IBM Watson, *Heal* Thyself

How IBM overpromised and underdelivered on AI health care

By ELIZA STRICKLAND ILLUSTRATIONS BY EDDIE GUY


MENISCUS AND
LIGAMENT TEARS

IN 2014, IBM OPENED swanky new headquarters for its artificial intelligence division, known as IBM Watson. Inside the glassy tower in lower Manhattan, IBMers can bring prospective clients and visiting journalists into the "immersion room," which resembles a miniature planetarium. There, in the darkened space, visitors sit on swiveling stools while fancy graphics flash around the curved screens covering the walls. It's the closest you can get, IBMers sometimes say, to being inside Watson's electronic brain.

One dazzling 2014 demonstration of Watson's brainpower showed off its potential to transform medicine using AI—a goal that IBM CEO Virginia Rometty often calls the company's moon shot. In the demo, Watson took a bizarre collection of patient symptoms and came up with a list of possible diagnoses, each annotated with Watson's confidence level and links to supporting medical literature.

Within the comfortable confines of the dome, Watson never failed to impress: Its memory banks held knowledge of every rare disease, and its processors weren't susceptible to the kind of cognitive bias that can throw off doctors. It could crack a tough case in mere seconds. If Watson could bring that instant expertise to hospitals and clinics all around the world, it seemed possible that the AI could reduce diagnosis errors, optimize treatments, and even alleviate doctor shortages—not by replacing doctors but by helping them do their jobs faster and better.

Outside of corporate headquarters, however, IBM has discovered that its powerful technology is no match for the messy reality of today's health care system. And in trying to apply Watson to cancer treatment, one of medicine's biggest challenges, IBM encountered a fundamental mismatch between the way machines learn and the way doctors work.

IBM's bold attempt to revolutionize health care began in 2011. The day after Watson thoroughly defeated two human champions in the game of *Jeopardy!*, IBM announced a new career path for its AI quiz-show winner: It would become an AI doctor. IBM would take the breakthrough technology it showed off on television—mainly, the ability to understand natural language—and apply it to medicine. Watson's first commercial offerings for health care would be available in 18 to 24 months, the company promised.

In fact, the projects that IBM announced that first day did not yield commercial products. In the eight years since, IBM has trumpeted many more high-profile efforts to develop AI-powered medical technology—many of which have fizzled, and a few of which have failed spectacularly. The company spent billions on acquisitions to bolster its internal efforts, but insiders say the acquired companies haven't yet contributed much. And the products that have emerged from IBM's Watson Health division are nothing like the brilliant AI doctor that was once envisioned: They're more like AI assistants that can perform certain routine tasks.

"Reputationally, I think they're in some trouble," says Robert Wachter, chair of the department of medicine at the University of California, San Francisco, and author of the 2015 book *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age* (McGraw-Hill). In part, he says, IBM is suffering from its ambition: It was the first company to make a major push to bring AI to the clinic. But it also earned ill will and skepticism by boasting of Watson's abilities. "They came in with marketing first, product second, and got everybody excited," he says. "Then the rubber hit the road. This is an incredibly hard set of problems, and IBM, by being first out, has demonstrated that for everyone else."

At a 2017 conference of health IT professionals, IBM CEO Rometty told the crowd that AI "is real, it's mainstream, it's here, and it can change almost everything about health care," and added that it could usher in a medical "golden age." She's not alone in seeing an opportunity: Experts in computer science and medicine alike agree that AI has the potential to transform the health care industry. Yet so far, that potential has primarily been demonstrated in carefully controlled experiments. Only a few AI-based tools have been approved by regulators for use in real hospitals and doctors' offices. Those pioneering products work mostly in the visual realm, using computer vision to analyze images like X-rays and retina scans. (IBM does not have a product that analyzes medical images, though it has an active research project in that area.)

Looking beyond images, however, even today's best AI struggles to make sense of complex medical information. And encoding a human doctor's expertise in software turns out to be a very tricky proposition. IBM has learned these painful lessons in the marketplace, as the world watched. While the company isn't giving up on its moon shot, its launch failures have shown technologists and physicians alike just how difficult it is to build an AI doctor.

