Data Overload: Data, Journalism, and COVID-19

April 21, 2020

So again, welcome to the webinar. My name's Todd Wallack. I'm spending the year as a Berkman Klein Nieman fellow at Harvard and I've been a data journalist at The Boston Globe for about seven years, as well as an investigative reporter working both with the Boston Globe's Spotlight Team and the rest of the newsroom.

Caroline Chen covers health care for ProPublica, and she previously was a reporter at Bloomberg News.

Armand Emamdjomeh is a graphics assignment editor for The Washington Post, and he was previously a deputy director of data visualization at The Los Angeles Times.

And I was excited to have this group of people talk about the issues that journalists face dealing with data because we all have some different expertise. I'm sort of a generalist looking at data trying to find, mine it for all sorts of types of stories. Caroline is more of a specialist in health care and will have more expertise in health care data. And Armand has lots of experience in visualizing data.

And it seems like there's been tons of interest and challenges in looking at COVID-19 data. People have been trying to track it by day and time to see trends and whether it's getting worse or better, as well as geographically. But there also have been a lot of challenges, such as obstacles trying to obtain the data. So you've seen a lot of headlines about that, particularly at a local level, or getting more details on the data, such as on race or age or other details about people affected. And questions about the accuracy and reliability of the data. ProPublica and The New York Times and others have written a lot of stories raising questions about the accuracy and the challenges comparing one area to another because of difficult variances in who's tested, how accurate the tests are, how accurate death counts are, and other issues.

So I'm going to start by asking a number of questions, and at about 12:35, will switch to questions from the audience. So feel free to start tossing in questions as we speak. And about halfway through, we'll start going to audience questions and finish at 1:00.

OK. I want to start off just by asking Armand and Caroline, what data have you seen readers most interested in?

Could I hop in here? I think early on the questions were just where is the disease spreading, right? So I think obviously, especially in the US, as the virus first started to hit, everybody just wanted to know case counts. And then I think that started to soon overlay with a concern about deaths. And so I think that continues to be of interest, cases and deaths. And then I would say where are the tests and testing capabilities, testing capacities. And I think now there's an

understanding that there are two types of tests. There's the diagnostic tests. Those are the PCR, the swab tests. And now the new incoming antibody tests.

And I think the more sophisticated readers are starting to gain an understanding of what do the numbers mean as we're starting, just like right now, this week, starting to see studies come out with some numbers around these antibody studies. And there are already furious debates around those study results and whether or not those are meaningful.

So I see those as layering, right? We continue to want to care about case counts. And we continue to want to care a lot about death and you know, segmentation of those. So demographics and race and who are being affected. And these are layering as we go.

And Armand, what have you seen?

Yeah, I completely agree, definitely along the same lines of that pattern. It has been case counts and where there were reported cases and reported outbreaks, you know, deaths. And now, I think the one thing I can add to what Caroline said, is there's been interest in the trends that are being reported by states as well.

So we've made steps to show what does this data look like over time, of course noting the caveats in the data and how it's being reported and recorded by the states.

Oh, and I forgot to mention, of course, there's the whole conversation around supplies, PPE, ventilators. It's obviously been very hard. That's always a moving target depending on whether you're talking a local level, national level. You can never put a nail on how much PPE there is at a given hospital, state, but there is always interest in that question of supplies.

And I'm also curious about how easy or difficult has it been to obtain all the data for your stories and graphics?

I can talk about that. I think, you know, the data at a national level is basically nonexistent. Most things are reported at the state level, so that makes means know you have to either rely on an aggregator or aggregate the data yourself, you know, going to all these different state sites, figuring out where they report it, how they report it, in what format. Also noting that this data, what the states are reporting, also changes over time and what platforms they're using to report it.

I sent out a tweet a few days ago that was like, what if you were reporting a live election, but you were building your rig for reporting the results as the election was happening? And what everywhere was reporting as the results was also changing, and they were changing how they report it as well. So it has been extraordinarily difficult in that sense, to build things that don't constantly break, and build data flows that are actually kind of stable, given the fact that what is being reported is moving under them as well.

