Estimates of COVID-19 Cases Across Canadian Provinces

David Benatia, CREST – ENSAE
Raphael Godefroy, Université de Montréal
Joshua Lewis, Université de Montréal*

May 2020

Abstract

This paper estimates population infection rates from coronavirus disease 2019 (COVID-19) across four Canadian provinces from late March to early May 2020. The analysis combines daily data on the numbers of conducted tests and diagnosed cases with a sample selection model that corrects for non-random testing. The estimated population infection rates were 1.7 - 2.6 percent in Quebec, 0.7 - 1.4 in Ontario, 0.5 - 1.2 percent in Alberta, and 0.2 - 0.4 percent in B.C over the sample period. The results suggest widespread undiagnosed COVID-19 infection. For each identified case by mid-April, we estimate that there were roughly 12 population infections.

*Benatia: CREST (UMR 9194), ENSAE, Institut Polytechnique de Paris, 5 Avenue Henry Le Chatelier, 91120 Palaiseau, France (e-mail: david.benatia@ensae.fr), Godefroy: Department of Economics, Université de Montréal, 3150, rue Jean-Brillant, Montreal, QC, H3T1N8 (email: raphael.godefroy@umontreal.ca), Lewis: Department of Economics, Université de Montréal, 3150, rue Jean-Brillant, Montreal, QC, H3T1N8 (email: joshua.lewis@umontreal.ca).

1 Introduction

The first cases of coronavirus disease 2019 (COVID-19) in Canada were documented in late January, 2020, and by May 5 more than 63,000 cases had been reported (CSSE, 2020). Because testing has been limited to a small fraction of the population and infected individuals with mild or no symptoms may not seek testing, however, there is potential for widespread undocumented infections.[1]

This paper estimates population infection rates for COVID-19 across four Canadian provinces – Quebec, Ontario, Alberta, and B.C. – from late March to early May. The analysis is based on the methodology developed in Benatia, Godefroy and Lewis (2020) that corrects observed infection rates among tested individuals for non-random sampling to calculate infection rates in the overall population.[2] To implement the procedure, we estimate the relationship between the number of tests and the share of positive tests. This gradient is informative for the severity of selection bias. For example, a *negative* slope indicates *positive* selection bias, since individuals who are most frequently tested have the highest probability of infection. Once this gradient has been estimated, we can recover unbiased estimates of the population infection rate.

A key challenge for the estimation procedure is endogeneity in the supply of testing. For example, if policymakers expand testing in response to increases in underlying disease prevalence, our estimation strategy would underestimate the selection bias gradient and thus overestimate total population infections. To address this concern, we focus on high frequency day-to-day changes in the number of completed tests across U.S. states and Canadian provinces.[3] Because there is little scope for evolution in

[1] See Dong et al. (2020); Lu et al. (2020); Hoehl et al. (2020); Pan et al. (2020); Bai et al. (2020).

[2] Our methodology builds on insights from prior empirical work in economics on the problem of sample selection bias. See Heckman (1976); Heckman (1979); Heckman, Lalonde and Smith (1999); Blundell and Costa Dias (2002); Das, Newey and Vella (2003); Newey (2009).

[3] To improve the precision of the estimates, we include both Canadian provinces with U.S. states to estimate the selection bias gradient. Once this gradient has been estimated, however, our calculations for underlying disease prevalence rely solely on the shares of positive cases across Canadian provinces.

disease prevalence from one day to the next, daily changes in testing should be orthogonal to changes in population infection rates. To further validate this assumption, we estimate models that control for province and state fixed effects, thereby allowing for daily exponential growth in disease prevalence that is specific to each jurisdiction.

We find wide cross-province differences in both the levels of population infection rates and their trends. Average population infection rates that ranged from 0.3 percent in B.C. to 3 percent in Quebec. Infection rates in B.C. declined modestly over the sample period. In Ontario, infection rates rose from early to mid-April and subsequently declined. Meanwhile, Quebec and Alberta experienced increases in population infection rates over the month of April. These trends need not reflect increases in the number of newly infected individuals, since our population infection rates capture *both* newly infected individuals and those with continued detectable viral load over the sample period.[4]

Our results also suggest widespread undetected COVID-19 infection across Canadian provinces. We calculate that for every diagnosed case there were 12 population infections in mid-April. These ratios range from 8.6 in B.C. to 14.8 in Ontario. These estimates are comparable to recent evidence on the rates of undetected infection in the United States and internationally (Perkins et al., 2020; Johndrow, Lum and Ball, 2020; Verity et al., 2020; Ferguson et al., 2020).

This paper provides new evidence on overall population infection rates for COVID-19 in Canada. Our findings complement evidence for COVID-19 prevalence, nationwide. Using survey results for COVID-19 symptoms, Reid (2020) finds that more than 100,000 households reported COVID-like symptoms after adjusting for seasonal

[4]There is an extended period over which individuals may test positive for COVID-19. PCR testing has identified cases days before symptom onset and detected continued viral RNA presence more than three weeks after symptom onset (Huang et al., 2020; Cai et al., 2020; Zhou et al., 2020). Often times these positive cases occur among individuals who are no longer symptomatic, and it is believed that they reflect lingering viral material that no longer poses a risk of transmission.

influenza rates. The results do not account for potentially large numbers of asymptomatic infections. Meanwhile, Verity et al. (2020) combines assumptions regarding the age-adjusted case fatality rate with COVID-related deaths to estimate total population infections in Canada on March 31. These estimates indicate that case detection rate was just 5 percent through March. Our analysis provides the first provincial-level estimates. Given wide cross-provinces differences in per capita testing, official case counts may mask important geographic differences in the severity of the outbreak. Indeed, whereas the official case count in Quebec was 55 percent higher than Ontario, our results show that gap in total cases was less than 20 percent.

