Observed and estimated prevalence of Covid-19 in Italy: Is it possible to estimate the total cases from medical swabs data?

F. Bassi – Department of Statistical Sciences, University of Padova, Italy,
G. Arbia, Department of Statistical Sciences, Catholic University of the Sacred Hearth, Milano, Italy,
P.D. Falorsi, Italian National Statistical Institute.

Abstract: *During the Covid-19 pandemic in Italy, official data are collected with medical swabs following a pure convenience criterion which, at least in an early phase, has privileged the exam of patients showing evident symptoms. However, there are evidences of a very high proportion of asymptomatic patients (e. g. Aguilar et al., 2020; Chugthai et al, 2020; Li, et al., 2020; Mizumoto et al., 2020a, 2020b and Yelin et al., 2020). In this situation, in order to estimate the real number of infected (and to estimate the lethality rate), it should be necessary to run a properly designed sample survey through which it would be possible to calculate the probability of inclusion and hence draw sound probabilistic inference. Some researchers proposed estimates of the total prevalence based on various approaches, including epidemiologic models, time series and the analysis of data collected in countries that faced the epidemic in earlier time (Brogi et al., 2020). In this paper, we propose to estimate the prevalence of Covid-19 in Italy by reweighting the available official data published by the Istituto Superiore di Sanità so as to obtain a more representative sample of the Italian population. Reweighting is a procedure commonly used to artificially modify the sample composition so as to obtain a distribution which is more similar to the population (Valliant et al., 2018). In this paper, we will use post-stratification of the official data, in order to derive the weights necessary for reweighting them using age and gender as post-stratification variables, thus obtaining more reliable estimation of prevalence and lethality*.

1. Introduction

In the recent period of pandemic emergency, most statistical analyses on Covid-19 focused primarily on finding the best forecasting model to be able to anticipate the number of epidemic cases at a national and local level (see for a review, for example, Ceylan, 2020). Comparatively less attention has been paid in the literature to a detailed descriptive analysis of the publicly available data. In Italy, there are two major sources of the data on the diffusion of Covid-19: the Istituto Superiore di Sanità - Italian National Institute of Health (INIH) and the Dipartimento di Protezione Civile - Civil Protection Department (CPD). Both sources, at the moment, do not disclose individual data on the epidemic. In particular, CPD publishes daily the number of positive patients, of deaths and of individuals tested by swabs at national, regional, and provincial level. In addition, INIH released till the end of May 2020 a biweekly report with infections and deaths disaggregated by gender and age class at a national level. Positive cases were also disaggregated by age (but not by gender) in each of the 20 Italian regions. So both sources do not provide any information about the asymptomatic patients, thus limiting the usefulness of these data in view of calculating interesting epidemic parameters such as the *lethality rate*. However, a careful look at these data, especially if compared with the dimension and the demographic structure of the actual population of our country, might give important information to better understand the effects of the virus. The aim of this paper is to exploit the available information to estimate the prevalence and the lethality of the virus in the total Italian population. In particular, we use the data published by the INIH and the CPD (henceforth the "official data") with reference to the positive patients by Covid-19 and to the number of deaths at May 26 2020 – unfortunately after this data the INIH stopped publishing the regional data of infections by age. The information on the demographic structure of the Italian population, refers instead, to 1 January 2019 and it is taken from the National Statistical Institute website (www.istat.it/en)

Our estimation makes use of all available information on the number of positive patients and on swabs at national and regional level. As it is well known, these data suffer from the severe limitation of being observed without a proper sample design and so they can be seen as a convenience sample that cannot be used to

draw probabilistic inference. To try and reduce such distortion, we propose to post-stratify the convenience sample using gender and age as post-stratification variables in order to obtain a dataset which is closer to a representative sample of the Italian population, under some reasonable assumptions.

On 25 May 2020, the Italian Statistical Institute, in collaboration with the Italian Miinstry for Health, started a sample survey to estimate the seroprevalence of Covid-19 in the country. The sample is composed 150,000 individuals living in 2,015 Italian municipalities and it is representative of the Italian population by gender, age and economic activity sector. Selected people were contacted via telephone by the Italian Red Cross's regional centers to arrange an appointment for a blood sample to be taken at one of the authorized laboratories and to answer some questions on health conditions, eventual previous infection by the virus, contacts with Coronavirus positive patients. The regional government should inform each participant residing in their territory of the result of the test. If the test was positive, the person in question should be put temporarily into isolation at home and contacted by the regional health service or the local ASL health authority to do a nasopharyngeal swab to verify contagiousness.

