scientific reports

OPEN

Seroprevalence of SARS-CoV-2 antibodies in Saint Petersburg, Russia: a population-based study

Anton Barchuk[1,6,7✉], Dmitriy Skougarevskiy[1], Kirill Titaev[1], Daniil Shirokov[2,8], Yulia Raskina[1], Anastasia Novkunkskaya[1], Petr Talantov[3], Artur Isaev[5], Ekaterina Pomerantseva[4], Svetlana Zhikrivetskaya[4], Lubov Barabanova[2] & Vadim Volkov[1]

Properly conducted serological survey can help determine infection disease true spread. This study aims to estimate the seroprevalence of SARS-CoV-2 antibodies in Saint Petersburg, Russia accounting for non-response bias. A sample of adults was recruited with random digit dialling, interviewed and invited for anti-SARS-CoV-2 antibodies. The seroprevalence was corrected with the aid of the bivariate probit model that jointly estimated individual propensity to agree to participate in the survey and seropositivity. 66,250 individuals were contacted, 6,440 adults agreed to be interviewed and blood samples were obtained from 1,038 participants between May 27 and June 26, 2020. Naïve seroprevalence corrected for test characteristics was 9.0% (7.2–10.8) by CMIA and 10.5% (8.6–12.4) by ELISA. Correction for non-response decreased estimates to 7.4% (5.7–9.2) and 9.1% (7.2–10.9) for CMIA and ELISA, respectively. The most pronounced decrease in bias-corrected seroprevalence was attributed to the history of any illnesses in the past 3 months and COVID-19 testing. Seroconversion was negatively associated with smoking status, self-reported history of allergies and changes in hand-washing habits. These results suggest that even low estimates of seroprevalence can be an overestimation. Serosurvey design should attempt to identify characteristics that are associated both with participation and seropositivity.

Serological surveys in the midst of COVID-19 pandemic address the issue of underestimation of the number of cases registered officially with RT-PCR using material from nasopharyngeal swabs[1,2]. They use blood antibody tests that are markers of past infection. WHO recommends serological surveys to monitor COVID-19 spread[3]. However, estimates from serological surveys can be also biased. Estimates can be distorted by non-response bias, non-representativeness of the study sample, and imperfect test characteristics. Previous serological surveys so far have all but focused on the former[4–10]. This poses a significant problem when some observed factors that influence the decision to participate in the survey may be also associated with test results[11]. Non-response or self-selection bias has been widely acknowledged in descriptive epidemiology[12–15]. In particular, it has been predominantly addressed in seroprevalence surveys of HIV[16].

In this paper we present seroprevalence estimates coming from the first cross-sectional data of our longitudinal study with serial sampling to assess the spread COVID-19 in Saint Petersburg, Russia conducted between May 27 and June 26 2020. St. Petersburg is the second largest city in the country and fourth largest in Europe with the population of approximately 5.2 mln. The first case in the city was registered on 5 March, 2020 and 36,667 cases (7.1 per 1000) were reported as of 31 August, 2020. The study of the spread of COVID-19 in St. Petersburg was established to estimate the extent of epidemic in a population-based manner, and, to the best of our knowledge, this was the first COVID-19 serological survey in the country. Our primary aim was to compare naïve and non-response bias-adjusted seroprevalence to show the utmost importance of rigorous serosurvey designs. We report

[1]Institute for Interdisciplinary Health Research, European University at St. Petersburg, Shpalernaya Ulitsa, 1, St. Petersburg, Russia 191187. [2]Clinic "Scandinavia" (LLC Ava-Peter), Ilyushina Ulitsa, 4-1, St. Petersburg, Russia 197372. [3]The Russian Academy of Sciences Commission for Counteracting the Falsification of Scientific Research, Leingradsky Prospekt, 14, Moscow, Russia 119991. [4]Center of Genetics and Reproductive Medicine GENETICO LLC, Ulitsa Gubkina, 3-1, Moscow, Russia 119333. [5]Human Stem Cells Institute, Ulitsa Gubkina, 3-2, Moscow, Russia 119333. [6]N.N. Petrov National Research Medical Center of Oncology, Pesochny, Leningradskaya Ulitsa, 68, St. Petersburg, Russia 197758. [7]Health Sciences, Faculty of Social Sciences, Tampere University, Arvo Ylpön katu 34, 33520 Tampere, Finland. [8]ITMO University, Kronverksky Prospekt, 49, St. Petersburg, Russia 197101. ✉email: abarchuk@eu.spb.ru

how various observable characteristics of individuals shift the naïve prevalence estimates when accounted for and carefully address possible sources of bias. Finally, we provide observable characteristics of surveyed individuals that are associated with risk of seroconversion in a population-based study.

Methods

Study design and participants. The St. Petersburg COVID-19 study is a population-based epidemiological survey of random sample from the adult population to assess the seroprevalence of anti-SARS-CoV-2 antibodies. The study is conducted as a longitudinal study with serial sampling from the same individuals. The study involved one phone-based survey followed by an individual invitation to the clinic, one paper-based survey, and blood sample collection for antibody testing. Interviews were carried out between May 21, 2020 and June 25, 2020. Blood samples were collected between May 27, 2020 and June 26, 2020.

Eligible individuals were adults residing in St. Petersburg older than 18 years and recruited using the random digit dialling (RDD) method. RDD was accompanied by the computer assisted telephone interviewing (CATI) in order to collect information on both individuals who accepted and declined invitation for testing. Residents of St. Petersburg are almost universal mobile phone users, with 99.5% of households having mobile phones as of 2016 (see Supplementary Appendix Table A3). Participants from six distant districts of the city located too far away from the test site were excluded leaving 12 central districts of the city with population of approximately 4.3 mln. The full study protocol is available online (https://eusp.org/sites/default/files/inline-files/EU_SG-Russian-Covid-Serosurvey-Protocol-CDRU-001_en.pdf).

