Infection Control & Hospital Epidemiology (2019), **40**, 314–319
doi:10.1017/ice.2018.343

Original Article

Automated data mining of the electronic health record for investigation of healthcare-associated outbreaks

Alexander J. Sundermann MPH, CIC[1,2], James K. Miller PhD[3], Jane W. Marsh PhD[1], Melissa I. Saul MS[4], Kathleen A. Shutt MS[1], Marissa Pacey[1], Mustapha M. Mustapha MBBS, PhD[1], Ashley Ayres BS, CIC[2], A. William Pasculle ScD[5], Jieshi Chen MS[3], Graham M. Snyder MD, SM[2], Artur W. Dubrawski PhD[3] and Lee H. Harrison MD[1]

[1]The Microbial Genomic Epidemiology Laboratory, Infectious Diseases Epidemiology Research Unit, University of Pittsburgh School of Medicine and Graduate School of Public Health, Pittsburgh, Pennsylvania, [2]Department of Infection Prevention and Control, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, [3]Anton Laboratory, Carnegie Mellon University, Pittsburgh, Pennsylvania, [4]Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania and [5]Department of Pathology, University of Pittsburgh, Pittsburgh, Pennsylvania

Abstract

Background: Identifying routes of transmission among hospitalized patients during a healthcare-associated outbreak can be tedious, particularly among patients with complex hospital stays and multiple exposures. Data mining of the electronic health record (EHR) has the potential to rapidly identify common exposures among patients suspected of being part of an outbreak.

Methods: We retrospectively analyzed 9 hospital outbreaks that occurred during 2011–2016 and that had previously been characterized both according to transmission route and by molecular characterization of the bacterial isolates. We determined (1) the ability of data mining of the EHR to identify the correct route of transmission, (2) how early the correct route was identified during the timeline of the outbreak, and (3) how many cases in the outbreaks could have been prevented had the system been running in real time.

Results: Correct routes were identified for all outbreaks at the second patient, except for one outbreak involving >1 transmission route that was detected at the eighth patient. Up to 40 or 34 infections (78% or 66% of possible preventable infections, respectively) could have been prevented if data mining had been implemented in real time, assuming the initiation of an effective intervention within 7 or 14 days of identification of the transmission route, respectively.

Conclusions: Data mining of the EHR was accurate for identifying routes of transmission among patients who were part of the outbreak. Prospective validation of this approach using routine whole-genome sequencing and data mining of the EHR for both outbreak detection and route attribution is ongoing.

(Received 16 September 2018; accepted 3 December 2018)

Healthcare-associated outbreaks caused by serious bacterial pathogens cause substantial morbidity and mortality and add to healthcare costs.[1,2] Detection of outbreaks can be difficult in large hospitals where bacterial transmission may go unnoticed for prolonged periods.[3] Investigation and control of a hospital outbreak requires the identification of the route of transmission among patients suspected of being part of the outbreak so that intervention measures can be implemented. This task can be labor intensive for outbreaks that involve complex patients who have long stays, multiple transfers within the hospital, and multiple procedures. Multiple transmission routes responsible for hospital outbreaks have included transmission from environmental contamination; colonized healthcare personnel; medical procedures using contaminated devices; and contaminated medications, solutions, or other medical therapies.[4,5]

Widespread availability of the electronic health record (EHR) offers the potential to use automated data mining tools to find common exposures among hospitalized patients during outbreak investigations. Many epidemiologically relevant variables are readily available in the EHR, including patient location in the hospital, procedures performed, therapies received, and contact with individual healthcare personnel. Data mining, the process of identifying patterns in large data sets, has the potential to be useful for identifying common exposures in the EHR during hospital outbreak investigations. Furthermore, whole-genome sequencing (WGS) has become increasingly available; this method discriminates pathogens at the genetic level.[6–8] Genomic data from patient bacterial isolates have the potential to aid in the data mining and outbreak investigation process.[9] We are developing a surveillance system called Enhanced Detection System for Healthcare-Associated Transmission (EDS-HAT); it will combine prospective WGS surveillance of clinical isolates of key hospital-associated bacterial pathogens

Author for correspondence: Lee H. Harrison, Email: lharriso@pitt.edu

*The affiliations have been updated since original publication. An erratum notice detailing this change was also published (DOI: 10.1017/ice.2019.84).

