Title: Rapid statistical methods for inferring intra- and inter-hospital transmission of nosocomial pathogens from whole genome sequence data

One Sentence Summary: We present rapid statistical methods for distinguishing intra- versus inter- hospital transmission of bacterial pathogens using whole genome sequence data; these methods do not require the use of SNP thresholds or the generation and interpretation of phylogenetic trees.

Authors: Marianne Aspbury^{1,2†}, James Sciberras^{1,4†}, Jukka Corander³, Sion C. Bayliss⁴, Tjibbe Donker^{5,6}, Edward J. Feil^{4†*}, Richard James^{1‡}

Affiliations:

¹Department of Physics and Centre for Networks and Collective Behaviour, University of Bath, Bath, UK.

²Department of Natural Sciences, University of Bath, Bath, UK.

³Department of Biostatistics, University of Oslo, N-0317, Oslo, Norway.

⁴The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK

⁵The National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, UK.

⁶Nuffield Department of Medicine, University of Oxford, UK

*To whom correspondence should be addressed: E.Feil@bath.ac.uk

†‡ These authors contributed equally.

Abstract:

Whole genome sequence (WGS) data for bacterial pathogens can provide evidence as to the source of nosocomial infection, and more specifically the ability to distinguish between intra- and inter-hospital transmission. This is currently achieved either through using SNP thresholds, which can lack statistical robustness, or by constructing phylogenetic trees, which can be computationally expensive and difficult to interpret. Here we compare two alternative statistical approaches using 1022 genomes of methicillin resistant *Staphylococcus aureus* (MRSA) clone ST22. In 71% of cases both methods predict the same hospital origin, which is also supported by the ML tree. Robust assignments are divided approximately equally between intra-hospital transmission and inter-hospital transmission. Our approaches are rapid and produce intuitive output that could inform on immediate infection control priorities, as well as providing long-term data on inter-hospital transmission networks. We discuss the strengths and weakness of our methods, and the generalisability of this approach.

Main Text:

Introduction

Whole-genome sequencing (WGS) has proven utility for molecular epidemiological surveillance of nosocomial pathogens, and provides the power to track the emergence and spread of high-risk clones within single health-care settings, or on regional, national or global scales (1–7). The recent development of community-oriented database platforms, analytical methods and data visualisation tools (8, 9) provide the critical infrastructure for WGS data to assume a pivotal role in guiding infectious disease management policies. These platforms provide powerful community resources for epidemiological surveillance and detailed retrospective studies. However, significant challenges remain with respect to reconstructing local transmission pathways and source attribution within a time frame that can inform on immediate infection control priorities. There is an urgent need for the establishment of a national platform based on WGS data that can provide real-time statistical probabilities of intra-hospital transmission indicative of an emerging outbreak, versus inter-hospital transmission via staff movement or patient referral (10, 11).

Conventional epidemiological information such as date of infection, a patient’s current or past ward / hospital, and detailed clinical data continue to underpin infection control efforts. However, the use of large representative reference WGS databases, combined with local capacity for routine high throughput sequencing, provides substantial additional power to resolve transmission networks and define outbreaks (5, 12). Currently, epidemiological inference from WGS data almost always involves the reconstruction of a phylogenetic tree; for recent examples see (5, 13). Whilst this is an established approach for retrospective studies, the utility of building and interpreting trees for real-time infection control is more limited. In addition to being computationally expensive and thus potentially time-consuming, phylogenetic trees are not always statistically robust, and their interpretation is not always simple. Phylogenetic clusters are often defined on pragmatic grounds, and any given query isolate may fall awkwardly between two clusters, or else within a single cluster containing isolates from multiple other hospitals. The use of SNP thresholds for distinguishing between intra- versus inter-hospital transmission offers a practical short-cut, but one which either provides no phylogenetic context, little statistical rigour or measures of uncertainty, or else will require regular re-evaluation in the light of rapidly expanding databases. Moreover, in those cases where SNP thresholds rule out intra-hospital transmission this approach provides no direct evidence as to the likely source of the isolate.

Here we address this critical methodological gap by proposing and validating two rapid independent statistical approaches, one heuristic and one Bayesian, each of which generate intuitive graphical output, in a standardized format, regarding the probability of intra- versus inter-hospital transmission. The methods both predict the most likely hospital origin of a query isolate by comparison to a reference database, and require no capacity or experience in building or interpreting phylogenetic trees. We validated the methods using a large (>1000 genomes) WGS dataset of the important *Staphylococcus aureus* hospital-acquired (HA)-MRSA clone ST22 (EMRSA-15), representing multiple hospitals in the UK and Ireland and isolated over a ten year window (2001-2010) (2, 6, 14). We consider the robustness of these methods with respect to sampling evenness, intra-hospital diversity and temporal breadth.

The methods are broadly consistent with each other and with inferences based on a maximum-likelihood (ML) tree. Discrepancies between the methods either represent subtle yet important

strengths and weaknesses of each method, or else a paucity of signal in the data. The analysis provides robust hospital assignments, defined as being consistent for both methods and with the ML tree, for over two-thirds of the 91 isolates tested against the reference database. Approximately half of these robust assignments point to intra-hospital transmission, and half to inter-hospital transmission. Surprisingly, almost 40% of inter-hospital events are long-range transmissions between “referral regions”. Finally, we discuss the utility of these methods for infection control and, more broadly, in reconstructing transmission networks from large WGS datasets. We also consider the potential challenges in deploying our approach to other species and datasets.