The *Jeopardy!* victory in 2011 showed Watson's remarkable skill with natural-language processing (NLP). To play the game, it had to parse complicated clues full of wordplay, search massive textual databases to find possible answers, and determine the best one. Watson wasn't a glorified search engine; it didn't just return documents based on keywords. Instead it employed hundreds of algorithms to map the "entities" in a sentence and understand the relationships among them. It used this skill to make sense of both the *Jeopardy!* clue and the millions of text sources it mined.

"It almost seemed that Watson could understand the meaning of language, rather than just rec-

PROJECT: Oncology Expert Advisor

MD Anderson Cancer Center partnered with IBM Watson to create an advisory tool for oncologists. The tool used natural-language processing (NLP) to summarize patients' electronic health records, then searched databases to provide treatment recommendations. Physicians tried out a prototype in the leukemia department, but MD Anderson canceled the project in 2016—after spending US $62 million on it.

So Far, Few Successes

IBM began its effort to bring Watson into the health care industry in 2011. Since then, the company has made nearly 50 announcements about partnerships that were intended to develop new AI-enabled tools for medicine. Some collaborations worked on tools for doctors and institutions; some worked on consumer apps. While many of these alliances have not yet led to commercial products, IBM says the research efforts have been valuable, and that many relationships are ongoing. Here's a representative sample of projects.

DATE		IBM PARTNER	PROJECT	CURRENT STATUS
2011	Feb.	Nuance Communications	Diagnostic tool and clinical-decision support tools	No tools in use
	Sept.	WellPoint (now Anthem)	Clinical-decision support tools	No tools in use
2012	March	Memorial Sloan Kettering Cancer Center	Clinical-decision support tool for cancer	Watson for Oncology
	Oct.	Cleveland Clinic	Training tool for medical students; clinical-decision support tool	No tools in use
2013	Oct.	MD Anderson Cancer Center	Clinical-decision support tool for cancer	No tool in use
2014	March	New York Genome Center	Genomic-analysis tool for brain cancer	No tool in use
	June	GenieMD	Consumer app for personalized medical advice	No app available
	Sept.	Mayo Clinic	Clinical-trial matching tool	Watson for Clinical Trial Matching
2015	April	Johnson & Johnson	Consumer app for pre- and postoperation coaching; consumer app for managing chronic conditions	No apps available
	April	Medtronic	Consumer app for personalized diabetes management	Sugar.IQ app
	May	Epic	Clinical-decision support tool	No tool in use
	May	University of North Carolina, others	Genomic-analysis tool for cancer	Watson for Genomics
	July	CVS Health	Care-management tool for chronic conditions	No tool in use
	Sept.	Teva Pharmaceuticals	Drug-development tool; consumer app for managing chronic conditions	No tool in use; no app available
	Sept.	Boston Children's Hospital	Clinical-decision support tool for rare pediatric diseases	No tool in use
	Dec.	Nutrina	Consumer app for personalized nutrition advice during pregnancy	No app available
	Dec.	Novo Nordisk	Consumer app for diabetes management	No app available
2016	Jan.	Under Armour	Consumer app for personalized athletic coaching	No app available
	Feb.	American Heart Association	Consumer app for workplace health	No app available
	April	American Cancer Society	Consumer app for personalized guidance during cancer treatment	No app available
	June	American Diabetes Association	Consumer app for personalized diabetes management	No app available
	Oct.	Quest Diagnostics	Genomic-analysis tool for cancer	Watson for Genomics from Quest Diagnostics
	Nov.	Celgene Corp.	Drug-safety analysis tool	No tool in use
2017	May	MAP Health Management	Relapse-prediction tool for substance abuse	No tool in use

ognizing patterns of words," says Martin Kohn, who was the chief medical scientist for IBM Research at the time of the *Jeopardy!* match. "It was an order of magnitude more powerful than what existed." What's more, Watson developed this ability on its own, via machine learning. The IBM researchers trained Watson by giving it thousands of *Jeopardy!* clues and responses that were labeled as correct or incorrect. In this complex data set, the AI discovered patterns and made a model for how to get from an input (a clue) to an output (a correct response).