Cite this article: Sundermann AJ, *et al.* (2019). Automated data mining of the electronic health record for investigation of healthcare-associated outbreaks. *Infection Control & Hospital Epidemiology*, 40: 314–319, https://doi.org/10.1017/ice.2018.343

© 2019 by The Society for Healthcare Epidemiology of America. All rights reserved.

https://doi.org/10.1017/ice.2018.343 Published online by Cambridge University Press

followed by prospective data mining of the EHR to rapidly identify potential outbreaks and their routes of transmission. The purpose of this study was to develop and validate data mining tools using retrospective outbreaks, with the ultimate goal of utilizing these tools as a component of EDS-HAT.

Methods

Study setting

This study was conducted at the University of Pittsburgh Medical Center-Presbyterian Hospital (UPMC), an adult medical and surgical tertiary-care hospital with 762 total beds, 150 critical care unit beds, >32,000 yearly inpatient admissions, and >400 solid organ transplants per year. The UPMC eRecord EHR system has >29,000 active users, including >5,000 physicians affiliated with UPMC, and it comprises >3.6 million unique electronic patient records. UPMC uses Cerner Millenium PowerChart (Cerner, Kansas City, MO) and EpicCare (Epic Systems, Madison WI) as the backbone of its inpatient and outpatient EHR systems, respectively. The University of Pittsburgh Institutional Review Board approved this study.

Characterization of retrospective outbreaks from 2011 to 2016

The subject of this study were outbreaks that occurred during 2011–2016 and that had been previously characterized using molecular typing and traditional epidemiologic methods to identify the transmission route. The routine infection prevention practice at the time was to notify the Microbial Genomic Epidemiology Laboratory (MiGEL) of suspected outbreaks caused by bacterial pathogens so that molecular subtyping could be performed. For each patient suspected of being included in the outbreak, the bacterial isolate was obtained from the clinical microbiology laboratory. For *Clostridium difficile*, which is diagnosed at our institution by culture-independent diagnostic testing, the nucleic acid amplification test–positive stool specimen was cultured for *C. difficile*. For identification of the common exposure responsible for individual outbreaks, our infection prevention team analyzed the health records of patients included in the outbreak to identify the responsible routes of transmission (eg, shared locations/staff, shared procedures/operations, or shared medications). Some outbreak investigations utilized environmental or device cultures to confirm the route of transmission. The transmission route defined by infection prevention was used as the gold standard for comparison with transmission routes identified by the data mining algorithm.

During the study period, our primary method for molecular characterization of bacterial isolates was pulsed-field gel electrophoresis (PFGE). To be considered part of the outbreak, patient isolates had to have 85% band similarity by PFGE. In 2016, WGS replaced PFGE. Based on our experience using WGS for hospital outbreaks, a cutoff of ≤20 single-nucleotide polymorphisms (SNPs) was used to define genetically related patient isolates.

Extraction and processing of EHR data for data mining

All inpatient, emergency room, and same-day surgery encounters between January 1, 2011, and December 31, 2016, were identified through an EHR data repository that accepts data from Cerner and EpicCare as full-text medical records and integrates information from central transcription, laboratory, pharmacy, finance, administrative, and other departmental databases.[10] For each encounter, we obtained microbiology reports and charge transactions from the data repository. To maintain patient confidentiality, a research database was established in which each patient was assigned a study identification number (ID) using De-ID software (De-ID Data, Philadelphia, PA). Criteria for exemption from informed consent by the university's institutional review board were met.

