How Next-Generation Sequencing and Multiscale Data Analysis Will Transform Infectious Disease Management

Theodore R. Pak and Andrew Kasarskis

Icahn Institute and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York

Recent reviews have examined the extent to which routine next-generation sequencing (NGS) on clinical specimens will improve the capabilities of clinical microbiology laboratories in the short term, but do not explore integrating NGS with clinical data from electronic medical records (EMRs), immune profiling data, and other rich datasets to create multiscale predictive models. This review introduces a range of "omics" and patient data sources relevant to managing infections and proposes 3 potentially disruptive applications for these data in the clinical workflow. The combined threats of healthcare-associated infections and multidrug-resistant organisms may be addressed by multiscale analysis of NGS and EMR data that is ideally updated and refined over time within each healthcare organization. Such data and analysis should form the cornerstone of future learning health systems for infectious disease.

Keywords. whole genome sequencing; hospital-acquired infections; healthcare-associated infections; electronic medical records; multiscale analysis.

Next-generation sequencing (NGS) and "big data" analysis techniques may transform our understanding of diseases that have a complex inherited component, such as cancer, diabetes, and heart failure. Perhaps even more significant is the impact these technologies will have on the management of infectious diseases, which have discrete, identifiable causes that can be isolated, cultured, and tested against drugs in vitro as part of a standard clinical workflow. Despite steady technological improvements in each step, this workflow's principles have not changed for a century [1, 2].

Our capacity to acquire "omics" data about infections is increasing exponentially. Nanoscale parallelization of DNA sequencing has precipitously dropped the cost per base pair of finished genomes while increasing throughput, and the cost of sequencing and assembling a bacterial genome trends below $100 [2]. PacBio RS sequencing has increased median read lengths to over 10 kbp, facilitating rapid, automated finishing of genomes for outbreak pathogens [3, 4]. Recent studies have used "omics" experimental techniques such as Luminex cytokine assays, RNA sequencing, and mass cytometry to characterize immune responses to infection or vaccination with remarkable precision. Potential applications of this range from classifying acute respiratory infections in children [5] to predicting immunogenicity of a vaccine [6].

Many public databases curate and disseminate "omics" data relevant to infectious disease (Table 1), but most lack significant clinical metadata. Increasing adoption of electronic medical records (EMRs) can potentially mitigate this problem because they typically include data on demographics, medications, laboratory results, and more. However, with many different stakeholders entering EMR data, automatically extracting certain facts (eg, "this patient had the flu last Tuesday") is often difficult. Nevertheless, high-accuracy methods

Received 30 April 2015; accepted 24 July 2015; electronically published 6 August 2015.

Correspondence: Andrew Kasarskis, PhD, Icahn Institute and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1498, New York, NY 10029 (andrew.kasarskis@mssm.edu).

Clinical Infectious Diseases® 2015;61(11):1695–702

DOI: 10.1093/cid/civ670

Table 1. Examples of Public Bioinformatics Databases That May Be Leveraged for Multiscale Analysis of Infectious Disease[a]

Database Focus	For General Research	For Infectious Disease	
		Multipathogen	Pathogen-Specific
Genomes	• NCBI Nucleotide (GenBank/RefSeq) • ENA/EMBL • DDBJ	• ViPR • NMPDR • PATRIC • EuPathDB	• Influenza Research Database • Tuberculosis Database • LANL: Databases for HIV, HCV, and HFV
Gene products and functionality	• UniProt • KEGG	• Pathogen-Host Interaction Database • Antibiotic Resistance Genes Database • Comprehensive Antibiotic Resistance Database	
Expression and immune profiles	• GEO • ArrayExpress	• ImmPort	

Citations for individual databases can be found in the Supplementary Data.

Abbreviations: DDBJ, DNA Data Bank of Japan; ENA/EMBL, European Nucleotide Archive/European Molecular Biology Laboratory; EuPathDB, Eukaryotic Pathogen Database; GEO, Gene Expression Omnibus; HCV, hepatitis C virus; HFV, hemorrhagic fever viruses; HIV, human immunodeficiency virus; ImmPort, Immunology Database and Analysis Portal; KEGG, Kyoto Encyclopedia of Genes and Genomes; LANL, Los Alamos National Laboratory; NCBI, National Center for Biotechnology Information; NMPDR, National Microbial Pathogen Data Resource; PATRIC, Pathosystems Resource Integration Center; ViPR, Virus Pathogen Database and Analysis Resource.