Yeah, I would say that there are certain things that, just by the nature of the pandemic, are going to be constantly changing. So for example, testing capacity. I've done a lot of reporting around testing and testing capacity. And just by the nature of what's happening, that is changing

constantly. So whether that's nationally, whether that's locally, if you're trying to say what is the testing capacity of my state, that number is going to be constantly changing. And it should be, right? Because we have been constantly ramping up testing capacity.

So for any reporter to try to get a beat on that, on trying to inform their local readers, they'd have to constantly update that. Is it possible to actually get an accurate number at any point in time? I think that is technically possible, but your number's going to be outdated like within an hour, even at a specific lab.

So I have, at certain points, been able to be like, I nailed the number. It's already old. Is there any point in even doing that? Yeah, I think there is. It's a worthy exercise to try to get a ballpark and to track trends for readers. So there have been times where I've tried to do that for specific stories, but it is a frustrating exercise, and I've really encouraged other reporters to really try to explain to readers where you got this number, what has gone into this number, and how long of a shelf life the number will have, and really try to show your work to your readers more than I normally would. So I think some of that is inherent.

There are other things, though, where you can only-- your information is only as good as where you get it from. So for example, the WHO puts out daily situation reports. That is the only way you can really get a source for international case counts, right? But the WHO's information is only as good as the countries from which it comes from. So I keep repeatedly explaining this to people, that the WHO has a recommended way for what they count as a positive case. And they say that it is if you test positive with a PCR based test.

For the longest time, China just decided that they were only going to count as positive someone who had a positive PCR test and symptoms. They were not counting people who had a positive PCR case but no symptoms. So they weren't counting asymptomatic cases. There's nothing the WHO can do about that. There's nothing anybody can do about that. And then after a while, trying to change that. So you needed to know that about China. And I mean, that's deeply frustrating. You can't get everybody to report the same way. And you need to have those caveats in your reporting.

And this also trickles down to like 50 states, or 56 states and territories, all doing it in their same way as well.

Right.

That's got to make comparisons really tricky, when everyone has a different way of reporting the data, tracking the data. There are different rules on who gets tested and what gets counted.

It sure does. And again, I think you can only be clear about the caveats of this is entirely dependent on what's being reported and how it's being reported.

Yeah, I've been very, very cautious about comparisons.

Got it. And are there any other problems you've noted in the data that people should be aware of?

I have been very careful-- or I've been encouraging reporters in my newsroom and trying to explain to the public just to be aware of what the definitions are of numbers that get thrown around. So one thing, for example, I've been trying to explain a lot to lay readers is what actually is the fatality rate, right?

And there's a big gap, I think, between what the public wants to know, which is, you know, if I get infected will I die, and what is reported as the case fatality rate, right?

So the case fatality rate is the number of reported deaths divided by the number of lab confirmed infections. So everybody knows in the US, it had been really hard-- it continues to be really hard in many places-- to get tested in the first place. And a lot of places are not testing unless you're really, really sick.

So that denominator is going to be much smaller than the actual number of infections. So especially early on in the United States, the case fatality rate was something like 10%. Because we just weren't testing a lot of people. And you think of it as an iceberg model. Like, the deaths are usually the easiest to find and count, especially early on in a pandemic. This always happens in a pandemic. And the people who are asymptomatically infected are the hardest to find in the first place.

But again, the average lay reader, they just want to know if I get infected, will I die? And they're looking at that number that's been reported in your headline, and they're just looking at that and being like, if I get infected, that's my chance of dying. And we have such a huge responsibility as reporters to explain that number and not just throw things around in headlines.

So I think there are a lot of numbers like this. As a science and health reporter, I feel like we have a lot of responsibility to explain to people, so that are not and are. These are the rate of infection, the chance you have of infecting other people, the average number of people you'll infect. It's a process of understanding, and this is what I'm trying to get across to my readers. There is so much we are still learning that we don't know yet, and we cannot present this as set in stone.

And building on that, mentioning all cases being kind of a difficult fraction to divide against. Like, the deaths number also is slippery. We've seen stories highlighting this in recent days, and it's been something we've been kind of saying for a while. It's like not every death is being accurately categorized too. Recently New York City added some-- what was it? Like 3,700? I forget the exact number of deaths that were classified as probably COVID-19. And you know, if it's happening in New York City, probably there is some fraction of cases that's being categorized throughout or never even recorded. So that number is slippery as well.