Procedures. RDD was carried out using area prefixes of mobile phone numbers to include only mobile phone users in St. Petersburg. The individuals who had answered the call were asked to answer 25 questions on demographics, marital status, education level, income level, past history of illnesses, travelling abroad, household size, social contacts, and visits to public places during lockdown (see full questionnaire in the study protocol). Refusal to participate in blood sampling was also recorded. We have also randomly incentivized respondents to participate in the study by offering complimentary taxi transit to and from the clinic test site for approximately 25% of those who agreed to go through CATI.

Those who had agreed to take part in antibody testing were later contacted by the clinic call center and were assigned an appointment date for blood sampling. The participants signed informed consent forms and filled out additional paper-based survey forms in the clinic on the day of the visit. Forms included question on the medical history, history of allergies, smoking, alcohol consumption, chronic diseases and medication taken regularly. Blood sampling started on May 27, 2020 and was planned for two weeks but was prolonged till June 26, 2020 because of low participation rates.

Laboratory tests. We assessed anti-SARS-CoV-2 antibodies using two tests. Serum samples were tested using chemiluminescent microparticle immunoassay (CMIA) Abbott Architect SARS-CoV-2 IgG on the Abbott ARCHITECT i2000sr platform (Abbott Laboratories, Chicago, USA) that detects immunoglobulin class G (IgG) antibodies to the nucleocapsid protein of SARS-CoV-2 (cutoff for positivity 1.4). In addition to that blood samples were also tested by enzyme-linked immunosorbent assay (ELISA) using CoronaPass total antibodies test (Genetico, Moscow, Russia) that detects total antibodies (cutoff for positivity 1.0) and is based on recombinant receptor binding domain of the spike protein of SARS-CoV-2 (Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA). We simultaneously report seroprevalence based on CMIA and ELISA.

Sample size. Initial sample size of 1550 participants was calculated assuming prevalence of 20% and test sensitivity (100%) and specificity (99.6%) for our CMIA test with sampling error was 2% using a 95% confidence interval (see Supplementary Appendix Fig. A1)[17]. After receiving the preliminary results (for 500 individuals), we reduced the sample size by assuming 10% prevalence that gave us a target sample size of 882 participants, that was rounded to 1000 participants.

Statistical analysis. The primary aim of the study was to assess the seroprevalence of antibodies to SARS-CoV-2 in serum samples based on CMIA tests and ELISA tests accounting for non-response bias and test characteristics (sensitivity and specificity). Seroprevalence was defined as the proportion of those tested positive to all participants. Non-response was assessed by comparison of answers provided during the CATI by those visited the test site and all other surveyed.

To understand the direction of non-response bias in our data we estimated a binomial probit regression of individual agreement to participate in the study and offer his/her blood sample on observable characteristics. We used this fitted model to compute conditional probability to participate in the study (holding all but one variable at mean levels at a time). Our bivariate probit model is formally introduced in Statistical Appendix).

We analyse variables obtained from CATI and the clinic paper-based survey (ordered or unordered factor variables), and results of antibody tests (binary variables). Participant age was split into groups (18–34, 35–49, 50–64, or 65 years old).