Charge transaction data are in a data repository as charge codes that contain patient account numbers, dates of service, cost centers, transaction codes, charge quantities, and charge amounts. Charge codes include any medication, procedure, location, or other billable item in our hospital and therefore encompass a large number of possible hospital exposures for each patient. Multiple charge codes can represent exposure to a single instrument; therefore, charge codes for key procedures (eg, endoscopic retrograde cholangiopancreatography [ERCP] and bronchoscopy) were collapsed into a single variable group that represented that exposure. For example, ERCP has 8 Current Procedure Terminology codes: 43260, ERCP diagnostic including collection of specimens; 43261, ERCP with biopsy; 43278, ERCP with ablation of tumors, etc. All were combined into a single variable called "ERCP," although each charge code was also analyzed individually.

Data mining of the electronic health record (EHR)

The data mining program was designed using a case-control approach based upon the genotyping results using patient EHR data unrelated to the outbreaks as controls.[11] Case patients were defined as those who had clinical isolates with the same strain by PFGE or WGS, as defined above. Controls were patients who were hospitalized during the 30 days before the case patients' culture date and did not test positive for the genetically related bacterial species. Hospital exposures were then compared for cases and controls.

The data mining program was run on all 9 previously identified outbreaks identified by the infection prevention team at our institution during 2011–2016 to determine the sensitivity of the algorithm for identifying the correct transmission route. The transmission route was deemed to be correct if the route was ranked in the 3 most likely routes of transmission and/or had odds ratios >1 with significant *P* values. Preventable infections were calculated based upon a hypothetical 7- or 14-day infection prevention intervention from the date of the positive culture assuming that the data mining program had been running in real time and that effective interventions were enacted (eg, removal of contaminated equipment, disinfection of environment, and/or enhanced precautions). Outbreaks were deemed nonpreventable if there were only 2 isolates in the outbreak.

We scored possible common routes of transmission within an outbreak according to the formula:

$$S = a\,\ln(a/r) + (r-a)\,\ln(1-a/r) - a\,\ln(\gamma),$$

where a is the number of case patients exposed, r is the number of patients exposed overall (ie, case patients who are part of the outbreak and control patients who are not), and γ is a parameter that balances the positive and negative evidence ($\gamma = 1e{-}4$). For a given set of case patients, each patient can be said to have been infected through the hypothesized common route or by intermediate transmission (ie, via transmission from another case patient). If we take θ to be the unknown probability a patient becomes infected upon exposure to the hypothetical route and γ to be the probability a patient is infected by intermediate transmission (ie, by some other means such as patient-to-patient transmission), then the likelihood of observing a particular set of case patients is proportional to $\theta^b(1-\theta)^{r-b}\gamma^{r-b}$, where b is the number of case patients

https://doi.org/10.1017/ice.2018.343 Published online by Cambridge University Press

Table 1. Characteristics of Outbreaks[a]

No.	Date	Organism	Cluster: No. of Related Isolates	Molecular Typing Method	Duration of Transmission, Days	Transmission Route	Infections Potentially Prevented: 7-Day Intervention	Infections Potentially Prevented: 14-Day Intervention
1	Feb 12	*Acinetobacter baumannii*	3	PFGE	19	Trauma ICU	1	1
2	Mar 13	*Klebsiella pneumoniae*	A: 28	PFGE	865	ERCP	20	20
			B: 2		3	ERCP	0[b]	0[b]
			C: 2		13	ERCP	0[b]	0[b]
			Total: 32					
3	Jun 15	*K. pneumoniae*	7	PFGE	29	Bronchoscope	5	3
4	Jul 15	*Pseudomonas aeruginosa*	8	PFGE	42	Bronchoscope	5	4
5	Aug 15	*A. baumannii*	5	PFGE	80	Medical ICU	3	2
6	Dec 15	*P. putida*	3	PFGE	1	Bronchoscope	0	0
7	Apr 16	*K. pneumoniae*	9	PFGE & WGS	39	Gastroscope	5	3
8	Jun 16	*Clostridium difficile*	A: 2	WGS	4	Trauma Floor	0[b]	0[b]
			B: 2		15	Post Anesthesia Unit	0[b]	0[b]
			C: 2		35	Pulmonology Floor	0[b]	0[b]
			Total: 6					
9	Sep 16	*C. difficile*	4	WGS	67	Medical ICU	1	1
							Total: 40	Total: 34