And I think when talking about fatality rates and that kind of thing, rather than just talking about one big number, we've been trying to, when we have the data available, to at least break it down a little bit better into segments of the population or report the comorbidities that studies have been reporting. So it's not like a flat 3.2% or whatever it would be. It depends on a lot of factors that are related to the individual.

It definitely sounds challenging when there are questions about and uncertainty about both the numerator and the denominator when you're trying to calculate rates. My sense is that these are problems that data journalists, and journalists in general, encounter when trying to get data. It's often hard to get one clean database at a national level or global level, where often aggregating it from lots of different places. And each place might have different ways of counting the numbers and reporting the numbers. And the data can be messy. Is there anything different that you're finding in dealing with COVID-19 data? Or does it reflect challenges you've faced doing other types of stories?

This is more philosophical. But you know, these numbers are being reported by states and by countries and everywhere very precisely, but in its nature, it's a very imprecise count. So there is this weird situation.

You know, inaccurate but precise is one type of data classification. And I think that's where we are now. It's like you're throwing-- you're taking shots at a dartboard, and they're all landing in a very exact, same similar place, but you're off somewhere. You're not actually hitting the dartboard. It's like somewhere off of the wall because you're throwing the darts kind of blindfolded. But we have very precise counts.

Yeah. One thing I've seen, and I guess my-- I know this is a really hard thing to do, especially if you have an editor that's pushing you, is to resist the urge to write. Because what I do see is that health departments, as they release data, are refining as they go. And I think this is because they are also figuring out what they need to release.

So for example, to give a very specific example, New York City started out by giving test-- they were only reporting by burrough. And then a lot of people were like, well, that's not enough information. And they were getting a lot of criticism. And then they started releasing-- it wasn't quite by neighborhood. It was by this very strange not quite zip code, not quite neighborhood.

It was percentage of positive, but they didn't have any raw numbers. There were no numerator, no denominator. It was percentage. And I was like, well, I can't do anything with that. Because if you say that in this zone, it was 66% positive, that could mean that you only did three tests there, and two people tested positive. That's meaningless. But I did see some news organizations write a story on that. And I was like, that's a bit dangerous.

And then, I think like within a week, they then re-released numbers, which were by zip code and had numerator and denominator. They had way more information. And then you could write a more meaningful story. And then New York City has continued to update and iterate and give more and more granular information.

So I do think that there is a benefit to kind of waiting. Because I've seen, more than I've ever seen before in any other outbreak I've covered, sort of health departments iterate as they go with the data that they're releasing. And I actually see, because this is happening across the country, actually reporters, I think, be able to push health departments and be able to say, hey, you know, Ohio released this information. Florida, why aren't you releasing this information? And be able to sort of push departments off of each other.

And I think in a similar theme, I think it's really sometimes dangerous to write a story off of a preprint. I do think it's really great that scientists are, researchers are moving quickly and sharing information on MetaArchive and bioRxiv and not waiting to go through that whole process. But then it's not peer reviewed, right? So this puts you in a really dangerous position as a reporter to have to write a story off of a non-peer reviewed study.

So I think one of my goals is to never let a preprint walk alone, as in you don't write a story on a preprint by itself. You try to let it go in concert with other studies and look for a trend. Or at least let lots and lots and lots of people comment on it, and don't just write a story on this.

So this is happening right now with all these antibody studies, right? Like, Stanford put out its preprint on its antibody serosurvey. And there were a lot of stories that got written really quickly. And then in the next day, there has been the critique wave of like, was it a good survey? Was it biased? You know, all of that stuff. And I just wish that a lot of reporters might have waited a little bit.

And now there is the Los Angeles serosurvey. And I think you could have maybe waited and collected a bunch of these studies and maybe done one thoughtful story in one go, or at least gotten a lot more outside voices than you normally would before writing that one story. Because they aren't peer reviewed. So you do have to treat preprints differently.