Results

Assignment of intra- versus inter hospital transmission

Preliminary analyses confirmed substantial geographical structure data (see Methods). The ML tree for all 1022 isolates is given in Figure 1. The colours of the inner ring and tree branches correspond to referral clusters and the colours of the outer ring refer to hospitals. These colour assignments are consistent with all subsequent figures, and supplemental figures showing individual clades are all taken from this tree. We took each of the 91 isolates from 2010 in chronological order (earliest first) as query isolates, and used our methods to compare these against the database of all remaining isolates from 2001-2009. Once a query isolate had been analysed, it was then added to the reference database. We compared the outputs of the heuristic and Bayesian methods (see Methods) against each other, against the ML tree and against the known origin of each isolate at the level of hospital and referral cluster. We provide examples of the outputs in the supplementary material to illustrate each of the cases we discuss below (Figures S1-S8). Figure 2 summarises these comparisons. At the hospital level, 32 of the 91 query isolates (35%) gave consistent predictions for both methods, which also concurred with the position of the isolates on the ML tree, and with the actual hospital origin. Thus, about one third of the 91 isolates from 2010 are unambiguously associated with one specific hospital. At the level of referral cluster, this figure increases to almost two-thirds ($n=57$; 62%). In many cases the signal from the tree is unambiguous, in that the query strain is embedded within a clade consisting of isolates representing a single hospital (see example in Figure S1). However, both methods are also robust in cases where the query isolate happens to be closely related to isolates that do not correspond to the majority hospital of the other isolates in the clade, (e.g. Figure S2), or when the query is closely related to only a single isolate corresponding to the originating hospital (e.g. Figure S3).

There are 49 cases (54%) where neither method produces an output consistent with the actual hospital origin (Figure 2). Of these, there are 33 cases (36% of total) where the two methods consistently predict the same hospital origin, but the predicted hospital is different from the actual hospital origin. These are likely to represent genuine inter-hospital transmission events. Thirteen of these 33 cases (39%) represent transmission events between referral clusters, whilst the remaining 20 cases (61%) represent transmission between hospitals in the same referral cluster. Transmission events supported by consistency between the two methods are also consistent with the phylogenetic analysis. An example of an inter-hospital transmission across referral clusters is given in Figure S4. An example of an inter-hospital transmission within a referral cluster is given in Figure S5.

Thus, at the hospital level, the 91 query isolates are divided into three roughly equal groups: 35% are unambiguously associated with the hospital from which they were sampled and represent intra-hospital transmission, 36% appear to be likely inter-hospital transmission events (of which 39% represent transmission between referral clusters and 61% transmission between hospitals within the same referral clusters), and 29% are of more uncertain origin as the two methods produce inconsistent outputs. At the level of referral cluster, 62% of the isolates are clearly associated with the referral cluster from which they were sampled, 13 (14%) are likely transmission events between referral clusters (both methods agree but differ from the actual origin), and 24% are of more uncertain origin.

Considering cases where the methods are inconsistent

In order to compare the performance of the two methods in differing scenarios, we then considered those cases where they produced inconsistent outputs. At the hospital level, there are five cases where the Bayesian output is consistent with the actual hospital origin but the heuristic output is not, and another 5 cases where the reverse is true (heuristic output is consistent with actual hospital origin but Bayesian output is not). This points to similar performance for the two methods, but indicates that each method may be more reliable under specific circumstances. At the level of referral cluster there are 10 cases where the heuristic output is consistent with the actual origin (but different from the Bayesian output) and only 3 cases where the reverse is true. In order to understand why the methods occasionally gave inconsistent results, we examined each of these cases in turn in the context of the phylogenetic tree. First, we considered the 5 cases where the output from the Bayesian method was consistent with the actual hospital origin but the heuristic method was not. In 3 of these cases, the output from the heuristic was also inconsistent with the known origin at the referral cluster level. These cases share two properties. First, the signal from the phylogenetic tree is ambiguous; that is, no single hospital is associated with isolates that are obviously closely related to the query. Second, there are very few isolates from the hospital from which the query was recovered in the reference database, but those that are present are closely related to the query. In such circumstances, the Bayesian method will predict the originating hospital on the grounds that it is more poorly represented than the alternatives, whilst the heuristic method, similar to a casual inspection of the tree, will not explicitly consider differences in sampling depth between the hospitals. An example is given in Figure S6. In this case the Bayesian method predicts the originating hospital as East1A, which is consistent with the actual origin of the query, whereas the heuristic predicts a hospital from a completely different referral cluster (SouthEastC).

Some of the isolates from 2010 represent hospitals that are not represented in earlier isolates. This means that as the query isolates from 2010 are added sequentially to the database in chronological order, there are a number of cases where the actual originating hospital of the query is not represented in the database. In these cases the heuristic cannot assign a probability to this hospital, whereas the prior implemented in the Bayesian method enables it to choose this hospital (see Methods). Thus, this represents another scenario whereby the Bayesian is more likely to predict the actual originating hospital than the heuristic method (see Fig S7 for an example).

The incorporation of sampling depth and evenness in the Bayesian method may provide some additional power in the examples above, but in other cases might produce misleading results.

This is illustrated by the 5 cases where the heuristic output is consistent with the originating hospital, but the Bayesian output is not. This can happen when a single isolate from a different hospital is somewhat related to the query isolate, a situation that might arise through a transmission event independent of the query isolate. If the hospital represented by this isolate is poorly sampled in the database, then the Bayesian method may favour this hospital even if there are isolates from more well-sampled hospitals that are more closely related to the query. If multiple distinct lineages are circulating within a well-sampled hospital, then this hospital will be even more disfavoured by the Bayesian method, as this will lower the aggregate relatedness of isolates from that hospital to the query. In contrast, the heuristic method is blind to intra-hospital diversity and will tend to simply assign the hospital corresponding to the closest isolates, regardless of the presence of more diverged sub-groups within that hospital (Figure S8).

We then considered the 8 cases where the Bayesian and heuristic methods were inconsistent with each other and also with the actual origin of the isolate at the RC level. In 4 of these cases, the position of the query isolate on the tree is consistent with an inter-hospital transmission event. Whilst both methods detect that the query isolates do not clearly correspond to the hospital (or referral cluster) from which they were sampled, they differ in their predictions of where the query isolate originated from. These differences again reflect the level of intra-hospital diversity and the difference in sampling depth between hospitals. The different scenarios are illustrated by the hypothetical trees in Figure 3. In sum, the Bayesian method will favour hospitals that are poorly represented in the database, but are represented by isolates closely related to the query isolate, over hospitals that are well sampled and contain diverse isolates, some of which are closely related to query and others of which are completely unrelated to the query (eg cases 2 and 3 in Figure 3). The heuristic will simply tend to predict hospitals represented by the isolates that are most closely related to the query, regardless of whether those hospitals also contain completely unrelated isolates or how well sampled those hospitals are.