Our empirical framework complements existing methods used to estimate population infection rates in the United States and internationally (Ferguson et al., 2020; Perkins et al., 2020; Li et al., 2020*b*; Riou et al., 2020*a*; Johndrow, Lum and Ball, 2020; Javan, Fox and Meyers, 2020; Verity et al., 2020). One approach has been based on the Susceptible Infectious Removed (SIR) epidemiological model, which calibrates parameters to the specific characteristics of the SARS-CoV-2 pandemic to estimate current and future infections. A challenge for this approach is the large uncertainty regarding the relevant parameter values for the virus, and the fact that the parameter values will evolve as societies take different measures to reduce transmission. Other research has relied on Bayesian modelling to infer past disease prevalence from observed COVID-19 deaths. While, these models require fewer assumptions regarding the underlying parameter values, because they 'scale up' observed deaths to estimate population infections, small differences in the assumed case fatality will have substantial effects on the estimates. Given considerable uncertainty regarding the true case fatality, and the fact that COVID-19 related deaths may be undercounted during the course of the pandemic, these estimates may fail to capture the overall extent of population infection.[5]

[5]See Riou et al. (2020*b*); Han et al. (2020); Wu et al. (2020); Clay, Lewis and Severnini (2018, 2019); Katz and Sanger-Katz (2020); Prakash and Hall (2020).

2 Data

Our analysis draws on daily data on total test results (positive plus negative) and positive tests across Canadian provinces and U.S. states for the period March 31 to May 5. Provincial data were obtained from the Epidemiological Data from the COVID-19 Outbreak in Canada project (Berry et al., 2020). This project is conducted by a team of researchers from the University of Toronto and the University of Guelph and provides information on cases and testing across provinces based on publicly available information from government reports and news media. We exclude days in which there were identified changes in provincial reporting standards and days in provincial health authorities did not release information on completed tests.[6] In addition, we use information on the number of positive tests by age group, which is available on provincial health departments, and provincial population estimates from Statistics Canada (2020). We supplement these data with information on the total tests results and positive cases across U.S. states for the same time period from the COVID Tracking Project, a site launched by journalists from The Atlantic that publishes high-quality data on the outbreak across U.S. states (Meyer, Kissane and Madrigal, 2020).

Figure 1 reports the daily tests and positive cases across the four provinces. Daily testing was fairly stable in Quebec throughout April. In contrast, there were substantial increases in daily testing in both Ontario and Alberta, and to a lesser extent in B.C.

3 Methodology

We estimate the following regression model across provinces and states i, on day, t:

[6]There are a large number of missing observations from B.C. due to the periodic release of testing information by the provincial health authorities.

$$\log \frac{s_{i,t}}{n_{i,t}} - \log \frac{s_{i,t-1}}{n_{i,t-1}} = \alpha_1 \left[e^{\beta \frac{n_{i,t}}{pop_i}} - e^{\beta \frac{n_{i,t-1}}{pop_i}} \right] + \alpha_2 \left[e^{2\beta \frac{n_{i,t}}{pop_i}} - e^{2\beta \frac{n_{i,t-1}}{pop_i}} \right] + \alpha_3 \left[e^{3\beta \frac{n_{i,t}}{pop_i}} - e^{3\beta \frac{n_{i,t-1}}{pop_i}} \right] + u_{i,t} \quad (1)$$

where $n_{i,t}$ is the number of tests, $s_{i,t}$ is the number of positive tests, and pop_i is the province (state) population. The term $u_{i,t}$ is an error with mean zero and unknown variance.

Equation (1) captures the relationship between daily changes in the number of conducted tests and daily changes in the fraction of positive cases, derived from the theoretical framework described in Benatia, Godefroy, and Lewis (2020) (see Appendix A). We estimate equation (1) by non-linear least squares, allowing for heteroskedastic errors.

Our identifying assumption requires strict exogeneity of the error term with respect to changes in the number of daily tests: $E(u_{i,t}|\Delta n_{i,t}) = 0$. This assumption ensures that that the errors are uncorrelated with any function of $\Delta n_{i,t}$. In practice, this assumption requires that daily changes in underlying population disease prevalence must be unrelated to daily changes in testing. This assumption is supported by at least two pieces of evidence. First, given the short time interval, there is limited scope for disease evolution, which occurs more gradually depending on the characteristics of the virus and population behaviour. Second, in robustness tests, we controls for province (state) fixed effects, which allow for jurisdiction-specific exponential growth in underlying disease prevalence from one day to the next. These controls do not affect the main coefficient estimates.

We also require that the population coefficient estimates (α_i, β) be similar across jurisdictions. This assumption requires that decisions regarding how to prioritize tests

were made similarly across provinces and U.S. states. Although states had latitude to implement their own diagnostic testing procedures, the guidance laid out for testing prioritization by the CDC was broadly similar to the policies implemented across Canadian provincial health departments (CDC, 2020). Because policy decisions regarding testing of elderly populations may have differed across jurisdictions, we also report estimates based solely on cases among individuals under age 70.

Finally, our analysis depends on the quality of diagnostic testing, and systematic false negative test results may affect the population disease prevalence estimates (Liu et al., 2020; Ai et al., 2020; Yang et al., 2020). Because our analysis focuses on day-to-day variation, however, changes in the rates of misdiagnosis should not be systematically related to changes in the number of implemented tests. As a result, these errors should not bias the coefficient estimates, but may reduce precision through classical measurement error (Wooldridge, 2002).

After estimation equation (1), we can recover estimates of the total population infection rate, $\hat{P}_{i,t}$, in province i at date t, using the predicted values from the regression and setting $n = pop_i$ according to the following equation:

$$\hat{P}_{i,t} = \exp\left\{ \log(s_{i,t}) + \sum_{k=1}^{3} \hat{\alpha}_k \left(e^{k\hat{\beta}} - e^{k\hat{\beta}\frac{n_{i,t}}{pop_i}} \right) \right\} \tag{2}$$

We then used the Delta-method to estimate the confidence interval for $\hat{P}_{i,t}$.

4 Results

Table 1, Panel A reports the estimates for equation (1) across three time periods: March 31 - April 7, April 14 - 21, and April 28 - May 5. We estimate the model separately for all ages (cols. 1, 3, 5) and excluding cases among individuals over age 70 (cols. 2, 4, 6). Consistent with the theoretical framework, we find large estimates

of $\hat{\beta}$ ranging from -1,092 to -1,391, which implies that the sample selection in testing approaches zero as the number of tests approaches the total population size. We also find alternating signs on the coefficient $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$, consistent with the power series approximation developed in Benatia, Godefroy and Lewis (2020).

