry
- Smoking test
- Spirometry
- *and some other*

**The resulting BIA database**:

2.35 million meas. records

# New source of fraud BIA data (identified in 2015)

- Serial measurement of the same person under the guise of different

| Sex | RTime | Age, yrs | R50 | X50 | PA, grad | Height, cm | BM, kg | WC, cm | HC, cm | BMI, kg/m2 |
|---|---|---|---|---|---|---|---|---|---|---|
| m | 27.09.2012 10:25 | 36 | 474.03 | 84.63 | 10.12 | 173.7 | 74 | 81 | 98 | 24.53 |
| m | 27.09.2012 10:26 | 58 | 472.84 | 83.94 | 10.07 | 156 | 78 | 80 | 102 | 32.05 |
| f | 27.09.2012 10:26 | 37 | 473.18 | 81.55 | 9.78 | 156 | 70 | 70 | 98 | 28.76 |
| m | 27.09.2012 10:27 | 36 | 471.34 | 82.76 | 9.96 | 159 | 70 | 77 | 109 | 27.69 |
| f | 27.09.2012 10:27 | 41 | 470.73 | 82.35 | 9.92 | 158 | 70 | 70 | 108 | 28.04 |
| m | 27.09.2012 10:28 | 59 | 470.28 | 81.99 | 9.89 | 169 | 65 | 65 | 84 | 22.76 |
| m | 27.09.2012 10:29 | 27 | 471.64 | 79.85 | 9.61 | 159 | 56 | 63 | 87 | 22.15 |
| m | 27.09.2012 10:29 | 59 | 470.01 | 81.38 | 9.82 | 163 | 66 | 79 | 89 | 24.84 |
| f | 27.09.2012 10:30 | 21 | 469.65 | 81.15 | 9.80 | 158 | 56 | 78 | 91 | 22.43 |
| f | 27.09.2012 10:31 | 47 | 470.86 | 79.12 | 9.54 | 158 | 56 | 74 | 87 | 22.43 |
| f | 27.09.2012 10:31 | 46 | 469.45 | 80.71 | 9.76 | 172 | 56 | 69 | 87 | 18.93 |
| f | 27.09.2012 10:45 | 63 | 618.57 | 73.97 | 6.82 | 166.4 | 73.2 | 86 | 115 | 26.44 |
| m | 27.09.2012 10:46 | 66 | 620.02 | 72.21 | 6.64 | 159 | 56 | 71 | 87 | 22.15 |
| m | 27.09.2012 10:47 | 52 | 620.86 | 71.05 | 6.53 | 156 | 57 | 66 | 87 | 23.42 |
| m | 27.09.2012 10:47 | 59 | 621.96 | 70.43 | 6.46 | 158 | 56 | 71 | 87 | 22.43 |
| f | 27.09.2012 10:48 | 48 | 622.04 | 69.84 | 6.41 | 157 | 63 | 72 | 89 | 25.56 |
| f | 27.09.2012 10:49 | 56 | 622.21 | 69.39 | 6.36 | 158 | 74 | 80 | 108 | 29.64 |
| f | 27.09.2012 10:50 | 28 | 622.25 | 68.80 | 6.31 | 156 | 54 | 70 | 99 | 22.19 |
| f | 27.09.2012 10:51 | 29 | 622.65 | 68.40 | 6.27 | 159 | 56 | 83 | 91 | 22.15 |

# Classification of the incorrect BIA data

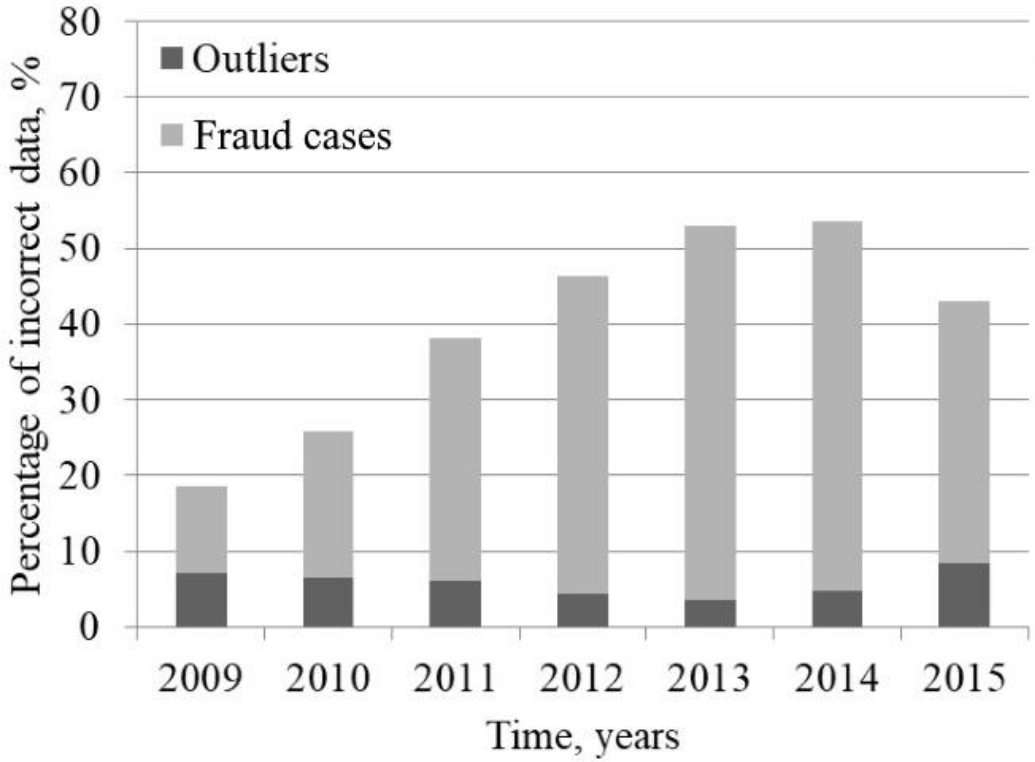
- Measurement (+typesetting) errors
- Fraud cases
  - Software emulation of measurement
  - Measurement of the electronic verification module instead of a patient
  - Serial measurements of the same person under the guise of different