The temporal breadth of the database

Finally, we examined how the temporal spread of the reference database affects the predictions made by the two methods. This was motivated by the observation that those SNPs observed earlier in the sampling window also tend to be the most widely distributed among the hospitals (Figure S13). This means that older SNPs will, on average, be less informative than those SNPs that have arisen more recently. We even considered it possible that the inclusion of very old isolates in the database may result in the predictions becoming less reliable by worsening the signal to noise ratio. We therefore repeated the analysis using the 91 isolates from 2010 in turn as query isolates against the reference database, but consecutively removing the isolates from the earliest year in the database (ie first removing all the isolates from 2001, then all the isolates from 2002 and so on). In each case we recorded the fraction of the isolates in which the predictions agreed with the actual origin (at the hospital and referral cluster level), and the mean score of the predicted hospital (ie the strength of signal). Whilst this analysis does not provide any clear indication that the inclusion of older isolates significantly confounds the final predictions of the methods, the strength of signal of the predicted hospital or referral cluster is weakened with the inclusion of older isolates when using the Bayesian method (Figure S9)

Discussion

Whole genome sequencing (WGS) has rapidly become a central tool for the molecular epidemiology of bacterial pathogens. In addition to shedding light on transmission dynamics, WGS data can provide predictions of antibiotic resistance profiles (15, 16) and, potentially, virulence capability (17, 18), within a time-frame sufficiently short to inform on treatment options. Here we describe two statistical approaches that further exploit WGS data by distinguishing cases of infection arising through intra-hospital transmission, from those more likely to have been imported from outside the hospital (inter-hospital transmission). In 71% of the 91 cases tested the two methods were consistent, and the predicted source concurred with the phylogenetic context of the query isolate. Thus both methods perform well in cases where the phylogenetic signal is relatively strong, and the provision of intuitive output in real-time has significant implications for infection control priorities. If, in a given hospital, the majority of cases were found to be due to intra-hospital transmission, then resources should be prioritised to re-evaluating infection control procedures, and identifying putative sources of infection including individual health-care workers (HCWs). If consecutive cases of intra-hospital transmission occurred over a short period, and the isolates possessed identical “hospital profiles” from our procedures, then this would be strongly indicative of an outbreak. In contrast, if the majority of cases were imported from other hospitals, then screening on admission and/or prior to referral should be prioritised.

In addition to providing an invaluable tool for real-time epidemiology, the data produced by these methods provide useful summary statistics regarding the relative frequency of intra- versus inter-hospital transmission. Given the strong spatial structuring evident from an overall inspection of the ML tree, it might have been expected that intra-hospital transmission is much more frequent than inter-hospital transmission, but our analyses suggests this is not the case. Approximately half of the robustly assigned cases were consistent with intra-hospital transmission and half with inter-hospital transmission. Moreover, of the inter-hospital transmission events, 39% were long-range events between hospitals in different referral clusters. These observations are consistent with the recent report by Auguet et al, which also highlighted the significance of inter-hospital transmission of *S. aureus* ST22 (19). Whilst all of these transmission events were corroborated by an inspection of the tree, this validation step took far longer than generating the output, and compiling summary statistics from a brute-force inspection of the tree alone would be a painstaking and subjective process. The utility of our methods to rapidly catalogue multiple transmission events could pave the way towards constructing transmission networks within and between hospitals based on the frequency of actual events. This contrasts with previous studies based on broad measures of differentiation between bacterial populations circulating in different hospitals (14).

In 29% of the cases we examined, our two methods did not produce consistent results, and in these cases the phylogenetic signal tended to be more ambiguous. Although our methods appear to be reasonably robust to sampling unevenness and temporal breadth, an obvious limitation lies in the fact that the reference database does not represent all of the major hospitals in the UK and Ireland, and some hospitals are only represented by a very small number of isolates. Whereas the Bayesian method can account for cases where the originating hospital of a query isolate is not present in the reference database, the heuristic method cannot, and this can explain some of the discrepancies. An alternative possibility is that the strains did not directly derive from a hospital at all, but from the community. Coll et al have recently demonstrated frequent transmission within the community and between the community and hospitals for *S. aureus* ST22 in the East of England(13). We also note that pre-existing inter-hospital transmission events present in the

database might complicate the assignment of subsequent events, and that there may be rare cases that are essentially intractable regardless of the reference database used or the sophistication of the method.

Whilst presenting two alternative methods provides a useful means by which to gauge their reliability (in terms of consistency between the methods) and their relative strengths and weaknesses, ultimately the use of dual methods is sub-optimal and adds unnecessary complexity. How then to choose which method is most suitable? Whilst our results suggest comparable performance of the two methods, one possibility going forward would be to use the heuristic method to obtain the best possible priors to inform the Bayesian method. These priors could be updated at regular intervals, thus using the heuristic method to optimise the Bayesian method as the reference database grows. Larger, more representative reference databases will be generated over time, and this will mitigate many of the limitations we have identified in this study and will help to further optimise the methodology.

A major question remains as to how generalizable our approach is to other pathogenic species, or even other clones of *S. aureus*. The ability to reliably reconstruct recent transmission chains within and between hospitals is dependent upon the epidemiological characteristics of the clone/species in question. Multidrug resistant clones that transmit primarily through the healthcare system will be much more amenable to this approach than more susceptible clones that are frequently carried and transmitted asymptotically in the community and/or commonly occupy environmental niches. Whilst transmission of MRSA ST22 has recently been shown to occur within the community in the UK, it has also been by far the most common hospital-acquired MRSA in this country over the last two decades or so (3, 6), and has spread world-wide (20). Other *S. aureus* clones, such as USA300, are more frequently associated with transmission within the community (21, 22), and the approach as described will have more limited utility for this clone.

Other major pathogens, such as *E. coli* and *Klebsiella*, may be even less suitable due to frequent recombination, and transmission within the community and environment. However, for such species the method should still have utility when specific nosocomial antibiotic resistant clones are considered. For example, *K. pneumoniae* ST258 is a major carbapenemase producing Enterobacteriaceae (CPE) clone that has clonally disseminated over a short time period throughout the Italian healthcare network (23, 24) and elsewhere (25, 26), and preliminary analysis using an ST258 dataset confirms the usefulness of our approach for this clone (data not shown).