Figure 1 presents scatterplots of the relationship between daily changes in per capita testing the share of positive tests across states and provinces for the three time periods. The downward sloping relationships imply that larger day-to-day increases in the number of conducted tests are associated with decreases in the share of positive tests. A symptom of selection bias is that variables that have no structural relationship with the dependent variable may appear to be significant (Heckman, 1979). So, these patterns strongly suggest non-random testing, since daily changes in testing should be unrelated to population disease prevalence except through a selection channel.

Table 2 reports the results for Quebec, Ontario, Alberta, and B.C. that adjust observed COVID-19 case rates for non-random testing based on the procedure described in Section 3. Column 2 report the estimates for all age population infection rates on April 4, April 18, and May 2, along with heteroskedasticity robust 95 percent confidence intervals. Column 4 reports the average estimates for the three time periods March 31 - April 4, April 14 - 18, and April 28 - May 2. These latter averages mitigate sampling error in the daily prevalence estimates, which depend on the observed share of positive tests on any particular day.

The results reveal widespread disparities in COVID-19 prevalence across provinces. Population infection rates range from more than 2 percent in Quebec to less than 0.4 percent in B.C. Trends in infection rates differed significantly across provinces. Infection rates in B.C. declined modestly over the sample period. In Ontario, infection rates rose from early to mid-April and subsequently declined. Meanwhile, Quebec and Alberta experienced steady increases in population infection rates over the sample

period.

Columns 3 and 5 report the estimated population infection rates among individuals below age 70. These estimates will not be influenced by specific policies regarding the testing of elderly population and residents of senior residential facilities that may have differed across provinces.[7] For Alberta and B.C., the results are similar to overall population prevalence. Meanwhile, the estimates are systematically lower in Ontario and Quebec, particularly in the latter periods. These results are consistent with the timing of the shift in testing of elderly facilities in these two provinces (CBC, April 8, 2020; Jones, April 22, 2020).

Table 3 explores the robustness of the main estimates to controls for province and state fixed effects. These controls allow for growth rates in underlying disease prevalence that are specific to each locality, to account for the fact that the true infection rate may evolve even with a 24-hour period. Because the intercepts are allowed to differ across each jurisdiction, they also account for variation in the daily evolution of the disease across states and provinces due to differing enforcement of social distancing or other location-based determinants of disease spread. The results (reported in cols. 2, 4, 6, and 8) are virtually identical to the baseline estimates. Moreover, the augmented model tends to produce more precise confidence intervals.

To interpret the findings, we calculate the total population infections in mid-April and compare them to the number of diagnosed cases in each province. We multiply the average estimated prevalence from April 14-18 (Table 2, Panel B, col. 4) by the total province population and compare them to the cumulative diagnosed cases by April 23. The gap in the two periods captures the typical five-day incubation period, to account for the fact that individuals may not seek testing until symptom onset (Li et al.,

[7]To derive these estimates, we subtract the number of cases among the elderly from the total daily cases and total daily tests across provinces. Because we lack data on *total* tests by age group, negative tests among the elderly are included in the denominator, so these estimates should be interpreted as a lower bound estimate for disease prevalence.

2020*a*; Lauer et al., 2020). In principal, these numbers capture two distinct measures of COVID-19 spread: current infections versus cumulative infections. Nevertheless, given limited COVID-19 infection prior to mid-March and the fact that viral presence is detectable by PCR testing three weeks after initial symptom onset (Cai et al., 2020; Zhou et al., 2020), population infection rates in mid-April are likely to be similar to cumulative infections since onset.

Table 4 presents the results. We find widespread undetected population infection. By April 23, 41,371 cases had been identified across the four provinces, however our estimates suggest that there were more than a half million infected individuals. In Quebec and Alberta, there were 11 to 12 population infections for each diagnosed case. In Ontario, we find that there were 15 population infections per diagnosed case. These gaps to align with differences in testing across provinces – Alberta (27 per 1,000) and Quebec (22 per 1,000) versus Ontario (13 per 1,000). Meanwhile, B.C. had the smallest fraction of undetected cases, despite conducting just 14 tests per 1,000 population. This discrepancy can likely be attributed to the fact that the scope of the outbreak was substantially more limited in B.C., allowing officials to better identify clusters of cases.

5 Discussion

This paper provides new evidence on the population prevalence of COVID-19 in Quebec, Ontario, Alberta, and B.C. from late March to early May. Our analysis adapts a sample selection model approach developed in Benatia, Godefroy and Lewis (2020). We find widespread population infection that exceed official reported cases by factors of 9 to 15 across provinces.

Our findings are comparable to recent prevalence estimates from the U.S. and coun-

tries in Western Europe. The estimated infection rates in Quebec are similar to those from the United Kingdom (2.7%), and several U.S. states (Pennsylvania – 2.4%, Rhode Island – 2.4%, and Massachusetts – 3.4%). Meanwhile, the rates in Ontario are similar to Austria (1.1%), Denmark (1.1%), Vermont (1.4%), Virginia (1.4%), and Idaho (1.5%) in early April.[8] Our results are also consistent with recent evidence from serological testing across several U.S. jurisdictions that show widespread undetected infection by mid-April (Bhattacharya et al., 2020; Goodman and Rothfeld, April 23, 2020; Conarck and Chang, April 24, 2020).

Our analysis provides a complement to existing methods used to estimate population infection rates. These approaches either require strong assumptions on unknown disease parameters, or accurate measurement of COVID-related deaths, which may be undercounted over the course of the pandemic. Our estimation approach builds on standard econometric techniques and relies on a transparent identification assumption that is likely to hold in many contexts. As high frequency data on testing and positive become more widely available at finer geographic units, this approach could be applied to estimate population infection rates at the city or district level.

As physical distancing policies are relaxed, it will be essential that policymakers have access to timely data on infection rates. Given the potential for widespread undiagnosed infection, the expansion of randomized population-based PCR testing may play a key role in identifying localized outbreaks. Meanwhile, widespread implementation of serological testing will help identify the large numbers of individuals with some level of immunity to the virus.

[8]See (Ferguson et al., 2020; Johndrow, Lum and Ball, 2020; Javan, Fox and Meyers, 2020; Benatia, Godefroy and Lewis, 2020).

Acknowledgements

This study was supported by funding from the Social Sciences and Humanities Research Council (Grant: SSHRC 430-2017-00307).

References

Ai, T., Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia. 2020. "Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report on 1014 Cases." *Radiology*, DOI: 10.1148/radiol.2020200642.