[a] Not an exhaustive list.

for extracting infectious phenotypes such as influenza-like illness [7], unclear human immunodeficiency virus (HIV) status [8], and community-acquired pneumonia [9] have been demonstrated, and consortia such as eMERGE (Electronic Medical Records and Genomics) are standardizing comparison, validation, and deposition of these algorithms into a central repository [10].

The marriage of real-time digital clinical information with "omics" technology creates the opportunity to increase the precision of clinical decision making and challenges us to quickly design and execute bioinformatics analyses. Predictive modeling of infectious disease that incorporates EMR data is still rare, although one recent study generated a social network for hospital-acquired infection from EMR data using recorded contacts between patients and caretakers [11]. Another found that statistical analysis of EMR data produces risk factors for *Clostridium difficile* infection (CDI) that outperform models based only on medically recognized risks [12]. Likely because of the difficulty of integrating data across so many levels, no published studies have yet bridged predictive modeling on EMR data with pathogen genome sequences or other "omics" data from individual patients. Yet, for infectious disease, this is exactly what will fulfill the vision of a rapid-learning health system [13, 14] that converts the informational byproducts of healthcare recorded by practitioners into evidence for future decision making. Whereas EMR data holds details of the clinical process and outcomes, "omics" data tie it back to pathophysiology and the precise strain and host–pathogen interactions present in each patient. Together, they can fuel a "learning engine" that integrates heterogeneous data into new clinical insights, interventions, and therapies. We will discuss how to leverage current bioinformatics software to build such an engine, and how this engine will be able to attack currently insurmountable problems in the field.

THE GENOMIC CLINICAL MICROBIOLOGY LABORATORY

Previous reviews [1, 2] have proposed that cheap sequencing technology will transform clinical microbiology, while acknowledging technical and informational barriers to adoption. Whole-genome sequencing via NGS provides ultimate resolution for epidemiological studies of transmission and relatedness, and may soon be cost-effective for routine use [1, 2]. For pathogen identification, however, NGS is unlikely to usurp robotic culturing systems (eg, Vitek and BD Phoenix) or newer mass spectrometry systems by cost and sensitivity comparisons alone, although it can lower turnaround time for difficult-to-culture organisms and identify novel or rarely seen pathogens [1, 15]. Because susceptibility or resistance of an organism to drugs is in principle fully encoded in its genetic material [2, 16], NGS can also lower turnaround times for drug susceptibility testing of slow-growing organisms, such as *Mycobacterium tuberculosis* [17] and HIV type 1 [18]. This strategy should only expand as fuller catalogs of genomic variants that cause drug resistance are compiled for other pathogenic organisms.

Leveraging Existing Bioinformatics Tools

An oft-mentioned hurdle [1, 2] for widespread use of NGS in clinical microbiology is the lack of readily accessible software for converting these data into species identifications, phylogenies,

and drug susceptibilities. However, many mature open-source bioinformatics solutions for individual components of these problems exist, and connecting these components into a pipeline is therefore a tractable software engineering exercise. Examples for most subtasks are listed in Table 2. As NGS use by clinical microbiology laboratories becomes more commonplace, we might anticipate full-fledged genomic clinical microbiology software packages to become widely available.

This expectation has 3 foreseeable shortcomings. The first is that current tools are tied to centrally curated repositories of evidence. Although proponents of genomic clinical microbiology often envision encyclopedic databases hosted by international consortia [1, 2], human curation is expensive and inefficient at scale, and many infectious diseases are locale-specific phenomena. Models based on pooled data may fail to reflect variation between healthcare delivery regions [19, 20]; for instance, a recent fitness model of H3N2 influenza based on international genomic surveillance data creates predictions only at the resolution of clades spanning multiple continents [21]. Because implementation of NGS in a healthcare institution's microbiology laboratory produces copious sequencing data not easily shared through public databases, institutions should prepare to manage repositories of local evidence and predictive models that work specifically for them. Over time, as data exchange interfaces are developed, institutions could form consortia to generalize analyses, which is a strategy that has successfully increased the power of human genome-wide association studies [22, 23].

