Title

Regression-based modeling of pairwise genomic linkage data identifies risk factors for healthcare-associated infection transmission: Application to carbapenem-resistant *Klebsiella pneumoniae* transmission in a long-term care facility

Authors

Hannah Steinberg, MPH[1,2]; Timileyin Adediran, PhD, MPH, CIC[3]; Mary K. Hayden, MD[4]; Evan Snitkin, PhD[3,5*]; Jon Zelner, PhD[1,2*]

*Co-Senior Authors

Affiliations

[1]University of Michigan School of Public Health, Department of Epidemiology
[2]University of Michigan School of Public Health, Center for Social Epidemiology and Population Health
[3]University of Michigan, Department of Microbiology and Immunology
[4]Rush University Medical Center, Division of Infectious Diseases
[5]University of Michigan, Department of Medicine, Division of Infectious Diseases

Corresponding Author

Hannah Steinberg
hsteinb@umich.edu

Abstract

Background: Pathogen whole genome sequencing (WGS) has significant potential for improving healthcare-associated infection (HAI) outcomes. However, methods for integrating WGS with epidemiologic data to quantify risks for pathogen spread remain underdeveloped.

Methods: To identify analytic strategies for conducting WGS-based HAI surveillance in high-burden settings, we modeled patient- and facility-level transmission risks of carbapenem-resistant *Klebsiella pneumoniae* (CRKP) in a long-term acute care hospital (LTACH). Using rectal surveillance data collected over one year, we fit three pairwise regression models with three different metrics of genomic relatedness for pairs of case isolates, a proxy for transmission linkage: 1) single-nucleotide variant genomic distance, 2) closest genomic donor, 3) common genomic cluster. To assess the performance of these approaches under real-world conditions defined by passive surveillance, we conducted a sensitivity study including only cases detected by admission surveillance or clinical symptoms.

Results: Genomic relatedness between pairs of isolates was associated with room sharing in two of the three models and overlapping stays on a high-acuity unit in all models, echoing previous findings from LTACH settings. In our sensitivity analysis, qualitative findings were robust to the exclusion of cases that would not have been identified with a passive surveillance strategy, however uncertainty in all estimates also increased markedly.

Conclusions: Taken together, our results demonstrate that pairwise regression models combining relevant genomic and epidemiologic data are useful tools for identifying HAI transmission risks.

Key Messages

- Whole genome sequencing of healthcare associated infections (HAI) is becoming more common and new methods are necessary to integrate these data with epidemiologic risk factors to quantify transmission drivers.
- We demonstrate how pairwise regression models, in which the outcome of a regression model represents genomic similarity between a pair of isolates, can identify known transmission risk factors of carbapenem-resistant *Klebsiella pneumoniae* in a long-term acute care facility.
- Pairwise regression models could be used with rich epidemiologic data in other settings to identify risk factors of endemic HAI transmission.

Key Words

Carbapenem-resistant *Kleabsiella pneumoniae*, microbial genomics, epidemiology, transmission models, healthcare-associated infections, transmission risk factors

Acknowledgements

We thank Josh Warren for his consultation on genomic pairwise regression modeling for infectious disease transmission studies. This research was supported by the National Institutes of Health [R01 AI148259]. H.S., E.S, and J.Z. conceptualized the study and developed methodology. H.S. conducted analyses, created visualizations, and wrote the original draft. M.H. provided data and guidance in interpreting results. T.A. consulted throughout the project. All authors reviewed and edited this manuscript. Authors have no competing interests to report.

Introduction

Despite intensive research and scrutiny, healthcare-associated infections (HAIs) remain among the most frequent adverse events occurring in health facilities throughout the United States and the world. Improvements in broadly-effective infection prevention interventions such as hand hygiene and environmental cleaning, and targeted interventions such as pathogen decolonization, have been attributed with recent reductions in HAIs.[1] Still, it is estimated that on any given day 1 in 31 hospital patients in the US has at least one HAI.[1] HAIs are among the top 10 causes of death in the US and are associated with billions of dollars in excess healthcare costs.[2] Antibiotic resistance is common in healthcare pathogens, and can make these infections harder to treat.

Colonization typically precedes infection, and many HAI prevention strategies work by interrupting transmission of colonizing pathogens. However, a better understanding of drivers of transmission and of which patients are more likely to transmit or acquire colonization could help in developing more effective interventions to reduce HAIs.

With the increased availability and falling cost of whole genome sequencing (WGS), there has been increased interest in the use of WGS to understand transmission pathways in healthcare settings.[3] However, many studies of transmission in healthcare settings are descriptive in nature, e.g. identifying shared exposures among individuals with genomic linkage, but not quantitatively evaluating whether exposures are shared more than would be expected by chance. Thus, understanding risk factors for transmission and identifying putative targets for improved infection prevention will require more rigorous methods that integrate genomic and epidemiologic data to quantitatively identify transmission risk factors. While this has been done to some degree in outbreak settings,[4,5] there is still a need for methods applicable in high-prevalence endemic settings, where the constant importation of resistant organisms makes delineating transmission links challenging, even with genomic data.