Bai, Y., L. Yao, T. Wei, F. Tian, D. Jin, and et al. 2020. "Presumed Asymptomatic Carrier Transmission of COVID-19." *JAMA*, doi:10.1001.

Benatia, D., R. Godefroy, and J. Lewis. 2020. "Estimating COVID-19 Prevalence in the United States: A Sample Selection Model Approach." medRxiv Working Paper.

Berry, I., J. Soucy, A. Tuite, and D. Fisman. 2020. "Open Access Epidemiological Data and an Interactive Dashboard to Monitor the COVID-19 Outbreak in Canada." CMAJ: DOI: 10.1503/cmaj.75262.

Bhattacharya, J., E. Bendavid, B. Mulaney, N. Sood, S. Shah, and et al. 2020. "COVID-19 Antibody Seroprevalence in Santa Clara County, California." Working Paper.

Blundell, R., and M. Costa Dias. 2002. "Evaluation Methods for Non-experimental Data." *Fiscal Studies*, 21(4): 427–468.

Cai, J., J. Xu, D. Lin, Z. Yang, L. Xu, and et al. 2020. "A Case Series of Children with 2019 Novel Coronavirus Infection: Clinical and Epidemiological Features." *Clinical Infectious Disease*, doi: 10.1093/cid/ciaa198.
CBC

CBC. April 8, 2020. "Quebec Premier Says All Patients and Staff to be Tested at Long-term Care Homes." `https://www.cbc.ca/news/canada/montreal/quebec-covid-19-april-8-1.5525861`.
CDC

CDC. 2020. "Centers for Disease Control and Prevention: Coronavirus (COVID-19)." `https://www.cdc.gov/coronavirus/2019-ncov/index.html`.

Clay, K., J. Lewis, and E. Severnini. 2018. "Pollution, Infectious Disease, and Mortality: Evidence from the 1918 Spanish Influenza Pandemic." *Journal of Economic History*, 78(4): 1179–1209.

Clay, K., J. Lewis, and E. Severnini. 2019. "What Explains Cross-City Variation in Mortality during the 1918 Influenza Pandemic? Evidence from 440 U.S. Cities." *Economics and Human Biology*, 35: 42–50.

Conarck, B., and D. Chang. April 24, 2020. "Miami-Dade Has Tens of Thousands of Missed Coronavirus Infections, UM Survey Finds." `https://www.miamiherald.com/news/coronavirus/article242260406.html`.

CSSE, JHU. 2020. "Johns Hopkins Center for Systems Science and Engineering. Coronavirus COVID-19 Global Cases." `https://coronavirus.jhu.edu/`.

Das, M., W. Newey, and F. Vella. 2003. "Nonparametric Estimation of Sample Selection Models." *The Review of Economic Studies*, 70(1): 33–58.

Dong, Y., X. Mo, Y. Hu, X. Qi, F. Jiang, and et al. 2020. "Epidemiological Characteristics of 2143 Pediatric Patients with 2019 Coronavirus Disease in China." *Pediatrics*, doi: 10.1542.

Ferguson, N., D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba, and G. Cuomo-Dannenburg. 2020. "Impacts of Non-pharmaceutical Interventions to Reduce COVID-19 Mortality and Healthcare Demand." London: Imperial College COVID-19 Response Team.

Goodman, D., and M. Rothfeld. April 23, 2020. "1 in 5 New Yorkers May Have Had COVID-19, Antibody Tests Suggest." `https://www.nytimes.com/2020/04/23/nyregion/coronavirus-antibodies-test-ny.html`.

Han, Y., J. Lam, V. Li, P. Guo, Q. Zhang, A. Wang, J. Crowcroft, S. Wang, J. Fu, Z. Gilani, and J. Downey. 2020. "The Effects of Outdoor Air Pollution Concentrations and Lockdowns on COVID-19 Infections in Wuhan and Other Provincial Capitals in China." Working Paper.

Heckman, J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economics and Social Measurement*, 5(4): 475–492.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*, 4(7): 153–162.

Heckman, J., R. Lalonde, and J. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics*, ed. O. Ashenfelter and D. Card, 1866–2097. Amsterdam:North-Holland.

Hoehl, S., H. Rabenau, A. Berger, M. Kortenbusch, J. Cinatl, and et al. 2020. "Evidence of SARS-CoV-2 Infection in Returning Travelers from Wuhan, China." *New England Journal of Medicine*, 382: 1278–1280.

Huang, R., J. Xia, Y. Chen, C. Shan, and C. Wu. 2020. "A Family Cluster of SARS-CoV-2 Infection Involving 11 Patients in Nanjing, China." *The Lancet Infectious Disease*, doi: 10.1016/S1473-3099(20)30147-X.

Javan, E., S. Fox, and L. Meyers. 2020. "Probability of Current COVID-19 Outbreaks in All US Counties." Working Paper.

Johndrow, J., K. Lum, and P. Ball. 2020. "Estimating SARS-CoV-2 Positive Americans using Deaths-only Data." Working Paper.

Jones, A. April 22, 2020. "Coronavirus: Ontario to Test All Long Term Care Residents, Staff for COVID-19." `https://globalnews.ca/news/6852825/ontario-test-all-long-term-care-residents-staff-coronavirus/`.

Katz, J., and M. Sanger-Katz. 2020. "Deaths in New York City are More than Double the Usual Total." New York Times. `https://www.nytimes.com/interactive/2020/04/10/upshot/coronavirus-deaths-new-york-city.html`.

Lauer, S., K. Grantz, Q. Bi, F. Jones, Q. Zheng, H. Meredith, A. Azman, N. Reich, and J. Lesser. 2020. "The Incubation Period of Coronavirus Disease 2019 (COVID-19) from Publicly Reported Confirmed Cases: Estimation and Application." *New England Journal of Medicine*, DOI: 10.7326/M20-0504.

Li, Q., X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, M. Med, K. Leung, and E. Lau. 2020*a*. "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus – Infected Pneumonia." *New England Journal of Medicine*, DOI: 10.1056/NEJMoa2001316.

Li, R., S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. 2020*b*. "Substantial Undocumented Infection Facilitates the Rapid Dissemination of Novel Coronavirus (SARS-Cov2)." *Science*, 10.1126/science.abb3221.