Long before Watson starred on the *Jeopardy!* stage, IBM had considered its possibilities for health care. Medicine, with its reams of patient data, seemed an obvious fit, particularly as hospitals and doctors were

AI's First Forays Into Health Care

Doctors are a conservative bunch—for good reason—and slow to adopt new technologies. But in some areas of health care, medical professionals are beginning to see artificially intelligent systems as reliable and helpful. Here are a few early steps toward AI medicine.

ROBOTIC SURGERY
Currently used only for routine steps in simple procedures like laser eye surgery and hair transplants.

CLINICAL-DECISION SUPPORT
Hospitals are introducing tools for applications like predicting septic shock, but they haven't yet proved their value.

IMAGE ANALYSIS
Experts are just beginning to use automated systems to help them examine X-rays, retina scans, and other images.

VIRTUAL NURSING
Rudimentary systems can check on patients between office visits and provide automatic alerts to physicians.

GENETIC ANALYSIS
With genome scans becoming a routine part of medicine, AI tools that quickly draw insights from the data are becoming necessary.

MEDICAL ADMINISTRATION
Companies are rushing to offer AI-enabled tools that can increase efficiency in tasks like billing and insurance claims.

PATHOLOGY
Experimental systems have proved adept at analyzing biopsy samples, but aren't yet approved for clinical use.

MENTAL HEALTH
Researchers are exploring such applications as monitoring depression by mining mobile phone and social media data.

switching over to electronic health records. While some of that data can be easily digested by machines, such as lab results and vital-sign measurements, the bulk of it is "unstructured" information, such as doctor's notes and hospital discharge summaries. That narrative text accounts for about 80 percent of a typical patient's record—and it's a stew of jargon, shorthand, and subjective statements.

Kohn, who came to IBM with a medical degree from Harvard University and an engineering degree from MIT, was excited to help Watson tackle the language of medicine. "It seemed like Watson had the potential to overcome those complexities," he says. By turning its mighty NLP abilities to medicine, the theory went, Watson could read patients' health records as well as the entire corpus of medical literature: textbooks, peer-reviewed journal articles, lists of approved drugs, and so on. With access to all this data, Watson might become a superdoctor, discerning patterns that no human could ever spot.

"Doctors go to work every day—especially the people on the front lines, the primary care doctors—with the understanding that they cannot possibly know everything they need to know in order to practice the best, most efficient, most effective medicine possible," says Herbert Chase, a professor of medicine and biomedical informatics at Columbia University who collaborated with IBM in its first health care efforts. But Watson, he says, could keep up—and if turned into a tool for "clinical decision support," it could enable doctors to keep up, too. In lieu of a *Jeopardy!* clue, a physician could give Watson a patient's case history and ask for a diagnosis or optimal treatment plan.

Chase worked with IBM researchers on the prototype for a diagnostic tool, the thing that dazzled visitors in the Watson immersion room. But IBM chose not to commercialize it, and Chase parted ways with IBM in 2014. He's disappointed with Watson's slow progress in medicine since then. "I'm not aware of any spectacular home runs," he says.

He's one of many early Watson enthusiasts who are now dismayed. Eliot Siegel, a professor of radiology and vice chair of information systems at the University of Maryland, also collaborated with IBM on the diagnostic research. While he thinks AI-enabled tools will be indispensable to doctors within a decade, he's not confident that IBM will build them. "I don't think they're on the cutting edge of AI," says Siegel. "The most exciting things are going on at Google, Apple, and Amazon."