Right. And interestingly, of course, none of our articles are peer reviewed. So I'm curious what process you go through to make sure that your own interpretations and analysis are sound before publishing.

I just run preprints by way, way more people than I normally would. If something's already published in a journal, I know that it's gone through that peer review process. If it hasn't been, I will run it by a lot more outside experts than I normally would and just go that extra mile and really ask myself, do I have to write this now? Can I wait for it to go through that peer review process?

And you can ask the author. Sometimes they'll say, oh yeah, this has already been accepted by JAMA or The Lancet or whatever. And that gives me an extra measure of confidence. If that's the case, that's helpful to know. And if not and it's like, this is such an important study that I need to write about it right now, then I get all those outside voices. I try to get many independent outside voices that are from a number of different institutions, get all their critiques. And if all of them are really, really negative, then again, I have to ask, why am I writing about this study in the first place? The bar just gets so much higher if it's not in a journal and hasn't gone through the peer review process.

And I assume, Caroline, even when ProPublica or The Post or others are doing their own analysis, we do the same thing. We go to outside experts and say, here are the numbers I'm calculating. Does my methodology make sense? Is there a good explanation for these conclusions? Rather than just posting something on Twitter or throwing it on our website, we first normally talk to experts first.

Yeah, exactly. And you know, there's a bit of self analysis in here too. Like looking at what we call data smells. Does what's in the data question your basic assumptions? Does it show an opposite trend to what you're expecting? Are there massive gaps or negative values where there shouldn't be? It's kind of like sanity checking the data as well.

And similarly, I know there are questions about different models that organizations are using. A lot of people are looking at the University of Washington model. It has a website that's very easy to use, predicting when peaks are going to be for hospitalizations and other issues. But there are lots of other models it seems, and there are questions about what variables go into each model, how the numbers are calculated. And they can produce conflicting results. So that has to be challenging to deal with.

Yeah, so I did a whole column on forecasting and predictions earlier on, which was partly for reporters and partly for the public. And I think, again, the question really is, who is your audience? And who are you writing for? And I try to keep that at the back of my mind. Because I think there's a difference here.

If you are writing for really a lay public, again, you have to remind them, is this an estimation? And I was talking to, for that particular column, I was talking to an epidemiologist. And I said, you know, I was reading this sentence that somebody had written about their particular model. And I said, it seemed awfully specific, where they said that this means that in New York-- this was back in early March-- that last week there were, it was something like 1,583 to 2,000, blah blah blah. It was like down to the digit number of people infected.

And I was like, I read that sentence and I feel like it gives a lay audience this sense that you can be that precise and calculate down to a single digit how many people are infected. And for me, as a writer, I would never give that level of precision. Because it signals something to a reader. I would round and use the words "around."

And I said, what does this say to you as an epidemiologist? And it was really interesting because she said, I like seeing that sort of precision. Because from one epidemiologist to another, I can then go and redo his model and make sure that our numbers match exactly. So it's very useful from one researcher to another.

But I agree with you, for a lay audience, that's not the message we want to send. Because I said, what is the takeaway you would want for a lay audience? She said the takeaway I would want a lay audience to hear is it's not 400 and it's not a million. You're in the low thousands.

So really, that's kind of the question that I always-- when I'm talking to someone who's doing modeling, I say, what is the takeaway you would want for a lay audience? And really, she said with models, you need to be thinking in orders of magnitude. And I think that our responsibility as reporters is to then say, OK, so I'm going to give an orders of magnitude type of number to my readers.

Got it. And I'm also curious, are there any mistakes that you see lots of people repeatedly making that bug you? One I see all the time is people say, oh, there have been four million people tested,

as if there's been four million tests. But some of the tests require multiple samples. People could have been tested multiple times. So there are different numbers.

I also see people say, oh, there are this many cases when it's number of confirmed cases. And there are other studies showing there are probably many times more people who've been infected but haven't been tested.

Yeah, the one that you just mentioned, Todd, I think is the one that I've seen most often just in talking with people and hearing that like, oh, this place has only five cases. It's like, well, no. I mean, yes, but no. That is just being what's reported and what's being conveyed, being reported by the states. And again, that comes back to what Caroline was just saying about this precision, implying that we know there are 526 cases in this county in Illinois or something.

