Computational Pathology

An Emerging Definition

David N. Louis, MD; Georg K. Gerber, MD, PhD; Jason M. Baron, MD; Lyn Bry, MD, PhD; Anand S. Dighe, MD, PhD; Gad Getz, PhD; John M. Higgins, MD; Frank C. Kuo, MD, PhD; William J. Lane, MD, PhD; James S. Michaelson, PhD; Long P. Le, MD, PhD; Craig H. Mermel, MD, PhD; John R. Gilbertson, MD; Jeffrey A. Golden, MD

Advances in high-throughput laboratory and health information technologies are revolutionizing the disciplines of pathology and laboratory medicine. The ability to extract clinically actionable knowledge using computational methods from complex, high-dimensional laboratory and clinical (digital) data, thereby yielding more precise diagnoses, disease stratification, and selection of patient-specific treatments, will clearly be a significant and important realization in the delivery of health care. Pathologists, who are at the nexus of diagnostic data, models of disease pathogenesis, and clinical correlation, are ideally positioned to provide leadership in the emerging "big data" era of medical care. We thus propose a vision for a new discipline of computational pathology.

We define *computational pathology* as an approach to diagnosis that incorporates multiple sources of raw data (eg, clinical electronic medical records; laboratory data, including "-omics"; and imaging); extracts biologically and clinically relevant information from those data; uses mathematical models at the levels of molecules, individuals, and populations to generate diagnostic inferences and predictions; and presents that clinically actionable knowledge to customers through dynamic and integrated reports and interfaces, enabling physicians, patients, laboratory personnel, and other health care system stakeholders to make the best possible medical decisions (Figure). This vision goes beyond an information technology or informatics-centric view and leverages the core competency of pathology—the understanding of disease processes at the molecular, individual, and population levels, and the ability to effectively integrate and communicate clinically actionable knowledge.

Realization of this vision will require changes in the practice of pathology, which is currently an order-driven, *observational* discipline. Our clinical laboratories generate more observations than any individual can reasonably interpret for clinical care. Although the data we provide to clinicians are largely reported as unique and independent results, the number of potentially meaningful relationships among combinations of observations is astronomic, and the workup (the process and order of the observations) has become increasingly complex and expensive. The understanding and high-level interpretations of those relationships remain in the mind of the pathologist and associated clinician, with the knowledge gap constantly increasing, which is not a sustainable situation.

Four driving factors have now created opportunities to move the field forward: (1) pathology data are increasingly digital and remain the most detailed and structured sets of information data in patient health records; (2) laboratory information systems are increasingly powerful, flexible, and integrated, allowing more complex analyses and interfaces with other diagnostic systems; (3) pathologists have access to the entire (digital) clinical record of patients, which gives them the ability to correlate laboratory data with clinical status and endpoints to develop foundational medical and biological knowledge; and (4) large-scale, clinical, phenotyping data are increasingly being collected and stored in structured databases, which enhance the capacity of pathologists to query and integrate information across many subjects to drive population-based analyses.

By developing the tools to harness those drivers, pathology can improve delivery of medically actionable knowledge over time and across populations and increase the efficiency of health care delivery. In this manner, the discipline of pathology can move from observation alone to a combination of observation, truly integrative interpretation, and longitudinal workup of patients. Although the parts of computational pathology exist, the whole does not. Thus, below, we describe those components and the people and infrastructure within the specialty that are needed to bring the parts together into a coherent whole to realize the vision of computational pathology.

