# **Supplementary Material for** *"Estimating the Attributable Disease Burden and Effects of Inter-Hospital Patient Sharing on Clostridium difficile Infections"*

## Contents


**1 Network autocorrelation model** **2**
- 1.1 Network autocorrelation model . . . 2
- 1.2 Attributable proportion of CDIs . . . 3

**2 Sensitivity Analyses** **5**
- 2.1 Including In-degree . . . 5
- 2.2 Principal diagnoses . . . 5
- 2.3 Internal consistency . . . 5
- 2.4 Change in diagnostic procedures . . . 6

**3 Additional information on data** **7**
- 3.1 Variation in covariates . . . 7
- 3.2 Temporal correlations of patient sharing . . . 7

# 1 Network autocorrelation model

Consider the context where one has a response variable of interest measured on a set of units, a set of covariates that help explain the variation in the response, and further the response of one unit may impact the response of other units to which it is connected. The set of connections and the units being connected form a network, and this network induces dependencies between the response variable measured on the units. Newcomb (1951) describes the network effect in the conext of a social network thus:

> Any observable behavior [e.g., a displayed position on an issue] is not only a response (on the part of a subject) which is to be treated as a dependent variable; it is also a stimulus to be perceived by others with whom the subject interacts, and thus to be treated as an independent variable.

In the context of CDI cases and patient sharing, the number of CDIs (response) in one hospital may affect the number of CDIs at hospitals to which it sends patients. Patient sharing will of course not be the only driver of a hospital's CDI rate, and hence there are salient covariates such as patient population characteristics that must be considered.

## 1.1 Network autocorrelation model

### Notation

The relevant quantities derived from the data are as follows. Let $n$ denote the number of hospitals, and $\boldsymbol{y}$ the $n \times 1$ response vector of monthly average number of CDI cases. Let $X$ denote the $n \times p$ matrix of covariates, where in our context $p = 5$: intercept, monthly average number of admissions, median length of stay, median number of diagnoses per inpatient visit, and proportion of patients over 65. Let $A$ denote the $n \times n$ adjacency matrix capturing the patient transfer network, where the $(i, j)^{th}$ entry $A_{ij}$ is the average number of patients transferred from hospital $j$ to hospital $i$ (note that it is often conventional to denote this quantity instead by $A_{ji}$). Finally, let $D$ denote the $n \times n$ diagonal matrix such that the $i^{th}$ diagonal entry is the monthly average number of admissions.

The parameters of the network autocorrelation model are as follows. Let $\boldsymbol{\beta}$ denote the $p \times 1$ vector of regression coefficients corresponding to the covariates $X$. Let $\rho$ denote the network contagion effect, which dictates how the network affects the response. Finally let $\sigma^2$ denote the residual variance.

### Statistical Model

The network autocorrelation model we implemented posits that the number of CDI cases for hospital $i$ depends on the number of patients transferring into $i$ in conjunction with the CDI rates of the hospitals of origin for those transfer patients. Specifically, the statistical model can be written as

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + \rho \sum_{j \neq i} A_{ij} \cdot \frac{\boldsymbol{y}_j}{D_{jj}} + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \overset{iid}{\sim} N(0, \sigma^2), \qquad i = 1, 2, \ldots, n \tag{1}$$

or in matrix notation

$$\boldsymbol{y} = X\boldsymbol{\beta} + \rho AD^{-1}\boldsymbol{y} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \overset{iid}{\sim} N(\mathbf{0}, \sigma^2 I_n). \tag{2}$$

Equivalently, we can consider the joint distribution of the response vector $\boldsymbol{y}$ as

$$\boldsymbol{y} \sim N\Big(M^{-1}X\boldsymbol{\beta}, \sigma^2 M^{-1}(M')^{-1}\Big) \tag{3}$$
$$\text{where } M := (I - \rho AD^{-1}).$$

This joint distribution provides explicitly the likelihood which can then be maximized using standard numerical optimization techniques in order to find the maximum likelihood estimators (MLEs). Inference can then be made using standard likelihood theory. Estimation was done using the sna package in R (Butts, 2016).

## Interpretation

To understand the network effect, one must better understand the key quantity $\rho$, as this coefficient corresponds to how the network affects the response variable of interest. First note that the CDI rate of hospital $j$, which we will denote as $r_j$ ($0 \leq r_j \leq 1$), equals $\boldsymbol{y}_j/D_{jj}$. This is the quantity being summed in (1) over those hospitals sending patients to hospital $i$. More specifically, for each hospital $j$ sending patients to hospital $i$, we are adding to the $i^{th}$ response the quantity $\rho$ times $A_{ij}r_j$, or in other words, the CDI rate of hospital $j$ times the number of patients (on average) being sent from $j$ to $i$ all multiplied by $\rho$. Hence the monthly average number of CDI cases at hospital $i$ increases as (i) the number of patients transferred in increases, (ii) the contamination of those hospitals which are sending $i$ patients increases, and (iii) $\rho$ increases. Also note that if hospital $j$ does not send any patients to $i$ and/or if $j$ does not have any CDIs (i.e., $r_j = 0$), then $j$ will not increase the mean number of CDI cases at hospital $i$. A direct interpretation of $\rho$ comes from noting that for a hospital $j$ with CDI rate $r_j$, every $1/r_j$ patients transferred out will lead to a subsequent increase of $\rho$ in number of CDIs at other hospitals. That is,

$$\rho A_{ij} r_j = \rho \frac{1}{r_j} r_j = \rho.$$

For example, if a hospital with CDI rate equal to 0.01 then transfers 100 patients to hospital $i$, we would expect the average monthly number of CDIs at $i$ to increase by $\rho$.

