Technical Details for Delphi's Covid-Like Illness Survey, Run in Partnership with Facebook

Delphi Group, Carnegie Mellon University

April 17, 2020

1 Survey questions

Starting April 6, 2020, Facebook began directing a random sample of its users each day to our (CMU-run) survey. This survey starts with the first 4 questions:

1. In the past 24 hours, have you or anyone in your household had (yes/no for each):
    a. Fever (of 100 degrees or higher)
    b. Sore throat
    c. Cough
    d. Shortness of breath
    e. Difficulty breathing
2. How many people in your household (including yourself) are sick (fever, along with at least one other symptom from the above list)?
3. How many people are there in your household in total (including yourself)?
4. What is your current ZIP code?

There is a 5th question, which asks: How many additional people in your local community that you know personally are sick (fever, along with at least one other symptom from the above list)? We are still studying how to best use the information from answers to this question (it casts a "wider net" and therefore potentially provides an interesting perpsective beyond the household one.) There are also many other questions that follow, which go into more detail on symptoms and demographics. These are of interest to other researchers, but are not of interest to us for the purposes of the current system. The full survey can be found at: https://cmu.ca1.qualtrics.com/jfe/preview/SV_8zYl1sFN6fAFIxv?Q_SurveyVersionID=&Q_CHL=preview. (This links to a preview version, and no answers will be recorded.)

2 ILI and CLI indicators

Influenza-like illness or ILI is a standard indicator, and is defined by the CDC as: fever along with sore throat or cough. From the list of symptoms from Q1 on our survey, this means a AND (b OR c).

Covid-like illness or CLI is not a standard indicator, though from our discussions with the CDC, it seems reasonable to define it as: fever along with cough or shortness of breath or difficulty breathing. From the list of symptoms from Q1 on our survey, this means a AND (c OR d OR e).

CLI is (by definition) symptom-based, and most certainly does *not* mean a confirmed case of covid-19, but we are still interested in estimating percent CLI in the population at the finest possible temporal and geographic resolutions, because we believe it should be *predictive of covid-related hospitalizations in the days that follow*, similar to what has been well-established with percent ILI and influenza-related hospitalizations. Given its large overlap and therefore its small measurement cost (the fact that measuring it requires adding just one part, part b, to the survey), it seems worth measuring ILI as well. We may be able to use it to calibrate/adjust what we are seeing, in terms of CLI, with respect to a "background" influenza signal.

3 Household ILI and CLI

For a single survey, we are interested in the quantities:

- $X =$ the number of people in the household with ILI;
- $Y =$ the number of people in the household with CLI;
- $N =$ the number of people in the household.

Note that N comes directly from the answer to Q3, but neither X nor Y can be computed directly (because Q2 does not give an answer to the precise symptomatic profile of all individuals in the household, it only asks how many individuals have fever and at least one other symptom from the list). We hence estimate X and Y with the following simple strategy. Consider ILI, without a loss of generality (we apply the same strategy to CLI). Let Z be the answer to Q2.

- If the answer to Q1 does not meet the ILI definition, that is, a AND (b OR c) does not evaluate to true, then we report $X = 0$.
- If the answer to Q1 does meet the ILI definition, that is, a AND (b OR c) does evaluate to true, then we report $X = Z$.

This can only "over count" (result in too large estimates of) the true X and Y. This happens when some members of the household experience ILI and others experience CLI. In this case, for both X and Y, our simple strategy would return the sum of the ILI and CLI cases. (However, given the extremely high overlap between the definitions ILI and CLI, it is reasonable to believe that an individual would have both, or neither—and much more commonly neither. Therefore we do not "over counting" phenomenon is not practically very problematic.)

4 Estimating percent ILI and CLI

Let x and y be the number of people with ILI and CLI, respectively, over a given time period and in a given location (for example, the time period being a day, and a location being a particular MSA or metropolitan statistical area). Let n be the total number of people in this location. We are interested in estimating the true ILI and CLI proportions, which we denote by p and q, respectively:

$$p = \frac{x}{n} \quad \text{and} \quad q = \frac{y}{n}. \tag{1}$$

We estimate p and q across 4 different temporal-spatial aggregation schemes:

1. daily, at the MSA (metropolitan statistical area) level;
2. daily, at the HRR (hospital referral region) level;
3. daily, at the county level;
4. weekly, at the county level.

Note that these spatial aggregations are possible as we have the ZIP code of the household from Q4 of the survey. Our current rule-of-thumb is to discard any estimate (whether from daily-MSA, daily-HRR, daily-county, or weekly-county) that is comprised of less than 100 survey responses. The 4th aggregation, weekly-county, is therefore valuable because it "opens up" several counties that may have been too sparse on individual days.

In a given temporal-spatial unit (for example, daily-MSA), let X_i and Y_i denote number of ILI and CLI cases in the household, respectively (computed according to the simple strategy described in Section 3), and let N_i denote the total number of people in the household, in survey i, out of m surveys we collected. Then, our estimates of p and q are

$$\hat{p} = \frac{1}{m}\sum_{i=1}^{m}\frac{X_i}{N_i} \quad \text{and} \quad \hat{q} = \frac{1}{m}\sum_{i=1}^{m}\frac{Y_i}{N_i}. \tag{2}$$

Their estimated standard errors are:

$$\widehat{\mathrm{se}}(\hat{p}) = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\left(\frac{X_i}{N_i} - \hat{p}\right)^2} \quad \text{and} \quad \widehat{\mathrm{se}}(\hat{q}) = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\left(\frac{Y_i}{N_i} - \hat{q}\right)^2}. \tag{3}$$

5 Inverse probability weighting

When Facebook sends a user to our survey, it generates a random ID number and sends this to us as well. Once the user completes the survey, we pass this ID number back to Facebook to confirm completion, and in return receive a weight—call it w_i for user i. (To be clear, the random ID number that is generated is completely meaningless for any other purpose than receiving said weight, and does not allow us to access any information about the user's Facebook profile, or anything else whatsoever.)

