# Technical Details of Delphi's Survey-Based Real-Time Indicators of COVID-19 Activity

Delphi Group, Carnegie Mellon University

May 21, 2020

## 1 Background

Starting April 6, 2020, we began running symptom surveys through several popular websites. In particular, each day from this date onwards, our survey partners direct a random sample of their users each day to our survey, hosted on a CMU-Qualtrics site. This survey starts with the first 5 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?
5. 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)?

Beyond these 5 questions, there are also many other questions that follow in the survey, which go into more detail on symptoms and demographics. These are primarily of interest to other researchers, but could still be useful for our purposes. The full survey can be found here. (This links to a preview version, and no answers will be recorded.) As of this writing, the median number of survey responses per day is about 72,000.

## 2 Signal names

As of this writing, the available `signal` names in the Epidata API are:

(a) `raw_ili` and `raw_cli`: estimates of the percentage of people with ILI (influenza-like illness), respectively CLI (COVID-like illness), as described in Section **??**;

(b) `raw_hh_cmnty_cli` and `raw_nohh_cmnty_cli`: estimates of the percentage of people who know someone in their local community, including their household and not including their household, respectively, with CLI, as described in Section 6;

(c) all of the above, but where `ili` is replaced by `wili`, and `cli` by `wcli`: estimates formed by applying inverse probability weighting to reflect the US population, as described in Section 7;

(d) all of the above, but where `raw` is replaced by `smoothed`: estimates formed by pooling together the last 7 days of data, as described in Section 9.

# 3 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).

# 4 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 degree of overlap between the definitions of ILI and CLI, it is reasonable to believe that an individual would have both, or neither—with of course neither being much more common here. Therefore we do not consider this "over counting" phenomenon practically problematic.

# 5 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 particular day, and a location being a particular county). Let $n$ be the total number of people in this location. We are interested in estimating the true ILI and CLI percentages, which we denote by $p$ and $q$, respectively:

$$p = 100 \cdot \frac{x}{n} \quad \text{and} \quad q = 100 \cdot \frac{y}{n}. \tag{1}$$

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

1. daily, at the county level;
2. daily, at the MSA (metropolitan statistical area) level;
3. daily, at the HRR (hospital referral region) level;
4. daily, at the state 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 at a county, MSA, HRR, or state level) that is comprised of less than 100 survey responses.

In a given temporal-spatial unit (for example, daily-county), 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 4), 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 $q$ (see Appendix A.1 for motivating details) are:

$$\hat{p} = 100 \cdot \frac{1}{m} \sum_{i=1}^{m} \frac{X_i}{N_i} \quad \text{and} \quad \hat{q} = 100 \cdot \frac{1}{m} \sum_{i=1}^{m} \frac{Y_i}{N_i}. \tag{2}$$

Their estimated standard errors are:

$$\widehat{\text{se}}(\hat{p}) = 100 \cdot \frac{1}{m} \sqrt{\sum_{i=1}^{m} \Big(\frac{X_i}{N_i} - \hat{p}\Big)^2} \quad \text{and} \quad \widehat{\text{se}}(\hat{q}) = 100 \cdot \frac{1}{m} \sqrt{\sum_{i=1}^{m} \Big(\frac{Y_i}{N_i} - \hat{q}\Big)^2}, \tag{3}$$

simply the sample standard deviations of the estimators in (2) (treating $m$ as fixed).

# 6 Estimating percent "community CLI"

Over a given time period, and in a given location, let $u$ be the number of people who know someone in their community with CLI, and let $v$ be the number of people who know someone in their community, outside of their household, with CLI. With $n$ denoting the number of people total in this location, we are interested in the percentages:

$$a = 100 \cdot \frac{u}{n} \quad \text{and} \quad b = 100 \cdot \frac{y}{n}. \tag{4}$$

We will estimate $a$ and $b$ across the same 4 temporal-spatial aggregation schemes as before.

For a single survey, let:

- $U = 1$ if and only if a positive number is reported for Q2 or Q5;
- $V = 1$ if and only if a positive number is reported for Q2.

In a given temporal-spatial unit (for example, daily-county), let $U_i$ and $V_i$ denote these quantities for survey $i$, and $m$ denote the number of surveys total. Then to estimate $a$ and $b$, we simply use:

$$\hat{a} = 100 \cdot \frac{1}{m} \sum_{i=1}^{m} U_i \quad \text{and} \quad \hat{b} = 100 \cdot \frac{1}{m} \sum_{i=1}^{m} V_i. \tag{5}$$

Their estimated standard errors are:

$$\widehat{\text{se}}(\hat{a}) = 100 \cdot \sqrt{\frac{\hat{a}(1-\hat{a})}{m}} \quad \text{and} \quad \widehat{\text{se}}(\hat{b}) = 100 \cdot \sqrt{\frac{\hat{b}(1-\hat{b})}{m}}, \tag{6}$$

which are the plug-in estimates of the standard errors of the binomial proportions in (5) (treating $m$ as fixed).

Note that $\sum_{i=1}^{m} U_i$ is the number of survey respondents who know someone in their community with *either ILI or CLI*, and not CLI alone; and similarly for $V$. Hence $\hat{a}$ and $\hat{b}$ in (5) will generally overestimate $a$ and $b$ in (4). However, given the extremely high overlap between the definitions of ILI and CLI, as we explained at the end of Section 4, we do not consider this to be practically very problematic.