In addition to the lack of standard frameworks for integrating genomic with epidemiologic data, additional barriers to current methods (e.g. SNV-based regression models,[6] machine learning algorithms,[7] and probabilistic transmission models[8]) include the requirement of single nucleotide variant (SNV) cutoffs to infer transmission,[6,7] needing data on uninfected controls,[7] and models that are complex[7] and/or require many assumptions about the transmission system hindering their generalizability.[8] Additionally, models that do not account for the disproportionate effect of super-spreaders on model outcomes may overestimate confidence in risk-factor estimates.[9] Recent work has shown that pairwise models that utilize individual, pairwise, and contextual data to describe the genetic relatedness of pathogen isolates in an endemic setting[9] can identify drivers of transmission with fewer assumptions and computational needs than some previous studies and do not require data on non-cases or SNV cutoffs. This method involves a regression model in which the outcome is a measure of genetic similarity between a pair of isolates, and covariates are assessed for their influence on genetic similarity, which can be considered in many cases a proxy for transmission.

In this analysis, we evaluate the use of pairwise models to describe how carbapenem-resistant *Klebsiella pneumoniae* (CRKP), an important healthcare-associated pathogen, is transmitted in a long-term acute care hospital (LTACH). CRKP's high prevalence in LTACHs make delineating transmission pathways complex with traditional epidemiologic methods. Although we have limited epidemiologic information in this dataset, we are able to identify known transmission risk factors with our models and hope this study can serve as an example of how to conduct this type of analysis in settings with richer epidemiologic data.

In addition to evaluating different approaches for incorporating genomic relatedness into pairwise statistical models, we also assess the sensitivity of these models to case capture. Sampling strategy may be important in understanding CRKP transmission as asymptomatic colonization with CRKP is a common precursor to invasive infection[10] and potentially important in intra-facility spread,[11] yet most facilities do not screen for asymptomatic colonization. To understand the impact of the sampling scheme on identification of transmission risk factors we evaluated models with a more passive surveillance strategy for detection of carriers.

Methods

Study population.

CRKP surveillance samples were collected via rectal swab on admission and every two weeks from June 2012-June 2013 for all patients (n = 937 unique patients) in an LTACH in Chicago, Illinois (USA).[12] Average daily patient census was 98 (SD: 7.4), and the median length of stay was 27 days (IQR: 17, 44). The mean age of patients was 60.5 years (SD: 15.8), and 43.1% of patient-days were for ventilated patients. This study was approved by the institutional review boards at Rush University Medical Center (Chicago, IL, USA) and the University of Michigan (Ann Arbor, MI, USA). Informed consent was waived.

Surveillance samples were cultured and unique colony morphologies were identified to species. Ertapenem disks were used to screen isolates for CRKP and a confirmatory PCR was conducted to detect *bla*$_{\text{KPC,}}$ the sole carbapenemase gene associated with CRE in the region during the study period. To capture contact patterns, each patient's daily room and floor locations were recorded. Antibiotic usage over time was also recorded for each patient. Whole genome sequences were obtained for all positive isolates and recombination-filtered core genome alignments were produced for each sequence type.[13] For this analysis, only sequence type 258 (ST258) isolates, the most common sequence type in the LTACH (70% of cases), were used.

Regression models of pairwise genomic relatedness.

We constructed pairwise regression models in which each observation was a pair of CRKP-positive patients with the individual in the pair who tested positive first being considered the donor and the other individual the recipient. We assessed three different measures of genomic relatedness to be used as outcomes in these models: (1) a log-linear model of core genome single nucleotide variant (SNV) distances between the two isolates, (2) a logistic regression model with a binary outcome indicating whether the potential donor was the most closely related potential donor for a given recipient (based on SNV distance), and (3) a logistic regression model with a binary outcome indicating if the pair's isolates were previously determined to be in the same genomic cluster using a threshold-free clustering method.[13] Each of these measures assess the extent to which cases are linked by transmission, with lower genomic distance or cluster co-membership indicating a higher likelihood of direct transmission.

Only donor-recipient pairs where the two individuals overlapped in the facility during their pairwise exposure period (from the last time the potential donor tested negative, or time of admission for admission-positive donors, to the first time the recipient tested positive) were included in the models. If multiple isolates were available for a patient, only the closest related isolates (based on SNV distances) for each pair of patients were used. To account for the influence of unusually infectious individuals, all models included a random intercept term for each potential donor. Covariates in the models included whether the pair shared time on the same floor (and which floor) or in the same room during their exposure period, time between sample collection of positive cultures in weeks (a measure of similarity of colonization timing), dichotomous antibiotic receipt by the donor during the exposure period (stratified into carbapenem and non-carbapenem groups), and time period within the study (broken up into quarters). Statistical analyses were conducted using *R* v.4.2.2.[14]

Evaluation of the effect of serial sampling on risk-factor estimates.

