ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Infectious Disease Research in the Era of Big Data

Peter M. Kasson[1,2]

[1]Department of Biomedical Engineering and Department of Molecular Physiology, University of Virginia, Charlottesville, Virginia 22908, USA; email: kasson@virginia.edu

[2]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, 752 37 Uppsala, Sweden

Annu. Rev. Biomed. Data Sci. 2020. 3:43–59

First published as a Review in Advance on April 7, 2020

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-121219-025722

Keywords

big data, infectious disease, personalized medicine, transmission modeling

Abstract

Infectious disease research spans scales from the molecular to the global—from specific mechanisms of pathogen drug resistance, virulence, and replication to the movement of people, animals, and pathogens around the world. All of these research areas have been impacted by the recent growth of large-scale data sources and data analytics. Some of these advances rely on data or analytic methods that are common to most biomedical data science, while others leverage the unique nature of infectious disease, namely its communicability. This review outlines major research progress in the past few years and highlights some remaining opportunities, focusing on data or methodological approaches particular to infectious disease.

1. WHAT MAKES INFECTIOUS DISEASE SPECIAL?

The key distinguishing factor for infectious disease from an analytic point of view is that much of it is also communicable between hosts, whether via person-to-person spread, zoonotic or epizootic spread, or spread via intermediates in the natural or built environment. Treating such transmission then becomes a critical opportunity in the analysis of infectious disease data that differs from noncommunicable diseases. Transmission can be represented either (*a*) explicitly, as in reconstructions of host-to-host transmission chains to guide public-health interventions (1–3); (*b*) implicitly, as in outbreak predictions and analyses that do not directly predict transmission chains (4, 5); or (*c*) as the underlying spatiotemporal correlation structure to shape predictions of disease likelihood, drug resistance, clinical course, etc. based on related cases (6–8). Here, we focus on the aspects and approaches that distinguish infectious disease research from noncommunicable disease research. There are many important infectious disease studies that follow lines similar to noncommunicable disease studies, such as genome-wide association studies of infectious disease susceptibility (9), but those are left for other reviews.

2. METHODS: MIXING DATA SCIENCE BEST PRACTICES AND PROBLEM-SPECIFIC METHODS

To a large degree, infectious disease research and analytics have followed general methodological trends in the biomedical data science field, for instance, use of logistic regression, decision trees, factor analysis, and, more recently, deep learning approaches (10–14). However, the transmissible nature of infectious disease has led to the development of some methods particular to the field and the increased application of others relative to noncommunicable disease research. Phylogeography is not particular to infectious disease—for example, humans move and have progeny—but it has had particular application in connecting pathogen sequence data with spatiotemporal data to more accurately model infectious disease spread (15, 16). Similarly, temporal networks (graphs that have time-varying edges) have been used to model the transient nature of contacts involved in infectious disease transmission (17). Because pathogens interact with the adaptive immune system, the use of immunologic similarity as a distance metric has provided important insights into the expected efficacy of antigenic memory and has provided a widely used tool to visualize antigenic evolution (18). As richer and denser data become available, we anticipate the development of additional statistical inference methods tuned to the particular features of infectious disease data.

3. NEW TYPES OF DATA

Much recent progress and excitement in infectious disease research have come from the availability of new, often large-scale datasets. Some of these data have proven amenable to analysis with well-established statistical learning methods, while others have inspired the development of new analytic methods or given new applicability to methods that were less frequently utilized. Here, we classify large-scale infectious disease studies according to the following rubric, schematized in **Figure 1**. (*a*) Big computation: Recent advances in computing power have led to the generation of large-scale datasets from computer modeling, often mechanistic in nature, that can be analyzed for insight into infectious disease. Here, we focus again on infectious disease–specific concerns, particularly pathogen fitness and drug resistance. (*b*) Big experiments or large-scale perturbational studies: Methods such as deep mutational scanning and host factor screens have generated large-scale experimental datasets on infectious disease fitness that have yielded mechanistic insights into replication, pathogenesis, and transmissibility. (*c*) Large-scale clinical data: Clinical datasets that are

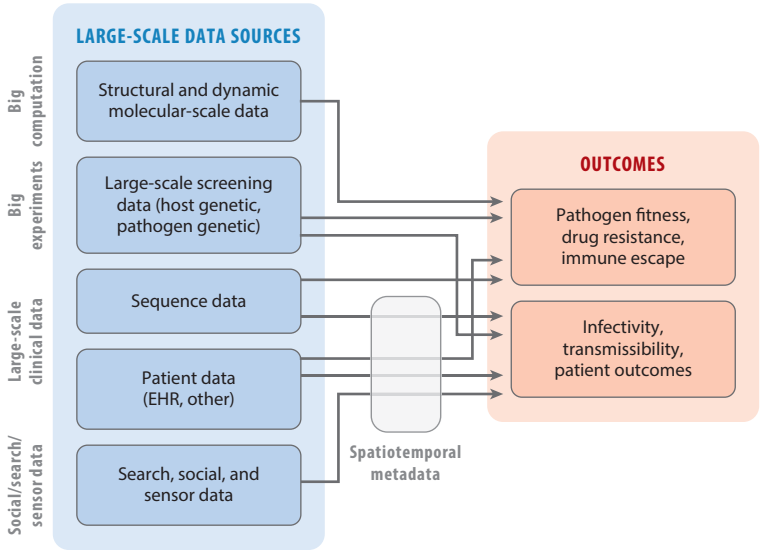

Figure 1

Infectious disease data sources and outcomes measurements. Here, we organize large-scale data sources (*blue*) into big computation (large-scale computational modeling and simulation); big experiments (large, systematic screening data such as host or pathogen genetic screens); large-scale clinical data including both surveillance and diagnostic sequencing data, as well as electronic health record (EHR) and in-hospital sensor data; and informal out-of-hospital data sources such as search, social, and sensor data. These can be used in models that are then validated (*red*) against measures of pathogen fitness, such as drug resistance and immune escape, or of patient or animal outcomes, such as infectivity, transmissibility, and patient health status. The communicable nature of infectious disease means that spatiotemporal metadata are of immense utility in interpreting sequence, patient, and social data.

either large in that they involve many patients or very deep in terms of information per sample have been a major area of development. This includes deep sequencing of isolates to infer transmission chains, disease spread, and other data, as well as use of more traditional hospital laboratory and electronic health record (EHR) data. An exciting new area in this regard is the use of in-hospital sensor data. (*d*) Social, search, and sensor data: Data from social media, search, and other Internet activity, as well as sensor data such as from mobile phones, provide a rich and largely collateral source of information on infectious disease outbreaks and can be used to inform and monitor public health interventions. This expanding area also raises new ethical questions about user/patient privacy and public health.