In the secondary analyses we also assessed seroprevalence by week based on the date of interview and the date of blood sampling. In subgroup analysis we first compared seroprevalence estimates corrected for non-response between different groups of individuals based on their answers in CATI. To explore individual risk factors for test positivity and obtain prevalence ratios we estimated a generalised linear model with Poisson distribution and a log link restricted to data from participants who completed clinic paper-based survey. We have entertained the possibility to use robust variance-covariance matrix in our adjusted prevalence ratio analysis. However, such
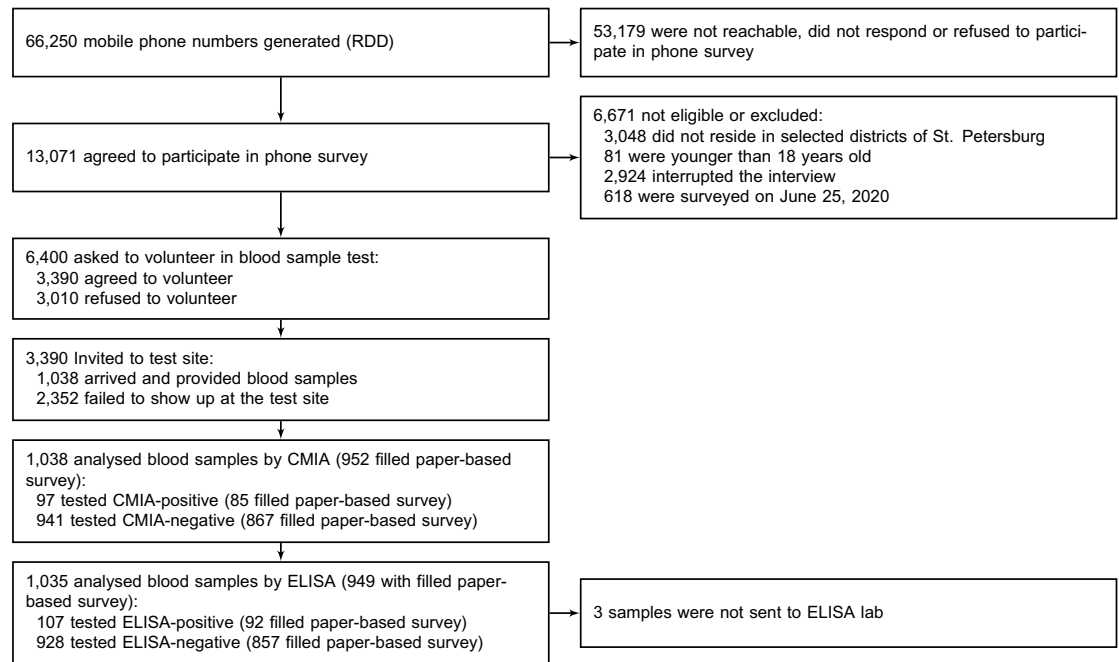
```
┌─────────────────────────────────────────┐         ┌──────────────────────────────────────────┐
│ 66,250 mobile phone numbers generated    │ ──────→ │ 53,179 were not reachable, did not respond │
│ (RDD)                                     │         │ or refused to participate in phone survey  │
└─────────────────────────────────────────┘         └──────────────────────────────────────────┘
                   │
                   ↓
┌─────────────────────────────────────────┐         ┌──────────────────────────────────────────┐
│ 13,071 agreed to participate in phone    │ ──────→ │ 6,671 not eligible or excluded:            │
│ survey                                    │         │   3,048 did not reside in selected         │
└─────────────────────────────────────────┘         │     districts of St. Petersburg            │
                   │                                 │   81 were younger than 18 years old        │
                   ↓                                 │   2,924 interrupted the interview          │
┌─────────────────────────────────────────┐         │   618 were surveyed on June 25, 2020       │
│ 6,400 asked to volunteer in blood sample │         └──────────────────────────────────────────┘
│ test:                                     │
│   3,390 agreed to volunteer               │
│   3,010 refused to volunteer              │
└─────────────────────────────────────────┘
                   │
                   ↓
┌─────────────────────────────────────────┐
│ 3,390 Invited to test site:              │
│   1,038 arrived and provided blood        │
│     samples                               │
│   2,352 failed to show up at the test     │
│     site                                  │
└─────────────────────────────────────────┘
                   │
                   ↓
┌─────────────────────────────────────────┐
│ 1,038 analysed blood samples by CMIA      │
│ (952 filled paper-based survey):          │
│   97 tested CMIA-positive (85 filled      │
│     paper-based survey)                   │
│   941 tested CMIA-negative (867 filled    │
│     paper-based survey)                   │
└─────────────────────────────────────────┘
                   │
                   ↓
┌─────────────────────────────────────────┐         ┌──────────────────────────────────────────┐
│ 1,035 analysed blood samples by ELISA     │ ──────→ │ 3 samples were not sent to ELISA lab       │
│ (949 with filled paper-based survey):     │         └──────────────────────────────────────────┘
│   107 tested ELISA-positive (92 filled    │
│     paper-based survey)                   │
│   928 tested ELISA-negative (857 filled   │
│     paper-based survey)                   │
└─────────────────────────────────────────┘
```

Figure 1. Flow chart of participants' progress through the St. Petersburg seroprevalence study.

adjustment narrowed the confidence intervals rendering our adjusted estimates less conservative[18]. For this reason we report confidence intervals from the unadjusted variance-covariance matrix.

In sensitivity analysis we explored how inclusion of different sets of observable characteristics of individuals (namely, travel history, face mask use, public transport use, visits to public places and others) in the model that corrected seroprevalence for non-response influenced the results. We also applied alternative definitions of seroprevalence (test combination either favouring sensitivity or specificity). To account for possible sample non-representativeness in sensitivity analysis we computed raking weights to match the survey age group and educational attainment proportions in 2016 representative survey of adult city population (see Supplementary Appendix Table A3 for description of this survey and the target proportions). R package anesrake was used to compute the weights[19]. We then estimated seroprevalence on re-weighted data.

We treated refusals to answer certain phone or paper-based survey questions as missing data, for this reason the results onwards are considered after listwise deletion of observations with missing variables.

All reported seroprevalence results were also corrected for test characteristics using the manufacturer's validation data—sensitivity (100% and 98.7%) and specificity (99.6% and 100%) for CMIA and ELISA test, respectively[20]. Standard errors were computed with delta method. Detailed description of statistical analysis is provided in Statistical Appendix).

Data sharing. All analyses were conducted in R[21] with the aid of GJRM package[22], study data and code is available online (https://github.com/eusporg/spb_covid_study20).

Ethical considerations and study registration. The study was approved by the Research Planning Board of European University at St. Petersburg (on May 20, 2020) and the Ethic Committee of the Clinic "Scandinavia" (on May 26, 2020). All research was performed in accordance with the relevant guidelines and regulations. Informed consent was obtained from all participants of the study. The study was registered with the following identifiers: Clinicaltrials.gov (NCT04406038, submitted on May 26, 2020, date of registration—May 28, 2020) and ISRCTN registry (ISRCTN11060415, submitted on May 26, 2020, date of registration—May 28, 2020).

Results

Participation rates. Between May 21 and June 25, 2020 66,250 individuals were reached using RDD. Of 13,071 respondents agreed to participate in the CATI 6,671 were excluded for various reasons (see Fig. 1). The resulting 6,400 individuals responded to CATI questionnaire (see Supplementary Appendix Table A2 for details regarding missing records on variables of interest). The respondents were representative of the city population in terms of their gender, employment status, and household size, but were younger than the adult city population as of 2016 and had higher levels of educational attainment (see Supplementary Appendix Table A3).