Unlikely in other countries, where similar surveys were successful (see, as an example, the Spanish experience as in Pollàn, 2020), in Italy some difficulties were encountered especially in obtaining respondents cooperation. The period of information collection was extended till 15 July and the present moment results are not yet available. Some press sources indicate a collaboration rate of around 50% of the sample. This result suggests that, at least for Italy, some other statistical methods might be useful in order to estimate the diffusion of the Covid-19 infection at country level. This paper proposes a strategy based on data collected on patients tested by nasopharyngeal swabs and post-stratification sampling techniques.

The rest of the paper is organized as follows. Section 2 is devoted describe the prevalence and the lethality as they emerge from the officially released data. Section 3 presents an estimation of the prevalence of Covid-19 in Italy based on the post-stratification. Section 4 concludes.

2. A descriptive analysis of the prevalence and deaths in Italy

Table 1 reports the absolute frequencies and percentages of infected people and of deaths for Covid-19 distinguishing by gender and age (10-year classes). It also contains the implied lethality rate measured as the ratio between the number of deaths and the number of infected people. This information, published twice a week by the INIH, shows the way in which positive cases are distributed in the various demographic groups. Table 2 reports the observed prevalence and the mortality rates, i.e., respectively, the proportion of infected people and of deaths referring to the consistency of each class in the population measured as the ratio between the number of infected (and deaths) and the total susceptible population.

Table 1. Positive patients, deaths and lethality rate by Covid-19 by gender and age, 26 May 2020, official data. Source: Official data by INIH.

Age	Positive patients				Deaths				Lethality rate		
	Male	Female	Total	Proportion of males	Male	Female	Total	Proportion of males	Male	Female	Total
0-9	1,015	903	1,918	52.92%	1	2	3	33.33%	0.10%	0.22%	0.16%
10-19	1,724	1,718	3,442	50.09%	0	0	0	0.00%	0.00%	0.00%	0.00%
20-29	5,685	7,240	12,925	43.98%	6	3	9	66.67%	0.11%	0.04%	0.07%
30-39	8,026	9,902	17,928	44.77%	35	19	54	64.81%	0.44%	0.19%	0.30%
40-49	12,513	17,427	29,940	41.79%	184	62	246	74.80%	1.47%	0.36%	0.82%
50-59	19,210	22,221	41,431	46.37%	778	215	993	78.35%	4.05%	0.97%	2.40%
60-69	18,494	12,385	30,879	59.89%	2,299	676	2,975	77.28%	12.43%	5.46%	9.63%
70-79	19,033	14,107	33,140	57.43%	5,566	2,283	7,849	70.91%	29.24%	16.18%	23.68%
80-89	16,497	24,027	40,524	40.71%	6,593	4,801	11,394	57.86%	39.96%	19.98%	28.12%
90+	3,818	14,783	18,601	20.53%	1,556	2,873	4,429	35.13%	40.75%	19.43%	23.81%
Total	106,015	124,713	230,728	45.95%	17,018	10,934	27,952	60.88%	16.05%	8.77%	12.11%

Table 2. Observed prevalence and mortality rate in the Italian population by gender and age, 27 May 2020: Source: Official data by INIH.

Age	Positive patients			Deaths		
	Male	Female	Total	Male	Female	Total
0-9	0.04%	0.04%	0.04%	0.00%	0.00%	0.00%
10-19	0.06%	0.06%	0.06%	0.00%	0.00%	0.00%
20-29	0.18%	0.24%	0.21%	0.00%	0.00%	0.00%
30-39	0.23%	0.28%	0.25%	0.00%	0.00%	0.00%
40-49	0.27%	0.37%	0.32%	0.00%	0.00%	0.00%
50-59	0.42%	0.47%	0.44%	0.02%	0.00%	0.01%
60-69	0.53%	0.32%	0.42%	0.07%	0.02%	0.04%
70-79	0.70%	0.44%	0.56%	0.20%	0.07%	0.13%
80-89	1.18%	1.11%	1.14%	0.47%	0.22%	0.32%
90+	1.82%	2.62%	2.40%	0.74%	0.51%	0.57%
Total	0.36%	0.40%	0.38%	0.06%	0.04%	0.05%

