Novel Approaches for Healthcare Outbreak Detection and Investigation

by

Alexander John Sundermann

BS in Microbiology, University of Rochester, 2013

MPH in Infectious Diseases and Microbiology, University of Pittsburgh, 2014

Submitted to the Graduate Faculty of the

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Public Health

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Alexander John Sundermann

It was defended on

March 25, 2022

and approved by

Maria Mori Brooks, PhD
Professor of Epidemiology and Biostatistics, Co-Director, Epidemiology Data Center
Graduate School of Public Health, University of Pittsburgh

Elise M Martin, MD, MS
Associate Medical Director of Infection Prevention and Hospital Epidemiology
School of Medicine, University of Pittsburgh

Mark S. Roberts, MD, MPP
Distinguished Professor of Health Policy and Management
Director, Public Health Dynamics Lab
Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Lee H. Harrison, MD
Professor of Medicine and Epidemiology
Director, Center for Genomic Epidemiology
Graduate School of Public Health, University of Pittsburgh

Novel Approaches for Healthcare Outbreak Detection and Investigation

Alexander John Sundermann, DrPH

University of Pittsburgh, 2022

Methods for detecting outbreaks in healthcare settings have remained unchanged for many years. Often this involves the use of geo-temporal clustering which looks for an increase in the number of expected infections within a small timeframe in a confined location. This approach often misses transmission where it did occur and mis-identifies transmission where it did not occur. Additionally, other routes of potential transmission, such as shared providers or procedures, are often not considered. These data are readily available within the electronic health record (EHR). Traditional infection prevention methods often use whole genome sequencing (WGS) at the end of an outbreak to confirm or refute its presence, referred to as reactive sequencing.

The objective of this dissertation is to create and evaluate the Enhanced Detection System for Healthcare-Associated Transmission (EDS-HAT), which better detects and investigates outbreaks compared to traditional infection prevention methods. EDS-HAT combines WGS surveillance with machine learning (ML) of the EHR. The creation of EDS-HAT was performed in three steps. First, we performed a systematic review of institutions performing WGS surveillance and/or machine learning of EHR data to obtain a better understanding of EDS-HAT's use and implications. We found that very few institutions were performing WGS surveillance or machine learning of EHR, yet both had profound impact on outbreak detection and investigation. Second, we developed and trained a proof-of-concept ML algorithm on past, well-described outbreaks that occurred at our institution. We found that the algorithm could accurately identify the correct transmission route on the second patient in all but one outbreak. Lastly, we performed

two years of WGS surveillance to directly compare to traditional infection prevention practice. Based on those results, EDS-HAT, if run in real time, could potentially identity otherwise undetected outbreaks, prevent many infections, save money, and be substantially more effective than traditional infection prevention practice. Overall, our findings support the use of WGS and machine learning of the EHR to detect and investigate outbreaks. If implemented in real-time, EDS-HAT represents a potential paradigm shift in infection prevention to increase patient safety.

Table of Contents

List of Tables

List of Figures

Preface

There are many individuals who have guided me and contributed significantly to my research. First, I want to thank my advisor Dr. Lee Harrison for his constant support, willingness to always teach, and for his ability to always entertain my thoughts and ideas. His excitement for this research over the years has inspired me throughout my career to persist. I'd also like to thank my committee members for their guidance and input as mentors in my research.

I'd like to thank everyone who I've worked with in the Microbial Genomic Epidemiology Laboratory and colleagues within the division of infectious diseases. I've met so many wonderful people in the lab who have always taken the time to teach me new skills and encourage me in my research. I'd also like to thank my colleagues in the University of Pittsburgh Medical Center Presbyterian Department of Infection Prevention who have always supported me throughout my DrPH program. I'd also like to thank my friends and family, especially my parents who have always been supportive for my research and interests within school.

Lastly, to my wife, Hannah. Your unwavering support and dedication are unmeasurable. You have always pushed me to my greatest potential, and I am so grateful.

1.0 INTRODUCTION

1.1 HEALTHCARE-ASSOCIATED INFECTIONS

Healthcare-associated infections (HAIs) are an unfortunately common occurrence within hospitals. The Centers for Disease Control and Prevention (CDC) estimates that one in 31 patients in any day has at least one HAI.[1] Moreover, HAIs are a source of significant morbidity and mortality that are preventable.[2] HAIs in the United States are tracked through the National Healthcare Safety Network (NHSN) which has created surveillance definitions of what defines a healthcare-associated infection for the purposes of standardization across healthcare facilities.[3] Reporting of HAIs are a requirement of Centers for Medicare and Medicaid Services for facility reimbursement that are benchmarked by performance.[4]

In recent years, the CDC and NHSN have shown that there have been significant advances in infection prevention for reduction in HAIs. However, the COVID-19 pandemic has stopped or even reversed some of that progress according to recent CDC data.[5] The increase may be attributed to the demand on hospital staffing, strict isolate requirements from COVID-19, and cohorting of critically ill patients.

Sources of HAIs may result from the patient's flora, environment, contaminated equipment or medication, or other healthcare providers.[6] The CDC has created prevention and intervention recommendations based upon types of infections, the organisms, and mode of transmission.[7] For example, central line blood stream infections may be caused by skin flora on the patient which can enter the body and cause infection during the insertion of a central line. The CDC provides evidence-based recommendations on the prevention of this by creating a sterile environment,

cleaning the site of insertion, and protecting the dressing for any contamination.[8] However, some HAIs may result from outbreaks of contaminated medication or transmission within hospital units which requires a detailed process of investigation elucidate the cause and intervention to prevent further spread.[9]

1.2 HEALTHCARE-ASSOCIATED OUTBREAKS

1.2.1 Detecting and Investigating Outbreaks

An outbreak within a healthcare setting may refer to a sudden increase of infections compared to what is normally seen in a certain time period.[10] However, there is no clear guidance on defining endemic levels of infections and over what time frame. To detect outbreaks, institutions often rely on using geo-temporal clustering of infections, which utilizes both space and time aspects. This method looks to see what patients with the same infections have shared unit location commonalities, often on the same unit. Similarly, the temporal aspect will examine if the patients are on that unit at the same time. Geo-temporal clustering has strong evidence of transmission given the potential for the pathogen to be transmitted is high if both the source patient and susceptible patient are present on the same unit at the same time.

Often, clinicians will see an increase of infections on their hospital unit within a relatively short time frame. It is then that the clinician may inform infection prevention leadership of a suspected outbreak or transmission. Similarly, the infection prevention department may utilize NHSN surveillance definitions as a benchmark and tool for detecting outbreaks of infections.

These surveillance methods give infection prevention departments a relatively loose threshold for detecting an outbreak.[11]

Once an outbreak is suspected, the infection prevention department can initiate an outbreak investigation. This often entails creating a line list of patients, manually reviewing patient charts for shared commonalities, performing staff interviews, performing audits of clinical practice, and taking environmental cultures. Once a hypothesis is formed, the infection prevention department can initiate an intervention, monitor for additional infections, and tailor interventions based on subsequent cases.

1.2.2 Whole Genome Sequencing

An infection prevention department may choose to perform whole genome sequencing on bacterial isolates from suspected outbreak patients. This is referred to as reactive whole genome sequencing, given that the investigation and interventions are often concluded as the sequencing is being performed. Whole genome sequencing can provide ‘genetic fingerprinting’ by discerning which isolates are potentially transmitted by looking at genome mutations, or single nucleotide polymorphisms (SNPs). If two patients have the same organism with a low SNP difference, this likely indicates that one patient transmitted to the other or there is a common source. Whereas patients with a high SNP difference may indicate their infections are unrelated, or not transmitted to each other. Reactive WGS can assist in confirming the initial hypothesis or even by refuting the presence of a clonal outbreak.

1.2.3 Limitations with Current Approaches

There are many limitations to this current approach of outbreak detection and investigation. First, the use of geo-temporal clustering can both miss transmission and misidentify transmission.[12] Geo-temporal clustering does not consider alternative transmission pathways such as shared equipment, procedures, medication, or providers moving throughout the hospital. Second, relying on the use of NHSN definitions may also miss transmission. These definitions are meant for surveillance purposes and may miss clinically-defined infections or colonization that do not meet the full surveillance definition. Lastly, the use of reactive sequencing is limited in its use given it occurred well after the outbreak started and restricted to only the isolates selected often by geo-temporal clustering.

1.2.4 Opportunities for Advancement

Decrease costs for the use of WGS has provided an opportunity to proactively use WGS as a surveillance tool. What was once nearly thousands of dollars for sequencing a single bacterial isolate can now be done for under $100. The cost lowers the threshold needed to achieve a cost benefit in preventing infections by whole genome sequencing.

Additionally, of electronic health records and machine learning algorithms have become more available and broadly used. Historically, patient charts had been kept on paper as a barrier to large-scale data analysis. Large troves of data are waiting to be analyzed now with the development and implementation of electronic health records.

Together, these tools could help infection prevention departments overcome the limitations of current outbreak detection methods. Leveraging the implementation of such a tool would require

a detailed cost-benefit analysis in which the utility would be studied. The purpose of this dissertation is to analyze that tool.

2.0 MANUSCRIPT 1: WHOLE GENOME SEQUENCING SURVEILLANCE AND MACHINE LEARNING FOR HEALTHCARE OUTBREAK DETECTION AND INVESTIGATION: A SYSTEMATIC REVIEW AND SUMMARY

Alexander J. Sundermann,[1-3] Jieshi Chen,[4] James K. Miller,[4] Elise M Martin,[2,5] Graham M. Snyder,[2,5] Daria Van Tyne,[2] Jane W. Marsh,[1,2] Artur Dubrawski,[4] and Lee H. Harrison[1-3]

1. Microbial Genomic Epidemiology Laboratory, Center for Genomic Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.
2. Division of Infectious Diseases, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA.
3. Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
4. Auton Lab, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
5. Department of Infection Prevention and Hospital Epidemiology, UPMC Presbyterian, Pittsburgh, Pennsylvania, USA

2.1 ABSTRACT

Background: Whole genome sequencing (WGS) has traditionally been used in infection prevention to confirm or refute the presence of an outbreak after it has occurred. Due to decreasing costs of WGS, an increasing number of institutions have been utilizing WGS-based surveillance.

Additionally, machine learning (ML) or statistical modeling to supplement infection prevention practice have also been used. We systematically reviewed the use of WGS surveillance and machine learning to detect and investigate outbreaks in healthcare settings.

Methods: We performed a PubMed search using separate terms for WGS surveillance and/or machine learning technologies for infection prevention through March 15, 2021.

Results: Of 767 studies returned using the WGS search terms, 42 articles were included for review. Only 2 (4.8%) studies were performed in real-time, and 39 (92.9%) only studied one pathogen. Nearly all (41, 97.6%) studies found genetic relatedness between some isolates collected. Across all studies, there were 525 outbreaks detected among 2,837 related isolates (average 5.4 isolates/outbreak). 35 (83.3%) studies only utilized geo-temporal clustering to identify outbreak transmission routes. Of 21 studies returned using the ML search terms, 4 were included for review. In each study, ML aided outbreak investigations by complementing methods to gather epidemiologic data and automating identification of transmission pathways.

Conclusions: WGS surveillance is an emerging method that can enhance outbreak detection. ML has the potential to identify novel routes of pathogen transmission. Broader incorporation of WGS surveillance into infection prevention practice has the potential to transform the detection and control of healthcare outbreaks.

2.2 INTRODUCTION

Whole genome sequencing (WGS) for infection prevention has traditionally been used in reaction to a suspected outbreak, usually at the end of an investigation to confirm or refute the outbreak's presence. In contrast, WGS surveillance of selected healthcare-associated pathogens

regardless of whether an outbreak is suspected can be used to identify outbreaks that are not detected by traditional hospital epidemiologic methods. High costs and needed infrastructure for implementation have been historic barriers to widespread use of WGS surveillance. However, the cost of WGS has fallen, and the expansion of genomic surveillance due to COVID-19 may enable healthcare institutions to establish WGS surveillance programs for other pathogens. Additionally, our work and studies from Australia have found cost-benefits to implementing a WGS surveillance program with effective intervention.[13]

Although WGS surveillance is effective in identifying transmission, it does not provide information on the responsible transmission route, which is crucial for interrupting an outbreak. Traditional epidemiologic methods for identifying where transmission occurs have relied on geotemporal clustering within the hospital, which is inadequate for identifying more complex patterns of transmission.[14,15] Automated analysis of electronic health records (EHRs) creates an opportunity to use machine learning or statistical modeling approaches for determining the outbreak transmission routes identified by WGS surveillance.[16–20] These automated approaches would assist hospital infection prevention departments by providing systematic methods to investigate outbreaks and identify transmission routes.

In this systematic review, we provide a summary of prior studies utilizing WGS surveillance in healthcare settings for outbreak detection, as well as the use of machine learning and statistical modeling technologies for identifying transmission routes. The purpose of this review is to summarize the current literature in this field, identify barriers to widespread implementation, and synthesize current knowledge on this topic to help guide decision-making about implementation of WGS surveillance.

2.3 METHODS

Two search terms were utilized in PubMed with no beginning date up until March 15, 2021 [Figures 1 & 2]. The WGS surveillance terms "(whole genome sequenc*) AND (surveillance OR routine) AND (healthcare OR hospital) AND transmission" returned 767 results. Article abstracts were screened to exclude studies that were solely community-based, non-infection related, utilized non-WGS methods (e.g., older molecular subtyping methods such as pulsed-field gel electrophoresis), or only utilized reactive WGS in response to suspected outbreaks. Genomic and epidemiologic data on organisms, number of isolates sequenced, percent of isolates that were related, number of outbreaks, and epidemiological links were extracted and summarized. Articles were excluded if the data were not sufficiently detailed for extraction.

The machine learning search terms utilized were "("electronic health record" OR "electronic medical record" OR "artificial intelligence" OR "AI" OR "ML" OR "model") AND (outbreak OR transmission) AND ("data mining" OR "machine learning") AND (infection OR infectious) AND ("healthcare-associated" OR "hospital-associated" OR "healthcare-acquired" OR "hospital-acquired")" and returned 21 results. Article abstracts were screened to exclude infection prediction and outcome studies. Data on the methodology and findings of outbreak and transmission detection models were extracted and summarized.

2.4 RESULTS

There were 42 articles on WGS surveillance included in the final review.[15–17,21–59] Of these studies, only 2 employed machine learning or statistical modeling to investigate transmission,

which were also captured in the ML search. From 2013-2016, there was only one article per year, with a substantial increase thereafter [Figure 3]. Most studies were from the United States (12), United Kingdom (10), Australia (5), Germany (4), Japan (2); China, Denmark, Finland, France, India, Italy, Netherlands, Spain, Sweden, and Thailand had one study each.

The duration of WGS surveillance varied substantially by study, with a median of 12 months and a range of 1-73 months [Table S1]. Only 2 (4.8%) studies were performed in real-time; all other studies were performed retrospectively. Thirty-nine and three studies included single or multiple pathogens, respectively [Table 1]. *Staphylococcus aureus* was the most commonly studied organism (12, 28.6%) with four additional organisms present in >2 studies (nine *Klebsiella pneumoniae*, seven *Clostridioides difficile*, six *Enterococcus faecium*, three *Pseudomonas aeruginosa*). Organisms selected for sequencing (e.g., by anatomic site of infection, multi-locus sequence type, antibiotic resistance phenotype) were diverse across studies.

Criteria for defining genetic relatedness were also highly variable between studies, and were generally based on the number of single nucleotide polymorphism (SNP) differences between genomes [Table S1]. Among organisms present in >2 studies, *C. difficile* had the most consistent SNP cutoff at 2 SNPs, with one study that used 10 SNPs to identify related isolates [Figure 4]. *S. aureus* had the widest distribution of SNP cut-offs, ranging from 7 to 50 SNPs.

An analysis of the proportion of sequenced isolates that were determined to be genetically related to one another in each study revealed an average of 23.8% of isolates (range 0%-61%). There were 525 outbreaks detected among 2,837 related isolates (average 5.4 isolates/outbreak). 41 (97.6%) studies found some level of genetic relatedness between the sequenced isolates.

We examined the methods employed to identify the responsible transmission routes for outbreaks that were detected by WGS. The majority of studies (35, 83.3%) restricted attempts to

identify transmission routes to the same hospital unit during a defined time period.[21,23–39,41–50,52–54,56–59] Only 7 (16.7%) studies examined other possible routes such as medical procedures or healthcare workers.[15–17,22,40,51,55]

Several studies were notable for uncovering otherwise unidentified transmissions, which is the main goal of WGS surveillance. Sullivan et al[55] were prompted by an outbreak in a neonatal intensive care unit (NICU) to retrospectively investigate all MRSA bloodstream infections for 16 months. Their investigation uncovered isolates related to the NICU outbreak from adult patients in a separate tower. Further investigation revealed shared ventilators between the adult unit and the NICU, which was believed to have caused transmission. Separately, Roy et al.[52] performed sequencing of influenza A H1N1for 6 months and found that traditional infection prevention practice falsely identified outbreaks, while WGS surveillance data were able to connect cases that were previously not believed to be epidemiologically related. Lastly, Berbel Caban et al.[22] utilized WGS surveillance of MRSA over two years and found multiple undetected outbreaks within two New York City hospitals. One cluster of 24 isolates from 16 patients spanned 21 months and nine different hospital wards with patterns of shared healthcare workers. In this study, the authors emphasized the limitations of investigating only geo-temporal clustering in outbreak detection and investigation.

There were 4 articles within the ML search terms included in the final synthesis, 2 of which overlapped in the WGS surveillance search terms.[17,40,60,61] Table 2 summarizes the methods and limitations of each study. One study utilized imputation of cultures to model transmission dynamics from environmental sink contamination, 2 studies used Bayesian methods to model transmission, and one study combined WHONET and SaTScan tools to detect outbreaks. All of these studies implemented tools to supplement outbreak detection or investigation, yet each study

also noted the importance of manual or expert input to further investigate the transmissions or outbreaks detected. An example of this is the study by Satchel et al[61] in which they found 45 outbreaks of which only six were confirmed by IP investigation. Yet the authors state that the tool helped to streamline investigation efforts which reduced time spent by IP.

2.5 DISCUSSION

In this systematic review, we synthesized studies that demonstrate the utility of WGS surveillance in finding cryptic outbreaks in healthcare settings. Nearly all studies (97.6%) found outbreaks, but few (4.8%) utilized machine learning or statistical modeling methods to investigate transmission routes. WGS surveillance, while uncommon but increasingly utilized, aided infection prevention practice in these studies by uncovering outbreaks and enabling intervention.