While not exhaustive, this categorization is designed to be illustrative and informative as to the new opportunities and outstanding challenges for large-scale data analysis in infectious disease research. We treat each of these categories in turn, discussing the analytic opportunities and challenges associated with each and touching on some major methods used and insights gained.

4. BIG COMPUTATION: LARGE-SCALE SIMULATION DATASETS

The continued increase in computational capacity has enabled simulation not only of larger problems but also of larger parameter spaces, whether of mutation, transmission, or otherwise. As a result, simulation results have themselves become datasets for statistical learning. One example is the simulation of >100 mutants of a bacterial drug-resistance enzyme. Predictions from the

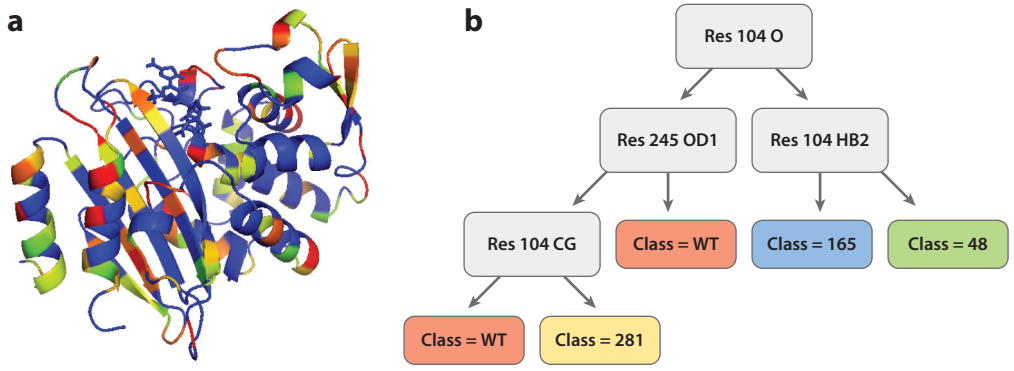

Figure 2

Use of structural and dynamic data to predict drug resistance. (*a*) Large-scale simulations were performed of 125 mutants of a CTX-M9 bacterial β-lactamase, and mutations were scored by their predicted ability to increase drug resistance (*blue*, low; *red*, high). (*b*) Decision trees were then trained to predict which structural features in the validated mutants transmitted the mutations to the drug resistance enzyme active site. The decision tree shows residues and atoms in the active site the positions of which reported mutations at allosteric sites [wild type (WT), L48, T165, S281]. This example shows how large-scale computational models can inform mechanistic understanding. Data replotted from Reference 19.

simulation are then correlated to experimental phenotypic or enzymatic assays, and statistical learning is performed on the simulation results to identify not only the mutant but also the particular structural features modulating drug resistance in the experimental assays (19). Analogous work that is large scale in simulating protein conformational motions but smaller scale in mutations has also yielded insight into new ligand-binding mutations in antibiotic resistance (20). These are interesting analytic problems because the feature space includes a rich set of structural and dynamic outputs of the simulations that can be used to extract mechanistic insights. Statistical methods used include mutual information feature selection, decision trees, and time series Markov modeling approaches developed for such simulations (20, 21). Studies to date have primarily focused on either active site mutations or allosteric mutations controlling drug resistance (**Figure 2**) as a path toward predicting not-yet-observed resistance mechanisms and including such mechanisms in new drug development efforts to improve robustness against mutational resistance. Since the simulations represent substantial computational effort, there is much room for further innovation in the development of statistical learning approaches to systematically mine on a large scale these computationally expensive datasets for insight. Other large-scale experimental datasets that could be well paired with simulation include deep mutational scanning and similar large-scale mutational studies (22–25). Analyzing simulation datasets matched with experimental or clinical outcomes is particularly ripe for additional insights and is relatively underexplored.

Modeling disease transmission, evolution, and immunity has also advanced in richness beyond traditional SEIR (susceptible/exposed/infected/recovered) models. Large-scale models incorporating pathogen genetic drift, pathogen evolution, travel and contact data derived from air and ground transport data, and even immunity (26–31) offer a greatly increased expressive capacity compared to simple earlier models. The critical questions, of course, are whether enough data are available to constrain such complex models and to rigorously evaluate them, as well as what insights can be extracted. Some examples include fitting spatially resolved and genetic drift–capable agent-based models against global H3 or H1 influenza data (32) to test relationships between global migration and antigenic drift. Similar models have been used to test the impact of immunity on pathogen evolution (33), helping resolve fundamental questions in the dynamics and evolution of different pathogens. As more large-scale datasets become available for parameterization, fitting,

and evaluation, we expect the utility of such models to increase both in understanding emergent biological phenomena and in guiding policy decisions such as vaccination or antimicrobial use strategies.

5. BIG EXPERIMENTS: LARGE-SCALE PERTURBATIONAL STUDIES

Recent advances in high-throughput experiments have enabled studies of a comprehensiveness not previously possible to examine determinants of pathogen fitness, virulence, and other key observables. Broadly speaking, these have been made possible by technological advances in either high-throughput perturbation or high-throughput measurement and consist of systematic sweeps of pathogen genetic or chemical changes, host changes, or large and complex observational datasets, particularly on human clinical samples. We discuss some outstanding examples of each of these below.

5.1. Pathogen Genetic Screens

Systematic genetic perturbations have long been a major tool in microbiology, exemplified by the Keio collection of nonessential gene knockouts in *Escherichia coli* (34). Recent progress has enabled more rapid generation of such genetic libraries for different pathogens and their assessment in both cell culture and, to a more limited extent, animal models of infection (35, 36). Plasmid-based and phase-based approaches enable generation of large variant libraries in bacteria, but the resulting bacteria are frequently perturbed from an infection biology standpoint in more ways than simply the target genetic variant. For viruses, the small genome size and availability of reverse-genetic systems to efficiently produce progeny virions from genomic material have facilitated deep mutational scanning analyses, where the fitness impact of all single point mutations in a viral gene can be quantitatively assessed under multiple replicative conditions (35). In influenza and more recently other viruses, this has enabled systematic prospective mapping of the replicative fitness landscape for key viral genes (**Figure 3**), as well as the effects on immune escape (35, 37–39). This level of resolution is theoretically feasible although technically more challenging in other pathogens, offering a rich data source to analyze genetic and protein fitness landscapes.