Table 2. Selected Published Bioinformatics Software Packages or Databases That Address Specific Steps of Clinical Microbiology Tasks Using Next-Generation Sequencing Data[a]

Problem Domain	Software or Database
Strain typing	• Multilocus sequence typing database
De novo assembly from long reads	• Celera • Hierarchical Genome Assembly Process
Species identification	
From clonal sample	• NCBI BLAST • GenBank • Other genome databases in Table 1
From nonclonal sample	
Meta-assembly	• AMOS • MIRA • MetaVelvet
Clustering and species annotation	• MEGAN • MG-RAST
Maximum likelihood phylogeny trees	• BEAST • RAxML • ClonalFrame • ClonalOrigin
Whole-genome alignment	
For SNP calling	• Mummer • Mugsy
For structural variant calling	• Mauve
Gene annotation	
Bacterial	• GLIMMER • RAST
Drug resistance in bacteria	• ResFinder • ARG-ANNOT
Other	• Influenza Virus Sequence Annotation Tool

Citations for individual databases can be found in the Supplementary Data.

Abbreviations: AMOS, A Modular, Open-Source assembler; ARG-ANNOT, Antibiotic Resistance Gene–ANNOTation; BEAST, Bayesian Evolutionary Analysis Sampling Trees; BLAST, Basic Local Alignment Search Tool; GLIMMER, Gene Locator and Interpolated Markov Modeler; MEGAN, MetaGenome Analyzer Database; MG-RAST, Metagenomics Rapid Annotation Using Subsystem Technology; MIRA, Mimicking Intelligent Read Assembly; NCBI, National Center for Biotechnology Information; RAxML, Randomized Axelerated Maximum Likelihood; RAST, Rapid Annotation using Subsystem Technology; SNP, single-nucleotide polymorphism.

[a] Not an exhaustive list. Well-established tools are available for many specific subtasks.

A second shortcoming is that current pathogen annotation tools primarily make predictions using the simplistic criterion of sequence similarity. Machine learning (ML) algorithms could eventually integrate a wider array of genotypic features extractable from pathogen genomes—variant calls, putative gene and motif annotations, and more—and train holistic models that predict phenotypes. A "top-down," integrative model predicting limited phenotypes from genotyping for *Mycoplasma genitalium* is available [24]; top-down predictions of virulence, however, add the substantial complexity of host interactions. Therefore, genome-wide ML models of virulence have mostly been "bottom-up," blind to mechanistic knowledge, and oriented toward even smaller-genome pathogens with considerable genomic surveillance data. ML on viral sequence features has predicted more effective antiretroviral combinations for HIV [25–27], genetic markers for host selectivity within families of viruses [28], and optimal strain selection for H3N2 influenza vaccines [21]. In general, given the explosion in available data, significant untapped potential remains for ML-based models that predict virulence, transmissibility, and drug resistance from pathogen genotypes.

The third shortcoming is that for many common pathogens, these models are still limited by the paucity of clinical metadata linked to sequenced pathogens. Pathogen phenotypes accessible directly from EMRs include prognostic variables, such as length of stay and disposition, and laboratory results, such as drug susceptibilities. Although laboratory information systems typically do not forward nonclinical results (eg, growth curves) to EMRs, data exported from the laboratory information systems can help define richer phenotypes. For some diseases, EMRs will contain laboratory results that directly reflect infection severity (eg, viral load for hepatitis C virus and HIV) [29], whereas other diseases will require more complex criteria [7, 9, 30]. Natural language

processing of physician notes will facilitate the extraction of complex, high-accuracy clinical phenotypes from the EMR [7, 31]. Routine NGS of specimens and EMR data on drugs prescribed and administered will enable ad hoc studies crossing pathogen genotypes against interventions and outcomes. Richer characterization of particular host–pathogen encounters may be provided by immune and molecular profiling of selected patients, as well as animal experiments that establish individual pathogen genetic associations and molecular mechanisms. Biomarkers derived from such data [5, 6] could enhance predictive models built on a zealous integration of NGS and EMR data.