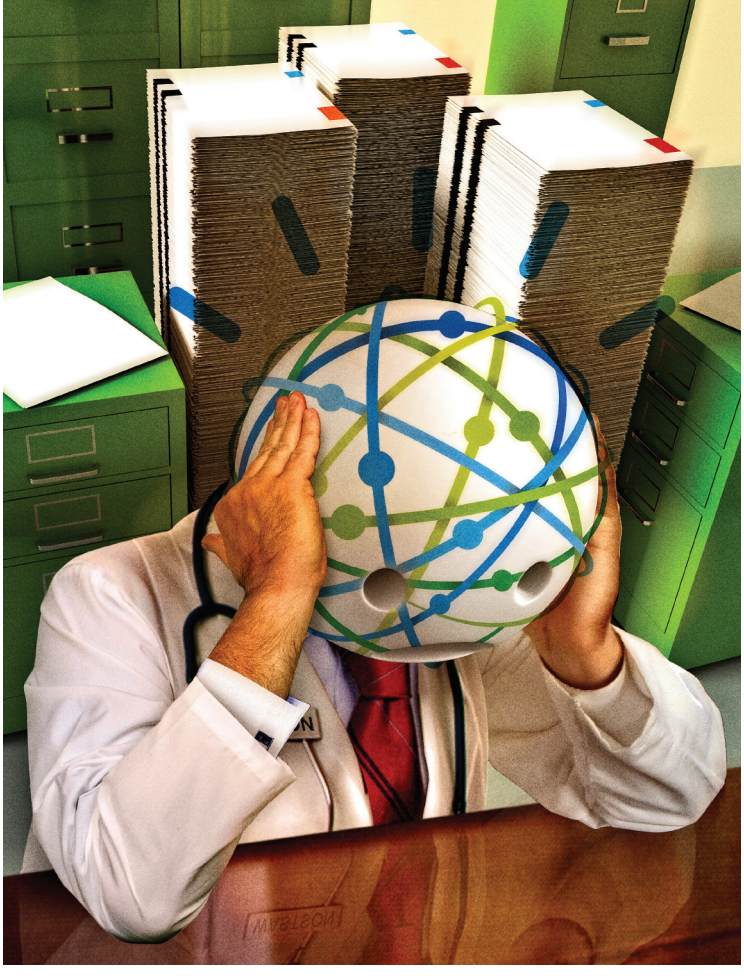
As for Kohn, who left IBM in 2014, he says the company fell into a common trap: "Merely proving that you have powerful technology is not sufficient," he says. "Prove to me that it will actually do something useful—that it will make my life better, and my patients' lives better." Kohn says he's been waiting to see peer-reviewed papers in the medical journals demonstrating that AI can improve patient outcomes and save health systems money. "To date there's very little in the way of such publications," he says, "and none of consequence for Watson."

In trying to bring AI into the clinic, IBM was taking on an enormous technical challenge. But having fallen behind tech giants like Google and Apple in many other computing realms, IBM needed something big to stay relevant. In 2014, the company invested US $1 billion in its Watson unit, which was developing tech for multiple business sectors. In 2015, IBM announced the formation of a special Watson Health division, and by mid-2016 Watson Health had acquired four health-data companies for a total cost of about $4 billion. It seemed that IBM had the technology, the resources, and the commitment necessary to make AI work in health care.

Today, IBM's leaders talk about the Watson Health effort as "a journey" down a road with many twists and turns. "It's a difficult task to inject AI into health care, and it's a challenge. But we're doing it," says John E. Kelly III, IBM

"Diagnosis is not the place to go. That's something the experts do pretty well. It's a hard task, and no matter how well you do it with AI, it's not going to displace the expert practitioner."

–AJAY ROYYURU, IBM's vice president of health care and life sciences research



senior vice president for cognitive solutions and IBM research. Kelly has guided the Watson effort since the *Jeopardy!* days, and in late 2018 he also assumed direct oversight of Watson Health. He says the company has pivoted when it needs to: "We're continuing to learn, so our offerings change as we learn."