Studies utilizing WGS surveillance have primarily relied on geo-temporal linkage to identify transmission routes. Restricting investigations to geo-temporal linkage fails to identify potential transmission by procedures that are performed in areas of the hospital other than patient nursing units or healthcare workers, as shown in some of the studies in this review. Some studies stated the limitations of relying solely on geo-temporal parameters for identifying the transmission route for related isolates. Regardless, WGS surveillance enabled many of these studies to uncover substantial and significant previously undetected outbreaks that likely impacted patient outcomes and associated healthcare costs.

The vast majority of studies were retrospective in nature, which limits the potential impact of WGS surveillance on healthcare epidemiology and infection prevention. If performed in real-time, IP teams have an opportunity to perform an investigation, such as audit practices, collect

environmental cultures, and interview staff, which is not possible in retrospective studies. Further, we found that many studies focused on one pathogen, which is less sensitive for detecting outbreaks than WGS surveillance of multiple pathogens. It is possible, for example, for a single transmission route to lead to spread of multiple pathogens.

Substantial investment and infrastructure are needed to establish real-time WGS surveillance. Healthcare institutions must have appropriate laboratory capacity, bioinformaticians, and genomic epidemiologists to interpret the data. A recent paper by Parcell et al[62] discussed barriers to instituting a WGS surveillance program for outbreak detection from an economic and system-wide perspective. Indeed, it is often difficult to prove estimates of cost-effectiveness when considering prevention interventions, but two studies have demonstrated the cost-effectiveness of WGS surveillance programs.[13,19]

We identified very few studies on the utility of ML or statistical modeling methods for identification of outbreak transmission routes by WGS surveillance. In our experience, ML adds value in detecting transmission routes that do not involve geotemporal clustering such as invasive procedures, healthcare workers, and outbreaks separated by units and prolonged in time.[17,18,20] The use of ML in combination with WGS surveillance is clearly an understudied area of healthcare epidemiology and infection prevention. Barriers such as interoperability of electronic health records and adoption of WGS surveillance prevent the implementation of such programs. However, adoption of public health WGS surveillance for COVID-19 may expedite the use of this technology by healthcare institutions.

The combination of prospective WGS surveillance, EHR data, and ML has the potential to dramatically transform the paradigm of outbreak detection and investigation for infection prevention and control by identifying outbreaks quicker and enabling early intervention to halt

transmission. This approach will both improve patient safety and reduce healthcare costs. However, healthcare institutional investment into establishing WGS surveillance programs will be key to expansion and implementation of this approach.

2.6 TABLES AND FIGURES

Table 1. Studies by date, organism and outbreaks detected utilizing WGS surveillance

Year	First Author	Organism(s)	Type	Unique Isolates	Related (%)	Outbreaks Detected	No Epi Link (%)
2021	Meredith[46]	SARS-CoV-2	.	299	159 (53.2)	35	35 (22)
2021	Miles-Jay[47]	*E. coli*	ST131, H30	126	17 (13.5)	8	9 (52.9)
2021	Rose[51]	*S. aureus*	Methicillin-resistant	56	15 (26.8)	7	7 (46.7)
2020	Berbel Caban[22]	*S. aureus*	Methicillin-resistant	224	33 (14.7)	8	.
2020	Cremers[24]	*S. aureus*	Methicillin-sensitive	84	40 (47.6)	14	0 (0)
2020	Gona[32]	*K. pneumoniae*	.	80	39 (48.8)	10	14 (35.9)
2020	Hammerum[35]	*K. pneumoniae*	.	103	36 (35)	13	11 (30.6)
2020	Marmor[43]	*E. cloacae*	.	63	7 (11.1)	1	0 (0)
		C. freundii	.		10 (15.9)	1	0 (0)
2020	Neumann[48]	*E. faecium*	Vancomycin-resistant	111	.	.	.
2020	Sundermann[17]	*P. aeruginosa*	ST27	882	31 (3.5)	10	1 (3.2)
2020	Sundermann[16]	*E. faecium*	Vancomycin-resistant, ST1471	439	10 (2.3)	.	1 (10)
2020	Tsujiwaki[56]	*S. aureus*	Methicillin-resistant	57	19 (33.3)	5	0 (0)
2019	Eigenbrod[26]	*A. baumannii*	.	39	15 (38.5)	4	5 (33.3)
2019	Eyre[30]	*C. difficile*	.	299	43 (14.4)	6	20 (46.5)
2019	García-Fernández[31]	*C. difficile*	.	367	41 (11.2)	6	34 (82.9)
2019	Hall[34]	*S. aureus*	Methicillin-resistant	55	27 (49.1)	12	8 (29.6)
2019	Harada[36]	*K. pneumoniae*	Bloodstream infections	140	2 (1.4)	1	2 (100)
2019	Jakharia[38]	*C. difficile*	.	45	4 (8.9)	2	4 (100)
2019	Kossow[39]	*S. aureus*	Methicillin-resistant	.	8	1	0 (0)
2019	Mathur[45]	*K. pneumoniae*	Colistin-resistant	21	8 (38.1)	4	0 (0)

Year	First Author	Organism(s)	Type	Unique Isolates	Related (%)	Outbreaks Detected	No Epi Link (%)
2019	Roy[52]	Influenza	A H1N1	36	5 (13.9)	2	2 (40)
2019	Sherry[53]	*Enterobacteriaceae*	Carbapenemase-producing	291	53 (18.2)	12	8 (15.1)
2019	Stenmark[54]	*S. capitis*	Bloodstream infections	46	12 (26.1)	6	12 (100)
2019	Sullivan[55]	*S. aureus*	Methicillin-resistant	141	28 (19.9)	4	2 (7.1)
2019	van Beek[57]	*K. pneumoniae*	Carbapenemase-producing, ST512	.	20	2	4 (20)
2019	Wang[58]	*C. striatum*	.	91	18 (19.8)	6	3 (16.7)
2019	Ward[15]	*S. aureus*	.	953	85 (8.9)	28	65 (76.5)
		E. faecium	.	86	13 (15.1)	5	9 (69.2)
		P. aeruginosa	.	118	2 (1.7)	1	2 (100)
		K. pneumoniae	.	100	0 (0)	0	0 (0)
2018	Auguet[21]	*S. aureus*	Methicillin-resistant	610	261 (42.8)	90	13 (5)
2018	Donskey[25]	*C. difficile*	.	66	12 (18.2)	4	4 (33.3)
2018	Houldcroft[37]	Adenovirus	.	43	6 (14)	2	0 (0)
2018	Kwong[40]	*K. pneumoniae*	Carbapenemase-producing	86	53 (61.6)	4	10 (18.9)
2018	Leong[41]	*E. faecium*	Vancomycin-resistant	80	10 (12.5)	2	3 (30)
2018	Martin[44]	*C. difficile*	.	640	227 (35.5)	.	69 (30.4)
2018	Wendel[59]	*A. baumannii*	.	36	20 (55.6)	2	2 (10)
2017	Coll[23]	*S. aureus*	Methicillin-resistant	1465	785 (53.6)	173	187 (23.8)
2017	Eyre[29]	*C. difficile*	.	652	128 (19.6)	.	.
2017	Gorrie[33]	*K. pneumoniae*	.	106	17 (16)	5	0 (0)
2017	Raven[49]	*E. faecium*	Bloodstream infections	293	93 (31.7)	6	.
2016	Elbadawi[27]	*K. pneumoniae*	Carbapenemase-producing	46	4 (8.7)	1	0 (0)
2015	Roach[50]	*S. epidermidis*	.	178	56 (31.5)	10	.
		P. aeruginosa	.	44	7 (15.9)	3	.
		E. faecium	.	36	13 (36.1)	3	.
		S. aureus	.	118	4 (3.4)	2	.
		E. faecalis	.	72	6 (8.3)	3	.
		S. maltophilia	.	58	2 (3.4)	1	.
2014	Long[42]	*S. aureus*	.	305	0 (0)	0	.
2013	Eyre[28]	*C. difficile*	.	957	333 (34.8)	.	152 (45.6)

Table 2. Studies utilizing machine learning or modeling to detect outbreaks or transmission

Year	First Author	Machine Learning or Model Method	Utility & Findings	Limitations
2018	Lensing[60]	Imputation of clinical and environmental cultures to model transmission dynamics	Aided in targeted environmental cleaning to decolonize plumbing systems and reduce the risk of transmission of carbapenem-resistant *Enterobacteriaceae* based upon learned positivity.	Requires expert knowledge of plumbing system and prior data on colonization
2018	Kwong[40]	Bayesian transmission modeling using Markov chain Monte Carlo	Assisted in transmission modeling of KPC-producing *K. pneumoniae* to determine if spread resulted from inter-facility or intra-facility transmission.	Does not provide specific details in sequence of transmission within a complex outbreak
2020	Sundermann[17]	Bayesian inference with case-control methodology to describe transmission sources	Scanned electronic health record and provided statistical output for possible transmission routes beyond but including geo-temporal clustering.	Requires robust mapping of electronic health record charge codes
2017	Stachel[61]	WHONET-SaTScan	Utilizes space-time permutation scan statistics to identify potential outbreaks.	Low positive predictive value creating a high number of “false alarms”

3.0 MANUSCRIPT 2: AUTOMATED DATA MINING OF THE ELECTRONIC HEALTH RECORD FOR INVESTIGATION OF HEALTHCARE-ASSOCIATED OUTBREAKS

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

1. Infectious Diseases Epidemiology Research Unit, University of Pittsburgh School of Medicine and Graduate School of Public Health, Pittsburgh, Pennsylvania
2. Auton Laboratory, Carnegie Mellon University, Pittsburgh, Pennsylvania
3. Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania
4. Department of Infection Control and Hospital Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania
5. Department of Pathology, University of Pittsburgh, Pittsburgh, Pennsylvania

3.1 ABSTRACT

Background: Identifying routes of transmission among hospitalized patients during a healthcare-associated outbreak can be tedious and difficult, particularly for patients with complex

hospital stays. Data mining (DM) of the electronic medical record (EMR) 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 2012-2016 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 DM of the EMR to identify the correct route of transmission, 2) when 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 on the eighth patient in one outbreak, and the second patient in all other outbreaks. Up to 40 or 34 infections (71% or 60% of infections, respectively) could have been prevented if EMR DM had been implemented in real-time, assuming initiation of effective intervention within 7 or 14 days of outbreak onset, respectively.

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

3.2 INTRODUCTION

Healthcare-associated outbreaks caused by serious bacterial pathogens cause substantial morbidity and mortality and add to healthcare costs.[63,64] Detection of outbreaks can be difficult in large hospitals where bacterial transmission may go unnoticed for prolonged periods of time.[65] Investigation and control of a hospital outbreak requires identifying the route of transmission

among patients suspected of being part of the outbreak. This task can be burdensome and 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 been described and include transmission from environmental contamination; colonized healthcare personnel; during medical procedures using contaminated devices; and through contaminated medications, solutions or other medical therapies.[66,67]

The wide-spread availability of the electronic medical record (EMR) offers the potential to use automated data mining tools to find common exposures among hospitalized patients during outbreak investigations. Many relevant epidemiologically-important variables are readily available in the EMR, 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 EMR during hospital outbreak investigations. Furthermore, whole genome sequencing (WGS) has become an increasingly available method that discriminates pathogens at the genetic level.[68–70] Genomic data from patient bacterial isolates has the potential to aid in the data mining and outbreak investigation process.[71] We are developing a system that, in combination with WGS surveillance of clinical isolates of key hospital-associated bacterial pathogens, prospectively monitors the EMR to identify potential outbreaks and their routes of transmission. The purpose of this study was to develop and validate data mining tools to accomplish this goal using well-characterized outbreaks from 2011-2016 at our institution.

3.3 METHODS

3.3.1 Study Setting

This study was conducted at the University of Pittsburgh Medical Center-Presbyterian Hospital (UPMC), an adult medical/surgical tertiary care hospital with 762 total beds, 150 critical care unit beds, more than 32,000 yearly inpatient admissions, and over 400 solid organ transplants per year. The UPMC eRecord EMR system has more than 29,000 active users, including more than 5,000 physicians affiliated with UPMC, and comprises more than 3.6 million unique electronic patient records. UPMC uses Cerner PowerChart and EpicCare as the backbone of its inpatient and outpatient EMR systems, respectively.

3.3.2 Characterization of retrospective outbreaks from 2011 to 2016

During the period of 2011-2016, routine infection prevention practice 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.*

During the study period, our primary method for molecular characterization of bacterial isolates other than *C. difficile* 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, whole genome

sequencing replaced PFGE and a cut-off of ≤ 20 single nucleotide polymorphisms (SNPs) was used to define genetically related patient isolates.

For identification of the common exposure responsible for individual outbreaks, our infection prevention team analyzed the medical records of patients included in the outbreak to identify the responsible routes of transmission (e.g., shared locations/staff, shared procedures/operations, or shared medications). Some outbreak investigations utilized environmental cultures to confirm routes 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.

3.3.3 Extraction and processing of EHR data for data missing

All inpatient, emergency room, and same day surgery encounters between January 1, 2011 and December 31, 2016 were identified through an electronic medical record data repository that contains full-text medical records and integrates information from central transcription, laboratory, pharmacy, finance, administrative, and other departmental databases.[72] For each encounter, we obtained microbiology reports and charge transactions from the data repository. To maintain patient confidentiality, each patient was assigned a studyid using De-ID software (De-ID Data Corp, Philadelphia, PA). Criteria were met for exemption from informed consent by the university's Institutional Review Board.

Charge transaction data are in the EMR as charge codes. Multiple charge codes can represent exposure to a single instrument; therefore, charge codes for key procedures (e.g., endoscopic retrograde cholangiopancreatography [ERCP] and bronchoscopy) were collapsed into a single variable group that represented that exposure. For example, ERCP has 8 CPT codes (e.g.,

43260: ERCP; diagnostic, including collection of specimens. . .; 43261: ERCP with biopsy; 43278: ERCP; with ablation of tumors. . .) and were all combined into a single variable called “ERCP”, although each charge code was also analyzed individually.

3.3.4 Data missing of the electronic health record (EHR)

The data-mining program was designed using a case-control approach based upon the genotyping results using patient EMR data that are non-related to the outbreaks as controls. 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 same time period who 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 that were identified by infection prevention 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 top three possible 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 intervention from the date of the positive culture assuming the data-mining program had been running in real-time and appropriate interventions were enacted (removal of contaminated equipment, disinfection of environment, and/or enhanced precautions). Outbreaks were deemed non-preventable if there were only two isolates in the WGS grouping.

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 (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.[20] We take γ=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 (i.e. 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 (i.e. by some other means such as patient-to-patient transmission), then the likelihood of observing a particular set of case patients is proportional to $\theta^b(1-\theta)^{r-b}\gamma^{-b}$, where b is the number of case patients infected by the route. We arrive at the formula S by maximizing this expression in θ, which occurs at θ=b/r, and b, which occurs at either 0 or a. Since b=0 is a degenerate solution, it is disregarded. The final score is the log of this maximum likelihood.

The score above represents an unnormalized log-likelihood. Since it is not normalized, it is suitable for ranking routes but not comparable across time as the number of case patient changes. We therefore estimate 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, during development of the approach it became clear 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 EMR and therefore could not possibly 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.

3.4 RESULTS

The characteristics of the 9 outbreak investigations during the study period are shown in Table 1. For some investigations, the molecular typing revealed several separate clusters. For example, for investigation No. 2, there were 2 clusters involving 2 isolates each. For two (22%) *C. difficile* outbreaks (Nos. 8 and 9), epidemiologic investigation revealed that transmission occurred in the nursing units where the patients resided. Three (33%) investigations involved *Klebsiella pneumoniae*, one of which represented a polyclonal ERCP-related outbreak (No. 2),[3] and one each involved bronchoscopy (No. 3) and gastroscopy (No. 7). Two *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), *P. putida* (No. 6) were also considered to involve bronchoscopy as the source.

The data-mining program detected the correct routes of transmission on the eighth patient of the ERCP outbreaks and all other previous outbreaks on the second positive isolate of each outbreaks' 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 two to six (OR=28.7, p=0.02) on the second case (Figure 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=29.1 p=0.021). Table 1 displays the transmission routes which were both determined independently by infection prevention and 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 total, for the 2011-2016 outbreak requests, potentially 40 or 34 infections (71% or 60% of possible

preventable infections, respectively) could have been prevented based upon the 7- or 14-day intervention.

3.5 DISCUSSION

In this study, data mining of the EMR correctly identified transmission routes by the eighth patient of one outbreak and the second patient in all other eight outbreaks. If run in conjunction with routine molecular typing, up to 40 infections (71% of possible preventable infections) could have been prevented, assuming that proper intervention had occurred. Our results provide proof of concept that automated data mining can correctly identify routes of exposure in hospital outbreak investigations.

To our knowledge, this is the first reported study that combines molecular typing results and automated data mining of the EMR in hospital outbreak settings to identify routes of bacterial transmission. Current infection prevention methods rely on the infection preventionists or other clinicians to recognize an increased number of infections or geographic clustering of cases. Outbreaks that involve common hospital pathogens and/or less obvious transmission routes can be more difficult to detect. This is exemplified by our ERCP outbreak (No. 2), which was only identified 10 months after it began because it involved a common organism (K. pneumoniae) and patients admitted to multiple inpatient units after outpatient ERCP.3 Substantial time and labor must be spent on attempting to identify possible routes of transmission which may include sending isolates for WGS after the outbreak has already expanded.

There are several potential advantages of automated data mining over traditional approaches to hospital outbreak investigations. First, the EMR 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 non-outbreak 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 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. Second, the intervention delay of 7 or 14 days is based on hypothetical timelines that considered the time required to perform WGS, analyze data and enact appropriate interventions (e.g. removing a device from use, targeted environmental cleaning, 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 included a limited number of EMR variables in our analysis. However, in subsequent iterations we plan to expand the variables that are studied. Finally, automated data mining of the

EMR does not obviate the need for traditional “shoe leather” epidemiology for outbreak investigations. Additional efforts will often be required such as 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.

