pp. 1–15 (2017)

HCViewer: software and technology for quality control and processing raw mass data of preventive screening

O. A. Starunova[*], S. G. Rudnev[†*], and V. I. Starodubov[*]

Abstract — A prototype automated system for the analysis and processing raw data of mass population screening in the Russian health centers (HCs), software program HCViewer, is developed. In it, essential elements of big data, including data filtering (i.e., detection and removal of outliers and fraud cases), statistics and spatial visualization are implemented. Testing HCViewer on the datasets of knowingly reliable, unreliable, and partially reliable bioimpedance body composition data has shown high diagnostic specificity (95.2–98.9%) and sensitivity (94.5–99.2%) of the filtering criteria. Retrospective analysis of the HCs' bioimpedance database revealed rapid growth in the proportion of fraud cases, from 5–7% in 2009 to 43–45% in 2013–2014 (with the correction for false alarm rate), and the high level of heterogeneity in data quality, so that 80.3% of the incorrect data were generated in 20% of the HCs. This indicates inefficient management of the Russian preventive screening system and suggests incomparability and, therefore, uselessness of the related health statistics from some regions without a preliminary filtering of the underlying raw data. Our results show the potential utility of HCViewer for the dynamical quality control of the preventive screening data, as well as for the online health monitoring and modelling potential effects of various control measures aimed at the prevention of non-communicable diseases. After application of the selection criteria, a cross-sectional 2009–2015 database of bioimpedance measurements of 1.27 million individuals aged 5–96 years was formed for the assessment of population health. On the whole, fraud prevention should become the responsibility of the Russian health care system to ensure an efficiency of its function.

Keywords: Health care fraud, preventive screening, bioelectrical impedance analysis, mass population data, big data, software.

MSC 2010: 62-07, 62P10, 68T35, 92D30

Non-communicable diseases (NCDs), including cardiovascular, chronic lung diseases, cancer, and diabetes mellitus, are the leading cause of premature mortality worldwide entailing significant human and economic losses [38]. In Russia, the probability of dying between ages 30 and 70 from four main NCDs is one of the highest in the world and amounts to 29.9% [38] while the age-standardized rate of NCDs mortality remains 2–3 times higher than in industrial Western European countries [12]. This means the importance of the development and implementation of the efficient strategies to combat NCDs.

[*]Federal Research Institute for Health Organization and Informatics, Ministry of Health of the Russian Federation, Moscow 127254, Russia

[†]Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow 119333, Russia. E-mail: sergey.rudnev@gmail.com

This work was supported by the Russian Science Foundation grant No. 14-15-01085.

Figure 1. The Russian network of health centers (HCs): geographical distribution.

The morbidity and mortality of NCDs depend on nutritional state, body composition, physical activity level, smoking status, alcohol consumption and other objectively evaluated factors [22, 36]. So, the leading role in reducing NCDs mortality is given to prevention and timely identification of patients at risk during screening [6, 15]. Besides the assessment of risks, screening studies provide useful information on the biological variability, spatial differences and temporal trends of morphological and physiological parameters depending on age, sex, ethnicity, as well as social, demographic, climatic and other factors [1, 8].

Due to the launching of the automated data collection system, the Russian network of health centers (HCs), which was established in 2009 and now constitutes an important part of the Russian preventive screening system, represents the only source of preventive screening raw mass data at the national level [32]. Currently, this network comprises nearly 800 HCs (of them, about 740 stationary and 60 mobile) evenly distributed throughout Russia according to the population density (Fig. 1).

Among other screening methods routinely used in the HCs, such as anthropometry, blood pressure and chemistry, angiologic and cardiologic screening, pulse oximetry and spirometry, the bioelectrical impedance analysis of body composition (BIA) is utilized. BIA deals with passive electrical properties of biological tissues which characterize their ability to oppose electric current flow [11]. The body composition parameters, such as body fat (BF), fat-free mass (FFM), and relative body fatness (%BF), are obtained using measured impedance values, e.g. the resistance and reactance, and some additional variables, such as age, sex, height, and body mass [17–19]. At present, BIA is the most common method of body composition assessment in epidemiological studies [4, 7, 9]. In relation to NCDs, BIA provides a number of specific markers for diagnosing obesity, metabolic syndrome risk, sarcopenia, malnutrition, physical disability risks, and other conditions [13, 19, 39].

According to current regulations, reimbursement of services provided by the HCs depends solely on the total number of examined patients which is prescribed by the attendance plan. (In most federal subjects, the payments are made through the system of compulsory medical insurance.) This practice is potentially opposed

to the interests of data quality, and in the case of low attendance or overestimated attendance plan may create an environment with the high risk of fraud. Additional problem arising from known elevated rate of the HCs' staff turnover is the uneven level of the personnel training which is another obstacle to ensure data quality. (On the importance of quality control and reliability of anthropometric measurement data collected by health workers, see, e.g., [14].)

Indeed, in our previous work, the presence of significant ever-growing proportion of incorrect BIA data from the HCs, including outliers and fraud data, was observed, with the overall initial estimate being from 11.2% in 2010 to 28.1% in 2014 [33]. This estimate did not take into account the correction for false positive results since the properties of the utilized filtration algorithm were not studied. On the other hand, an additional source of the incorrect data has been identified recently with the potential to increase significantly the earlier estimate.

Preliminary check and filtration of raw data is, therefore, a prerequisite for any objective health monitoring based on preventive screening data from the HCs. The federal database of HCs contains the enormous volume of data [32] which is hardly possible to process manually. So, in view of the need to timely identify and eliminate outliers and fraud data from the data stream generated by the HCs, automated filtration and processing of the HCs' data is necessary, suggesting an idea of using big data techniques [27].