3,390 of surveyed individuals agreed to receive a phone call from the clinic and schedule a visit for antibody testing. Between May 27 and June 26, 2020 only 1038 individuals that satisfied eligibility criteria visited the clinic and provided blood samples (16.2% and 30.6% of those who were interviewed and agreed to participate in serosurvey, respectively). The rest declined the invitation or did not show up at the test site. 1038 CMIA tests

Regressors included in bivariate probit model	CMIA				ELISA			
	Number of participants		Seroprevalence (95% CI)		Number of participants		Seroprevalence (95% CI)	
	Interviewed	Tested	Naïve	Single imputation	Interviewed	Tested	Naïve	Single imputation
Demographic characteristics	6400	1038	9.0% (7.2–10.8)	8.7% (7.0–10.5)	6397	1035	10.5% (8.6–12.4)	10.1% (8.3–12.0)
Demographic and socioeconomic characteristics	6063	999	9.2% (7.4–11.1)	9.0% (7.0–11.0)	6061	997	10.8% (8.8–12.7)	10.7% (8.6–12.9)
Characteristics associated with seropositivity	6267	1026	9.0% (7.2–10.8)	7.1% (5.6–8.7)	6264	1023	10.5% (8.6–12.4)	8.6% (6.9–10.3)
Demographics, socioeconomic status and characteristics associated with seropositivity	5953	990	9.2% (7.4–11.1)	7.4% (5.7–9.2)	5951	988	10.8% (8.8–12.7)	9.1% (7.2–10.9)

Table 1. SARS-CoV-2 seroprevalence estimates from bivariate probit models with different sets of individual characteristics for non-response correction. "Demographic characteristics" means the following variables: individual age group (18–34, 35–49, 50–64, 65+ years old) and sex. "Socioeconomic characteristics" means the following variables: higher education status and higher self-reported income level. "Characteristics associated with seropositivity" means the following variables: history of illness in the last 3 months, history of COVID-19 testing, whether respondent lives alone, change in hand washing habits during pandemic, week of the phone interview, and city district. All models include a variable indicating random offer of taxi transportation to and from the clinic test site for interviewed participants. All estimates are corrected for tests characteristics (see Statistical appendix for details).

and 1035 ELISA tests were eventually performed on eligible individuals. The clinic-visiting participants have also filled out 965 clinic paper-based survey forms.

652 (62.8%) of 1,038 participants were women; 396 (38.2%) were aged 18–34 years, 357 (34.4%) were aged 35–49 years, 218 (21.0%) were aged 50–64 years, and 67 (6.5%) were older than 65 years, the majority of participants lived in multiple-person households, 843 (81.2%) (see Supplementary Appendix Table A2 for summary statistics on phone survey respondents and tested individuals).

In the course of the study we observed the gradual attrition of participants. Compared with the individuals who limited their participation to the CATI, participants who took part in antibody testing were younger, more likely to be female, report a higher education level, experience illnesses in the previous 3 months, report a history of previous COVID-19 testing and a change in their hand-washing habits during the epidemic. Our attempt to randomly incentivize respondents to take part in the study by offering taxi did not reach its purpose (see Supplementary Appendix Fig. A2a).

Seroprevalence estimates.   Between May 27 and June 26, 2020, 115 positive results were reported by any test (97 positive tests out of 1038 were reported by CMIA and 107 positive tests out of 1035 were reported by ELISA). 30 of these 115 (26.1%) individuals with any positive test result did not report any symptoms of past illnesses in the previous 3 months. Naïve seroprevalence corrected for test specificity and sensitivity was 9.0% (95% CI 7.2–10.8) by CMIA and 10.8% (8.8–12.7) by ELISA (see Table 1). When we accounted for non-response bias with respect to demographic and socioeconomic characteristics our seroprevalence point estimates did not change considerably. Inclusion of characteristics associated with seroprevalence as regressors in our single imputation model shifted point estimates of seroprevalence downwards and after adjustment for all aforementioned characteristics in the model seroprevalence was 7.4% (95% CI 5.7–9.2) for CMIA and to 9.3% (7.4–11.2) for ELISA.

Secondary subgroup analysis.   Seroprevalence was similar between men and women and was slightly lower in the older (65+) age group (see Table 2). The seroprevalence was higher for individuals who reported past history of illnesses—15.1% (95% CI 11.6–18.6) for CMIA and 20.0% (95% CI 14.8–25.2) for ELISA) compared to those who did not (3.8% (95% CI 2.1–5.5 for CMIA and 7.4% (95% CI 5.4–9.3 for ELISA). It was also higher for individuals who reported past history of COVID-19 tests, but was slightly lower in individuals who reported that they started washing hands more often since the onset of pandemic and lived alone. There was noticeable variation in seropositivity between city districts (see Fig. 2).

We observed a slight increase in seroprevalence by the week of the phone interview (see Fig. 3a) and by the week of the blood draw (see Fig. 3a).

Our secondary analysis of participants who filled out clinic paper-based survey forms revealed additional covariates associated with seroconversion. It was negatively associated with smoking status with prevalence ratios 0.46 (95% CI 0.22–0.87) and 0.34 (95% CI 0.14–0.72) (PR for current smokers vs non-smokers based on CMIA and ELISA, respectively), and self-reported history of allergies with prevalence ratios 0.54 (95% CI 0.30–0.90) and 0.53 (95% CI 0.28–0.93) (see Table 3).

Sensitivity analysis.   Alternative definitions of seroprevalence (test combination either favouring sensitivity or specificity) did not qualitatively change the effect of non-response bias (see Supplementary Appendix Table A4). Seroprevalence estimates obtained on re-weighted survey data (based on age group and education attainment level) were similar to estimates from the main analysis (see Supplementary Appendix Table A5).