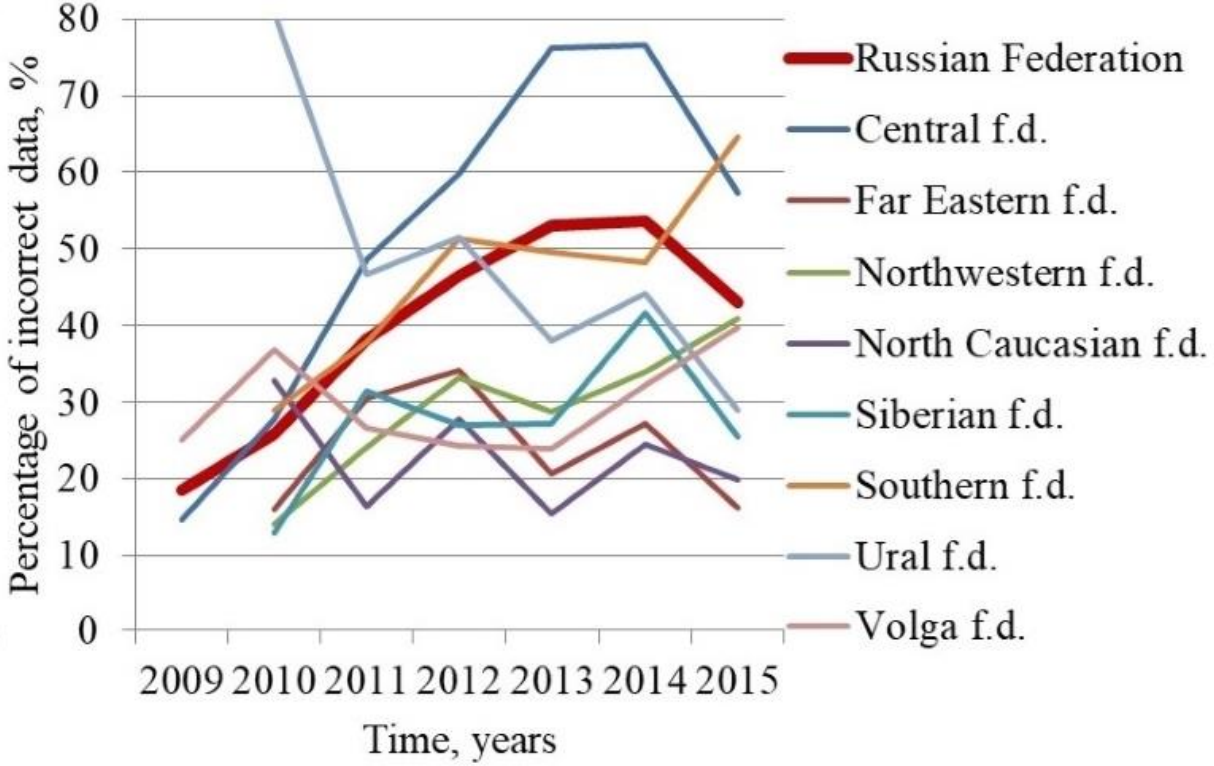
Inclusion selection criteria to detect measurement errors, adults

| Ht, cm | BM, kg | BMI, kg/m$^2$ | BF, kg | %BF | R50, Ohm | Xc50, Ohm | PA, grad |
|---|---|---|---|---|---|---|---|
| 130–210 | 35–150 | 12–55 | 0.5–75 | 0–55 | 250–1000 | 20–150 | 3.0–10.2 |

Starunova OA, Rudnev SG, Starodubov VI (2017) HCViewer: software and technology for quality control and processing raw mass data of preventive screening. Russ J Numer Anal Math Model 32(5): 315-326

# Dynamics of the incorrect BIA data

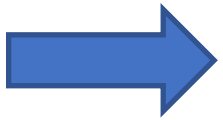



- Relatively low, but significant percentage of measurement errors (5.2% on average)
- Fraud growing rapidly (39.6% on average)
- Insufficiency of the existing control measures, need for automated data quality control
- Potential controllability: 80% of the incorrect data were generated in 20% of the HCs

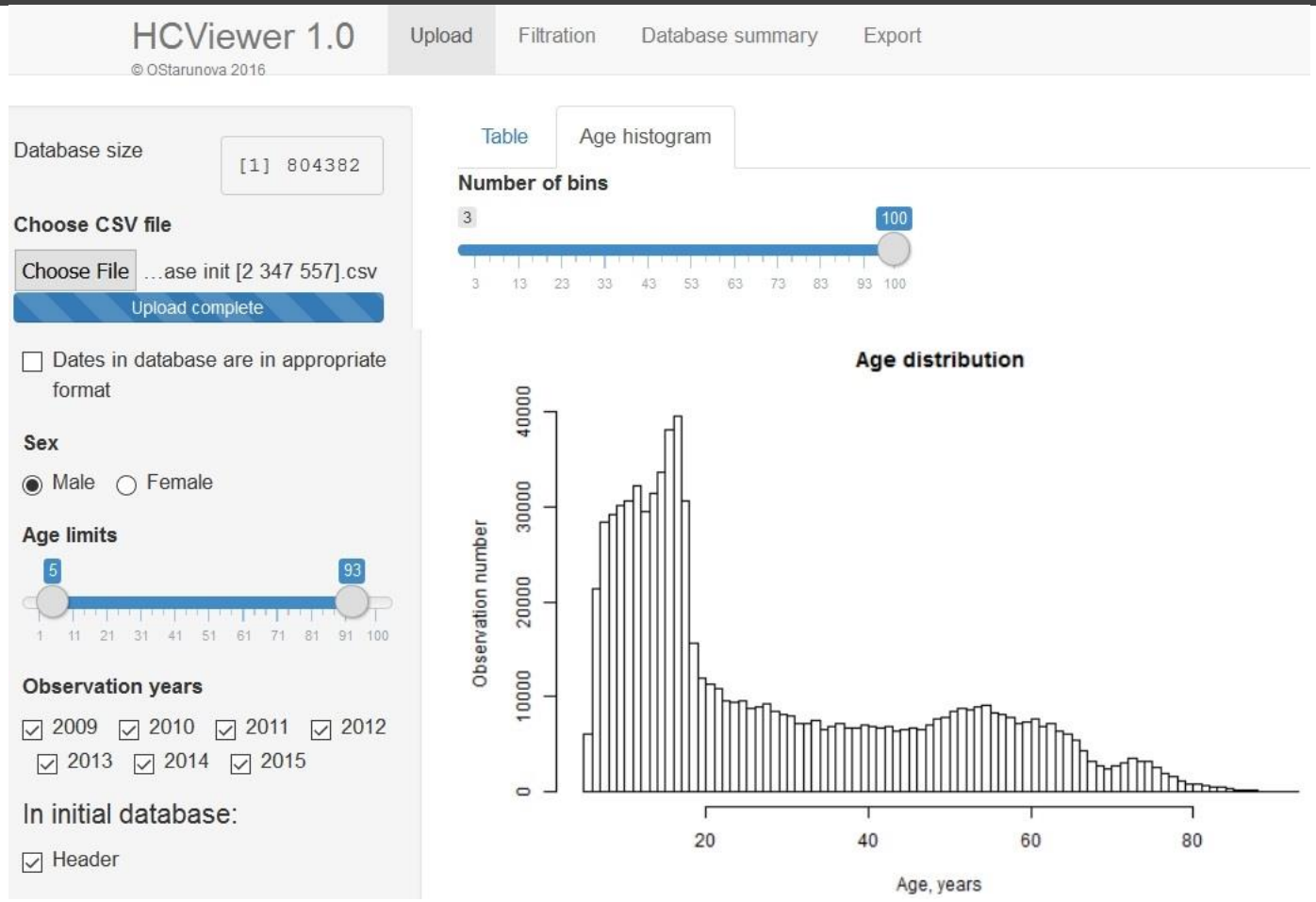
# Motivation of using Benford's analysis

- Strict selection criteria for data filtering
- Known structure of the incorrect BIA data
- Availability of the other HCs' data awaiting for analysis
- Fraud is adaptive



