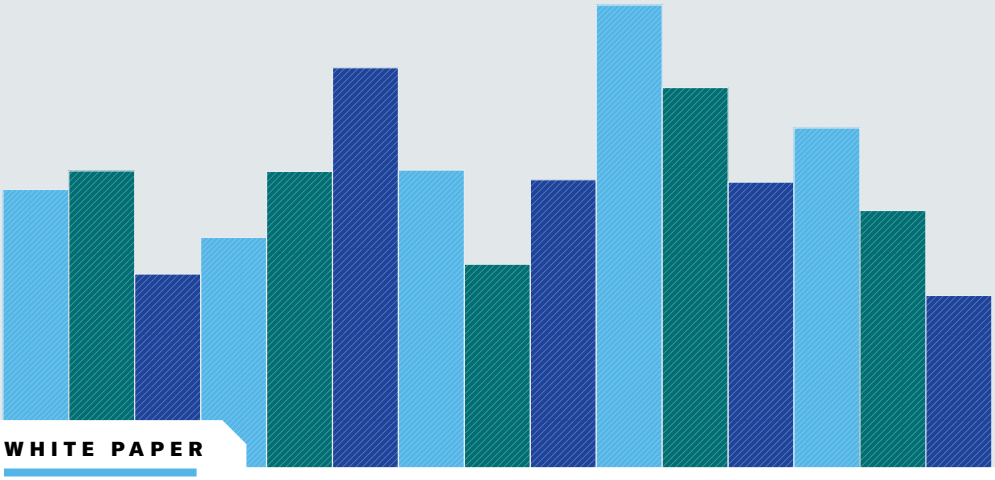
Harvard Business Review
ANALYTIC SERVICES



WHITE PAPER

Navigating the Data Deluge:

Turning Data into Insights across the Health Care and Life Sciences Industries

Sponsored by

NTT DATA

SPONSOR PERSPECTIVE

In our increasingly digitized world, today's most successful businesses are driven by data. They gather and analyze information from a myriad of sources, using what they learn to optimize operations, accelerate innovation, and make informed strategic decisions. Limitless opportunities to harness the power of data exist across the business spectrum, but perhaps no other industry can use it more meaningfully than the life sciences and health care sector.

With an unprecedented bounty of information at their fingertips, life sciences and health care professionals can understand the human condition like never before. From empowering consumers to take charge of their own well-being, to discovering effective new treatments, to providing a more convenient and affordable health care experience, data can reveal pathways to better outcomes for organizations and individuals alike. Thinking bigger, when high-quality and relevant health care data is democratized across businesses, governments, non-profits, and academia, it fuels our collective quest to become a healthier and happier society.

But, as it turns out, achieving data mastery is much easier said than done. Before organizations can make the most of the data they collect, they must be able to decipher it, understand its value, and seamlessly integrate it into their daily operations. All the while, they must tiptoe through a minefield of patient privacy rights, security concerns, and government regulations.

We at NTT DATA are proud to play a role in helping organizations navigate these complexities, and this report from Harvard Business Review Analytic Services serves as an excellent entry point to the discussion. It contains insightful commentary from respected experts in the field, as well as real-life examples from companies that are turning the ship in the right direction.

There is much work to be done but much to look forward to as data science enables a healthier future for us all. Thanks for reading.

Tanvir Khan

President, Health Plan and Life Sciences

NTT DATA

Navigating the Data Deluge:
Turning Data into Insights across the Health Care and Life Sciences Industries

Recent technological advances across the health care and life sciences industries are producing an almost inexhaustible supply of data that, when leveraged, can be used to improve patient outcomes. This real-world data (RWD) can be derived from multiple sources, including what is traditionally considered health data, such as electronic health care and hospital records, wearable devices, or applications on smartphones. RWD can also be gathered from other sources, such as internet usage that sheds light on behavioral patterns among populations.

