Editorial

The data deluge: The information explosion in medicine and science

Richard L. Byyny, MD, FACP

The explosive increase in the amount and flow of information and data represents an important professional challenge for those of us in medicine and science.

Today, data sets are measured in petabytes (10^{18} bytes), and data is so efficiently gathered and stored that it presents a major challenge when evaluating its reliability, extracting its useful information, and using it effectively to improve our understanding of science and medicine.

Five years ago there was already 276 billion gigabytes (Gb) of digital information, and about 19 billion Gb of analog information. The data continues to accumulate and physicians continue to struggle with how to organize it and use it to understand science, to prevent disease, and to better serve the suffering. As Nobel Prize winner Tom Cech notes, we are still in the contemporary Dark Ages when trying to access and utilize the available data and information being produced and stored.

Physicians and other scientists are good and getting better at producing data. But we must become proficient—with or without the help of technology—at mining and managing the data in ways that will allow us to use it to maximum effect.

Throughout history, changes in technology have often increased the production of information and its dissemination. When we advanced from verbal communication to written records we could slowly produce manuscripts and books that some could read and learn from. In concert with the development of writing were archives, repositories of tablets and other permanent records, which evolved to become libraries. But the ability to easily disseminate collected information had to await the fifteenth century.

Around 1439, Johannes Gutenberg invented the printing press, resulting in a dramatic increase in the spread of information at a more reasonable cost. Even then, many lamented the problem of too much information.

The integration of the rational sciences with medicine in the 1700s and 1800s built the foundation for scientific medicine. During this period the pursuit of observational science evolved and the study and understanding of anatomy progressed to pathologic anatomy and the identification of the relationship between clinical symptoms and signs to postmortem findings and disease. Physicians developed new instruments and methods to study diseases and patients. Jenner discovered the efficacy of cowpox vaccination to prevent smallpox. The discovery that quinine was a specific treatment for malaria occurred during the same period.

The concepts of cell theory, cellular pathology, physiology, and pathophysiology were established. Anesthesia and antisepsis dramatically improved surgery and its outcomes. The germ theory of disease was put forth and microbial agents causing disease were isolated, identified, and characterized. This was followed by the development of antimicrobial drugs, the use of antitoxins, and the development of more vaccines to prevent disease.

