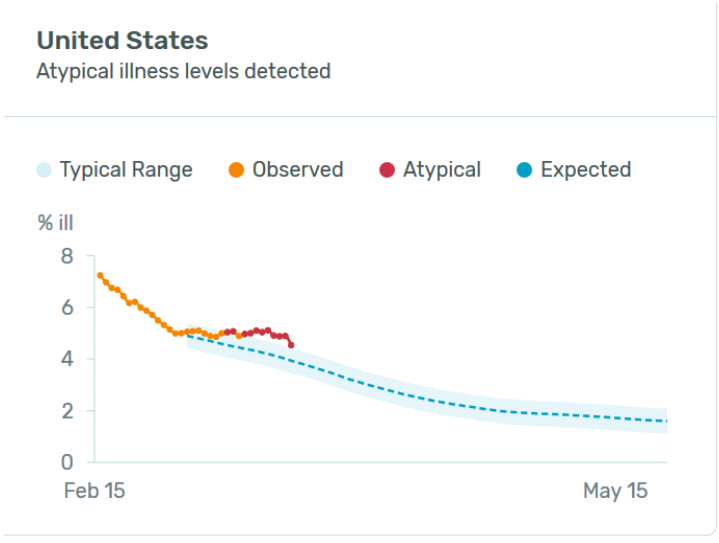
3/21/20

Kinsa is a company that makes smart thermometers. When someone takes their temperature with one of these thermometers, that data is automatically sent to Kinsa. They also collect other data, including location, age, and gender. There are about 1 million of these thermometers out there in the United States, according to this [this recent NYTimes article] which I came across this morning.

This could be really useful for tracking coronavirus. Kinsa has already been using these data to track the flu in the US. They published a [paper] about it in 2018, and say that they're often able to spot trends and new hot spots days before the CDC does. They have started applying it to coronavirus, and say that their data identified a spike in the number of people with fevers in South Florida a few days before testing found that coronavirus spike. They now have a coronavirus dashboard type website [here], with some technical background behind it [here] (including links to their flu work).

It seems like it should also be possible to use their data to estimate the total number of coronavirus cases in the US. [They] have a graph on their website which shows a spike in feverish flu-like-illness above the expected baseline in early to mid March, which they attribute to coronavirus. It seems like it should be possible to turn that graph into an estimated number of excess cases, and then into an estimate of the number of Americans infected with coronavirus, but I don't see anything from them which tries to do that estimation.



graph from https://healthweather.us/ accessed March 21, 2020

So I decided to give it a try.

My calculations (and the data they use) for estimating the number of Americans infected with coronavirus based on this graph are all in this spreadsheet, but I'll lay out my process here in a way that I hope you can follow even without looking at that spreadsheet.

First step is making sense of the scale that their graph uses. The y-axis is labeled "% ill", and after a little bit of searching it appears to be an estimate of a percentage from ILINet. The CDC has a program where various healthcare facilities tell them each week how many patients they had that week, and how many of those patients had influenza-like-illness (ILI). The CDC then reports each week what percentage of those patients had ILI, and uses that percentage as one of the main ways they track the flu situation. In their 2018 paper, the Kinsa researchers built a model to try to predict that percentage based on the data that they were getting from their thermometers (such as how many of their thermometers gave a reading above 100F). And they are now feeding the incoming data into that model, and using the output of that model as their main way to report the prevalence of feverish symptoms (which are now indicators of COVID-19 as well as influenza).

So I need to translate this ILI measure into a number of illnesses nationwide. I initially tried to search for a canonical translation from the CDC, as well as more details about how this ILI data are collected which might let me do a first-principles calculation. That turned up some useful things, like this downloadable data from the CDC of what the weekly ILI percentage has been during the 2019-20 flu season, but it was slow going. Then I noticed that the CDC also frequently reports estimates of the total number of flu cases in the US so far this flu season. I couldn't find weekly data of these estimates on the CDC site, but a google search for million flu turned up a bunch of articles with headlines like "US Flu Cases Reach 29 Million: Have We Hit Peak Season?" I found 8 of those articles from this flu season (dates Jan 27 - Mar 13), with # of flu cases increasing from 15M to 36M. I plugged their dates & # of flu cases into a spreadsheet, and found that it made something very close to a straight line with a slope of 0.44M flu cases per day. And the ILI percentage estimates during this period range from 5.0 to 6.7, with an average of 5.8, which suggests that when 5.8% of outpatients have ILI then the number of flu cases is growing at about 0.44 million per day. Which I think means that each 1 percentage point increase in ILI corresponds to about 75,000 new flu cases per day. (It's possible that this last step is wrong and I'm supposed to subtract off some baseline level of feverish illness which is not due to the seasonal flu, in which case 1 percentage point of ILI might correspond to 100,000 or 125,000 new flu cases per day.)