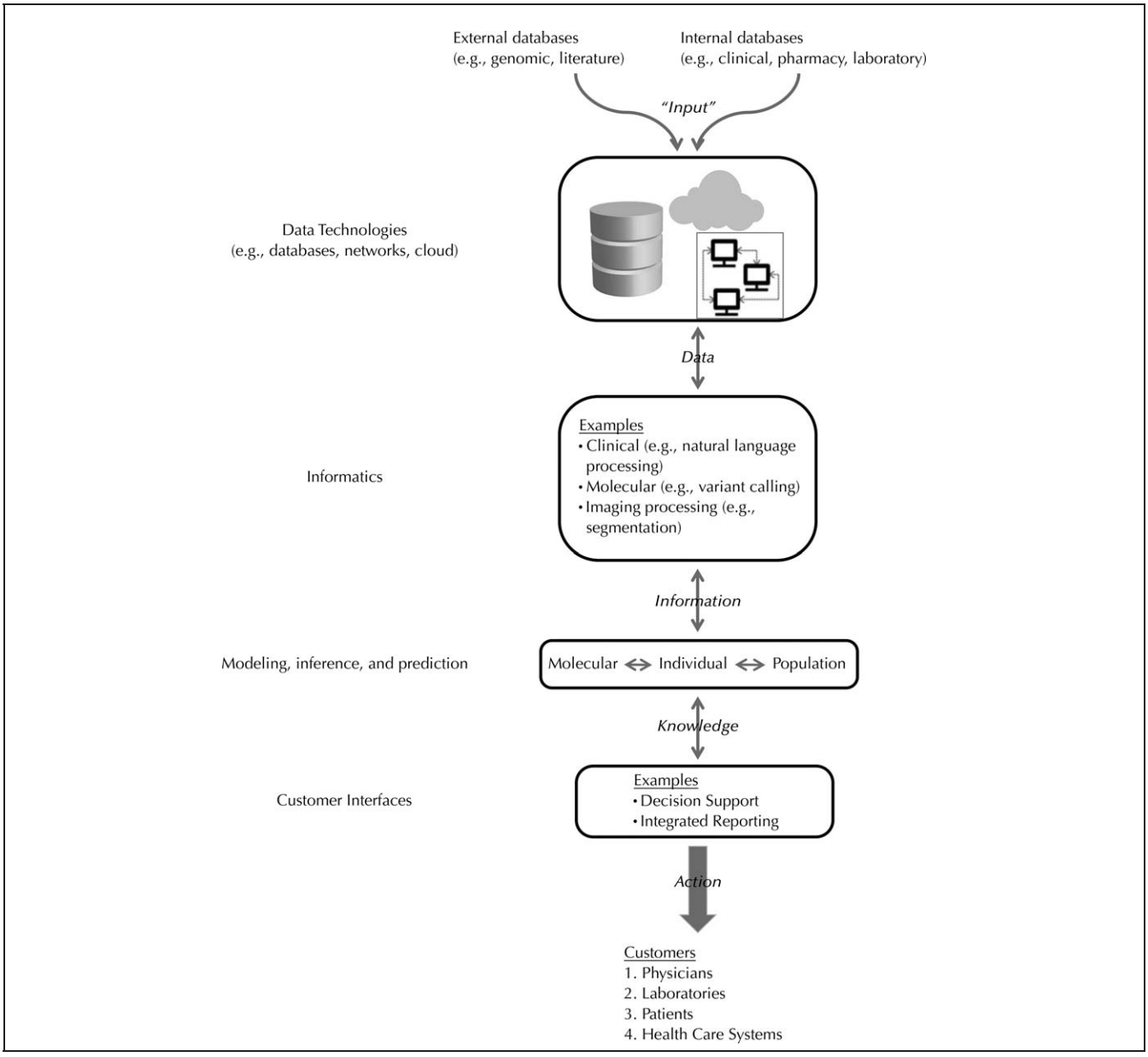
Accepted for publication February 3, 2014.

From the Partners Program in Computational Pathology and the Department of Pathology, Massachusetts General Hospital, Boston (Drs Louis, Baron, Dighe, Getz, Higgins, Michaelson, Le, Mermel, and Gilbertson); the Department of Pathology, Brigham and Women's Hospital, Boston (Drs Gerber, Bry, Kuo, Lane, and Golden); the Departments of Pathology (Drs Louis, Gerber, Baron, Bry, Dighe, Getz, Kuo, Lane, Michaelson, Le, Mermel, Gilbertson, and Golden) and Systems Biology (Dr Higgins), Harvard Medical School, Boston; and the Broad Institute, Cambridge, Massachusetts (Drs Getz, Le, and Mermel). Drs Louis, Gilbertson, and Golden are co-senior authors.