Note. PFGE, pulsed-field gel electrophoresis; WGS, whole-genome sequencing; ICU, intensive care unit; ERCP, endoscopic retrograde cholangiopancreatography.
[a]The correct transmission route was identified by the data mining program for all outbreaks.
[b]Only 2 isolates; cannot prevent any infections.

infected by the route. We arrive at the formula S by maximizing this expression in θ, which occurs at $\theta = b/r$, and b, which occurs at either 0 or a. Because $b = 0$ is a degenerate solution, it is disregarded. The final score is the log of this maximum likelihood.

The score above represents a nonnormalized log-likelihood. Because it is not normalized, it is suitable for ranking routes but is not comparable across time as the number of case patient changes. Therefore, we estimated an extreme value statistic as the probability a route would score at least as highly as its observed score under the assumption that the case patients were uniformly randomly sampled (the null hypothesis). This P value is estimated numerically using importance sampling from the observed data.

Researchers were initially blinded to the true routes of transmission in this analysis. However, it became clear during the development of the approach that this significantly reduced our ability to identify and correct data processing and modeling problems. For example, the charge codes for gastroscope procedures initially were not properly extracted and grouped from the EHR and therefore could not be identified in the analysis. On review, the correct charge codes for procedures using gastroscopes were grouped together as described above. The analysis was rerun, and the correct exposure route was identified.

Results

The characteristics of the 9 outbreak investigations during the study period are shown in Table 1. For some investigations, molecular typing revealed several separate clusters. For example, for investigation no. 2, there were 2 clusters involving 2 isolates each. For 2 *C. difficile* outbreaks (nos. 8 and 9; 22%), epidemiologic investigation revealed that transmission occurred in the nursing units where the patients resided. Furthermore, 3 investigations (33%) involved *Klebsiella pneumoniae*, 1 of which represented a polyclonal ERCP-related outbreak (no. 2),[3] and 1 each involved bronchoscopy (no. 3) and gastroscopy (no. 7).[12] In addition, 2 *Acinetobacter baumanni* investigations were determined to have been transmitted in intensive care units (nos. 1 and 5). One outbreak each of *Pseudomonas aeruginosa* (no. 4) and *P. putida* (no. 6) were also considered to involve bronchoscopy as the source.

All but 1 outbreak had ranges of 2–9 case patients with control populations of ~4,000 to 56,000 patients (data not shown), depending on the duration of the outbreak. The remaining outbreak (no. 2) had 28 case patients with 130,000 control patients. Overall, outbreaks had 185 average charge transactions per patient, 2 average room transactions per patient, and 263,777 total possible routes of transmission explored by the data mining program.

The data mining program detected the correct routes of transmission on the eighth patient of the ERCP outbreak and all other previous outbreaks on the second positive isolate of each outbreak's respective timeline. For example, for investigation no. 4, *Pseudomonas aeruginosa* transmission related to a bronchoscope, the bronchoscopy procedure was detected in 100% of cases from case 2 to case 6 (OR, 138; $P = .02$) on the second case (Fig. 1). Figure 2 shows investigation no. 3, *Klebsiella pneumoniae* transmission related to a bronchoscope. The bronchoscope is persistently ranked the highest plausible transmission route starting at the second patient (OR = 140; $P = .021$). Table 1 displays the transmission routes that were both determined independently by infection prevention and by the data mining program.