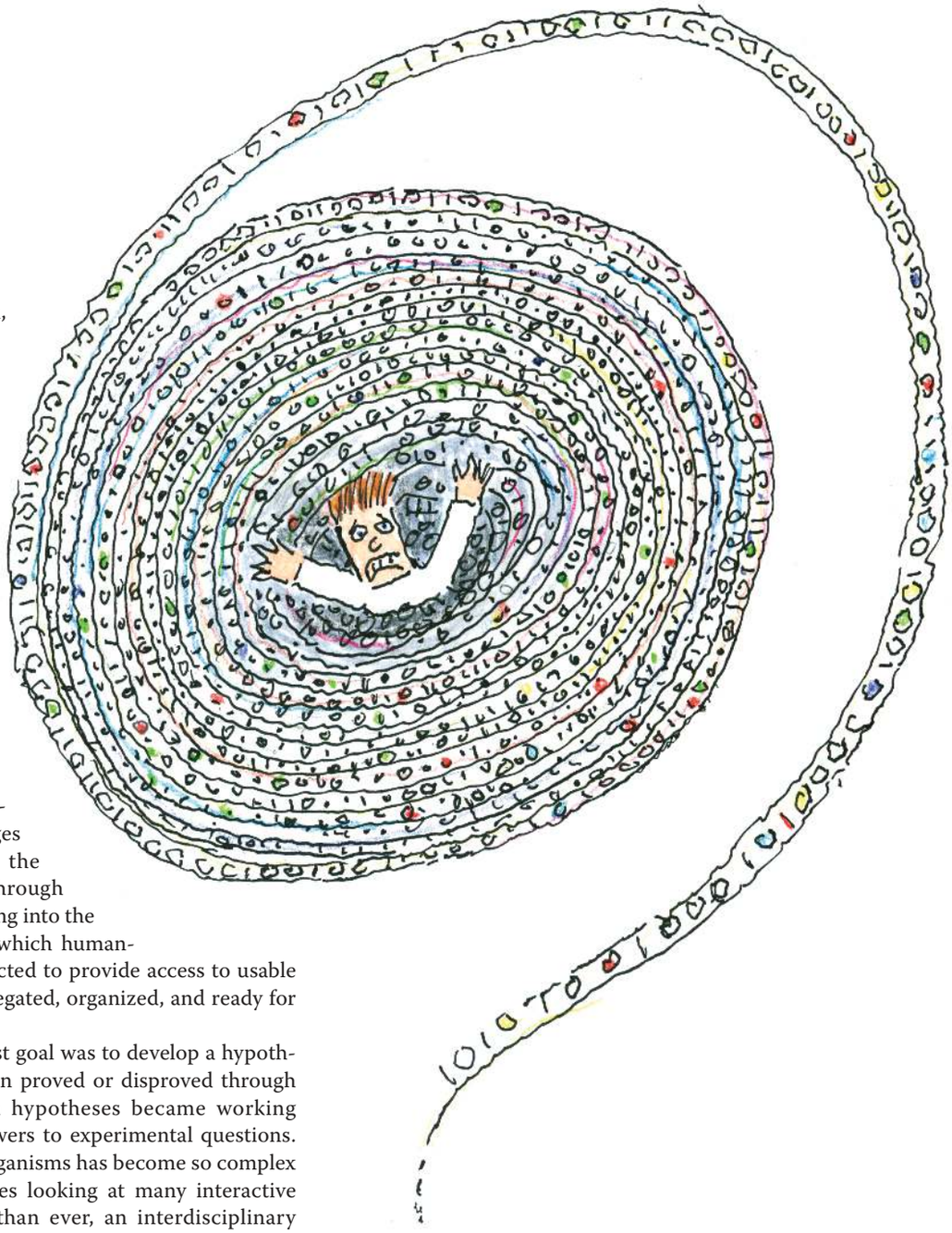
The twentieth century brought a dramatic rise in the publication of scientific journals and monographs, most of which were not critically reviewed. However, most physicians had no access to the available medical literature.

Sir William Osler, the author of *The Principles and Practice of Medicine*, the leading textbook of the early twentieth century, was very concerned about the increase in the medical literature. He worried about the lack of quality, limited access, and how it would be used. Osler and his colleague, John Shaw Billings, one of the innovators and leaders in medical librarianship,* worked together to further the development of medical libraries.

For most of the last fifty years, physicians and scientists have retrieved needed information on paper: they subscribed to journals, filled filing cabinets or their office floors with commonly referenced papers, or visited libraries. Biomedical scientists developed hypotheses about a gene, RNA, protein, receptor, or pathway, and performed experiments that resulted in huge advances in our understanding of health and disease, along with ever increasing data.

Throughout the twentieth and now twenty-first centuries, the flow of information has been increasing at nearly exponential rates, until it now threatens to drown us in data. By 2008, more than 5000 biology, chemistry, and medical journals were being published. PubMed listed one million articles. Publication of randomized controlled trials, the gold

*Billings was the creator of the Index Medicus, modernized the library of the Surgeon General's Office of the Army, and was the first director of the New York Public Library.



standard of clinical research, has increased rapidly since the first study was published in the 1940s, such that it is estimated that just to keep up on reading RCTs for the ten most common diagnoses in a field would mean reading twelve publications per week.

The development of computers and the Internet—instant access to virtually any and all information—has fundamentally changed the way knowledge is gathered, stored, and disseminated. With more than a billion people online and ten billion pages of information available on the Internet, we have evolved through Web 1.0 and 2.0 and are heading into the "semantic web," Web 3.0, in which human-computer interaction is projected to provide access to usable "metadata"—data that is aggregated, organized, and ready for analysis.