The authors have no relevant financial interest in the products or companies described in this article.

doi: 10.5858/arpa.2014-0034-ED

Reprints: Jeffrey A. Golden, MD, Department of Pathology, Brigham & Women's Hospital, 75 Francis St, Boston MA 02115 (e-mail: jagolden@partners.org).



Computational pathology: components and flow from data to clinically or biologically relevant information to diagnostic or prognostic knowledge and, ultimately, to action that improves the health of individuals and the efficiency of the health care system.

THE PARTS

Our envisioned schematic of computational pathology (depicted in the Figure, with examples given in the Table) features 5 major parts: (1) core sources of input *data* from internal sources, such as the laboratory information system and electronic health record, and from external sources, such as the scientific and clinical trials literature and publicly available genomic databases; (2) data technologies, which store core data in structured databases and distribute data using efficient network designs to compute nodes for downstream *analyses*; (3) informatics pipelines, which transform data into biologically and clinically relevant *information*, such as diagnostic categories from narrative patient notes, mutations from genomic data, or segmentation of individual cells or tissue features from imaging data; (4) modeling, inference, and prediction algorithms, which help to transform relevant information into clinically actionable *knowledge*, including diagnoses and prognoses, and which capture our understanding of pathogenesis, from molecular, to individual patient, to population levels; and (5) customer interfaces, which effectively communicate integrated diagnostic and prognostic knowledge to our customers and support their decision-making processes, ultimately leading to *action* to improve individual and population health or systemwide efficiency. These interfaces will provide dynamic and interactive knowledge to physicians, patients, laboratory personnel, and other health care system stakeholders.

Input Data Sets

We envision leveraging a wide scope of data sources that will serve as the basis for downstream analyses. Those data sources include ones internal and external to the health care

Examples of Computational Pathology Utility			
Process	**Current Practice**	**Potential Enhancements**	**Hypothetic Application**
Test selection and ordering	• Clinicians manually decide which tests to order • Test selection is often not fully individualized • Cost effectiveness is usually given only limited consideration • Key tests may be overlooked	• Computational algorithms suggest a "personalized" testing strategy based on statistical algorithms that integrate patient data and domain knowledge • When clinically permissible, algorithms may structure the sequence of testing to minimize expected costs • Algorithms suggest tests that may have been overlooked	• Currently, patients with suspected genetic abnormalities often have testing performed for many genes simultaneously, leading to an often expensive and inefficient workup • Computational pathology algorithms could look at patient characteristics to determine the most likely or most clinically important diagnoses and suggest testing for those first
Quality control for specimen analysis	• Evaluation of QC is largely binary: based on a few data points, an assay is classified either as in control (ie, performing to specifications) or out of control • Clinically significant calibration shifts/trends or preanalytic errors may go undetected. In addition, between-QC events are often undetected and can lead to erroneous results reporting	• Computational pathology algorithms use all available information (eg, QC, patient results, moving averages, delta checks, correlation with patient characteristics, etc) to continuously assess the quality of a given assay • Algorithms detect abnormalities in analytic performance that would be undetected by conventional QC • Algorithms detect results that are likely the result of preanalytic error • Anatomic pathology QC becomes more quantitative	• Inaccurate results because of analytic variability and preanalytic errors are not released to the patient record • Eliminates the potential for medical errors and inappropriate workups based on inaccurate results
Follow-up testing	• If the results of first-line tests indicate the need for further testing, clinicians must manually order additional studies • This may require an additional phlebotomy or patient visits • "Reflex testing" protocols are useful but not patient-specific	• Computational pathology algorithms would automatically add appropriate follow-up studies to the patient specimens and may request additional specimens likely to be needed upfront • Reflex testing would be patient-specific and greatly expanded in scope • Clinicians could order condition-specific workups without specifying the assay	• A clinician could order a "bleeding workup" on a patient with a possible coagulation disorder and send several tubes of blood; the algorithms would use patient-specific features to sequentially workup the specimens to arrive at a diagnosis
Results interpretation	• Laboratory reports consist primarily of individual measurements and observations • Clinicians must then manually integrate and interpret many discrete data elements to arrive at a diagnosis	• Algorithms generate patient-specific interpretive information • The algorithms may identify subtle patterns beyond those that would be identified manually by even the most astute clinicians • The laboratory report may not only include raw data but also diagnostic information and, in some cases, therapeutic recommendations	• A patient with high-normal levels of liver transaminases, slight but unconcerning changes in red blood cell indices, and some other laboratory results that appear unremarkable may currently receive a clean bill of health, whereas future computational pathology algorithms may identify that as a pattern suggestive of early hemochromatosis-related organ damage