As most facilities do not have robust serial sampling strategies like our study facility implemented, we re-ran all models including only patients who tested positive on admission or had a positive CRKP test as part of clinical evaluation outside of the colonization study to examine the influence of serial sampling on the ability to make inferences on transmission dynamics.

Results

Prevalence of colonization and infection with CRKP in a single LTACH.

In total, 255 individuals were colonized with at least one strain of CRKP during the study period (with an average prevalence of 32% throughout the year),[13] 180 of whom (70% of those colonized) were colonized with strain ST258. Of the 180 patients colonized with CRKP ST258, 87 (48%) were positive on admission, 72 (40%) had CRKP detected via clinical testing, and 54 (30%) were detected after admission during serial sampling and never had a clinical CRKP isolate. There were 37 genomic clusters of ST258 (2-16 isolates per cluster) previously identified in this study population with a threshold-free cluster detection approach that clustered each CRKP isolate acquired at the LTACH to the importation isolate with which it shared the greatest number of variants.[13]

Genetic relatedness was greatest among CRKP pairs sharing a room or floor.

The median pairwise SNV distance between all ST258 isolates who overlapped in the facility was 53 (IQR: 38-86); for closest donor pairs and same-cluster pairs this value was 5 (IQR: 1-24) and 3 (IQR: 1-6), respectively, which are consistent with a previous study identifying 21 SNVs as an appropriate cutoff for ST258 intra-facility transmission[15] (**Table 1, Figure 1**). Room and floor sharing (particularly Floor D which housed the high acuity unit) during the pairwise exposure period was more common among closest-donor pairs and same cluster pairs than for all possible pairs (**Table 1**).

Twenty percent (95% CI: 9%-38%) of pairs who shared a room during their exposure window had CRKP isolates in the same cluster, and 15% (95% CI: 5%-31%) were closest donor pairs. Pairs who shared a floor but not a room during their exposure period were in the same cluster 7% of the time (95% CI: 5%-8%) and contained a closest potential donor 5% of the time (95% CI: 4%-6%). By contrast, pairs that did not overlap in the same room or floor during their exposure period were in the same genomic cluster only 3% of the time (95% CI: 2%-4%) and the closest potential donor to their recipient 3% of the time (95% CI: 2%-4%) (**Figure 2**). Fifty-eight percent (95% CI: 50%-67%) of closest donor pairs were in the same genomic cluster, while only 3% (95% CI: 2%-3%) of non-closest donor pairs were in the same cluster.

Pairwise models suggest room sharing, residing on the floor that housed the high acuity unit, and shorter time lags between colonization detection are associated with genetic relatedness. In all pairwise models shared time on either Floor A or Floor D was associated with increased pairwise genomic relatedness, with Floor D (which contained the facility's high acuity unit) having the larger effect on pairwise genomic relatedness, suggesting there may be more intra-floor risk on floors where patients require more intensive care (**Table 2**). In Model 1 (pairwise distance), both patients residing on Floor D was associated with SNV distances 44% (95% CI: 38% - 49%) closer. In Model 2 (closest donor), sharing time on floor D was associated with 7 (95% CI: 4-13) times greater odds of being the closest potential donor. In Model 3 (same transmission cluster), sharing time on floor D was associated with 10 (95% CI: 6-18) times greater odds of being in the same cluster. When collapsing the effect of floor sharing into a single covariate, sharing a floor during their exposure period was a significant predictor of genetic relatedness of CRKP isolate pairs in all models (**Supplemental Table 1**).

Other factors, such as sharing a room during a pairs' exposure period, a shorter lag between positive cultures, and the exposure period occurring in the fourth quarter of the study period were positively associated with isolate similarity in each model, although certainty varied. Model 3 identified all three of these factors as significantly related to cluster comembership, Model 1 captured two of these factors (shared room and time period) as significantly associated with SNV distances, and Model 2 failed to show a statistically significant effect of any of these factors on the likelihood of being the closest potential donor **(Table 2)**. Antibiotic exposure of the donor during the pairwise exposure period was not a meaningful predictor of CRKP relatedness in any of the models.

Case and admission-positive only models underestimate key risk factors.

Although our study utilized serial surveillance for asymptomatic carriage, most facilities have only clinical culture isolates available, with some also testing for CRKP colonization on admission. Of the 180 patients colonized with ST258 CRKP in our study, 30% would never have been identified if only clinical and admission screening cultures were conducted, and 61% of closest potential transmission pairs (defined as in Model 2) would have been missed. When we excluded these patients from our analyses, culture date difference remained a significant predictor of genomic similarity, but room and floor sharing was not significant in any of the three models (although the qualitative direction of coefficients remained unchanged) (**Supplemental Table 2**).