In the past, a scientist's first goal was to develop a hypothesis. That hypothesis was then proved or disproved through simple experiments. Proven hypotheses became working theories, providing valid answers to experimental questions. Today, the science of living organisms has become so complex that any investigation requires looking at many interactive processes, such that, more than ever, an interdisciplinary systems approach is needed to understand biology and the science of medicine. Data from multiple sources is coming at us in bigger pieces, faster, and cheaper. In genomics, for example, the amount of data has increased from about 100 Gb per year in 2006 at a cost of $20,000 per megabyte (Mb), to about 100,000 Gb in the year 2010, at a cost of $200 per Mb. Other areas of medicine show similar explosions in data, including that of diagnostic imaging, in which data sets of up to 1000 petabytes are not unheard of.

As noted above, traditional science is based on developing hypotheses and then designing experiments to confirm or disprove them, a process in many ways the antithesis of data mining. Data mining more resembles longitudinal population-based studies in which cohorts of people are followed over a period of time to identify associated predictors of disease.

Asking the appropriate question and assessing the appropriate databases is critical in designing longitudinal studies, as it is in data mining.

So how will we deal with the data deluge? As in Osler's era, medical scientists partner with librarians—which these days include computational scientists and technology experts—to develop new ways to store and retrieve information in forms that are useful. How do we present data in ways that allow us to grasp the essential and useful information and ignore the rest?

Key to the challenge of being able to use the flood of information that is threatening to overwhelm us will be the development and use of "intelligent agent software," programs that can automate commonly performed tasks and learn from their interactions with people. Such software could conceivably identify unrecognized opportunities to analyze data, solve problems, bring in interdisciplinary expertise, and integrate and prioritize diverse data sources in large, complex, and distributed information systems. To be truly useful, we would need the agents to know:

- What parts of particular sets of information are relevant to a specific individual and the current situation
- Which medical references pertain to a specific patient's condition
- To which web sites a physician should refer a patient for relevant information
- How to recognize potential unexpected relationships between the diverse information sources.

But it doesn't stop there. We would also need new tools and biomedical curators to categorize the data with common and integrated languages. Data collected should be curated and organized in a commonly agreed-upon format, then submitted to repositories that will allow interconnections among data sets. Data needs to become knowledge.

Until that happens, we need an effective way to take things in. I believe just-in-time learning is currently the most effective approach. Almost fifty years ago, one of my teachers and mentors, Dr. Telfer Reynolds, explained to me his strategy for continuous learning in medicine. He kept a black book in his lab coat pocket. When he discovered something he didn't know about medicine or a patient—which seemed rare—he would write it down in his book. Once a week, he would go to the Los Angeles Medical Society Library. He would start at the top of his list of questions and look up the information needed to answer his question, as well as other pertinent literature. At closing time, he would tear out his list, crumple it, and throw it in the trash so he could start a new list for the next week. Dr. Reynolds' use of just-in-time learning for specific reasons—to diagnose a problem or to teach students—meant that he was that much more likely to remember what he had just learned.

Today, just-in-time learning plays an increasingly important role. Information is most useful applied at the right time. Dr. Reynolds knew that learning is more likely to be useful, remembered, and teachable if it is tied to a problem or event. As a clinical scientist, practitioner, educator, and learner I, too, have long believed in just-in-time learning. Fortunately, the development of the internet and World Wide Web has greatly facilitated just-in-time access to information and data.

Although I have many issues with the use and utility of proprietary electronic health records and systems, it does provide one major advantage for just-in-time learning. While sitting with a patient and wondering about a diagnosis, test, or treatment I can immediately go to the online library and find the answer.

Before overloading your brain, recognize that not everything in medicine changes rapidly, if at all. New diseases appear infrequently. The clinical manifestations of most diseases change slowly, if at all. The symptoms in the history and physical findings, although varying from patient to patient, are usually consistent over time. And the physician's use of deductive reasoning to reach a conclusion from the clinical information doesn't change. If I work hard to learn what doesn't change rapidly in medicine and continue to practice the skills and use that knowledge, I will have a good, reliable, and persistent foundation of knowledge to draw from in caring for patients. I can then look up information "just-in-time" to answer questions that arise that I don't know or that may have changed recently.