5.2. Host Genetic Screens

The development of screens and libraries to systematically perturb host genetics has provided another key capability in the study of infectious disease. At least within cell culture models, insertional mutagenesis in human haploid cell lines, RNA interference screens, and more recently CRISPR-based screens have permitted large-scale testing of host factors required for infection. These screens are often complicated by cellular and screen heterogeneity that can cause low gene-to-gene concordance levels between assays (40), but improved analytic methodologies have resulted in substantial new insights, identifying both known and novel entry factors, other essential genes for pathogen replication, and host-restriction factors (41–45). Such approaches have been critical in identifying novel factors either positively or negatively involved in the replication of viruses such as Ebola, influenza, and dengue. Analogous interference or single-gene-reconstitution experiments have identified key permissive or restrictive host factors for *Listeria* and *Coxiella* infection (46–48). These assays obviously apply most directly to intracellular pathogens, but the extension of such approaches to other models of infection can yield insights into the host genetics of other diseases as well. These screening approaches provide some of the best unbiased methods to probe genetic effects on infection in cell culture models; they are thus a good complement to more targeted, hypothesis-driven approaches.

Figure 3

An experimental map of replicative fitness for all single–amino acid mutations of influenza hemagglutinin [genetic background A/WSN/1933(H1N1)]. Letters represent amino acids, and the size of each letter is proportional to the fitness of the corresponding variant. Plot adapted with permission from Reference 35.

Large-scale perturbational analyses that are not specific to infectious disease have also generated important datasets and yielded fundamental insights, at least in cell culture systems. Examples of these are phenotypic screens, combining chemical screens with genetic screens to probe drug–gene interactions, and leveraging protein–protein interaction or drug–metabolome interaction data (49, 50). Other techniques that are not explicitly perturbational and not particular to infectious disease, such as high-dimensional cell profiling, high-content imaging, and high-dimensional profiling of innate and adaptive immune responses in human clinical samples, also yield large, often complex datasets with the potential for great insight into infectious disease and host response (50–53).

6. LARGE-SCALE CLINICAL DATA

Recent years have seen huge strides in the availability and accessibility of clinical and surveillance data for analysis. Route availability of pathogen whole-genome sequencing has permitted both deeper analyses of pathogen genetic features and much more precise measures of spread and transmission; hospital laboratory and telemetry data also provide a rich source for analysis of infectious disease spread, drug resistance, and patient prognostic factors. Finally, innovative use of sensor data in the clinical setting has made possible what were previously painstaking epidemiological studies on a much larger-scale, systematic, and routine basis.

6.1. Whole-Genome Pathogen Sequencing

As in many areas of biology, the increased accessibility of whole-genome pathogen sequencing has had a tremendous impact on infectious disease research. One clear metric of its utility can be seen in the public health arena, where multiple public health agencies have transitioned to routine whole-genome sequencing of all positive samples for several diseases (54–56). This is facilitated by the smaller genome size of most pathogen genomes than those of higher eukaryotes, but it is also driven by the utility to epidemiological research, infection control, and, increasingly, direct clinical utility.

Because many pathogens exhibit substantial genetic variation, whole-genome sequencing has found direct application in determining transmission chains—using genomic and spatiotemporal data to infer either direct transmission or closely related transmission in an outbreak as opposed to multiple unrelated introductions of a pathogen. This is often performed via phylogenetic clustering. However, the choice of algorithm for such clustering depends on the desired purpose. For instance, different algorithms produce superior results for monitoring pathogen clade/strain evolution or for modeling transmission (57). Some of this methodological divergence also derives from the typically sparse nature of sequence sampling; it is possible that given arbitrarily many pathogen sequences per space-time unit, these problems would be reduced (1, 2). Alternate approaches utilize Bayesian estimates of spatial probabilities in conjunction with phylogenetic tree reconstruction to model phylogeographic spread—movement of sequences in space and time (15, 58, 59). Insight from these approaches has been applied at both the global scale—examining worldwide patterns of viral genetic drift (32, 60)—and the local scale—transmission patterns in individual communities and households (61, 62). Surprisingly, not all spatiotemporally correlated infection events are clonally related; work of this nature has shown that there are many independent introductions of influenza into a region and that multiple co-occurring infections even within a single household can sometimes represent independent infection events (62).

Pathogens such as RNA viruses display high rates of mutation and thus substantial genetic variation within a single infection of a single host. Aided by large-scale deep sequencing, this has

permitted both clinical and experimental studies of how host-to-host transmission functions—how stringent the transmission bottlenecks are, how severe the founder effects are upon infection of a new host, and how stereotyped the process of diversification is within a host. Although initially disparate estimates of transmission bottlenecks for influenza seem to be converging on a range of 1–4 distinct genomes per human-to-human transmission (62–64), additional modifiers potentially include biological variables such as strain, route of transmission, and whether the virus is well adapted to both hosts. This work addresses a set of critical questions for infection biology: For a given pathogen, how great is the within-host genetic diversity and how strict is the transmission bottleneck and resulting founder effect in between-host transmission? The variation in number of genomes transmitted with regard to biological variables likely reveals important information about barriers to and efficiency of viral infection under different conditions. The similarity or divergence of within-host genetic populations in consecutive hosts illuminates the rate of stochastic drift and the stringency of selection (33, 61, 62, 65, 66). Analogous studies have also yielded data on the within-host diversification and transmission of bacterial infections, which can be expected to function differently from RNA respiratory viruses (67, 68).

In addition to host-to-host and host-to-environment movement of complete pathogen genomes, bacterial pathogens also contain mobile genetic elements, both plasmid and chromosomally integrated. Antibiotic and toxin resistance are frequently encoded on these mobile elements, and the increase in whole-genome sequencing of patient and environmental isolates in hospitals has led to an awareness that drug resistance can be spread on relatively short timescales in a fashion that is neither proliferation of a particular resistant strain nor fixation of a mutation but instead interspecies transmission of a mobile genetic element (69–71). One example of this was a multispecies outbreak of carbapenem-resistant *Enterobacteriaceal* infections where resistance-mediating plasmids spread across genetically disparate bacterial strains and multiple species (72). This was appreciated much more readily because of the availability of large-scale whole-genome sequencing and then motivated more traditional epidemiological and infection-control investigations to identify reservoirs for this exchange and institute interventions to reduce the spread of resistant organisms to patients (3). Although a large number of studies have developed methods to predict drug resistance from pathogen genomic sequences, the slower turnaround time of whole-genome sequencing (and plasmidome sequencing for bacteria) relative to drug susceptibility testing means that this has not entered common clinical practice, despite good results from prospective testing (73). One early exception was *Mycobacterium tuberculosis*, where slow growth of the organism and therefore long waiting times for susceptibility testing led some public health authorities to institute routine whole-genome sequencing for susceptibility testing (54); this is now being implemented for a greater number of notifiable pathogens tracked by national centers (56).