The diagnostic tool, for example, wasn't brought to market because the business case wasn't there, says Ajay Royyuru, IBM's vice president of health care and life sciences research. "Diagnosis is not the place to go," he says. "That's something the experts do pretty well. It's a hard task, and no matter how well you do it with AI, it's not going to displace the expert practitioner." (Not everyone agrees with Royyuru: A 2015 report on diagnostic errors from the National Academies of Sciences, Engineering, and Medicine stated that improving diagnoses represents a "moral, professional, and public health imperative.")

In an attempt to find the business case for medical AI, IBM pursued a dizzying number of projects targeted to all the different players in the health care system: physicians, administrative staff, insurers, and patients. What ties all the threads together, says Kelly, is an effort to provide "decision support using AI [that analyzes] massive data sets." IBM's most publicized project focused on oncology, where it hoped to deploy Watson's "cognitive" abilities to turn big data into personalized cancer treatments for patients.

In many attempted applications, Watson's NLP struggled to make sense of medical text—as have many other AI systems. "We're doing incredibly better with NLP than we were five years ago, yet we're still incredibly worse than humans," says Yoshua Bengio, a professor of computer science at the University of Montreal and a leading AI researcher. In medical text documents, Bengio says, AI systems can't understand ambiguity and don't pick up on subtle clues that a human doctor would notice. Bengio says current NLP technology can help the health care system: "It doesn't have to have full understanding to do something incredibly useful," he says. But no AI built so far can match a human doctor's comprehension and insight. "No, we're not there," he says.

IBM's work on cancer serves as the prime example of the challenges the company encountered. "I don't think anybody had any idea it would take this long or be this complicated," says Mark Kris, a lung cancer specialist at Memorial Sloan Kettering Cancer Center, in New York City, who has led his institution's collaboration with IBM Watson since 2012.

PROJECT: Cognitive Coaching System

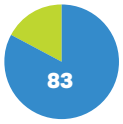
The sportswear company Under Armour teamed up with Watson Health to create a "personal health trainer and fitness consultant." Using data from Under Armour's activity-tracker app, the Cognitive Coach was intended to provide customized training programs based on a user's habits, as well as advice based on analysis of outcomes achieved by similar people. The coach never launched, and Under Armour is no longer working with IBM Watson.

The effort to improve cancer care had two main tracks. Kris and other preeminent physicians at Sloan Kettering trained an AI system that became the product Watson for Oncology in 2015. Across the country, preeminent physicians at the University of Texas MD Anderson Cancer Center, in Houston, collaborated with IBM to create a different tool called Oncology Expert Advisor. MD Anderson got as far as testing the tool in the leukemia department, but it never became a commercial product.

Both efforts have received strong criticism. One excoriating article about

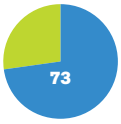
DO YOU AGREE?
Several studies have compared Watson for Oncology's cancer treatment recommendations to those of hospital oncologists. The concordance percentages indicate how often Watson's advice matched the experts' treatment plans.

Bumrungrad International Hospital, Thailand: 83% concordance



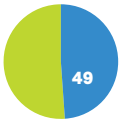
211 patients with breast, colorectal, gastric, and lung cancer

Manipal Comprehensive Cancer Center, India: 73% concordance



638 patients with breast cancer

Gachon University Gil Medical Center, South Korea: 49% concordance



656 patients with colon cancer

Watson for Oncology alleged that it provided useless and sometimes dangerous recommendations (IBM contests these allegations). More broadly, Kris says he has often heard the critique that the product isn't "real AI." And the MD Anderson project failed dramatically: A 2016 audit by the University of Texas found that the cancer center spent $62 million on the project before canceling it. A deeper look at these two projects reveals a fundamental mismatch between the promise of machine learning and the reality of medical care—between "real AI" and the requirements of a functional product for today's doctors.

Watson for Oncology was supposed to learn by ingesting the vast medical literature on cancer and the health records of real cancer patients. The hope was that Watson, with its mighty computing power, would examine hundreds of variables in these records—including demographics, tumor characteristics, treatments, and outcomes—and discover patterns invisible to humans. It would also keep up to date with the bevy of journal articles about cancer treatments being published every day. To Sloan Kettering's oncologists, it sounded like a potential breakthrough in cancer care. To IBM, it sounded like a great product. "I don't think anybody knew what we were in for," says Kris.

