A Statistical Approach to Batched Prevalence Testing for Coronavirus

by William J. Berger[1], Konrad R. Dabrowski[2], Jake A. Robinson[2], and Adam C. Sales[3]

As cases of novel coronavirus mount, the ability to conduct expeditious prevalence testing becomes paramount. A statistical approach to batched prevalence testing offers a more rapid and efficient means of monitoring at-risk populations.

Early papers offer straightforward models of group testing samples in order to quickly determine the presence of biomarkers in large groups [3, 5, 7]]. We are currently faced with a problem similar in nature, requiring urgent prevalence testing surpassing even the HIV epidemic [8, 9]. We propose that in many cases testing for novel coronavirus can be done more expeditiously and at lower cost by batching samples together. This proposal goes beyond earlier work by more explicitly modeling cost constraints, making findings particularly helpful to resource poor localities. There are two primary levels at which such a batched method might be implemented. First, batching can be used to expedite clinical panels, thereby triaging the diagnostic process. Here, group testing requires a halving procedure where initial samples are aliquoted, such that sample A can be batched with the group and sample B is saved to be evaluated individually were the batched sample to return positive [2]. Second, batching can be employed for prevalence testing in order to improve population-level estimates of infection. We write here to show how these aims might be met simultaneously.

Ab initio, assume no false negatives or false positives. The aim is to test some number of people N, in groups of size n. We will assume that the cost to test a batch of samples, c_b, is no less than the cost of testing an individual sample, c_i. Indeed, there are a number of aspects of the PCR diagnostic panel that must be implemented on each sample, making perfect streamlining impossible. The cost of implementing the batched technique will be c_b in the case that the batch tests negative, and $c_b + nc_i$ in the case where the batch tests positive. This is because if even one member of the group tests positive, each person's B sample will need to be run individually. Assuming each batch is a random subset of the

[1] University of Pennsylvania
[2] Temple University
[3] University of Texas at Austin

population, the probability that the group will test negative is $(1-p)^n$, where p is the prevalence in the local population.

The expected cost of one batched test is:

$$\mathbb{E}[\text{batch cost}] = c_b + nc_i(1-(1-p)^n) \tag{1}$$

Given N total people, N/n batched tests will yield a total expected cost of:

$$\mathbb{E}[\text{total cost}] = N\left(\frac{c_b}{n} + c_i(1-(1-p)^n)\right) \tag{2}$$

Under individual testing, the total cost is Nc_i, so the expected cost of batching relative to individual testing is:

$$\mathbb{E}[\text{relative cost}] = \frac{c_b/c_i}{n} + (1-(1-p)^n) \tag{3}$$

Figure 1 plots $\mathbb{E}$[relative cost] (3) for a range of values for p. In the first panel $c_b = c_i$, and in the second and third c_b is two and three times c_i, respectively.[4] These latter two more accurately reflect the realities of implementing RT-qPCR from respiratory specimens. PCR isolation of viral RNA is a labor intensive process, even prior to cDNA synthesis and amplification of target sequences [1]. There are thus inevitably process-based hurdles which limit the value of c_b/c_i on the high end. The bolded horizontal line running parallel to the x-axis reflects the break-even point, where batch and individual testing are equally efficient in expectation. Above the line individual testing is preferable, and below batched testing is.

Though the four values of underlying prevalences are instructive, they are primarily illustrative as worldwide incidence exhibits considerable heterogeneity. Regardless of peak infection levels in particular regions though, prevalence testing will remain crucial for early monitoring and detection of the virus in emergent contexts. A region with a population of 10 million with 10,000 cases—roughly the total number of cases reported in South Korea—would translate to a prevalence of 0.001. Early prevalence or sentinel testing could assume an even lower rate of infection, e.g. reported rates in India and Brazil at the end of March, both substantially below a 0.0001 level.

[4] The cost of a batch test c_b presumably depends on the size of the batch, n; we set it to a constant for the sake of simplicity.