Liu, J., X. Xie, Z. Zhong, W. Zhao, C. Zheng, and F. Wang. 2020. "Chest CT for Typical 2019-nCoV Pneumonia: Relationship to Negative RT-PCR Testing." *Radiology*, DOI: 10.1148/radiol.2020200330.

Lu, X., L. Zhang, H. Du, J. Zhang, Y. Li, and et al. 2020. "SARS-CoV-2 Infection in Children." *New England Journal of Medicine*, doi: 10.1056.

Meyer, R., E. Kissane, and A. Madrigal. 2020. "The COVID Tracking Project." `https://covidtracking.com/`.

Newey, W. 2009. "Two-Step Series Estimation of Sample Selection Models." *Econometrics Journal*, 12(S1): S217–S229.

Pan, X., D. Chen, Y. Xia, X. Wu, T. Li, and et al. 2020. "Asymptomatic Cases in a Family Cluster with SARS-CoV-2 Infection." *The Lancet Infectious Disease*, 20(4): 410–411.

Perkins, A., S. Cavany, S. Moore, R. Oidtman, A. Lerch, and M. Poterek. 2020. "Estimating Unobserved SARS-CoV-2 Infections in the United States." medRxiv Working Paper.

Prakash, N., and E. Hall. 2020. "Doctors and Nurses Say More People are Dying of COVID-19 in the US than We Know." Buzzfeed. `https://www.buzzfeednews.com/article/nidhiprakash/coronavirus-update-dead-covid19-doctors-hospitals`.

Reid, A. 2020. "The Incidence of COVID-19 Infection in Canada? New Survey Points to over 100,000 Households." Report: Angus Reid Institute.

Riou, J., A. Hauser, M. Counotte, and C. Althaus. 2020*a*. "Adjusting Age-Specific Case Fatality Rates during the COVID-19 Epidemic in Hubei, China, January and February." medRxiv Working Paper.

Riou, J., A. Hauser, M. Counotte, C. Margossian, G. Konstantinoudis, N. Low, and C. Althaus. 2020*b*. "Estimation of SARS-CoV-2 Mortality during the Early Stages of and Epidemic: A Modelling Study in Hubei, China and Norther Italy." Working Paper.
Statistics Canada

Statistics Canada. 2020. "Population Estimates on July 1, by Age and Sex (Table 17-10-0005-01." Statistics Canada. DOI: 10.25418/1710000501-eng.

Verity, R., L. Okell, I. Dorigatti, P. Winskill, C. Whittaker, and et al. 2020. "Estimates of the Severity of Coronavirus Disease 2019: A Model-based Analysis." *Lancet Infectious Disease*, doi: 10.1016/S1473-3099(20)30243-7.

Wooldridge, J. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA:MIT Press.

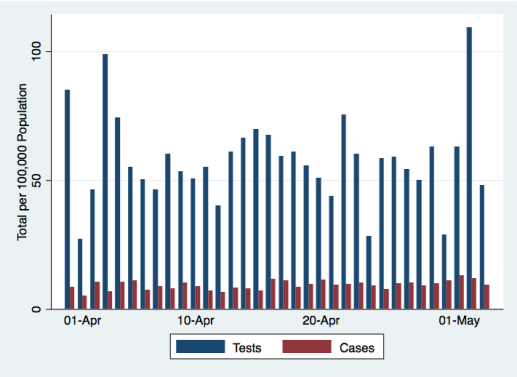
Wu, X., R. Nethery, B. Sabath, D. Braun, and F. Dominici. 2020. "Exposure to Air Pollution and COVID-19 Mortality in the United States." Working Paper.

Yang, Y., M. Yang, C. Shen, F. Wang, J. Yuan, J. Li, M. Zhang, Z. Wang, and L. Xing. 2020. "Evaluating the Accuracy of Different Respiratory Specimens in the Laboratory Diagnosis and Monitoring the Viral Shedding of 2019-nCoV Infections." medRxiv Working Paper.

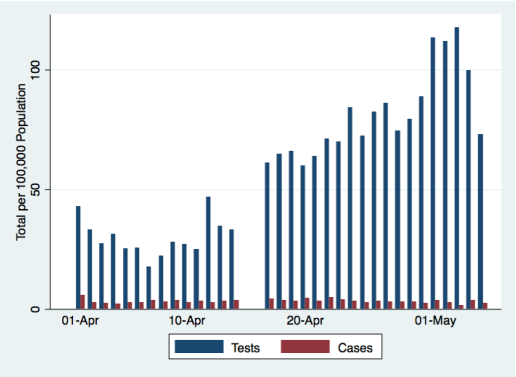
Zhou, F., T. Yu, R. Du, G. Fan, Y. Liu, and et al. 2020. "Clinical Course and Risk Factors for Mortality of Adult Inpatients with COVID-19 in Wuhan, China." *Lancet*, doi: 10.1016/S0140-6736(20)30566-3.

Figures and Tables

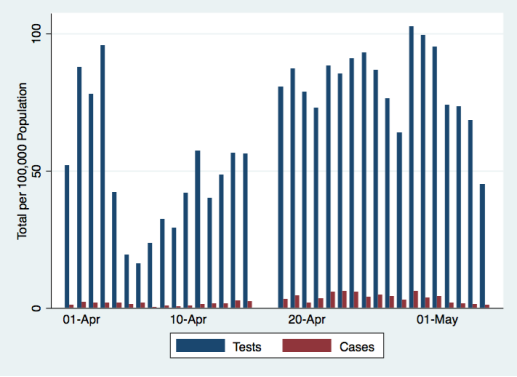
Figure 1: Daily Testing and New Cases Across Provinces