Our aim was to develop software for quality control and processing raw mass data of preventive screening in the HCs, to reanalyze the statistics of incorrect BIA data and to form an updated BIA measurements database for the assessment of population health.

1. Data and methods

Our main data source, a copy of the federal database of HCs [32], as of May 2015, was stored in SQL format on the server system R-IT Data Mill having 24 Tb of total disk space, installed operating system Microsoft Windows Server 2012 R2, and relational database management system Microsoft SQL Server 2014. The total volume of the database amounted to 94.8 Gb. Of these data, only BIA measurement data were used in our study. As described in [32], due to our previous activities [28] and also to the incompleteness of the federal database of HCs, two other BIA datasets were directly obtained from the HCs: the 2010–2012 data that was collected in accordance with the letter from the Russian Health Ministry No. 14-1/10/2-3200 of October 24, 2012, and the 2013–2015 data that was collected in accordance with the letter from the Federal Research Institute for Health Organization and Informatics No. 7-5/434 of July 2, 2015. After removing duplicates, these data were added to the federal database of HCs.

To ensure comparability, only measurement data from the same type of the BIA meter, ABC-01 'Medas' (SRC Medas, Russia), were considered. All measurements were made according to the conventional tetrapolar scheme using disposable bioadhesive Ag-AgCl ECG electrodes, mainly Schiller Biotabs (Shiller AG, Switzerland)

or F3001ECG (FIAB SpA, Italy). The combined initial dataset of the cross-sectional bioimpedance data stored in CSV format was formed. It contained 2,347,557 records with the results of BIA measurements of individuals aged 5–96 years from 335 HCs representing 62 out of 85 federal subjects of Russia [27].

After the exclusion of records that did not contain data on patient's age, sex, height, body mass, resistance or reactance value at electric current frequency 50 kHz (R50, Xc50), as well as records with patient ages below 5 years, we arrived at the intermediate BIA database containing 2,294,846 records (791,621 males and 1,503,225 females) ready to apply selection criteria to incorrect data.

Two major types of incorrect data were considered: outliers and fraud cases. In adults (patient ages 18+), stationary intervals for the admissible values of the BIA parameters independent of age and sex were used to detect outliers, see Table 1. For this, the data on height (Ht), body mass (BM), body mass index (BMI), body fat (BF), relative body fatness (%BF), resistance (R50), reactance (Xc50), and the impedance phase angle (PA) were considered. For the younger ages, upper and lower bounds of the intervals were put equal to a constant or determined as a piecewise constant functions, according to year of age, by linking the parameter value at 18 years to child and adolescent percentiles to account for age-dependent changes (usually within the limits of 3.5–4.5 standard deviations around mean values). The data outside the range were considered outliers.

Based on our previous work [27, 28, 34], three types of fraud cases were identified:

1) software emulation of measurement (this type of fraud in the database was associated with the exact R50 values 555.5 Ohm or 444 Ohm);

2) measurement of electronic verification module instead of a patient (the module is provided by the manufacturer for regular check of the BIA meter to ensure its optimal performance and functionality). Accordingly, the measured values of R50 and Xc50 lying simultaneously within the respective intervals 387–391 Ohm and 38–48 Ohm were considered fraud cases;

3) multiple measurements of the same person under the guise of different. Three empirical schemes to detect this type of fraud were used. For this, the records in the database were sorted by the time of measurement with the individual HCs' data considered sequentially, one by one. First, the adjacent records were considered fraud cases if the respective two consecutive values of R50 and Xc50 were different by no more than 1% and 7% simultaneously (these values represent the assumed maximal values of the measurement errors in the case that the measurer does not tend to strictly follow the prescribed BIA measurement protocol). Second, fraud cases were identified if the time interval between two consecutive BIA measurements was below a certain threshold chosen equal to 1.5 minutes (see below). Third, tak-

Table 1. Inclusion selection criteria used to detect outliers in the HCs' BIA data for the adult age.

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

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
f	27.09.2012 10:51	62	623.55	68.19	6.24	159	69	79	91	27.29
m	27.09.2012 10:52	23	624.57	68.01	6.21	158	59	75	89	23.63
f	27.09.2012 10:52	42	624.25	67.72	6.19	157	56	78	98	22.72
f	27.09.2012 10:53	53	625.63	67.63	6.17	158	63	78	85	25.24
f	27.09.2012 10:54	26	626.35	67.58	6.16	166	63	74	91	22.86
f	27.09.2012 10:55	55	626.36	67.48	6.15	159	56	63	87	22.15
f	27.09.2012 12:45	35	679.78	81.21	6.81	166.5	58.1	71	93	20.96
f	27.09.2012 12:47	60	680.22	80.58	6.76	158	78	70	98	31.25
m	27.09.2012 12:47	20	681.08	80.41	6.73	156	51	69	90	20.96
f	27.09.2012 12:48	28	683.75	77.72	6.48	158	56	70	98	22.43
m	27.09.2012 12:49	28	681.86	80.21	6.71	159	58	70	98	22.94
m	27.09.2012 12:49	17	684.39	77.60	6.47	158	50	65	89	20.03
m	27.09.2012 12:50	18	682.63	80.05	6.69	156	51	70	90	20.96
f	27.09.2012 12:51	38	682.58	79.94	6.68	158	70	80	105	28.04

Figure 2. A fragment of suspicious (fraudulent) measurement data from the HCs' BIA database: a series of multiple measurements of the same person under the guise of different.

ing into account the sufficient smoothness of the natural statistical distributions of the bioimpedance data (see, e.g., [2]), the range of measured R50 values for each HC was divided into 100 equal intervals (bins). If the number of measurements in some bin (i.e., the column height) exceeded 50, and was at least 1.5 times higher than the arithmetic mean of the 4 neighbouring bins (taken two on the left and right), then all the data in this bin were marked as fraudulent. Outliers that met the selection criteria for fraud cases were considered fraud cases.