In sum, here we demonstrate that it is possible, and indeed desirable, to draw inferences from WGS data regarding local epidemiology without recourse to phylogenetic analysis or SNP thresholds. We believe the methods as presented provide an important first step towards a more efficient exploitation of WGS for real-time infection control, and for building a more detailed picture of the transmission dynamics of multidrug resistant clones through health-care networks.

Materials and Methods

The isolates

Our analysis was based on the dataset of Donker et al. (14) and consists of 1022 isolates of *S. aureus* ST22 recovered from cases of invasive disease from 46 hospitals from England, Wales, Scotland, Northern Ireland and the Republic of Ireland from 2001 to 2010. The majority of the isolates were collected through the British Society of Antimicrobial Chemotherapy (BSAC) bacteraemia resistance surveillance programme, supplemented with additional isolates retrospectively collected from nine hospitals in the East of England over the same time-frame (14).

We grouped the English hospitals into referral clusters, as previously defined by Donker et al (14) on the basis of geographic proximity and patient movement, as this provides a useful additional geographical scale. For hospitals outside of England, where patient referral data is not available, we consider country or province (Scotland, Wales, Northern Ireland, Republic of Ireland) as a single referral cluster. To preserve anonymity, hospital names were coded, as described previously (14). Figure 4(A) shows the borders of the English referral regions, and Figure 4(B) shows the number of isolates used in this analysis from each referral region broken down by year. There is an approximately even representation of isolates across the different years. Figure 4(C) shows the number of isolates broken down by referral region and hospital. The large excess of isolates from the East of England reflects the inclusion of isolates from the second retrospective study.

The single-nucleotide polymorphisms (SNPs)

The short-read Illumina data was mapped to the ST22 reference genome MRSA252, and the regions spanning the major non-core elements (e.g. SCCmec) were excluded as described previously (2, 14). We discounted all monomorphic sites, and all SNP sites where the frequency of N (ambiguous base) was >1% to ensure quality control. We also removed a small number (n = 24) of SNP sites known to be associated with antibiotic resistance as these are known to be prone to homoplasy, and all tri-allelic (n = 144) and tetra-allelic (n = 4) as these may represent sequencing errors or sites under unusual selection pressure. This left a total of 23,686 bi-allelic SNPs which were carried forward to downstream analyses. Of these, 18,217 were singletons where the minority base is only present in a single isolate, and 5469 were present in >1 isolate, with the most common appearing in 454 of the 1022 isolates (Figure S10). Although singleton SNPs are not phylogenetically informative, we retained them in the reference database as it is possible they will be observed in query isolates, and thereby provide additional signal. The positions of the non-singleton SNPs are given in Figure S11; there is no indication that they are non-randomly clustered within the alignment. The mean number of SNPs per isolate was 53.7 (range 2-131).

Preliminary analyses

We carried out some preliminary analysis to examine the degree of geographical clustering of the SNPs. A maximum likelihood tree based on the SNP data confirmed that the ST22 isolates are clustered at the levels of hospital, referral region and country (Figure 1). We further examined

the degree of clustering on the hospital level by plotting, for each SNP, the number of isolates in which that SNP was found against the number of hospitals. We compared these data with a null distribution based on the number of hospitals in which each SNP is expected to occur, given their frequency, under the assumption that SNPs are randomly distributed among the hospitals (Figure S12). This analysis revealed that SNPs tend to be distributed among far fewer hospitals than would be expected by chance, thus confirming a strong geographical signal within the SNP data. However, as discussed by Donker *et al*, numerous inter-hospital transmission events are still evident from the phylogeny(14).

Previous work has shown that the HA-MRSA ST22 (EMRSA-15) clone emerged from a point source in the English west-midlands in the mid-1980s (3). Continued expansion, combined with occasional long-distance transmission, throughout our sampling period (2001-2010) explains why older SNPs are more common, thus likely to be more widely distributed throughout the health-care network (Figure S13). This implies that the time point at which SNPs are first observed should inversely correlate with the strength of the discriminatory signal for predicting hospital origin; that is, more recent SNPs will be restricted to fewer hospitals and will therefore be more informative.

The statistical methods

Having carried out a preliminary analysis of a quality controlled dataset, we then developed and tested statistical methodologies for predicting hospital origin. For the English hospitals we also considered referral cluster, and for isolates from Scotland, Wales, The Republic of Ireland and Northern Ireland we considered these national regions as the higher order geographical unit. Our approach was to consider each of the 91 isolates from 2010 as query isolates. Each of these isolates were compared in turn chronologically against the reference database consisting of all isolates from 2001-2009 plus all previously compared 2010 isolates. The predictions were then compared to each other, against the context of the isolates in the maximum likelihood tree, and against the actual origin of the isolates. For each comparison we considered origin both at the level of hospital and referral region / country.

To describe each method we labelled hospitals with an index $h = 1, \dots, H$, referral clusters with an index $r = 1, \dots, R$ and variable sites with an index $s = 1, \dots, S$. The reference database of N isolates was stored (Fig. S2) as an $S \times N$ binary matrix $\underline{D}$ whose element D_{si} is 1 if there was a SNP (the rarer nucleotide) at site s in isolate i and is 0 otherwise. The same coding was used for the query isolate, stored in the binary $S \times 1$ vector $\underline{x}$, containing $\sigma = \sum_{s=1}^S x_s$ SNPs. To compare $\underline{x}$ with the n_h isolates in the database that were collected from hospital h , we analysed the $S \times H$ matrix of SNP counts $\underline{C}$, whose element C_{sh} is the number of isolates from h that have a SNP at site s ($0 \leq C_{sh} \leq n_h$). In each classifier we compared $\underline{x}$ with $\underline{C}$ to assign a score y_h indicating the probability that $\underline{x}$ comes from (or is consistent with isolates taken from) hospital h . In principle, the only job of a classifier is to find the class $\hat{h}$ which has the largest score y_h , but we will present and analyse the full set of scores to assess and compare our methods. When the classes of interest were referral clusters, not individual hospitals, we further aggregated isolates in the database to produce a $S \times R$ count matrix whose element

$C_{sr} = \sum_{h \in R} C_{sh}$, and assigned scores y_r to each of the R referral clusters. For clarity we present detailed methods only for the case of classifying to the level of hospital.