But maybe that's on us too. I know the instinct is to try and report the data to the granularity we have available. But maybe there are better ways in that we do report the data that implies more of this imprecision about the data. And that's something I think we can ask ourselves and address as we try to put together these pages that are tracking the spread of the disease or whatever.

Another one is just people being exposed to types of scales and visuals that they're not used to seeing. So like, we're seeing a lot more logarithmic scales than we're used to. And they don't chart things-- you know, growth doesn't look the same way on a log scale than a linear scale. But if you're looking at it and think you're on a linear scale, then you might think things are declining or flattening out when actually, that is very much not the case.

Yeah, I think, Todd, you picked up on my biggest pet peeve, which is people not paying attention to units, right? And I've kind of been-- this has been my soapbox rant for the longest time. It's like please, try to get your units in people.

Because I think, again, that is what readers care about, right? They see a million tests, and they think that that is a million people. When you say, we're rolling out a million tests, they will automatically think that is a million people who can get tested. And depending on the type of testing, this is absolutely confusing. Like the CDC test, you had to divide by two. The Abbott test, the rapid test, it is one test per person.

So depending on which test you are doing, it is a different equation. And it really is a reporter's responsibility to figure out what the heck is being said. And it is a way for, frankly, for officials to inflate numbers. And it's the only way to really get an apples to apples comparison, is if we get a testing capacity in people. And so I think that's a journalist's responsibility, to always get the units in people. That way, we can compare state by state, country by country.

So I do think that that is a mistake-- well, a mistake, or I think a confusion-- that annoys me when I see that. And yeah, I think just not explaining that everything should be like, this is a reported number of deaths or reported number of cases at this point in time, as Armand said. I think those are really common.

I think also just-- this is more philosophical-- is just presuming we know things. I mostly see this frankly on Twitter and on TV, but just this air of we know what to do. Like, if this state just did this, then we would solve the crisis. No. Nobody knows what to do. We have only known this, like humanity has only known this virus since January. Well, I mean, in China it was a little bit earlier than that. But in the US, we haven't really known it that long.

And every time I dig into this, whether it is on really understanding how it is transmitted, or I recently was doing a lot of reporting on doctors struggling to understand how best to use ventilators, how to best treat critically ill patients. Everybody is struggling to do their best by patients and to really understand what to do.

And so I think there are no easy answers in this crisis. And I think you can give-- I think this is a failure of communication both by our officials and also actually, by journalists, when we make it sound like there is an obvious or easy answer, and failing to acknowledge that, to some degree, we are all still learning. And so that irks me, whenever it comes across as like, well, obviously.

That sounds good. Why don't we go to-- questions from the audience are starting to pile up. One that's been upvoted the most is from a Berkman fellow, BaoBao Zhang, who wondered how you feel about non-experts weighing in with their own analysis on Medium or Twitter or elsewhere and non-journalists. And not all of those people do what journalists do, which is going to experts first to vet their conclusions.

I definitely feel like, you know, it's a free society. And that's what platforms like Medium exist for. So you're specifically citing Tomas Pueyo's "The Hammer and the Dance." I think it's fine if people want to publish, and I think that they definitely find their own audiences.

I do think that things like that sometimes are-- I think they find their own audiences, basically. I had a lot of people actually send me that specific post and be like, I cannot understand this. Can you write a version that handholds a little bit more?

Because I think the part where oftentimes, experts who are experts in their field, whether they're data scientists or-- I see this a lot, where like, a clinician or somebody will be writing. They tend to use a lot of jargon and don't break it down to the degree that I tend to try to do. And some people do it. Some people are fantastic communicators naturally. But I think that's a tendency. I tend to see a lot of jargon.

And so I think there is a place for them, and then I think sometimes, the shortcoming is that I think they're not trained to be able to use the language that helps them reach as many people as they could and to give as much context as, I think, a journalist would know how to do. That's my off the top of my head answer.

That's good. There's also a question about how do we deal with issues where we publish an article based on data, and then the data changes, or the information changes. This probably comes up all the time with health care studies, Caroline, where a new study comes that contradicts a past study. Or a study's been retracted. So how do we deal with this one? People are still passing around the old article or chart based on old, outdated data and information.