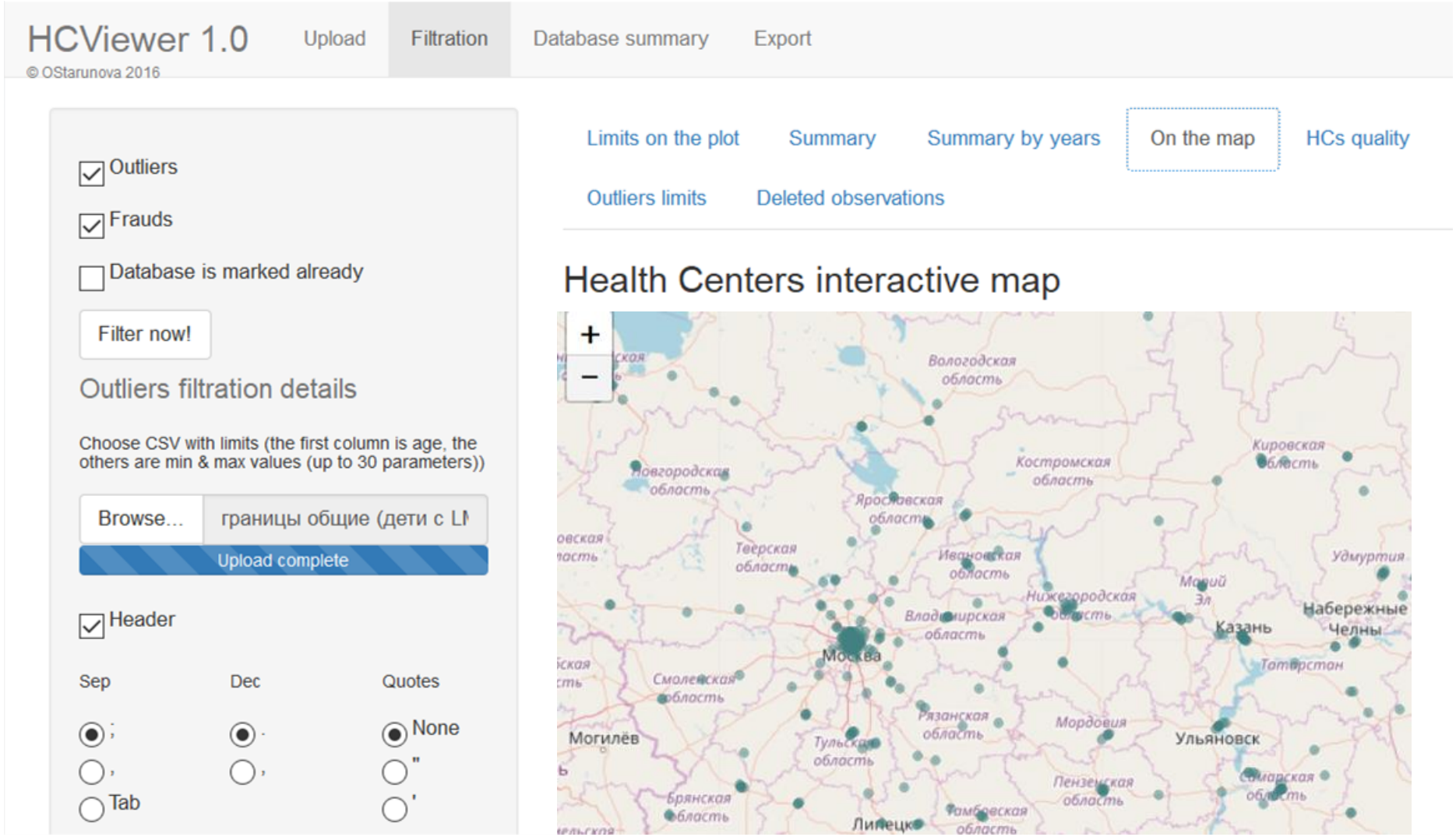
Need for flexible algorithms to identify potentially suspicious data

# HCViewer: software for data filtration and analysis

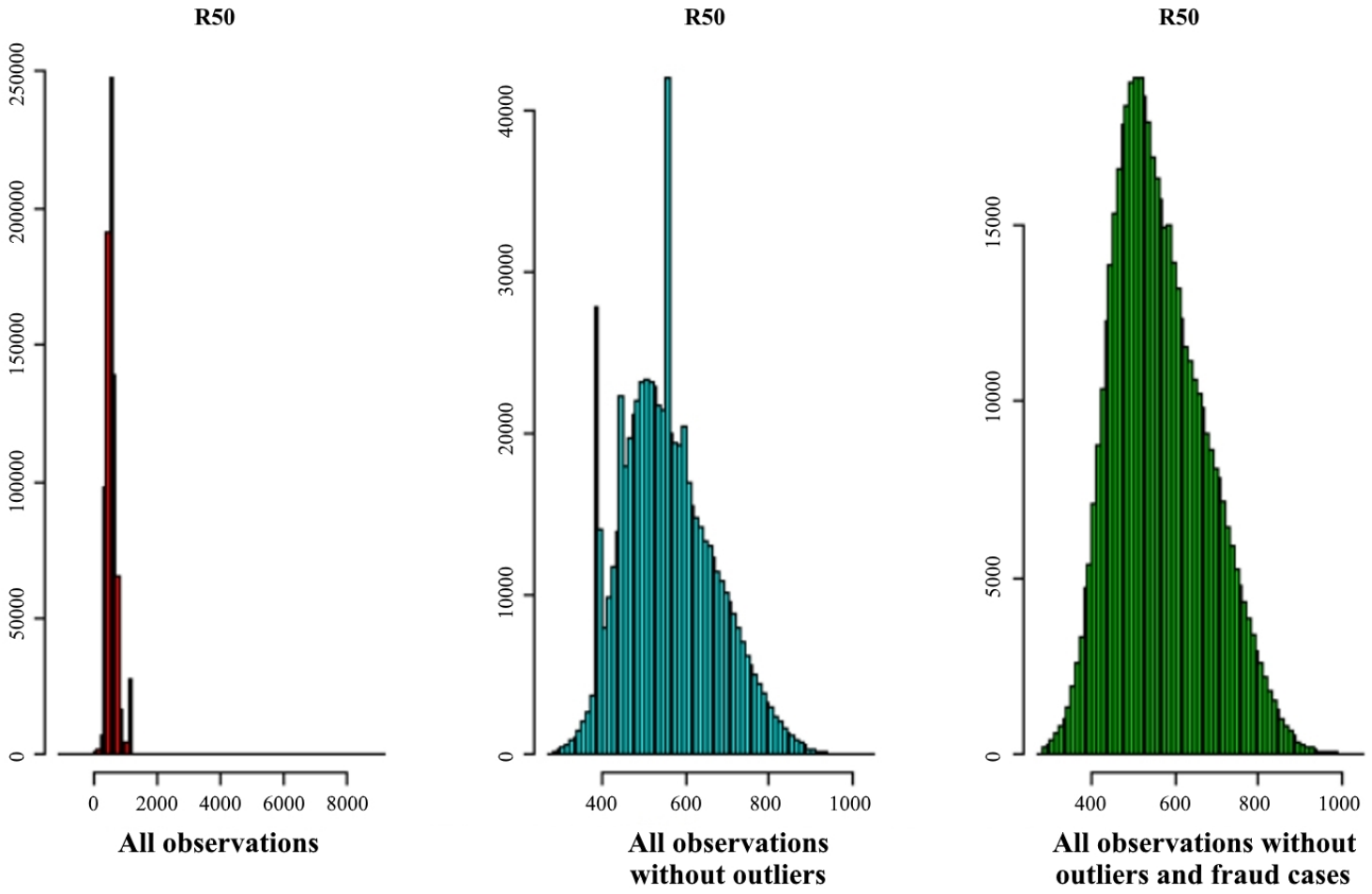



Starunova OA, Rudnev SG, Starodubov VI (2017) HCViewer: software and technology for quality control and processing raw mass data of preventive screening. Russ J Numer Anal Math Model 32(5): 315-326