Discussion

Using regression models of pairwise genomic relatedness, we were able to identify risk factors for CRKP transmission in an endemic LTACH setting. Although certainty varied, regardless of the metric of genomic relatedness employed, sharing a room or having an overlapping stay on the same floor, especially the floor that included the high acuity unit, predicted shorter genomic distances and a higher likelihood of membership in the same cluster. This is consistent with studies showing increased risk of CRKP infection among those with more intense care needs (e.g. fecal incontinence, mechanical ventilation) and those exposed to infected roommates.[16–18] This work suggests that decolonization and other infection prevention efforts should be focused on close within-facility contacts of CRKP patients, with particular attention to high-acuity patients who have higher illness severity and are likely to have more medical interventions and direct hands-on contact with staff.

The availability of WGS data from colonization isolates gave us the ability to evaluate how individual, dyadic, and contextual factors predicted the genetic similarity of CRKP isolates, a proxy for transmission risk. Using these WGS data, we quantified the relatedness of CRKP isolate pairs in three ways: SNV distance, closest potential donor, and cluster co-membership. Each of these measures of relatedness yielded risk-factor estimates which were consistent with each other as well as existing literature on CRKP transmission. This suggests that the measure of relatedness used in pairwise models may be flexible. It may be important, however, to consider the sampling strategy and transmission dynamics of the pathogen of interest when selecting a pairwise metric. For instance, if serial sampling was not conducted and it is unlikely that direct transmission pairs have been identified, a closest donor approach may not be sensible. Or, if the pathogen of interest has a well-established SNV cutoff to determine cluster co-membership, using the criteria of a pair meeting that cutoff may be used as the model outcome. In our study population, it appears that a threshold-free cluster comembership model identifies transmission risk factors with the most certainty compared to a closest donor or SNV distance models.

Sensitivity analyses revealed that excluding data from serial surveillance isolates reduced our ability to identify the risk factors highlighted using the full dataset. This likely reflected the decreased number of cases overall in the reduced dataset as well as missed direct transmission links, highlighting the importance of serial culture surveillance as a tool for identifying transmission risk factors. When patients who did not have a clinical CRKP isolate during our study period and were negative on admission were excluded from analyses, 61% of probable transmission pairs were missed and our ability to detect an effect of room and floor sharing on genetic relatedness was weakened.

In addition to room and floor sharing, our models revealed that pairs are less likely to be closely genetically related if the time between collection of positive samples is longer. This suggests that a susceptible patient is more likely to get CRKP from someone who has more recently acquired CRKP than someone who has been colonized for a longer time. It could also indicate intra-host evolution between the time of acquisition and transmission. Two of our three models also suggested that individuals colonized during the last quarter of our study period were more closely related to their potential donors than those infected in the first period. This could be an artifact of model setup (as the study period progressed, the number of potential donors increased), the result of the introduction of a new strain into the facility with different transmission patterns, or more intra-facility transmission in this period. However, incidence of CRKP within the facility appeared to decrease throughout the study period,[12] and thus this result may indicate the onward transmission of fewer strains within the facility resulting in those infected appearing to be more closely related to each other than if many strains were circulating due to a bottleneck effect. Lastly, although antibiotic exposure has been associated with CRKP acquisition risk in healthcare facilities,[16,17] our results do not provide evidence that antibiotic exposure is associated with a change in the number of transmissions generated by a colonized or infected LTACH resident. However, the very high prevalence of antibiotic use in our patient population may have hindered detection of their impact on transmission.

Due to data availability and methodological constraints, our study can be considered to have the following limitations. First, we did not have access to information on patient-level procedures, devices, and healthcare worker exposures, which all may play a role in transmission and could help determine specific mechanisms increasing transmission risks, specifically on Floor D. However, we were able to identify known risk factors of CRKP transmission in an LTACH and hope this study will serve as a template for facilities that may have more detailed data available. Additionally, given that only 58% of closest donor pairs were in the same genomic cluster, it is likely we are missing some direct transmission links of CRKP in this LTACH. Thus, even the closest donor model may not be completely representative of direct transmission between two patients, and this may be why some transmission risks were not identified in the closest donor model. However, given our strategy of serially sampling every patient in the facility, there is a high probability of direct transmission links being represented in our model outcomes, so risk factors for direct transmission should be picked up even if not all transmission pairs are present in the data. This is supported by our results corresponding with known CRKP risk factors. Additionally, we are not only interested in direct transmission links but also transmission patterns of certain clusters of isolates, both of which could be identified in our models and helpful in infection prevention interventions. Finally, we chose to only include patient-pairs who had overlapping stays in the facility, and thus were unable to identify if sequential room occupation[18] was a risk factor in the LTACH. We chose to limit our pairs to those who had overlapping stays to limit the dataset to more likely direct and staff-mediated transmission scenarios, as we did not find sequential room sharing to be a common transmission source in previous work with this facility.[13]

A caveat of our study design is that it includes demographic and contextual data on only CRKP positive patients. Thus, all inferences are conditional on both members of a pair being colonized. Accordingly, the epidemiologic risk factors identified in our results should be interpreted as driving genomic similarity between isolates from colonized individuals, as opposed to an individual's risk of colonization. This aspect of our study, however, makes it more accessible, as in many community settings, public health datasets only contain case data, and thus methods that necessitate uninfected controls would be infeasible.