Watson learned fairly quickly how to scan articles about clinical studies and determine the basic outcomes. But it proved impossible to teach Watson to read the articles the way a doctor would. "The information that physicians extract from an article, that they use to change their care, may not be the major point of the study," Kris says. Watson's thinking is based on statistics, so all it can do is gather statistics about main outcomes, explains Kris. "But doctors don't work that way."

In 2018, for example, the FDA approved a new "tissue agnostic" cancer drug that is effective against all tumors that exhibit a specific genetic mutation. The drug was fast-tracked based on dramatic results in just 55 patients, of whom four had lung cancer. "We're now saying that every patient with lung cancer should be tested for this gene," Kris says. "All the prior guidelines have been thrown out, based on four patients." But Watson won't change its conclusions based on just four patients. To solve this problem, the Sloan Kettering experts created "synthetic cases" that Watson could learn from, essentially make-believe patients with certain demographic profiles and cancer characteristics. "I believe in analytics; I believe it can uncover things," says Kris. "But when it comes to cancer, it really doesn't work."

PROJECT: Sugar.IQ

Medtronic and Watson Health began working together in 2015 on an app for personalized diabetes management. The app works with data from Medtronic's continuous glucose monitor, and helps diabetes patients track how their medications, food, and lifestyle choices affect their glucose levels. The FDA-approved app launched in 2018.

The realization that Watson couldn't independently extract insights from breaking news in the medical literature was just the first strike. Researchers also found that it couldn't mine information from patients' electronic health records as they'd expected.

At MD Anderson, researchers put Watson to work on leukemia patients' health records—and quickly discovered how tough those records were to work with. Yes, Watson had phenomenal NLP skills. But in these records, data might be missing, written down in an ambiguous way, or out of chronological order. In a 2018 paper published in *The Oncologist*, the team reported that its Watson-powered Oncology Expert Advisor had variable success in extracting information from text documents in medical records. It had accuracy scores ranging from 90 to 96 percent when dealing with clear concepts like diagnosis, but scores of only 63 to 65 percent for time-dependent information like therapy timelines.

In a final blow to the dream of an AI superdoctor, researchers realized that Watson can't compare a new patient with the universe of cancer patients who have come before to discover hidden patterns. Both Sloan Kettering and MD Anderson hoped that the AI would mimic the abilities of their expert oncologists, who draw on their experience of patients, treatments, and outcomes when they devise a strategy for a new patient. A machine that could do the same type of population analysis—more rigorously, and using thousands more patients—would be hugely powerful.

But the health care system's current standards don't encourage such real-world learning. MD Anderson's Oncology Expert Advisor issued only "evidence based" recommendations linked to official medical guidelines and the outcomes of studies published in the medical literature. If an AI system were to base its advice on patterns it discovered in medical records—for example, that a certain type of patient does better on a certain drug—its recommendations wouldn't be considered evidence based, the gold standard in medicine. Without the strict controls of a scientific study, such a finding would be considered only correlation, not causation.

Kohn, formerly of IBM, and many others think the standards of health care must change in order for AI to realize its full potential and transform medicine. "The gold standard is not really gold," Kohn says. AI systems could consider

many more factors than will ever be represented in a clinical trial, and could sort patients into many more categories to provide "truly personalized care," Kohn says. Infrastructure must change too: Health care institutions must agree to share their proprietary and privacy-controlled data so AI systems can learn from millions of patients followed over many years.

According to anecdotal reports, IBM has had trouble finding buyers for its Watson oncology product in the United States. Some oncologists say they trust their own judgment and don't need Watson telling them what to do. Others say it suggests only standard treatments that they're well aware of. But Kris says some physicians are finding it useful as an instant second opinion that they can share with nervous patients. "As imperfect as it is, and limited as it is, it's very helpful," Kris says. IBM sales reps have had more luck outside the United States, with hospitals in India, South Korea, Thailand, and beyond adopting the technology. Many of these hospitals proudly use the IBM Watson brand in their marketing, telling patients that they'll be getting AI-powered cancer care.