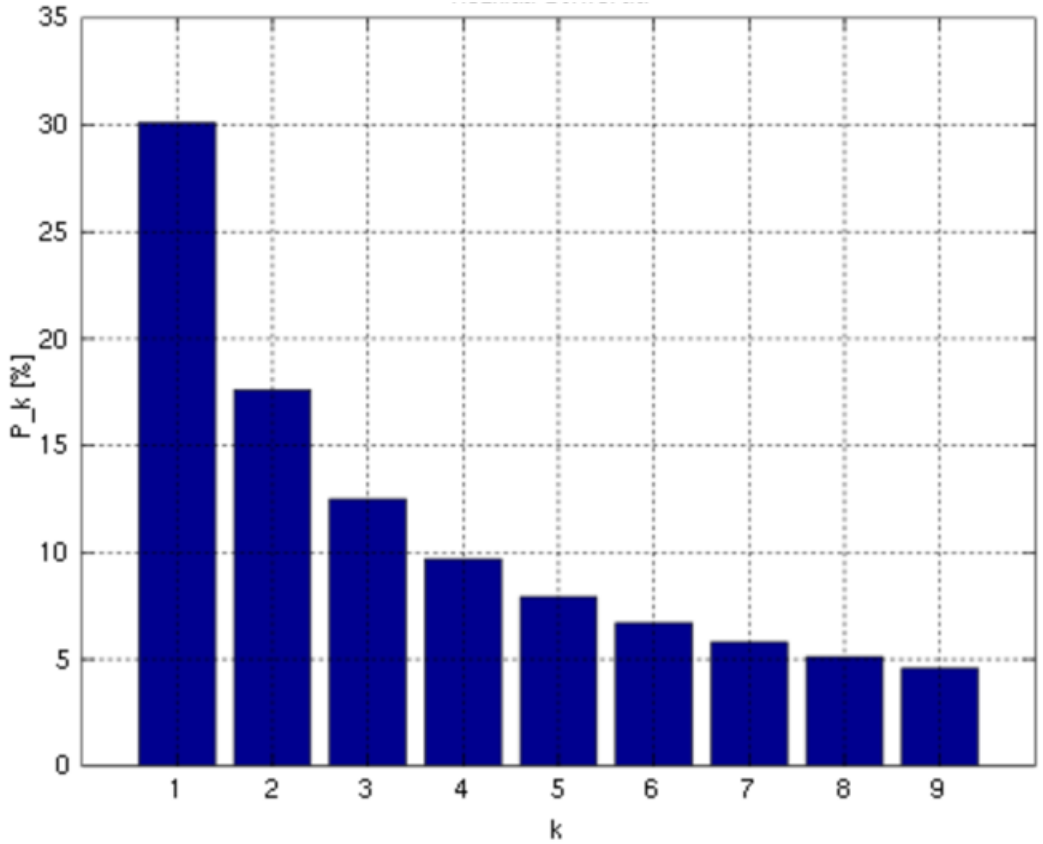
# HCViewer: the HCs' interactive map



# HCViewer: sequential data filtering



# Benford's law




$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right), d = \overline{1,9}$$

The distribution of first significant digits of different biological and social datasets, according to Benford's law. Each bar represents a digit, and the height of the bar is the percentage of numbers that start with that digit.


Newcomb S (1881) Note on the frequency of use of different digits in natural numbers. Am J Math 4:39-40

Benford F (1938) The law of anomalous numbers. Proc Am Phil Soc 78:551-572

Berger A, Hill T (2011) A basic theory of Benford's Law. Probability Surveys 8:1-126

# Benford's law application: population size for 198 countries

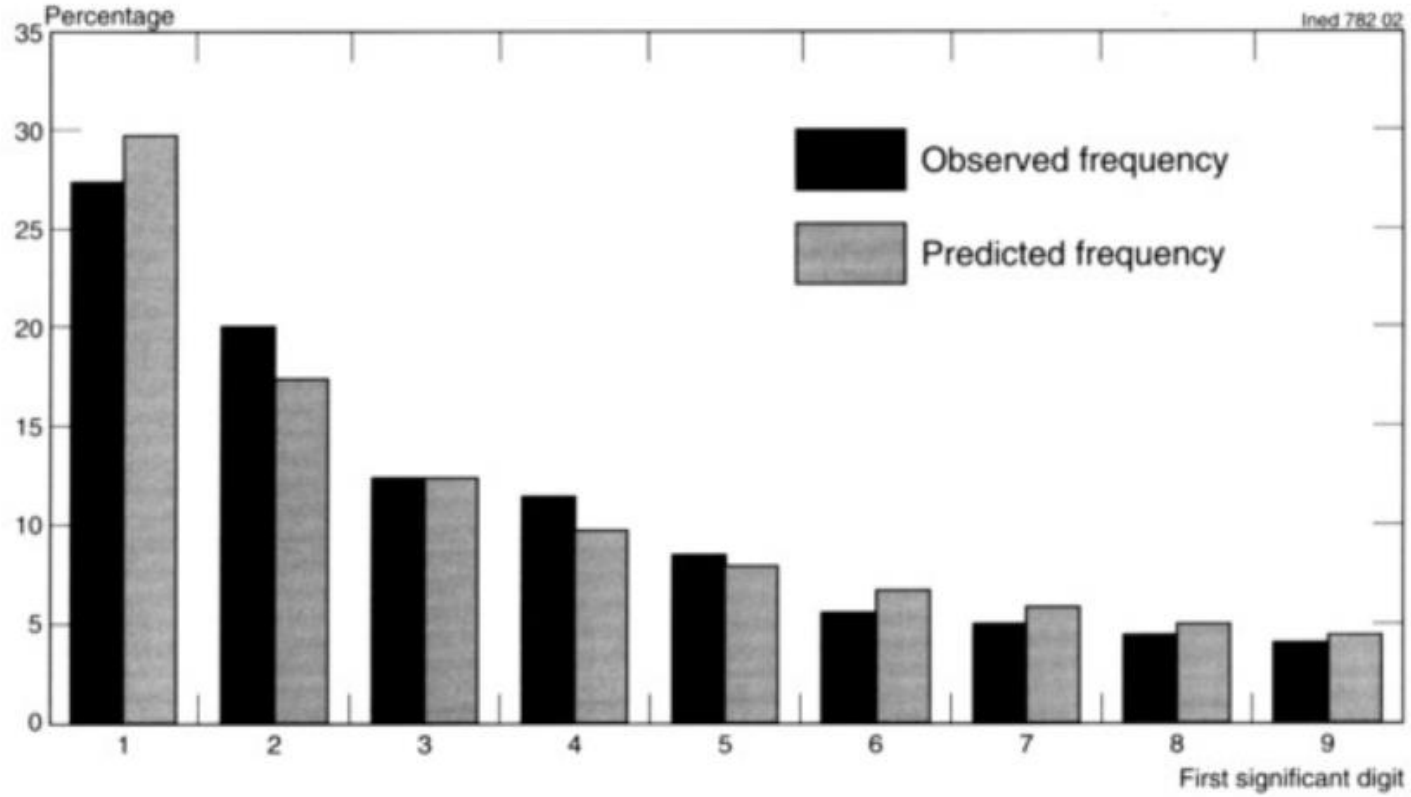



Figure 1.–Observed frequency of the first significant digit of the 1997 population size for 198 countries and predicted frequency according to Benford's law