3.6 TABLES AND FIGURES

Table 3. Characteristics of outbreaks. The correct transmission route was identified by the data mining program for all outbreaks

No.	Date	Organism	Cluster: No. related isolates	Molecular typing method	Duration of transmission (days)	Transmission Route	Infections prevented: 7 day intervention	Infections prevented: 14 day intervention
1	Feb-12	*A. baumannii*	3	PFGE	19	Trauma ICU	1	1
2	Mar-13	*K. pneumoniae*	A: 28	PFGE	865	ERCP	20	20
			B: 2		3	ERCP	0*	0*
			C: 2		13	ERCP	0*	0*
			TOTAL: 36					
3	Jun-15	*K. pneumoniae*	10	PFGE	29	Bronchoscope	5	3
4	Jul-15	*P. aeruginosa*	10	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	16-Jun	*C. difficile*	A: 2	WGS	4	Trauma Floor	0*	0*
			B: 2		15	Post Anesthesia Unit	0*	0*
			C: 2		35	Pulmonology Floor	0*	0*
			TOTAL: 6					
9	16-Sep	*C. difficile*	4	WGS	67	Medical ICU	1	1
							TOTAL: 40	**TOTA**L: 34

*only 2 isolates; cannot prevent any infections

PFGE, pulsed field gel electrophoresis; WGS, whole genome sequencing; ICU, intensive care unit; ERCP, endoscopic retrograde cholangiopancreatography

(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

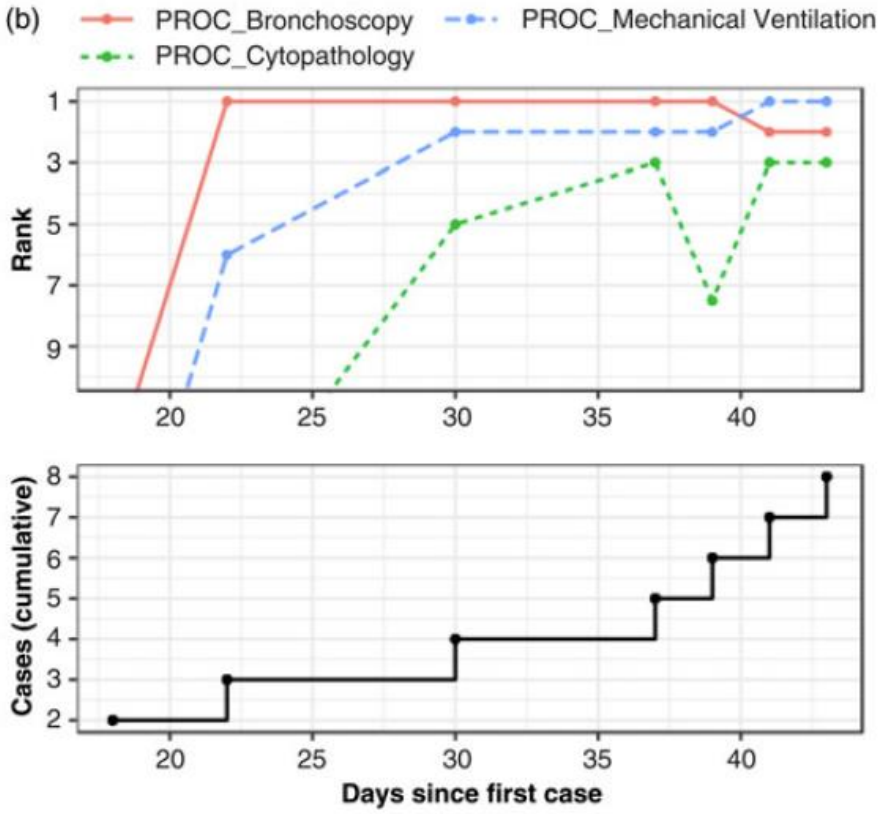


Figure 1. Transmission route ranking for outbreak No. 4: Pseudomonas aeruginosa from a contaminated bronchoscope

(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

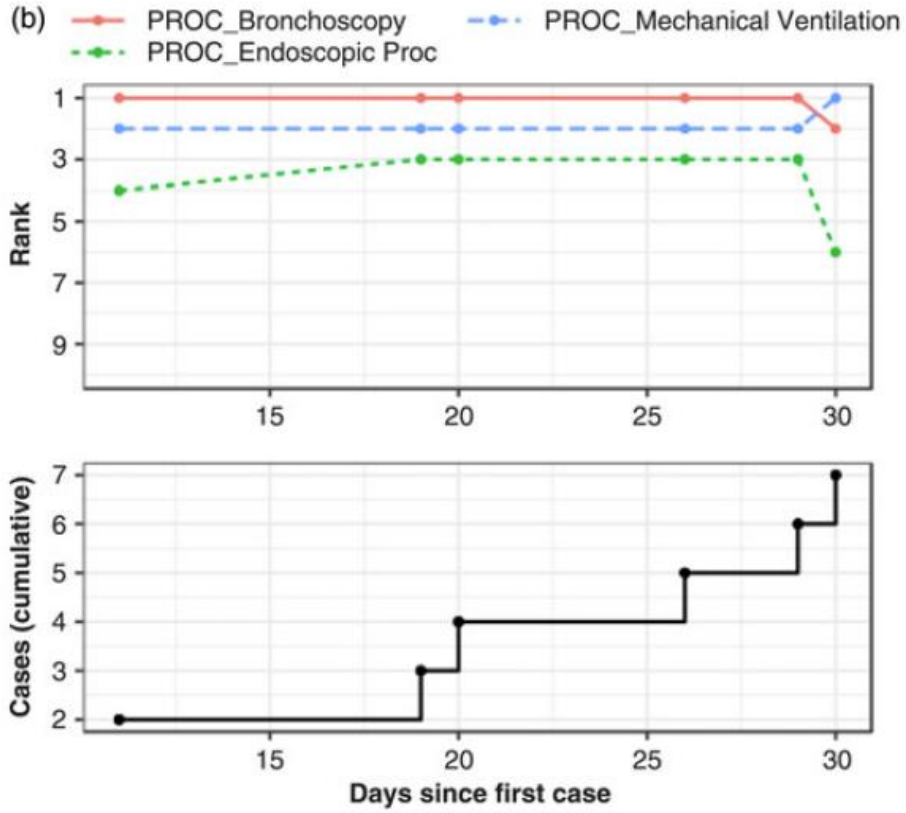


Figure 2. Transmission route ranking for outbreak no. 3: Klebsiella pneumoniae from a contaminated bronchoscope

4.0 MANUSCRIPT 3: WHOLE-GENOME SEQUENCING SURVEILLANCE AND MACHINE LEARNING OF THE ELECTRONIC HEALTH RECORD FOR ENHANCED HEALTHCARE OUTBREAK DETECTION

Alexander J. Sundermann, MPH, CIC,[1-3] Jieshi Chen, MS,[4] Praveen Kumar, B.Tech,[5] Ashley M. Ayres, BS, CIC,[6] Shu-Ting Cho, MS,[2] Chinelo Ezeonwuka, MSc,[1,2] Marissa P. Griffith, BS,[1,2] James K. Miller, PhD,[4] Mustapha M. Mustapha, MBBS, PhD,[1,2] A. William Pasculle, ScD,[7] Melissa I. Saul, MS,[8] Kathleen A. Shutt, MS,[1,2] Vatsala Srinivasa, MPH,[1,2] Kady Waggle, BS,[1,2] Daniel J. Snyder, MSc,[9] Vaughn S. Cooper, PhD,[9] Daria Van Tyne, PhD,[2] Graham M. Snyder, MD, SM,[2,6] Jane W. Marsh, PhD,[1,2] Artur Dubrawski, PhD, MSc,[4] Mark S. Roberts, MD,[5,8] and Lee H. Harrison, MD[1-3]

6. Microbial Genomic Epidemiology Laboratory, Center for Genomic Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.
7. Division of Infectious Diseases, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA.
8. Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
9. Auton Lab, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
10. Department of Health Policy and Management, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
11. Department of Infection Control and Hospital Epidemiology, UPMC Presbyterian, Pittsburgh, Pennsylvania, USA
12. Department of Pathology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

13. Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
14. Department of Microbiology and Molecular Genetics, and Center for Evolutionary Biology and Medicine, University of Pittsburgh School of Medicine, Pennsylvania, USA

4.1 ABSTRACT

Background: Most hospitals use traditional infection prevention (IP) methods for outbreak detection. We developed the Enhanced Detection System for Healthcare-Associated Transmission (EDS-HAT), which combines whole genome sequencing (WGS) surveillance and machine learning (ML) of the electronic health record (EHR) to identify undetected outbreaks and the responsible transmission routes, respectively.

Methods: We performed WGS surveillance of healthcare-associated bacterial pathogens from November 2016 to November 2018. EHR ML was used to identify the transmission routes for WGS-detected outbreaks, which were investigated by an IP expert. Potential infections prevented were estimated and compared to traditional IP practice during the same period.

Results: Of 3,165 isolates, there were 2,752 unique patient isolates in 99 clusters involving 297 (10.8%) patient isolates were identified by WGS; clusters ranged from 2-14 patients. At least one transmission route was detected for 65.7% of clusters. During the same time, traditional IP investigation prompted WGS for 15 suspected outbreaks involving 133 patients, for which transmission events were identified for 5 (3.8%). If EDS-HAT had been running in real-time, 25-63 transmissions could have been prevented. EDS-HAT was found to be cost-saving and more effective than traditional IP practice, with overall savings of $192,408 - $692,532.

Conclusion: EDS-HAT detected multiple outbreaks not identified using traditional IP methods, correctly identified the transmission routes for most outbreaks, and would save the hospital substantial costs. Traditional IP practice misidentified outbreaks for which transmission did not occur. WGS surveillance combined with EHR ML has the potential to save costs and enhance patient safety.

4.2 INTRODUCTION

Approaches for healthcare outbreak detection have remained essentially unchanged for decades.[68] When an outbreak is suspected, a method to establish genetic relatedness such as whole genome sequencing (WGS) may be performed. This approach can miss outbreaks and falsely identify suspected outbreaks that are refuted by WGS.

Although WGS surveillance has been useful for identifying otherwise undetected transmission events, identifying the responsible transmission route has had limited success. This is because investigations have focused primarily on geotemporal clustering which can miss complex transmission routes.[73,74]

In late 2016 we began development of the Enhanced Detection System for Healthcare-Associated Transmission (EDS-HAT), which combines WGS surveillance with machine learning (ML) of the electronic health record (EHR) to detect outbreaks and identify their routes of transmission.[75–79] We have found EHR ML useful for transmission routes that cannot be identified by traditional means.[75,76,78]

EDS-HAT was run with an at least six-month lag between infection and WGS so that its performance could be compared to our practice of using WGS in reaction to suspected outbreaks. We conducted a detailed analysis of EDS-HAT compared to traditional IP practice.

4.3 METHODS

4.3.1 Study Setting

This study was performed at the University of Pittsburgh Medical Center-Presbyterian Hospital (UPMC), an adult tertiary care hospital with 758 total beds, 134 critical care beds, and over 400 annual solid organ transplants. An independent chronic care facility with 32 beds is physically imbedded within UPMC. Transfer of patients between this facility and UPMC is common. Ethics approval was obtained from the University of Pittsburgh Institutional Review Board.

4.3.2 Isolate Collection

A description of the outbreak detection process is shown in Figure 1. For WGS surveillance, we collected select bacterial pathogens isolated from clinical specimens between November 2016 and November 2018: *Acinetobacter species*, *Pseudomonas* species, extended-spectrum beta-lactamase-producing [ESBL] *Escherichia coli*, *Klebsiella* species, *Clostridioides difficile*, ESBL Enterobacter species, vancomycin-resistant *Enterococcus* [VRE], methicillin-resistant *Staphylococcus aureus* [MRSA], *Stenotrophomonas* species, *Serratia* species,

Burkholderia species, *Legionella* species, *Providencia* species, *Proteus* species, and *Citrobacter* species. These pathogens were selected because they cause serious infections and healthcare-associated outbreaks. For *Clostridioides difficile*, we performed culture of stool specimens that were culture-independent diagnostic test-positive for *C. difficile*. Inclusion criteria were hospital admission or observation ≥3 days before the culture date and/or a recent inpatient or outpatient encounter in the 30-days before the culture date.

4.3.3 Whole-Genome Sequencing

WGS was performed on the NextSeq 500 platform (Illumina, San Diego, CA). Reads were assembled with SPAdes v3.13,[80] annotated with Prokka v1.14,[81] and multi-locus sequence types (STs) were assigned using PubMLST typing schemes (https://github.com/tseemann/mlst).[82]

Pairwise core genome single nucleotide polymorphisms (cgSNP) differences were calculated using snippy v4.3.0 (https://github.com/tseemann/snippy) within species STs having ≥2 isolates. Genetically related clusters were assigned using initial SNP cutoffs using hierarchical clustering with single linkage.[76,77] Based on our experience and the literature,[74,76,77,83–90] clusters were defined as isolates from >1 patient having ≤15 pairwise cgSNPs for all species except for *C. difficile*, for which ≤5 pairwise cgSNPs were used to identify clusters. For this organism, we defined clusters as all isolates that were within 0-2 cgSNPs regardless of whether a transmission route was identified and included cases that were within 3-5 cgSNPs of one another only if we could identify a statistically significant transmission route detected at 0-2 cgSNPs.

4.3.4 Extraction and Processing of Electronic Health Record Data

All patient encounters including inpatient, emergency room, and same day surgery were mined for charge transaction codes, clinical microbiologic data, admission data, discharge data, and length of stay.[75] Charge transaction codes were included because they reflect many types of exposures associated with transmission, such as medical procedures, medical services, and medications. Data were assigned a unique identification number using De-ID software (De-ID Data, Philadelphia, PA). The names of healthcare workers who signed clinical notes were also extracted and de-identified. Procedures with multiple charge codes were aggregated into groups for transmission route analysis.

4.3.5 Machine Learning Algorithm

A ML algorithm based on point estimates for model parameters and incorporating case-control methodology was used.[75,78] Case patients were defined as those with clinical isolates that clustered by WGS as defined above and control patients where all patients who were hospitalized in the 30-days prior to a case patients' culture date and did not have a positive result for the genetically related strain. Only route exposures on or prior to a case patient's culture date were considered.

The ML algorithm scores each outbreak by the maximum log-likelihood ratio of observing the case infections given that exposure to the principal transmission route probabilistically causes infection over the likelihood of a non-transmission explanation. A constant patient-to-patient transmission likelihood is added for each case infection not exposed to the principal transmission route. Empirical p-values are computed by estimating the likelihood of a higher outbreak score

given that no relationship exists between the case patients. This is done by sampling random sets of patients of equal size and computing their outbreak score maximized over routes. Importance sampling is used to improve efficiency of this process. Model parameters were fit using nine historical outbreaks between 2012-2016, which are separate from the analysis presented in this manuscript (Table S1). Parameter estimation was accomplished by transforming the outbreak detection problem into logistic regression as previously described.[75,78]

Transmission routes for clustered isolates with statistically significant odds ratios (OR) ($p<0.05$) from the algorithm for category types (e.g., procedures, locations, and providers) underwent manual EHR review for accuracy and biological plausibility. The manual EHR review was performed by an experienced infection preventionist (AJS), who subsequently reviewed the findings with two senior investigators (LHH and GMS), all who have experience in hospital epidemiology and outbreak investigation. The purpose of the manual EHR review was to determine the most likely transmission route predicted by the ML algorithm or investigate routes of transmission that that were not identified by the algorithm. For some clusters, more than one transmission route was considered plausible (e.g., transmission from a medical device with subsequent hospital unit-based transmission).

4.3.6 Clinical and Economic Modeling

Clinical and economic impact analysis was conducted from a hospital's perspective. The analysis utilized the transmission network of outbreaks, effectiveness of IP interventions by transmission route and time needed to implement IP interventions to estimate the expected number of transmissions under EDS-HAT, based on the method we previously described.[79] Since the effectiveness of IP interventions can decrease with time, we estimated lower and upper impact

boundaries, with the true value likely between these estimates. For the lower boundary, we assumed that effectiveness would decline linearly and measured effectiveness from the time when the IP team first intervened. The effect of subsequent IP interventions that would have been implemented whenever an additional patient was infected through same route was ignored. For the upper boundary, intervention effectiveness was assumed to remain constant. For outbreaks with more than one plausible transmission route, we weighted routes by the OR generated by the ML algorithm. If any route was missed by ML but detected by manual EHR review, we conservatively assigned the lowest OR score. Additionally, we performed a downstream cluster analysis to calculate the number of preventable infections if an intervention based on one outbreak could potentially prevent another outbreak using the same IP effectiveness parameters. For example, if EDS-HAT detected an outbreak in a hospital unit and an intervention was implemented, theoretically that intervention could prevent a subsequent outbreak.

Outcomes were incremental costs per transmission averted, number of readmissions averted, and lives saved. Probabilistic sensitivity analysis was conducted to assess the impact of uncertainty in parameter values of EDS-HAT. Data sources are described in Table S2. All costs were adjusted to 2020 using the medical component of the Consumer Price Index.[91] Costs and benefits were discounted at 3%. Readmissions at 7- and 30-days post-discharge were recorded. EHR review was performed to ascertain if readmissions were attributable to the infection; attributable readmissions were incorporated into the economic impact analysis.

4.3.7 Traditional Infection Prevention Practice

WGS was performed in reaction to IP requests (reactive WGS) for suspected outbreaks. For the two-year study period, the number of outbreaks detected by EDS-HAT versus traditional IP practice was determined.

4.4 RESULTS

Of 3,165 clinical isolates that underwent WGS, 2,752 unique patient isolates were clustered by ST. A total of 297 (10.8%) isolates representing 99 distinct, genetically related clusters ranging in size between 2-14 isolates were identified (Figure 2, Table 1). 269 (90.6%) of isolates were from inpatient cultures, 27 (9.1%) were from the emergency room, and 1 (0.3%) was from an outpatient visit. EDS-HAT detected potential transmission routes for 65 (65.7%) clusters containing 221 (74.4%) of the related isolates (Table S3). No significant transmission routes were detected by the EDS-HAT ML algorithm or manual review in the remaining 34 clusters, which ranged in size from 2-5 patients and contained 76 isolates. A brief description of high-impact or notable outbreaks and transmission routes detected by EDS-HAT ML is described in Table 2 while Table S3 describes all outbreaks.

4.4.1 Outbreaks detected by traditional IP practice

During the study period, our IP department requested reactive WGS for 15 suspected and potentially actionable outbreaks while EDS-HAT was running in parallel (2 *A. baumannii*, 1

Burkholderia cepacia, 6 *C. difficile*, 1 *K. pneumoniae* 3 *S. marcescens*, 2 *S. maltophilia*) involving 133 patients. Of these 15 suspected clusters, 5 (3.8%) patient isolates from 2 clusters (*A. baumannii* and *S. maltophilia*) were found to be genetically related. Of these 5 patients with related isolates, 2 of the transmissions involving *A. baumannii* were also detected by EDS-HAT.