	CMIA			ELISA		
	Number of participants		Seroprevalence (95% CI)	Number of participants		Seroprevalence (95% CI)
	Interviewed	Tested		Interviewed	Tested	
Overall	5953	990	7.4% (5.7–9.2)	5951	988	9.1% (7.2–10.9)
Age groups						
18–34	2228	388	7.8% (5.2–10.5)	2227	387	11.3% (8.1–14.4)
35–49	1916	342	6.5% (4.0–9.0)	1915	341	7.4% (4.8–10.1)
50–64	1159	199	10% (6.0–14.0)	1159	199	10.8% (6.6–15.0)
65+	650	61	4.1% (0.0–8.8)	650	61	3.1% (0.0–7.2)
Sex						
Female	3505	623	7.5% (5.5–9.6)	3505	623	8.7% (6.5–10.9)
Male	2448	367	7.3% (4.6–9.9)	2446	365	9.5% (6.6–12.5)
Higher education						
No	1928	169	7.4% (3.8–10.9)	1927	168	9.7% (5.7–13.7)
Yes	4025	821	7.5% (5.7–9.3)	4024	820	8.7% (6.8–10.6)
Higher income						
No	3402	491	6.6% (4.4–8.8)	3402	491	8.6% (6.2–11)
Yes	2551	499	8.5% (6.1–11.0)	2549	497	9.7% (7.1–12.3)
Respondent lives alone						
No	4857	805	8.0% (6.0–9.9)	4855	803	9.8% (7.7–12.0)
Yes	1096	185	5.1% (2.0–8.1)	1096	185	5.5% (2.5–8.6)
History of illness in the last 3 months						
No	4047	548	3.8% (2.1–5.5)	4046	547	5.0% (3.1–7.0)
Yes	1906	442	15.1% (11.6–18.6)	1905	441	17.6% (13.9–21.3)
History of COVID-19 testing						
No	5038	762	5.4% (3.7–7.1)	5036	760	7.2% (5.2–9.1)
Yes	915	228	18.6% (13.6–23.6)	915	228	19.4% (14.4–24.5)
Change in hand washing habits during pandemic						
No	2029	279	9.6% (6.3–12.9)	2029	279	11.8% (8.2–15.4)
Yes	3924	711	6.3% (4.5–8.1)	3922	709	7.6% (5.7–9.6)

Table 2. Seroprevalence of SARS-CoV-2 in subgroups of participants. All estimates are from the model that includes demographics, socioeconomic status and characteristics associated with seropositivity. All estimates are corrected for test sensitivity and specificity (see Statistical appendix for details).
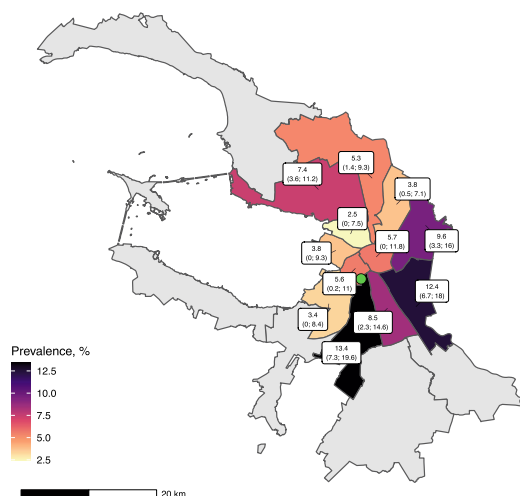
Figure 2. Prevalence estimates by district. This map shows CMIA-based prevalence estimates corrected for participation bias by surveyed districts with 95% CIs in parentheses. Green dot is the clinictest site location. Remote districts excluded from survey are in grey. This map was created with the aid of ggplot2[23], sf[24], and ggspatial[25] packages in R[21]using OpenStreetMap data[26].

(a) Prevalence by phone interview week

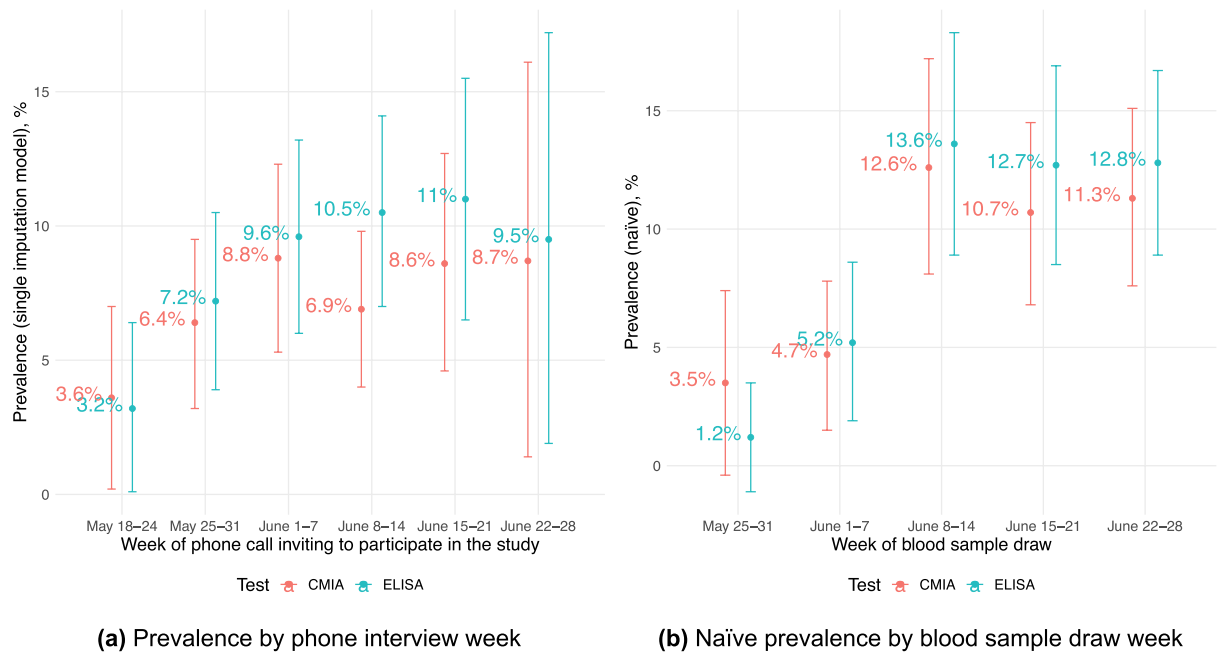**(b)** Naïve prevalence by blood sample draw week

Figure 3. Prevalence estimates over time.