As WGS pathogen data become more widely available and inexpensive to obtain, genomic data have an important role to play in routine surveillance of pathogens such as CRKP. Our analysis shows that these data can provide insights in high-prevalence settings which would not be accessible otherwise. Our results also underscore that the choice of genomic relatedness may be important but is also somewhat flexible and is likely to vary by context and goal of the surveillance activity. For example, infection prevention efforts targeted at mitigating the spread of novel drug-resistant variants may utilize different outcomes than those focused on identifying generic transmission risk factors of endemic pathogens. The approach outlined in this analysis requires few assumptions including no arbitrary SNV cutoffs, no uninfected controls, and only modest computational power and suggests that routine WGS-based surveillance may allow for earlier detection and facility-specific intervention in nosocomial outbreaks of CRKP and other pathogens causing significant morbidity and mortality in vulnerable, hospitalized populations.

References

1. Magill, S. S. *et al.* Changes in Prevalence of Health Care–Associated Infections in U.S. Hospitals. *N. Engl. J. Med.* **379**, 1732–1744 (2018).
2. Agency for Healthcare Research and Quality. Health Care-Associated Infections. https://www.ahrq.gov/professionals/quality-patient-safety/patient-safety-resources/resources/hais/index.html.
3. Quainoo, S. *et al.* Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clin. Microbiol. Rev.* **30**, 1015–1063 (2017).
4. Harris, S. R. *et al.* Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *Lancet Infect. Dis.* **13**, 130–136 (2013).
5. Snitkin, E. S. *et al.* Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Sci. Transl. Med.* **4**, 148ra116 (2012).
6. Martin, J. S. H. *et al.* Patient and Strain Characteristics Associated With Clostridium difficile Transmission and Adverse Outcomes. *Clin. Infect. Dis.* **67**, 1379–1387 (2018).
7. Sundermann, A. J. *et al.* Whole-Genome Sequencing Surveillance and Machine Learning of the Electronic Health Record for Enhanced Healthcare Outbreak Detection. *Clin. Infect. Dis.* **75**, 476–482 (2022).
8. Eyre, D. W. *et al.* Probabilistic transmission models incorporating sequencing data for healthcare-associated Clostridioides difficile outperform heuristic rules and identify strain-specific differences in transmission. *PLOS Comput. Biol.* **17**, e1008417 (2021).

9. Warren, J. L., Chitwood, M. H., Sobkowiak, B., Colijn, C. & Cohen, T. Spatial modeling of M. tuberculosis transmission with dyadic genetic relatedness data. *Biometrics* **79**, 3650–3663 (2023).

10. Kontopoulou, K. *et al.* The clinical significance of carbapenem-resistant Klebsiella pneumoniae rectal colonization in critically ill patients: from colonization to bloodstream infection. *J. Med. Microbiol.* **68**, 326–335 (2019).

11. Spencer, M. D. *et al.* Whole Genome Sequencing detects Inter-Facility Transmission of Carbapenem-resistant Klebsiella pneumoniae. *J. Infect.* **78**, 187–199 (2019).

12. Hayden, M. K. *et al.* Prevention of Colonization and Infection by Klebsiella pneumoniae Carbapenemase-Producing Enterobacteriaceae in Long-term Acute-Care Hospitals. *Clin. Infect. Dis.* **60**, 1153–1161 (2015).

13. Hawken, S. E. *et al.* Threshold-free genomic cluster detection to track transmission pathways in health-care settings: a genomic epidemiology analysis. *Lancet Microbe* **3**, e652–e662 (2022).

14. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing (2022).

15. David, S. *et al.* Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread. *Nat. Microbiol.* **4**, 1919–1929 (2019).

16. Mills, J. P., Talati, N. J., Alby, K. & Han, J. H. The Epidemiology of Carbapenem-Resistant *Klebsiella pneumoniae* Colonization and Infection among Long-Term Acute Care Hospital Residents. *Infect. Control Hosp. Epidemiol.* **37**, 55–60 (2016).

17. Swaminathan, M. *et al.* Prevalence and Risk Factors for Acquisition of Carbapenem-Resistant Enterobacteriaceae in the Setting of Endemicity. *Infect. Control Hosp. Epidemiol.* **34**, 809–817 (2013).

18. Wu, Y.-L. *et al.* Exposure to infected/colonized roommates and prior room occupants increases the risks of healthcare-associated infections with the same organism. *J. Hosp. Infect.* **101**, 231–239 (2019).