An example of suspicious measurement data in the HCs' BIA database indicating the presence of the 3rd-type fraud, as identified by an expert, is shown in Fig. 2. These data of the consecutive measurements of, presumably, individuals of both sexes different in age, height, body mass and other anthropometric characteristics, show slow changes in the successive values of the bioimpedance resistance (R50) and reactance (Xc50) values which is typical for individual physiological variation. Note that, normally, the duration of the BIA measurement procedure of one person using the ABC-01 'Medas' BIA meter lasts, at least, 5–7 minutes (taking into account the time needed for patient preparation). In the presence of several patients in

the queue (for example, when examining an organized group), the time to measure one person can be reduced. Anyway, based on our experience, we assumed that, with the BIA meter used, it is impossible to perform the consecutive measurements of any two different people in less than 1.5 minutes. Actually, the data shown in Fig. 2 represent three series of repeated measurements of one person with each series performed at a frequency of less than 1.5 minutes.

To check the performance of HCViewer, six additional BIA measurement datasets and two subsets of the 'fraudulent' HCs' BIA data were used to assess diagnostic effectiveness of the selection criteria for outliers and fraud cases. Diagnostic specificity of the selection criteria (i.e., the percentage of non-outliers and non-fraud cases among the knowingly reliable data) was estimated using the following data:

– on 3399 ethnically Russian children and adolescents aged 7–18 years from the European part of Russia (Moscow city, Arkhangelsk, Arkhangelsk region, and Elista) collected by trained anthropologists from Moscow University [16];

– on 3934 ethnically Russian adults and adolescents aged 16–86 years from Eastern Siberia (Krasnoyarsk region) collected by trained anthropologists from Krasnoyarsk State Medical Academy [31];

– on 538 adult tuberculosis (TB) patients from the Volga federal district of Russia (Cheboksary, Ioshkar-Ola, Saransk, and Ulyanovsk) collected by one of the authors in 2012 ($n = 240$) [29] and in 2015 ($n = 298$), unpublished data;

– on 104 staff members of TB dispensaries from the Volga federal district of Russia collected in 2015 by one of the authors, unpublished data.

The samples of TB patients and TB dispensaries staff members were included as representing the opposing groups with respect to body mass index and state of nutrition.

Diagnostic sensitivity of the selection criteria (i.e., the percentage of outliers and fraud cases among the knowingly unreliable data) was estimated using the following data:

– repeated self-measurements executed by one of the authors ($n = 100$);

– repeated measurements of the BIA meter's electronic verification module ($n = 100$);

– randomly selected data sequence from the 'fraudulent' HC X ($n = 524$) containing at least 418 (79.8%) of fraud records as was directly established by an expert;

– randomly selected data sequence from the 'fraudulent' HC Y ($n = 487$) containing at least 262 (53.8%) of fraud records as was directly established by an expert.

To develop HCViewer, the RStudio [26] open source integrated development environment for R [23] was used with the following statistical packages: shiny [25] to implement the user interface, sp [3] for visualization of geographical distributions of data, ggplot2 [37] for graphing, and leaflet [24] for interactive maps.

2. Results

On the Upload tab of the developed software program HCViewer 1.0, the user can import screening data stored in CSV format, select a subgroup of interest accord-

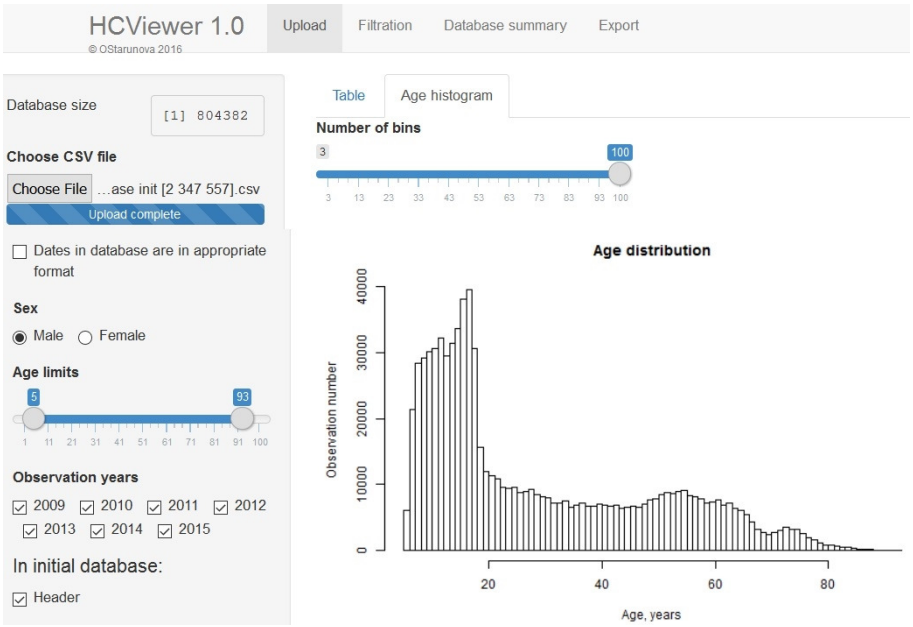

Figure 3. HCViewer: general information and the age distribution of the group under study, data for males (a screenshot of the user interface).

ing to sex, age and observation year, and visualize age distribution of the resulting sample. In Fig. 3, a screenshot of the user interface is shown with the age distribution for males in the initial BIA database (n=2,347,557).