In all our analyses the 91 query isolates from 2010 were classified in chronological order. Once classified an isolate became part of the database, by adding $\underline{x}$ as an extra column to the database $\underline{D}$, and associating $\underline{x}$ with whichever hospital and referral cluster it was collected from. The size of the database, and the number of classes, therefore increased as we tested the 2010 isolates. At the outset, $N = 932$, $H = 36$, $R = 15$; by the end $N = 1021$, $H = 46$, $R = 16$, as in 2010 there were isolates from 10 hospitals new to the database and from one new referral cluster. The number of variable sites remained constant ($S = 23,686$) throughout our analysis.

Method 1: A simple heuristic framework (“heuristic”)

We first developed a simple heuristic approach by weighting the presence of a SNP in a given hospital according to the inverse of its overall frequency, reasoning that any classifier giving equal weight to all sites might be prone to having the most useful information swamped by a mass of less differentiating information.

This was achieved by building a classifier that scores only the σ sites that are SNPs in the query isolate $\underline{x}$ and allocates more weight to SNPs previously seen in few hospitals. We scaled the SNP counts C_{sh} (at those sites at which $x_s = 1$) by the total number of isolates in the database that have a SNP at site s to produce a measure,

$$z_{sh} = \frac{C_{sh}}{\sum_{h=1}^H C_{sh}},$$

of the localisation of a given SNP. A SNP that has only ever appeared in one hospital scores a maximum $z_{sh} = 1$ for that hospital, and 0 for all other hospitals. A SNP seen in equal numbers of isolates in two hospitals would produce a z -score of 0.5 for each of those hospitals, and so on. A SNP seen in many hospitals will produce a low z -score for that SNP in each of those hospitals.

The total weight of z -scores for each SNP is 1, so the normalised hospital score for this heuristic classifier is simply

$$y_h = \frac{1}{\sigma} \sum_{s=1}^S x_s z_{sh}.$$

[The x_s appears in this expression to ensure that only SNPs present in $\underline{x}$ are counted.]

Method 2: A Bayesian framework (“Bayesian”)

Our second approach uses conventional naïve Bayes classification (27), in which the hospital scores are posterior probabilities $y_h = L_h P_h / \Omega$, comprising a likelihood function L_h (the probability that query isolate $\underline{x}$ contains the 0s and 1s it does, given that it comes from h) and

any prior expectation P_h we have that $\underline{x}$ comes from h . The normalisation constant

$$\Omega = \sum_h L_h P_h.$$

As there are only two possible states at each site (SNP or non-SNP) we used a Binomial likelihood function, with a Beta prior (hyper-parameters a and b) to estimate the probability that an isolate from hospital h has a SNP at site s as

$$\hat{p}_{sh} = \frac{C_{sh} + a}{n_h + a + b}.$$

Following Gelman et al (28) we used the database itself to fix the hyperparameters $a = 0.045$ and $b = 19.8$ by fitting to the mean and variance of the observed distribution of the number of SNPs per site (Figure S1). The estimated probability that there is not a SNP at site s in hospital h is $(1 - \hat{p}_{sh})$. These two probabilities are sufficient to construct the likelihood function L_h . In a naïve Bayes classifier, each of the contributing pieces of evidence (sites in our case) is treated as independent, so our likelihood function is $L_h = \prod_{s=1}^S (\hat{p}_{sh})^{x_s} (1 - \hat{p}_{sh})^{(1-x_s)}$.

To help us interpret the level of support for a given classification we introduced a dummy class $h = 0$ into the database. This comprised a single, invented, isolate, with a typical number of SNPs (50), none of which aligns with any of the SNPs in the database. The likelihood L_0 is then easily computed, as there are 0 SNPs in S sites, and all of the 50 ‘extra’ SNPs are mismatches. All of the H true hospitals have their likelihood L_h adjusted for the presence of the dummy class by the addition of 50 ‘extra’ non-SNP matches.

In the Bayesian classifier presented here P_h is the product of two prior distributions. The first reflects that fact that there is uneven sampling of the hospitals in the database (Figure 3), so we might expect that a randomly-chosen query isolate is more likely to come from a highly-sampled hospital. The second prior is a simple attempt to account for the known geographical clustering of patient referrals and the MRSA genome in this dataset. Donker et al. (14) estimated the number of ‘candidate introductions’ - isolates that look out-of-place in the phylogenetic tree - in the same database used here. We used the frequencies of these candidate introductions to parameterise a 3-level distribution that asserts our prior expectation that the mostly likely hospital (and referral cluster) for a query isolate is the hospital (referral cluster) it was taken from. We note in passing that for most query isolates, neither of our priors makes much difference to the classification; by far the biggest source of contrast between hospitals comes from the likelihood function. The exception is the situation where the query isolate comes from a hospital that is new to the database, as happens 10 times in our test. In such cases we set the baseline score either to that of the dummy hospital, or to the likelihood L_r of the origin referral cluster if there are already isolates in the database from r .

For practical reasons we computed the log-likelihoods $\log L_h = \sum_{s=1}^S \{x_s \log \hat{p}_{sh} + (1 - x_s) \log (1 - \hat{p}_{sh})\}$ and display the Bayesian log-scores $\log y_h = \log L_h + \log P_h - \log \Omega$. All methods were implemented in R and output was generated for each individual query in seconds on a standard laptop.

All scripts are readily available at <https://github.com/marianne-aspbury/hospital-origin>.