6.2. Analysis of Laboratory and Telemetry Data

With widespread adoption of EHRs, there is an opportunity to use EHR data to improve infectious disease detection and care. We delineate three categories of such use: infectious disease surveillance, quality-of-care improvement, and personalized medicine or systems designed to improve clinical care of an individual patient. As discussed below, the first two of these areas are more developed than the third. We also note that, although EHRs have become common, not all healthcare systems have highly portable records between healthcare providers, so much work has concentrated either on a single healthcare setting, on using insurance data in non-single-payer systems, or on using single-payer EHR data when available.

ICD-9 (International Classification of Diseases, Ninth Revision) codes, ICD-10 codes, and their modifications are the most straightforward way to parse EHR data. Not surprisingly, this has been the starting point for many data-analytic approaches, including syndromic surveillance

for outbreak detection. One example is a large-scale deployment of ICD-10-based surveillance for acute respiratory infection in Germany (74). Given the limitations of ICD coding, natural language processing of narrative text entry fields in the health record is a richer area for processing. Some early large-scale deployments of syndromic surveillance used manual coding of free text (4), but this has largely been replaced in more recent work by natural language processing approaches (75, 76). However, provider and facility variability in both narrative text coding and laboratory data complicates large-scale use of such methods across many hospitals (75). Despite these challenges and the historic use of more manual methods such as sentinel sites and patient questionnaires, there remains substantial interest and value in reliable, automated syndromic surveillance for outbreak detection. Despite the deployment of some functional systems, there remains a substantial gap between validation of retrospective analyses and the routine use of real-time automated syndromic surveillance from EHR data.

Infection-control and quality-of-care analyses utilize much of the same analysis as outbreak detection but have somewhat relaxed immediacy requirements. Obviously, interrupting or preventing transmission of healthcare-associated infections is of the greatest benefit to patients, but systematically identifying clusters and spurring earlier and more systematic follow-up and corrective action are of substantial value to hospitals and benefit to patients. The attention to infection-control analytics stems from two factors: First, since some EHR data are less immediate (discharge diagnosis, insurance claims, etc.), permitting a time lag before analysis improves data quality. Second, in many healthcare systems including the United States, hospitals have a strong financial incentive to reduce healthcare-associated infections. There has thus been substantial scholarship and methodological innovation in this area. Careful traditional epidemiology has been a cornerstone of infection-control and quality-of-care improvements; the hope is that more advanced analyses will further enhance the utility of such work. We highlight the work of Wiens and colleagues in developing logistic regression predictors for *Clostridium difficile* infection in individual patients based on automatic EHR data extraction that can readily be specialized to the data features available in a given hospital and outperform curated predictors (77, 78). However, to our knowledge, such tools have primarily been used in a retrospective manner, while maximal benefit will be derived from prospective use for risk stratification and intervention.

Predicting individual patient risk for healthcare-associated infections straddles the line between infection-control measures designed to learn from patient outcomes to systematically improve care for future patients and analyses designed to improve care for the patient at hand. This latter category, effectively comprising decision support systems based on EHR data in the area of infectious disease, represents an area of personalized medicine and a substantial opportunity for real-time data analytics. HIV risk has been a particularly active area for EHR-based analytics, particularly in developing better targeted HIV screening (79) and better targeting of pre-exposure HIV prophylaxis (80). These have used a mix of manual feature coding, natural language processing for ingestion of free text, and structured data extraction from EHRs.

Such work has primarily treated infectious disease risk as a static phenomenon, learning features associated with risk, rather than leveraging the spatiotemporal correlation imposed by disease transmissibility. We anticipate that further research combining baseline static predictors with localized risk estimates will further improve the accuracy and utility of these methods. The feasibility and utility of such transmission-aware models will depend on sampling density: The more patients there are in a transmission chain who have EHR data available, the greater the utility will be for models that incorporate explicit or implicit transmission.

In addition to infectious disease transmission, much work has been devoted to predicting patient decompensation due to infection, namely sepsis prediction in an inpatient setting (81–85). Most work in this regard leverages both hospital laboratory and telemetry data to predict

sepsis before it is clinically evident, giving providers time to respond. Initial work on improved patient-specific alarms from telemetry data (81) has matured into more event-specific learning. As would be expected, initial research used retrospective prediction (82, 85), but some systems have been placed into prospective testing with good results or ongoing trials (83, 84) (**http://www.clinicaltrials.gov/** identifiers NCT03655626 and NCT03960203). This represents an exciting example of clinical deployment of decision support systems to improve patient outcomes.

6.3. Other Sensor Data in a Healthcare Setting

In-hospital sensor data have proven to be another rich source for infectious disease transmission studies. Despite the growth of transmitting devices in routine use (the so-called Internet of Things), most studies have explicitly placed transmitting tags on individuals, locations, or objects and analyzed the resulting data to help connect contact patterns to disease transmission and inform recommendations for improved infection-control practices (86). Such data also provide useful adjuncts to traditional microbiological data, permitting higher-resolution time series and spatial modeling of contact networks in conjunction with other information (87). As discussed below, similar approaches have been taken outside the healthcare setting, particularly with informal data sources. Here, we distinguish explicitly interventional studies in a controlled healthcare environment from the use of informal data sources in uncontrolled environments.

6.4. Summary

Large-scale clinical data offer a variety of feature-rich information for analyzing pathogen transmission, dynamics, drug resistance, and clinical outcomes. Progress in generating and ingesting datasets for clinical whole-genome sequencing, patient records, and the like has already enabled substantial discoveries, such as better estimates of transmission bottlenecks, pathogen evolution, and unexpected reservoirs for drug-resistant organisms. The use of such analyses in real-time systems aimed at improving clinical care for individual patients is an area for future growth—there has been exciting progress in recent years, but most systems are deployed either retrospectively or as time-limited research projects rather than being in routine clinical use. The latter represents an exciting frontier, and the increasing availability of all kinds of healthcare data—sequencing, laboratory, physician-entered, and sensor data—provides rich material for improved research on infectious disease and the development of more sophisticated predictive and analytic models. Because of the transmissibility of infectious disease, models that incorporate online learning based on recent data are particularly useful in this arena compared to noncommunicable diseases. In general, real-time decision support models will have to be rigorously validated and shown to be efficacious, easy to use, and cost effective in improving patient care.