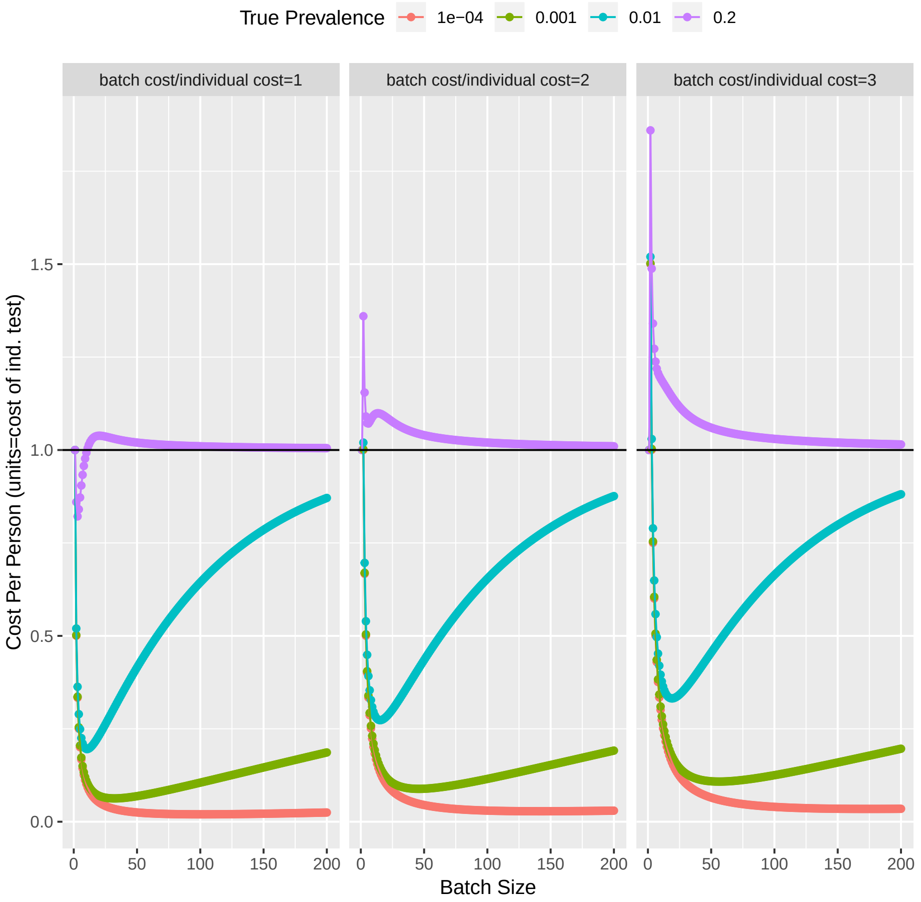

Figure 1: Cost schedule for batched testing for various prevalence levels

These results indicate that batched testing provides the greatest benefits for at-risk populations where the epidemic is in its early stages. Populous and dense regions in the developing world, which currently have low levels of the disease despite large populations, would particularly stand to benefit from a batched approach. Savings offered by group testing would also reduce the cost burden on such developing countries, allowing for a wider net to be cast in monitoring future outbreaks.

While rates are still sufficiently low in many North American cities to warrant batched prevalence testing, the technique may also prove advantageous in the near-term were coronavirus to wane in the North America and Europe, only for a second wave to emerge upon reinfection from regions with a delayed in-

p	c_b/c_i	Optimal Batch Size	Savings
0.0001	1	101	0.98
	2	142	0.97
	3	175	0.97
0.001	1	32	0.94
	2	46	0.91
	3	56	0.89
0.01	1	11	0.80
	2	15	0.73
	3	19	0.67
0.2	1	3	0.18
	2	1	0.00
	3	1	0.00

Table 1: Optimal batch sized for various prevalences and cost ratios

fection schedule [6]. Batched prevalence testing could then serve as a crucial early warning of reinfection in countries where the disease had previously been contained.

Table 1 shows optimal batch sizes n, and corresponding values of $1 - \mathbb{E}$[relative cost], or the expected proportion of total individual testing costs saved due to batching. Efficiency is greatly enhanced at low levels of prevalence. With prevalence at or above 20% it becomes impossible to support batched prevalence testing. Even at prevalence levels of 1%, samples could only be grouped in batches of between 10 and 20 samples, under optimal theoretical conditions. When batched prevalence testing is employed in at-risk, low prevalence regions, however, batching can theoretically reduce testing costs by greater than 90%. Though estimates which assume perfect sensitivity and specificity are unrealistic, we would not expect levels of viral RNA in batched runs to be substantially more dilute than individual samples, since the qPCR technique exponentially amplifies the signal on the front-end. This amplification allows for batching at the extraction phase, thereby eliminating redundancies all through the PCR process. Indications of no false negatives at levels of $10^{0.5} copies/\mu L$ provide reassurance against the worry of false negatives [1]. Still, performing multiple runs of batched tests (though it increases value of c_b/c_i) is another way to mitigate these concerns while maintaining efficiency according to the cost schedule above.

As we noted, there are a number of practical factors which limit the efficiency of batched testing respiratory samples from suspected coronavirus patients. For one, given early testing bottlenecks in the U.S., such a procedure would be of limited utility given that the prevalence of the disease among those to whom the test is administered is high. Furthermore, additional supplementary processes coming on line further increase the rate and capacity at which testing can be conducted [4]. Regardless, prevalence testing will remain crucial in public health's fight against the virus. Not only will base rates of the disease remain important to monitor in North America and Europe, but batched prevalence testing provides special advantages for resource poor countries expecting to face the next wave of infections.

References

[1] Centers for Disease Control. CDC 2019-novel coronavirus (2019-nCoV) real-time RT-PCR diagnostic panel, March 2020. URL `https://www.fda.gov/media/134922/download`.

[2] C. L. Chen and W. H. Swallow. Using group testing to estimate a proportion, and to test the binomial model. *Biometrics*, pages 1035–1046, 1990.

[3] R. Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[4] Food and Drug Administration. Coronavirus (covid-19) update: FDA issues first emergency use authorization for point of care diagnostic, March 2020. URL `shorturl.at/vNP35`.

[5] M. H. McCrady. The numerical interpretation of fermentation-tube results. *The Journal of Infectious Diseases*, pages 183–212, 1915.

[6] J. Qiu. Covert coronavirus infections could be seeding new outbreaks. *Nature*, March 2020.

[7] M. Sobel and P. A. Groll. Binomial group-testing with an unknown proportion of defectives. *Technometrics*, 8(4):631–656, 1966.

[8] H. Tamashiro, A. Fauquex, D. Heymann, J. Emmanuel, P. Sato, and

W. Maskill. Reducing the cost of hiv antibody testing. *The Lancet*, 342 (8863):87–90, 1993.

[9] X. M. Tu, E. Litvak, and M. Pagano. Screening tests: Can we get more by doing less? *Statistics in Medicine*, 13(19-20):1905–1919, 1994.