Abbreviation: QC, quality control.

system or local institution. Primary internal data would be from anatomic pathology and the clinical pathology laboratories, including textual pathology reports, analyte values, and nucleic acid sequences. High-resolution digital imaging data from in vivo microcopy, specimen imaging, and whole-slide imaging will be particularly rich but challenging data sources. Because of the need for a standardized, consistent input (ie, glass slides) for the capture and downstream processing of digital images, the corollary to developing the ability to use digital data will be engineering complete systems for automating histologic processing to the point of acquiring digital images. Further complexities relate to analyses of molecular data at all levels (epigenetic, genetic, expression, etc.) and for many technical parameters (quality of material, quantity of material from small specimens, possible multiple data sets from different biopsies to capture heterogeneity of within tissue samples). Additional sources of internal data will include institutional, clinical, electronic health records, including structured, clinical, phenotyping databases; pharmacy, administration (registration, billing), and hospital operation (quality metrics, length of stay) records; and radiology imaging. External data sources will encompass a broad array of databases at multiple scales, including molecular pathways and gene annotations, deidentified individual genomes, the medical literature, and clinical trials at the population level.

"Curating" such data to ensure quality control will be an ongoing challenge that will benefit from pathology expertise. The success of this endeavor will be facilitated by the development, implementation, and acceptance of effective

data modeling and exchange standards in health care. These standards are currently an active area of development in both pathology and clinical informatics, and significant progress has been made in establishing them. Importantly, because pathology systems feed large amounts of critical information to virtually every other clinical system in health care, pathology is well positioned to drive and demand implementation of effective standards. Complete data standardization is not easy and will eventually need to include measures of data quality, completeness, precision, and definition of the shared information, as well as the models of laboratory practice in which the data are shared.

Data Technologies

Efficient storage and access to the large and complex data sets described in "The Parts," will be essential and will require database and networking technologies. Standard, relational databases, which have been used since the 1970s, are unlikely to be adequate for this task. Newer technologies, including compressed, object-oriented, and distributed architectures, will enable storage of massive data sets, such as DNA sequences and images; encoding of complex relationships among different data types; and efficient retrieval. Well-designed networks will also be needed to efficiently transport data within the computational pathology environment, which will necessarily include many compute nodes that may be local or cloud-based. Database and network technologies will need to adhere to strict security standards to ensure patient privacy.

Informatics

The disciplines of clinical informatics, bioinformatics, and image processing involve transformation of raw data into clinically or biologically relevant *information*, which can then be used by the downstream modeling and inference algorithms described below ("Modeling and Inference"). For instance, pathology and radiology reports and clinic notes contain largely "free text" data. Clinical informatics methods that use search and natural language processing algorithms can extract meaningful, structured data from those reports, such as diagnostic categories or defined clinical endpoints, and such abilities will be augmented by an increased use of structured data reporting in clinical medicine. Analogously, bioinformatics methods can be used to identify variants or mutations from raw sequencing data. Image processing methods can be used to segment histology or other types of images into individual cells or tissue regions, or to derive morphological measurements.