Percentages in the last row of Table 2 indicate that this measure of prevalence is higher for females than for males. This is a rather new evidence. On April 9 2020, for example, the prevalence was higher for males (0,24% vs. 0,21%) while on April 16, it was equal for the two genders (0,29%). An explanation of this result could be that in the last days observed in our reference period, even patients with light symptoms were tested and resulted positive to the infection and women showing less severe symptoms were examined in a larger percentage. In the absence of conclusive results, it is worthwhile to monitor accurately this phenomenon over time to verify if this tendency is confirmed.

Figure 1. Prevalence in the Italian population by gender and age, 26 May 2020. Source: Official data by INIH

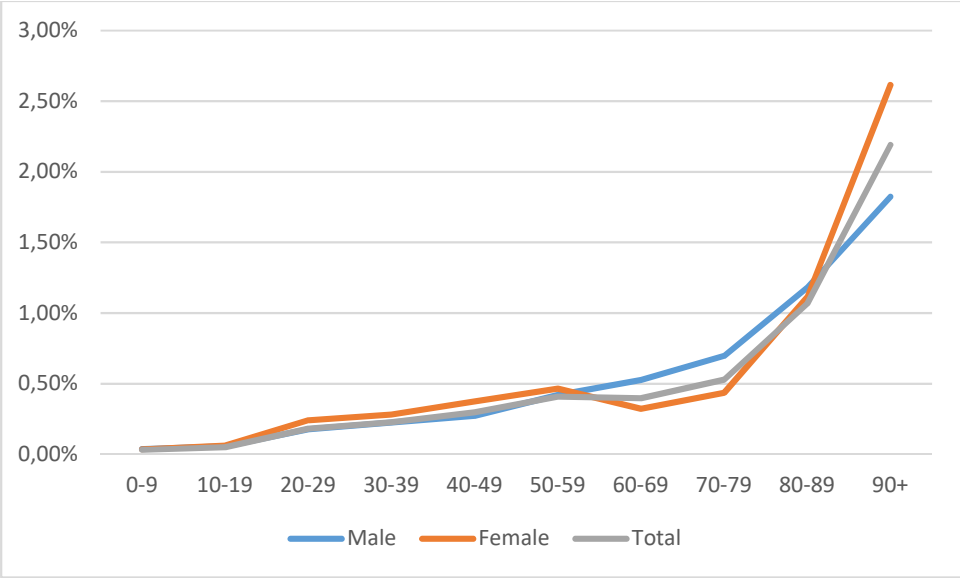
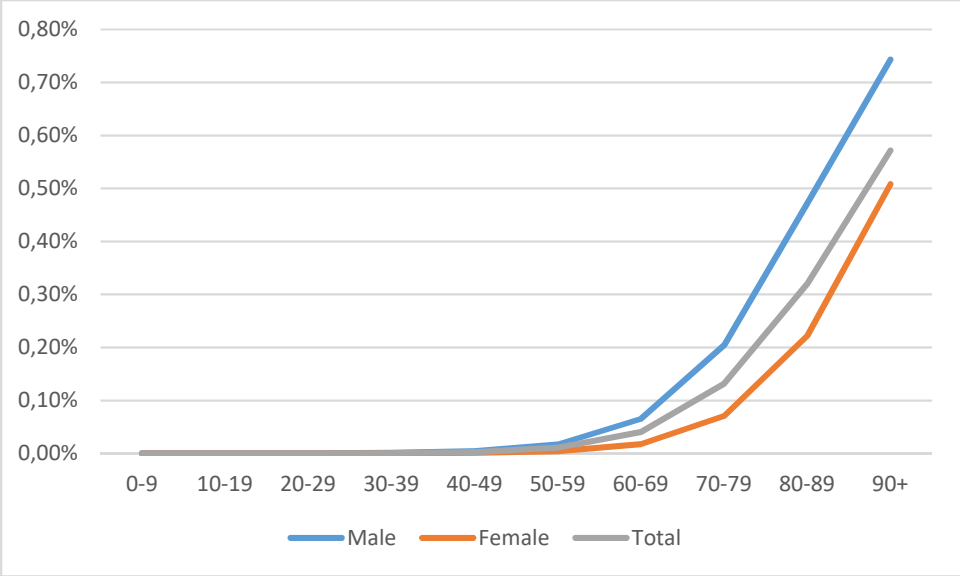


Figure 1 clearly shows that prevalence in the sample increases with age for both genders. Overall, the prevalence is higher for women than for men. More specifically, this is true for people younger than 59 years and for those older than 90, while for the other age classes we observe an opposite trend. Another interesting result emerging from the official data, is that women with an age between 60 and 80 show a lower risk of infection than women at least 10 years younger.