## 1.2 Attributable proportion of CDIs

The network autocorrelation model provides a natural counterfactual framework to estimate the proportion of CDIs due to the network effect. Here we focus on the sum of the expected monthly average number of CDI cases, i.e., the expected monthly average number of CDI cases across the state of California. Specifically, we looked at the ratio given by

$$R := \frac{\mathbb{E}(\mathbb{1}'\boldsymbol{y}|\rho = 0)}{\mathbb{E}(\mathbb{1}'\boldsymbol{y}|\rho \neq 0)} = \frac{\mathbb{1}'X\boldsymbol{\beta}}{\mathbb{1}'M^{-1}X\boldsymbol{\beta}}, \tag{4}$$

where $\mathbb{1}$ denotes the $n \times 1$ vector of 1's. The MLE of this quantity $\widehat{R}$ is trivial to obtain by plugging in the MLEs of $\boldsymbol{\beta}$ and $\rho$. Inference can be obtained using the delta method to find the asymptotic normal distribution of the estimator $\widehat{R}$. Specifically $\widehat{R}$ asymptotically follows a normal distribution given by

$$\begin{aligned} \widehat{R} &\sim N\big(R, \nabla R \Sigma \nabla R'\big), \\ \nabla R &:= \begin{pmatrix} \frac{(\mathbb{1}'M^{-1}X\boldsymbol{\beta})\mathbb{1}'X - (\mathbb{1}'X\boldsymbol{\beta})\mathbb{1}'M^{-1}X}{(\mathbb{1}'M^{-1}X\boldsymbol{\beta})^2} & -\frac{(\mathbb{1}'X\boldsymbol{\beta})\mathbb{1}'M^{-1}AD^{-1}M^{-1}X\boldsymbol{\beta}}{(\mathbb{1}'M^{-1}X\boldsymbol{\beta})^2} \end{pmatrix} \end{aligned} \tag{5}$$

where $\Sigma$ is the covariance matrix of the MLEs of $\begin{pmatrix} \boldsymbol{\beta}' & \rho \end{pmatrix}$.

# 2 Sensitivity Analyses

## 2.1 Including In-degree

As a check, we reran the model with in-degree as a covariate to ensure that our network term in the NAM was in fact estimating a contagion effect rather than acting as a proxy for in-degree. In-degree was statistically insignificant (p-value: 0.59) while the network effect maintained statistical significance (p-value: 0.028). As an example of a contrast to this, we performed the same analysis as above but using monthly average number of acute myocardial infarctions, a non-infectious condition. Without in-degree included, the network contagion effect was significant (p-value: $< 0.001$), but when in-degree was included, in-degree was significant (p-value: $< 0.001$) while the contagion effect no longer was significant (p-value: 0.81). In fact, the sign switched on this term with the inclusion of in-degree, serving as a warning to researchers implementing our proposed approach to ensure their NAM is not serving as a proxy for in-degree.

## 2.2 Principal diagnoses

A principal diagnosis of CDI is nearly always present on admission, and hence is not a hospital acquired infection (HAI). We anticipated that the patient transfers only affected the number of HAIs, and hence there should not be a statistically significant relationship between the patient transfer network and the monthly average number of principal CDIs at a hospital. Our data allowed us to disambiguate between principal and secondary diagnoses of CDI, and hence we ran a sensitivity analysis replicating that performed for the main manuscript, except replacing the response variable with that constructed from principal CDI diagnoses. This analysis failed to reject the null hypothesis that there was no network effect (p-value: 0.29).

## 2.3 Internal consistency

To see if our results were internally consistent, we split the data geographically into what is typically considered Northern California (NCA) and Southern California (SCA) and reran the analyses. To make this partition we used the latitude line 35°47′28”, which divides our healthcare facilities into groups of 163 (NCA) and 222 (SCA). Despite losing power by halving the number of healthcare facilities in each of the two subsamples, we nevertheless obtained very similar and statistically significant estimates of the network effect. These estimates along with 1-sided p-values are provided below.

| | $n$ | $\hat{\rho}$ | p-value |
|---|---|---|---|
| Full data | 385 | 4.99 | $< 0.001$ |
| Northern CA | 163 | 4.14 | $< 0.001$ |
| Sourthern CA | 222 | 5.41 | $< 0.001$ |