We can use these weights to adjust our estimates of the true ILI and CLI proportions so that they are representative of the US population (rather than the Facebook population), according to a state-by-age-gender stratification of the US population from the 2018 Census March Supplement. In more detail, we receive $w_i = 1/\pi_i$, where π_i is an estimated probability (produced by Facebook) that an individual with the same state-by-age-gender profile as user i would take our CMU survey.

The adjustment we make follows the simple inverse probability weighting strategy (also called Horwitz-Thompson estimation). As before, in a given temporal-spatial unit (for example, daily-MSA), let X_i and Y_i denote number of ILI and CLI cases in the household, respectively (computed according to the simple strategy described in Section 3), and let N_i denote the total number of people in the household, in survey i, out of m surveys we collected. Also let $w_i = c/\pi_i$ denote the weight that accompanies for survey i, where $c > 0$ is chosen so that these weights have been self-normalized over the temporal-spatial unit of interest (meaning $\sum_{i=1}^m w_i = 1$). Then, our adjusted estimates of p and q are:

$$\hat{p}_w = \sum_{i=1}^m \frac{X_i}{N_i} w_i \quad \text{and} \quad \hat{q}_w = \sum_{i=1}^m \frac{Y_i}{N_i} w_i. \tag{4}$$

Their estimated standard errors are:

$$\widehat{\mathrm{se}}(\hat{p}) = \sqrt{\sum_{i=1}^m \Big(\frac{X_i}{N_i} - \hat{p}_w\Big)^2 w_i^2} \quad \text{and} \quad \widehat{\mathrm{se}}(\hat{q}) = \sqrt{\sum_{i=1}^m \Big(\frac{Y_i}{N_i} - \hat{q}_w\Big)^2 w_i^2}. \tag{5}$$

A Appendix: details behind the choice of estimator

Suppose there are h households total in the underlying population, and for household i, denote $\theta_i = N_i/n$. Then note that the quantities of interest, p and q in (1), are

$$p = \sum_{i=1}^h \frac{X_i}{N_i} \theta_i \quad \text{and} \quad q = \sum_{i=1}^h \frac{Y_i}{N_i} \theta_i.$$

Let $s \subseteq \{1, \dots, h\}$ denote sampled households, with $m = |s|$, and suppose we sampled household i with probability $\theta_i = N_i/n$ proportional to the household size. Then unbiased estimates of p and q are simply

$$\hat{p} = \frac{1}{m} \sum_{i \in s} \frac{X_i}{N_i} \quad \text{and} \quad \hat{q} = \frac{1}{m} \sum_{i \in s} \frac{Y_i}{N_i}, \tag{6}$$

which are same as in (2).

Note that we can again rewrite our quantities of interest as

$$p = \frac{\mu_x}{\mu_n} \quad \text{and} \quad q = \frac{\mu_y}{\mu_n},$$

where $\mu_x = x/h$, $\mu_y = y/h$, $\mu_n = n/h$ denote the expected number people with ILI per household, expected number of people with CLI per household, and expected number of people total per household, respectively, and h denotes the total number of households in the population. Suppose that instead of proportional sampling, we sampled households uniformly, resulting in $s \subseteq \{1, \dots, h\}$ denote sampled households, with $m = |s|$. Then the natural estimates of p and q are instead plug-in estimates of the numerators and denominators in the above,

$$\hat{p} = \frac{\bar{X}}{\bar{N}} \quad \text{and} \quad \hat{q} = \frac{\bar{X}}{\bar{N}} \tag{7}$$

where $\bar{X} = \sum_{i \in s} X_i/m$, $\bar{Y} = \sum_{i \in s} Y_i/m$, and $\bar{N} = \sum_{i \in s} N_i/m$ denote the sample means of $\{X_i\}_{i \in s}$, $\{Y_i\}_{i \in s}$, and $\{N_i\}_{i \in s}$, respectively.

Whether we consider (6) or (7) to be more natural—mean of fractions or fraction of means, respectively—depends on the sampling model: if we are sampling households proportional to household size, then it is (6); if we are sampling household uniformly, then it is (7). We settled on the former (equivalently, we settled on (2)) based on both conceptual and empirical supporting evidence.

- Conceptually, though we do not know the details, we have reason to believe that Facebook offers an essentially uniform random draw of eligible users—those 18 years or older—to take our survey. In this sense, the sampling is done proportional to the number of "Facebook adults" in a household: individuals 18 years or older, who have a Facebook account. Hence if we posit that the number of "Facebook adults" scales linearly with the household size, which seems to us like a reasonable assumption, then sampling would still be proportional to household size. (Notice that this would remain true no matter how small the linear coefficient is, that is, it would even be true if Facebook did not have good coverage over the US.)

- Empirically, we have computed the distribution of household sizes (proportion of households of size 1, size 2, size 3, etc.) in the Facebook survey data thus far, and compared it to the distribution of household sizes from the census. These align quite closely, also suggesting that sampling is likely done proportional to household size.