Figure 2. Proportion of deaths in the Italian population by gender and age, 7 May 2020. Source: Official data by INIH.



The composition indices reported in Table 3 compare the distribution of patients by gender and age with the distribution of the same variables in the Italian population. A composition lower than 1 indicates that individuals in that class are present among positive patients in a lower proportion with respect to the whole population. In contrast, a value greater than 1 provides an opposite indication. If we compare the composition indexes in the same age group between the two genders we find a confirmation of the evidences shown in Figure 1.
Finally, Figure 2 clearly shows that the proportion of deaths increases with age. In this case, the situation is much worse for males than for females in all age classes (see also composition indices in Table 3).

Table 3. Composition indices for patients and deaths with reference to the Italian population, 7 May 2020. Source: Official data by INIH.

	Patients		Deaths	
Age	Male	Female	Male	Female
0-9	0.1075	0.1012	0.1044	0.0007
10-19	0.1603	0.1708	0.1654	0.0000
20-29	0.4905	0.6714	0.5777	0.0032
30-39	0.6250	0.7808	0.7024	0.0170
40-49	0.7550	1.0390	0.8979	0.0692
50-59	1.1629	1.2902	1.2279	0.2934
60-69	1.4600	0.8972	1.1665	1.1306
70-79	1.9345	1.2085	1.5406	3.5243
80-89	3.2756	3.0838	3.1591	8.1550
90+	5.0555	7.2496	6.6566	12.8350

3. Post-stratified estimation of prevalence lethality of Covid-19 in Italy

In the reference period of our analyses, various guidelines have been followed to test individuals with swab to ascertain the presence of Covid-19 infection. In the early phase of the epidemic, the World Health Organization (WHO) recommended to test only patients with at least three specific symptoms and reporting contact with the infection. In a second moment of time, since a large proportion of positive patients did not report any symptoms, these guidelines were relaxed and even people with light symptoms (or only with contact with other infected patients), were examined. Moreover, the procedures to access to swab can be different in the 20 Italian regions. These considerations, together with the evidences found in the literature

of a high proportion of asymptomatic patients (e. g. Aguilar et al., 2020; Chugthai et al, 2020; Li, et al. , 2020; Mizumoto et al., 2020a, 2020b and Yelin et al., 2020), suggest that prevalence in the Italian population could be much higher than what appears from the official data.

In order to estimate the total number of infected by Covid-19, it should be necessary to run a properly designed sample survey through which it is possible to calculate the probability of inclusion and hence draw sound probabilistic inference. Such a survey has a high costs and it needs time to be realized. While waiting for the results of such a desirable survey, some researchers proposed various approaches to estimate the prevalence. They are based on epidemiologic models, time series models and the analysis of data collected in countries that faced the epidemic before like for Italy and Europe countries, China and Korea (see, e. g., Brogi et al., 2020).

In this paper, we propose to try and approximate the true prevalence of Covid-19 in Italy by exploiting the official data published by the INIH (Istituto Superiore di Sanità, 2020a and 2020b) and the CPD, but reweighted them so as to obtain something closer to a representative sample of the Italian population. The procedure of reweighting (Valliant et al., 2018) is used when, in most cases due to measurement error, a sample is not representative of the reference population. More specifically, reweighting is a procedure to artificially modify the sample composition, in the phase of data analysis, so as to obtain a distribution which is closer to the population. In its simplest form, reweighting assigns appropriate weights to each sample unit where weights can be defined on the basis of the inclusion probability (if known) or on the basis of available information on the population. In this case, we call it post-stratification (Holt and Smith, 1979; Little, 1993). In this last case, after choosing one or more stratification variables, whose distribution is known in the population, the sample units are weighted with the ratio between the theoretical proportion in the population and the observed proportion in the sample. In this paper, we will use post-stratification to analyse the official data sample, using age and gender proportion in the Italian population as post-stratification variables.