4.4.2 Clinical and economic impact analysis

EDS-HAT could have prevented 25 (lower bound) to 63 (upper bound) transmissions. Moreover, 3.1-8.0 fewer 30-day attributable readmissions and 1.6-3.3 fewer deaths would have occurred had EDS-HAT been running in real time. Under EDS-HAT, the increase in cost of sequencing would be offset by cost savings in costs of treating infections, resulting in overall cost savings of $192,408 to $692,532 over the study period. EDS-HAT was found to be a more-effective and cost-saving program than traditional IP practice by providing savings of $7,745 - $10,939 for each transmission averted. Based on the lower bound estimates, EDS-HAT remained cost-saving and more effective in various independent scenarios: when the time needed for effective intervention was increased to 21 days, proportion of time spent towards outbreak detection under EDS-HAT was doubled (20%), effectiveness against procedures and healthcare workers was reduced to 30% (relative risk = 0.7), duration after which IP intervention's effectiveness would become zero was reduced to 13 weeks for all transmission routes except instruments, or the proportion of untreated cases was increased to 70% for respiratory, 50% for urine, 25% for wound or 10 % for stool. In probabilistic sensitivity analysis, EDS-HAT was cost-saving and more effective than traditional IP practice alone in more than 88% of simulations in lower and 99% in upper bound scenarios (Figure 3, Table S4).

4.5 DISCUSSION

In this study, we demonstrate the value of combining WGS surveillance with ML of the EHR for enhanced hospital outbreak detection. EDS-HAT detected consequential outbreaks and transmission routes that were undetected by traditional IP practice, whereas the latter mostly identified suspected outbreaks that were not confirmed by reactive WGS. Both components of EDS-HAT are essential: WGS surveillance is used to "connect the dots" between seemingly unrelated patients to signal an outbreak and ML, in combination with review by an IP expert, then identifies the responsible transmission route. In our study, we found that 10.8% of sequenced isolates were related which is in line with other studies of WGS surveillance.[30,31,41,47,83,87]

The results of our clinical and economic impact analysis suggest that, had it been running in real time, EDS-HAT would be highly cost-saving. The cost of sequencing one bacterial isolate is low ($70) relative to the high costs of treating a single, potentially-preventable infection (e.g., over $24,000 for *Pseudomonas* pneumonia). Recent budget and clinical impact analyses of WGS surveillance of multidrug-resistant pathogens in Australia also demonstrated that this approach is cost-saving.[92,93] Our analysis showed costs savings despite our conservative modeling assumptions which included the effectiveness of various types of interventions and the fact that we did not consider the cost of personal protective equipment and other costs associated with isolation precautions of patients. By using this conservative approach, we likely underestimated the true impact and cost savings of EDS-HAT.

The inability to demonstrate transmission routes that do not involve geotemporal clustering is a serious limitation of previous studies of WGS surveillance for outbreak detection in hospitals.[73,74] EDS-HAT overcomes this limitation by incorporating EHR ML.[94–96] Outbreaks that were detected exclusively by EDS-HAT tended to involve common hospital pathogens that lacked

geographic clustering and had transmission routes that were not readily apparent on manual EHR review. For example, the interventional radiology VRE outbreak identified a newly discovered procedural vulnerability, the outbreak of *Pseudomonas aeruginosa* affirmed known risks of endoscopy, outbreaks in the chronic care facility highlighted the problem of high risk transmission in this vulnerable patient population, the outbreak associated with wound care highlighted operational susceptibilities in the nature of care provided, and the cluster of MRSA associated with EEG and specific providers shows how EDS-HAT can detect unusual and specific routes.

Implementation of real-time WGS surveillance and ML of the EHR will require investment in healthcare infrastructure; the results of our economic analysis provide evidence that implementation can be cost-saving for hospitals that perform reactive WGS. Parcell et al highlight barriers to implementation and methods for integration into infection prevention practice.[62] We view EDS-HAT as complementary to infection prevention practice because it alerts of possible outbreaks, which prompts additional investigating and intervention. EDS-HAT requires input from infection preventionists to evaluate the transmission routes that are generated and determine what interventions are needed.

There are several limitations to our study. First, it is unlikely that all outbreaks and outbreak patients were captured in this study, because, for example, some infected patients may not have cultures taken or cultures may have been negative because of recent antibiotic administration. In addition, our exclusion of cultures during the first three days of hospitalization likely led us to miss transmission events. Second, we did not include surveillance swabs, meaning that we likely missed transmission events for, for example, VRE. Third, the retrospective nature of the study did not allow us to investigate and confirm potential transmission routes for some of our outbreaks; this limitation can be alleviated and the potential impact will likely increase when EDS-HAT is run in

real-time. Fourth, during this two-year evaluation, we had fewer transmissions identified by traditional IP practice at our institution than usual.[75,97,98] However, EDS-HAT would likely have detected any IP-identified outbreak more quickly. Fifth, our economic modeling of real-time interventions may not reflect true intervention effectiveness and timeliness. However, we adjusted for both conservative and loose parameters to estimate the true effectiveness in between those bounds. Sixth, we did not account for potential asymptomatic carriage of urinary and wound cultures in our model. However, IP would intervene regardless of clinical presentation given it would aid in interrupting future transmission. In addition, many of these positive cultures are treated and, therefore, incur costs, whether the treatment is appropriate or not. Finally, we included only a limited number of pathogens in WGS surveillance because of feasibility and cost and therefore likely missed outbreaks caused by other pathogens.

Advances in microbial genomics and bioinformatics, digitalization of healthcare data, and machine learning technology have made enhanced outbreak detection in hospitals feasible. Taken together, our results suggest that EDS-HAT represents a potential paradigm shift in how outbreaks are detected in hospitals. If instituted in real time, this approach can reduce healthcare-related costs and significantly improve patient safety.

4.6 FIGURES AND TABLES

Table 4. EDS-HAT isolates sequenced and attributable readmissions

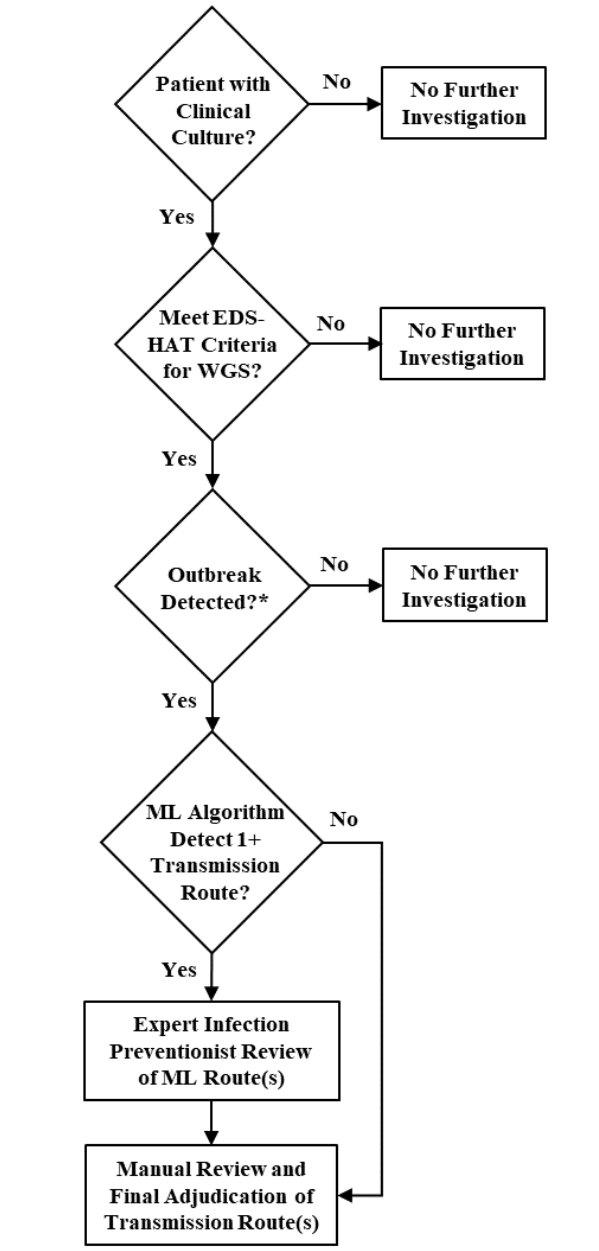
	Sequenced				Attributable Readmissions	
Species	**Collected**	**Unique Patient Isolates**	**No. Related (%)**	**Clusters**	**7-day**	**30-day**
Acinetobacter species	83	72	12 (16.7)	3	1	1
Burkholderia species	12	12	0 (0)	0	0	0
Citrobacter species	126	118	2 (1.7)	1	0	0
Clostridioides difficile	558	524	80 (15.3)	21	2	10
Escherichia coli (ESBL)	170	149	10 (6.7)	4	0	1
Klebsiella species (ESBL, not *pneumoniae*)	25	20	0 (0)	0	0	0
Klebsiella pneumoniae (ESBL)	111	102	27 (26.5)	8	0	1
Legionella species	1	1	0 (0)	0	0	0
Methicillin-resistant *Staphylococcus aureus*	425	365	39 (10.7)	18	1	5

	Sequenced				Attributable Readmissions	
Species	**Collected**	**Unique Patient Isolates**	**No. Related (%)**	**Clusters**	**7-day**	**30-day**
Proteus species	151	140	2 (1.4)	1	0	0
Providencia species	14	13	0 (0)	0	0	0
Pseudomonas aeruginosa	881	693	31 (4.5)	10	2	3
Pseudomonas species (not aeruginosa)	28	27	0 (0)	0	0	0
Serratia species	181	173	14 (8.1)	7	1	3
Stenotrophomonas species	127	114	4 (3.5)	2	0	0
Vancomycin-resistant *Enterococcus faecalis*	17	17	0 (0)	0	0	0
Vancomycin-resistant *Enterococcus faecium*	247	212	76 (35.8)	24	5	16
Total	**3165**	**2752**	**297 (10.8)**	**99**	**12**	**40**

Table 5. High-impact or notable outbreaks detected by EDS-HAT

Outbreak	Details
Vancomycin-resistant *Enterococcus faecium* outbreak associated with interventional radiology (IR) and injection of sterile contrast[6]	This outbreak involved ten initial patients and was ongoing when it was discovered. The EDS-HAT ML algorithm identified IR as a significant transmission route (OR 43.8; p-value <0.01; 95% confidence interval [CI], 5.6 to 346). Nine patients, including three with bacteremia, were identified as having IR procedures involving unsterile practices in the preparation of contrast. Safe practices and enhanced environmental cleaning were implemented and no additional IR-associated infections occurred. Subsequently, transmission of the outbreak strain occurred among four patients on shared hospital units.
Pseudomonas aeruginosa outbreak associated with gastroscopy[5]	This outbreak comprised six patients housed on different units over seven months. Two patients had bacteremia, three had pneumonia, and one had a urinary tract infection. The EDS-HAT ML algorithm detected gastroscopy as a significant route for four patients (OR 300.6; p-value <0.01; 95% CI, 15.8 to 5690.5) with a fifth patient who did not have a charge code that reflected the gastroscopy procedure but who had a clinical note reflecting the procedure that was identified on manual EHR review. A post-disinfection gastroscope culture performed as part of routine IP practice was positive for *P. aeruginosa*; the isolate was sequenced and belonged to the outbreak, confirming gastroscopy as the responsible transmission route.
Outbreaks of multiple pathogens at the imbedded chronic care facility	EDS-HAT ML identified 11 clusters involving 38 patients over 22 months, with a range 2-9 total patients per cluster; 25 (65.8%) patients had this facility as a plausible transmission route. Pathogens included *C. difficile* (6 clusters), *K. pneumoniae* (1 cluster), MRSA (1 cluster), *P. aeruginosa* (2 clusters), and VRE (1 cluster). Three *C. difficile* patients in three clusters were subsequently transferred to our institution and had unit-based commonalities with three additional patients who later developed *C. difficile* infection suggesting continuing transmission.
Outbreaks of multiple pathogens on an intensive care unit (ICU)	There were 12 clusters with 57 patients (range 2-14), of whom 28 (49.1%) had a single ICU stay identified by EDS-HAT ML as the potential transmission route. Organisms included *C. difficile* (3 clusters involving 10 patients), *K. pneumoniae* (3 clusters involving 16 patients), *P. aeruginosa* (1 cluster involving 3 patients),

Outbreak	**Details**
	Serratia marcescens (1 cluster involving 2 patients), and VRE (4 clusters involving 26 patients).
C. difficile outbreaks associated with wound care	There were 9 *C. difficile* clusters, ranging in size from 2-12 patients. Of 52 patients, 29 (55.8%) had wound care service identified as a potential transmission route, with exposures occurring 1-92 days (mean 16 days, median 9 days) before the positive test for *C. difficile*. This consult service involved nurses providing management of sacral pressure ulcer wounds.
MRSA infections associated with electroencephalography (EEG)	This cluster consisted of two patients with culture dates separated by 8 days. The EDS-HAT ML algorithm identified EEG as a transmission route. Manual EHR review determined that both patients had a bedside EEG performed on the same day on separate units by the same physician and technician, two and ten days before positive culture dates.



*As described in methods
EDS-HAT: Enhanced Detection System for Healthcare-Associated Transmission; WGS: Whole Genome Sequencing; ML: Machine Learning; IP: Infection Preventionist

Figure 3. Flow diagram of the EDS-HAT outbreak detection process, from clinical culture through adjudication of transmission route(s)

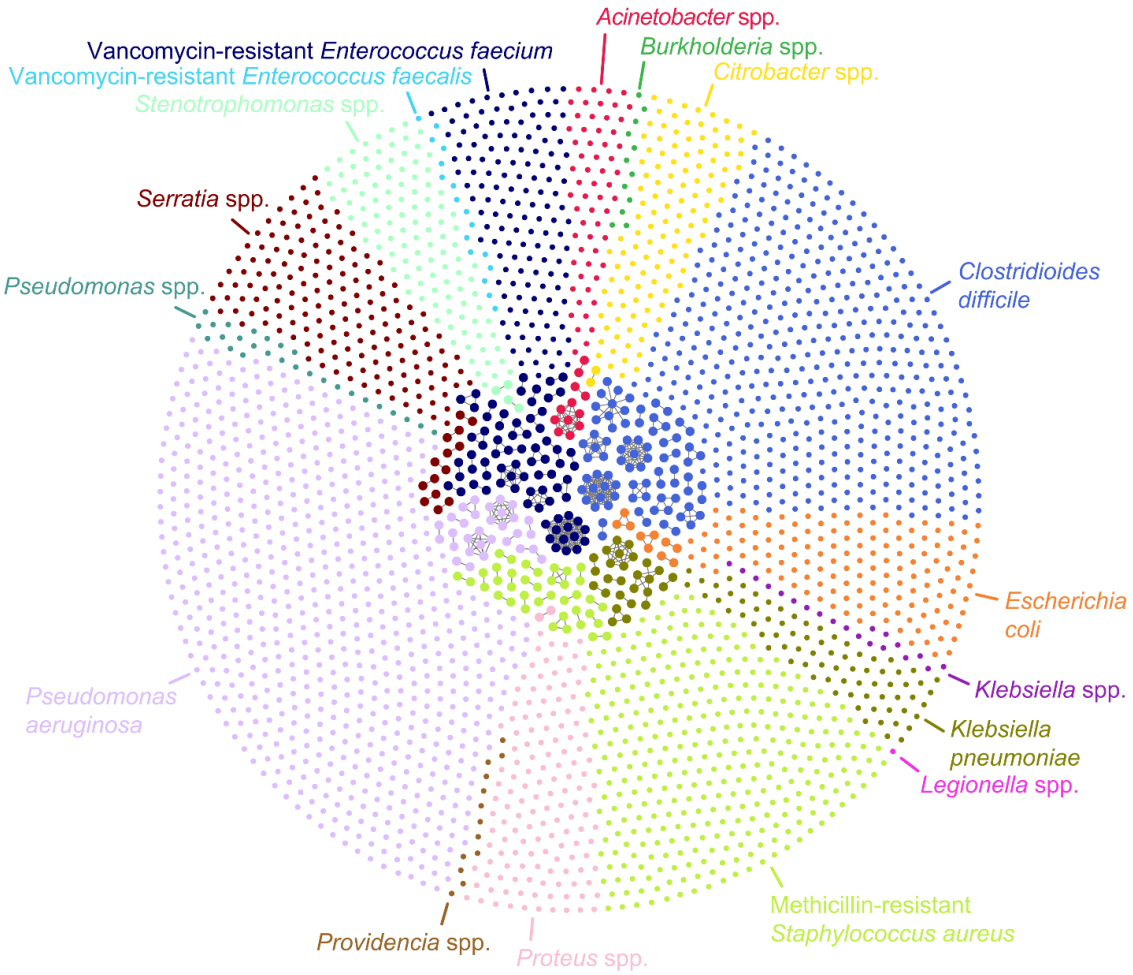


Figure 4. Cluster network of EDS-HAT isolates sequenced, grouped by bacterial species. The outer circle shows patient isolates that are not genetically related. The inner circle shows outbreaks of genetically related isolates as defined by cgSNP cut-offs describe

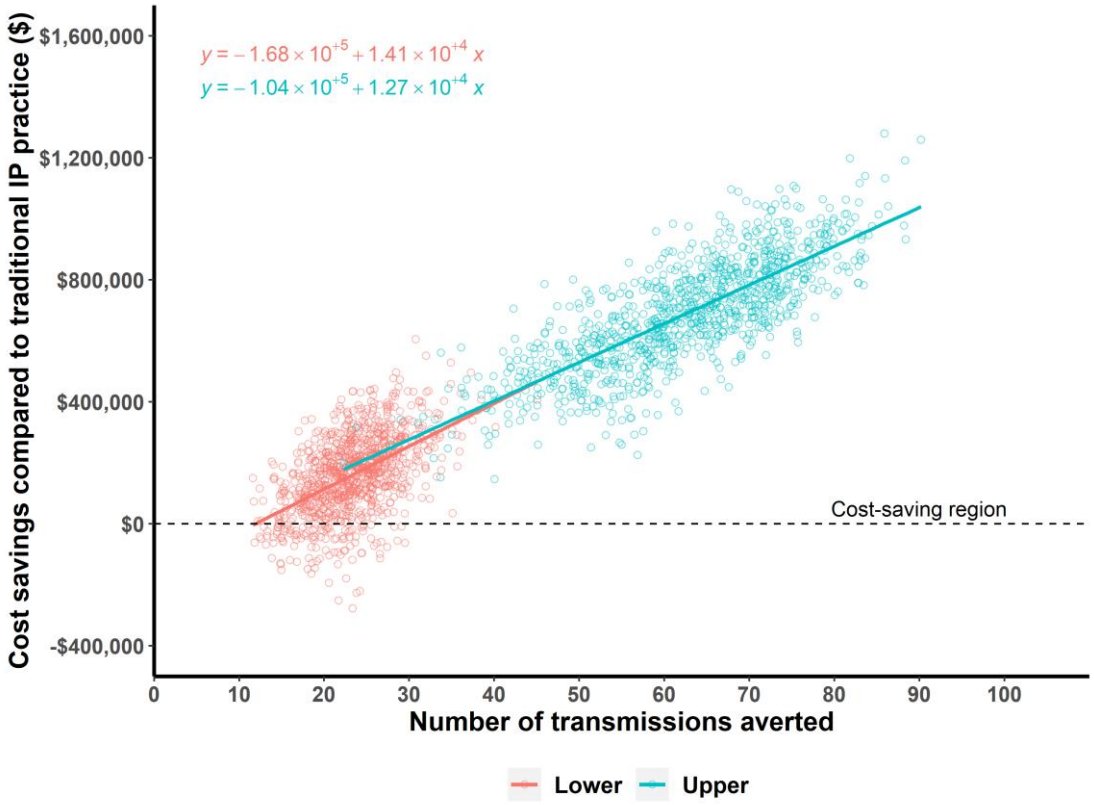


Figure 5. EDS-HAT cost-savings and effectiveness plot for estimated lower and upper bound boundaries (see Methods). Cost-savings of EDS-HAT was examined by estimated costs associated with number of transmissions averted, using 1,000 simulations in probabilistic s

5.0 CONCLUSION

5.1 MAJOR FINDINGS

Whole genome sequencing surveillance and machine learning of the electronic health record has the potential to significantly enhance healthcare outbreak detection and investigation. The decrease costs of WGS and availability of EHR data has provided an opportunity for healthcare systems to leverage these data to improve patient safety. The evidence presented in this dissertation supports its use through multiple analysis: a systematic literature review of the potential impact of WGS surveillance, a proof of concept design of EDS-HAT, and application of the tool in our hospital over two years.