Increasing EMR phenotype information associated with pathogen genomes will spur a new generation of pathogenicity and risk models based on genomic data. Ideally, these models can drive a "learning engine" that integrates heterogeneous input data from an encounter with an infected patient and predict outcomes for possible interventions. Predictions can be delivered to physicians via clinical decision support systems that complement EMR functions by suggesting relevant actions within a patient's electronic chart. The closing of the EMR–NGS–EMR loop (Figure 1) should be the ultimate goal of bioinformatics pipelines for genomic clinical microbiology, because this would maximize the utility of data created for clinical encounters, continuously turning yesterday's observations and outcomes into evidence for tomorrow's predictions [13, 14].

This sounds ambitious, but we can look to analogous software designed as subcomponents of learning healthcare systems to anticipate likely costs and avenues for development. The i2b2 (Informatics for Integrating Biology and the Bedside) platform [23] and its counterpart SCILHS (Scalable Collaborative Infrastructure for a Learning Health System) [32] are vendor-agnostic solutions for extracting and unifying data across EMRs for reuse in cohort design and robust meta-analysis. The eMERGE consortium stimulated the creation of SHARPn (Strategic Health IT Advanced Research Projects) for normalization and natural language processing of EMR data [33] and CLIP-MERGE (Clinical Implementation of Personalized Medicine Through Electronic Health Records and Genomics) for automated pharmacogenomics alerts [34]. For these examples, working software was created after 1–5 years of development with $100 000–$10 million of annual public grant funding [23, 32–34]. If the aforementioned open-source software is leveraged, an equal scale of public funding and collaboration among academic medical centers could make similar strides toward the proposal in Figure 1. A modular framework allowed i2b2 to expand in scope organically after initial release [23, 32], suggesting that successful strategies should first aim for simple but clinically useful tasks such as identifying species and transmissions while anticipating the addition of more complex analyses via plugins and community contributions. In short, a reasonable investment in scrupulous software engineering could produce the seeds of a learning health system for infectious disease within the decade.

IMPACT ON CLINICAL MANAGEMENT

Three concrete applications of this strategy address urgent global problems in infectious disease. One problem is rising antimicrobial resistance, which the World Health Organization names as one of the 3 greatest threats to human health [35]. Care providers overusing antimicrobials and fomenting resistance in subclinical carriers are partly to blame, with recent studies estimating the fraction of misuse to be between one-quarter and one-half of all treatments [36]. Multidrug resistance increases the morbidity and mortality of healthcare-acquired infections (HAIs), which have an incidence of 1.7 million cases per year in the United States and an estimated annual cost of more than $30 billion [37] that dwarfs the likely cost of any informatics-based preventive efforts. The sobering threat of extensively drug-resistant community-circulating organisms, some of which have therapeutic failure rates of 25%–29% [38], alters the risk analysis for hospital procedures once considered routine and calls for comprehensive new strategies for management.

Identifying High-Risk Patients for HAI

Infection control for HAIs depends on identifying high-risk patients and applying isolation precautions or reducing known risk factors during their hospital course. For CDI, the most frequently reported nosocomial infection in the United States, many questions about how infections are acquired and how to manage at-risk patients remain [39]. The prevailing notion that infections are mostly transmitted person-to-person within hospitals [40] conflicts with recent NGS evidence that sources of infection are more diverse [41], suggesting a greater role for asymptomatic colonized patients and environmental sources.

Each healthcare system represents a unique milieu of person-to-person contact networks, contaminated surfaces, microbiomes, and asymptomatic colonization that contributes to the risk of CDI. Data from EMRs and NGS can prove or disprove transmission between patients and unlock the secrets of modifiable risk factors in this chaotic environment. ML algorithms predicting individual risk of CDI for a large hospital performed better (area under the receiver-operating characteristic curve [AUC] = 0.81) when operating on >10 000 unconstrained EMR variables rather than curated variables for known risk factors [12]. Similar ML models based on EMR data between 2009 and 2014 for The Mount Sinai Hospital in New York City, encompassing 192 000 patients and 1366 CDI diagnoses, show equal performance (AUC = 0.80) and draw out associations not typically published for CDI. These may be unique to Mount Sinai's environment and include respiratory failure (odds ratio [OR], 8.3; 95% confidence interval [CI], 6.6–10.3), nutritional