In our analysis, in order to correct the observed sample, we need to introduce some working assumptions which are required because all disaggregated information is not available. Indeed, as already said, data by gender and age are available only for positive patients, and not for those that are negative at the swab. As a consequence, in order to simulate the characteristics of all patients tested by oropharyngeal swab till 26 May 2020, we proceeded as follows.

First of all, we considered the positive patients disaggregated into age classes in each of the 20 Italian regions, an information provided twice a week by the INIH (Istituto Superiore di Sanità 2020b). Furthermore, the total number of people subjected to pharyngeal swab was derived using the information provided by the percentage of positive tests in each region supplied by the CPD. Lacking the appropriate age disaggregation we assumed that this percentage is constant in all age classes. The simulated sample of patients subjected to swabs is then distributed into the two genders by using the sex ratios observed among positive patients (see Table 1, column 5).

Secondly, we calculated the post-stratification weights referring to the distribution by age and gender of the Italian population at the last available date which was January 1st, 2019.

Finally, we re-estimate the prevalence after post-stratification.

The aggregate out-coming value from this operation is prevalence estimated equal to 11.69%, a number that, reported to the total population of Italy, reveals that 7.054.118 people could have been affected by Covid-19 in the country as of 26 May 2020. Furthermore, lethality rate is also re-estimated on the post-stratified sample leading to a rate of 1,56% and the median age of positive patients of 52 years.

Table 4 reports the estimated prevalence by age in the Italian population and the consequent estimated lethality rate. Even if it is known that the vast majority of swabs were obtained from symptomatic, this estimate of prevalence partly corrects for all those patients who have not been subjected to pharyngeal

swab for various reasons, mainly for presenting light symptoms (pauci-symptomatic) or for being asymptomatic.

Table 4. Estimated prevalence and lethality in the Italian population by age after post-stratification.

Age	Prevalence	Lethality
0-9	6.61%	0.02%
10-19	7.11%	0.00%
20-29	9.43%	0.01%
30-39	11.57%	0.04%
40-49	15.75%	0.10%
50-59	16.20%	0.30%
60-69	13.08%	1.22%
70-79	11.67%	3.28%
80-89	7.43%	4.06%
90+	1.85%	3.47%
Total	11.69%	1.56%

By comparing our estimation with the current estimates based on unweighted (official) data, we observe that prevalence in age classes is much higher than that calculated with official data, with the only remarkable exception of the class of people with 90 years and over (see Table 2). This result shows that there might be a much higher proportion of people infected than those measured with test, possibly due to the fact that many positive patients do not even know of their condition because they have very light symptoms or no symptoms at all (Lavezzo et al., 2002). Our estimates of prevalence have the obvious effect to produce a much lower lethality rate in all age classes (except 90 and over) with respect that calculated with the uncorrected official data. For what concerns the lethality rate, we calculated it reporting the number of death in ach age class to the number of estimated infections. We assume that deaths caused by Coronavirus are all correctly reported.

4. Summary of results and concluding remarks

In this report, we analyze data on Covid-19 infections in Italy with reference to the consistency and demographic structure of the Italian population. Looking at the official data, published by the INIH and by the CPD, in general for male patients the risk of infection increases with age. In contrast, the dynamics of infection is peculiar for female patients: it increases until the age of 50, then decreases until 80 and increases again for older people. Furthermore, the observed prevalence is higher for females than for males until the age of 59. After that age women show a lower prevalence and the distance from males increases with age. Finally, until the beginning of April 2020, men represented the majority among positive patients while afterwards the proportion of women was continuously increasing. The dynamic of the infection by Covid-19 in female patients deserves further study, both in terms of observation over time and with epidemiologic and statistical analyses.

Secondly, starting from official data on Covid-19, we estimate the prevalence of the infection in the Italian population assuming that only a small proportion of patients could access to pharyngeal swab test. Lacking adequate data disaggregation (for example the number of patients subjected to swab by gender and age) we had to make some working assumptions to which the results obtained are strongly depending. However, we believe that our methodology represents a reasonable approximation while waiting for more reliable data obtained with a properly designed national sample survey and that it could be further improved if more data were made available. In particular, it would be important to have the availability of disaggregated data, at least for gender and age classes, so as to avoid to impose our restrictive hypotheses of uniformity and to

obtain better estimates of prevalence and lethality in each specific group of the population. This would enable us to identify the categories that are more exposed to the risk of infection and to support a system of active surveillance also in the period of recession of the pandemic.