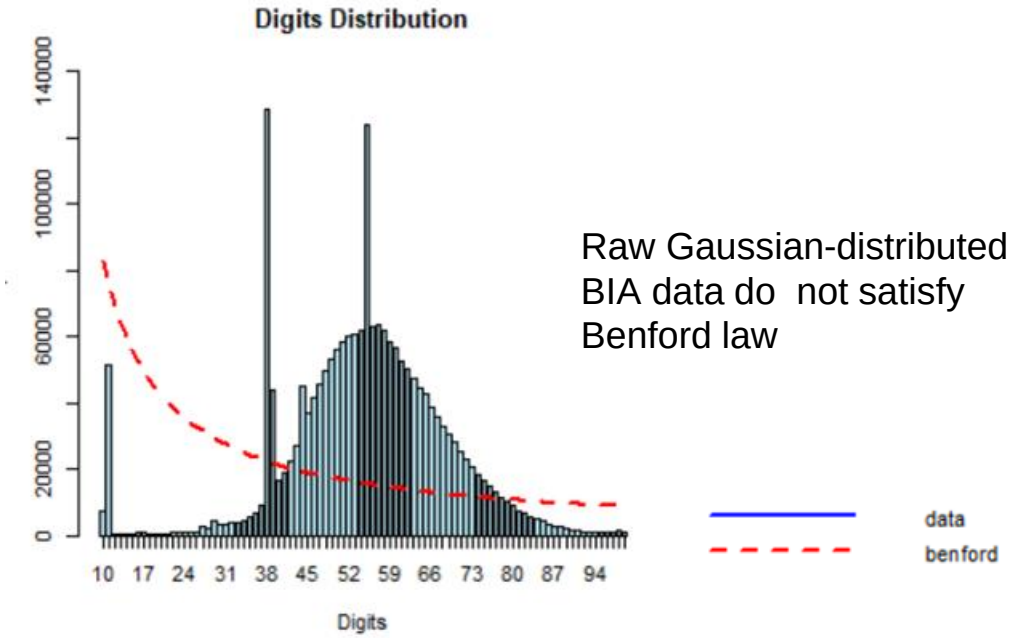
*Source*: INED 1997.

[1] Sandron A (2002) Do populations conform to the Law of Anomalous Numbers. Population. 57(4):755-761

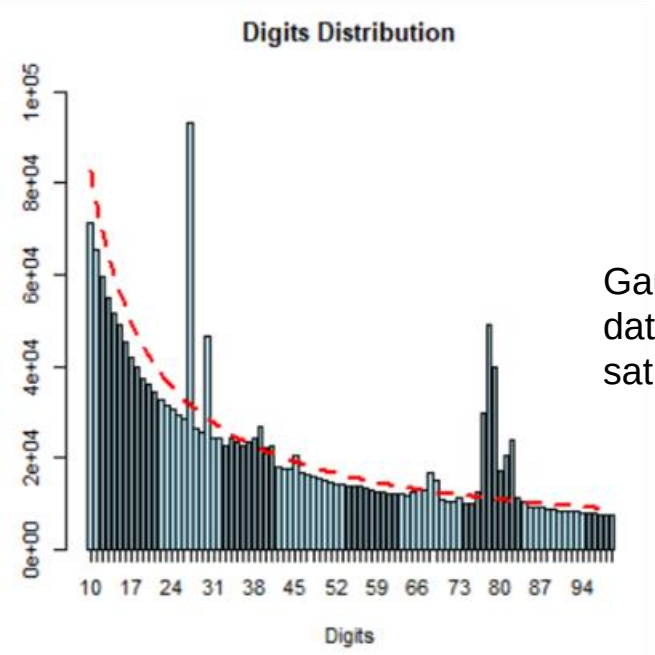
# Benford analysis: main idea

The main purpose is to find out where the dataset deviates from Benford's Law and to identify suspicious data for further verification.

For raw bioimpedance parameters data we can’t use this law directly [1], since the data should be distributed in the wide range. So, we raised the R50 data to the $10^{th}$ power to enlarge the difference between min and max.




Raw Gaussian-distributed BIA data do not satisfy Benford law

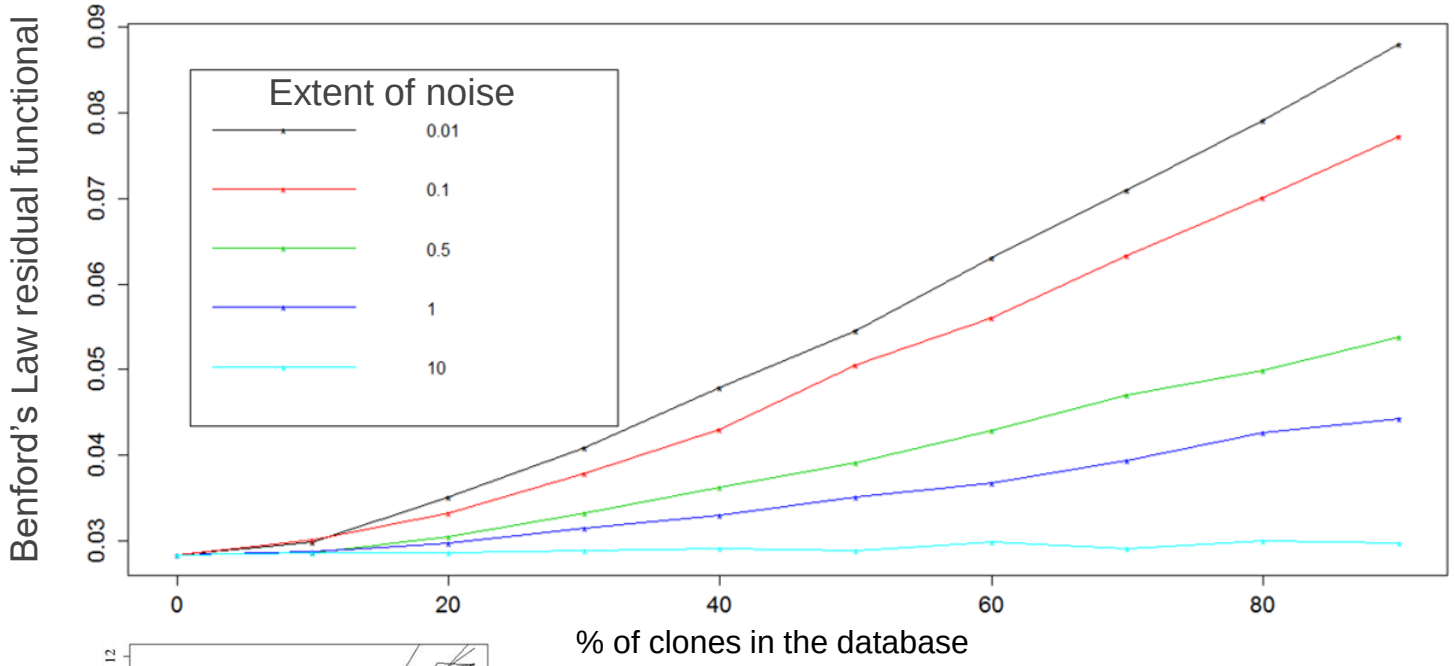


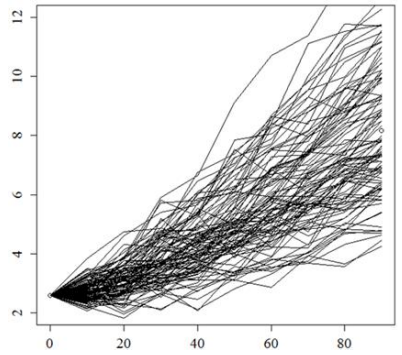

Gaussian-distributed BIA data to the 10th power satisfy Benford law



[1] Formann A K (2010) The Newcomb-Benford Law in Its Relation to Some Common Distributions. PloS One. 5(5): e10541

# Benford analysis: preliminary result




Individual runs of the experiment

The database (N=914) of knowingly reliable BIA data was considered [2].

