Improving Knowledge Through Next Generation Data Linkages

Cancer research is awash with databases that capture widely different aspects of the disease, from tumor samples and genomic sequencing to clinical study results, sociodemographics, and billing. This information typically gets warehoused in unconnected data sets, investigated by different types of researchers, and published in journals specific to their focus of research. Michaela Dinan, PhD, Associate Professor of Epidemiology (Chronic Diseases) and Co-Leader of the Cancer Prevention and Control Research Program at Yale Cancer Center, sees that as a lost opportunity. Her research demonstrates that when disparate data sets are pushed into conversation with each other, they can disclose new insights about cancer, cancer care, and the healthcare system.

"If you can think of novel ways to use data that have been around a long time," said Dr. Dinan, "you can make real contributions to the field."

Her most recent contribution was published in *JAMA Network Open* in October 2021. The paper describes a pilot study that investigated how breast cancer screening impacts clinical, genomic, and sociodemographic factors associated with newly diagnosed breast cancer. Dr. Dinan and colleagues did this by combining and cross-analyzing information from separate databases to create a first-in-kind linkage of genomic data with real-world, population-level diagnoses of breast cancer.

The National Cancer Institute (NCI) hosts the Surveillance, Epidemiology, and End Results (SEER)-Medicare Program and collects the information through the linkage of two distinct data sets. The SEER dataset provides cancer incidence and survival, including detailed information such as each tumor's stage of diagnosis and histology. The database also includes general socio-economic information such as income and education levels based on zip codes. The NCI then links data from Medicare claims, which include the medical care that an individual has had over time, such as cancer tests and treatments. All patients have their identity protected by extensive checks and balances to ensure that no individual patient can ever be identified from the research. The novelty of the project is that Dr. Dinan and her colleagues combined this SEER-Medicare data with physical tumor specimens from the SEER-Residual Tissue Repositories (RTRs) and then conducted gene expression analysis.

"This was the first study to link physical tumor samples for these patients in this SEER database to Medicare claims data, and to create one novel data set," said Dr. Dinan. "If you only look at one data set, you're not getting the whole picture."