The set of parameters for data filtration, and the selection criteria, are to be determined on the Filtration tab. Here, the admissible values of the parameters according to age can be loaded from a text file. Also on this tab, the user can choose the minimal value of the difference between the consecutive values of R50 and Xc50, respectively, at which the data will be considered as real measurements of patients (the default tolerance limits for fraud detection are 1% for R50 and 7% for Xc50).

After application of the selection criteria, a cross-sectional 2009–2015 database of BIA measurements of 1,268,005 individuals (411,323 males and 856,682 females) aged 5–96 years was obtained. The database accounted for 0.86% of the current Russian population. These data are originated from 325 HCs representing all federal districts of Russia and 62 out of 85 federal subjects in them. The data are presented in Fig. 4, which was constructed on the Filtration tab of the user interface.

In Table 2, the distribution of the incorrect HCs' BIA data according to federal district of Russia is presented. One can see that the percentage of incorrect data was large and varied significantly between federal districts: from 21.8% in the North Caucasian district to 60.0% in the Central district. In each federal district, the number of fraud cases exceeded the number of outliers. The leader in the number of the incorrect data, the Central federal district, has shown the highest fraud-to-outlier ratio of about 20:1 suggesting the most unfavorable situation with fraud cases and a relatively high quality of real BIA measurements in this district. The Southern

Table 2. Distribution of outliers and fraud cases in the HCs' BIA data by federal districts of Russia, 2009–2015.

Federal district	Incorrect data, abs			Incorrect data, %			All available data, abs
	Outliers	Fraud cases	Total	Outliers	Fraud cases	Total	
Central	30,365	631,659	662,024	2.8	57.2	60.0	1,104,248
Far Eastern	6,198	23,399	29,597	5.8	21.7	27.5	107,785
Northwestern	13,550	40,872	54,422	7.1	21.5	28.6	190,338
North Caucasian	3,769	10,817	14,586	5.6	16.2	21.8	66,962
Siberian	18,273	39,872	58,145	8.7	20.0	28.7	202,870
Southern	15,220	25,525	40,745	18.5	31.0	49.5	82,317
Ural	3,867	30,601	34,468	4.8	38.3	43.1	79,905
Volga	26,611	106,242	132,853	5.8	23.1	28.9	460,418
All federal districts	117,853	908,987	1,026,840	5.2	39.6	44.8	2,294,843

federal district, the second in the percentage of incorrect data, has shown maximal proportion of outliers suggesting high frequency of the measurement errors.

The percentage of the incorrect BIA data for males (48.0%) was slightly higher than for females (43.0%). The HCs were essentially heterogeneous in the percentage of the incorrect data, with 80.3% of outliers and fraud cases being generated by 20% of the HCs.

Of 908,987 fraud records detected in the HCs' BIA database (see Table 2), 62,659 (6.9%) were classified as the software emulation of measurement (1st-type fraud), and 136,455 (15.0%) as the measurement of the electronic verification module (2nd-type fraud). Of the much more prevalent 3rd-type fraud cases (i.e., multiple measurements of the same person under the guise of different), 596,320 were identified as the measurements performed unrealistically quickly (faster than 1 time in 1.5 minutes), and 743,007 as satisfying to, at least, one of the remaining filtering criteria showing close correlation between the empirical schemes of the 3rd-type fraud detection criteria.

The major part of the HCs' fraud cases in the Central federal district of Russia was generated in Moscow city, but the percentage of fraud cases was high in Tambov, Voronezh, and Bryansk regions as well (see Table 3). The most reliable BIA data, as judged by the low percentage of outliers and fraud cases, were generated by the HCs of Tver, Yaroslavl, Kostroma, and Vladimir regions. (However, the total number of measurements in these regions was relatively small as compared to Moscow.) After exclusion of the BIA data originated from Moscow, the rest of the data from the Central federal district have shown intermediate proportion of the incorrect data as compared to other federal districts (34.9%, including 5.4% outliers and 29.5% fraud cases).

The percentage of outliers did not change significantly with time while the proportion of fraud cases increased rapidly, from 11.5% in 2009 to 48.8–49.3% in 2013–2014 (see Fig. 5, left panel). In general, we observed steady growth in the proportion of the incorrect data between 2009 and 2013–2014 with the maximal growth rates in the Central and Southern federal districts while the other federal

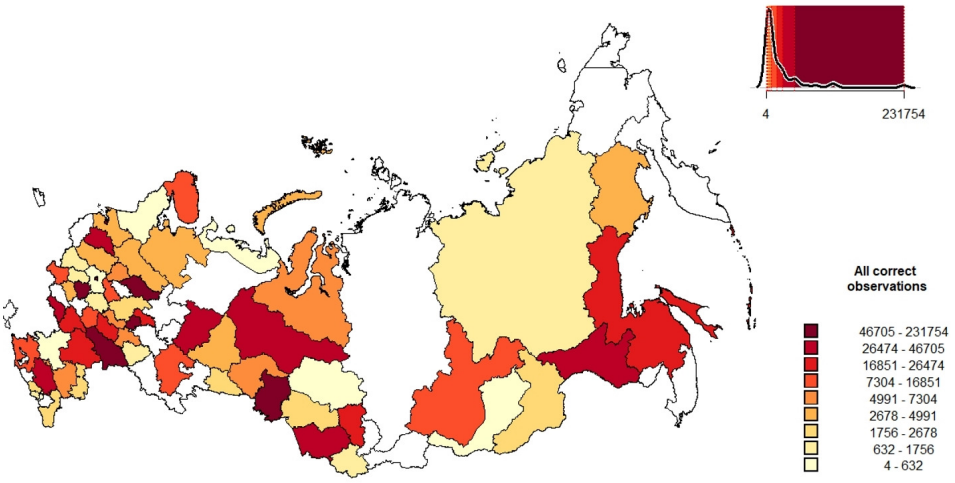

Figure 4. HCViewer: geographical distribution of the selected bioimpedance data from the HCs ($n = 1,268,003$) by federal subjects of Russia.