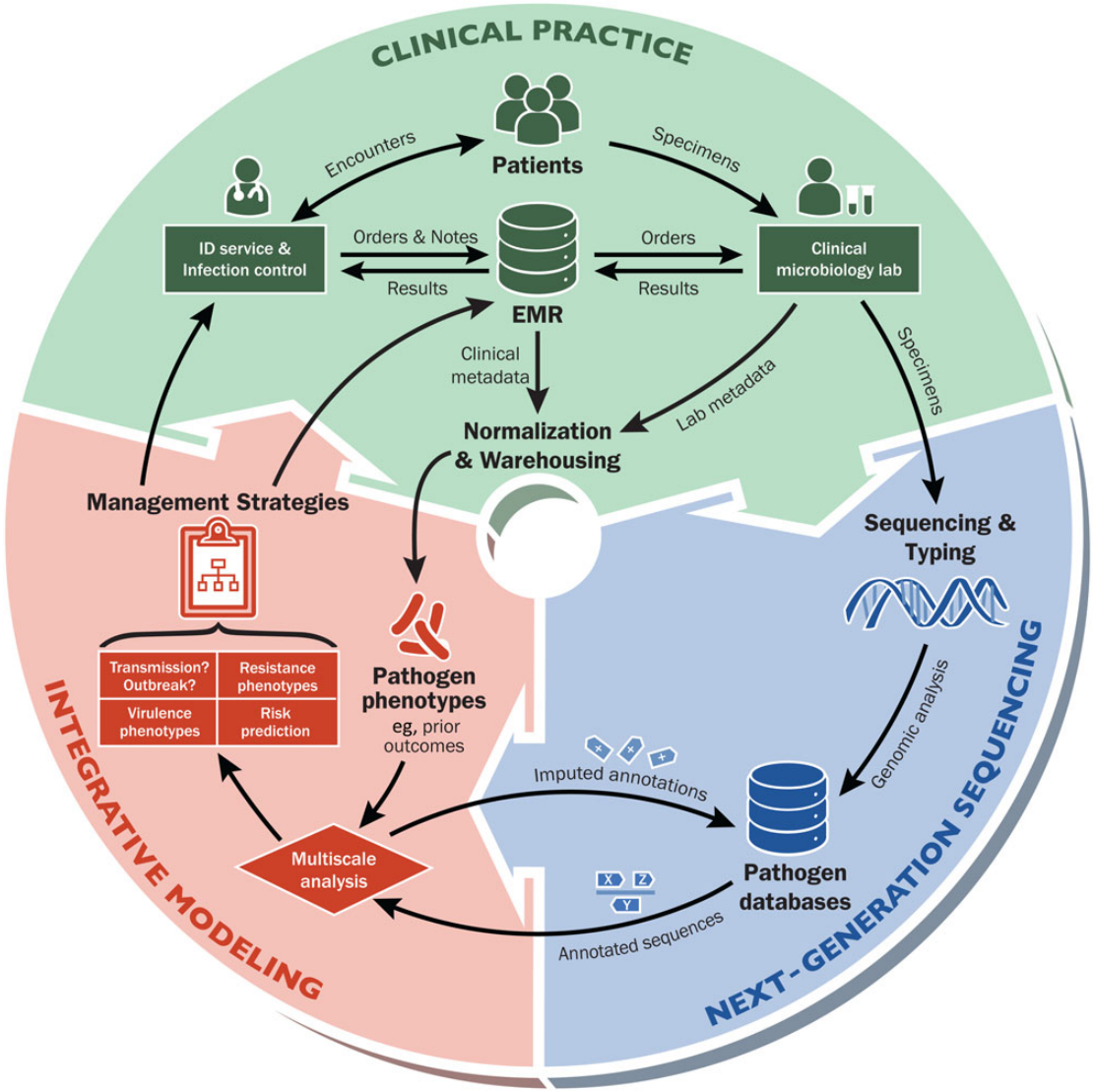

Figure 1. A learning health system for infectious diseases. Next-generation sequencing (NGS) technologies now permit routine genomic analysis of clinical microbiology specimens. When integrated with pathogen phenotypes derived from clinical metadata in electronic medical records (EMRs) and laboratory metadata, we can generate predictive models for pathogen transmission, outbreaks, drug resistance, virulence, and risk factors for infection or critical outcomes that are specific to the health system and its patient population. If management strategies are formulated from these predictions and sent to infectious disease (ID) physicians and hospital infection control, a continuous loop of data analysis, application, and model refinement is created.

irregularity (OR, 6.6; 95% CI, 4.7–8.6), and pancytopenia (OR, 4.4; 95% CI, 3.1–5.5) (Timothy O'Donnell, personal communication).