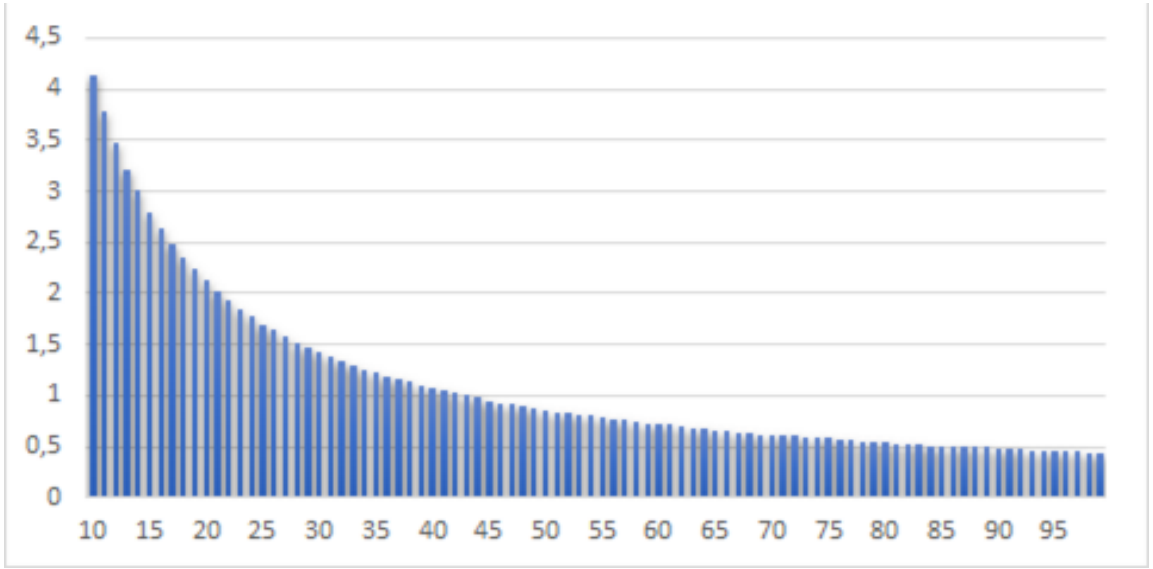
We selected one patient's record in the database and replaced some part of records by this one (shown on the X axis). Then we slightly noised clones' data (the extent of noise is shown on plot by color) and considered how much data would deviate from the Benford law (mean value by 1000 experiments).

The residual function increased monotonically with the % of clones in the database.


[1] Starunova OA (2015) / Proc. 10th Int. workshop 'Science and innovation-2015' (03-12 July, 2015). Ioshkar-Ola, Volga State Technol. Univ.
[2] Anisimova AV, Godina EZ, Nikolaev DV, Rudnev SG (2016) *IFMBE Proceedings*

# Two significant digits distribution

**First and second significant digits distributions, %**



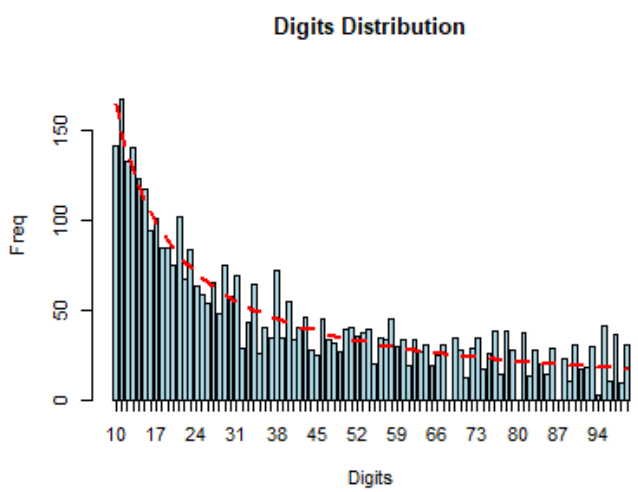

| First significant digit | Second significant digit: 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ≈ Σ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4.14 | 3.78 | 3.48 | 3.22 | 3.00 | 2.80 | 2.63 | 2.48 | 2.35 | 2.23 | 30.10 |
| 2 | 2.12 | 2.02 | 1.93 | 1.85 | 1.77 | 1.70 | 1.64 | 1.58 | 1.52 | 1.47 | 17.61 |
| 3 | 1.42 | 1.38 | 1.34 | 1.30 | 1.26 | 1.22 | 1.19 | 1.16 | 1.13 | 1.10 | 12.49 |
| 4 | 1.07 | 1.05 | 1.02 | 1.00 | .98 | .95 | .93 | .91 | .90 | .88 | 9.69 |
| 5 | .86 | .84 | .83 | .81 | .80 | .78 | .77 | .76 | .74 | .73 | 7.92 |
| 6 | .72 | .71 | .69 | .68 | .67 | .66 | .65 | .64 | .63 | .62 | 6.69 |
| 7 | .62 | .61 | .60 | .59 | .58 | .58 | .57 | .56 | .55 | .55 | 5.80 |
| 8 | .54 | .53 | .53 | .52 | .51 | .51 | .50 | .50 | .49 | .49 | 5.12 |
| 9 | .48 | .47 | .47 | .46 | .46 | .45 | .45 | .45 | .44 | .44 | 4.58 |
| ≈ Σ | 11.97 | 11.39 | 10.88 | 10.43 | 10.03 | 9.67 | 9.34 | 9.04 | 8.76 | 8.50 | |

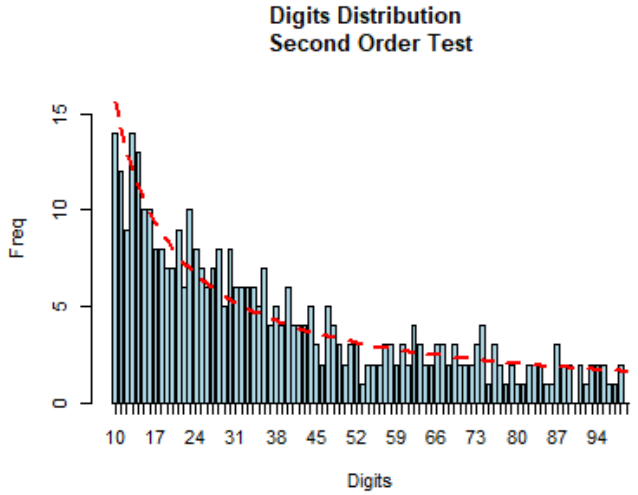
Joenssen D W (2013) Two Digit Testing for Benford's Law. Proceedings of the ISI World Statistics Congress, 59th Session in Hong Kong. Available at: http://www.statistics.gov.hk/wsc/CPS021-P2-S.pdf