7. SOCIAL, SEARCH, AND SENSOR DATA

Informal data sources, particularly from social media, search queries, and incidental sensor data outside the healthcare setting, present an important yet challenging area for infectious disease analytics. These data are high frequency and often high coverage, yet can display correlated noise in ways that are difficult to model. For tasks such as outbreak surveillance and transmission modeling, the frequency and coverage of informal sources help overcome weaknesses of formal data sources that can be lower frequency, involve time delays, or have much lower sampling coverage.

7.1. Search Data

The canonical example of using search data for infectious disease detection was Google Flu Trends (later Google Disease Trends, including dengue) (88). This service used signals from Google

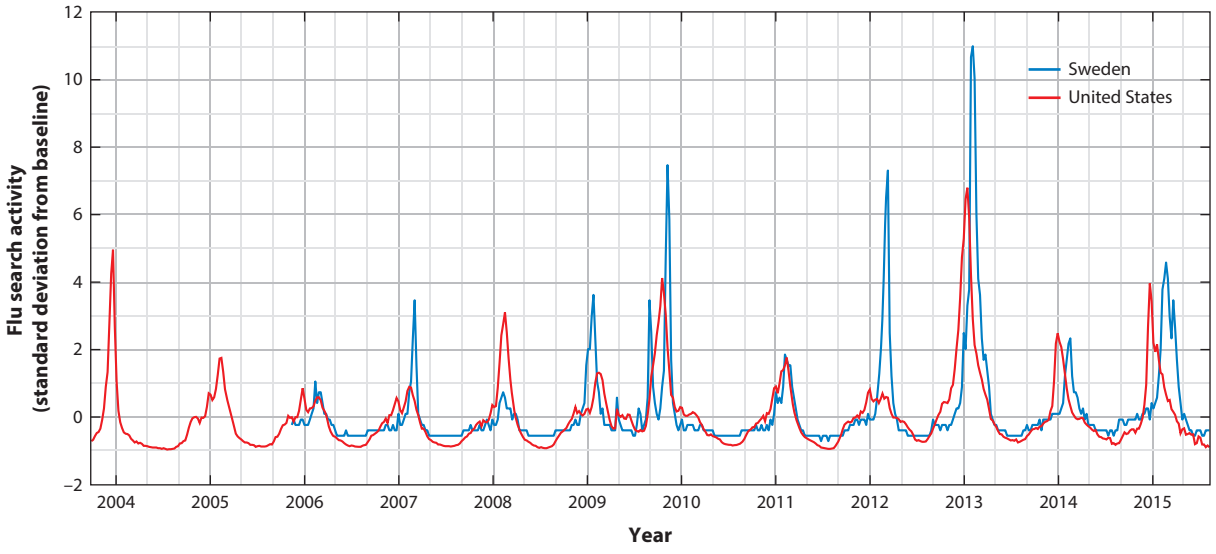

Figure 4

Google Flu Trends predictions for several influenza seasons plotted for the United States and Sweden. Data from Google Flu Trends (**http://www.google.org/flutrends**), used with permission.

search data to estimate influenza-like illness incidence in near-real time (**Figure 4**), providing information much more rapidly than the authoritative estimates from the Centers for Disease Control and Prevention (CDC). These predictions later showed some systematic overestimation versus CDC estimates and have been criticized by some as an example of big tech hubris (89). Nonetheless, more recent third-party evaluations of the dengue data and reanalyses of the influenza data with new statistical models have shown utility for some situations (90, 91), and Google Disease Trends opened the use of search and social data for infectious disease surveillance. Canada, for instance, routinely uses an automated alert system based primarily on news data with ongoing efforts to expand to social media data (92). Similarly, HealthMap is a prominent monitoring system using traditional news sources (93).

7.2. Social Media Data

The use of social media data brings both new possibilities and new concerns, particularly surrounding the use of individually identifiable data that may be private or semiprivate. Perhaps for that reason, most work on outbreak detection using social media has focused on public Twitter data. In the wake of Google Flu Trends, several papers analyzed Twitter data for influenza-like illness outbreak detection or dengue detection (94–96). More recent work has tended to be multimodal, integrating social media data with other signals, perhaps recognizing that such adjunctive approaches yield the greatest accuracy at this time (5). Simply using social media data as geospatial signals with assorted content/contextual analysis is perhaps most straightforward (97–99). However, explicitly taking into account social network structure to analyze risk and potential spread of infectious diseases is an area that is less well explored, substantially more complex, potentially richer as a novel data source, and more fraught with ethical considerations. Clearly, online social network structure does not trivially map to infectious disease transmission risk, as many individuals have online contacts who are not physically proximal, but social media information likely contains rich signals about risk profiles and transmission of contagious diseases.

The legal and ethical ramifications of such information have been explored more extensively for physical contacts and risk information in the context of, for instance, public health follow-ups of notifiable infectious diseases (100–102), but even there the legal framework varies substantially by jurisdiction, and online social network information contains increased technical, legal, and ethical complexities (103). This will be a challenge for the field in future years.

7.3. Sensor Data

In addition to formal sensor placement in the healthcare setting, informal sensor data in uncontrolled environments constitute rich data sources for infectious disease studies. With high mobile phone uptake in much of the world, many individuals now carry mobile location sensors. Similarly, transmitting devices built into other objects provide a wealth of signals that could be analyzed for infectious disease research. In addition to the privacy implications of such work, there are analytic challenges posed by confounders in such data (104). However, they have been utilized for out-of-hospital retrospective or observational studies of disease risk or epidemic analysis (105–108). Despite a set of associated confounders, the increasing availability of sensor and locational data constitutes a potentially fruitful frontier in monitoring potential community transmission of infectious diseases and providing individualized risk assessments in the future.