The first chapter provides a systematic review of institutions that published on utilizing WGS surveillance and/or machine learning to investigate and detect outbreaks. The results of the review conclude that institutions who have implemented WGS surveillance have been able to find previously undetected outbreaks, better understand transmission dynamics within their facility, and better estimate the true rates of pathogen transmission. Yet the latter part of the review shows that not many institutions have significantly leveraged the availability of EHR data to complement investigations. The combination of these technologies may be superior to using one single technology alone. The use of WGS surveillance is emerging with more publications each year. Institutions should look to prepare to or adopt this technology.

The next chapter provided a proof of concept for EDS-HAT by showing WGS and a data mining algorithm can early detect previously well-described outbreaks at our institution. Additionally, using past outbreaks that went undetected often for months was vital in the teaching

the machine learning model. This enabled us to set model parameters for an algorithm that was based upon well-defined outbreaks for application in a larger analysis.

Lastly, the final chapter evaluated a full-scale implementation of EDS-HAT over a 2-year period in comparison with traditional infection prevention methods. We found, if implemented in real-time, EDS-HAT would be highly cost-savings and prevent transmissions and deaths in patients from HAIs. Moreover, we compared these results to the traditional infection prevention results which showed superiority in EDS-HAT.

Taken together, the evidence presented in this dissertation strong supports the further research and possible implementation of WGS surveillance and machine learning of the EHR. This research is able to depict the clear limitations of traditional infection prevention methods while showing the utility in new, emerging technologies that can significantly enhance healthcare institutions' ability to detect and stop transmission.

5.2 FUTURE DIRECTIONS

A real-time application of EDS-HAT is necessary to fully evaluate its effectiveness in detecting and stopping the spread of outbreaks. As of the submission of this dissertation, our institution has started WGS surveillance and is in the process of building a real-time process for machine learning of EHR data. Our initial findings on the use of WGS surveillance support that of our two-year analysis findings in that it is able to detect outbreaks at two patients and often when traditional infection prevention practices has not detected anything. However, a formal analysis of this approach will entail measuring subsequent cases of the same route in an outbreak after an

intervention by the infection prevention department. A successful tool will show that outbreaks can be terminated through interventions on the implicated transmission route.

Large datasets for a real-time application of EDS-HAT will require a graphical user interface to better interact and explore the data compared to more manual evaluation methods within this dissertation. The development of such interface will be key to expansion in other hospitals within a system given the nature of patient movement as well. Other institutions that are seeking to implement a program like EDS-HAT should consider an interface for evaluation.

This dissertation was performed at one large, academic, tertiary care hospital which results may not be generalizable to all healthcare settings given the acuity of patients and nature of invasive procedures. Other institutions should carry out additional studies to independently conclude their findings of a WGS surveillance tool and/or use of EHR data for outbreak detection. We will seek to expand EDS-HAT as a real-time tool at our hospital and subsequently to other hospitals if proven effective in real-time. We aim to evaluate its impact at each hospital, with the goal of improving patient safety.

EDS-HAT enables institutions to better and more accurately detect outbreaks. However, often the implemented interventions remain unchanged. For example, when a unit-based transmission route is suspected, an infection preventionist performs education, hand hygiene audits, and interview staff. This does not change with detection of highly suspected transmission via EDS-HAT. Additional studies are being planned to explore the application of targeted environmental cultures into the workflow of a real-time EDS-HAT program. This approach has the potential to better detect the exact mechanism of unit-based transmission. For example, if a unit-based transmission were suspected, and targeted environmental cultures supported

transmission via a reusable medical equipment, resources could be focused on enhancing cleaning of that equipment and thus increasing intervention effectiveness.

Additionally, this dissertation has discussed the role and limitations of geo-temporal clustering as a primary approach for outbreak detection. Often, institutions assume that patients are only infected and contagious with bacterial pathogens for a short duration (<30 days) after the initial clinical infection. Therefore, a patient will not be assumed to be a potential source of infection months after the initial infection. Preliminary EDS-HAT data show that patients can have repeated infections with the same bacterial strain hundreds of days after the initial infection. This evidence would refute traditional infection prevention practice and may change approaches for managing chronically-infected patients to prevent transmission to others. Additional studies in this topic are being planned as well.

Further, surveillance definitions of infections have been created by the CDC National Healthcare Safety Network to streamline workflows, standardize across institutions, and track the overall burden of healthcare-associated infections. Preliminary data from EDS-HAT show that the National Healthcare Safety Network's definitions do not capture >50% of transmissions within our institution. These findings are concerning given healthcare institutions often rely off of these definitions to track issues as well as for the use of national statistics on the burden of HAIs. Additional studies are being conducted to fully understand the limitations of these definitions at our institution that are likely relevant to other healthcare facilities.

6.0 PUBLIC HEALTH SIGNIFICANCE

Healthcare-associated infections are a global issue that cause high morbidity and mortality. Many of these HAIs are becoming more resistant to antimicrobials which heightens the need to detect outbreaks and prevent HAIs. HAIs also contribute to transmission and outbreaks within healthcare settings. Healthcare institutions often rely on using busy bedside clinicians to detect suspected outbreaks. After initiation of a suspected outbreak, infection preventionists are then tasked with a full investigation and the occasional use of reactive WGS. This process is time consuming and requires a significant allocation of resources. Moreover, oftentimes the conclusion of these investigations provide no actionable evidence to intervene or even confirmation of transmission.

Often stated in quality improvement investigations is “it’s how we’ve always done it.” Healthcare institutions may be significantly underestimating the amount of infection transmission by continuing to use antiquated approaches for outbreak detection and investigation. Public health requires institutions to constantly adapt to a changing landscape and new approaches. EDS-HAT aims provide the knowledge base on the use of emerging, new technologies so that healthcare institutions can incorporate new approaches for solving problems.

When applying this work to a national or global landscape, it is important to think about a feasible expansion and rollout to other healthcare institutions. How will underserved areas without sequencing capabilities apply this program? How do you incentivize healthcare institutions to actually uncover cryptic transmission and issues in their facilities? How do you create a data system across multiple centers that could communicate the genomic findings quickly? These are all questions of public health significance that should be considered. Clearly communication of

the benefits and providing incentives for healthcare facilities via health policy will be key factors. Policy makers should highlight the persistent burden that HAIs have set on healthcare facilities and use this research and others to show a need to shift health policy to preventative measures rather than reactive. Additionally, further research is needed on the ethical use and implications of WGS surveillance. Questions should address how to properly counsel patients and healthcare workers involved in transmission.

In conclusion, the results of this dissertation show the traditional method of outbreak detection often misidentifies suspected transmission and misses a significant amount of transmission where it did occur. Adoption of WGS surveillance and EHR machine learning may be able to help healthcare institutions to alleviate the detection burden from clinicians, provide more accurate outbreak investigations, reduce the incidence of healthcare associated infections, and thus improve patient safety.

Appendix A Tables Whole Genome Sequencing Surveillance and Machine Learning for Healthcare Outbreak 1 Detection and Investigation: A Systematic Review and Summary

Appendix Table 1. Details of studies utilizing whole genome sequencing surveillance

Year	First Author	ML/DM/ Model	Real-Time	Duration (Months)	Sequence Method	SNP Cutoff	Methods Used for Epi Connections	Organism(s)	Type	# Unique Isolates	# Related	% Related	# Clusters	Avg. # / Cluster	# No Epi Link	% No Epi Link	Transmission Route
2020	Sundermann	Yes	No	24	WGS	15	Machine learning, procedures, geo-temporal, providers	*P. aeruginosa*	Sequence type 27	882	31	3.5	10	3.1	1	3.2	Gastroscopes
2020	Sundermann	No	No	12	WGS	15	Geo-temporal, procedures	*E. faecium*	Vancomycin-resistant, sequence type 1471	439	10	2.3	Not Provided	NA	1	10.0	Interventional radiology
2021	Meredith	No	Yes	1	WGS	0	Geo-temporal	SARS-CoV-2	.	299	159	53.2	35	4.5	35	22.0	Ward-based transmission
2021	Rose	No	No	6	WGS	40	Geo-temporal, procedures	*S. aureus*	Methicillin-resistant	56	15	26.8	7	2.1	7	46.7	Ward-based transmission
2020	Tsujiwaki	No	No	12	WGS	10	Geo-temporal	*S. aureus*	Methicillin-resistant	57	19	33.3	5	3.8	0	0.0	Ward-based transmission
2021	Miles-Jay	No	No	48	WGS	14	Geo-temporal	*E. coli*	Sequence Type 131 H30	126	17	13.5	8	2.1	9	52.9	Hospital-based transmission Potential procedure
2020	Gona	No	No	12	WGS	20	Geo-temporal	*K. pneumoniae*	.	80	39	48.8	10	3.9	14	35.9	Ward-based

Year	First Author	ML/DM/ Model	Real-Time	Duration (Months)	Sequence Method	SNP Cutoff	Methods Used for Epi Connections	Organism(s)	Type	# Unique Isolates	# Related	% Related	# Clusters	Avg. # / Cluster	# No Epi Link	% No Epi Link	Transmission Route
																	transmission
2020	Hammerum	No	No	54	MLST	NA	Geo-temporal	*K. pneumoniae*	.	103	36	35.0	13	2.8	11	30.6	Ward-based transmission
2020	Cremers	No	No	24	WGS	3	Geo-temporal	*S. aureus*	Methicillin-sensitive	84	40	47.6	14	2.9	0	0.0	Ward-based transmission
2019	Mathur	No	No	22	WGS	10	Geo-temporal	*K. pneumoniae*	Colistin-resistant	21	8	38.1	4	2.0	0	0.0	Ward-based transmission
2020	Marmor	No	No	60	WGS	49	Geo-temporal	*E. cloacae*	.	63	7	11.1	1	7.0	0	0.0	Ward-based transmission
								C. freundii	.		10	15.9	1	10.0	0	0.0	Ward-based transmission
2019	Hall	No	No	3	WGS	39	Geo-temporal	*S. aureus*	Methicillin-resistant	55	27	49.1	12	2.3	8	29.6	Ward-based transmission
2019	Sullivan	No	No	16	WGS	7	Geo-temporal, equipment	*S. aureus*	Methicillin-resistant	141	28	19.9	4	7.0	2	7.1	Ward-based transmission and ventilators
2019	Eigenbrod	No	No	38	WGS	14	Geo-temporal	*A. baumannii*	.	39	15	38.5	4	3.8	5	33.3	Ward-based transmission
2019	Stenmark	No	No	360	WGS	2	Geo-temporal	*S. capitis*	Bloodstream infections	46	12	26.1	6	2.0	12	100.0	Unknown
2019	Sherry	No	No	60	WGS	23	Geo-temporal	Enterobacteriaceae	Carbapenemase-producing	291	53	18.2	12	4.4	8	15.1	Ward-based transmission
2019	Wang	No	No	7	WGS	2	Geo-temporal	*C. striatum*	.	91	18	19.8	6	3.0	3	16.7	Ward-based

Year	First Author	ML/DM/ Model	Real-Time	Duration (Months)	Sequence Method	SNP Cutoff	Methods Used for Epi Connections	Organism(s)	Type	# Unique Isolates	# Related	% Related	# Clusters	Avg. # / Cluster	# No Epi Link	% No Epi Link	Transmission Route
																	transmission
2017	Coll	No	No	12	WGS	50	Geo-temporal	*S. aureus*	Methicillin-resistant	1465	785	53.6	173	4.5	187	23.8	Hospital-based or community-based transmission
2019	Jakharia	No	No	12	WGS	10	Geo-temporal	*C. difficile*	.	45	4	8.9	2	2	4	100	NA
2019	García-Fernández	No	No	36	WGS	2	Geo-temporal	*C. difficile*	.	367	41	11.2	6	6.8	34	82.9	Ward-based transmission
2019	Ward	No	No	12	WGS	12	Geo-temporal, procedures, providers	*S. aureus*	.	953	85	8.9	28	3.0	65	76.47	Ward-based
						10		*E. faecium*	.	86	13	15.1	5	2.6	9	69.23	Ward-based, providers
						30		*P. aeruginosa*	.	118	2	1.7	1	2.0	2	100.00	NA
						15		*K. pneumoniae*	.	100	0	0.0	0	0.0	0	0	NA
2019	Roy	No	No	3	WGS	3	Geo-temporal	Influenza	A H1N1	36	5	13.9	2	2.5	2	40.0	Ward-based transmission
2019	Eyre	No	No	6	WGS	2	Geo-temporal	*C. difficile*	.	299	43	14.4	6	7.2	20	46.5	Ward-based transmission
2018	Wendel	No	No	12	MLST	NA	Geo-temporal	*A. baumannii*	.	36	20	55.6	2	10.0	2	10.0	Ward-based transmission
2018	Houldcroft	No	No	66	WGS	Not Provided	Geo-temporal	Adenovirus	.	43	6	14.0	2	3	0	0.0	Ward-based transmission
2018	Donskey	No	No	6	WGS	2	Geo-temporal	*C. difficile*	.	66	12	18.2	4	3	4	33.3	Ward-based transmission
2018	Leong	No	No	36	WGS	10	Geo-temporal	*E. faecium*	Vancomycin-resistant	80	10	12.5	2	5	3	30.0	Ward-based

Year	First Author	ML/DM/ Model	Real-Time	Duration (Months)	Sequence Method	SNP Cutoff	Methods Used for Epi Connections	Organism(s)	Type	# Unique Isolates	# Related	% Related	# Clusters	Avg. # / Cluster	# No Epi Link	% No Epi Link	Transmission Route
																	transmission
2018	Kwong	Yes	Yes	48	WGS	30	Geo-temporal, procedures	*K. pneumoniae*	Carbapenemase-producing	86	53	61.6	4	13.3	10	18.9	Ward-based transmission
2018	Auguet	No	No	4	WGS	10	Geo-temporal	*S. aureus*	Methicillin-resistant	610	261	42.8	90	2.9	13	5.0	Ward-based transmission
2017	Gorrie	No	No	12	WGS	25	Geo-temporal	*K. pneumoniae*	.	106	17	16.0	5	3.4	0	0.0	Ward-based transmission
2016	Elbadawi	No	No	4	PFGE / WGS	2	Geo-temporal	*K. pneumoniae*	Carbapenemase-producing	46	4	8.7	1	4	0	0.0	Ward-based transmission
2019	Harada	No	No	4	WGS	Not Provided	Geo-temporal	*K. pneumoniae*	Bloodstream infections	140	2	1.4	1	2.0	2	100.0	Unknown
2019	Kossow	No	No	6	MLST	Not Provided	Geo-temporal	*S. aureus*	Methicillin-resistant	Not Provided	8	NA	1	8.0	0	0.0	Ward-based transmission
2019	van Beek	No	No	57	cgMLST	Not Provided	Geo-temporal	*K. pneumoniae*	Carbapenemase-producing, sequence type 512	Not Provided	20	NA	2	10.0	4	20.0	Ward-based transmission
2013	Eyre	No	No	42	WGS	2	Geo-temporal	*C. difficile*	.	957	333	34.8	Not Provided	NA	152	45.6	Ward-based transmission
2018	Martin	No	No	20	WGS	2	Geo-temporal	*C. difficile*	.	640	227	35.5	Not Provided	NA	69	30.4	Ward-based transmission
2015	Roach	No	No	12	WGS	40	Geo-temporal	*S. epidermidis*	.	178	56	31.5	10	5.6	Not Provided	NA	Ward-based transmission
								P. aeruginosa	.	44	7	15.9	3	2.3	Not Provided	NA	

Year	First Author	ML/DM/ Model	Real-Time	Duration (Months)	Sequence Method	SNP Cutoff	Methods Used for Epi Connections	Organism(s)	Type	# Unique Isolates	# Related	% Related	# Clusters	Avg. # / Cluster	# No Epi Link	% No Epi Link	Transmission Route
								E. faecium	.	36	13	36.1	3	4.3	Not Provided	NA	
								S. aureus	.	118	4	3.4	2	2.0	Not Provided	NA	
								E. faecalis	.	72	6	8.3	3	2.0	Not Provided	NA	
								S. maltophilia	.	58	2	3.4	1	2.0	Not Provided	NA	
2014	Long	No	No	6	WGS	40	Geo-temporal	*S. aureus*	.	305	0	0	0	0	Not Provided	NA	NA
2017	Raven	No	No	73	WGS	23	Geo-temporal	*E. faecium*	Bloodstream infections	293	93	31.7	6	15.5	Not Provided	NA	Ward-based transmission
2020	Berbel Caban	No	No	24	WGS	15	Geo-temporal, procedures, providers	*S. aureus*	Methicillin-resistant	224	33	14.7	8	4.1	Not Provided	NA	Unit-based transmission Vascular access device
2017	Eyre	No	No	12	WGS	2	Geo-temporal	*C. difficile*	.	652	128	19.6	Not Provided	NA	Not Provided	NA	Hospital-level transmission
2020	Neuman n	No	No	12	WGS	Not Provided	Geo-temporal	*E. faecium*	Vancomycin-resistant	111	Not Provided	NA	Not Provided	NA	Not Provided	NA	Ward-based transmission

Appendix B Tables & Figures: Whole-Genome Sequencing Surveillance and Machine Learning of the Electronic Health Record for Enhanced Healthcare Outbreak Detection

Appendix Table 2. List of model parameters

Model Parameters
1. Log probability of non-HAT infection: 1e-4
2. Log probability of within-outbreak transmission by means other than the "route" e.g. secondary transmission pathways patient to patient: 1e-2
3. Single nucleotide polymorphism distance: range from 0 to 15
4. Analysis window length: Given an outbreak of isolates, start from 30 days before the first culture date, and end on the last culture date.