Informatics algorithms are commonly structured into computational pipelines, which consist of a series of algorithms in which the output of an earlier algorithm acts as an input into the next algorithm in a series. A pipeline is an important engineering principle that creates modularity, allowing for more efficient development, testing, and deployment of systems through separate subsystems—each performing a specific task—which ultimately link to perform a complex analysis. Areas in which computational pipelines are being actively used include genomic analyses and evaluation of digital imaging data from histologic and other microscopic preparations. These pipelines and their associated data typically require substantial computing power and are often deployed on large computational clusters with hundreds to thousands of compute nodes, which may be local or, increasingly, cloud-based (ie, they use on-demand, rented computers and storage). Implementation of informatics pipelines in computational pathology will require their integration within existing information systems and other supporting infrastructure. Additionally, many pipelines have been designed for research applications, and the rigorous engineering, documentation, testing, and regulatory filings required for use in the clinical laboratory are significant hurdles, but ones that pathologists are ideally qualified to drive.

Modeling and Inference

Modeling is the heart of computational pathology and the most important feature distinguishing this new discipline from classic informatics. A computational model describes internal interactions and measurable behaviors of a real-world system, often using formalized, mathematical language. Computational models will move pathology from observation and data management to the encoding of our understanding of disease processes in precise, testable formats. Good models allow data to be represented in interpretable ways, to hide complexities where possible, and ultimately, to make better predictions.

Models are typically built using experimental or expert observational knowledge. Pathologists are well positioned to provide the expert knowledge needed to build those models, given our understanding of biomedical systems at multiple levels of detail. In all aspects, the critical analysis and development of new models create opportunities for pathologists to work with statisticians, computer scientists, and other computationally focused experts, as well as to gain more knowledge and capacity to leverage computational thinking in our understanding of disease pathogenesis and our rendering of diagnoses. However, for many diseases, their pathogenesis is incompletely or even poorly understood. In those cases, machine-learning methods, which adaptively construct models based on data, can be employed to derive phenomenologic models from data. For example, modern Bayesian machine-learning methods provide a principled framework to balance detailed understanding and phenomenologic models, by incorporating prior knowledge when available and adaptively learning from data the parts of a model for which expert knowledge is sparse. All models will require extensive validation before clinical use. Validated models derived from accepted pathophysiologic mechanisms may appeal to clinicians' intuition and might be more readily integrated into patient management, whereas models inferred from machine-learning methods could require additional validation or demonstration of improved outcomes before being generally accepted by clinicians.

Another important feature of models is that they may describe behaviors of a system at different levels of detail and abstraction, each of which will be relevant in different contexts and depending on the state of knowledge of the system. In computational pathology, models at molecular (from single cells to tissues), patient-specific, and population levels will all be important. Molecular-level models will be useful to improving our understanding of disease processes and to identifying relevant biomarkers and "drugable" pathways. Patient-specific models will be essential for realizing the potential of personalized or precision medicine, allowing us to provide tailored interpretations of clinical information within the context of each individual patient. Population-level models are critical for reducing inefficiencies and for effectively allocating resources in pathology laboratories and in the broader health care

system. For instance, untoward effects of new medications could be detected across a population if an automated system were to identify a novel relationship between pharmacy (eg, prescription of a new medication) and laboratory (eg, changing renal function) data—likely at an earlier point than if such individual events progressed to their endpoints (eg, renal toxicity) and were then collated together by a series of physicians.

Probabilistic methods will be critical for rigorously evaluating the performance of computational pathology models. Statistical inference is the process of drawing conclusions or *knowledge* from a mathematical model of data, which are subject to random variations; *prediction* is the process of making an inference about an unknown outcome. Since biomedical data are inherently noisy because of measurement errors and sampling variations across time, space, and populations, statistical inferences and predictions, more than deterministic, rules-based approaches, will be essential for drawing meaningful conclusions from real-world biomedical data.