"Every day, each of us is generating an enormous amount of data on physical and behavioral health," says Rajendra Pratap Gupta, chairman of the United Nations Internet Governance Forum (IGF) Dynamic Coalition on Digital Health. "We are all, by default, part of the world's biggest clinical trial without even knowing it."

The challenge—and the opportunity—for health care and life sciences organizations is to access and analyze this wealth of data to form real-world evidence (RWE) that can be used to advance medical care. But the task of amassing all this data and examining it critically can be overwhelming. The wide variety of data, from disparate sources and in different formats, presents challenges for data aggregation, curation, ingestion, and standardization. Simply having access to data does not ensure that its value can be harnessed; organizations need the technology and skills to find meaningful patterns in the data and draw conclusions that can support decision making.

There are many stakeholders involved across the health care value chain, from regulators to various life sciences organizations and consumers

HIGHLIGHTS

Simply having access to data does not ensure that its value can be harnessed; organizations need the technology and skills to **find meaningful patterns in the data** and **draw conclusions that can support decision making**.

A challenge inherent in the use of sensitive health care data is **how to anonymize it to preserve privacy while still retaining useful information**—and advances in artificial intelligence have made this possible.

Organizations have made great strides in **overcoming the challenges of data democratization**, and such efforts are changing the ways in which health care is devised, developed, and delivered—leading to tangible results in **public health**, **chronic disease management**, and **innovation**.

themselves, raising not only logistical issues but also ethical ones, especially when it comes to patient privacy. Access to data involves data sharing and collaboration—data democratization across the health care value chain—meaning organizations have to engage many constituencies to ensure that their efforts to connect with and use data bear fruit. "We are living in a data-hungry world," says Dr. Nimita Limaye, research vice president, life sciences R&D strategy and technology at International Data Corp. (IDC), a Needham, Mass.-based technology-focused market research and advisory firm. "There is a saying that data is the new gold, and this is true. But I'd say it's more than that. I think of data as the new foreign exchange that connects health care and life sciences."

Greater access to data from even non-health sources can help advance public health, says Lisa Gualtieri, a professor in public health and community medicine at Tufts University School of Medicine. "By accessing internet search patterns, it is possible to improve the way we predict and respond to disease outbreaks," she says. "The key issue is how to share and use that information. For example, emergency [public safety] departments could make sure personnel are available, perhaps a pharmaceutical company could ramp up manufacturing, or local pharmacies could stock up on certain products."

This Harvard Business Review Analytic Services report explores recent trends and advances in data democratization across the health care and life sciences industries. It examines how organizations are adopting RWD and RWE in all facets of their businesses and the tools, technologies, and approaches that are helping them do so. It also provides examples of how such approaches have led to better patient outcomes, and highlights best practices and factors vital for data democratization success.

To be sure, in an increasingly complex and fast-changing landscape, data democratization is emerging as a key area in health care and life sciences. "Particularly with the acceleration of digitization seen during the Covid-19 pandemic, data has been flowing in very aggressively—the variety, volume, and speed have grown exponentially," says IDC's Limaye. "Given this context, all business decisions now have to be data-driven."

Defining Data Democratization

Democratizing data, or expanding access to it, can enhance decision making and accelerate innovation. But as Limaye explains, there is a contradiction at play when putting this into practice. "The health care and life sciences industries want to democratize access to data so that all stakeholders, and more people within organizations, have access to it," she says. "But at the same time, governance, control, and privacy regulations are stricter than ever, which, of course, is critical for sensitive data. On the one hand, we need greater transparency, openness, trust, and data sharing; on the other, we need tighter control."

"There is a saying that data is the new gold, and this is true. But I'd say it's more than that. I think of data as the new foreign exchange that connects health care and life sciences," says Dr. Nimita Limaye, research vice president, life sciences R&D strategy and technology at International Data Corp.

The key to overcoming this contradiction, according to Limaye, lies not only in data access but also in the use of technology to analyze the data. "Data democratization is about empowering the end user to work with data in a meaningful way, in a timely manner. It depends not only on access to data but also on providing access to tools to fully leverage that data in a safe and compliant manner," she says.