"As a tool, Watson has extraordinary potential. I do hope that the people who have the brainpower and computer power stick with it. It's a long haul, but it's worth it."

—MARK KRIS, lung cancer specialist, Memorial Sloan Kettering Cancer Center, New York City

In the past few years, these hospitals have begun publishing studies about their experiences with Watson for Oncology. In India, physicians at the Manipal Comprehensive Cancer Center evaluated Watson on 638 breast cancer cases and found a 73 percent concordance rate in treatment recommendations; its score was brought down by poor performance on metastatic breast cancer. Watson fared worse at Gachon University Gil Medical Center, in South Korea, where its top recommendations for 656 colon cancer patients matched those of the experts only 49 percent of the time. Doctors reported that Watson did poorly with older patients, didn't suggest certain standard drugs, and had a bug that caused it to recommend surveillance instead of aggressive treatment for certain patients with metastatic cancer.

↗ POST YOUR COMMENTS at https://spectrum.ieee.org/watson0419

These studies aimed to determine whether Watson for Oncology's technology performs as expected. But no study has yet shown that it benefits patients. Wachter of UCSF says that's a growing problem for the company: "IBM knew that the win on *Jeopardy!* and the partnership with Memorial Sloan Kettering would get them in the door. But they needed to show, fairly quickly, an impact on hard outcomes." Wachter says IBM must convince hospitals that the system is worth the financial investment. "It's really important that they come out with successes," he says. "Success is an article in the *New England Journal of Medicine* showing that when we used Watson, patients did better or we saved money." Wachter is still waiting to see such articles appear.

Sloan Kettering's Kris isn't discouraged; he says the technology will only get better. "As a tool, Watson has extraordinary potential," he says. "I do hope that the people who have the brainpower and computer power stick with it. It's a long haul, but it's worth it."

Some success stories are emerging from Watson Health—in certain narrow and controlled applications, Watson seems to be adding value. Take, for example, the Watson for Genomics product, which was developed in partnership with the University of North Carolina, Yale University, and other institutions. The tool is used by genetics labs that generate reports for practicing oncologists: Watson takes in the file that lists a patient's genetic mutations, and in just a few minutes it can generate a report that describes all the relevant drugs and clinical trials. "We enable the labs to scale," says Vanessa Michelini, an IBM Distinguished Engineer who led the development and 2016 launch of the product.

Watson has a relatively easy time with genetic information, which is presented in structured files and has no ambiguity—either a mutation is there, or it's not. The tool doesn't employ NLP to mine medical records, instead using it only to search textbooks, journal articles, drug approvals, and clinical trial announcements, where it looks for very specific statements.

IBM's partners at the University of North Carolina published the first paper about the effectiveness of Watson for Genomics in 2017. For 32 percent of cancer patients enrolled in that study, Watson spotted potentially important mutations not identified by a human review, which made these patients good candidates for a new drug or a just-opened clinical trial. But there's no indication, as of yet, that Watson for Genomics leads to better outcomes.

The U.S. Department of Veterans Affairs uses Watson for Genomics reports in more than 70 hospitals nationwide, says Michael Kelley, the VA's national program director for oncology. The VA first tried the system on lung cancer and now uses it for all solid tumors. "I do think it improves patient care," Kelley says. When VA oncologists are deciding on a treatment plan, "it is a source of information they can bring to the discussion," he says. But Kelley says he doesn't think of Watson as a robot doctor. "I tend to think of it as a robot who is a master medical librarian."

Most doctors would probably be delighted to have an AI librarian at their beck and call—and if that's what IBM had originally promised them, they might not be so disappointed today. The Watson Health story is a cautionary tale of hubris and hype. Everyone likes ambition, everyone likes moon shots, but nobody wants to climb into a rocket that doesn't work. ■