Table 3. Distribution of outliers and fraud cases in the HCs' BIA data from the Central federal district of Russia, by federal subjects (2009–2015).

Federal subject	Incorrect data, abs			Incorrect data, %			All available data, abs
	Outliers	Fraud cases	Total	Outliers	Fraud cases	Total	
Belgorod region	1,650	7,617	9,267	3.4	15.6	19.0	48,829
Bryansk region	1,097	18,044	19,141	3.1	51.5	54.6	35,048
Kaluga region	61	107	168	5.5	9.6	15.1	1,111
Kostroma region	3,791	2,190	5,981	7.1	4.0	11.0	54,278
Moscow city	13,135	536,256	549,391	1.7	68.6	70.3	781,145
Moscow region	115	87	202	14.0	10.6	24.6	822
Orel region	313	127	440	10.0	4.1	14.1	3,119
Ryazan region	125	489	614	9.9	38.9	48.8	1,258
Smolensk region	254	747	1,001	10.1	29.7	39.8	2,517
Tambov region	1,716	16,766	18,482	6.6	64.9	71.5	25,831
Tula region	6,138	10,295	16,433	8.8	14.8	23.6	69,757
Tver region	169	261	430	3.2	4.8	8.0	5,398
Vladimir region	248	828	1,076	2.8	9.2	12.0	8,959
Voronezh region	1,425	37,355	38,780	2.5	63.7	66.2	58,610
Yaroslavl region	128	490	618	1.7	6.5	8.2	7,566
All federal subjects	30,365	631,659	662,024	2.8	57.2	60.0	1,104,248

districts (e.g., Volga and Ural ones) showed different patterns of change (see Fig. 5, right panel). In 2015, the federal subject producing the highest number of fraud cases, the Moscow city, was severely under-represented. If the Moscow data were available, then the observed 2013–2014 total percentage of the incorrect data would remain.

With the selection criteria used, HCViewer showed high diagnostic specificity (95.2–98.9%) for the total amount of the incorrect data on the datasets of knowingly

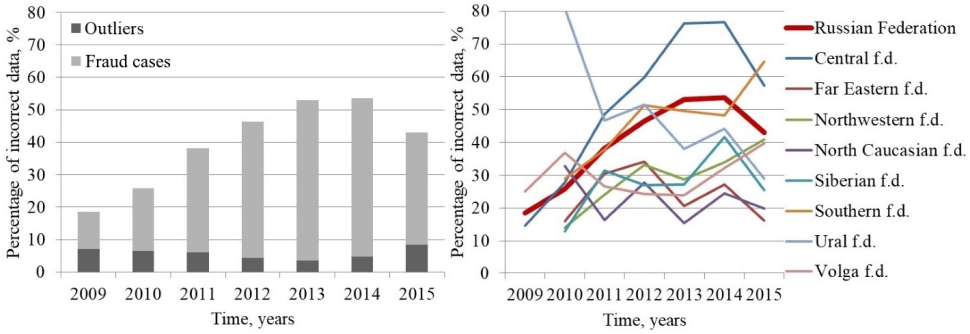

Figure 5. Dynamics of the percentage of outliers and fraud cases in the HCs' BIA database ($n = 2,294,843$, left panel), and the percentage of the incorrect BIA data by observation year and federal district of Russia (right panel).

Table 4. The performance of HCViewer on the datasets of knowingly reliable data.

BIA data	Outliers		Fraud cases		Total incorrect	
	abs	%	abs	%	abs	%
Healthy children and adolescents, $n = 3399$ [16]	4	0.12	33	0.97	37	1.09
Healthy adults and adolescents, $n = 3934$ [31]	27	0.69	163	4.14	190	4.83
TB adults, $n = 538$	2	0.37	20	3.72	22	4.09
TB dispensaries staff, $n = 104$	0	0.00	2	1.92	2	1.92

reliable bioimpedance data (see Table 4). Diagnostic specificity for fraud cases in healthy children and adolescents' group [16] (95.9%) was lower as compared to healthy adults and adolescents' group [31] (99.0%) partially reflecting greater variability of the BIA data in the former group. Also, only small percentage of outliers (less than 1%) was detected in each group suggesting good performance of HCViewer for outliers. The main part of outliers in the healthy adult and adolescents' group [31] was actually associated with the high values of the impedance phase angle due to measurement errors. High diagnostic specificity of HCViewer on the datasets of TB patients and TB dispensaries staff suggests good performance on the unusual samples significantly different from the general population.

The data presented in Table 4 suggest that the typical rate of false-positive results on the HCs' BIA data with the filtration criteria used could be of the order of 4–5%. With this correction for false alarm rate, the results presented in Table 2 give us final estimate of the proportion of fraud cases in the HCs' BIA data being from 5–7% in 2009 to 43–45% in 2013–2014 with an average of 35%. So, we obtain a new estimate of the total proportion of outliers and fraud cases in the HCs' BIA data which is about 40%. The relatively high proportion of outliers (5.1%) in the HCs' BIA data (see Table 2) coupled with the low proportion of outliers in our datasets of knowingly reliable data (Table 4), suggest a significant level of measurement errors arising during BIA measurements in the HCs thus indicating the potential importance of regular training and accreditation of the HCs' personnel.