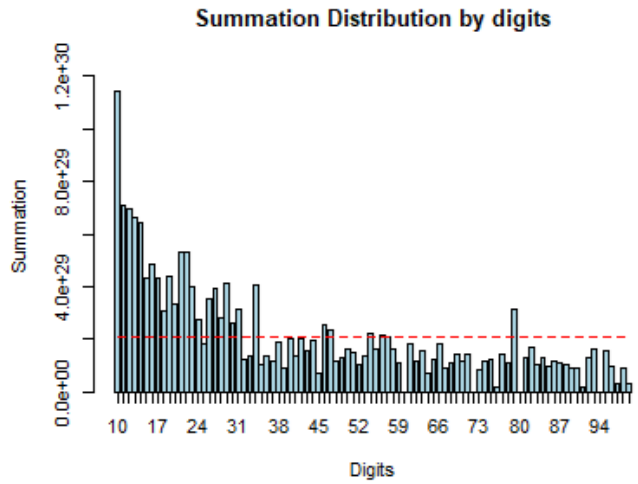
# Benford analysis in R: package ‘BenfordTests’

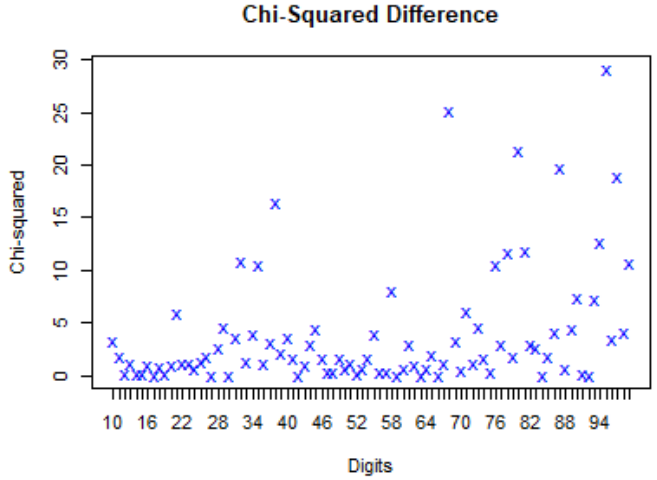

Joenssen D W (2013) Two Digit Testing for Benford’s Law. Proceedings of the ISI World Statistics Congress, 59th Session in Hong Kong. Available at: http://www.statistics.gov.hk/wsc/CPS021-P2-S.pdf

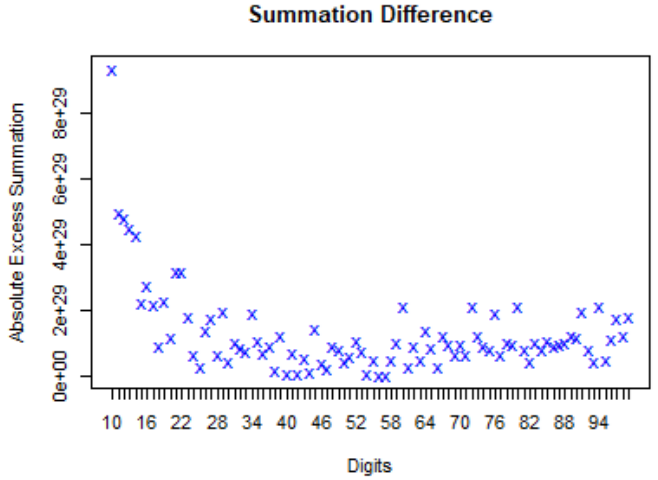
# Example 1: reliable data from the Health Center X