Yeah, oh. What a nightmare it is right now with the situation. So one thing that I am doing now, even more so than I normally do, is I am aggressively dating my information.

In my stories, on my sentence, I'll say like, this fact, as of Wednesday afternoon, according to the Association of Public Health Labs, the US had a testing capacity of a million tests as of Wednesday afternoon. Because literally by Thursday morning, the number is going to change.

So I try to tag as much of my information as possible. I'm linking a lot more aggressively than I normally do and also adding the date and time stamp. So whenever whoever comes along to my articles sees that information, they will know as of when that information was true.

So unfortunately, some people are not going to read that carefully. But at least the time stamp will be next to that. So I cannot go back and update my articles constantly. But at least the information that somebody reads will have a time stamp next to that. So I think that's probably the best thing you can do. And then yes, update as you go.

And I think, again, this is where the language that you use at the time you write also helps you write. Because I also say use language like, at this time, scientists understand this to be x. So when I was working on a column about asymptomatic transmission, there was a lot of language I had in there which was like, as of now, scientists understand that whatever.

So again, there's a date at the top of my article. I'm using a lot of language that indicates I'm giving you the best of understanding at this time. And then I'm also linking to studies and putting language in that's like, as of this interview that I did on this date, this is what I was told. So I think all of that in combination, hopefully, even if a reader comes along later, will know that that was information that was current at the time that I wrote that article in. And I think that's the best you can do.

Yeah. And from a data standpoint, we can either build our pages and apps to plug into live data that updates, so that you are seeing updated data as of the times stamp at the top of the page or right on the chart or whatever. Again, we try to be transparent about when that data is updating.

Or like Caroline says, we can build it statically, with like Illustrator, or just save it as the static SVG and have to make very clear that this is data as of x. Otherwise, we've been in situations when we're trying to publish a story, and we just have to keep updating the charts like five times because the data keeps changing as we're writing the story.

Yeah, and other more subtle things. So ProPublica normally does really long, sort of deep dive investigations. And actually, our social folks are used to just retweeting our stories forever. Because we often are doing such long, retrospective investigations that you could retweet our stories like two years from now, and there's no reason why somebody couldn't read them again later. And we've completely reconfigured that. So they no longer will retweet a story because they know the information could be totally old. So even thinking about that, like your social strategy. They will check in and be like, can I still tweet this story from our main account? Is that information still new? Like, thinking about that kind of thing.

And then obviously, if there's some really major new information. Like for example, if I had written a whole column on asymptomatic transmission, and there's some really major information that's really relevant to know, I will put an update at the top of that story. So being selective.

Both good points. Next question is from Eva Wolfangel, who's a Knight Science Journalism fellow, who asks about the fact that researchers often try to communicate uncertainty, and I guess there are two challenges journalists face. One is how to communicate that same uncertainty. And then there's also the question, do we undercut our own stories and reporting and data when we communicate that uncertainty? Or are people just going to say, oh, it's an estimate. It has such a wide range. It has a margin of error. You can't really rely on it. So how do you deal with those challenges?

I mean, I try to convey that in describing the process of science, right? So just to give a very specific example, in the column I was working on on asymptomatic transmission, there was a part where I talked about how new studies have shown that viral load is actually higher at the start of the disease, course of disease for COVID-19, which means that you could be more contagious even before symptoms started.

But I went out of my way to explain how this is unexpected because for COVID-19's close cousins, the coronavirus cousins SARS and MERS, you were most contagious, you are the highest viral load, in the middle of the course of disease, when your symptoms were highest.

And I think just explaining that, which would be why your natural presumption would be-- the original presumption was that COVID-19 would behave the same way. It's something that any reader can understand, that naturally, you'd look at historic models, and you'd expect it to behave the same way.

And I think trying to explain the process of science helps, and I feel like I just over explain. And I think showing that uncertainty, or even just saying things like-- I just was working on a story about ventilator use, and I had a clinician give me a number. And then he called me back and was like, you know that number I gave you? I know you hate this, journalists hate this, but it might change. And I was like, no, no, no. That's fine. That's fine, and I appreciate that you wanted to clarify that.