What does change rapidly in medicine includes diagnostic strategies, technologies, and therapies. These areas require constant attention and continuous learning. Make it a habit to stay current in advances within these areas. Although not everything changes all the time, many things are changing. For issues too complex for this strategy, you can and should rely on subspecialty consultants—those whose depth and breadth of knowledge are more profound. It is also important to recognize that patients now have access to much of the same information as their physicians, and can bring useful or confusing information to bear on their ailments.

It is estimated that 12,000 new articles and 300 randomized controlled trials are added to Medline each week, and that new medical articles appear at a rate of one every twenty-six seconds. We clearly need a plan to keep up as well as we can.

Here is my proposed strategy:

1. Read the literature to attain, maintain, or improve knowledge and/or medical competence.

2. Maintain your curiosity and inquisitiveness—with an appropriate degree of skepticism.

3. Information is most helpful when used to answer questions about a patient's condition, pathobiology, diagnosis,

therapy, or prognosis.

4. Pick a place to start. This will depend on your own knowledge of a topic. If your knowledge is limited, start with general textbooks written by experts and then move to more specialized textbooks, including online textbooks. Once you have satisfactorily increased your understanding of the problem or issue, move on to critical reviews and original studies and research articles for more in-depth knowledge or to answer specific questions.

5. Develop a strategy for evaluating research articles and studies to determine if the article is of high quality and if the information will be useful. It is often useful to quickly scan or read the title, abstract, introduction, and conclusion to determine if the information is relevant to your practice, patients, knowledge, or teaching. If the scan of the article is positive then spend more time on the stated hypothesis, study design, results, analysis, assessment, and discussion.

6. Critically appraise the research and article content. I recommend the *JAMA* series on "Step-by-Step Critical Appraisal" that describes how to critically appraise medical literature. These have been published in *JAMA* for many years, with emphasis in each article about different types of research and how the articles should be critically reviewed.

Critical appraisal is a proven systematic process used to identify the strengths and weaknesses of a research article and to assess the usefulness and validity of the reported findings. It is the "basic science" of evidence-based medicine. When critically reviewing the medical literature it is also helpful to begin with a list of questions, as in the table.

Critical Appraisal of the Medical Literature: List of Useful Questions	
Is the study's research question relevant?	Was the study design appropriate for the research question?
Is the topic relevant to a question or to one's own field of work?	Did the study design and methods address the question?
Does the study, if valid, add anything new?	Are there important sources of bias or interpretation in the study?
Are there stated incremental advances of value?	How were participants selected and allocated?
What type of research question does the study pose?	How was data collected?
What is the stated hypothesis?	Did the study follow the protocol?
Who is the population of patients or subjects studied?	Was the analysis and assessment rational, appropriate and valid?
What are the measurable parameters or outcomes of interest?	Is the sample size sufficient for validation?
Is it a study related to diagnosis, therapy, frequency of events, prognosis, or something else?	Does the data justify the conclusions?
What is the study design? – Meta-analysis of randomized controlled trials – A randomized controlled trial – A cohort study, prospective or retrospective – A case-control study – An observational study – A descriptive study – A systematic or historical review	Are there sources of potential conflicts of interest? Are the findings clinically or scientifically relevant?

A key part of our professional responsibility is continuous learning to improve our knowledge, skills, and our practice of the art of medicine. Information overload makes this much more difficult. It is ironic that we have exchanged what was a lack of access to medical information in Osler's time for the contemporary problem of drowning in data.

I don't doubt that Osler would have embraced the abundance of information and the new technologies for finding and spreading it. But he, like many of us today, would have recognized and worried about the dilemma we face in distinguishing the useless from the useful, and in deciding how to put the useful to best use. So even though challenging, it's up to us to make sense of and organize the vast knowledge available to become more worthy to serve the suffering.

Dr. Byyny (AΩA, University of Southern California, 1964) is the Executive Director of Alpha Omega Alpha and editor of *The Pharos*. His address is:

525 Middlefield Road, Suite 130
Menlo Park, California 94025
E-mail: r.byyny@alphaomegaalpha.org