Potential infections prevented are shown in Table 1 based upon a 7- or 14-day intervention period given the delay in plausible intervention with real-time WGS and data mining analysis. In

https://doi.org/10.1017/ice.2018.343 Published online by Cambridge University Press

(a) Transmission Route: Bronchoscopy

Days since first case	Cases (cumulative)	% Cases exposed	% Controls exposed	Rank	p-value	OR*	95% Confidence Interval
18	2	100.0%	3.5%	13	2.20E-02	138	(7, 2884)
22	3	100.0%	3.5%	1	7.70E-04	192	(10, 3732)
30	4	100.0%	3.6%	1	3.20E-05	241	(13, 4486)
37	5	100.0%	3.7%	1	1.20E-06	284	(16, 5143)
39	6	100.0%	3.7%	1	3.80E-08	337	(19, 5995)
41	7	85.7%	3.7%	2	1.70E-07	158	(19, 1314)
43	8	87.5%	3.7%	2	5.80E-09	181	(22, 1478)

*0.5 was added to each cell for comparisons that had a zero cell

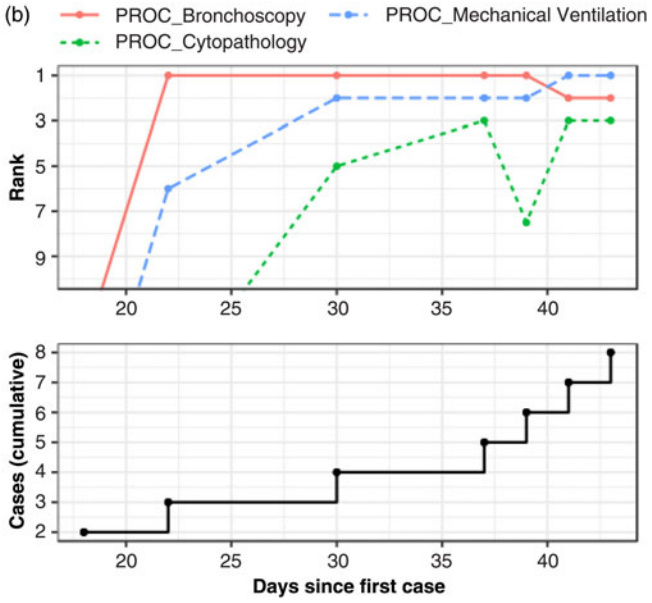

Fig. 1. Transmission route ranking for outbreak no. 4: *Pseudomonas aeruginosa* from a contaminated bronchoscope. Panel A: Results for bronchoscopy, showing timing of cases, the proportion of cases who were found in the EHR to have been exposed to bronchoscopy before illness onset, the percent of the control population that was found in the EHR to have been exposed to bronchoscopy, the score, the ranking relative to other exposures, the *P* value and odds ratio. Panel B: Graphical depiction of relative ranking of bronchoscopy and 2 other ranking exposures (top figure) and the cumulative number of cases (bottom figure), both over time.

total, for the 2011–2016 outbreak requests, potentially 40 or 34 infections (78% or 66% of possible preventable infections, respectively) could have been prevented based on intervention implementation at 7 or 14 days.

Discussion

In this study, data mining of the EHR correctly identified transmission routes by the eighth patient of 1 outbreak and the second patient in all other outbreaks studied. If run in conjunction with routine molecular typing, up to 40 infections (78% of possible preventable infections) could have been prevented assuming that proper intervention had occurred. To our knowledge, this is the first reported study that combines molecular typing results and automated data mining of the EHR in a hospital outbreak setting to identify routes of bacterial transmission. Our results provide proof of concept that automated data mining can correctly identify routes of exposure in hospital outbreak investigations.

Automated data mining has several potential advantages over traditional approaches to hospital outbreak investigations. First, the EHR can be rapidly scanned for common exposures among patients with complex hospitalizations. Second, automated data mining allows rapid assessment of the strength of association of suspected exposures. In this study, we incorporated a case-control study design to identify outbreak transmission routes, which is similar to the approach that is used in traditional outbreak investigations. We are currently refining this approach to allow the infection preventionist to easily select and explore the most appropriate control population within the hospital. For example, to identify the route of transmission during an outbreak that occurs on a single nursing unit, the most appropriate control population may be nonoutbreak patients on the same unit. Both approaches have the potential to substantially decrease the number of hours required for outbreak investigations and to allow infection prevention personnel with limited outbreak investigation expertise to conduct relatively sophisticated investigations.