References and Notes

1. S. Willems, S. Kampmeier, S. Bletz, A. Kossow, R. Köck, F. Kipp, A. Mellmann, Whole-Genome Sequencing Elucidates Epidemiology of Nosocomial Clusters of *Acinetobacter baumannii*, *J. Clin. Microbiol.* **54**, 2391–2394 (2016).
2. S. Reuter, M. E. Török, M. T. G. Holden, R. Reynolds, K. E. Raven, B. Blane, T. Donker, S. D. Bentley, D. M. Aanensen, H. Grundmann, E. J. Feil, B. G. Spratt, J. Parkhill, S. J. Peacock, Building a genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic of Ireland, *Genome Res.* **26**, 263–270 (2016).
3. M. T. G. Holden, L.-Y. Hsu, K. Kurt, L. A. Weinert, A. E. Mather, S. R. Harris, B. Strommenger, F. Layer, W. Witte, H. de Lencastre, R. Skov, H. Westh, H. Zemlicková, G. Coombs, A. M. Kearns, R. L. R. Hill, J. Edgeworth, I. Gould, V. Gant, J. Cooke, G. F. Edwards, P. R. McAdam, K. E. Templeton, A. McCann, Z. Zhou, S. Castillo-Ramírez, E. J. Feil, L. O. Hudson, M. C. Enright, F. Balloux, D. M. Aanensen, B. G. Spratt, J. R. Fitzgerald, J. Parkhill, M. Achtman, S. D. Bentley, U. Nübel, A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic, *Genome Res.* **23**, 653–664 (2013).
4. F. Tagini, G. Greub, Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review, *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 2007–2020 (2017).
5. E. S. Snitkin, S. Won, A. Pirani, Z. Lapp, R. A. Weinstein, K. Lolans, M. K. Hayden, Integrated genomic and interfacility patient-transfer data reveal the transmission pathways of multidrug-resistant *Klebsiella pneumoniae* in a regional outbreak, *Sci. Transl. Med.* **9**, eaan0093 (2017).
6. D. M. Aanensen, E. J. Feil, M. T. G. Holden, J. Dordel, C. A. Yeats, A. Fedosejev, R. Goater, S. Castillo-Ramírez, J. Corander, C. Colijn, M. A. Chlebowicz, L. Schouls, M. Heck, G. Pluister, R. Ruimy, G. Kahlmeter, J. Åhman, E. Matuschek, A. W. Friedrich, J. Parkhill, S. D. Bentley, B. G. Spratt, H. Grundmann, European SRL Working Group, Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive *Staphylococcus aureus* in Europe, *MBio* **7** (2016), doi:10.1128/mBio.00444-16.
7. K. Zhou, M. Lokate, R. H. Deurenberg, M. Tepper, J. P. Arends, E. G. C. Raangs, J. Lo-Ten-Foe, H. Grundmann, J. W. A. Rossen, A. W. Friedrich, Use of whole-genome sequencing to trace, control and characterize the regional expansion of extended-spectrum β -lactamase producing ST15 *Klebsiella pneumoniae*, *Sci. Rep.* **6**, 20840 (2016).
8. S. Argimón, K. Abudahab, R. J. E. Goater, A. Fedosejev, J. Bhai, C. Glasner, E. J. Feil, M. T. G. Holden, C. A. Yeats, H. Grundmann, B. G. Spratt, D. M. Aanensen, Microreact: visualizing and sharing data for genomic epidemiology and phylogeography, *Microb Genom* **2**, e000093

(2016).

9. M. C. J. Maiden, M. J. Jansen van Rensburg, J. E. Bray, S. G. Earle, S. A. Ford, K. A. Jolley, N. D. McCarthy, MLST revisited: the gene-by-gene approach to bacterial genomics, *Nat. Rev. Microbiol.* **11**, 728–736 (2013).

10. M. Ciccolini, T. Donker, R. Köck, M. Mielke, R. Hendrix, A. Jurke, J. Rahamat-Langendoen, K. Becker, H. G. M. Niesters, H. Grundmann, A. W. Friedrich, Infection prevention in a connected world: the case for a regional approach, *Int. J. Med. Microbiol.* **303**, 380–387 (2013).

11. M. J. Ray, M. Y. Lin, R. A. Weinstein, W. E. Trick, Spread of Carbapenem-Resistant Enterobacteriaceae Among Illinois Healthcare Facilities: The Role of Patient Sharing, *Clin. Infect. Dis.* **63**, 889–893 (2016).

12. L. Senn, O. Clerc, G. Zanetti, P. Basset, G. Prod'homme, N. C. Gordon, A. E. Sheppard, D. W. Crook, R. James, H. A. Thorpe, E. J. Feil, D. S. Blanc, The Stealthy Superbug: the Role of Asymptomatic Enteric Carriage in Maintaining a Long-Term Hospital Outbreak of ST228 Methicillin-Resistant *Staphylococcus aureus*, *MBio* **7**, e02039–15 (2016).

13. F. Coll, E. M. Harrison, M. S. Toleman, S. Reuter, K. E. Raven, B. Blane, B. Palmer, A. R. M. Kappeler, N. M. Brown, M. Estée Török, J. Parkhill, S. J. Peacock, Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community, *Sci. Transl. Med.* **9**, eaak9745 (2017).

14. T. Donker, S. Reuter, J. Scriberras, R. Reynolds, N. M. Brown, M. E. Török, R. James, E. O. E. M. R. Network, D. M. Aanensen, S. D. Bentley, M. T. G. Holden, J. Parkhill, B. G. Spratt, S. J. Peacock, E. J. Feil, H. Grundmann, Population genetic structuring of methicillin-resistant *Staphylococcus aureus* clone EMRSA-15 within UK reflects patient referral patterns, *Microb Genom* **3**, e000113 (2017).

15. A. C. Schürch, W. van Schaik, Challenges and opportunities for whole-genome sequencing-based surveillance of antibiotic resistance, *Ann. N. Y. Acad. Sci.* **1388**, 108–120 (2017).

16. P. Bradley, N. C. Gordon, T. M. Walker, L. Dunn, S. Heys, B. Huang, S. Earle, L. J. Pankhurst, L. Anson, M. de Cesare, P. Piazza, A. A. Votintseva, T. Golubchik, D. J. Wilson, D. H. Wyllie, R. Diel, S. Niemann, S. Feuerriegel, T. A. Kohl, N. Ismail, S. V. Omar, E. G. Smith, D. Buck, G. McVean, A. S. Walker, T. E. A. Peto, D. W. Crook, Z. Iqbal, Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*, *Nat. Commun.* **6**, 10063 (2015).

17. M. Laabei, R. Massey, Using functional genomics to decipher the complexity of microbial pathogenicity, *Curr. Genet.* **62**, 523–525 (2016).

18. T. D. Read, R. C. Massey, Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology, *Genome Med.* **6**, 109 (2014).