Customer Interfaces

A critical part of computational pathology will be delivering complex, multidimensional, and dynamic diagnostic and prognostic knowledge to a diverse customer base that includes clinicians, technical personnel in the laboratory, patients, and various other stakeholders in the health care system. Delivery of this knowledge must go beyond the current, static pathology report. Thus, computational pathology will need to take on an user interface design. Interfaces will incorporate interactive elements, allowing the customer to query and obtain multiple views of the diagnostic and prognostic knowledge provided. For instance, a report on a patient with an identified genetic condition may be tailored in a specific manner for a cardiologist versus an orthopedic surgeon or genetic counselor. Interactive elements would help to place knowledge about the individual patient within context, providing links to other patient-specific information, such as prior relevant findings, or further supporting information, including medical evidence from published clinical trials and general or specialty-specific guidelines. Notably, the "customer" of the pathology report may also be a machine; the pathology report will almost certainly be the input to a range of computational, display, and archiving activities. It must be in a form supporting human readability and structured for automated analyses. Existing standards, such as the HL7 CDA (clinical document architecture) and health care PDF (portable document format; ASTM–Association for Information and Image Management BP-01-08), support such reporting.

The ultimate goals of customer interfaces developed by computational pathologists will be to support the best possible decisions regarding patient health and the overall efficacy of the health care system. Thus, to be most useful, customer interfaces must also incorporate formal decision-theory algorithms that capture the utility of different decisions based on the customer's particular perspective. Elements of a decision-support system include pretest and posttest probabilities, the implications of a false-positive or false-negative in utility or cost to the customer, possible options ("constraining the decision"), providing a measure of accuracy and precision (the error bars or posterior distributions), and the ability to follow outcomes to inform future decisions. Traditional areas for decision support at the interface of pathology and clinical medicine have been used for guiding test ordering and interpretation. Our vision for computational pathology extends such functions to operate in more-complex areas, combining multiple sources of information. For example, to support clinicians in selecting antiretroviral therapies, pathogen identification, susceptibility testing, and genomic information, such as viral population genotypes and host pharmacogenomic markers, could be combined with patient phenotypic characteristics (such as $CD4^+$ T-cell counts and disease status) and prior information derived from population outcomes as well as from the patient's prior history to generate predicted profiles of responsiveness to different drug regimens. In this manner, a clinical decision-support module would enhance the diagnostic value of laboratory results from test selection to results interpretation to clinical action.

COMPUTATIONAL PATHOLOGY IMPROVES THE PRACTICE OF PATHOLOGY AND CREATES OPPORTUNITIES FOR PATHOLOGISTS IN INSTITUTIONAL SYSTEMS

The components of computational pathology must also be applied to improving the pathology enterprise itself. The tools of the discipline will be directly applicable to a wide range of laboratory activities and goals, including improving diagnostic accuracy, optimizing the health of the population under our care, and improving laboratory efficiency with reduced costs. Productivity and value must be assessed within an interdependent health care system. These interdependencies are important and when understood and exploited can be powerful because laboratory goals are seldom independent of the larger health care system (eg, decision-support models will partially depend on the cost, accuracy, and value that the health care system places on the time to diagnosis).

It is thus important that computational pathology include, at its core, not only data sources, models, and decision-support systems for the diagnostic process but also components for business intelligence and process analysis, and these modules need to interact. Just as a diagnostic decision-support system can include clinical data, it could also include process modeling or financial data. These capabilities would bring value to the system being analyzed, whether the task relates to management of patients, laboratories, or populations. In fact, one of the most important challenges in pathology today is the development of models that relate diagnostic laboratory output to global enterprise parameters such as health and financial outcomes. Indeed, analyses that could operate in near real time and derive value at the population level, such as tracking of infectious diseases or measuring test use and costs within the hospital or clinic, are some of the applications likely to have the greatest immediate effect.

Operationalizing this vision will require tools and user interfaces for pathologists and technical staff to access needed components of systems and resulting outputs. Current laboratory information systems provide minimal functionality to support this vision; that is, what is available to the pathologist today pales in comparison to what is available in other fields, such as financial analysis and process engineering. Thus, there will be a need to create a workstation/computational environment that connects the pathologist to models, decision systems, analytics, data sources, and robust reporting capabilities. With this

connection, the practicing pathologist will be able to oversee and lead the diagnostic process and contribute to the development of new computational tools.