Our study and approach have several limitations. First, only outbreaks that had been detected by traditional epidemiologic approaches were included. This limitation could have resulted in missing other patients with genetically related isolates who should have been included as cases, thus leading to both an underestimate of the magnitude of the outbreak and having the patients incorrectly included in our control population. Despite this limitation, data mining still identified the correct transmission routes. We anticipate that we can largely overcome this limitation in the future by implementing WGS surveillance of key hospital pathogens. Second, the intervention delay of 7 or 14 days was based on hypothetical timelines that considered the time required to perform WGS, analyze data, and enact interventions (eg, removing a device from use, targeted environmental cleaning, and/or staff education). Regardless, a conservative delay of 14 days for effective interventions still demonstrated 34 potential infections prevented across a relatively small number of outbreaks. Third, we did not include

https://doi.org/10.1017/ice.2018.343 Published online by Cambridge University Press

(a) Transmission Route: Bronchoscopy

Days since first case	Cases (cumulative)	% Cases exposed	% Controls exposed	Rank	p-value	OR*	95% Confidence Interval
11	2	100.0%	3.4%	1	2.10E-02	140	(7, 2930)
19	3	100.0%	3.6%	1	6.50E-04	186	(10, 3600)
20	4	100.0%	3.6%	1	2.40E-05	241	(13, 4479)
26	5	100.0%	3.8%	1	8.70E-07	280	(16, 5073)
29	6	100.0%	3.7%	1	3.10E-08	337	(19, 5987)
30	7	85.7%	3.7%	2	1.20E-07	156	(19, 1302)

*0.5 was added to each cell for comparisons that had a zero cell

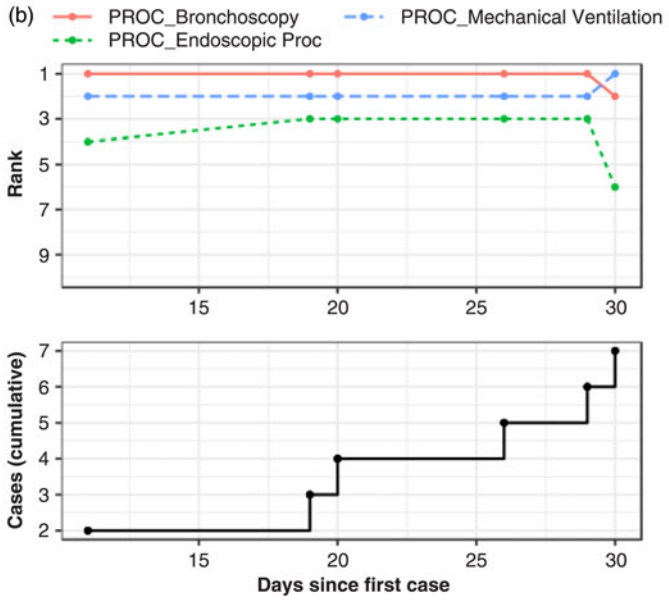

Fig. 2. Transmission route ranking for outbreak no. 3: *Klebsiella pneumoniae* from a contaminated bronchoscope. Panel A: Results for bronchoscopy, showing timing of cases, the proportion of cases who were found in the EHR to have been exposed to bronchoscopy before illness onset, the percent of the control population that was found in the EHR to have been exposed to bronchoscopy, the score, the ranking relative to other exposures, the *P* value and odds ratio. Panel B: Graphical depiction of relative ranking of bronchoscopy and 2 other ranking exposures (top figure) and the cumulative number of cases (bottom figure), both over time.

some variables, such as exposure to specific healthcare workers. In subsequent iterations, we plan to expand the number of variables that are included. Fourth, our analysis focused on single-modal transmission routes. For example, outbreak no. 2 was not detected until the eighth patient because some of the patients early in the outbreak did not have an ERCP procedure, which suggested another transmission route early in the outbreak. This initially diminished our ability to detect ERCP as the major transmission route. However, had WGS surveillance been in place in real time, the outbreak might have been detected sooner. Fifth, the creation of the data mining algorithm required unblinding of the investigators for some outbreaks. However, this was required during this developmental phase of EDS-HAT, which is now undergoing prospective validation. Finally, automated data mining of the EHR does not obviate the need for traditional "shoe leather" epidemiology for outbreak investigations. Additional efforts will often be required such as manual review of the EHR to obtain information about exposure details and culturing of an implicated device or direct observations of suspected procedures based on the results of this automated approach.