# 7 Inverse probability weighting

Notice that the estimates defined in last two sections actually reflect the percentage of individuals with ILI and CLI, and individuals who know someone with CLI, with respect to the population of US users from our survey partners. (To be precise, the estimates from Section 5 actually reflect the percentage individuals with ILI and CLI, with respect to the population of US users from our survey partners *and* their households members). In reality, our estimates are even

further skewed by the propensity of people in the population of US users from our survey partners to take our survey in the first place.

When our survey partners send a user to our survey, they generate a random ID number and sends this to us as well. Once the user completes the survey, we pass this ID number back to our survey partner 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 profile on our partner's site, 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—adjusting both for the differences between the US population and US users from our survey partners (according to a state-by-age-gender stratification of the US population from the 2018 Census March Supplement), and for the propensity of a user to take our survey in the first place. In more detail, we receive $w_i = 1/\pi_i$, where $\pi_i$ is an estimated probability (produced by our survey partners) that an individual with the same state-by-age-gender profile as user $i$ would take our CMU survey. The adjustment we make follows a standard inverse probability weighting strategy (this being a special case of importance sampling).

**Adjusting percent ILI and CLI.** As before, in a given temporal-spatial unit (for example, daily-county), let $X_i$ and $Y_i$ denote number of ILI and CLI cases in the household, respectively (computed according to the simple strategy from Section 4), 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 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:

$$\hat{p}_w = 100 \cdot \sum_{i=1}^m w_i \frac{X_i}{N_i} \quad \text{and} \quad \hat{q}_w = 100 \cdot \sum_{i=1}^m w_i \frac{Y_i}{N_i}, \tag{7}$$

with estimated standard errors:

$$\widehat{\text{se}}(\hat{p}_w) = 100 \cdot \sqrt{\sum_{i=1}^m w_i^2 \Big(\frac{X_i}{N_i} - \hat{p}_w\Big)^2} \quad \text{and} \quad \widehat{\text{se}}(\hat{q}_w) = 100 \cdot \sqrt{\sum_{i=1}^m w_i^2 \Big(\frac{Y_i}{N_i} - \hat{q}_w\Big)^2}, \tag{8}$$

the delta method estimates of variance associated with self-normalized importance sampling estimators in (7).

**Adjusting percent "community CLI".** As before, in a given temporal-spatial unit (for example, daily-county), let $U_i$ and $V_i$ denote the indicators that the survey respondent knows someone in their community with CLI, including and not including their household, respectively, for survey $i$, out of $m$ surveys collected. Also let $w_i$ be the self-normalized weight that accompanies survey $i$, as above. Then our adjusted estimates of $a$ and $b$ are:

$$\hat{a}_w = 100 \cdot \sum_{i=1}^m w_i U_i \quad \text{and} \quad \hat{b}_w = 100 \cdot \sum_{i=1}^m w_i V_i. \tag{9}$$

with estimated standard errors:

$$\widehat{\text{se}}(\hat{a}_w) = 100 \cdot \sqrt{\sum_{i=1}^m w_i^2 (U_i - \hat{a}_w)^2} \quad \text{and} \quad \widehat{\text{se}}(\hat{b}_w) = 100 \cdot \sqrt{\sum_{i=1}^m w_i^2 (V_i - \hat{b}_w)^2}, \tag{10}$$

the delta method estimates of variance associated with self-normalized importance sampling estimators in (9).

# 8 Mega-county estimates

Daily, for each state, we take all counties that do not have have sufficient sample sizes to warrant their own estimates—meaning, they have less than 100 survey responses—and we combine all of their their data together, to compute one "mega-county" estimate. (We call this a "rest of state" estimate on our COVIDcast map tool.) Effectively, this allows us form estimates by pooling data across many of the rural counties in a state (often, these rural counties will be small in population and have insufficient sample sizes individually). Note that, through our construction here, the composition of each mega-county (in terms of the individual contributing counties) can change from day to day.

# 9 Temporal pooling

Additionally, we consider estimates formed by pooling data over time. To be specific, daily, for each location, we pool all data available in that location over the last 7 days, and recompute everything described in the last four sections: unweighted and weighted estimates for counties, mega-counties, MSAs, HRRs, and states. This significantly alleviates sparsity issues at the county and metro levels, because, after pooling, far more locations will have sufficient data—at least 100 survey responses—to warrant their own estimates. Aside from improving sparsity, it can also be seen as a simple form of temporal smoothing.

# A Appendix: Additional details

## A.1 Details behind the choice of estimator (2)

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

$$\hat{p} = \frac{1}{m}\sum_{i \in S} \frac{X_i}{N_i} \quad \text{and} \quad \hat{q} = \frac{1}{m}\sum_{i \in S} \frac{Y_i}{N_i}, \tag{11}$$

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,

$$\hat{p} = \frac{\bar{X}}{\bar{N}} \quad \text{and} \quad \hat{q} = \frac{\bar{X}}{\bar{N}} \tag{12}$$

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 (11) or (12) 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 (11); if we are sampling household uniformly, then it is (12). 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 our survey partners offer 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 "[website] adults" in a household: individuals 18 years or older, who have an account with our survey partners. Hence if we posit that the number of "[website] 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 our survey partners 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 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.