Alexandra Ebert, chief trust officer at MOSTLY AI, a technology company based in Vienna that helps businesses unlock their big data assets while preserving privacy, believes that data democratization extends beyond the borders of a particular organization. "A common definition of data democratization involves giving more people within an entity or organization access to data. But a broader definition takes this one step further, where the benefits of opening up data access apply across the borders of an organization, the value chain, and across nations, too," she explains. "Society can benefit tremendously from data democratization in the health care space, and organizations that venture into data sharing and collaboration with other partners in the data-rich ecosystem also stand to benefit. Data democracy is about making it easy to access data and finding ways in which the value from that data can be fairly distributed across the different stakeholders."

Ebert describes an initiative from the European Commission's Joint Research Centre that aims to facilitate open synthetic data sharing across private- and public-sector organizations. "The European Union has access to the data of 50 million European citizens who have cancer. But the EU faces the same problem as any organization that has a wealth

of data; only a small group of people have access to it, and the data was never shared publicly—and this stalls policymaking, research, and innovation,” she says. “This initiative aims to make this valuable, granular data available to researchers and policymakers across Europe, in a privacy-safe manner, to facilitate research and allow data-driven policymaking.”

Indeed, data from non-health sources can contribute to public health. “For example, by looking at the number of searches for terms such as cough, fever, and congestion or purchases of tissues and other products—or even longer sessions on Netflix—it is possible to predict an increase in disease in a specific geography,” says Tufts’ Gualtieri. Similarly, she notes that other types of data could be used to understand changing patterns in population behavior that may impact public health.

Other definitions of data democratization involve giving patients more control over and access to their own health care data. “In theory, data democratization is data of the people, by the people, for the people. But in practice, it’s data of the people by the health care providers for the tech, finance, and health care industries,” the United Nations IGF’s Gupta asserts. “Solving the disconnect between these two definitions is the crux of reaching the full potential of data democratization.”

Gupta believes an approach to data democratization that puts data in the hands of the patient can help manage multiple chronic diseases. “Digital therapeutics have a huge role to play in this regard,” he explains. “With the right tools, patients will have a say in their treatment, will be far more compliant, and can save on health care costs. And the insurers will benefit, as hospitalizations could decrease. We have the technology, such as natural language processing and image processing, to build digital therapeutic devices, and, of course, machine learning is improving all the time. To make digital therapeutics an accessible, affordable, viable option, we need greater collaboration between all sorts of companies, such as tech startups and health care providers, but most of all, we need a mindset that gives the patient access to data.”

Dr. Satish Rath, a physician and digital health scientist, is a partner in Bengaluru, India, for Life Science Vision Capital (LSV), a London-based multi-stage investment firm focused on digital health, life sciences, health care, medical technology, and social impact investing, and he envisions a future where the combination of technology, collaboration across organizations and governments, and greater data democratization could revolutionize health care. “Imagine if every person had a secure cloud application on their mobile phone that stored all their personal health records, including claims data or medication purchased from a pharmacy, and this information was combined with behavioral data such as physical activity,” he explains. “This data set could be built incrementally and automatically over the course of a lifetime.”

An integral part of this scenario is that the data belongs solely to the patient. “This wealth of data would belong to the patient, and if they need treatment, they could give consent

“In theory, data democratization is data of the people, by the people, for the people. But in practice, it’s data of the people by the health care providers for the tech, finance, and health care industries. Solving the disconnect between these two definitions is the crux of reaching the full potential of data democratization.”

Rajendra Pratap Gupta, chairman of the United Nations Internet Governance Forum Dynamic Coalition on Digital Health

to the health care provider to access it for the duration of the treatment. They could also, if they wanted to, give explicit consent to contribute this data for research purposes," says Rath. "An application of this kind could ensure seamless access to diagnosis and treatment and boost research capabilities globally."

Whether organizations are focused on greater access to data internally or collaboratively across the value chain, they are likely to encounter common obstacles on their data democratization journey. Such challenges include establishing interoperability, ensuring privacy, and developing governance models. But technology is available to overcome these challenges.