We also envision computational pathology as broadly introducing quantitative modeling and analytic approaches to diagnosis, from molecules to individual patients to populations. This vision will require different systems, which must, in turn, be interfaced with other systems and data sources. Because many health care systems are now transitioning to large electronic medical records, this is an ideal opportunity to consider the needs of diagnostic systems residing in pathology departments. To do so requires substantial involvement by pathologists on leadership committees designing and overseeing information system implementations. This involvement by pathologists in enterprise-level information systems discussions is an essential step in keeping the discipline vibrant and near the center of clinical management in the future.

RESOURCES

Development of computational pathology as a discipline will require a long-term commitment, with a timeline on the order of years. As such, computational pathology must necessarily start with well-defined use cases to develop requirements and specification that will form the basis for longer-term plans and strategies. Development of the discipline will also clearly require considerable resources—certainly more resources than a single department can muster. As such, the development of computational pathology must be a collaborative undertaking, including academic and commercial partners. The academic side of the equation will involve not only pathologists but also researchers and clinical subspecialty departments. The commercial side will involve companies in information technology to biotech and to diagnostics, from large-market capitalization firms to small start-up companies. Interest and funding from government sources will also be an important aspect, with some funding opportunities already existing.

EDUCATION

Integration of computational methods and thought processes into training programs will be essential. Indeed, the current dearth of pathologists with training in quantitative disciplines is a significant hurdle that must be addressed. It is clear that future pathologists will require a fundamental understanding and ability to use computational principles in their practices. In turn, significant academic opportunities will exist for the development and integration of successful curricula.

WHAT'S IN A NAME?

Computational Pathology (rather than *computational diagnostics* or *computational medicine*) puts the discipline firmly within the field of pathology and highlights the unique capability of pathologists to model and understand disease at multiple levels of detail, from the molecular to the individual patient to populations, and to elucidate hidden aspects of the disease process not readily intuited just from observational analyses. Further, the name conveys our conviction that this new discipline has as much potential to help us *understand* diseases (ie, through accurate modeling of physiology and pathophysiology) as to diagnose them.

CONCLUSION

Continuing advancements in high-throughput laboratory and computational technologies provide the impetus for proactively forming a new subspecialty of computational pathology. These activities will transcend individual academic centers, commercial entities, and private practices. The components outlined will need to be built and integrated, requiring collaborations across clinical and research departments, institutions, and industry partners (from information technology to laboratory equipment makers). In the Boston, Massachusetts, area, this has started as a diverse set of collaborations between academic departments (eg, the Departments of Pathology at Massachusetts General Hospital and Brigham & Women's Hospital and the Departments of Medicine at these same institutions), between departmental and central information technology groups (eg, Partners HealthCare [Boston, Massachusetts]), between institutions (eg, Massachusetts General Hospital and Massachusetts Institute of Technology), and between hospitals and companies (ranging from the electronics industry to diagnostic laboratories to small start-up companies). Moreover, it will require multiple sources of funding and must affect reimbursed aspects of clinical care.

Although beyond the capabilities of an individual pathology department, the content behind the models and tools leveraged in computational pathology should nonetheless be developed by *pathologists* so that essential models of disease, biology, populations, laboratories, and diagnostic decisions are appropriately incorporated. Broader adoption of computational methods and quantitative thinking will improve our capacity to harness health care data in its many forms and ensure that the practice of pathology remains a specialty in which the knowledge and understanding of disease mechanisms remains firmly aligned with means to improve diagnostics for individual patients and across broader populations.

We recognize that the creation of computational pathology will take time and sustained vision. Moreover, it will require dramatic cultural and intellectual changes in the field of pathology. Some may doubt whether these changes will ever happen. To such doubters, we say keep in mind the advice of the architect Daniel Burnham, who said, "Make no little plans. They have no magic to stir men's blood...." Thus, although we understand that creation will take time, we look forward to getting the blood stirring on a real-life version of the above vision.