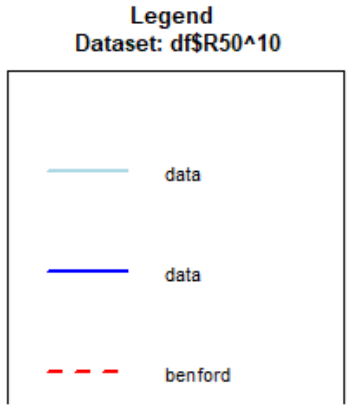


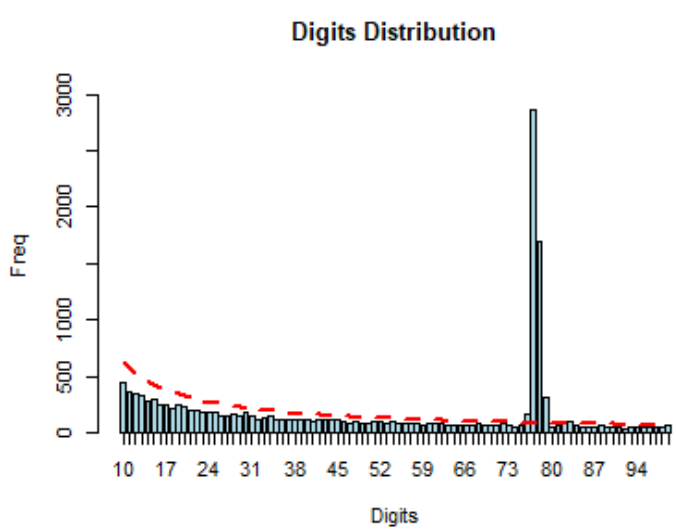



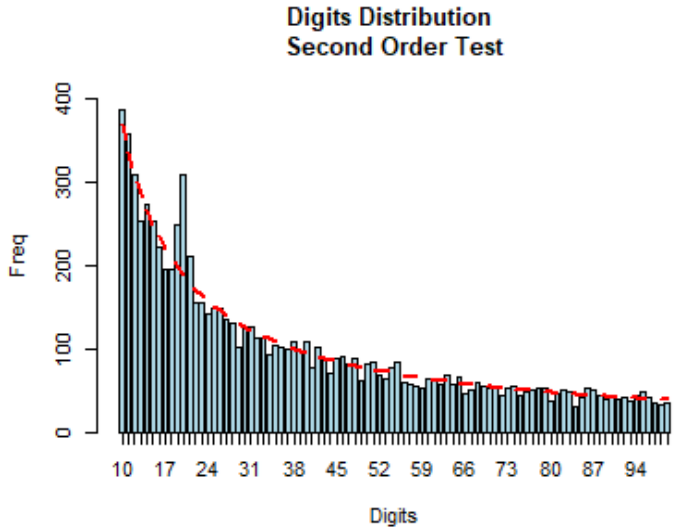



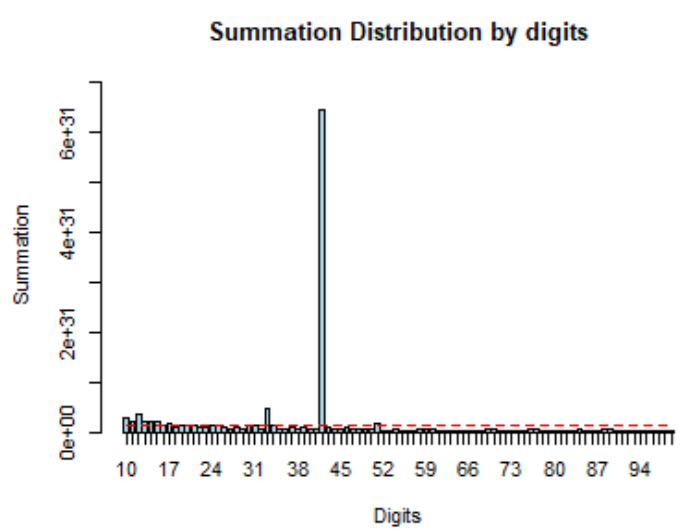



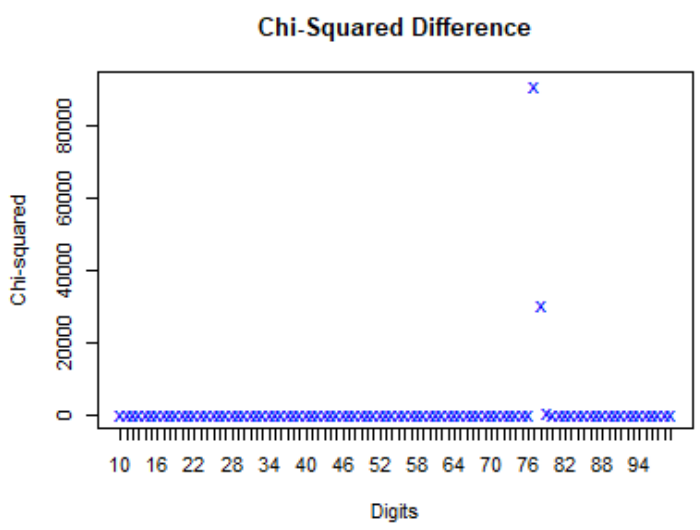



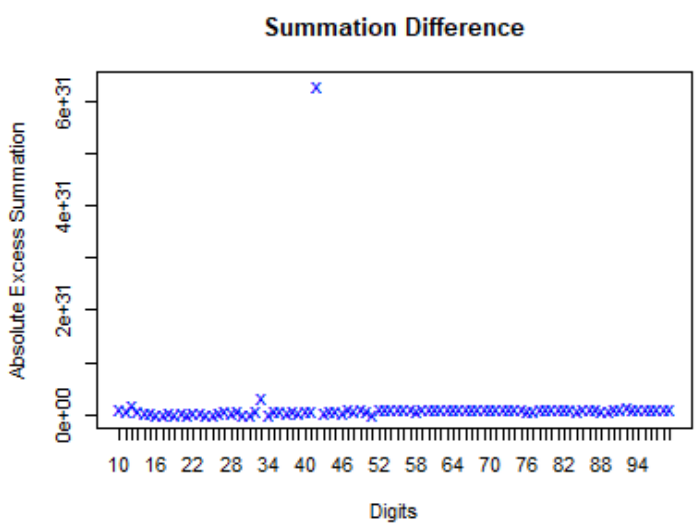
# Example 2: fraudulent data from the Health Center Y

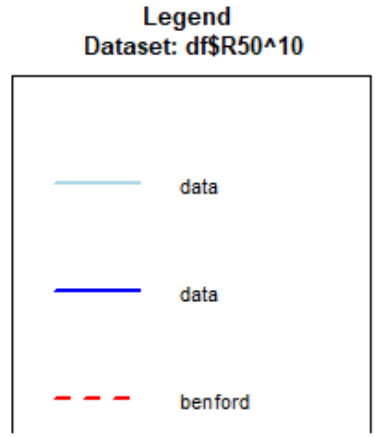


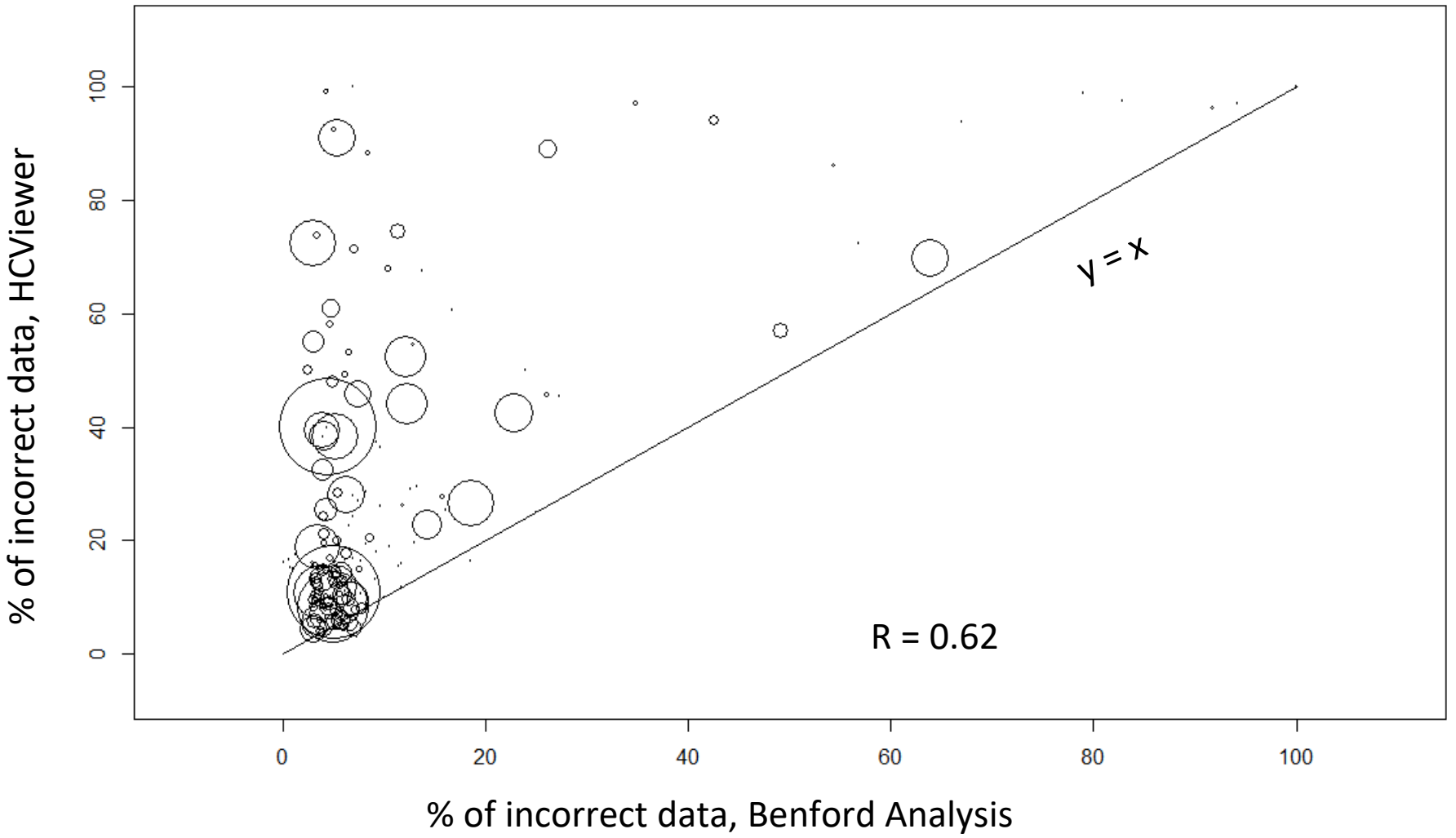



# The HCs bioimpedance data: Benford analysis vs HCViewer



# Summary

- Fraud and measurement errors significantly compromise the HCs' data of preventive screening which makes it difficult to use them for the analysis of population health
- Rapid growth of the proportion of fraud cases suggests insufficiency of the existing control measures
- Software HCViewer is developed for automated quality control of the HCs' data
- High sensitivity and specificity of the expert-derived empirical selection criteria for the BIA measurement data (Starunova et al., 2017)
- Benford analysis:
  - Correlated well with the empirical selection criteria for the BIA data
  - Can be used to identify, at least, some types of the incorrect HCs' data

# HCViewer: prospects

- Implementation of Benford analysis and machine learning
- Online health monitoring, more analysis and visualization options

# Acknowledgements


- Russian Science Foundation (grant no. 14-15-01085)
- Russian Health Ministry
  - Department of Information Technologies and Communications
  - Department of Health and Sanitary-Epidemiological Well-Being
- IT company SofTrust (Belgorod)

# Thanks for your attention!