Discussion

Our study aimed to assess the spread of epidemic in the fourth largest European city—St. Petersburg. This is the first population-based serological survey estimating COVID-19 spread in Russia and one of the few representative population-based studies in Europe Although the seroprevalence estimate varied based on the test used and type of correction applied, the total number of population with detectable antibodies was still far lower than the proportion needed for herd immunity. Overall seroprevalence in the range between 7% and 10% was in line with the results obtained from the previous studies and provides evidence of the similar epidemic development across the world with less than one tenth of population affected in the first months[5,6].

To the best of our knowledge, this is the first seroprevalence survey of COVID-19 that applied correction based on characteristics that are associated with the risk of seropositivity in combination with incentivised participation. Early COVID-19 serological surveys are likely to exhibit high sampling error because of recruitment methods[27]. Population based studies with random sampling relied on probability weighting obtained from the comparison with the source population[5–7]. Our findings show that even low estimates of seroprevalence (around or below 10%) obtained in population surveys can be an overestimation in populations with high risk of non-response bias.

We detected only a slight change in the estimate of seroprevalence when we corrected our estimated for non-response bias with respect to demographic or socioeconomic characteristics, but far more significant difference was detected when several behavioural characteristics were included in models and applied in the correction. In general, our analysis shows that naïve estimates that do not account for the non-response bias tend to drive prevalence estimates upward. In contrast to the findings in the literature examining the non-response bias in HIV serosurveys, on average participants who are more likely to have antibodies are more likely to participate in COVID-19 surveys[16,28]. Participants with history of illness in the last 3 months or past history of tests for COVID-19 in the last 3 months were more likely to agree to antibody testing in our study probably seeking external confirmation.

In our sample of participants we did find only a slight age difference in the seropositivity rates, and there was no difference between men and women, which is in line with previous findings[6]. However, we observed several clear differences in seroprevalence estimates in a subgroup analysis. First of all, we detected an elevated seroprevalence in participants who reported history of illness and history of any COVID-19 test in the last 3 months, this association was seen regardless of the modelling approach. Second, seroprevalence was lower in participants who lived alone and reported that they started to wash their hands more often. Third, in the secondary analysis of participants who were tested we observed that seroprevalence was lower in current smokers compared to never smokers, it was also lower in participants who reported past history of allergies.

All associations revealed in our study should not be immediately regarded as causal due to limitations in the study design and analysis. History of testing and illness in the last 3 months can be easily interpreted. Seroprevalence among those reporting a history of COVID-19 testing was relatively low (around 20%), this can be explained by the high scale of testing in Russia since the onset of the epidemic. However, our study is not a direct evidence of the effectiveness of hand hygiene, as self-reported change in habits can reflect other differences between sub-populations. There is limited and conflicting evidence about the smoking rates in COVID-19 patients[29,30]. While our study is the one of the first that compared population-based seroprevalence estimates between smokers and non-smokers there is a need for more studies to confirm this finding[9]. There are many

	CMIA				ELISA			
	Crude PR	95% CI	Adjusted PR	95% CI %	Crude PR	95% CI	Adjusted PR	95% CI %
Age group								
18–34	1.00	Ref	1.00	Ref	1.00	Ref	1.00	Ref
35–49	0.66	(0.41–1.04)	0.59	(0.35–0.98)	0.82	(0.50–1.33)	0.79	(0.46–1.33)
50–64	1.00	(0.62–1.58)	1.00	(0.54–1.78)	1.34	(0.81–2.17)	1.38	(0.74–2.47)
65+	0.24	(0.04–0.77)	0.30	(0.02–1.45)	0.47	(0.11–1.29)	0.84	(0.13–2.89)
Male	1.14	(0.77–1.67)	1.07	(0.66–1.70)	1.04	(0.69–1.56)	0.93	(0.56–1.51)
Higher education	0.85	(0.54–1.41)	0.61	(0.36–1.06)	0.98	(0.60–1.71)	0.70	(0.41–1.29)
Higher income	1.07	(0.73–1.57)	1.05	(0.67–1.65)	1.17	(0.78–1.75)	1.11	(0.70–1.78)
Respondent lives alone	0.60	(0.32–1.02)	0.59	(0.28–1.09)	0.67	(0.36–1.16)	0.63	(0.30–1.19)
Respondent started to wash hands more often	0.63	(0.43–0.93)	0.58	(0.38–0.91)	0.65	(0.44–0.99)	0.64	(0.41–1.02)
Respondent travelled abroad in the last 3 months	1.05	(0.56–1.81)	0.84	(0.41–1.54)	0.98	(0.49–1.75)	0.73	(0.33–1.40)
History of COVID-19 testing	2.68	(1.82–3.92)	2.05	(1.30–3.20)	3.23	(2.16–4.81)	2.41	(1.51–3.81)
Cold symptoms in the last 3 months *	4.32	(2.70–7.19)	3.79	(2.30–6.54)	4.42	(2.71–7.57)	4.13	(2.45–7.34)
Smoking status								
Never smoked	1.00	Ref	1.00	Ref	1.00	Ref	1.00	Ref
Previous smoker	0.87	(0.53–1.37)	0.94	(0.55–1.54)	0.83	(0.50–1.33)	0.94	(0.55–1.57)
Current smoker	0.54	(0.27–0.97)	0.46	(0.22–0.87)	0.42	(0.19–0.81)	0.34	(0.14–0.72)
Alcohol consumption frequency								
Never	1.00	Ref	1.00	Ref	1.00	Ref	1.00	Ref
Monthly	1.21	(0.73–2.10)	1.31	(0.76–2.34)	1.11	(0.66–1.93)	1.19	(0.68–2.14)
Weekly or more often	0.92	(0.52–1.67)	0.96	(0.52–1.80)	0.82	(0.45–1.50)	0.94	(0.50–1.80)
Chronic diseases or medication use	0.86	(0.56–1.30)	0.84	(0.52–1.33)	0.77	(0.49–1.19)	0.69	(0.42–1.12)
Past history of allergies	0.53	(0.30–0.90)	0.54	(0.30–0.92)	0.50	(0.27–0.86)	0.53	(0.28–0.93)