Appendix Table 3. Data inputs for clinical and economic modeling

Variable	**Mean**	**95% CI[a]**	**Distribution**	**Source**
Effectiveness related parameters				
Time between culture date and implementation of IP intervention under EDS-HAT	9 days	3-17 days	Gamma	Assumption
Effectiveness (Relative risk) of intervening against transmission routes				
Healthcare worker	0.50	0.27 – 0.85	Lognormal	Assumption
Instrument	0.00		Not varied	Assumption
Inpatient unit	0.70	0.50 – 0.98	Lognormal	[1]

Variable	Mean	95% CI[a]	Distribution	Source
Procedure	0.10	0.02 – 0.34	Lognormal	Assumption
Unknown	1.00		Not varied	Assumption
Duration of IP intervention [b]				
Healthcare worker	26 weeks	7 – 57	Gamma	Assumption
Instrument	Always		Not varied	Assumption
Inpatient unit	26 weeks	7 – 57	Gamma	Assumption
Procedure	52 weeks	14 – 114	Gamma	Assumption
% colonized respiratory cultures [c]	49%	38% - 60%	Beta	[2]
Attributable mortality risk due to infection				
Pneumonia	0.143	0.142-0.145	Beta	[3]
Wound	0.028	0.028-0.029	Beta	[3]
Urinary tract	0.023	0.023-0.024	Beta	[3]
Bacteremia	0.123	0.122-0.125	Beta	[3]
Clostridioides difficile	0.030	0.029-0.031	Beta	[4]
Risk of 30-day readmission due to infection				
Pneumonia	0.100	0.038-0.187	Beta	Unpublished data
Wound	0.219	0.127-0.327	Beta	Unpublished data
Urinary tract	0.098	0.033-0.192	Beta	Unpublished data

Variable	Mean	95% CI[a]	Distribution	Source
Bacteremia	0.119	0.041-0.231	Beta	Unpublished data
Clostridioides difficile	0.125	0.062-0.205	Beta	Unpublished data
Cost related parameters [d]				
Annual salary of an IP professional	$95,700	$92,315 - $99,086	Normal	[5]
Number of IP professionals in IP team				
Traditional IP practice	8		Not varied	Unpublished data
EDS-HAT	8		Not varied	Assumption
% time spent on outbreak investigations				
Traditional IP practice	10%	3% - 22%	Beta	Unpublished data
EDS-HAT	10%	3% - 22%	Beta	Assumption
Cost of performing WGS per isolate				
Traditional IP practice	$70	$57 - $84	Gamma	Unpublished data
EDS-HAT	$70	$57 - $84	Gamma	Unpublished data
Number of isolates sequenced per year				
Traditional IP practice	129	105 – 155	Gamma	Unpublished data
EDS-HAT	1,300	1,058-1,567	Gamma	Unpublished data
Cost of treating infection [e, f]				
Klebsiella Pneumonia (J15.0)	$21,096	$18,292 - $24,096	Gamma	[6]
Pseudomonas Pneumonia (J15.1)	$24,301	$22,512 - $26,157	Gamma	[6]

Variable	Mean	95% CI[a]	Distribution	Source
MRSA Pneumonia (J15.212)	$22,700	$21,220 - $24,229	Gamma	[6]
Escherichia coli Pneumonia (J15.5)	$22,058	$18,027 - $26,490	Gamma	[6]
Pneumonia due to VRE (J15.8)	$19,021	$16,121 - $22,156	Gamma	[6]
Other Pneumonia (J15.6)	$13,277	$12,657 - $13,912	Gamma	[6]
Wound (T81.4XXA)	$16,970	$16,587 - $17,357	Gamma	[6]
Urinary tract (N39.0)	$7,815	$7,700 - $7,932	Gamma	[6]
Bacteremia (R78.81)	$13,172	$12,419 - $13,946	Gamma	[6]
C. difficile infection (A04.7)	$10,457	$10,148 - $10,771	Gamma	[6]

[a] The 95% CI column represents confidence interval for parameters whose estimates are sourced from published studies, while it represents uncertainty range for parameters (e.g. response time, cost of sequencing) whose estimates are either assumption-based or sourced from internal data (labelled as unpublished).
[b] The duration refers to the time at which effectiveness of IP intervention would reduce to zero assuming a linear decline. The effectiveness of interventions on instruments would never decline because the contaminated instrument would be discontinued from service.
[c] It was assumed that all positive cultures from wound, urine and blood represented infections, while 51% of positive respiratory cultures were assumed infections and remaining 49% were considered colonized.
[d] All costs were adjusted to 2020 using medical component of Consumer Price Index (CPI) obtained from Bureau of Labor Statistics.
[e] ICD 10 codes are provided in parenthesis

[f] Costs for treating infection on index hospitalization and subsequent HAI-related readmission, when applicable, were assumed same.

Certain variables such as effectiveness against instrument, number of IP professionals in IP team were considered fixed and hence not varied in probabilistic sensitivity analysis.
Abbreviations: EDS-HAT, Enhanced Detection System for Healthcare-Associated Transmission; ICD, International Classification of Diseases; IP, Infection Prevention; MRSA, Methicillin-resistant Staphylococcus aureus; WGS, Whole genome sequencing

Appendix Table 4. List of clusters detected by EDS-HAT

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
Vancomycin-resistant *E. faecium*	1471	E	14	640	10.6	Yes	Yes	No	Interventional Radiology Locations Routes IV Team Wound Care	Floor Unit Q	9 of the 10 initial patients all had interventional radiology exposure less than 21 days prior to their positive culture. The subsequent patient isolates all have exposures through unit-based routes, wound care visits, and IV team consult visits.
C. difficile	1	A	12	729	3.1	Yes	Yes	No	Floor Unit D Wound Care	Floor Unit Q Floor Unit J ICU F ICU I Floor Unit C	There are multiple 1-unit commonalities between 2 patient isolates, but no trending units among all patients. Wound care consult visits (8 patients) was a common transmission route.
C. difficile	8	A	11	570	6.7	Yes	Yes	No	Floor Unit D Wound Care IV Team	ICU A	There were unit commonalities on Floor Unit D and one occasion on ICU A. Wound care (5 patients) and IV Team consults (8 patients) were common.
C. difficile	1	C2	9	553	0.0	Yes	Yes	No	Wound Care IV Team Speech Therapy	Chronic Care Facility Floor Unit B	Three patients had exposure to the Chronic Care Facility prior to their positive culture dates. Wound care (5 patients), speech therapy (3 patients), and IV team consults (6 patients) were also potential routes of transmission.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
A. baumaunii	2	A	8	498	9.2	Yes	No	Yes	ICU A Providers	Floor Unit E	Patient 1 has no clear epidemiological links. Patients 2, 3, 4, and 8 all have stays on the same ICU A prior to their positive culture with Patient 2 and 3 having overlap on the unit. There was overlap for Patient 4, post-positive culture, on a general ward unit with Patient 5. Patients 4, 5, and 6 all saw one physician provider prior to their positive. Patient 6 saw this provider at the same time as another physician provider which subsequently saw Patient 7.
K. pneumoniae	258	A	7	613	17.9	Yes	Yes	No	Gastroscopy EEG Bronchoscopy	ICU G	Patients 1 and 2 had shared ICU G stays and the same gastroscope used in procedures. The next 3 patients were on ICU F housed near each other. The last 3 patients all had gastroscopy with the same gastroscopes.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
K. pneumoniae	307	D	7	223	6.4	Yes	Yes	Yes	Floor Unit O ICU G Bronchoscopy Providers	Floor Unit O	All but Patient 4 had ICU G exposure and all but Patients 2-4 had Floor Unit O exposure, some with overlap of positive patients. Patients 1 and 3 had the same bronchoscope prior to their culture date. Patients 1, 5-7 had ICU G BALs. All patients had visits by at least 1 of 4 ICU G service line providers prior to their culture date.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
P. aeruginosa	241	A	7	474	2.7	Yes	Yes	Yes	Floor Unit J Gastroscopy IV Team Provider	Floor Unit J	There are no known commonalities for patient 1. Patient 2 was on a separate ICU from Patient 3, however these ICUs share respiratory therapy staff and had the same provider after Patient 2's positive culture date. Patient 2 was on Floor Unit J after their positive culture date. Prior to their positive culture date, Patient 4 was on Floor Unit J with Patient 2 as well as Patient 6 with then Patient 4. Patients 2, 4, and 5 all saw the same physician consult provider prior to their positive date overlapping with Patients 2 and 4 in the same time period. Patients 3 and 5 were on unit Floor Unit J prior to their positive culture. The IV Team all visited Patients 3, 5, and 6 prior to their positive date. Lastly, Patient 2, after their positive culture date, underwent gastroscopy by a gastroscope which was then used on Patient 4 after reprocessing. Patients 6 and 7 both underwent gastroscopy with another gastroscope

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
											prior to their positive culture dates.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
C. difficile	1	C	6	222	1.7	Yes	Yes	No	Chronic Care Facility Wound Care IV Team Speech Therapy	ICU D	5 out of 6 patients had the Chronic Care Facility as an exposure with overlapping stays. Patient 1 also had an ICU D commonality with patient 5. Wound care (4 patients), speech therapy (4 patients), and IV team (6 patients) consults were common among patients as well.
P. aeruginosa	27	A	6	190	6.0	Yes	Yes	No	Gastroscopy Floor Unit J ICU H Bronchoscopy	Floor Unit E	Patients 1-4 all underwent gastroscopy with the same gastroscope A and on ICU H or Floor Unit J prior to testing positive. Patients 2 and 4 were roommates on Floor Unit E prior to Patient 4 testing positive. Patient 4 subsequently underwent a gastroscopy with a different scope B which was then used on Patient 5. There were no detected epidemiological links to Patient 6. Scope A was subsequently cultured and positive for the same strain.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
Vancomycin-resistant *E. faecium*	18	B	6	565	5.5	Yes	No	No	Floor Unit H ICU F Floor Unit A Floor Unit B ICU I ICU G		Five patients have unit stays on ICU F. There are multiple other unit commonalities ranging 2 to 3 patients.
Vancomycin-resistant *E. faecium*	736	B	5	321	7.4	No	No	No			
C. difficile	1	B3	4	6	0.0	Yes	Yes	No	Chronic Care Facility Wound Care		All patients were on the Chronic Care Facility prior to their positive culture dates. Wound care visits were also common among 2 of the patients.
C. difficile	10	A	4	531	5.0	Yes	Yes	No	Floor Unit D IV Team	ICU G	Patient 1 and 3 share a common bed exposure in ICU G. Patients 1 and 4 were both on Floor Unit D simultaneously many times over the cluster period (post-positive culture patient 1, pre-positive culture patient 4) including one occasion of patient 1 leaving a room and patient 4 immediately moving into that room. IV team visits were common among patients 2 and 3 within 20 days to their positive dates.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
C. difficile	110	A	4	131	0.5	Yes	Yes	No	Floor Unit D Floor Unit J ICU G Peripheral Vascular Lab IV Team	Chronic Care Facility	There are multiple unit commonalities among these patients. Visits by the IV team and the visits to the peripheral vascular lab was detected for 2 patients.
Methicillin-resistant *S. aureus*	105	A	4	120	7.0	Yes	Yes	Yes	ICU F IV Team Provider		3 of 4 patients had stays on a shared unit. Patients 3 and 4 had shared exposures to IV team consults.
Vancomycin-resistant *E. faecium*	736	A	4	572	13.2	Yes	Yes	No	ICU A Floor Unit Q IV Team	Floor Unit H	Patients 1-3 all share visits by the IV team. There are unit commonalities for Floor Unit Q, ICU A, and Floor Unit H prior to positive culture dates.
C. difficile	1	I	3	194	0.7	Yes	Yes	No	Floor Unit M Wound Care Interventional Radiology		The first 2 patients shared Floor Unit M as a potential source. The last 2 patients shared both wound care and interventional radiology suite visits as a potential transmission route.
C. difficile	54	A	3	41	4.0	Yes	Yes	Yes	ICU A IV Team	Providers	Patients 2 and 3 have concurrent stays on ICU A in adjacent rooms. Patient 1 only has stays Chronic Care Facility. Notably, Chronic Care Facility and ICU A share some providers. Additionally, patients 2 and 3 had visits from the IV team consults prior to their positive.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
E. coli	131	A	3	261	1.3	Yes	Yes	Yes	Floor Unit J ICU H Gastroscopy Bronchoscopy Providers		Patient 1 was on the ICU H/Floor Unit J service line prior to their positive culture. Patient 2 was subsequently also on ICU H/Floor Unit J service line prior to their positive culture. There was unit overlap with Patient 3. Patient 2 and 3 both underwent gastroscopy with the same scope prior to their positive culture (3, 28 days). All patients had ICU H bronchoscopy prior to their positives. Patients 2 and 3 were both seen by the same physician <30 days prior to their positive date.
E. coli	131	B	3	316	8.7	No	No	No			
K. pneumoniae	258	D	3	406	6.7	Yes	No	No		ICU A Floor Unit G	Patient 2 was on Floor Unit G with patient 3 prior to patient 3's positive culture date. All patients were on ICU A Patient 1 was positive on admit on ICU A Patient 2 moved into patient 1's room as soon as discharged from it.
Methicillin-resistant *S. aureus*	105	C	3	143	3.3	No	No	No			

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
P. aeruginosa	186	A	3	59	3.3	Yes	Yes	No	Gastroscopy Interventional Radiology	ICU G Floor Unit J ICU H	Patient 1, post-positive culture date, undergoes gastroscopy. This same gastroscope is subsequently used on patients 2 and 3 prior to their positive culture dates. All patients undergo an interventional radiology procedure prior to their positive. Patients 2 and 3 have overlapping stays on ICU H/Floor Unit J with bronchoscopies performed bedside and in the operating room with the same bronchoscopes.
P. aeruginosa	253	A	3	126	14.0	Yes	Yes	No	Gastroscopy Bronchoscopy Floor Unit J ICU H ICU G		The first 2 patients share stays on ICU H/Floor Unit J while the third patient shares a stay on ICU G with shared staff. There are multiple gastroscope and bronchoscopy procedures performed on all patients.
Vancomycin-resistant *E. faecium*	17	Z	3	157	6.7	Yes	Yes	No	Interventional Radiology ICU G Floor Unit O		Patients 2 and 3 share stays on Floor Unit O and ICU G together. Patients 1 & 2 both have interventional radiology procedures 20 days or less to their positive culture date.
Vancomycin-resistant *E. faecium*	17	O	3	274	10.0	No	No	No			

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
Vancomycin-resistant *E. faecium*	18	C	3	609	15.3	No	No	No			
Vancomycin-resistant *E. faecium*	172	.	3	278	12.7	No	Yes	No	IV Team		Patients 2 and 3 share visits by the IV team.
Vancomycin-resistant *E. faecium*	203	A	3	415	9.3	No	No	No			
Vancomycin-resistant *E. faecium*	736	F	3	144	8.7	Yes	Yes	No	ICU G Interventional Radiology Operating Room Artificial Heart	ICU H	All patients had stays on ICU G or ICU H which share equipment and staff. 2 patients had exposure to the same operating room the same staff. Artificial heart staff visits were common among these patients. Lastly, patients 1 and 3 had exposure to the interventional radiology suite less than 21 days prior to their positive culture date.
Vancomycin-resistant *E. faecium*	1471	G	3	57	6.0	Yes	No	No	Floor Unit D		Patients 2, post-positive, was in an adjacent room next to patient 3 prior to their positive culture date.
A. baumaunii	2	E	2	197	13.0	Yes	Yes	No	Floor Unit C ICU G IV Team	ICU A	Both patients were on ICU A a room apart with visits from the IV Team. Both patients also had Floor Unit C and ICU G stays prior to their positive cultures.
A. baumaunii	2	D	2	110	0.0	No	No	No			