OK, now on to the estimate. The graph shows that, on March 14, they observed 5.10% on the ILI index when they expected 4.31%, a 0.79 percentage point gap. Since 1 percentage point corresponds to about 75,000 new flu cases in a day, the 0.79 percentage point gap suggests that there were about 60,000 new cases that day above expected. It's tempting to call this the number of new coronavirus cases in the US on March 14, which would make this calculation almost done, but I'm hesitant to jump to that label. I'll go into more detail about why in a bit. For now let's instead call them something like "newly identified feverish flu-equivalent cases." Okay, that's a mouthful, I'll go with "feverish cases."

If I want to estimate the total number of feverish cases in the US up through March 14, there are two simplish ways to get that estimate from this number of 60,000 new cases that day. Option 1 is to do the same calculation for each of the previous days and add those up (and maybe add a few more cases to account for the ones before the first day that I included). If I do that back through March 2, I get a total of about 400,000 cases (and if I add another 70,000 for the pre March 2 cases that brings it up to 470,000). Option 2 is to make an assumption about the doubling rate. Let's say the number of cases doubles every 4 days, then each day it increases by 19%. So we can assume that those 60,000 new cases on March 14 are 19% of the total cases through March 13, so there were 315,000 total cases through March 13, plus those 60,000 new ones on the 14th brings it up to 375,000 total cases. Not too far off from the 470,000 that the other estimate gives. (By the way, that 70,000 pre March 2 cases was also based on the 4 day doubling period idea - it was the total number of new cases March 2-5, which would equal the number of prior cases if the number of cases doubles every 4 days.) Average these two estimates together and we get about 420,000 total feverish cases through March 14.

Now let's start to look more closely at what these numbers mean and where there might be something funny going on that this calculation is missing. First of all, there's something pretty odd about the graph. The excess cases seem to increase roughly linearly from March 2-16, then level off, and on March 20 they dip down. That is an odd shape for the number of daily new coronavirus infections to have. If the number of infections is increasing exponentially, then the number of daily new infections should also increase exponentially, not linearly. And it should keep going up and up and up, not level off or drop. The US has taken some measures to slow the spread of coronavirus, but it doesn't seem like it's done enough to stop the daily increase in new cases or bring it down (especially not as of March 17). Either this means that there's something screwy with the data and this whole attempted estimate is doomed, or it will get clarified as I think through more considerations.

One specific issue with estimation is that baseline number. I'm not entirely clear on where it came from. It basically matches the actual data at the end of February. Their methods page says that it's based on taking their ILI estimate prior to when the excess deaths started and projecting it forward, and that the lower & upper bound of the range they give form a 95% confidence interval. So I could redo the calculations using those two bounds… when I do I get a range of 23,000-97,000 new cases on March 14 (initial estimate was 60,000), and 86,000-810,000 total cases through March 14 (initial estimate was 420,000). 95% is a pretty wide confidence interval given all the other uncertainties here, and the mean of the two edge-of-range estimates is pretty close to the initial estimate, so I'm just going to ignore this issue and assume that it's fine that use their baseline number.

Another concern about the baseline is that it's probably estimating the number of flu cases based on historical data, but there have been huge and growing behavioral changes over the past couple weeks. Other countries have found that precautions related to coronavirus have led to fewer cases of the flu, and that seems likely to happen in the US too. If they don't take this

into account in their baseline estimate, then the actual number of flu cases will dip below their baseline estimate, increasingly over time.

So it seems like there could be a sweet spot when their data is most useful for estimating the number of coronavirus cases. At first coronavirus is too rare to be very visible in the overall fever trend given all the influenza cases. Then it becomes more common and becomes visible in the overall fever trend. Then people change their behavior to avoid getting coronavirus, and also get less influenza, and it's hard to track that dip in influenza cases so it's also hard to tease the coronavirus trend apart from the overall fever trend. If things go badly, maybe coronavirus will become such a dominant part of the overall trend that the influenza trend becomes irrelevant and it becomes easier again to track the coronavirus fevers. The odd pattern of linear growth and flattening and decline, noted a few paragraphs ago, matches this model. And their map lets you see the trend around particular counties, and it looks like that pattern of flattening and decline happened at different times in different regions. King county WA (Seattle area) peaked on March 9. Santa Clara county CA (SF Bay area) peaked March 6 and is now down below the baseline - my guess is that means that the number of prevented flu cases (due to social distancing, shelter-in-place, etc.) is now larger than the number of coronavirus cases.