A model-based decision support system would screen patients with higher CDI or asymptomatic colonization likelihood and allow earlier diagnosis and intervention. NGS-confirmed transmission events and interactions between people and equipment seen in the EMR and other data could extend this basic model to highlight common factors behind verified transmission and inform empiric, real-time modifications of infection control policy. Cross-sectional analysis by NGS-derived phenotypes and risk factors in the EMR would facilitate more precise clinical decision making, for instance, whether shortening patient time in intensive care units or decreasing use of provocative antibiotics would be more preventive within the local milieu. Short of a clinical trial that is probably infeasible to conduct, much less replicate across institutions, there is scant evidence for making these decisions at present, so a localized quantitative model can only help.

Earlier Detection of Outbreaks Inside and Outside the Hospital

Current infection control software suites such as VigiLanz Dynamic Monitoring Suite and TheraDoc Infection Control Assistant primarily issue outbreak alerts based on infection frequency thresholds. This could be rendered obsolete by routine NGS of clinical microbiology specimens, which determines with great precision whether a transmission event has occurred [1, 2]. A software system with access to EMRs and other hospital data could automatically search elements common between verified transmission cases (caregivers, equipment, or rooms) and alert staff to inspect these elements before they produce enough transmissions to trigger a frequency threshold alert. Given enough historical data, NGS could also help hospitals differentiate community- from hospital-acquired infections and thereby refine metrics used to evaluate infection control policies.

An active effort to sample the environment inside and outside the hospital could further extend the reach of this surveillance. Within the hospital, "problem spots" identified by earlier investigations could be resampled regularly via NGS to reevaluate the efficacy of infection control measures. The hospital also samples the pathogen ecosystem of the local population. Hospitals already report diagnoses of highly transmissible and dangerous infections to government authorities, and sharing NGS data for these cases would permit real-time assessment of where pathogens are coming from, how they are evolving, and where populations naive to a pathogen are located. Current mapping and surveillance efforts [42] would be vastly enhanced by rich phylogenetic information, allowing outbreaks across disparate regions to be linked [3, 4, 43]. Fine-grained, real-time tracking of infectious disease spread would better inform doctors diagnosing and treating new patients, field agents tracking cases and contacts, and health policy makers seeking preventive population measures.

Antimicrobial Stewardship

Decision support systems for empiric antibiotic therapy have been investigated for decades [44], but with the prevalence of antimicrobial resistance skyrocketing, the urgency to implement systems that specifically encourage restraint with antibiotics has increased [45]. Selective reporting is a common strategy that directs providers toward optimal therapies simply by omitting names of inappropriate drugs in susceptibility reports [46]. A more aggressive strategy pushes EMR alerts whenever physicians prescribe antibiotic treatment inconsistent with best practices [47].

These solutions ignore the power of the EMR to provide evidence that justifies or improves the antimicrobial stewardship interventions. For instance, although it is well accepted that antibiotic overuse increases the prevalence of resistance, current antimicrobial stewardship programs have demonstrated neither effects on patient outcomes nor even that decreased antibiotic treatment leads to decreased antibiotic resistance [45]. By integrating NGS and EMR data, these hypotheses could be investigated in minute detail within large patient cohorts. NGS can reveal and enumerate the genetic mechanisms of resistance circulating through a health system. By tracing the recurrence of pathogens in the local community, an NGS-equipped health system can determine whether patients receiving antibiotics have generated and transmitted drug-resistant mutants. Specific drug regimens can be correlated with the development of particular resistance mutations. Conversely, given enough longitudinal data, the efforts of an antimicrobial stewardship program can be validated by observing decreased emergence of resistance mutations to drugs prescribed more conservatively.