HCViewer has shown absolute performance for fraud detection (100% sensitivity) on the datasets of repeated measurements of the electronic verification module

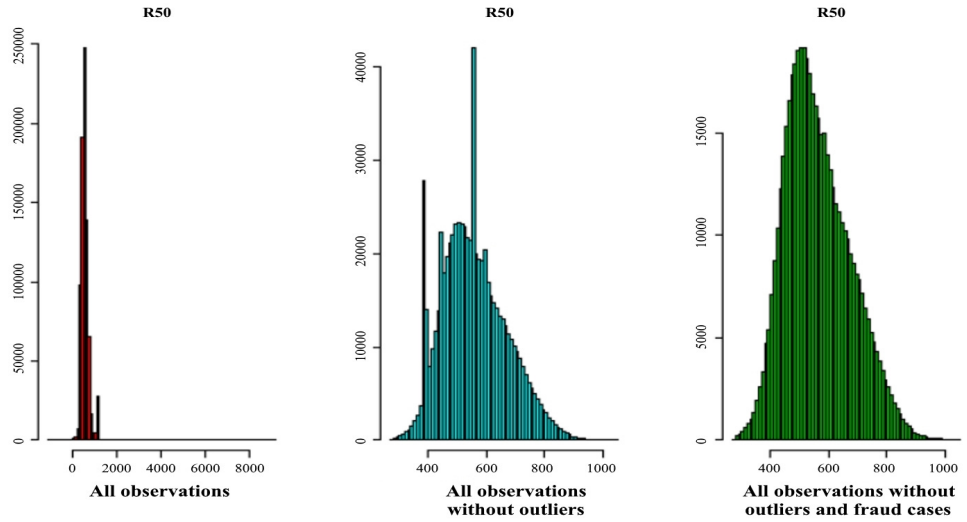

Figure 6. HCViewer: histograms of the resistance values (R50) for the initial BIA dataset (left panel) and after sequential removal of outliers (central panel) and fraud cases (right panel).

Table 5. The performance of HCViewer on the datasets of knowingly unreliable (or partially unreliable) data.

BIA data	Outliers		Fraud cases		Total incorrect	
	abs	%	abs	%	abs	%
Repeated self-measurements, $n = 100$	0	0	100	100	100	100
Repeated meas. of the verification module, $n = 100$	0	0	100	100	100	100
Data from the 'fraudulent' HC X, $n = 524$	4	0.76	452	86.3	456	87.0
Data from the 'fraudulent' HC Y, $n = 487$	5	1.03	285	58.5	290	59.5

and of repeated self-measurements executed by one of the authors. Also, by counting the number of true-positive cases on the subsets of fraud records of partially reliable data from the 'fraudulent' HCs X and Y, we observed high sensitivity (94.5% and 99.2%, respectively) of HCViewer for the 3rd-type fraud (Table 5).

Figure 6 illustrates consistent improvement of the data quality due to removal of outliers and fraud cases from the initial BIA dataset. Over the selection process, the frequency distribution of the parameter R50 evolved and finally adopted regular form thus suggesting higher degree of reliability of the selected data. The visualization option is accessible on the Database summary tab of the user interface.

The Export tab enables user to export selected data or annotated original data as a CSV file.

3. Discussion

Fraud remains a persistent problem and is increasing dramatically with the expansion of modern technology [5]. As an example, in the USA, where the health care system has become a major expenditure and now accounts for 17.1% of the GDP,

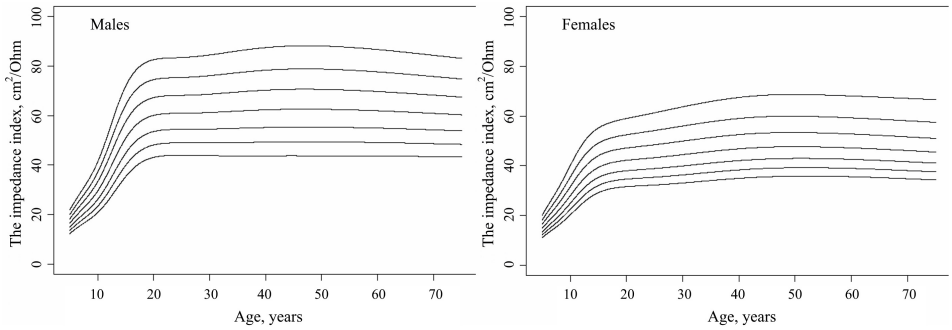

Figure 7. The impedance index ($Ht^2/R50$) growth charts (3rd, 10th, 25th, 50th, 75th, 90th, and 97th centiles) for the Russian general population according to the HCs' data for males (left panel) and females (right panel).

the losses due to outright fraud are estimated to be between 3 and 10%, from 60 to 170 billion dollars annually [20]. The same is applied to other countries [21], with estimated loss to health care fraud of 260 billion dollars per year, or approximately 6% of the global health care spending [10].

In the Russian health care system, fraud represents the largely unrecognized problem that can only worsen with the implementation of e-health. Our retrospective analysis of the raw mass BIA data from the national network of HCs showed that the absolute and relative numbers of fraud cases increased dramatically, from 5–7% in 2009 to 43–45% in 2013–2014. So, the final percentage of fraud cases greatly exceeded the above global average of 6% thus indicating inefficient management of the Russian preventive screening system and the urgent need to improve the efficiency of its function. One possible reason of the inefficiency was the neglect by the Russian health care system of the analysis of the raw mass data which represents a difficult task and can hardly be done manually due to an enormous volume of gathered information. Indeed, the official report on the activities of the HCs, the Form 68 of the sectoral statistical observation, is compiled using aggregated data from the HCs without any preliminary quality control. Similar practice is applied in the Russian clinical examination system of the adult population. Our results suggest incomparability and, therefore, uselessness of the related health statistics from some regions without a preliminary filtering of the underlying raw data.