Overcoming the Interoperability Challenge

According to Gupta, the greatest obstacles to data democratization are integrity, integration, and intelligence. "First, organizations must ensure their data is of high integrity," he says. "Second, they must find ways to integrate data from multiple sources in different formats—in other words, gain interoperability—and third, they need intelligent tools for analysis and insight. If they don't have integrity and interoperability, then the intelligence won't work."

The health care sector has developed shared frameworks that are advancing interoperability, such as the Fast Healthcare Interoperability Resources standard for exchanging health care information electronically and the Observational Medical Outcomes Partnership (OMOP) common data model designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence.

Many companies are using these frameworks and others to ensure a seamless exchange of information across various functions within their organization and with other organizations in the health care ecosystem. In addition, some companies are establishing centers of excellence (CoE) to advance data democratization.

Tokyo-based pharmaceutical company Takeda is one example. Sunil Dravida, global head of Takeda's Real-World Data Center of Excellence in Cambridge, Mass., describes the aim of the CoE as ensuring that the right data is available at the right time, in the right format, to all the constituents across the pharma landscape, spanning drug discovery, drug development, clinical trials, and medical affairs and operations. "Historical RWD is the basis on which Takeda can develop its drug pipeline and match our understanding of molecular structures and the history of natural disease with the ability to produce drugs that improve patients' lives," Dravida explains. "It also helps with post-market surveillance, such as detecting and reporting adverse events, or even predicting them through safety signal detection."

Takeda's CoE combines RWD from multiple sources and takes steps to ensure its quality, assess its provenance, and ensure interoperability. "If we don't have the data, we scan the landscape and find it," says Dravida. "We often use data providers and data aggregators to access electronic medical records, claims, lab data, genomics data, and biomarker data—there are so many players in the market that provide rich, granular data. It's important that the data we get in gives us longitudinality across the data set and meets our requirements for coverage of all the data points we need."

The next step is to ensure interoperability so that analysis can take place. "We have to bring all of this data together in a common model, without losing the provenance or lineage and without losing its value through transforming the data," Dravida continues. "Typically, we use OMOP as our standard to do so. We then use artificial intelligence [AI] and machine learning applications, as well as robotic process automation to aid analysis, particularly for highly complex data sets that need huge computational power."

An important element in this process is considering the end users of the data. "The CoE ultimately exists to provide a service to the end user, and we need to understand their skills and their needs," says Dravida. "Perhaps they are looking for self-service analytical tools, or dashboards, where they can access the data they need in a few clicks. If so, we empower them by providing the tools that they need. In other instances, data scientists may need to explore data sets beyond what is provided on a common platform, for instance, to delve into nuanced data applicable to a specific therapeutic area; if so, the CoE will provide access.

"All these elements fall under our understanding of data democratization—it comes down to having the right technology, processes, and people to accelerate RWE generation across the pharma landscape," he says.

Solving the Utility-Privacy Trade-off

A challenge inherent in the use of sensitive health care data is how to anonymize it to preserve privacy while still retaining useful information—and advances in AI have made this possible. MOSTLY AI's Ebert explains how AI-generated synthetic data can help organizations overcome what she terms "the utility-privacy trade-off." "Customer data is highly regulated, and the more sensitive the data, the more obligations organizations have to fulfill," she explains. "For instance, data privacy and security regulations, such as the U.S. Health Insurance Portability and Accountability Act or the General Data Protection Regulation in Europe, stipulate that data needs to be anonymous if it is to be used for research."

However, anonymizing data with traditional techniques such as masking or obfuscating can reduce its utility, as organizations need to delete or distort the majority of data

"We often use data providers and data aggregators to access electronic medical records, claims, lab data, genomics data, and biomarker data—there are so many players in the market that provide rich, granular data. It's important that the data we get in gives us longitudinality across the data set and meets our requirements for coverage of all the data points we need," says Sunil Dravida, global head of Takeda's Real-World Data Center of Excellence.

points because they could reveal an individual's identity. "With traditional anonymization, there is a trade-off between privacy and utility," Ebert notes. "Today, organizations need to anonymize rich, granular behavioral data assets. It is not uncommon for a data table to include 200, 300, or 1,000 columns per customer. If organizations use traditional anonymizing techniques—developed before the era of big data—that mask or delete many of these data columns, they don't leave any utility behind once the privacy threshold has been reached."