CONCLUSIONS

Routine access to pathogen genomic data will transform our ability to manage infections, but only if we can integrate this information with clinical and other data to power predictive models for critical outcomes. Assuming that the hurdles of cost, accuracy, and turnaround time can be addressed, which is likely given current trends, NGS will soon become a standard clinical microbiology procedure. The unprecedented specificity of this data will in the near term allow reconstruction of transmission networks inside and outside hospitals. In the far term, having rich clinical data linked to pathogen genotypes will permit predictions of prognosis, virulence, and drug susceptibility for active infections once NGS data are available. Incorporating these capabilities into a new clinical workflow that actively refines predictive models by adjusting to new data (Figure 1) should improve case management, risk prediction for HAIs, detection of outbreaks, and antimicrobial stewardship. The missing link in this transformation, and the goal for bringing it to fruition, is software that leverages best-of-breed existing tools, incorporates all relevant heterogeneous datatypes, builds on electronic phenotyping algorithms to scrub low-accuracy EMR data, and validates against gold standard clinical case review.

Healthcare institutions and researchers should recognize that a potent combination of NGS and EMR data will transform infectious disease management. The threats posed by multidrug resistance and healthcare-associated infections demand a revolution in management strategy. Predictive modeling grounded in rich, diverse molecular and clinical data will dramatically increase the precision of care and help hold these threats at bay.

Supplementary Data

Supplementary materials are available at *Clinical Infectious Diseases* online (http://cid.oxfordjournals.org). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the

sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Acknowledgments. We thank Deena Altman, Shirish Huprikar, and members of the Pathogen Surveillance Program at Mount Sinai for critical suggestions on the manuscript.

Financial support. Both authors were supported by the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai.

Potential conflict of interest. Both authors: No reported conflicts.