Table 3. Prevalence ratios for self-reported characteristics of tested individuals in phone and paper-based surveys. * – "Cold symptoms in the last 3 months" was used in the paper-based survey instead of "Past history of illness in the last 3 months" in the phone-based interview.

examples when smoking effects were subject to structural epidemiological biases[31]. Even if this association is causal, then behavioural or biological mechanisms should be explored. Smoking is a well-established risk factor for many diseases and it is likely linked to COVID-19 severity regardless of the risk of infection[29].

It is also tempting to immediately search for biological explanation that link allergy status and risk of infection[32]. However, we should be very cautious due to limitations of study design and other possible explanations, e.g. people who self-report being allergic may behave in a way to minimize risk of being infected. The question about allergy was very general in our paper-based survey, that also limits the value of this finding.

Important source of bias in serological studies is the performance and the nature of the serological tests[33]. Possible explanation of the difference in our study includes different classes of Ig analysed—IgG in case of CMIA and IgG+IgM+IgA in case of ELISA. However, given the total seroprevalence of not more than 10% it seems that lack of IgM and IgA in CMIA test can only partially explain the difference. A recent study showed that seroconversion started on day 5 after disease onset and IgG level rose even earlier than IgM[34]. Another possible explanation for different seroprevalence estimates of two tests is the nature of antigen. SARS-CoV-2 antibody responses specific to the Spike (S) and/or the nucleocapsid (N) proteins are equally sensitive in the acute infection phase[35]. However, as compared to anti-S antibody responses, those against the N protein appear to wane in the post-infection[36]. Recent evaluations of CMIA test used in our study reported sensitivity far below 100% reported by manufacturer. This may also explain the difference[37,38]. Independent validation of the serological assays used in our study is required. This validation should take into account that fact the sensitivity may be declining over time. Another source of underestimation is a proportion of infected that do not seroconvert. Straightforward adjustments for this sort of biases are not available without additional laborious testing[39].

Our study has several other important limitations. We are addressing seroprevalence in adults only, while previous studies also included participants younger than 18 years old[5,6]. We are reporting prevalence over the period of more than two months that may not reflect the point prevalence at the end of the study period. Our study had a relatively low participation rate given the existing propensity to answer phone calls in the city. However, the majority of phone numbers generated through random digit dial were not reached, rather than

declined to participate. Among 6,671 excluded 3,048 (45,7%) were actually ineligible. We assumed missingness at random for those who did not complete the interview or did not pick the phone. Comparison with the previous representative city survey showed that our sample was representative (see Supplementary Appendix Table A3). We have also excluded distant city districts from our sampling. Even though we observed statistically significant differences between by-district seroprevalence, the lion's share of city residents (about 4.3 mln of 5.2 mln) live in the surveyed districts. Our randomized incentivisation scheme was not successful because randomly assigned taxi offer was not associated with participation agreement and failed to become a valid exclusion restriction. In our main analysis we did not apply post-stratification methods adopted previously[5]. However, application of raking weights estimated to match targets from a representative survey of adult city population showed little to no changes in weighted seroprevalence estimates. We explained this by little to no association between seroconversion and age or education level. Finally, we report cross-sectional results but longitudinal data are needed to offer additional insights to immunity waning and prolonged defence against re-infection.

Conclusion. COVID-19 pandemic has already affected at least 300 000 residents of St. Petersburg that can be extrapolated to millions in the whole country. However the vast majority of population does not carry antibodies to SARS-CoV-2. This highlights the need for further high-quality population based studies that can provide evidence for measures to diminish the impact of the pandemic.