Application of HCViewer has revealed greater proportion of the incorrect data and fraud cases as compared to our previous results [33] even with the correction for false alarm rate which was not taken into account before. This is explained by the recently discovered new and powerful source of incorrect data, namely, multiple measurements of the same person under the guise of different.

In fact, when developing HCViewer, we used an empirical approach to detect incorrect data based on familiarity of the authors with the features of the bioimpedance method and instrumentation, as well as with the HCs' software (see, e.g., [28, 32]). Further development of HCViewer will aim at the implementation of other methods and algorithms, e.g., Benford's law (see [34]) for data validation and the model GAMLSS [35], which is currently embedding into HCViewer in the test mode, for

generation of population-based reference data. An example of the GAMLSS model application to the dataset of the selected HCs' bioimpedance data is shown in Fig. 7 where the centile reference curves for the impedance index $Ht^2/R50$ are presented depending on age and sex. (The impedance index is known to be the best single predictor of total body water [17] and is used in most BIA body composition prediction formulae [18, 30].) The obtained reference data will be utilized for intergroup comparisons, data standardization, and the assessment of health risks.

Clearly, subject to minimal revision, HCViewer can be used not only for the retrospective analysis, but also for the online monitoring of quality of the HCs' data and generation of health statistics (e.g., as an add-on to the federal database of HCs). On the whole, fraud prevention should become the responsibility of the Russian health care system to ensure an efficiency of its function.

4. Conclusion

The prototype automated system for the analysis and processing raw mass data of preventive screening in the HCs, the software program HCViewer, is developed. Application of HCViewer to the HCs' BIA data has revealed rapid growth in the proportion of fraud cases suggesting insufficiency of the existing control measures. The potential utility of using HCViewer for the dynamical quality control of the Russian preventive screening data and for the online health monitoring is shown.

Acknowledgements

The authors are grateful to the staff of the Human Auxology lab of the MSU Research Institute and Museum of Anthropology, and its Head, Prof. E. Z. Godina, and also to Prof. L. V. Sindeyeva of the Krasnoyarsk State Medical Academy, for the permission to use their original anthropometric data. The authors thank two anonymous reviewers for their helpful comments and suggestions to improve the manuscript.

References

1. D. Aune, A. Sen, M. Prasad, et al., BMI and all cause mortality: systematic review and non-linear dose-response meta-analysis of 230 cohort studies with 3.74 million deaths among 30.3 million participants. *BMJ* **353** (2016) i2156.
2. M. C. G. Barbosa-Silva, A. J. D. Barros, J. Wang, S. B. Heymsfield, and R. N. Pierson, Bioelectrical impedance analysis: population reference values for phase angle by age and sex. *Amer. J. Clin. Nutr.* **82** (2005) 49–52.
3. R. S. Bivand, E. Pebesma, and V. Gomez-Rubio, *Applied spatial data analysis with R*. Springer, New York, 2013. http://www.asdar-book.org/
4. A. Böhm and B. L. Heitmann, The use of bioelectrical impedance analysis for body composition in epidemiological studies. *Eur. J. Clin. Nutr.* **67** (2013) S79–S85.
5. R. J. Bolton and D. J. Hand, Statistical fraud detection: a review. *Stat. Sci.* **17** (2002) 235–249.