References

Aguilar, J. B., Faust, J. S. Westafer, L. M. and Gutierrez, J. B. (2020) Investigating the Impact of Asymptomatic Carriers on COVID-19, medXiv, doi: https://doi.org/10.1101/2020.03.18.20037994 .

Brogi, F., Guardabascio, B., Barcaroli, G. (2020=) Covid-19 in Italy: actual infected population, testing strategy and imperfect compliance, DOI: 10.13140/RG.2.2.18275.50729.

Ceylan, G. (2020) Estimation of COVID-19 prevalence in Italy, Spain, and France, *Science of the The Total Environment,* doi.org/10.1016/j.scitotenv.2020.138817.

Chughtai, A.A. and Malik, A.A., (2020). Is Coronavirus disease (COVID-19) case fatality ratio underestimated? *Global Biosecurity*, 1(3).

Cochran, W.G. (1977) *Sampling Techniques*. Wiley. New York.

Holt, D. and Smith, T. M. F. (1979) Post stratification, Series A, 142, 1, 33-46.

Istituto Superiore di Sanità (2020a) COVID-19 epidemic. 26 may 2020, national update.

Istituto Superiore di Sanità (2020b) COVID-19 epidemic. 26 may 2020, national update (Appendix).

Little, R. J. A. (1993) Post-Stratification: A Modeler's Perspective, *Journal of the American Statistical Association*, 88:423, 1001-1012, DOI: 10.1080/01621459.1993.10476368

Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G., Barzon, L., Del Vecchio, C., Rossi, L., Manganelli, R., Loregian, A., Navarin, N., Abate, D., Sciro, M., Merigliano, S., Decanale, E., Vanuzzo, M.C., Saluzzo, F., Onelia, F., Pacenti, M., Parisi, S., Carretta, G., Donato, D., Flor, L., Cocchio, S., Masi, G. Sperduti, A., Cattarino, L., Salvador, R. Gaythorpe, K.a.M., Brazzale, A.R., Toppo, S., Trevisan, M., Baldo, V., Donnelly, C.A., Ferguson, N.M., Dorigatti, I., Crisanti A**.** (2020) Suppression of COVID-19 outbreak in the municipality of Vo, Italy, *Nature*, https://doi.org/10.1038/s41586-020-2488-1.

Li, R., Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, Jeffrey Shaman (2020) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2), *Science* 16 Mar 2020, eabb3221, DOI: 10.1126/science.abb3221 .

Mizumoto, K., Kagaya, K., Zarebski, A., & Chowell, G. (2020a). Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, *25*(10), 2000180. https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180 .

Mizumoto, K., Katsushi, K. Zarebski, A. and Gerardo (2020b) Estimating the Asymptomatic Proportion of 2019 Novel Coronavirus onboard the Princess Cruises Ship, 2020, medRxiv, .https://doi.org/10.1101/2020.02.20.20025866doi.

Pollán, M., Pérez-Gómez, B., Pastor-Barriuso, R., Oteo, J., Hernán, M.A., Pérez-Olmeda, M., Sanmartín, J.L., Fernández-García, A., Cruz, I., Fernández de Larrea, N., Molina, M., Rodríguez-Cabrera, F., Martín, M., Merino-Amador, P., Paniagua, J.L., Muñoz-Montalvo, J.F., Blanco, F., & Yotti, R. (2020) Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study, *The Lancet*, doi.org/10.1016/ S0140-6736(20)31483-5.

Valliant, R., Dever, J.A., Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*, New York, Springer.

Yelin, I. , Noga Aharony, Einat Shaer-Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Areen Kuzli, Nagam Gandali, Tamar Hashimshony, Yael Mandel-Gutfreund, Michael Halberthal, Yuval

Geffen, Moran Szwarcwort-Cohen, Roy Kishony (2020) Evaluation of COVID-19 RT-qPCR test in multi-sample pools, medRxiv, 27 march, 2020.doi: https://doi.org/10.1101/2020.03.26.20039438.