So to estimate the total number of feverish cases on March 20, I'm not going to repeat the calculation that I did for March 14. Instead, I'll take the numbers for March 14 (which looks on the graph like the last day before national flattening set in) and project them forward, assuming a 4 day doubling rate since then. It would be better to do a bunch of separate regional estimates with different cutoff dates (since flattening apparently did set in before March 14 around Seattle & the Bay Area), but I'm going to run with this method. It implies an estimate of 1.2 million total feverish cases through March 20.

Another behavior change that seems relevant to their method has probably also happened over the past few weeks - people are measuring their temperature more often. That might also mess with their model. A simple way to count fevers from thermometer measurements - if a thermometer measures above 100F then it found a fever that, if it measures below 100F or isn't used that day then it didn't. That leaves some undetected fevers, where a person had a temperature above 100F but didn't measure it. If people are taking their temperature more often, those undetected fevers get detected, the fever count goes up (without any change in people's temperatures). I am not sure if their (more complicated) method is subject to this problem, or if they've noticed it and done anything to try to adjust for it. It would cause the "observed" line of the graph to be too high. I don't know how big this problem might be; I'm just going to ignore it for now.

Alright, so I have an estimate of number of "feverish cases" through March 20: 1.2 million. How many coronavirus cases is that?

One issue is that not everyone who gets infected with coronavirus gets a fever. Similarly, the CDC says "It's important to note that not everyone with flu will have a fever." The "feverish

cases" estimate is only counting the cases with a fever, and then maybe extrapolating to a larger number based on what fraction of flu cases involve a fever. So I think that the relevant numbers for adjusting the estimate are 1) when the CDC estimates "US Flu Cases Reach 29 Million" or whatever, what fraction of those flu cases involve a fever?, and 2) what fraction of people who are infected with coronavirus get a fever. With some help, or more time spent doing research, I could probably get better estimates of these. For now, I'm going to guess 0.9 for flu and 0.7 for coronavirus. 0.9 because I think (based on pages like this and this) that the CDC is reporting the number of symptomatic flu cases (and excluding the asymptomatic instances where a person is infected with the virus), and it seems like those would almost all involve fever. 0.7 just feels like a plausibleish guess. That implies that the estimate for coronavirus cases should be multiplied by 0.9/0.7, which increases that 1.2 million estimate by 30% up to 1.5 million.

A related issue is timing. Feverish cases can only be observed once they become feverish, not on the day of infection. This is much less relevant for a slower-moving disease like the flu, and much more relevant for a coronavirus which is doubling every 4 days or whatever. The important question here is at what point in the course of the illness does a case get counted towards Kinsa's ILI proportion estimate. A few hypotheses about this timing: 1) A case is being counted by Kinsa on the first day that their thermometer measures a fever, since that is how they're getting their data. 2) A case is counted by the CDC on the day that the patient goes to the healthcare facility with the flu, since that is how they are getting their ILI data, and that is probably typically a couple days after the patient first measures their own temperature above 100F, so the Kinsa model's estimate of ILI is based on temperature measures from a couple days earlier. 3) A case is counted on the day the person with the flu becomes infected; the CDC has some model to estimate the number of people who have been infected with the flu as of a particular date and releases flu count numbers based on that model. I don't know which of these (if any) is true; investigating this question more seems like a highly tractable action which could influence the bottom line estimate by a fair amount. For now I'll guess the intermediate option, #1, since it yields the middle estimate and it seems like the simplest thing for Kinsa to be doing to create their chart. If it takes 5 days an average between infection and developing a fever, that means that all of these estimates are actually estimating the total number of infections as of 5 days earlier. With a doubling time of 4 days, that implies that we should multiply the estimate by more than 2. That gives us 1.5 million x 2.4 = 3.6 million coronavirus infections as of March 20.

That's my final estimate with this method, for now. 3.6 million Americans infected with coronavirus, or 1.1% of the population, as of March 20.

(As I noted before, exact calculations without rounding are in this spreadsheet. I also did a couple sensitivity analyses and found that the doubling time doesn't actually influence this estimate that much. If a use a 5-day doubling time everywhere that I used 4 days in this writeup, the estimate becomes 2.8 million. With a 3-day doubling time, it's 5.9 million.)