8. OUTLOOK

From a statistical viewpoint, increased model complexity brings greater expressivity but requires more extensive training and validation data. Recent increases in the scope of infectious disease datasets have enabled the use of more expressive statistical models and thus greater scientific insight. Data sampling density is particularly critical for infectious disease because of two factors: transmissibility and immune- or treatment-mediated selection. In both of these cases, sparsely sampled data simply do not permit accurate modeling of transmission, immune escape, or drug resistance because the key links between distant data points are not resolved. Even underdetermined problems can be constrained with large-scale computational sampling, and progress has been made on several fronts in that regard, but large-scale datasets are becoming transformative in enabling definitive answers to questions of how pathogens are transmitted around the world, acquire drug resistance, and escape the immune system. We anticipate the availability of such datasets to result in both further methodological developments and fundamental advances in biological insights. In the clinical arena, increasing real-time availability of EHR data, in-hospital sensor data, and informal sources such as search, social, and mobile phone data provide both greatly increased data volumes and the potential to deploy systems that give real-time, patient-specific risk assessments and recommendations. This is an incredibly exciting front, but one that will require careful validation, harmonization across disparate technical and legal environments, and navigation of privacy concerns both inside and outside the hospital. Ongoing clinical trials, as well as out-of-hospital sensor projects such as Verily's Project Baseline, will help researchers explore potential approaches to this promising but tricky area, and the field will learn from their successes and failures. In all of these areas, infectious disease continues to be a productive discipline for advanced data analytics, and its unique features reward methods development as well as the application of current state-of-the-art techniques.[1]

[1]This article was written before the emergence of COVID-19. The scientific and public-health response to the pandemic has illuminated the capabilities and challenges discussed in the article in the context of other pathogens and once again illustrates the power of big data approaches to infectious disease research.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The author would like to thank Kim Branson, Sourav Haldar, Nenad Macesic, and Ricardo Ferreira for helpful discussions. Jesse Bloom kindly permitted reproduction of a figure from his work. P.M.K. was supported by a Wallenberg Academy Fellowship.

LITERATURE CITED

1. Worby CJ, O'Neill PD, Kypraios T, Robotham JV, De Angelis D, et al. 2016. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Ann. Appl. Stat.* 10:395–417
2. Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. 2017. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Comput. Biol.* 13:e1005495
3. Mathers AJ, Vegesana K, German Mesner I, Barry KE, Pannone A, et al. 2018. Intensive care unit wastewater interventions to prevent transmission of multispecies *Klebsiella pneumoniae* carbapenemase–producing organisms. *Clin. Infect. Dis.* 67:171–78
4. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. 2004. Syndromic surveillance in public health practice, New York City. *Emerg. Infect. Dis.* 10:858–64
5. McGough SF, Brownstein JS, Hawkins JB, Santillana M. 2017. Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. *PLOS Negl. Trop. Dis.* 11:e0005295
6. Hagenaars TJ, Donnelly CA, Ferguson NM. 2004. Spatial heterogeneity and the persistence of infectious diseases. *J. Theor. Biol.* 229(3):349–59
7. Parham PE, Ferguson NM. 2006. Space and contact networks: capturing the locality of disease transmission. *J. R. Soc. Interface* 3:483–93
8. Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. 2006. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* 312:447–51
9. Newport MJ, Finan C. 2011. Genome-wide association studies and susceptibility to infectious diseases. *Brief. Funct. Genom.* 10:98–107
10. Guo P, Liu T, Zhang Q, Wang L, Xiao J, et al. 2017. Developing a dengue forecast model using machine learning: a case study in China. *PLOS Negl. Trop. Dis.* 11:e0005973
11. Banerjee I, Yamauchi Y, Helenius A, Horvath P. 2013. High-content analysis of sequential events during the early phase of influenza A virus infection. *PLOS ONE* 8:e68450
12. Gomes AL, Wee LJ, Khan AM, Gil LH, Marques ET Jr., et al. 2010. Classification of dengue fever patients based on gene expression data using support vector machines. *PLOS ONE* 5:e11267
13. Liang ZH, Powell A, Ersoy I, Poostchi M, Silamut K, et al. 2016. CNN-based image analysis for malaria diagnosis. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, ed. T Tian, Y Wang, Q Jiang, X Hu, Y Liu, et al., pp. 493–96. New York: IEEE
14. Macesic N, Polubriaginof F, Tatonetti NP. 2017. Machine learning: novel bioinformatics approaches for combating antimicrobial resistance. *Curr. Opin. Infect. Dis.* 30:511–17
15. Bielejec F, Rambaut A, Suchard MA, Lemey P. 2011. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics* 27:2910–12
16. Holmes EC. 2008. Evolutionary history and phylogeography of human viruses. *Annu. Rev. Microbiol.* 62:307–28
17. Masuda N, Holme P. 2013. Predicting and controlling infectious disease epidemics using temporal networks. *F1000Prime Rep.* 5:6
18. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM. 2004. Mapping the antigenic and genetic evolution of influenza virus. *Science* 305(5682):371–76

19. Latallo MJ, Cortina GA, Faham S, Nakamoto RK, Kasson PM. 2017. Predicting allosteric mutants that increase activity of a major antibiotic resistance enzyme. *Chem. Sci.* 8:6484–92
20. Hart KM, Ho CMW, Dutta S, Gross ML, Bowman GR. 2016. Modelling proteins' hidden conformations to predict antibiotic resistance. *Nat. Commun.* 7:12965
21. Cortina GA, Hays JM, Kasson PM. 2018. Conformational intermediate that controls KPC-2 catalysis and beta-lactam drug resistance. *ACS Catal.* 8:2741–47
22. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* 31:1581–92
23. Deng Z, Huang W, Bakkalbasi E, Brown NG, Adamski CJ, et al. 2012. Deep sequencing of systematic combinatorial libraries reveals beta-lactamase sequence constraints at high resolution. *J. Mol. Biol.* 424:150–67
24. Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, et al. 2013. Capturing the mutational landscape of the beta-lactamase TEM-1. *PNAS* 110:13067–72
25. Romero PA, Tran TM, Abate AR. 2015. Dissecting enzyme function with microfluidic-based deep mutational scanning. *PNAS* 112:7159–64
26. Bedford T, Cobey S, Beerli P, Pascual M. 2010. Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLOS Pathog.* 6:e1000918
27. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, et al. 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 546:401–5
28. Kuhnert D, Stadler T, Vaughan TG, Drummond AJ. 2016. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Mol. Biol. Evol.* 33:2102–16
29. Luo S, Perelson AS. 2015. Competitive exclusion by autologous antibodies can prevent broad HIV-1 antibodies from arising. *PNAS* 112:11654–59
30. Smith AM, Adler FR, Ribeiro RM, Gutenkunst RN, McAuley JL, et al. 2013. Kinetics of coinfection with influenza A virus and *Streptococcus pneumoniae*. *PLOS Pathog.* 9:e1003238
31. Marceau V, Noel PA, Hebert-Dufresne L, Allard A, Dube LJ. 2011. Modeling the dynamical interaction between epidemics on overlay networks. *Phys. Rev. E* 84:026105
32. Bedford T, Riley S, Barr IG, Broor S, Chadha M, et al. 2015. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* 523:217–20
33. Han AX, Maurer-Stroh S, Russell CA. 2019. Individual immune selection pressure has limited impact on seasonal influenza virus evolution. *Nat. Ecol. Evol.* 3:302–11
34. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, et al. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* 2:2006.0008
35. Thyagarajan B, Bloom JD. 2014. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife* 3:e03300
36. Varble A, Albrecht RA, Backes S, Crumiller M, Bouvier NM, et al. 2014. Influenza A virus transmission bottlenecks are defined by infection route and recipient host. *Cell Host Microbe* 16:691–700
37. Haddox HK, Dingens AS, Bloom JD. 2016. Experimental estimation of the effects of all amino-acid mutations to HIV's envelope protein on viral replication in cell culture. *PLOS Pathog.* 12:e1006114
38. Setoh YX, Amarilla AA, Peng NYG, Griffiths RE, Carrera J, et al. 2019. Determinants of Zika virus host tropism uncovered by deep mutational scanning. *Nat. Microbiol.* 4:876–87
39. Dingens AS, Arenz D, Weight H, Overbaugh J, Bloom JD. 2019. An antigenic atlas of HIV-1 escape from broadly neutralizing antibodies distinguishes functional and structural epitopes. *Immunity* 50:520–32.e3
40. Snijder B, Sacher R, Ramo P, Liberali P, Mench K, et al. 2012. Single-cell analysis of population context advances RNAi screening at multiple levels. *Mol. Syst. Biol.* 8:579
41. Carette JE, Raaben M, Wong AC, Herbert AS, Obernosterer G, et al. 2011. Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. *Nature* 477:340–43
42. Carette JE, Guimaraes CP, Varadarajan M, Park AS, Wuethrich I, et al. 2009. Haploid genetic screens in human cells identify host factors used by pathogens. *Science* 326:1231–35
43. Marceau CD, Puschnik AS, Majzoub K, Ooi YS, Brewer SM, et al. 2016. Genetic dissection of *Flaviviridae* host factors through genome-scale CRISPR screens. *Nature* 535:159–63