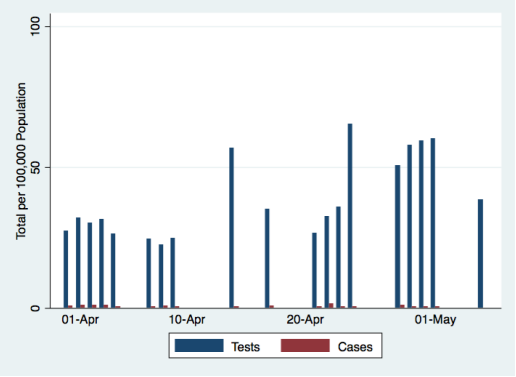
(a) Quebec



(b) Ontario



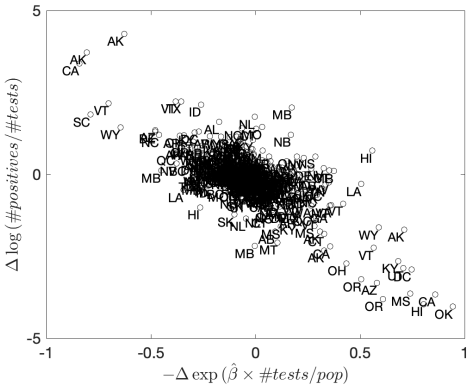
(c) Alberta



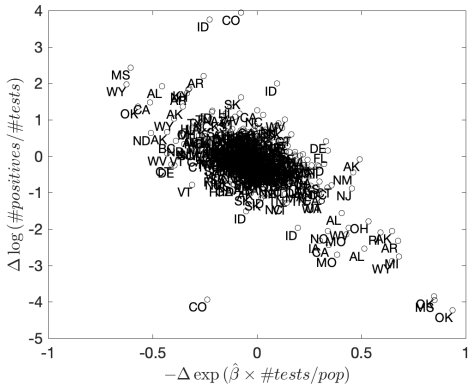
(d) B.C.

Notes: This figure reports the total daily coronavirus tests and the number of new cases per 100,000 population by province. The trends are based on data from (Berry et al., 2020). We exclude days in which there were identified changes in provincial reporting standards and days in provincial health authorities did not release information on completed tests.

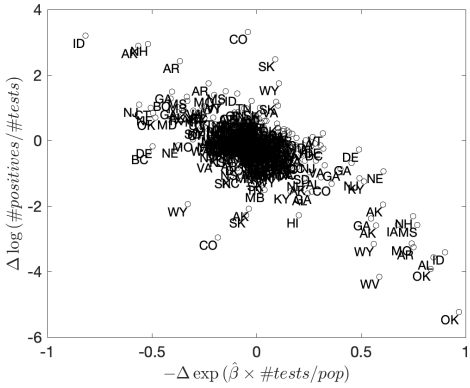
Figure 2: Daily Changes in Testing and the Share of Positive Cases



(a) Period 1: March 31 - April 7



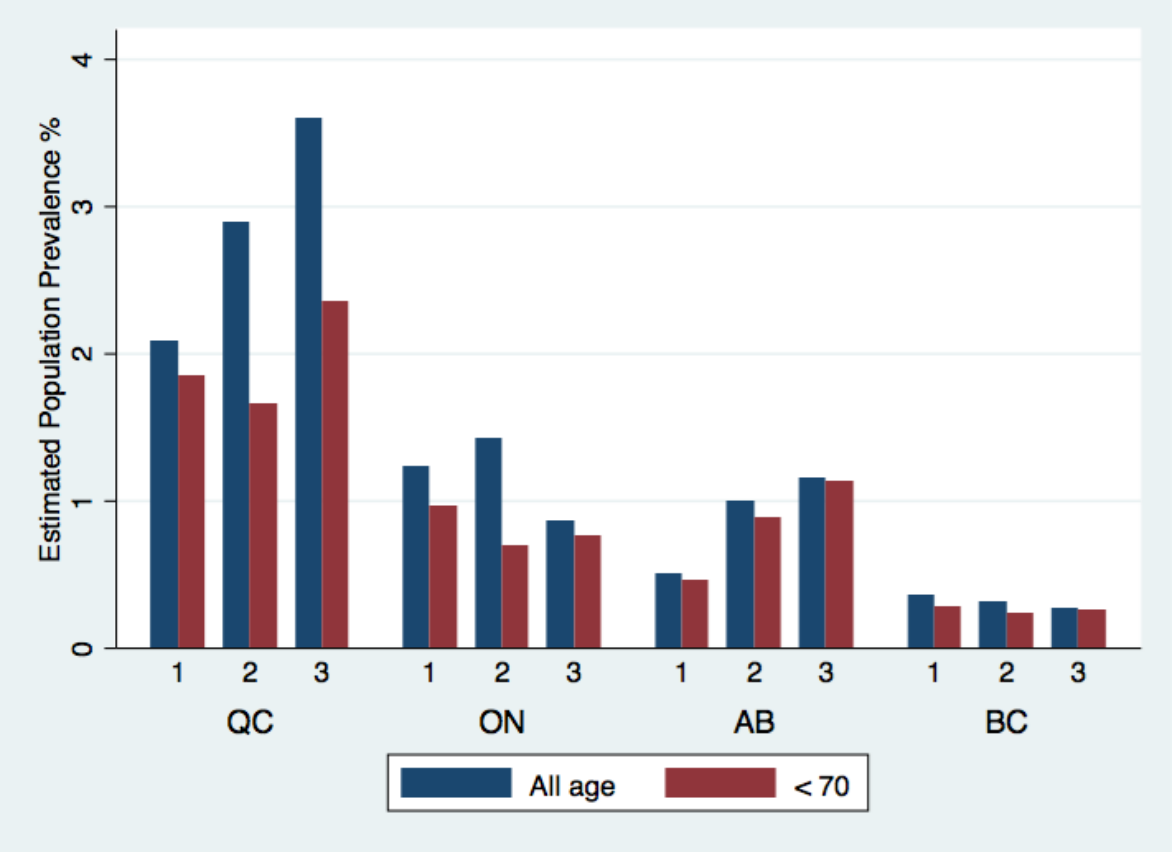
(b) Period 2: April 14 - 21



(c) Period 3: April 28 - May 5

Notes: This figure reports the relationship between daily changes in the exponential of per capita testing and daily changes in the log share of positive tests for the three time periods: March 31 - April 7, April 14 - 21, and April 28 - May 5. The relationship in each period is obtained using the coefficient of β derived from the main estimates of equation (1) (see Table 1, cols. 1, 3, 5).

Figure 3: Population COVID-19 Infection Rates by Province and Period



Notes: This figure reports average population infection rates across provinces for three different time periods: 1 - (March 31 - April 4); 2 (April 14 - April 18); 3 (April 28 - May 2). These average infection rates are obtained using the estimation procedure described in Section 3. All age population prevalence estimates are based on all cases and tests. To derive population prevalence for less than 70 year olds, we subtract the number of cases among the elderly from the total daily cases and total daily tests across provinces.