Tables and Figures

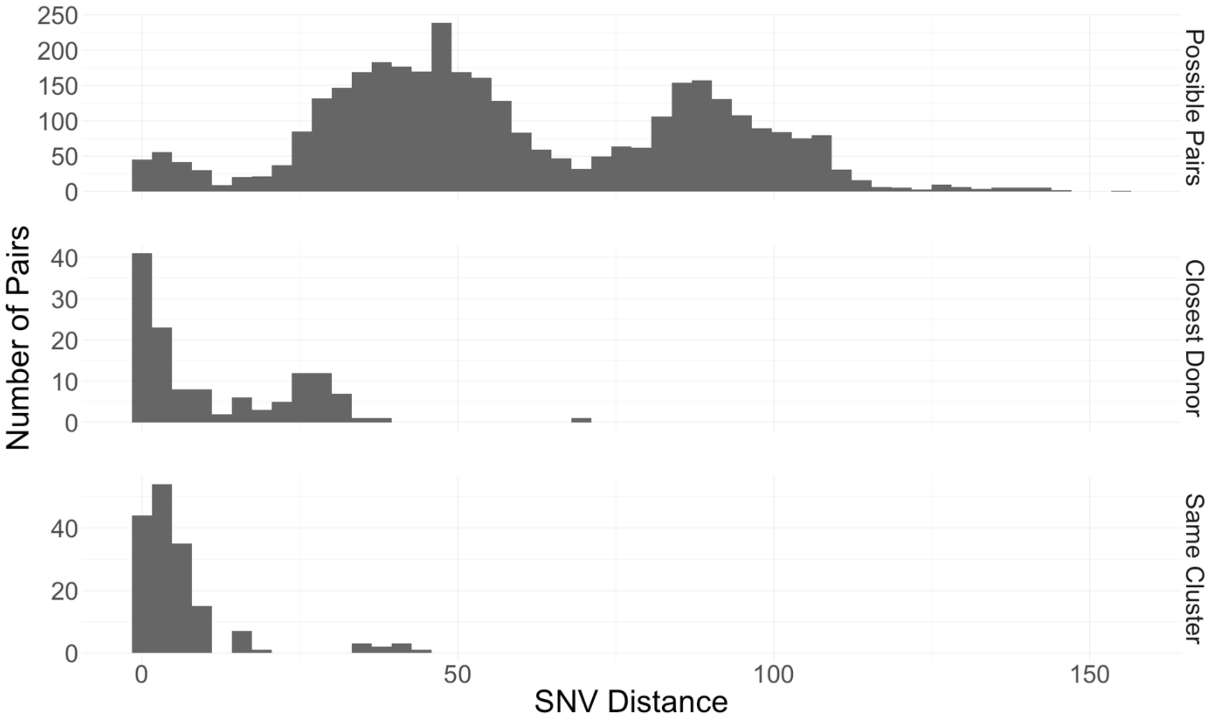

Figure 1. Distribution of single nucleotide variant (SNV) distances between different types of infectious/exposed pairs. From top to bottom, the panels show the distribution of SNV distances for 1) pairs who overlapped in the facility, 2) distances between each recipient and their genomically closest possible donor (based on core genome SNV distance) who overlapped in the facility, and 3) pairs who overlapped in the facility that also belong to the same genomic cluster.

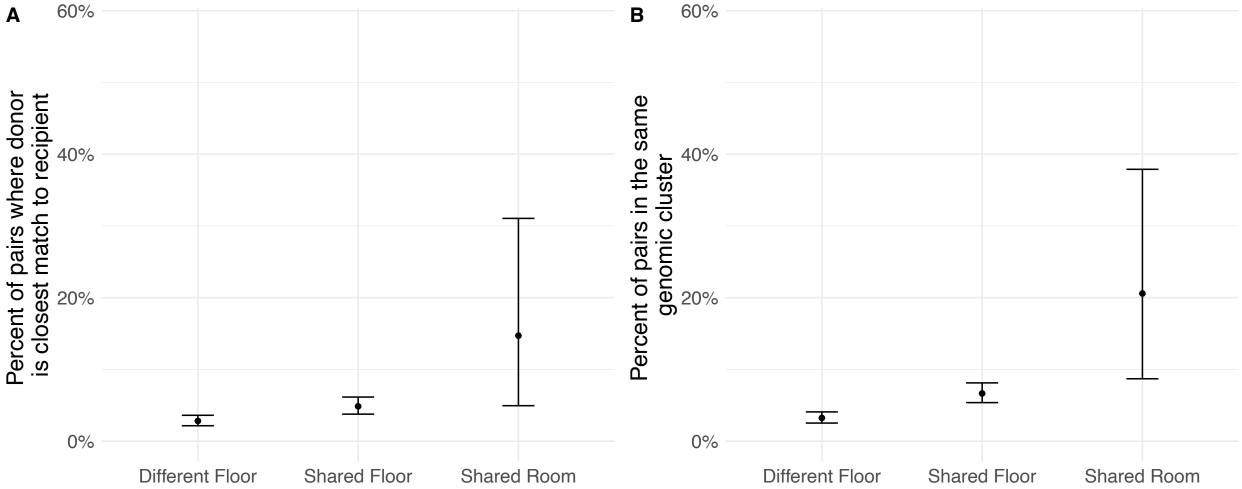

Figure 2. Percent of potential transmission pairs where (A) the infectious individual is the closest potential donor for the recipient and (B) the infectious and exposed patients are in the same genomic cluster. Each panel shows risks associated with residing on different floors in the same facility, residing on the same floor, and from sharing a room. Vertical bars represent 95% confidence intervals.