44. Krishnan MN, Ng A, Sukumaran B, Gilfoy FD, Uchil PD, et al. 2008. RNA interference screen for human genes associated with West Nile virus infection. *Nature* 455:242–45
45. Sessions OM, Barrows NJ, Souza-Neto JA, Robinson TJ, Hershey CL, et al. 2009. Discovery of insect and human dengue virus host factors. *Nature* 458:1047–50
46. Perelman SS, Abrams ME, Eitson JL, Chen D, Jimenez A, et al. 2016. Cell-based screen identifies human interferon-stimulated regulators of *Listeria monocytogenes* infection. *PLOS Pathog*. 12:e1006102
47. McDonough JA, Newton HJ, Klum S, Swiss R, Agaisse H, Roy CR. 2013. Host pathways important for *Coxiella burnetii* infection revealed by genome-wide RNA interference screening. *mBio* 4:e00606-12
48. Kuhbacher A, Emmenlauer M, Ramo P, Kafai N, Dehio C, et al. 2015. Genome-wide siRNA screen identifies complementary signaling pathways involved in *Listeria* infection and reveals different actin nucleation mechanisms during *Listeria* cell invasion and actin comet tail formation. *mBio* 6:e00598-15
49. El Zahed SS, Brown ED. 2018. Chemical-chemical combinations map uncharted interactions in *Escherichia coli* under nutrient stress. *iScience* 2:168–81
50. Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, et al. 2011. Phenotypic landscape of a bacterial cell. *Cell* 144:143–56
51. Pegoraro G, Misteli T. 2017. High-throughput imaging for the discovery of cellular mechanisms of disease. *Trends Genet*. 33:604–15
52. Roy Chowdhury R, Vallania F, Yang Q, Lopez Angel CJ, Darboe F, et al. 2018. A multi-cohort study of the immune factors associated with *M. tuberculosis* infection outcomes. *Nature* 560:644–48
53. Lee J, Boutz DR, Chromikova V, Joyce MG, Vollmers C, et al. 2016. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nat. Med.* 22:1456–64
54. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, et al. 2015. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect. Dis.* 15:1193–202
55. Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, et al. 2016. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir. Med.* 4:49–58
56. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, et al. 2016. Implementation of nationwide real-time whole-genome sequencing to enhance *Listeriosis* outbreak detection and investigation. *Clin. Infect. Dis.* 63:380–86
57. Han A, Parker E, Maurer-Stroh S, Russell C. 2018. Inferring putative transmission clusters with Phydelity. bioRxiv 477653. **https://doi.org/10.1101/477653**
58. Bielejec F, Lemey P, Baele G, Rambaut A, Suchard MA. 2014. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Syst. Biol.* 63:493–504
59. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchene S, Fourment M, et al. 2019. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* 15:e1006650
60. Langat P, Raghwani J, Dudas G, Bowden TA, Edwards S, et al. 2017. Genome-wide evolutionary dynamics of influenza B viruses on a global scale. *PLOS Pathog*. 13:e1006749
61. Poon LL, Song T, Rosenfeld R, Lin X, Rogers MB, et al. 2016. Quantifying influenza virus diversity and transmission in humans. *Nat. Genet.* 48:195–200
62. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, Lauring AS. 2018. Stochastic processes constrain the within and between host evolution of influenza virus. *eLife* 7:e35962
63. Poon LLM, Song T, Wentworth DE, Holmes EC, Greenbaum BD, et al. 2019. Reply to 'Reconciling disparate estimates of viral genetic diversity during human influenza infections'. *Nat. Genet.* 51:1301–3
64. Xue KS, Bloom JD. 2019. Reconciling disparate estimates of viral genetic diversity during human influenza infections. *Nat. Genet.* 51:1298–301
65. Wilker PR, Dinis JM, Starrett G, Imai M, Hatta M, et al. 2013. Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. *Nat. Commun.* 4:2636
66. Iqbal M, Xiao H, Baillie G, Warry A, Essen SC, et al. 2009. Within-host variation of avian influenza viruses. *Philos. Trans. R. Soc. Lond. B* 364:2739–47

67. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, et al. 2013. Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. *PLOS ONE* 8:e61319
68. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016. Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* 14:150–62
69. Stokes HW, Gillings MR. 2011. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS Microbiol. Rev.* 35:790–819
70. Zhang T, Zhang XX, Ye L. 2011. Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. *PLOS ONE* 6:e26041
71. Mathers AJ, Cox HL, Kitchel B, Bonatti H, Brassinga AK, et al. 2011. Molecular dissection of an outbreak of carbapenem-resistant *Enterobacteriaceae* reveals intergenus KPC carbapenemase transmission through a promiscuous plasmid. *mBio* 2:e00204-11
72. Mathers AJ, Stoesser N, Sheppard AE, Pankhurst L, Giess A, et al. 2015. *Klebsiella pneumoniae* carbapenemase (KPC)-producing *K. pneumoniae* at a single institution: insights into endemicity from whole-genome sequencing. *Antimicrob. Agents Chemother.* 59:1656–63
73. Roach DJ, Burton JN, Lee C, Stackhouse B, Butler-Wu SM, et al. 2015. A year of infection in the intensive care unit: prospective whole genome sequencing of bacterial clinical isolates reveals cryptic transmissions and novel microbiota. *PLOS Genet.* 11:e1005413
74. Kopke K, Prahm K, Buda S, Haas W. 2016. Evaluation einer ICD-10-basierten elektronischen Surveillance akuter respiratorischer Erkrankungen ($SEED^{ARE}$) in Deutschland [Evaluation of an ICD-10-based electronic surveillance of acute respiratory infections ($SEED^{ARI}$) in Germany]. *Bundesgesundheitsblatt Gesundheitsforsch. Gesundheitsschutz* 59:1484–91
75. Hripcsak G, Soulakis ND, Li L, Morrison FP, Lai AM, et al. 2009. Syndromic surveillance using ambulatory electronic health records. *J. Am. Med. Inform. Assoc.* 16:354–61
76. Ye Y, Tsui FR, Wagner M, Espino JU, Li Q. 2014. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. *J. Am. Med. Inform. Assoc.* 21:815–23
77. Wiens J, Campbell WN, Franklin ES, Guttag JV, Horvitz E. 2014. Learning data-driven patient risk stratification models for *Clostridium difficile*. *Open Forum Infect. Dis.* 1:ofu045
78. Wiens J, Guttag J, Horvitz E. 2014. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *J. Am. Med. Inform. Assoc.* 21:699–706
79. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. 2018. Using clinical notes and natural language processing for automated HIV risk assessment. *J. Acquir. Immune Defic. Syndr.* 77:160–66
80. Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE. 2019. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. *Lancet HIV* 6:E688–95
81. Zhang Y, Szolovits P. 2008. Patient-specific learning in real time for adaptive monitoring in critical care. *J. Biomed. Inform.* 41:452–60
82. Henry KE, Hager DN, Pronovost PJ, Saria S. 2015. A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.* 7:299ra122
83. Futoma J, Hariharan S, Sendak M, Brajer N, Clement M, et al. 2017. An improved multi-output Gaussian process RNN with real-time validation for early sepsis detection. arXiv:1708.05894 [stat.ML]
84. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. 2017. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir. Res.* 4:e000234
85. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, et al. 2016. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med. Inform.* 4:e28
86. Hornbeck T, Naylor D, Segre AM, Thomas G, Herman T, Polgreen PM. 2012. Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections. *J. Infect. Dis.* 206:1549–57
87. Voirin N, Payet C, Barrat A, Cattuto C, Khanafer N, et al. 2015. Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care hospital. *Infect. Control Hosp. Epidemiol.* 36:254–60

88. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457:1012–14
89. Lazer D, Kennedy R, King G, Vespignani A. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343:1203–5
90. Santillana M, Zhang DW, Althouse BM, Ayers JW. 2014. What can digital disease detection learn from (an external revision to) Google Flu Trends? *Am. J. Prev. Med.* 47:341–47
91. Klembczyk JJ, Jalalpour M, Levin S, Washington RE, Pines JM, et al. 2016. Google Flu Trends spatial variability validated against emergency department influenza-related visits. *J. Med. Internet Res.* 18:e175
92. Dion M, AbdelMalik P, Mawudeku A. 2015. Big data and the Global Public Health Intelligence Network (GPHIN). *Can. Commun. Dis. Rep.* 41:209–14
93. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J. Am. Med. Inform. Assoc.* 15:150–57
94. Signorini A, Segre AM, Polgreen PM. 2011. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLOS ONE* 6:e19467
95. Aramaki E, Maskawa S, Morita M. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1568–76. Stroudsburg, PA: Assoc. Comput. Linguist.
96. Broniatowski DA, Paul MJ, Dredze M. 2013. National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. *PLOS ONE* 8:e83672
97. Padmanabhan A, Wang SW, Cao GF, Hwang M, Zhang ZH, et al. 2014. FluMapper: a cyberGIS application for interactive analysis of massive location-based social media. *Concurr. Comput. Pract. Exp.* 26:2253–65
98. Adam NR, Shafiq B, Staffin R. 2012. Spatial computing and social media in the context of disaster management. *IEEE Intell. Syst.* 27:90–97
99. Middleton SE, Middleton L, Modafferi S. 2014. Real-time crisis mapping of natural disasters using social media. *IEEE Intell. Syst.* 29:9–17
100. Boehmer TK, Patnaik JL, Burnite SJ, Ghosh TS, Gershman K, Vogt RL. 2011. Use of hospital discharge data to evaluate notifiable disease reporting to Colorado's Electronic Disease Reporting System. *Public Health Rep.* 126:100–6
101. Keramarou M, Evans MR. 2012. Completeness of infectious disease notification in the United Kingdom: a systematic review. *J. Infect.* 64:555–64
102. Adams DA, Thomas KR, Jajosky RA, Foster L, Baroi G, et al. 2017. Summary of notifiable infectious diseases and conditions—United States, 2015. *MMWR Morb. Mortal. Wkly. Rep.* 64:1–143
103. Purtle J, Field RI, Hipper T, Nash-Arott J, Chernak E, Buehler JW. 2018. The impact of law on syndromic disease surveillance implementation. *J. Public Health Manag. Pract.* 24:9–17
104. Lee EC, Asher JM, Goldlust S, Kraemer JD, Lawson AB, Bansal S. 2016. Mind the scales: harnessing spatial big data for infectious disease surveillance and inference. *J. Infect. Dis.* 214:S409–13
105. Wesolowski A, Metcalf CJ, Eagle N, Kombich J, Grenfell BT, et al. 2015. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *PNAS* 112:11114–19
106. Buckee CO, Wesolowski A, Eagle NN, Hansen E, Snow RW. 2013. Mobile phones and malaria: modeling human and parasite travel. *Travel Med. Infect. Dis.* 11:15–22
107. Bengtsson L, Gaudart J, Lu X, Moore S, Wetter E, et al. 2015. Using mobile phone data to predict the spatial spread of cholera. *Sci. Rep.* 5:8923
108. Paz-Soldan VA, Stoddard ST, Vazquez-Prokopec G, Morrison AC, Elder JP, et al. 2010. Assessing and maximizing the acceptability of global positioning system device use for studying the role of human movement in dengue virus transmission in Iquitos, Peru. *Am. J. Trop. Med. Hyg.* 82:723–30