Table 1: Coefficient Estimates from Equation (1)

	Period 1: Mar 31 - Apr 7		Period 2: Apr 14 - 21		Period 3: Apr 28 - May 5	
	All age	< 70	All age	< 70	All age	< 70
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Baseline Model						
α_1	11.570	11.704	10.004	10.347	8.199	8.159
	(2.090)	(2.157)	(1.564)	(1.625)	(1.640)	(1.679)
α_2	-23.975	-24.327	-20.765	-21.675	-16.026	-15.960
	(3.946)	(4.064)	(3.155)	(3.260)	(3.393)	(3.472)
α_3	17.545	17.781	15.628	16.230	12.255	12.225
	(2.230)	(2.292)	(1.929)	(1.984)	(2.096)	(2.139)
β	-1390.815	-1381.209	-1608.010	-1578.140	-1107.816	-1092.915
	(156.032)	(156.412)	(204.343)	(193.004)	(182.211)	(182.714)
σ_u	0.516	0.527	0.579	0.593	0.608	0.609
	(0.017)	(0.018)	(0.020)	(0.021)	(0.022)	(0.22)
Observations	443	443	410	408	399	399
Panel B: Augmented Model with Province / State Fixed Effects						
α_1	11.313	11.422	10.045	10.333	8.026	7.994
	(1.512)	(1.561)	(1.115)	(1.153)	(1.155)	(1.181)
α_2	-23.469	-23.757	-20.818	-21.589	-15.540	-15.487
	(2.856)	(2.943)	(2.248)	(2.315)	(2.399)	(2.450)
α_3	17.261	17.454	15.649	16.162	11.866	11.844
	(1.614)	(1.660)	(1.378)	(1.414)	(1.491)	(1.518)
β	-1405.202	-1394.489	-1596.620	-1573.924	-1111.405	-1096.945
	(117.348)	(117.919)	(143.886)	(137.678)	(131.964)	(132.088)
σ_u	0.512	0.522	0.575	0.589	0.602	0.604
	(0.012)	(0.012)	(0.014)	(0.015)	(0.015)	(0.015)
Observations	443	443	410	408	399	399

Notes: This table reports the estimation of the coefficients from Equation (1). We estimate the model separately for each time period and for all age versus cases among individuals less than 70 years old. Panel A reports the coefficient estimates from the baseline model. Panel B reports the estimates from augmented models that include province and state fixed effects. Heteroskedasticity robust standard errors are reported in parentheses.

Table 2: Estimated Population Infection Rates for COVID-19

	Positive Tests (%)	Estimated Population Prevalence (%)		Ave. Estimated Pop. Prevalence (%)	
		All age	< 70	All age	< 70
	(1)	(2)	(3)	(4)	(5)
Panel A: COVID-19 Prevalence in Early April					
	April 4	April 4		March 31 - April 4	
Quebec	14.22	2.22 [1.03, 4.82]	1.95 [0.87, 4.35]	2.08	1.85
Ontario	7.31	0.86 [0.41, 1.79]	0.61 [0.29, 1.32]	1.23	0.96
Alberta	4.93	0.69 [0.33, 1.43]	0.63 [0.30, 1.34]	0.51	0.46
B.C.	2.51	0.23 [0.11, 0.49]	0.12 [0.05, 0.26]	0.36	0.28
Panel B: COVID-19 Prevalence in Mid-April					
	April 18	April 18		April 14 - 18	
Quebec	13.95	2.70 [1.52, 4.81]	2.56 [1.45, 4.53]	2.89	1.66
Ontario	5.93	1.21 [0.67, 2.18]	0.80 [0.44, 1.48]	1.42	0.7
Alberta	4.03	1.11 [0.59, 2.09]	1.05 [0.55, 2.00]	1.00	0.89
B.C.	2.79	0.43 [0.24, 0.75]	0.36 [0.20, 0.63]	0.31	0.24
Panel C: COVID-19 Prevalence in Early May					
	May 2	May 2		April 28 - May 2	
Quebec	10.87	2.91 [1.51, 5.63]	1.98 [1.01, 3.90]	3.60	2.35
Ontario	2.69	0.76 [0.39, 1.47]	0.75 [0.38, 1.48]	0.86	0.76
Alberta	2.56	0.60 [0.32, 1.13]	0.57 [0.30, 1.10]	1.16	1.13
B.C.	1.25	0.23 [0.13, 0.43]	0.24 [0.13, 0.45]	0.27	0.26

Notes: Column (1) reports the fraction of positive tests on the relevant day. Columns 2 - 3 report the coefficient estimates for population prevalence of COVID-19 based on the methodology described in Section 3. Heteroskedasticity robust 95% confidence intervals are reported in brackets. We report the results for all age prevalence and prevalence among individuals less than 70 years old. Column 4 - 5 report the average estimates for population prevalence of COVID-19 for the three time periods.