6. S. Boytsov and R. A. Potemkina, Preventive measures for public health in Russian Federation. *Eur. Heart J. Suppl.* **16** (2014) A84–A86.
7. W. C. Chumlea, S. S. Guo, R. J. Kuczmarski, et al., Body composition estimates from NHANES III bioelectrical impedance data. *Int. J. Obes.* **26** (2002) 1596–1609.
8. G. Danaei, M. M. Finucane, J. K. Lin, G. M. Singh, et al., National, regional, and global trends in systolic blood pressure since 1980: systematic analysis of health examination surveys and epidemiological studies with 786 country-years and 5.4 million participants. *Lancet* **377** (2011) 568–577.
9. F. M. Franssen, E. P. Rutten, M. T. Groenen, et al., New reference values for body composition by bioelectrical impedance analysis in the general population: results from the UK Biobank. *J. Amer. Med. Dir. Assoc.* **15** (2014) 448.e1–6.
10. Global Health Care Antifraud Network. *The Health Care Fraud Challenge: Worldwide, Health Care Fraud is a Lucrative, Though Illicit, Line of Work*. (2017) URL: http://www.ghcan.org/global-anti-fraud-resources/the-health-care-fraud-challenge/
11. S. Grimnes and O. G. Martinsen, *Bioimpedance and Bioelectricity Basics*. Academic Press, London, 2014.
12. N. F. Izmerov, G. I. Tikhonova, and T. Yu. Gorchakova, Mortality of working age population in Russia and industrial countries in Europe: trends of the last two decades. *Vestnik RAMN* **69** (2014) 121–126 (in Russian).
13. I. Janssen, R. N. Baumgartner, R. Ross, I. H. Rosenberg, and R. Roubenoff, Skeletal muscle cutpoints associated with elevated physical disability risk in older men and women. *Amer. J. Epidemiol.* **159** (2004) 413–421.
14. W. Johnson, N. Cameron, P. Dickson, S. Emsley, P. Raynor, C. Seymour, and J. Wright, The reliability of routine anthropometric data collected by health workers: A cross-sectional study. *Int. J. Nurs. Stud.* **46** (2009) 310–316.
15. E. Jopp, C. Scheffler, and M. Hermanussen, Prevention and anthropology. *Anth. Anz.* **71** (2014) 135–141.
16. V. A. Kolesnikov, S. G. Rudnev, D. V. Nikolaev, A. V. Anisimova, and E. Z. Godina, On a new protocol of the Heath-Carter somatotype assessment in the software for body composition bioimpedance analyzer. *Vestnik Moskovskogo Universiteta Ser. 23: Antropologia* **4** (2016) 4–13 (in Russian).
17. R. F. Kushner, D. A. Schoeller, C. R. Fjeld, and L. Danford, Is the impedance index (ht^2/R) significant in predicting total body water? *Amer. J. Clin. Nutr.* **56** (1992) 835–839.
18. U. G. Kyle, I. Bosaeus, A. D. DeLorenzo, P. Deurenberg, et al., Bioelectrical impedance analysis—part I: review of principles and methods. *Clin. Nutr.* **23** (2004) 1226–1243.
19. U. G. Kyle, I. Bosaeus, A. D. DeLorenzo, P. Deurenberg, et al., Bioelectrical impedance analysis—part II: utilization in clinical practice. *Clin. Nutr.* **23** (2004) 1430–1453.
20. J. Li, K.-Y. Huang, J. Jin, and J. Shi, A survey on statistical methods for health care fraud detection. *Health Care Manage. Sci.* **11** (2008) 275–287.
21. M. Mikkers, W. Sauter, and J. Boertjens (Eds.), *Healthcare Fraud, Corruption and Waste in Europe: National and Academic Perspectives*. Eleven International Publishing, Odense, 2017.
22. NCD Risk Factors Collaboration, Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19.2 million participants. *Lancet* **387** (2016) 1377–1396.
23. R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL: https://www.R-project.org/

24. RStudio and Inc., Leaflet: Create interactive web maps with the JavaScript 'leaflet' library, R package version 1.1.0, 2016. URL: http://rstudio.github.io/leaflet/

25. RStudio and Inc., Shiny: Web application framework for R, R package version 1.0.5, 2016. URL: http://shiny.rstudio.com/

26. RStudio Team, *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2016. URL: http://www.rstudio.com/

27. S. G. Rudnev, D. V. Nikolaev, K. A. Korostylev, O. A. Starunova, et al., Health Centres: technology to process mass data of preventive screening. *Social Aspects of Population Health* **46** (2015) 1. URL: http://vestnik.mednet.ru/content/view/716/30/lang,ru/ (in Russian).

28. S. G. Rudnev, N. P. Soboleva, S. A. Sterlikov, D. V. Nikolaev, et al., *Bioimpedance Study of Body Composition in the Russian Population*. Federal Research Institute for Health Organization and Informatics, Moscow, 2014 (in Russian).

29. S. G. Rudnev, S. A. Sterlikov, A. M. Vasil'eva, Zh. V. Elenkina, A. K. Larionov, and D. V. Nikolaev, Bioimpedance study of body composition in tuberculosis patients. *Tuberc. Lung. Dis.* **93** (2015) 33–40 (in Russian).

30. A. M. Silva, D. A. Fields, and L. B. Sardinha, A PRISMA-driven systematic review of predictive equations for assessing fat and fat-free mass in healthy children and adolescents using multicomponent molecular models as the reference method. *J. Obes.* (2013) 2013:148696.

31. L. V. Sindeyeva and S. G. Rudnev, Characteristics of age and sex-related variability of the Heath-Carter somatotype in adults and the possibility of its bioimpedance assessment (as exemplified by the Russian population of Eastern Siberia). *Morfologia* **151** (2017) 77–87 (in Russian).

32. V. I. Starodubov, S. G. Rudnev, D. V. Nikolaev, and K. A. Korostylev, The Federal Information Resource of health centers: current state and developmental perspectives. *Social Aspects of Population Health* **45** (2015) 1. URL: http://vestnik.mednet.ru/content/view/706/27/lang,ru/ (in Russian).

33. V. I. Starodubov, S. G. Rudnev, D. V. Nikolaev, and K. A. Korostylev, On the quality of preventive screening data in Health Centers and method to increase the efficiency of budget expenses. *Analytical Bulletin of the Federation Council of the Federal Assembly of the Russian Federation* **44** (2015) 43–49 (in Russian).

34. O. A. Starunova, On the method of verification of the Russian preventive screening data from the health centers. In: *Proc. 10th Int. Workshop 'Science and Innovation-2015' (03-12 July, 2015). Ioshkar-Ola, Volga State Technological University.* 2015, pp. 269–274 (in Russian).

35. D. M. Stasinopoulos and R. A. Rigby, Generalized additive models for location, scale and shape (GAMLSS) in R. *J. Stat. Software* **23** (2007) 1–46.

36. J. C. K. Wells and M. K. Shirley, Body composition and the monitoring of non-communicable chronic disease risk. *Global Health Epidemiol. Genomics* **1** (2016) e18.

37. H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2009.

38. World Health Organization, *Global Status Report on Noncommunicable Diseases 2014*. WHO, Geneva, 2014.

39. S. Zhu, Z. Wang, W. Shen, S. B. Heymsfield, and S. Heshka, Percentage body fat ranges associated with metabolic syndrome risk: results based on the third National Health and Nutrition Examination Survey (1988–1994). *Amer. J. Clin. Nutr.* **78** (2003) 228–235.