Both authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Köser CU, Ellington MJ, Cartwright EJP, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog **2012**; 8:e1002824.
2. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet **2012**; 13:601–12.
3. Chin C-S, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med **2011**; 364:33–42.
4. Rasko DA, Webster DR, Sahl JW, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany. N Engl J Med **2011**; 365:709–17.
5. Mejias A, Ramilo O. Transcriptional profiling in infectious diseases: ready for prime time? J Infect **2014**; 68(suppl 1):S94–9.
6. Querec TD, Akondy RS, Lee EK, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol **2009**; 10:116–25.
7. Silva JC, Shah SC, Rumoro DP, et al. Comparing the accuracy of syndrome surveillance systems in detecting influenza-like illness: GUARDIAN vs. RODS vs. electronic medical record reports. Artif Intell Med **2013**; 59:169–74.
8. Felsen UR, Bellin EY, Cunningham CO, Zingman BS. Development of an electronic medical record-based algorithm to identify patients with unknown HIV status. AIDS Care **2014**; 26:1318–25.
9. DeLisle S, Kim B, Deepak J, et al. Using the electronic medical record to identify community-acquired pneumonia: toward a replicable automated strategy. PLoS One **2013**; 8:e70944.
10. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc **2013**; 20:e206–11.
11. Cusumano-Towner M, Li DY, Tuo S, Krishnan G, Maslove DM. A social network of hospital acquired infection built from electronic medical record data. J Am Med Inform Assoc **2013**; 20:427–34.
12. Wiens J, Campbell W. Learning data-driven patient risk stratification models for *Clostridium difficile*. Open Forum Infect Dis **2014**; 1:1–9.
13. Kohane IS, Drazen JM, Campion EW. A glimpse of the next 100 years in medicine. N Engl J Med **2012**; 367:2538–9.
14. Committee on the Learning Healthcare System in America, Institute of Medicine. Best care at lower cost: the path to continuously learning health care in America. Washington, DC: The National Academies Press, **2014**.
15. Naccache SN, Peggs KS, Mattes FM, et al. Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. Clin Infect Dis **2015**; 60: 919–23.
16. Gordon NC, Price JR, Cole K, et al. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. J Clin Microbiol **2014**; 52:1182–91.
17. Boehme CC, Nabeta P, Hillemann D, et al. Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med **2010**; 363:1005–15.
18. Ram D, Leshkowitz D, Gonzalez D, et al. Evaluation of GS Junior and MiSeq next-generation sequencing technologies as an alternative to Trugene population sequencing in the clinical HIV laboratory. J Virol Methods **2015**; 212:12–6.
19. Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inform Assoc **2014**; 21:699–706.
20. Reis BY, Mandl KD. Integrating syndromic surveillance data across multiple locations: effects on outbreak detection performance. AMIA Annu Symp Proc **2003**.
21. Luksza M, Lässig M. A predictive fitness model for influenza. Nature **2014**; 507:57–61.
22. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med **2013**; 15:761–71.
23. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc **2012**; 19:181–5.
24. Karr JR, Sanghvi JC, Macklin DN, et al. A whole-cell computational model predicts phenotype from genotype. Cell **2012**; 150:389–401.
25. Zazzi M, Kaiser R, Sönnerborg A, et al. Prediction of response to antiretroviral therapy by human experts and by the EuResist data-driven expert system (the EVE study). HIV Med **2010**; 12:211–8.
26. Zazzi M, Incardona F, Rosen-Zvi M, et al. Predicting response to antiretroviral treatment by machine learning: the EuResist Project. Intervirology **2012**; 55:123–7.
27. Lengauer T, Sing T. Bioinformatics-assisted anti-HIV therapy. Nat Rev Microbiol **2006**; 4:790–7.
28. Raj A, Dewar M, Palacios G, Rabadan R, Wiggins CH. Identifying hosts of families of viruses: a machine learning approach. PLoS One **2011**; 6: e27631.
29. Norton B, Naggie S. The clinical management of HCV in the HIV-infected patient. Antivir Ther **2014**; doi:10.3851/IMP2910.
30. Klompas M, Haney G, Church D, Lazarus R, Hou X, Platt R. Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance. PLoS One **2008**; 3:e2626.
31. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ **2015**; 350:h1885.
32. Mandl KD, Kohane IS, McFadden D, et al. Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): architecture. J Am Med Inform Assoc **2014**; 21:615–20.
33. Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform **2012**; 45:763–71.
34. Gottesman O, Scott SA, Ellis SB, et al. The CLIPMERGE PGx Program: clinical implementation of personalized medicine through electronic health records and genomics-pharmacogenomics. Clin Pharmacol Ther **2013**; 94:214–7.
35. Infectious Diseases Society of America. The 10×'20 Initiative: pursuing a global commitment to develop 10 new antibacterial drugs by 2020. Clin Infect Dis **2010**; 50:1081–3.
36. McKellar MR, Fendrick AM. Innovation of novel antibiotics: an economic perspective. Clin Infect Dis **2014**; 59:S104–7.
37. Scott RD. The direct medical costs of healthcare-associated infections in U.S. hospitals and the benefits of prevention, **2009**. Available at: http://www.cdc.gov/hai/pdfs/hai/scott_costpaper.pdf. Accessed 14 July 2015.
38. Hirsch EB, Tam VH. Detection and treatment options for *Klebsiella pneumoniae* carbapenemases (KPCs): an emerging cause of multidrug-resistant infection. J Antimicrob Chemother **2010**; 65:1119–25.
39. Leffler DA, Lamont JT. Clostridium difficile infection. N Engl J Med **2015**; 372:1539–48.
40. Cohen SH, Gerding DN, Johnson S, et al. Clinical practice guidelines for *Clostridium difficile* infection in adults: 2010 update by the Society for Healthcare Epidemiology of America (SHEA) and the Infectious

Diseases Society of America (IDSA). Infect Control Hosp Epidemiol **2010**; 31:431–55.
41. Eyre DW, Cule ML, Wilson DJ, et al. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. N Engl J Med **2013**; 369:1195–205.
42. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med **2008**; 5:e151.
43. McAdam PR, Templeton KE, Edwards GF, et al. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant *Staphylococcus aureus*. Proc Natl Acad Sci U S A **2012**; 109:9107–12.
44. Leibovici L, Gitelman V, Yehezkelli Y, et al. Improving empirical antibiotic treatment: prospective, nonintervention testing of a decision support system. J Intern Med **1997**; 242:395–400.
45. Wagner B, Filice GA, Drekonja D, et al. Antimicrobial stewardship programs in inpatient hospital settings: a systematic review. Infect Control Hosp Epidemiol **2014**; 35:1209–28.
46. Doern CD. Integration of technology into clinical practice. Clin Lab Med **2013**; 33:705–29.
47. Kullar R, Goff DA, Schulz LT, Fox BC, Rose WE. The 'epic' challenge of optimizing antimicrobial stewardship: the role of electronic medical records and technology. Clin Infect Dis **2013**; 57:1005–13.