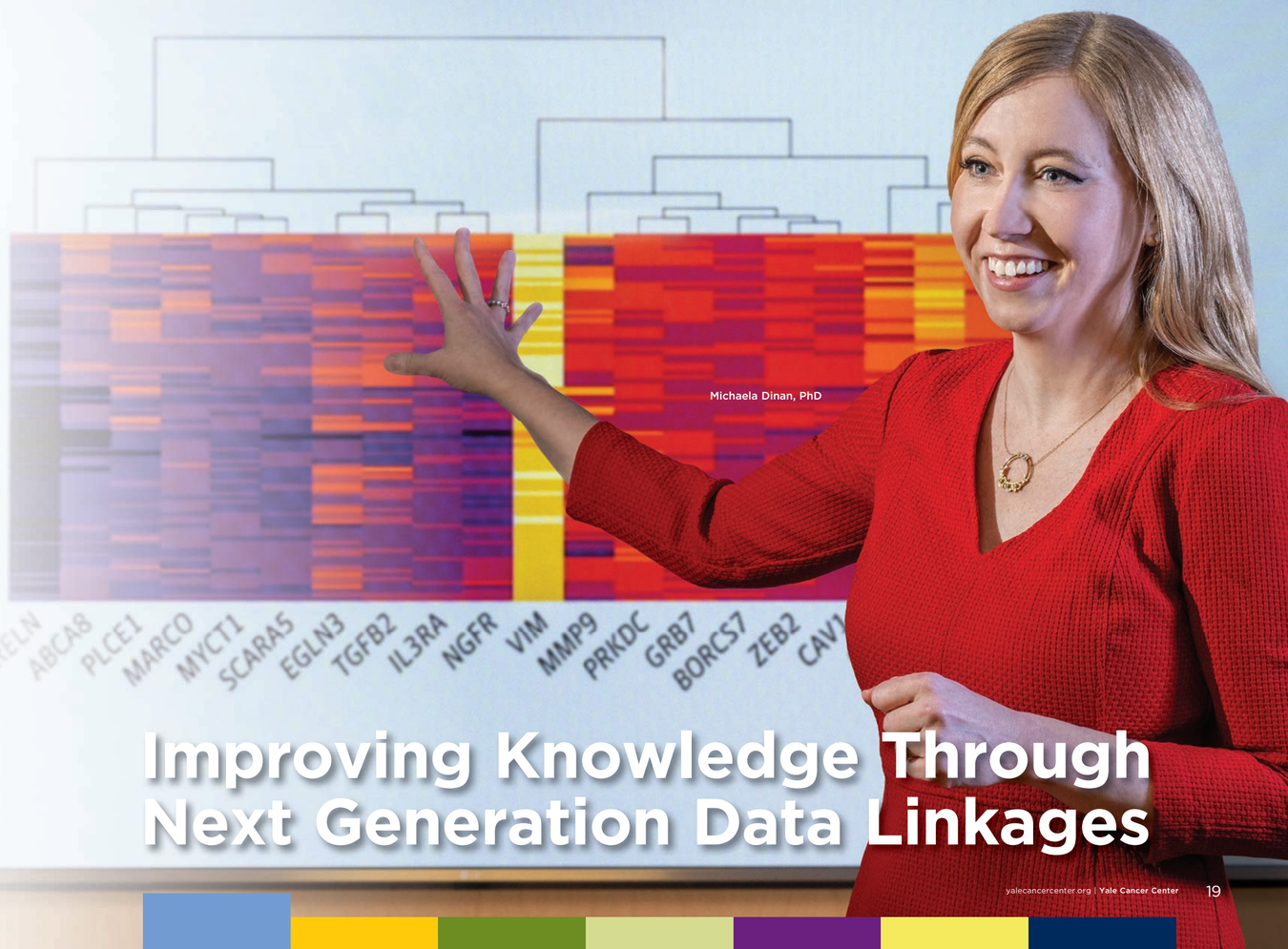
The combined data revealed that socioeconomic status and access to screening remained associated with mortality among patients with breast cancer. "That's probably our number one finding," said Dr. Dinan. "Our research suggests that living in resource-poor neighborhoods with less access to care may be important as well."

To link and cross-analyze databases might seem obvious in retrospect, but Dr. Dinan understands part of the reason it had not been done before. To collect and interpret the merged data took almost a decade. "There were lots of roadblocks," she said. "One of the main challenges in obtaining funding was the concern that 'It isn't feasible.' But now we can say it's possible because we've done it."

Dr. Dinan is now proposing the first-ever linkage between the SEER-Medicare databases and the SEER-Virtual Tissue Repository (VTR), which is a prospective, forward-facing version of the work done with the SEER-RTR. Dr. Dinan wants to mine the databases to answer two questions about the use of immunotherapy to treat renal cell carcinoma (RCC). About 20 percent of RCC patients have a "durable response" to these therapies, meaning a potential cure, but no one can predict who those patients will be. Second, between one to three percent of RCC patients have severe toxic reactions to immunotherapies. Again, no one knows beforehand who those patients will be.

This is where Dr. Dinan's methodology shows its value. Dr. Dinan will use the SEER-Medicare data to identify everybody who received immunotherapy for RCC and identify two cohorts of patients, one that shows evidence of a durable response and another that shows evidence of a severe autoimmune toxicity.

"So, we'll cherry pick these people," said Dr. Dinan. "Instead of waiting to see what happens in a clinical trial, we're going to find the outcome of interest first, and then go back and pull those patients to study what's different about them. We'll have their whole clinical profile from the SEER data, their whole treatment profile from Medicare claims, and then we'll use the SEER-VTR to do genomic sequence analysis on their tumors to see if we can figure out what's driving these rare events, whether a durable response or a severe reaction. This has huge implications for our ability to study rare events and rare cancers in the future."



Michaela Dinan, PhD