Table 1. Individual, dyadic, and contextual characteristics of pairs of CRKP infected or colonized patients. Results are shown for 1) all pairs who overlapped in the facility, 2) each recipient and their genomically closest possible donor (based on core genome SNV distance) who overlapped in the facility, and 3) pairs who overlapped in the facility that also belong to the same genomic cluster.

	Possible Pairs n = 3,500	**Closest Donor Pairs** n = 130	**Same Cluster Pairs** n = 165
Average SNV Distance	53 (38, 86)	5 (1, 24)	3 (1, 6)
Shared Room[1]	34 (1.0%)	5 (3.8%)	7 (4.2%)
Shared Floor (any)[1]	1,371 (39%)	70 (54%)	96 (58%)
Floor A	102 (2.9%)	6 (4.6%)	6 (3.6%)
Floor B	765 (22%)	21 (16%)	21 (13%)
Floor C	252 (7.2%)	10 (7.7%)	7 (4.2%)
Floor D (includes high acuity unit)	210 (6.0%)	32 (25%)	61 (37%)
Floor E	42 (1.2%)	1 (0.8%)	1 (0.6%)
Average Culture Date Difference (months)	1.40 (0.57, 3.27)	0.93 (0.43, 2.41)	0.47 (0.20, 1.37)
Antibiotic exposure in donor (any)[1]	2,961 (85%)	108 (83%)	140 (85%)
Carbapenem Antibiotic	1,417 (40%)	53 (41%)	61 (37%)
Non-Carbapenem Antibiotic	2,911 (83%)	106 (82%)	139 (84%)
Time Period[2]			
Period 1	568 (16%)	18 (14%)	15 (9.1%)
Period 2	771 (22%)	29 (22%)	23 (14%)
Period 3	992 (28%)	35 (27%)	33 (20%)
Period 4	1,169 (33%)	48 (37%)	94 (57%)

Medians with interquartile ranges are shown for continuous variables and counts with percentages are shown for categorical variables.

[1] During pairwise exposure period

[2] Time period when recipient first tested positive for CRKP

SNV = single nucleotide variant.

CRKP = carbapenem-resistant *Klebsiella pneumoniae*

Table 2. Drivers of variation in pairwise genomic relatedness as a function of individual and pair-level risk factors (n = 180 patients, 3500 pairs). Model 1 is a log-linear model of pairwise SNV distance as a function of individual and pairwise exposure risks. Coefficients are exponentiated and can be interpreted analogously to rate ratios, with values < 1 indicating smaller distances and > 1 indicating greater distances. Models 2 & 3 are logistic regression models characterizing changes in the odds that a given infectious case is the most closely related to the recipient (Model 2) or that the infectious case and exposed individual are in the same genomic cluster (Model 3). All results are adjusted for all covariates in the model, and a random effect for potential donor is included in the models.

	Model 1. SNV Distance Model	Model 2. Closest Donor Model	Model 3. Same Cluster Model
Intercept	**51.65 (46.27, 57.66)**	**0.02 (0.01, 0.04)**	**0.01 (0.00, 0.03)**
Shared Floor A[1]	**0.80 (0.70, 0.91)**	**3.19 (1.21, 8.42)**	**3.73 (1.32, 10.54)**
Shared Floor B[1]	1.00 (0.94, 1.06)	1.11 (0.63, 1.95)	1.22 (0.68, 2.18)
Shared Floor C[1]	1.08 (0.99, 1.18)	1.41 (0.66, 3.03)	1.08 (0.44, 2.68)
Shared Floor D[1] (Includes High Acuity Unit)	**0.56 (0.51, 0.62)**	**7.12 (3.84, 13.19)**	**10.1 (5.77, 17.8)**
Shared Floor E[1]	**1.32 (1.07, 1.62)**	1.19 (0.14, 10.28)	0.45 (0.04, 4.77)
Shared room[1]	**0.75 (0.60, 0.93)**	2.68 (0.78, 9.17)	**4.15 (1.07, 16.18)**
Culture date difference (30 days)	1.01 (1.00, 1.03)	0.93 (0.82, 1.06)	**0.77 (0.67, 0.90)**
Carbapenem antibiotic exposure of donor[1]	1.02 (0.96, 1.08)	1.16 (0.71, 1.89)	0.93 (0.55, 1.59)
Non-Carbapenem antibiotic exposure of donor[1]	1.03 (0.96, 1.11)	0.72 (0.40, 1.29)	0.95 (0.51, 1.79)
Time period (quarter 2 v quarter 1)	0.95 (0.87, 1.04)	1.25 (0.61, 2.59)	1.60 (0.69, 3.67)
Time period (quarter 3 v quarter 1)	0.98 (0.88, 1.10)	0.97 (0.44, 2.13)	1.85 (0.78, 4.42)
Time period (quarter 4 v quarter1)	**0.87 (0.76, 0.99)**	1.23 (0.54, 2.79)	**4.05 (1.66, 9.85)**

Estimates are exponentiated and 95% confidence intervals are in parentheses. Bolded values have confidence intervals that do not contain 1.