19. O. Tosas Auguet, R. A. Stabler, J. Betley, M. D. Preston, M. Dhaliwal, M. Gaunt, A. Ioannou, N. Desai, T. Karadag, R. Batra, J. A. Otter, H. Marbach, T. G. Clark, J. D. Edgeworth,

Frequent undetected MRSA ward-based transmission linked to patient sharing between hospitals, *Clin. Infect. Dis.* (2017), doi:10.1093/cid/cix901.

20. L.-Y. Hsu, S. R. Harris, M. A. Chlebowicz, J. A. Lindsay, T.-H. Koh, P. Krishnan, T.-Y. Tan, P.-Y. Hon, W. B. Grubb, S. D. Bentley, J. Parkhill, S. J. Peacock, M. T. G. Holden, Evolutionary dynamics of methicillin-resistant *Staphylococcus aureus* within a healthcare system, *Genome Biol.* **16**, 81 (2015).

21. P. Glaser, P. Martins-Simões, A. Villain, M. Barbier, A. Tristan, C. Bouchier, L. Ma, M. Bes, F. Laurent, D. Guillemot, T. Wirth, F. Vandenesch, Demography and Intercontinental Spread of the USA300 Community-Acquired Methicillin-Resistant *Staphylococcus aureus* Lineage, *MBio* **7**, e02183–15 (2016).

22. H.-H. Chang, J. Dordel, T. Donker, C. J. Worby, E. J. Feil, W. P. Hanage, S. D. Bentley, S. S. Huang, M. Lipsitch, Identifying the effect of patient sharing on between-hospital genetic differentiation of methicillin-resistant *Staphylococcus aureus*, *Genome Med.* **8**, 18 (2016).

23. V. Conte, M. Monaco, T. Giani, F. D’Ancona, M. L. Moro, F. Arena, M. M. D’Andrea, G. M. Rossolini, A. Pantosti, AR-ISS Study Group on Carbapenemase-Producing *K. pneumoniae*, Molecular epidemiology of KPC-producing *Klebsiella pneumoniae* from invasive infections in Italy: increasing diversity with predominance of the ST512 clade II sublineage, *J. Antimicrob. Chemother.* **71**, 3386–3391 (2016).

24. S. Gaiarsa, F. Comandatore, P. Gaibani, M. Corbella, C. Dalla Valle, S. Epis, E. Scaltriti, E. Carretto, C. Farina, M. Labonia, Others, Genomic epidemiology of *Klebsiella pneumoniae* in Italy and novel insights into the origin and global evolution of its resistance to carbapenem antibiotics, *Antimicrob. Agents Chemother.* **59**, 389–396 (2015).

25. S. A. Gomez, F. G. Pasteran, D. Faccone, N. Tijet, M. Rapoport, C. Lucero, O. Lastovetska, E. Albornoz, M. Galas, KPC Group, R. G. Melano, A. Corso, A. Petroni, Clonal dissemination of *Klebsiella pneumoniae* ST258 harbouring KPC-2 in Argentina, *Clin. Microbiol. Infect.* **17**, 1520–1524 (2011).

26. A. Gomez-Simmonds, A.-C. Uhlemann, Clinical Implications of Genomic Adaptation and Evolution of Carbapenem-Resistant *Klebsiella pneumoniae*, *J. Infect. Dis.* **215**, S18–S27 (2017).

27. J. Corander, Y. Cui, T. Koski, J. Sirén, Have I seen you before? Principles of Bayesian predictive classification revisited, *Stat. Comput.* **23**, 59–73 (2013).

28. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, *Bayesian data analysis* (CRC press Boca Raton, FL, 2014).

Acknowledgements: The ST22 data was generated as part of the UKCRC Translational Infection Research Initiative and the Medical Research Council (grant no. G1000803) with contributions to the grant from the Biotechnology and Biological Sciences Research Council (BBSRC), the National Institute for Health Research on behalf of the Department of Health, and the Chief Scientist Office of the Scottish Government Health Directorate (to Sharon Peacock). EJP is supported by the JPI-AMR project SpARK with funding contributions from the Medical Research Council (MR/R00241X/1), and the EPSRC (GCRF) project ReNEW (EP/P028403/1). MA received funding from the “STARS” programme, supported by the BBSRC, via the University of Bath Institute for Mathematical Innovation Summer Internship Scheme. JS was supported by a University of Bath studentship. TD is affiliated to the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England (PHE). Tjibbe Donker is based at University of Oxford. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England. Members of the East of England Microbiology Research Network are Benny Cherian (Basildon and Thurrock University Hospitals, Basildon, UK), Tony Elston (Colchester Hospital University, Colchester, UK), Richard Kent (Ipswich Hospital, Ipswich, UK), John Hayward (James Paget University Hospital, Great Yarmouth, UK), Louise Teare (Mid Essex Hospital Services, Chelmsford, UK), Juliet Foweraker (Papworth Hospital, Papworth Everard, UK), David Enoch (Peterborough and Stamford Hospitals, Peterborough, UK), Nada Elhag (Southend University Hospital, Westcliff-on-Sea, UK), Prema Singh (West Hertfordshire Hospitals, Watford, UK), and Rebecca Tilley (West Suffolk Hospital, Bury St Edmunds, UK).

Author Contributions: EJP, RJ Designed the study and wrote the manuscript. MA, JS wrote the scripts, carried out the analyses and contributed to the design of the study. JC, TD contributed to the statistical methodology and SB carried out the phylogenetic analysis.

Ethical Statement: Written informed consent from patients was not required as all bacterial isolates were collected, processed and stored as part of routine clinical care and/or bacteraemia surveillance programmes. The study protocol was approved by the National Research Ethics Service (reference 11/EE/0499), and by the Cambridge University Hospitals NHS Foundation Trust Research and Development Department (reference A092428)

Competing Interests: We have no competing interests to declare.

We are very grateful to Nick Priest for many helpful discussion, and to Matthew Holden for excellent feedback received as part of his examination of the PhD thesis of JS. We are also grateful for helpful discussions on this work at the “Permafrost” workshop 2018. All scripts are readily available at <https://github.com/marianne-asbury/hospital-origin>.

Figures

Figure 1: ML tree of all 1022 isolates. The colours of the inner ring and tree branches correspond to referral cluster. The colour of the outer ring corresponds to hospitals. For the positions of the referral clusters in England see Figure 4.