And in today's highly digitized environment, it is becoming easy to reidentify individuals from what may be assumed to be sufficiently anonymized data. "In the financial services sector, studies have shown that if you take only three supposedly anonymous credit card transactions, out of the hundreds of credit card transactions every customer performs per year, you can reidentify over 80% of customers," says Ebert. "This is the case even if the anonymized data only reveals the merchant and the date of the transaction—nothing more." This example illustrates the reidentification risk inherent in traditional anonymization, even if that effort removes sensitive data such as names or Social Security numbers to protect privacy.

Advances in AI—particularly for the generation of synthetic data—can help overcome this challenge. "Synthetic data retains the value of the data set while preserving privacy and making reidentification impossible. Synthetic data is generated by AI; it looks and feels like RWD but does not include data points from any actual individuals," Ebert explains. "The software trains on existing RWD and learns all of the patterns, correlations, and time dependencies in the data set. It then generates new synthetic patients and their synthetic medical records. And if you look at these synthetic data sets from a statistical point of view, you will find the exact same patterns, correlations, and structures, thereby removing the need to use RWD for sharing and analysis."

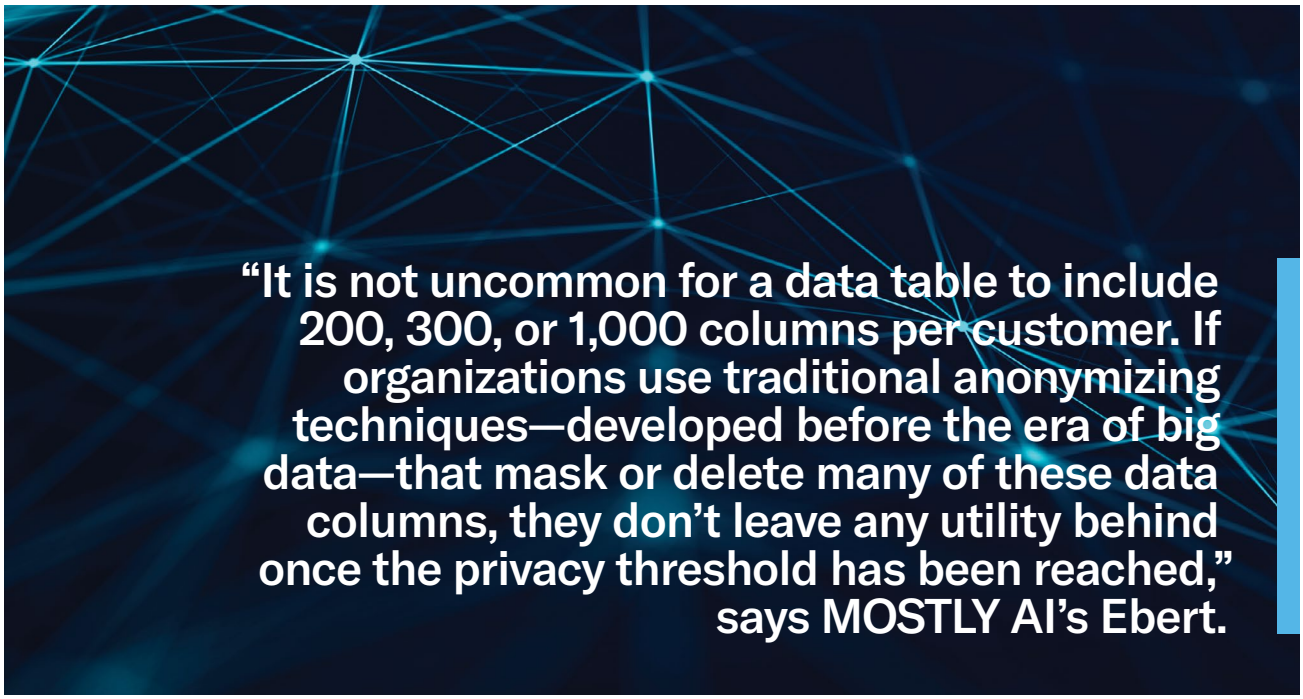
According to Ebert, synthetic data is a game changer, since it advances data democratization, simplifies legal compliance, and saves time. It also advances research and innovation. "In traditional data analytics, scientists need to know beforehand which columns of data they need to solve a particular challenge. By contrast, the purpose of applying AI is to find patterns or insights that you are not aware of yet, and for that, you need explorable data including all the available columns and data points," she says. "While traditional anonymization can only give you a handful of data points, AI-generated synthetic data preserves the patterns in RWD and reveals insights across huge amounts of data that are essential for innovation—without infringing on customers' privacy."