We have recently instituted WGS surveillance of key hospital bacterial pathogens to enhance outbreak detection in our hospital. If run in real time, routine WGS in combination with data mining has the potential to identify outbreaks earlier than traditional methods thus preventing a larger outbreak or, importantly, identify outbreaks that might not otherwise be detected. Prospective validation of this approach is underway.

Acknowledgments. We thank Chinelo Ezeonwuka and Sara Ohm for their assistance with the molecular typing.

Financial support. This study was funded in part by the National Institute of Allergy and Infectious Diseases (grant nos. R21Al109459 and R01AI127472).

Conflicts of interest. All authors report no conflicts of interest relevant to this article.

References

1. Magill SS, Edwards JR, Bamberg W, *et al.* Multistate point-prevalence survey of health care–associated infections. *N Engl J Med* 2014;370:1198–1208.
2. Scott RD. The direct medical costs of healthcare-associated infections in US hospitals and the benefits of prevention, 2009. Centers for Disease Control and Prevention website. http://www.cdc.gov/HAI/pdfs/hai/Scott_CostPaper.pdf. Published 2009. Accessed August 13, 2018.
3. Marsh JW, Krauland MG, Nelson JS, *et al.* Genomic epidemiology of an endoscope-associated outbreak of *Klebsiella pneumoniae* carbapenemase (KPC)–producing *K. pneumoniae*. *PLoS One* 2015;10:e0144310. doi: 10.1371/journal.pone.0144310.
4. Sood G, Perl TM. Outbreaks in health care settings. *Infect Dis Clin N Am* 2016;30:661–687.

5. Vonberg RP, Weitzel-Kage D, Behnke M, *et al.* Worldwide outbreak database: the largest collection of nosocomial outbreaks. *Infection* 2011; 39:29–34.
6. Peacock SJ, Parkhill J, Brown NM. Changing the paradigm for hospital outbreak detection by leading with genomic surveillance of nosocomial pathogens. *Microbiology* 2018;164:1213–1219. doi: 10.1099/mic.0.000700.
7. Heinrichs A, Argudin MA, De Mendonca R, *et al.* An outpatient clinic as a potential site of transmission for an outbreak of NDM-producing *Klebsiella pneumonia* ST716: a study using whole-genome sequencing. *Clin Infect Dis* 2018. doi: 10.1093/cid/ciy581.
8. Domman D, Chowdhury F, Khan Al, *et al.* Defining endemic cholera at three levels of spatiotemporal resolution within Bangladesh. *Nat Genet* 2018;50:951–955.
9. Pak TR, Kasarskis A. How next-generation sequencing and multiscale data analysis will transform infectious disease management. *Clin Infect Dis* 2015;61:1695–1702.
10. Yount RJ, Vries JK, Councill CD. The medical archival retrieval system: an information retrieval system based on distributed parallel processing. *Inform Process Manag* 1991;27:1–11.
11. Miller JK, Chen C, Sundermann AJ, Marsh JW, Saul MI, Shutt KA, Pacey M, Mustapha MM, Harrison LH, Dubrawski A. Statistical outbreak detection by joining medical records and pathogen similarity. Accepted manuscript for *J Biomed Inform.*
12. Parr A, Querry A, Pasculle A, Morgan D, Muto C. Carbapenem-resistant *Klebsiella pneumoniae* cluster associated with gastroscope exposure among surgical intensive care unit patients at University of Pittsburgh Medical Center. *Open Forum Infect Dis* 2016:3:248.

https://doi.org/10.1017/ice.2018.343 Published online by Cambridge University Press