So then I just added a line, a very short line, saying he added that it's early days, and more information will be gathered. And I think that's fine and a good indicator to readers that more studies are going to happen.

So I definitely feel like there are ways for writers to indicate that for their readers.

And from a visual standpoint, in terms of how to communicate uncertainty, look to the annual discussion every hurricane season about how to chart the likely path of a hurricane. It's like visuals. You want to give somebody something to look at that tries to convey the data as best as possible.

And I think in the case of this outbreak, the best we can do is work that into the chatter and the headline around the chart, the annotations. Say that it's reported cases or confirmed cases or reported deaths. Try and convey the uncertainty in what's around the chart, rather than the numbers, which are what's actually being reported and what we actually have to chart.

Got it. And I want to take on a question by Saul Tannenbaum, who asks about questions raised by COVID skeptics, who will often point out when we report deaths, argue that they're over counted or there are no COVID deaths, in extreme cases, and say, well, they're really dying from a heart attack. Or they're really dying from pneumonia. Or they're dying from some other cause. And yes, they tested positive for COVID, but that wasn't necessarily what caused their death. How do you deal with those types of questions?

That's interesting. I think that-- I don't know that that's a useful debate right now, right? I think all you can do-- because I think you can have that debate at either end. Because then you get into the debates on the people who are dying at home. Did they die of COVID? Did they die of not COVID? How do you then count the people who are the excess impact from COVID because they died at home because they didn't want to go in for help.

You know, I think there are so many swirling questions around deaths related to COVID that are going to be so hard to untangle. And I think as journalists, the only thing you can really do is just be really straight and really flat and be like, here are the number of people who died with a positive COVID test. And just leave it at that.

And then here are the number of people who died at home, and here is how it compares to the number of people who died at home last year at the same time. And show that gap if you are able to get that number from your state.

I just don't know that those debates are really helpful or getting into those weeds and trying to parse that is going to get anywhere at this point. Because you can have that same debate about the flu. Like, so and so had a positive flu test, but did they die of their underlying condition? Their pneumonia came from the flu, but they also had diabetes. What does that mean? I just don't know where you'll get with that.

And it reminds me of after Hurricane Maria, when they went and did studies of what did the excess mortality rate look at in Puerto Rico after Hurricane Maria?

I worked on the homicide report at The LA Times for several years. And the LA county coroner, if somebody was shot and then died like, say, 10 years later of complications from that gunshot wound, like eventual health impacts, it's still ruled a homicide because they died because of complications from that gunshot wound. So this is not just solely restricted to a COVID 19 debate. It's just mortality statistics in general.

Right. Another question that came up is, what is the most reliable source for COVID-19 data? I think there are at least a half a dozen sources of aggregated national data and a couple sources with global data.

Armand, go ahead.

Yeah, most reliable is the key. I mean, Johns Hopkins has really been putting tons of work into aggregating as much data as possible. You can take a scroll through their issues list on GitHub just to kind of get an idea of the volume of requests this has generated.

Of course, the World Health Organization and then I think a number of media organizations, including us, are also trying to aggregate at the US level, like state data and county data. I can't tell you which is the most accurate.

I would say just for US case counts, we mainly use Johns Hopkins data. We long ago gave up on the CDC, which is very unfortunate to have to say that. But they don't update on weekends, and they are like 24 hours behind on their weekday updates. So we use Johns Hopkins for just our daily case counts.

In terms of testing capacity, we mostly point to The COVID Tracking Project.

International, sometimes, depending on what it is, WHO or JHU. But again, it really depends on exactly what you're trying to get at.

Got it. And of course, sometimes for very local stories, there may be only one possible source of data coming from a county or coming from a hospital or somewhere.

And actually, going to your local health department directly is probably going to be the most up to date information, which will be even faster and more up to date than going to a site like JHU, frankly.

Got it. I have a question from Carolyn Schmidt. When you segment data and try to report it accurately, how do you also make sure it's accessible to a mass readership? And I personally, not being a health reporter, have found it difficult just dealing with all the acronyms and different terms that have come up in this area.