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
C. difficile	1	L	2	122	3.0	No	Yes	Yes	Provider IV Team		Both patients had visits by the IV team and by one physician performing a nerve block prior to these patients' positive test dates.
C. difficile	1	B4	2	190	4.0	Yes	Yes	No	Floor Unit H Wound Care		Both patients have Floor Unit H and wound care as common exposures.
C. difficile	1	B2	2	218	1.0	Yes	No	No	Floor Unit A	ICU A	Both patients have ICU A and Floor Unit A overlap.
C. difficile	1	B	2	709	4.0	Yes	No	No		Floor Unit O	Both patients have overlapping stays on Floor Unit O.
C. difficile	1	E	2	403	3.0	Yes	No	No		Outside Hospital Floor A	Patient 1, post-positive culture, had a stay on an outside hospital unit floor that patient 2 was housed on pre-positive culture.
C. difficile	1	F	2	15	0.0	No	No	No			
C. difficile	2	A	2	191	3.0	No	Yes	No	IV Team		Patients have IV Team visits 4- and 14-days prior to patient positives, respectively, as a commonality.
C. difficile	8	B	2	24	1.0	Yes	Yes	No	Floor Unit O Dialysis	Floor Unit A	Both patients had exposure on 2 units prior to their positive. These patients were in adjacent rooms in one unit. Further, patient 1 was on the dialysis unit after the patient's positive culture date shortly before patient 2 was on the renal unit pre-positive culture date.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
C. difficile	13	A	2	272	1.0	No	No	No			
C. difficile	17	A	2	36	3.0	Yes	No	No		Chronic Care Facility ICU A	Both patients were housed on ICU A and the Chronic Care Facility together.
C. difficile	43	B	2	200	1.0	No	No	No			
C. difficile	55	A	2	47	0.0	Yes	No	No	Chronic Care Facility		Both patients were on the Chronic Care Facility with the second patient just 2 days after the first patient (post-positive culture) left.
C. freundii	91	B	2	337	4.0	Yes	No	No	ICU E Floor Unit L		There are 2 common units among both patients, however there is a 9-month gap in between these stays.
E. coli	131	C	2	114	3.0	No	No	No			
E. coli	131	D	2	71	15.0	No	No	No			
K. pneumoniae	45	B	2	569	12.0	No	No	No			
K. pneumoniae	258	B	2	3	9.0	Yes	Yes	No	Floor Unit O ICU G IV Team		Both patients had concurrent stays together on Floor Unit O and ICU G. Both patients had an IV team visit 9- and 4-days prior to their positive date, respectively.
K. pneumoniae	258	E	2	31	15.0	Yes	No	No	Chronic Care Facility		Both patients were on the Chronic Care Facility together leading up to their positive culture dates.
K. pneumoniae	405	A	2	2	2.0	Yes	No	No	Floor Unit H		Both patients had concurrent exposure on Floor Unit H.
K. pneumoniae	2943	A	2	36	0.0	No	No	No			

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
Methicillin-resistant *S. aureus*	5	A	2	1	2.0	No	No	No			
Methicillin-resistant *S. aureus*	5	B	2	323	11.0	No	No	No			
Methicillin-resistant *S. aureus*	5	C	2	242	12.0	No	No	No			
Methicillin-resistant *S. aureus*	5	D	2	5	11.0	No	No	No			
Methicillin-resistant *S. aureus*	5	E	2	108	11.0	No	No	No			
Methicillin-resistant *S. aureus*	8	E	2	0	0.0	Yes	No	No	Floor Unit P	Operating Room	Both patients had overlapping stays on Floor Unit P together. Both patients also underwent procedures within the same operating room. These procedures had shared staff.
Methicillin-resistant *S. aureus*	8	A	2	4	6.0	Yes	No	No	Operating Room		Both patients have an operation in 2 operating rooms that share a common preparation space.
Methicillin-resistant *S. aureus*	8	F	2	58	1.0	Yes	Yes	No	Bronchoscopy	Chronic Care Facility ICU F	Both patients have shared unit exposure together on ICU F. Patient 1, post-positive culture date, was housed on the Chronic Care Facility with patient 2. Both patients also had bronchoscopy on ICU F.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
Methicillin-resistant *S. aureus*	8	B	2	8	3.0	No	Yes	No	EEG		Both patients have stays on separate units. Both patients underwent EEG on the same day which was 2- and 10- days prior to the patients' positive culture dates. Additionally, these EEGs were performed by the same technician and the same physician.
Methicillin-resistant *S. aureus*	8	C	2	53	0.0	Yes	No	No	Floor Unit O		Patient 1 was on unit Floor Unit O 1 month prior to patient 2.
Methicillin-resistant *S. aureus*	8	D	2	163	6.0	No	No	No			
Methicillin-resistant *S. aureus*	8	G	2	186	14.0	No	No	No			
Methicillin-resistant *S. aureus*	8	H	2	53	15.0	No	No	No			
Methicillin-resistant *S. aureus*	8	I	2	310	3.0	No	No	No			
Methicillin-resistant *S. aureus*	105	B	2	91	2.0	Yes	No	No	Floor Unit G	Floor Unit C ICU I	Both patients have exposure to Floor Unit G prior to positives. Post-positive, patient 1 is on ICU I and Floor Unit C prior to patient 1 being housed on these units. Additionally, patient 1 post-positive is in ICU I bed 6 which patient 2 later moves into.
Methicillin-resistant *S. aureus*	105	D	2	333	14.0	No	No	No			

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
P. aeruginosa	116	A	2	201	3.0	Yes	No	No		Operating Room	Patient 1, post-positive, has a procedure in an operating room. Patient 2, 1-month prior to their positive culture date, undergoes a procedure in the same operating room with the same staff.
P. aeruginosa	179	E	2	13	9.0	Yes	No	No	ICU F		Both patients were concurrently on ICU F in adjacent rooms prior to their positive culture date.
P. aeruginosa	253	B	2	30	7.0	Yes	Yes	No	ICU I Interventional Radiology	Chronic Care Facility	Both patients unit stays on ICU I. Post-positive, patient 1 moves to the Chronic Care Facility with patient 2, pre-positive. There was shared exposure to interventional radiology procedures for both patients prior to their positive culture dates.
P. aeruginosa	260	A	2	109	7.0	Yes	No	No	ICU D	Floor Unit I	Both patients were housed on ICU D and concurrently Floor Unit I.
P. aeruginosa	274	A	2	55	5.0	Yes	No	No		ICU D Operating Room	Both patients were admitted to the hospital for urology and trauma issues. Both patients underwent multiple irrigation and debridement procedures in the operating room utilizing the same staff. The patients had shared staff through ICU D.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
P. aeruginosa	1978	A	2	134	1.0	Yes	No	No	Chronic Care Facility		Both patients have concurrent stays on the Chronic Care Facility.
P. mirabilis	19	A	2	23	0.0	Yes	No	No		Floor Unit Q	Patient 1, post-positive, is roommates for 9-days with patient 2 on Floor Unit Q who is then positive within 2-weeks.
S. maltophilia	162	A	2	46	5.0	No	No	No			
S. maltophilia	172	A	2	19	12.0	Yes	No	No	ICU C		Both patients are on ICU C prior to their positive culture date in adjacent rooms.
S. marcescens	4	E	2	36	2.0	Yes	No	No		Floor Unit O	Patient 1, post-positive, is roommates for 4-days on Floor Unit O with patient 2 28-days prior to patient 2's positive culture date.
S. marcescens	6	E	2	35	11.0	No	No	No			
S. marcescens	13	B	2	61	4.0	Yes	No	No		ICU H	Patient 1 is has a stay on ICU H prior to positive. Patient 1 leaves this ICU H bed, and Patient 2 moves into this bed 2-days later.
S. marcescens	17	B	2	13	2.0	Yes	Yes	No		ICU G Bronchoscopy	Patients are housed in adjacent rooms on ICU G.
S. marcescens	19	A	2	144	14.0	Yes	No	No		Floor Unit I	Post-positive, patient 1 on unit Floor Unit I. 2-months later, patient 2 is on this unit 27 days prior to their positive culture date.

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
S. marcescens	20	A	2	65	3.0	Yes	Yes	No		Floor Unit J Gastroscopy	Both patients had gastroscopy with same gastroscope prior to their positive date. Patient 1 is on unit Floor Unit J and 6-days later patient 2 is on Floor Unit J.
S. marcescens	21	A	2	19	1.0	No	No	No			
Vancomycin-resistant *E. faecium*	17	U	2	469	9.0	Yes	No	No	ICU A		Both patients are on ICU A together in adjacent rooms.
Vancomycin-resistant *E. faecium*	17	Q	2	6	15.0	No	No	No			
Vancomycin-resistant *E. faecium*	17	S	2	46	12.0	No	No	No			
Vancomycin-resistant *E. faecium*	17	T	2	24	11.0	No	No	No			
Vancomycin-resistant *E. faecium*	18	H	2	178	2.0	No	Yes	No	US-guided drain		Both patients were on separate units. Both patients underwent ultrasound-guided drainage procedures of fluid collections for these isolates sequenced. Patient 1 had 1 colony of VRE while patient 2 had “rare” VRE in the culture result.
Vancomycin-resistant *E. faecium*	18	D	2	47	6.0	No	No	No			
Vancomycin-resistant *E. faecium*	412	D	2	24	4.0	No	No	No			

Organism	Sequence Type	Cluster	Size	Days between first and last case	Mean Pairwise SNPs	Unit-associated Route	Procedure-associated Route	Provider-associated Route	Machine Learning Routes	Manual Review Routes	Comment
Vancomycin-resistant *E. faecium*	736	J	2	103	6.0	Yes	No	No	Floor Unit D		Both patients were on Floor Unit D leading up to their positive culture date in adjacent rooms.
Vancomycin-resistant *E. faecium*	736	N	2	71	3.0	No	Yes	No	IV Team		Both patients were seen by the same IV team provider 1-day apart, both prior to their positive culture dates.
Vancomycin-resistant *E. faecium*	736	K	2	58	12.0	No	No	No			
Vancomycin-resistant *E. faecium*	1471	O	2	30	3.0	Yes	No	No	Floor Unit D ICU I Operating Room		Both patients have unit stays on Floor Unit D and ICU I. Additionally, both patients had an operating room procedure the same operating room.
Vancomycin-resistant *E. faecium*	1717	.	2	250	12.0	No	No	No			
Vancomycin-resistant *E. faecium*	1723	.	2	3	1.0	Yes	No	No		Chronic Care Facility	Both patients are on the Chronic Care Facility together for over 30-days prior to their positive culture dates.

Appendix Table 5. Clinical and economic modeling results

	Traditional IP practice	EDS-HAT	Change
Lower bound			
Number of transmissions	289.0	264.2	24.8 averted
Number of deaths	14.7	13.1	1.6 saved
Number of readmissions	36.8	33.7	3.1 averted
Total costs	$4,074,022	$3,881,614	($192,408)
IP program	$152,420	$152,420	$0
WGS costs	$17,764	$179,174	$161,410
Treating infections	$3,903,838	$3,550,020	($353,818)
Incremental cost per transmission averted for EDS-HAT = $7,745 saved for each transmission averted i.e. less costly and more effective			
Upper bound			
Number of transmissions	289.0	225.7	63.3 averted
Number of deaths	14.7	11.4	3.3 saved
Number of readmissions	36.8	28.8	8.0 averted
Total costs	$4,074,022	$3,381,490	($692,532)
IP program	$152,420	$152,420	$0
WGS costs	$17,764	$179,174	$161,410
Treating infections	$3,903,838	$3,049,896	($853,942)

	Traditional IP practice	EDS-HAT	Change
Incremental cost per transmission averted for EDS-HAT = $10,939 saved for each transmission averted i.e. less costly and more effective			

Appendix B.1 References

1. Anderson, D. J., et al., *Enhanced terminal room disinfection and acquisition and infection caused by multidrug-resistant organisms and Clostridium difficile (the Benefits of Enhanced Terminal Room Disinfection study): a cluster-randomised, multicentre, crossover study.* Lancet, 2017. **389**(10071): p. 805-814.

2. Martin, R. M., et al., *Molecular Epidemiology of Colonizing and Infecting Isolates of Klebsiella pneumoniae.* mSphere, 2016. **1**(5).

3. Klevens, R. M., et al., *Estimating health care-associated infections and deaths in U.S. hospitals, 2002.* Public Health Rep, 2007. **122**(2): p. 160-6.

4. CDC, *Nearly half a million Americans suffered from Clostridium difficile infections in a single year*. 2015, U.S. Department of Health and Human Services.

5. Landers, T., et al., *APIC MegaSurvey: Methodology and overview.* Am J Infect Control, 2017. **45**(6): p. 584-588.

6. AHRQ, *HCUPnet, Healthcare Cost and Utilization Project*. 2016: Rockville, MD.

Bibliography

1. HAI Data | CDC. Published October 20, 2021. Accessed March 14, 2022. https://www.cdc.gov/hai/data/index.html

2. Barrasa-Villar JI, Aibar-Remón C, Prieto-Andrés P, Mareca-Doñate R, Moliner-Lahoz J. Impact on Morbidity, Mortality, and Length of Stay of Hospital-Acquired Infections by Resistant Microorganisms. *Clin Infect Dis*. 2017;65(4):644-652. doi:10.1093/cid/cix411

3. About | NHSN | CDC. Published January 25, 2021. Accessed March 15, 2022. https://www.cdc.gov/nhsn/about-nhsn/index.html

4. CMS Requirements | NHSN | CDC. Published June 9, 2021. Accessed March 15, 2022. https://www.cdc.gov/nhsn/cms/index.html

5. Weiner-Lastinger LM, Pattabiraman V, Konnor RY, et al. The impact of coronavirus disease 2019 (COVID-19) on healthcare-associated infections in 2020: A summary of data reported to the National Healthcare Safety Network. *Infect Control Hosp Epidemiol*. 2022;43(1):12-25. doi:10.1017/ice.2021.362

6. Healthcare-Associated Infections | Healthy People 2020. Accessed March 15, 2022. https://www.healthypeople.gov/2020/topics-objectives/topic/healthcare-associated-infections

7. Preventing Healthcare-associated Infections | HAI | CDC. Published April 19, 2019. Accessed March 15, 2022. https://www.cdc.gov/hai/prevent/prevention.html

8. BSI | Guidelines Library | Infection Control | CDC. Published January 3, 2020. Accessed March 15, 2022. https://www.cdc.gov/infectioncontrol/guidelines/bsi/index.html

9. Outbreak Investigations in Healthcare Settings | HAI | CDC. Published August 9, 2021. Accessed March 15, 2022. https://www.cdc.gov/hai/outbreaks/index.html

10. Principles of Epidemiology | Lesson 1 - Section 11. Published December 20, 2021. Accessed March 15, 2022. https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section11.html

11. Webinar Recording: Local Health Department Access to the National Healthcare Safety Network - NACCHO. Accessed March 15, 2022. https://www.naccho.org/blog/articles/webinar-recording-local-health-department-access-to-the-national-healthcare-safety-network

12. Houlihan CF, Frampton D, Ferns RB, et al. Use of Whole-Genome Sequencing in the Investigation of a Nosocomial Influenza Virus Outbreak. *J Infect Dis*. 2018;218(9):1485-1489. doi:10.1093/infdis/jiy335

13. Gordon LG, Elliott TM, Forde B, et al. Budget impact analysis of routinely using whole-genomic sequencing of six multidrug-resistant bacterial pathogens in Queensland, Australia. *BMJ Open*. 2021;11(2):e041968. doi:10.1136/bmjopen-2020-041968

14. Sherry NL, Lee RS, Gorrie CL, et al. Pilot study of a combined genomic and epidemiologic surveillance program for hospital-acquired multidrug-resistant pathogens across multiple hospital networks in Australia. *Infect Control Hosp Epidemiol*. 2021;42(5):573-581. doi:10.1017/ice.2020.1253

15. Ward DV, Hoss AG, Kolde R, et al. Integration of genomic and clinical data augments surveillance of healthcare-acquired infections. *Infect Control Hosp Epidemiol*. 2019;40(6):649-655. doi:10.1017/ice.2019.75

16. Sundermann AJ, Babiker A, Marsh JW, et al. Outbreak of Vancomycin-resistant *Enterococcus faecium* in Interventional Radiology: Detection Through Whole-genome Sequencing-based Surveillance. *Clin Infect Dis*. 2020;70(11):2336-2343. doi:10.1093/cid/ciz666

17. Sundermann AJ, Chen J, Miller JK, et al. Outbreak of *Pseudomonas aeruginosa* Infections from a Contaminated Gastroscope Detected by Whole Genome Sequencing Surveillance. *Clin Infect Dis*. 2021;73(3):e638-e642. doi:10.1093/cid/ciaa1887

18. Sundermann AJ, Miller JK, Marsh JW, et al. Automated data mining of the electronic health record for investigation of healthcare-associated outbreaks. *Infect Control Hosp Epidemiol*. 2019;40(3):314-319. doi:10.1017/ice.2018.343

19. Kumar P, Sundermann AJ, Martin EM, et al. Method for Economic Evaluation of Bacterial Whole Genome Sequencing Surveillance Compared to Standard of Care in Detecting Hospital Outbreaks. *Clin Infect Dis*. 2021;73(1):e9-e18. doi:10.1093/cid/ciaa512

20. Miller JK, Chen J, Sundermann A, et al. Statistical outbreak detection by joining medical records and pathogen similarity. *J Biomed Inform*. 2019;91:103126. doi:10.1016/j.jbi.2019.103126

21. Tosas Auguet O, Stabler RA, Betley J, et al. Frequent Undetected Ward-Based Methicillin-Resistant *Staphylococcus aureus* Transmission Linked to Patient Sharing Between Hospitals. *Clin Infect Dis*. 2018;66(6):840-848. doi:10.1093/cid/cix901

22. Berbel Caban A, Pak TR, Obla A, et al. PathoSPOT genomic epidemiology reveals under-the-radar nosocomial outbreaks. *Genome Med*. 2020;12(1):96. doi:10.1186/s13073-020-00798-3

23. Coll F, Harrison EM, Toleman MS, et al. Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community. *Sci Transl Med*. 2017;9(413):eaak9745. doi:10.1126/scitranslmed.aak9745

24. Cremers AJH, Coolen JPM, Bleeker-Rovers CP, et al. Surveillance-embedded genomic outbreak resolution of methicillin-susceptible *Staphylococcus aureus* in a neonatal intensive care unit. *Sci Rep*. 2020;10(1):2619. doi:10.1038/s41598-020-59015-1

25. Donskey CJ, Sunkesula VCK, Stone ND, et al. Transmission of *Clostridium difficile* from asymptomatically colonized or infected long-term care facility residents. *Infect Control Hosp Epidemiol*. 2018;39(8):909-916. doi:10.1017/ice.2018.106

26. Eigenbrod T, Reuter S, Gross A, et al. Molecular characterization of carbapenem-resistant *Acinetobacter baumannii* using WGS revealed missed transmission events in Germany from 2012-15. *J Antimicrob Chemother*. 2019;74(12):3473-3480. doi:10.1093/jac/dkz360

27. Elbadawi LI, Borlaug G, Gundlach KM, et al. Carbapenem-Resistant *Enterobacteriaceae* Transmission in Health Care Facilities - Wisconsin, February-May 2015. *MMWR Morb Mortal Wkly Rep*. 2016;65(34):906-909. doi:10.15585/mmwr.mm6534a5

28. 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(13):1195-1205. doi:10.1056/NEJMoa1216064