[1] During pairwise exposure period

SNV = single nucleotide variant.

Supplementary Results

Supplemental Table 1. Results of pairwise regression models with shared floor as a single covariate (n = 180 patients, 3500 pairs). Model 1 is a log-linear model of pairwise SNV distance as a function of individual and pairwise exposure risks. Coefficients are exponentiated and can be interpreted analogously to rate ratios, with values < 1 indicating smaller distances and > 1 indicating greater distances. Models 2 & 3 are logistic regression models characterizing changes in the odds that a given infectious case is the most closely related to the recipient (Model 2) or that the infectious case and exposed individual are in the same genomic cluster (Model 3). All results are adjusted for all covariates in the model, and a random effect for potential donor is included in the models.

	Model 1. SNV Distance Model	Model 2. Closest Donor Model	Model 3. Same Cluster Model
Intercept	**49.66 (44.34, 55.61)**	**0.02 (0.01, 0.04)**	**0.01 (0.00, 0.03)**
Shared Floor[1]	**0.92 (0.88, 0.97)**	**2.10 (1.41, 3.14)**	**2.72 (1.82, 4.06)**
Shared Room[1]	**0.74 (0.59, 0.92)**	2.91 (0.93, 9.11)	**3.99 (1.18, 13.45)**
Culture Date Difference (per 30 days)	1.02 (1.00, 1.04)	**0.85 (0.75, 0.96)**	**0.68 (0.58, 0.79)**
Carbapenem antibiotic exposure of donor[1]	1.04 (0.97, 1.10)	1.05 (0.65, 1.70)	0.78 (0.46, 1.32)
Non-Carbapenem antibiotic exposure of donor[1]	1.03 (0.96, 1.11)	0.69 (0.39, 1.23)	0.89 (0.48, 1.63)
Time period (quarter 2 v quarter 1)	0.95 (0.87, 1.04)	1.38 (0.67, 2.81)	1.80 (0.80, 4.04)
Time period (quarter 3 v quarter 1)	0.98 (0.88, 1.10)	1.08 (0.50, 2.34)	2.02 (0.86, 4.75)
Time period (quarter 4 v quarter1)	**0.87 (0.76, 0.99)**	1.52 (0.68, 3.37)	**4.42 (1.84, 10.61)**

Estimates are exponentiated and 95% confidence intervals are in parentheses. Bolded values have confidence intervals that do not contain 1.

[1] During pairwise exposure period

SNV = single nucleotide variant.

Supplemental Table 2. Results of pairwise regression models excluding individuals who did not have a positive colonization isolate on admission or a CRKP isolate detected via clinical testing during the study period (n = 121 patients, 1354 pairs). Model 1 is a log-linear model of pairwise SNV distance as a function of individual and pairwise exposure risks. Coefficients are exponentiated and can be interpreted analogously to rate ratios, with values < 1 indicating smaller distances and > 1 indicating greater distances. Models 2 & 3 are logistic regression models characterizing changes in the odds that a given infectious case is the most closely related to the recipient (Model 2) or that the infectious case and exposed individual are in the same genomic cluster (Model 3). All results are adjusted for all covariates in the model, and a random effect for potential donor is included in the models.

	Model 1. SNV Distance Model	Model 2. Closest Donor Model	Model 3. Same Cluster Model
Intercept	**51.00 (43.35, 60.01)**	**0.03 (0.01, 0.10)**	**0.01 (0.00, 0.04)**
Shared Floor[1]	0.97 (0.91, 1.04)	1.57 (0.85, 2.89)	1.29 (0.63, 2.61)
Shared Room[1]	0.90 (0.67, 1.22)	2.07 (0.34, 12.61)	2.40 (0.21, 27.72)
Culture Date Difference (per 30 days)	**1.03 (1.01, 1.05)**	**0.84 (0.71, 0.99)**	**0.74 (0.60, 0.91)**
Carbapenem antibiotic exposure of donor[1]	0.97 (0.89, 1.07)	1.70 (0.84, 3.43)	1.52 (0.64, 3.65)
Non-Carbapenem antibiotic exposure of donor[1]	1.01 (0.89, 1.15)	0.41 (0.17, 1.02)	0.84 (0.27, 2.60)
Time period (quarter 2 v quarter 1)	1.06 (0.94, 1.20)	1.07 (0.36, 3.18)	1.20 (0.25, 5.66)
Time period (quarter 3 v quarter 1)	0.93 (0.80, 1.09)	1.21 (0.40, 3.66)	2.61 (0.58, 11.77)
Time period (quarter 4 v quarter1)	**0.82 (0.69, 0.98)**	2.16 (0.74, 6.29)	**7.83 (1.78, 34.39)**

Estimates are exponentiated and 95% confidence intervals are in parentheses. Bolded values have confidence intervals that do not contain 1.

[1] During pairwise exposure period

SNV = single nucleotide variant.