BIG DATA

Modern biomedical research and clinical care generate more data than ever. Information can include genetic, imaging and metabolic data from tens of thousands of patients. It may provide responses to treatment and side effects. These records are creating rich datasets, which can be mined with new artificial intelligence and machine learning tools.

Big data will provide an unprecedented opportunity to understand cancer at every level, from molecular signatures to nationwide statistics, and to help make treatment decisions based on the knowledge distilled from these massive collections.

Once available only to research institutions with extensive data storage and computing capabilities, these datasets are increasingly available to the wider researcher community through repositories such as NCI's Genomic Data Commons, which integrates data from large, collaborative projects such as The Cancer Genome Atlas and the Therapeutically Applicable Research to Generate Effective Therapies (TARGET) initiative.

These repositories provide efficient systems to securely store, share and analyze this information without compromising patient identity. Researchers are encouraged – sometimes required – to add their own results to these collections.

The availability of these public data collections has spurred development of software tools designed to store, process, analyze and visualize large datasets. In turn, this has led to a push to train a new generation of scientists in how to contribute to, and use, these data science assets. As these repositories mature, they will incorporate some of these shared tools for data analysis.

At the forefront of these tools and methodologies are machine learning and artificial intelligence approaches in which computer networks are programmed to rapidly analyze complex biomedical data and find hidden patterns.

For example, in what is known as quantitative imaging, software can detect and quantify abnormal features on a medical image that may be missed by the human eye, leading to an earlier or more accurate diagnosis.

In a research setting, these methods and algorithms are being employed to accurately predict which patients might benefit from a certain treatment based on a review of previous patient genomic data referenced against their responses to treatment.

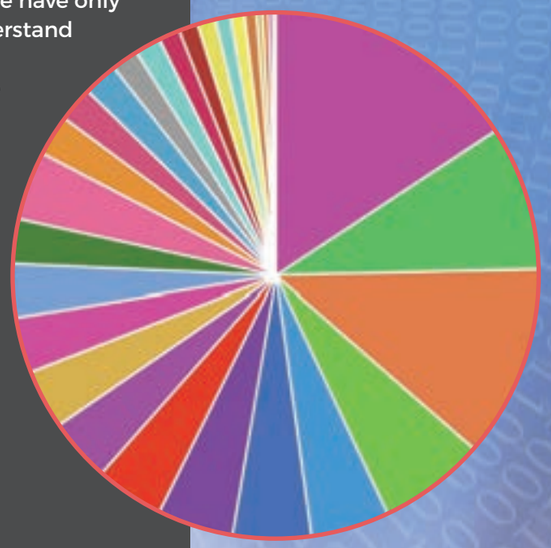
Another set of digital technologies is enabling patients to contribute data to clinical trials. Biometric sensors and fitness trackers, as well as health apps on patients' phones and mobile devices, allow researchers to obtain real-time reports on items such as pain and activity levels to supplement information collected during intermittent clinic visits.

Finally, in the realm of public health, more sophisticated ways to monitor cancer cases and management on a population level are becoming available.

An improved ability to extract and interpret data tucked into the health records of cancer patients while reliably removing identifiable patient information would do much to speed our ability to measure differences in cancer rates due to changes in cancer screening, therapies and healthcare policies.

These population data would also let us better detect health disparities, assess the outcomes of cancer prevention strategies and better formulate public policy to improve the health of broader populations.

Big data is already transforming cancer research and cancer care—but we have only just begun to understand what the data are telling us. Now, we are building on what we have learned and accelerating the transformation of data into meaningful advances for patients.



Caption: A pie chart representing data available in the NCI Genomic Data Commons (GDC) categorized by major site of origin. The GDC contains de-identified data from more than 32,000 patients with more than 60 different types of cancer gathered from 40 different projects.

Credit: NCI Genomic Data Commons