29. Eyre DW, Fawley WN, Rajgopal A, et al. Comparison of Control of *Clostridium difficile* Infection in Six English Hospitals Using Whole-Genome Sequencing. *Clin Infect Dis*. 2017;65(3):433-441. doi:10.1093/cid/cix338

30. Eyre DW, Shaw R, Adams H, et al. WGS to determine the extent of *Clostridioides difficile* transmission in a high incidence setting in North Wales in 2015. *J Antimicrob Chemother*. 2019;74(4):1092-1100. doi:10.1093/jac/dky523

31. García-Fernández S, Frentrup M, Steglich M, et al. Whole-genome sequencing reveals nosocomial *Clostridioides difficile* transmission and a previously unsuspected epidemic scenario. *Sci Rep*. 2019;9(1):6959. doi:10.1038/s41598-019-43464-4

32. Gona F, Comandatore F, Battaglia S, et al. Comparison of core-genome MLST, coreSNP and PFGE methods for *Klebsiella pneumoniae* cluster analysis. *Microb Genom*. 2020;6(4). doi:10.1099/mgen.0.000347

33. Gorrie CL, Mirceta M, Wick RR, et al. Gastrointestinal Carriage Is a Major Reservoir of *Klebsiella pneumoniae* Infection in Intensive Care Patients. *Clin Infect Dis*. 2017;65(2):208-215. doi:10.1093/cid/cix270

34. Hall MD, Holden MT, Srisomang P, et al. Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. *Elife*. 2019;8:e46402. doi:10.7554/eLife.46402

35. Hammerum AM, Lauridsen CAS, Blem SL, et al. Investigation of possible clonal transmission of carbapenemase-producing *Klebsiella pneumoniae* complex member isolates in Denmark using core genome MLST and National Patient Registry Data. *Int J Antimicrob Agents*. 2020;55(5):105931. doi:10.1016/j.ijantimicag.2020.105931

36. Harada S, Aoki K, Yamamoto S, et al. Clinical and Molecular Characteristics of *Klebsiella pneumoniae* Isolates Causing Bloodstream Infections in Japan: Occurrence of Hypervirulent Infections in Health Care. *J Clin Microbiol*. 2019;57(11):e01206-19. doi:10.1128/JCM.01206-19

37. Houldcroft CJ, Roy S, Morfopoulou S, et al. Use of Whole-Genome Sequencing of Adenovirus in Immunocompromised Pediatric Patients to Identify Nosocomial Transmission and Mixed-Genotype Infection. *J Infect Dis*. 2018;218(8):1261-1271. doi:10.1093/infdis/jiy323

38. Jakharia KK, Ilaiwy G, Moose SS, et al. Use of whole-genome sequencing to guide a *Clostridioides difficile* diagnostic stewardship program. *Infect Control Hosp Epidemiol*. 2019;40(7):804-806. doi:10.1017/ice.2019.124

39. Kossow A, Kampmeier S, Schaumburg F, Knaack D, Moellers M, Mellmann A. Whole genome sequencing reveals a prolonged and spatially spread nosocomial outbreak of Panton-Valentine leucocidin-positive meticillin-resistant *Staphylococcus aureus* (USA300). *J Hosp Infect*. 2019;101(3):327-332. doi:10.1016/j.jhin.2018.09.007

40. Kwong JC, Lane CR, Romanes F, et al. Translating genomics into practice for real-time surveillance and response to carbapenemase-producing *Enterobacteriaceae*: evidence from a complex multi-institutional KPC outbreak. *PeerJ*. 2018;6:e4210. doi:10.7717/peerj.4210

41. Leong KWC, Cooley LA, Anderson TL, et al. Emergence of Vancomycin-Resistant *Enterococcus faecium* at an Australian Hospital: A Whole Genome Sequencing Analysis. *Sci Rep*. 2018;8(1):6274. doi:10.1038/s41598-018-24614-6

42. Long SW, Beres SB, Olsen RJ, Musser JM. Absence of patient-to-patient intrahospital transmission of *Staphylococcus aureus* as determined by whole-genome sequencing. *mBio*. 2014;5(5):e01692-01614. doi:10.1128/mBio.01692-14

43. Marmor A, Daveson K, Harley D, Coatsworth N, Kennedy K. Two carbapenemase-producing *Enterobacteriaceae* outbreaks detected retrospectively by whole-genome sequencing at an Australian tertiary hospital. *Infect Dis Health*. 2020;25(1):30-33. doi:10.1016/j.idh.2019.08.005

44. Martin JSH, Eyre DW, Fawley WN, et al. Patient and Strain Characteristics Associated With *Clostridium difficile* Transmission and Adverse Outcomes. *Clin Infect Dis*. 2018;67(9):1379-1387. doi:10.1093/cid/ciy302

45. Mathur P, Khurana S, de Man TJB, et al. Multiple importations and transmission of colistin-resistant *Klebsiella pneumoniae* in a hospital in northern India. *Infect Control Hosp Epidemiol*. 2019;40(12):1387-1393. doi:10.1017/ice.2019.252

46. Meredith LW, Hamilton WL, Warne B, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis*. 2020;20(11):1263-1271. doi:10.1016/S1473-3099(20)30562-4

47. Miles-Jay A, Weissman SJ, Adler AL, Baseman JG, Zerr DM. Whole Genome Sequencing Detects Minimal Clustering Among *Escherichia coli* Sequence Type 131-H30 Isolates Collected From United States Children's Hospitals. *J Pediatric Infect Dis Soc*. 2021;10(2):183-187. doi:10.1093/jpids/piaa023

48. Neumann B, Bender JK, Maier BF, et al. Comprehensive integrated NGS-based surveillance and contact-network modeling unravels transmission dynamics of vancomycin-resistant *enterococci* in a high-risk population within a tertiary care hospital. *PLoS One*. 2020;15(6):e0235160. doi:10.1371/journal.pone.0235160

49. Raven KE, Gouliouris T, Brodrick H, et al. Complex Routes of Nosocomial Vancomycin-Resistant *Enterococcus faecium* Transmission Revealed by Genome Sequencing. *Clin Infect Dis*. 2017;64(7):886-893. doi:10.1093/cid/ciw872

50. Roach DJ, Burton JN, Lee C, et al. 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*. 2015;11(7):e1005413. doi:10.1371/journal.pgen.1005413

51. Rose R, Nolan DJ, Moot S, et al. Molecular surveillance of methicillin-resistant *Staphylococcus aureus* genomes in hospital unexpectedly reveals discordance between temporal and genetic clustering. *Am J Infect Control*. 2021;49(1):59-64. doi:10.1016/j.ajic.2020.06.180

52. Roy S, Hartley J, Dunn H, Williams R, Williams CA, Breuer J. Whole-genome Sequencing Provides Data for Stratifying Infection Prevention and Control Management of Nosocomial Influenza A. *Clin Infect Dis*. 2019;69(10):1649-1656. doi:10.1093/cid/ciz020

53. Sherry NL, Lane CR, Kwong JC, et al. Genomics for Molecular Epidemiology and Detecting Transmission of Carbapenemase-Producing *Enterobacterales* in Victoria, Australia, 2012 to 2016. *J Clin Microbiol*. 2019;57(9):e00573-19. doi:10.1128/JCM.00573-19

54. Stenmark B, Hellmark B, Söderquist B. Genomic analysis of *Staphylococcus capitis* isolated from blood cultures in neonates at a neonatal intensive care unit in Sweden. *Eur J Clin Microbiol Infect Dis*. 2019;38(11):2069-2075. doi:10.1007/s10096-019-03647-3

55. Sullivan MJ, Altman DR, Chacko KI, et al. A Complete Genome Screening Program of Clinical Methicillin-Resistant *Staphylococcus aureus* Isolates Identifies the Origin and Progression of a Neonatal Intensive Care Unit Outbreak. *J Clin Microbiol*. 2019;57(12):e01261-19. doi:10.1128/JCM.01261-19

56. Tsujiwaki A, Hisata K, Tohyama Y, et al. Epidemiology of methicillin-resistant *Staphylococcus aureus* in a Japanese neonatal intensive care unit. *Pediatr Int*. 2020;62(8):911-919. doi:10.1111/ped.14241

57. van Beek J, Räisänen K, Broas M, et al. Tracing local and regional clusters of carbapenemase-producing *Klebsiella pneumoniae* ST512 with whole genome sequencing,

Finland, 2013 to 2018. *Euro Surveill*. 2019;24(38). doi:10.2807/1560-7917.ES.2019.24.38.1800522

58. Wang X, Zhou H, Chen D, et al. Whole-Genome Sequencing Reveals a Prolonged and Persistent Intrahospital Transmission of *Corynebacterium striatum*, an Emerging Multidrug-Resistant Pathogen. *J Clin Microbiol*. 2019;57(9):e00683-19. doi:10.1128/JCM.00683-19

59. Wendel AF, Malecki M, Otchwemah R, Tellez-Castillo CJ, Sakka SG, Mattner F. One-year molecular surveillance of carbapenem-susceptible *A. baumannii* on a German intensive care unit: diversity or clonality. *Antimicrob Resist Infect Control*. 2018;7:145. doi:10.1186/s13756-018-0436-8

60. Julia L, Vilankar K, Kang H, Brown DE, Mathers A, Barnes LE. Environmental Reservoirs of Nosocomial Infection: Imputation Methods for Linking Clinical and Environmental Microbiological Data to Understand Infection Transmission. *AMIA Annu Symp Proc*. 2017;2017:1120-1129.

61. Stachel A, Pinto G, Stelling J, et al. Implementation and evaluation of an automated surveillance system to detect hospital outbreak. *Am J Infect Control*. 2017;45(12):1372-1377. doi:10.1016/j.ajic.2017.06.031

62. Parcell BJ, Gillespie SH, Pettigrew KA, Holden MTG. Clinical perspectives in integrating whole-genome sequencing into the investigation of healthcare and public health outbreaks - hype or help? *J Hosp Infect*. 2021;109:1-9. doi:10.1016/j.jhin.2020.11.001

63. Magill SS, Edwards JR, Bamberg W, et al. Multistate point-prevalence survey of health care-associated infections. *N Engl J Med*. 2014;370(13):1198-1208. doi:10.1056/NEJMoa1306801

64. Scott R. The Direct Medical Costs of Healthcare-Associated Infections in U.S. Hospitals and the Benefits of Prevention. *Centers for Disease Control and Prevention website*. Published online 2009:16.

65. 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(12):e0144310. doi:10.1371/journal.pone.0144310

66. Sood G, Perl TM. Outbreaks in Health Care Settings. *Infect Dis Clin North Am*. 2016;30(3):661-687. doi:10.1016/j.idc.2016.04.003

67. Vonberg RP, Weitzel-Kage D, Behnke M, Gastmeier P. Worldwide Outbreak Database: the largest collection of nosocomial outbreaks. *Infection*. 2011;39(1):29-34. doi:10.1007/s15010-010-0064-6

68. Peacock SJ, Parkhill J, Brown NM. Changing the paradigm for hospital outbreak detection by leading with genomic surveillance of nosocomial pathogens. *Microbiology,*. 2018;164(10):1213-1219. doi:10.1099/mic.0.000700

69. Heinrichs A, Argudín MA, De Mendonça R, et al. An Outpatient Clinic as a Potential Site of Transmission for an Outbreak of New Delhi Metallo-β-Lactamase-producing *Klebsiella pneumoniae* Sequence Type 716: A Study Using Whole-genome Sequencing. *Clin Infect Dis*. 2019;68(6):993-1000. doi:10.1093/cid/ciy581

70. Domman D, Chowdhury F, Khan AI, et al. Defining endemic cholera at three levels of spatiotemporal resolution within Bangladesh. *Nat Genet*. 2018;50(7):951-955. doi:10.1038/s41588-018-0150-8

71. Pak TR, Kasarskis A. How next-generation sequencing and multiscale data analysis will transform infectious disease management. *Clin Infect Dis*. 2015;61(11):1695-1702. doi:10.1093/cid/civ670

72. Yount RJ, Vries JK, Councill CD. The medical archival system: An information retrieval system based on distributed parallel processing. *Inf Process Manag*. Published online 1991. doi:10.1016/0306-4573(91)90091-Y

73. Sherry NL, Lee RS, Gorrie CL, et al. Pilot study of a combined genomic and epidemiologic surveillance program for hospital-acquired multidrug-resistant pathogens across multiple hospital networks in Australia. *Infection Control & Hospital Epidemiology*. Published online undefined/ed:1-9. doi:10.1017/ice.2020.1253

74. Ward DV, Hoss AG, Kolde R, et al. Integration of genomic and clinical data augments surveillance of healthcare-acquired infections. *Infect Control Hosp Epidemiol*. 2019;40(6):649-655. doi:10.1017/ice.2019.75

75. Sundermann AJ, Miller JK, Marsh JW, et al. Automated data mining of the electronic health record for investigation of healthcare-associated outbreaks. *Infect Control Hosp Epidemiol*. 2019;40(3):314-319. doi:10.1017/ice.2018.343

76. Sundermann AJ, Chen J, Miller JK, et al. Outbreak of *Pseudomonas aeruginosa* Infections from a Contaminated Gastroscope Detected by Whole Genome Sequencing Surveillance. *Clin Infect Dis*. Published online December 25, 2020. doi:10.1093/cid/ciaa1887

77. Sundermann AJ, Babiker A, Marsh JW, et al. Outbreak of Vancomycin-resistant *Enterococcus faecium* in Interventional Radiology: Detection Through Whole-genome Sequencing-based Surveillance. *Clin Infect Dis*. 2020;70(11):2336-2343. doi:10.1093/cid/ciz666

78. Miller JK, Chen J, Sundermann A, et al. Statistical outbreak detection by joining medical records and pathogen similarity. *J Biomed Inform*. 2019;91:103126. doi:10.1016/j.jbi.2019.103126

79. Kumar P, Sundermann AJ, Martin EM, et al. Method for economic evaluation of bacterial whole genome sequencing surveillance compared to standard of care in detecting hospital outbreaks. *Clin Infect Dis*. Published online May 5, 2020. doi:10.1093/cid/ciaa512

80. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021

81. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-2069. doi:10.1093/bioinformatics/btu153

82. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. 2010;11:595. doi:10.1186/1471-2105-11-595

83. Berbel Caban A, Pak TR, Obla A, et al. PathoSPOT genomic epidemiology reveals under-the-radar nosocomial outbreaks. *Genome Med*. 2020;12(1):96. doi:10.1186/s13073-020-00798-3

84. Jakharia KK, Ilaiwy G, Moose SS, et al. Use of whole-genome sequencing to guide a *Clostridioides difficile* diagnostic stewardship program. *Infect Control Hosp Epidemiol*. 2019;40(7):804-806. doi:10.1017/ice.2019.124

85. Gona F, Comandatore F, Battaglia S, et al. Comparison of core-genome MLST, coreSNP and PFGE methods for *Klebsiella pneumoniae* cluster analysis. *Microb Genom*. 2020;6(4). doi:10.1099/mgen.0.000347

86. Rose R, Nolan DJ, Moot S, et al. Molecular surveillance of methicillin-resistant *Staphylococcus aureus* genomes in hospital unexpectedly reveals discordance between temporal and genetic clustering. *Am J Infect Control*. 2021;49(1):59-64. doi:10.1016/j.ajic.2020.06.180

87. Marmor A, Daveson K, Harley D, Coatsworth N, Kennedy K. Two carbapenemase-producing *Enterobacteriaceae* outbreaks detected retrospectively by whole-genome sequencing at an Australian tertiary hospital. *Infect Dis Health*. 2020;25(1):30-33. doi:10.1016/j.idh.2019.08.005

88. Sherry NL, Lane CR, Kwong JC, et al. Genomics for Molecular Epidemiology and Detecting Transmission of Carbapenemase-Producing *Enterobacterales* in Victoria, Australia, 2012 to 2016. *J Clin Microbiol*. 2019;57(9):e00573-19. doi:10.1128/JCM.00573-19

89. Kwong JC, Lane CR, Romanes F, et al. Translating genomics into practice for real-time surveillance and response to carbapenemase-producing *Enterobacteriaceae*: evidence from a complex multi-institutional KPC outbreak. *PeerJ*. 2018;6:e4210. doi:10.7717/peerj.4210

90. Raven KE, Gouliouris T, Brodrick H, et al. Complex Routes of Nosocomial Vancomycin-Resistant *Enterococcus faecium* Transmission Revealed by Genome Sequencing. *Clin Infect Dis*. 2017;64(7):886-893. doi:10.1093/cid/ciw872

91. Medical care in US city average, all urban consumers, not seasonally adjusted. Bureau of Labor Statistics.

92. Gordon LG, Elliott TM, Forde B, et al. Budget impact analysis of routinely using whole-genomic sequencing of six multidrug-resistant bacterial pathogens in Queensland, Australia. *BMJ Open*. 2021;11(2):e041968. doi:10.1136/bmjopen-2020-041968

93. Genomic surveillance, characterization and intervention of a polymicrobial multidrug-resistant outbreak in critical care - PubMed. Accessed February 20, 2021. https://pubmed.ncbi.nlm.nih.gov/33599607/

94. Bartels MD, Larner-Svensson H, Meiniche H, et al. Monitoring meticillin resistant *Staphylococcus aureus* and its spread in Copenhagen, Denmark, 2013, through routine whole genome sequencing. *Euro Surveill*. 2015;20(17). doi:10.2807/1560-7917.es2015.20.17.21112

95. Mellmann A, Bletz S, Böking T, et al. Real-Time Genome Sequencing of Resistant Bacteria Provides Precision Infection Control in an Institutional Setting. *J Clin Microbiol*. 2016;54(12):2874-2881. doi:10.1128/JCM.00790-16

96. Price JR, Cole K, Bexley A, et al. Transmission of *Staphylococcus aureus* between health-care workers, the environment, and patients in an intensive care unit: a longitudinal cohort study based on whole-genome sequencing. *Lancet Infect Dis*. 2017;17(2):207-214. doi:10.1016/S1473-3099(16)30413-3

97. Marsh JW, Mustapha MM, Griffith MP, et al. Evolution of Outbreak-Causing Carbapenem-Resistant *Klebsiella pneumoniae* ST258 at a Tertiary Care Hospital over 8 Years. *mBio*. 2019;10(5). doi:10.1128/mBio.01945-19

98. Galdys AL, Marsh JW, Delgado E, et al. Bronchoscope-associated clusters of multidrug-resistant *Pseudomonas aeruginosa* and carbapenem-resistant *Klebsiella pneumoniae*. *Infect Control Hosp Epidemiol*. 2019;40(1):40-46. doi:10.1017/ice.2018.263