Figure 2 Summary of results: The consistency between the two methods, and the assignments where possible of intra-versus (A) versus inter- (B) hospital transmission for all 91 cases is summarised.

Figure 3 Differing scenarios that may result in inconsistencies between the methods. The position of the query is marked by an asterisk, and A and B represent two different hospitals. In case 1, both the heuristic and Bayesian methods will tend to predict hospital as the origin of the query isolate. In case 2, if hospital B is poorly sampled and unrelated genotypes are co-circulating in A, then the Bayesian method will predict B whilst the heuristic will predict A. Similarly, in case 3 the heuristic is likely to predict B, but the Bayesian method may predict A if the sample from B contains multiple genotypes. In case 4 both methods will tend to predict B.

Figure 4: Position of the referral regions in England (A), breakdown per year in the database (B) and number of isolates in each hospital by referral region (C). The colour codes for the hospitals are given in Figure 1. Colour codes for hospitals and referral regions are used for the supplemental figures.

List of supplementary material:

Supplementary Figures S1 – S8 comparative examples of the outputs from both statistical methods (see text and Figure 2).

Supplementary Figure S9 examining the impact of changing the temporal breadth of the database.

Supplementary Figures S10-S13 preliminary analyses of the data (see Methods).

A

B

C

Figure S1. The query isolate id (in bold) is 7565_8.25 was isolated from hospital WalesB (RC Wales). Inspection of the Tree (A) confirms that it falls within a clade of other isolates from the same hospital. This hospital was also the Predicted source using both the heuristic (B) and Bayesian (C) methods. RC = Referral cluster. H = Hospital. For a key to the colours see Figure 1.

Figure S2. The query isolate id (in bold) is 7564_8.74 was isolated from hospital NorthcentralA. Inspection of the Tree (A) confirms that it falls within a clade of other isolates from the same hospital, but is also closely related to isolate 7748_6.64 which originated from hospital East2G (presumably due to an inter-hospital transmission). Despite this, both the heuristic (B) and Bayesian (C) methods predicted NorthcentralA as the source of the query isolate. RC = Referral cluster. H = Hospital.

A

B

C

Figure S3. The query isolate id (in bold) is 7748_6.70 was isolated from hospital NorthernIrelandC. Inspection of the Tree (A) confirms that it is most closely related to isolate 7748_6.69 which was also isolated from hospital NorthernIreland C, but is not embedded in a large clade of isolates from this hospital. Both the heuristic (B) and Bayesian (C) methods predicted NorthernIrelandC as the source of the query isolate.

A

B

C

Figure S4. The query isolate id (in bold) is 7748_6.70 was isolated from hospital NorthCentralB. However, inspection of the Tree (A) confirms that it corresponds to a clade of isolates from Northern Ireland, indicating an inter-hospital transmission Event between referral clusters.

Both the heuristic (B) and Bayesian (C) methods predicted NorthernIrelandA as the source of the query isolate, which is Consistent with inter-hospital transmission. RC = Referral cluster. H = Hospital.

For a key to the colours see Figure 1.

Figure S5. The query isolate id (in bold) is 7083_1.38 was isolated from hospital East2G. However, inspection of the Tree (A) confirms that it corresponds to a clade of isolates from East2D, indicating an inter-hospital transmission event between hospitals belonging to the same referral cluster. Both the heuristic (B) and Bayesian (C) methods predicted East2D as the source of the query isolate, which is consistent with inter-hospital transmission. RC = Referral cluster. H = Hospital. For a key to the colours see Figure 1.

Figure S6. The query isolate id (in bold) is 7748_6.72 was isolated from hospital East1A. However, inspection of the Tree (A) does not provide a strong indication of the source of this isolate. The heuristic (B) predicts the origin as SouthEastC whilst the Bayesian (C) method predicted East1A, which is the actual origin of the isolate. This hospital was favoured in the Bayesian method because of only one other isolate from East1A was present in the database. RC = Referral cluster. H = Hospital. For a key to the colours see Figure 1.

Figure S7. 7564_8.79 was isolated from hospital NorthernIrelandB. As this was the earliest isolate recovered from this hospital It was not represented in the database, the heuristic was unable to predict it. However, the Bayseian hospital predicted this hospital due to the prior. RC = Referral cluster. H = Hospital.
For a key to the colours see Figure 1.

Figure S8. The query isolate id (in bold) is 7083_1.39 was isolated from hospital East2G and the tree (A) shows that it is Related to other isolates from this hospital, pointing to intra-hospital transmission. The heuristic (B) also predicts the origin as East2G. However, the the Bayesian (C) method predicted East1F on the basis that the only isolate corresponding to this hospital in the reference database was reasonably closely related to the query, an that multiple genotypes were present within the well-sample hospital East2G. RC = Referral cluster. H = Hospital. For a key to the colours see Figure 1.

Figure S9. Examining the impact of the temporal breadth of the database. At the hospital level, the percentage of predictions consistent with the actual origin increases slightly at the hospital level as older isolates are removed, but then drops after 2008 presumably due to a paucity of data in the reference database. At the RC level the removal of older isolates has no obvious impact on the Bayesian method, but decreases the consistency of predictions with the actual origin using the heuristic method. Regarding the mean score of the predicted origin, there is a slight increase at the hospital level as older isolates are removed for both methods, but this is more marked when only the later isolates remain for the Bayesian method. There is much more marked increase in the strength of signal at the RC level for the Bayesian method, whilst there is no clear effect using the heuristic method at this level.

Figure S10. The frequency distribution of SNPs. Data for the very few SNPs found in >20 are not shown

Figure S11. Occurrence of non-singleton SNPs (5469 positions/rows) across the 1022 isolates in the 2001-2010 database. At each position the minority nucleotides are black, and majority nucleotides are grey.

Figure S12. Each point represents a single SNP, and the figure shows, for each SNP, the number of isolates in which that SNP is observed against the number of hospitals. The red points are the observed data, the blue points are expected if the isolates were randomly distributed between hospitals (over 1000 simulations). This shows that SNPs are present in far fewer hospitals than expected, given their frequency in the database.

Figure S13. Plot showing relationship between the date at which SNPs are first observed, and their frequency in the database. More recently emerged SNPs are present at a lower frequency and in fewer hospitals, thus they have the greatest discriminatory signal with respect to hospital source.