Building Diverse Teams to Reduce Bias

As Ebert points out, bias and discrimination can make their way into AI, not because artificial intelligence is inherently biased, but because humans are. Organizations need to address this issue if they are to develop solutions or products that can fairly serve all consumers' needs. "Often, the data that's used to develop machine learning models or AI solutions carries historic human bias. One of the reasons for this is that organizations may have skewed data without even realizing it," she says.

The well-known case of Massachusetts Institute of Technology researcher, Joy Buolamwini, highlights the ways in which human bias can affect product outcomes. "Buolamwini was investigating facial recognition systems, and she noticed that the software she was researching could detect the faces of her light-skinned colleagues but not her own darker skin," Ebert explains. Various AI-powered facial recognition systems worked well on the faces of white men, were less accurate for white women, were far less accurate for black men, and fared the worst when identifying black women.

"The problem," Ebert continues, "was that the data used to develop and train the algorithms was skewed. It didn't include enough examples of minority groups. But it took a woman of color to notice it. Building diverse teams can help organizations to detect whether their data is representative across any number of categories such as race, gender, or financial measures. Fair synthetic data can help to create data that is fair by design, either by including more examples of minority groups or by correcting historical biases in the data—for example, by eliminating the gender pay gap—and



can make it easier for downstream models that are trained on this data not to learn bias in the first place."

Enhancing Collaboration

Limaye explains how advances in technologies such as federated learning—a machine learning technique that allows algorithms to be trained collaboratively without exchanging the data itself—fuel collaboration across the health care and life sciences value chain while scaling innovation. "By moving the training to the data instead of the data to the training, this technology can address the crucial bottleneck of sharing highly sensitive R&D data," she asserts. "When we think about data fluidity, an essential part of data democratization, federated learning platforms are going to make a huge difference. Adoption is still in its early days, but major pharma companies have already joined forces and embraced federated learning, particularly in the area of drug discovery."

A challenge when it comes to data sharing is that many organizations do not want to share proprietary data that they have invested heavily in and accumulated over many years. "Nobody wants to share their intellectual property [IP], and that's perfectly understandable," says Limaye. "Federated learning provides opportunities for organizations to benefit from intelligence and gain access to a wealth of data to train deep learning algorithms while keeping their IP private. Organizations can effectively harness better insights in their own data by using the collective power of accumulated large data sets across multiple organizations to train the algorithms, and then share the resulting upgraded intelligence tool—thereby accelerating and scaling innovation."

Data democratization is enhancing collaboration across professions, too. LSV's Rath believes that a combination of skills across diverse professions is essential for harnessing the power of data. "If we want to maximize the value of data, then we need expertise in the medical domain and in digital," he says. "I believe that in the future, clinicians will need digital skills, and analysts or data scientists will need an understanding of life sciences. There is greater collaboration across these professions at present, but for meaningful outcomes, I think we will change the way we think about training and adjust our perception of these skills as belonging to separate professions. This approach is important for data-based discovery and research and innovation."

