Dispatches from the Data Jungle of COVID-19

After 90 straight days of leading a team to untangle webs of pandemic data in all his waking hours, William Duck finally teared up – but not from his own exhaustion.

On August 21, his last day in CDC's COVID-19 response, Wil was saying goodbye to his IT team, who had volunteered to build from scratch the system that gleans numbers of COVID-19 cases, hospitalizations, and deaths. The hardships they had endured in the struggle to put the system together made Wil's emotions overflow.

Wil had joined the response on May 18, when CDC's COVID-19 case data system was new and still fragile. It had begun, with sputters, to produce data for CDC's internal use in April, but the crew had to work around the clock to process the information while patching barrages of system glitches with new programming code.

At the same time, CDC scientists urgently needed a robust stream of daily data to analyze how COVID-19 was spreading and who was most vulnerable to dying from it. States needed data to help them decide on measures to slow COVID-19's spread, and news outlets pounded the table for CDC to publish tolls of people sick or dead in the US from COVID-19.

"If data didn't make it out on a given day, we heard about it real quick," says Wil, who was deputy of data science and data management for the Case-Based Surveillance Section of the COVID-19 Data Analytics and Modeling Task Force.

The data originated from all 50 states and all tribal regions. Between them, they had roughly 50 different methods of collecting and formatting data, 50 legal policies on data, and 50 levels of detail on patients' age, race, and ethnicity, if they reported these details at all. Many states lacked ample staffing and IT infrastructure needed to gather and report data.

One of the country's largest states had just one employee to coordinate the submission of COVID-19 case data from the state's regions and then to send it to CDC. Wil had walked into a data jungle with the job of farming it into rows of crops in just weeks.

"The data landscape was never consistent. One state managed COVID-19 data on four IT systems that did not talk to each other. Another state still used paper forms and had an incredible backlog to feed into their system. States are also not mandated to give us their data; they share it voluntarily," Wil says.

When Wil joined the team, his colleague Kasey Diebold was devising an overarching coding concept to make the COVID-19 data from the 50-plus sources align. Wil was a good fit to hand off to because he had been fighting the same battle with public health data for decades.

Older state health data systems that don't align are the norm in the United States and urgently need modernizing into systems that work together. The COVID-19 crisis simply exacerbated the weakness, says Wil, who normally works as a health scientist in CDC's Center for Surveillance, Epidemiology, and Laboratory Services (CSELS).

Based on Kasey's concept, Wil's data management team continued ironing out misaligned data while building an algorithm that would take over most of that task.

"It took six intensive weeks," Wil says.

While they coded, snags kept coming. Some states, without notifying CDC, reformatted their state identification codes, triggering avalanches of duplicate COVID-19 cases in CDC's system.

"Suddenly, data submitted from one state showed 1,800 infants had died. We got with our task force's state coordination team and with the state and figured out that the birth dates were off by a century. Those patients were born in 1919 and 1920, not 2019 and 2020. It took three team members weeks to fix the problem," Wil says.

In June, a new problem threatened to derail the project. During Wil's stint, about 100 volunteer coders cycled through his response team, all working night and day. Overtime added up to person-weeks, and the pool of coders ran low while work piled higher than ever before.

"We were forced to automate manual data tasks, or we weren't going to be able to handle the workload," Wil says.

Wil helped establish a new team to automate the pipeline and continue refining it, and just after it began work, August 21 rolled around along with Wil's heartfelt goodbye via webcam.

The next day was his birthday. Wil woke up that morning, looked at the clock, then fell straight back to sleep.

https://www.cdc.gov/coronavirus/2019-ncov/communication/responder-stories/data-jungle.html