Table 3: Robustness Exercises: Fixed Effects Models

	Estimated Population Prevalence (%)				Ave. Estimated Pop. Prevalence (%)			
	All age		< 70		All age		< 70	
	Baseline	Add fixed effects	Baseline	Add fixed effects	Baseline	Add fixed effects	Baseline	Add fixed effects
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: COVID-19 Prevalence in Early April								
	April 4				March 31 - April 4			
Quebec	2.22 [1.03, 4.82]	2.32 [1.33, 4.08]	1.95 [0.87, 4.35]	2.04 [1.14, 3.67]	2.08	2.17	1.85	1.93
Ontario	0.86 [0.41, 1.79]	0.89 [0.52, 1.52]	0.61 [0.29, 1.32]	0.64 [0.37, 1.11]	1.23	1.28	0.96	1.00
Alberta	0.69 [0.33, 1.43]	0.72 [0.42, 1.22]	0.63 [0.30, 1.34]	0.65 [0.38, 1.14]	0.51	0.54	0.46	0.48
B.C.	0.23 [0.11, 0.49]	0.24 [0.14, 0.41]	0.12 [0.05, 0.26]	0.12 [0.07, 0.22]	0.36	0.38	0.28	0.29
Panel B: COVID-19 Prevalence in Mid-April								
	April 18				April 14 - 18			
Quebec	2.70 [1.52, 4.81]	2.67 [1.77, 4.03]	2.56 [1.45, 4.53]	2.55 [1.70, 3.82]	2.89	2.85	1.66	1.66
Ontario	1.21 [0.67, 2.18]	1.19 [0.78, 1.82]	0.80 [0.44, 1.48]	0.80 [0.52, 1.24]	1.42	1.40	0.7	0.76
Alberta	1.11 [0.59, 2.09]	1.09 [0.70, 1.72]	1.05 [0.55, 2.00]	1.05 [0.66, 1.66]	1.00	0.98	0.89	0.89
B.C.	0.43 [0.24, 0.75]	0.42 [0.28, 0.63]	0.36 [0.20, 0.63]	0.35 [0.23, 0.53]	0.31	0.31	0.24	0.23
Panel C: COVID-19 Prevalence in Early May								
	May 2				April 28 - May 2			
Quebec	2.91 [1.51, 5.63]	2.97 [1.87, 4.73]	1.98 [1.01, 3.90]	2.02 [1.25, 3.26]	3.60	3.66	2.35	2.39
Ontario	0.76 [0.39, 1.47]	0.77 [0.48, 1.24]	0.75 [0.38, 1.48]	0.78 [0.47, 1.24]	0.86	0.88	0.76	0.78
Alberta	0.60 [0.32, 1.13]	0.61 [0.39, 0.96]	0.57 [0.30, 1.10]	0.59 [0.37, 0.93]	1.16	1.18	1.13	1.16
B.C.	0.23 [0.13, 0.43]	0.24 [0.15, 0.37]	0.24 [0.13, 0.45]	0.24 [0.16, 0.37]	0.27	0.27	0.26	0.26

Notes: This table explores the sensitivity of the findings to controls for province (state) fixed effects. Columns 1 - 4 report the estimated population infection rates on the relevant date. Heteroskedasticity robust 95% confidence intervals are reported in brackets. Columns 5 - 8 report the average estimates for population prevalence of COVID-19 for the three time periods. Columns 1, 3, 5, and 7 report the baseline estimates, while columns 2, 4, 6, and 8 report the estimates based on augmented models that include province and state fixed effects.

Table 4: Diagnosed Cases and Estimated Total Cases of COVID-19

	Positive COVID-19 Tests, by April 23	Estimated Total COVID-19 Cases	Ratio of Total Cases to Positive Tests (2)/(1)	COVID-19 Tests per 1,000 Population
	(1)	(2)	(3)	(4)
Quebec	21,832	245,215	11.2	21.9
Ontario	13,995	206,845	14.8	13.4
Alberta	3,720	43,713	11.8	27.0
B.C.	1,824	15,721	8.6	13.5

Notes: Columns (1) reports the cumulative number of positive COVID-19 tests by April 23. Column (2) reports the total number of COVID-19 cases implied by the average estimated population prevalence from April 14 to April 18 (Table 2, Panel B, col. 4). Column (4) reports the cumulative number of COVID-19 tests by April 23 per 1,000 population.

A Appendix: Theoretical Framework from Benatia, Godefroy and Lewis (2020)

In this section, we present the theoretical framework developed in Benatia, Godefroy and Lewis (2020) to estimate COVID-19 prevalence. This framework motivates estimating equation (1).

A.1 Theory

To evaluate population disease prevalence, we developed a simple selection model for COVID-19 testing and used the framework to link observed rates of positive tests to population disease prevalence. We considered a stable population, denoting A and B as the numbers of sick and healthy individuals, respectively. Let p_n denote the probability that a sick person is tested and q_n the probability that a healthy person is tested, given a total number of tests, n. Thus, we have:

$$n = p_n A + q_n B,$$

and the number of positive tests is:

$$s = p_n A.$$

This simple framework highlights how non-random testing will bias estimates of the population disease prevalence. Using Bayes' rule, we can write the relative probability of testing as the following:

$$\frac{q_n}{p_n} = \frac{Pr(sick|n)/Pr(healthy|n)}{Pr(sick|tested,n)/Pr(healthy|tested,n)},$$

which is equal to one if tests are randomly allocated, $Pr(sick|tested,n) = Pr(sick|n)$. When testing is targeted to individuals who are more likely to be sick, we have $Pr(sick|tested,n) > Pr(sick|n)$ and $Pr(healthy|tested,n) < Pr(healthy|n)$, so the ratio will fall between zero and one. In this scenario, the ratio of sick to healthy people in the sample, $p_n A/q_n B$, will exceed the ratio in the overall population, A/B.

We specified the following functional form for the relative probability of testing:

$$\frac{q_n}{p_n} = \frac{1}{1 + e^{-a-bn}}, \tag{A.1}$$

The term $e^{-a-bn} > 0$ reflects the fact that testing has been targeted towards higher risk populations, with the intercept, $-a$, capturing the severity of selection bias when testing is limited. Meanwhile, the coefficient $b > 0$ identifies how selection bias decreases with n as the ratio q_n/p_n approaches one. Intuitively, as testing expands, the sample will become more representative of the overall population, and the selection bias will diminish.

Combining both equations, we have:

$$\log \frac{s}{n} = -\log\left(1 + \frac{1}{1 + e^{-a-bn}} \frac{B}{A}\right).$$

We used the fact that the ratio of negative to positive tests is much larger than one to make the following approximation:

$$\begin{aligned}
\log \frac{s}{n} &\approx -\log\left(\frac{1}{1 + e^{-a-bn}} \frac{B}{A}\right) \\
&\approx \log\left(1 + e^{-a-bn}\right) - \log \frac{B}{A} \\
&\approx \sum_{k=1}^{M} \frac{(-1)^{k-1} e^{-ka}}{k} e^{-kbn} - \log \frac{B}{A}
\end{aligned} \tag{A.2}$$

Given a change in the number of tests conducted in a particular population, n_1 to n_2, equation (2) implies the following change in the share of positive tests:

$$\log \frac{s_2}{n_2} - \log \frac{s_1}{n_1} \approx \sum_{k=1}^{M} \frac{(-1)^{k-1} e^{-ka}}{k} \left(e^{-kbn_2} - e^{-kbn_1}\right) \tag{A.3}$$

Our estimating equation (1) is based on a third-order approximation of equation (A.3), since higher order terms are found to be insignificant.