A Path to Data Democratization

Putting the patient first is among five best practices and key considerations that stakeholders across the health care and life sciences value chain can embrace to enhance their data democratization journeys. Here's the complete list:

Put the patient first. All stakeholders should ensure that efforts to democratize data and deliver the desired outcomes

"History shows that 19th-century medicine was about chemistry, 20th-century medicine was about physics, and 21st-century medicine is about technology and data—so medical science is now data science, and vice versa," says the United Nations IGF's Gupta.

put the patient at the heart of their strategy. "First and foremost, data democratization should be about engendering patients' trust by providing them with access to their personal health information. Whether in health care, or in life sciences, or where these fields come together—the overarching strategy should keep in mind that we are doing this for the patient," says Limaye. "That principle can cascade into all areas of strategy, such as defining what type of RWD you need for your specific therapeutic area, developing a data governance model, finding the right partners, and implementing the appropriate technology."

Understand complexity. "The field of health care and health care data is highly complex, and this means that one-size-fits-all solutions may not be effective," says Gualtieri. "What works for urban patients may not work for suburban or rural patients. Solutions based on the needs of a younger demographic may not be suitable for older patients who have more complex health needs. The challenge is to understand this complexity, examine all the components, and assess what solutions are not only cost-effective but also relevant to all the stakeholders involved, from doctors to organizations and patients."

Establish governance models early on. "When it comes to data sharing, establishing rules for partnerships is incredibly important," says Limaye. "Consider who will have access to the data, who it belongs to, how it will be used, and who will benefit. Also, what are the compliance requirements, and what happens if one player in the system does not comply? A risk-mitigation strategy needs to be thought out well in advance, and the consequences need to be detailed."

Think about data as a public good. "Organizations realize the value that data holds, but collectively we can do so much more if we build a mindset that data is a public good," says Rath. "For RWE that can aid advances in medical care, we need access to large volumes of varied RWD from multiple geographies. Greater collaboration and data sharing are the key to innovation and better care for all."

Incorporate diverse viewpoints. Diversity can foster more ethical decision making. "When it comes to ethics, we shouldn't burden data scientists with making a decision about what's right or wrong," says Ebert. "You need to have all the different departments of a company involved, including people from different cultural backgrounds. And ideally, organizations could include consumers on ethics boards. Ethical considerations are too important to let one team, or one profession, decide upon."

Conclusion

Data democratization in health care and life sciences is a complex area and presents challenges for organizations that are willing to embrace the journey. Organizations across the value chain are overcoming these challenges by developing standardized, shared frameworks that are a key component of democratizing data and ensuring a seamless exchange of information across various functions within an organization and with other organizations in the health care ecosystem. Others are establishing RWE centers of excellence that can generate insights and affect decision making across the clinical, pharma, and medical devices business continuum.

Organizations have made great strides in overcoming the challenges of data democratization, and such efforts are changing the ways in which health care is devised, developed, and delivered—leading to tangible results in public health, chronic disease management, and innovation.

What is clear is that data holds the key to the future of health care. "History shows that 19th-century medicine was about chemistry, 20th-century medicine was about physics, and 21st-century medicine is about technology and data—so medical science is now data science, and vice versa," asserts Gupta. "Data-backed technology is defining the future of health care."

Harvard Business Review
ANALYTIC SERVICES

ABOUT US

Harvard Business Review Analytic Services is an independent commercial research unit within Harvard Business Review Group, conducting research and comparative analysis on important management challenges and emerging business opportunities. Seeking to provide business intelligence and peer-group insight, each report is published based on the findings of original quantitative and/or qualitative research and analysis. Quantitative surveys are conducted with the HBR Advisory Council, HBR's global research panel, and qualitative research is conducted with senior business executives and subject matter experts from within and beyond the *Harvard Business Review* author community. Email us at hbranalyticservices@hbr.org.

hbr.org/hbr-analytic-services

MCCRE31030922