References

1. Koopmans, M. & Haagmans, B. Assessing the extent of SARS-CoV-2 circulation through serological studies. *Nat. Med.* (2020).
2. Goudsmit, J. The paramount importance of serological surveys of SARS-CoV-2 infection and immunity. *Eur. J. Epidemiol.* **1**, (2020).
3. World Health Organization and others. *Population-based age-stratified seroepidemiological investigation protocol for coronavirus 2019 (COVID-19) infection, 26 May 2020* (World Health Organization, Tech. Rep., 2020).
4. Silveira, M. F. *et al.* Population-based surveys of antibodies against SARS-CoV-2 in southern Brazil. *Nat. Med.* 1–4 (2020).
5. Stringhini, S. *et al.* Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study. *Lancet* (2020).
6. Pollán, M. *et al.* Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *Lancet* (2020).
7. Doi, A. *et al.* Estimation of seroprevalence of novel coronavirus disease (COVID-19) using preserved serum at an outpatient setting in Kobe. A cross-sectional study. *medRxiv* . https://doi.org/10.1101/2020.04.26.20079822 (2020).
8. Bryan, A. *et al.* Performance characteristics of the Abbott Architect SARS-CoV-2 IgG assay and seroprevalence testing in idaho. *J. Clin. Microbiol.* (2020).
9. Ward, H. *et al.* Antibody prevalence for SARS-CoV-2 in England following first peak of the pandemic: REACT2 study in 100,000 adults. *medRxiv* (2020). https://doi.org/10.1101/2020.08.12.20173690.
10. Xu, X. *et al.* Seroprevalence of immunoglobulin M and G antibodies against SARS-CoV-2 in China. *Nat. Med.* 1–3 (2020).
11. Van Loon, A. J. M., Tijhuis, M., Picavet, H. S. J., Surtees, P. G. & Ormel, J. Survey non-response in the Netherlands: effects on prevalence estimates and associations. *Ann. Epidemiol.* **13**, 105–110 (2003).
12. Stang, A. Nonresponse research: an underdeveloped field in epidemiology. *Eur. J. Epidemiol.* 929–931 (2003).
13. Hernán, M. A. Invited commentary: selection bias without colliders. *Am. J. Epidemiol.* **185**, 1048–1050 (2017).
14. Elwood, J. M. Commentary: on representativeness. *Int. J. Epidemiol.* **42**, 1014–1015 (2013).
15. Etter, J.-F. & Perneger, T. V. Analysis of non-response bias in a mailed health survey. *J. Clin. Epidemiol.* **50**, 1123–1128 (1997).
16. Mosha, N., Aluko, O., Todd, J., Machekano, R. & Young, T. Analytical methods used in estimating the prevalence of HIV/AIDS from demographic and cross-sectional surveys with missing data: a systematic review. *BMC Med. Res. Methodol.* **20**, 1–10 (2020).
17. Reiczigel, J., Földi, J. & Ózsvári, L. Exact confidence limits for prevalence of a disease with an imperfect diagnostic test. *Epidemiol. Infect.* **138**, 1674–1678 (2010).
18. Zou, G. A modified poisson regression approach to prospective studies with binary data. *Am. J. Epidemiol.* **159**, 702–706 (2004).
19. Pasek, J. Package 'anesrake'. *The Comprehensive R Archive Network* (2020).
20. SARS-CoV-2-Coronapass (2020). https://corona-pass.ru/elisa-coronapass-total. Accessed 17 Sept 2020.
21. Team, R. C. *et al.* R: A language and environment for statistical computing (2013).
22. Marra, G., Radice, R., Bärnighausen, T., Wood, S. & McGovern, M. A simultaneous equation approach to estimating HIV prevalence with nonignorable missing responses. *J. Am. Stat. Assoc.* **112**, 484–496 (2017).
23. Wickham, H. ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics **3**, 180–185 (2011).
24. Pebesma, E. Simple features for R: standardized support for spatial vector data. *R J.* **10**, 439 (2018).
25. Dunnington, D. ggspatial: Spatial data framework for ggplot2. *R Package* (2018).
26. OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org. https://www.openstreetmap.org (2020).
27. Levesque, J. & Maybury, D. A note on COVID-19 seroprevalence studies: a meta-analysis using hierarchical modelling. *medRxiv* (2020) https://doi.org/10.1101/2020.05.03.20089201.
28. Clark, S. & Houle, B. Validation, replication, and sensitivity testing of Heckman-type selection models to adjust estimates of HIV prevalence. *PLoS ONE* **9**, (2014).
29. Zhao, Q. *et al.* The impact of COPD and smoking history on the severity of COVID-19: a systemic review and meta-analysis. *J. Med. Virol.* (2020).
30. Guan, W.-J. *et al.* Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* **382**, 1708–1720 (2020).
31. Luque-Fernandez, M. A., Zoega, H., Valdimarsdottir, U. & Williams, M. A. Deconstructing the smoking-preeclampsia paradox through a counterfactual framework. *Eur. J. Epidemiol.* **31**, 613–623 (2016).
32. LeMessurier, K. S. *et al.* Allergic inflammation alters the lung microbiome and hinders synergistic co-infection with H1N1 influenza virus and Streptococcus pneumoniae in C57BL/6 mice. *Sci. Rep.* **9**, 1–15 (2019).
33. Bastos, M. L. *et al.* Diagnostic accuracy of serological tests for COVID-19: systematic review and meta-analysis. *BMJ* **370** (2020).
34. Kong, W.-H. *et al.* Serologic response to SARS-CoV-2 in COVID-19 patients with different severity. *Virol. Sin.* 1–6 (2020).
35. Perkmann, T. *et al.* Side by side comparison of three fully automated SARS-CoV-2 antibody assays with a focus on specificity. *medRxiv* (2020). https://doi.org/10.1101/2020.06.04.20117911.

36. Fenwick, C. *et al.* Changes in SARS-CoV-2 antibody responses impact the estimates of infections in population-based seropreva-lence studies. *medRxiv* (2020) https://doi.org/10.1101/2020.07.14.20153536.
37. Chew, K. L. *et al.* Clinical evaluation of serological IgG antibody response on the Abbott Architect for established SARS-CoV-2 infection. *Clin. Microbiol. Infect.* (2020).
38. Duggan, J., Brown, K., Andrews, N., Brooks, T. & Migchelsen, S. Evaluation of the Abbott SARS-CoV-2 IgG for the detection of anti-SARS-CoV-2 antibodies (2020). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/890566/Evaluation_of_Abbott_SARS_CoV_2_IgG_PHE.pdf. Accessed 16 Sept 2020.
39. Braun, J. *et al.* Presence of SARS-CoV-2 reactive T cells in COVID-19 patients and healthy donors. *medRxiv* (2020) https://doi.org/10.1101/2020.04.17.20061440.

Author contributions
AB, DSk, VV, KT, LB, DSh and PT conceived the study. AB, DSk and VV drafted the first version of the manu-script. KT, YR, AN, EP and DSh contributed to drafting sections of the manuscript. DS, AB and DSh did data analyses. SZ and EP did lab analyses. All authors participated in the study design, helped to draft the manuscript, contributed to the interpretation of data and read and approved the final manuscript.

Competing interests
AB reports personal fees from MSD and Biocad outside the submitted work. AI, EP and SZ report a pending patent for the test system (ELISA) for detecting antibodies specific to the SARS-COV-2 in a biological sample. Other authors have no conflict of interest to declare.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-92206-y.

Correspondence and requests for materials should be addressed to A.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.