I am not entirely sure I understand the question. Is this in terms of demographic or age breakdowns or types of data that we're looking at?

Maybe Carolyn can clarify, but certainly I know-- there's so many different angles and different publications of different audiences. So sometimes people write differently based on who their audiences is and try to keep that in mind.

Yeah Carolyn, please feel free to rewrite your question. I'm also not 100% sure what you mean by segment the data. But I'd be happy to try to answer that if you want to try to clarify.

There Is also a question from [? Magna ?] Cheney, whose a Nieman affiliate, asking about the best practice for archiving stories. So Caroline, you mentioned having a date stamp can be one way.

Yeah. The Guardian does this thing where they have a warning up really high, where they say like, warning, this story is like more than a year old. Or they have some sort of very visible warning up high, which I always appreciate whenever I see that. So that could be one way to do it.

OK. Michelle O'Neill asks, what can we report that's meaningful without having the basic data that we want? And I guess that comes up with when we want to say, you know, what states? Where are the hot spots? Or how is the US doing versus other countries when there are all these questions that we've brought up about how many people are actually infected given the differences in testing, and how many people have died given differences and counting. And because of all these uncertainties with numbers, it must be really challenging to figure out what we can really say with confidence.

Yeah, agreed. I think we need to make basic assumptions or adjustments when we can, you know, for denominators that don't exist or for other things that we don't consider reliable. Like on our pages that show the data that we know about cases and deaths throughout the country, instead of normalizing, we're looking at like known cases per population of the state, per population of the area. But again, we have to be clear that this is all just based on what's being reported.

OK. There was also a question with, what do you do when you don't have data or information, or you have conflicting data, so you can't be sure? Do you just avoid writing about it? Or do you write about it as best you can? It's certainly difficult to make charts when you don't have data.

Well, I do think that there is value to writing about the lack of information, especially when you feel like, say, your local health department is withholding information that should be public, right? So I think that there was, early on, a lot of good journalism being done about the need for demographic information, about who was being infected, right? And now we're getting a lot more of that information, which is pointing out big problems in who is being infected.

So I would not dismiss that as a possibility for where you start reporting on just the lack of good information, which can actually spur change and get you the information that you then want.

Great. So I think we just have time for one last question. Gina Pavone notes that there's talk of using an app for contact tracing. There's a project at MIT for that. And there's also been stories based on cell phone data that's been released. And Gina wonders how journalists deal with aspects of privacy or reporting on the challenges of releasing that data and using that type of sensitive data and making the topic accessible to a mass audience?

Yeah, I mean, I think this is a hot topic across a lot of different countries, a lot of different localities. And I think one, really understanding the nitty gritty of how it's going to be used is important. I think there are a lot of think pieces about these apps right now that I'm seeing, which ask a lot of philosophical and hypothetical questions. But I see way fewer stories that actually get into the innards of how they're going to be used, which would actually help answer some of these think pieces.

So I think that would be useful journalism to be done. It's much easier to be like, well, what about privacy? But then you don't actually know what's going to happen.

I think the other question though, which was raised with me, was some public health experts that asked, how is this going to actually intersect with the existing public health infrastructure? Because if there are a bunch of people who have downloaded an app and it's not talking to public health officials and not helping them do their actual work, that's also useless. So this needs to fit into the existing public health ecosystem.

So I think there's a lot of good reporting that can be done around that. And then again, this needs to fit into existing testing capacity. There's no point in having a great contact tracing app if you then can't test people and find out who is sick in the first place. So there's a lot of questions about do we have a very shiny looking object that doesn't mesh with the actual realities of needs? And I think helping people understand, your readers, actually understand how this app needs to fit into the actual workflow of continuing the virus. All of those things can be helpful to your readers.

And then ultimately, also just like the mathematics of how many people would actually need to download the app for this to be useful in achieving what it needs to do. Because there is a minimum number of people who need to have the app for a contact tracing app to work.

Got it. Thank you so much. Thanks for the panelists. And for everyone who tuned in, there will be a recording available in a couple days on the Berkman Klein Center event page. And there's also going to be a quick poll survey at the very end.

So thanks again for Armand and Caroline, and thanks for everybody who is watching.