## 2.4 Change in diagnostic procedures

During the period of this study, new diagnostic procedures and technologies such as the change from toxin to molecular tests potentially led to different rates of CDI (see, e.g., Polage et al., 2015). In order to test the sensitivity of our results to such a potential issue, we performed a simple changepoint analysis for every hospital. This changepoint analysis was done by assuming the monthly CDI counts followed a Poisson distribution with different rates on the two sides of the changepoint. We then tried all months ranging from 5 to 78 as change points and used BIC to select the best changepoint. We also compared the changepoint model to the model with no changepoint. 52 of the 385 hospitals did not have a detectable changepoint; Figure 1 shows graphically the changepoints of those hospitals which had a detectable changepoint. We then reran the NAM, where the response variable for each hospital was constructed by averaging over only the months prior to that hospital's changepoint. We then similarly reran the NAM a third time except using the months following the changepoints. The three NAMs (aggregated over time, pre-changepoint, post-changepoint) yielded results very similar to each other with equivalent conclusions regarding the network effect. The results are provided in Tables 1 and 2.

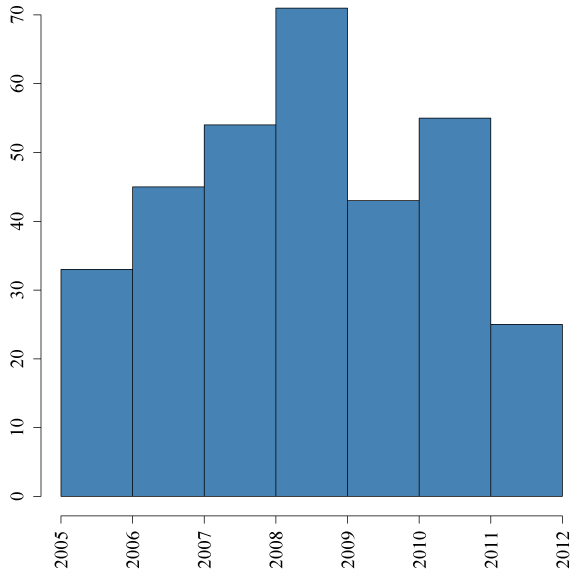


Figure 1: Estimated changepoints in CDI cases for the 333 hospitals with a detectable changepoint.

| | Est | LB | UB | p-value |
|---|---|---|---|---|
| Intercept | -3.5 | -5.6 | -1.4 | 0.00 |
| Total admissions | 0.0052 | 0.0048 | 0.0057 | $< 0.001$ |
| median LOS | -0.051 | -0.12 | 0.021 | 0.17 |
| median # dx | 0.42 | 0.25 | 0.60 | $< 0.001$ |
| % over 65 | 0.016 | -0.0051 | 0.037 | 0.14 |
| Geospatial effect | · | · | · | 0.74 |
| Network contagion effect | 5.0 | 2.7 | 7.3 | $< 0.001$ |
| % attributable to patient sharing | 12 | 6.4 | 18 | $< 0.001$ |

Table 1: Results from running a NAM using only the hospital specific pre-changepoint data.

| | Est | LB | UB | p-value |
|---|---|---|---|---|
| Intercept | -5.2 | -7.7 | -2.7 | $< 0.001$ |
| Total admissions | 0.0067 | 0.0062 | 0.0073 | $< 0.001$ |
| median LOS | -0.011 | -0.096 | 0.075 | 0.81 |
| median # dx | 0.54 | 0.33 | 0.74 | $< 0.001$ |
| % over 65 | 0.017 | -0.080 | 0.042 | 0.18 |
| Geospatial effect | · | · | · | 0.85 |
| Network contagion effect | 5.6 | 3.4 | 7.7 | $< 0.001$ |
| % attributable to patient sharing | 14 | 8.5 | 20 | $< 0.001$ |

Table 2: Results from running a NAM using only the hospital specific post-changepoint data.

# 3 Additional information on data

## 3.1 Variation in covariates

To better understand how hospitals' covariate values vary across months, we computed the within-hospital IQR, range, and standard deviation for the hospital-level covariates, i.e., the IQR, range, and sd for each hospital's time series corresponding to each covariate. The mean (median) values over the 385 hospitals are given in the table below.

| | IQR | Range | SD |
|---|---|---|---|
| Monthly total admissions | 107 (79) | 334 (263) | 74 (58) |
| Monthly median LOS | 0.968 (0.00) | 10.7 (1.00) | 1.66 (0.397) |
| Monthly median # dx | 1.50 (1.00) | 4.19 (3.00) | 1.06 (0.885) |
| Monthly % over 65 | 0.0543 (0.0391) | 0.188 (0.135) | 0.0410 (0.0288) |

## 3.2 Temporal correlations of patient sharing

Transfer rates are surprisingly steady across the 83 months of data. We computed the temporal product-moment correlations between the adjacency matrices for various lags. The boxplots of these correlations are provided below; for example, the left-most boxplot corresponds to the correlations between the 82 pairs of transfer networks one month apart, the

second left-most boxplot to the 81 pairs of transfer networks two months apart, etc. As can be seen, the correlations between the patient transfer networks are all very high, even a year apart. This provides support that it is reasonable to average across months to obtain a measure of connectivity through patient sharing among the hospitals.

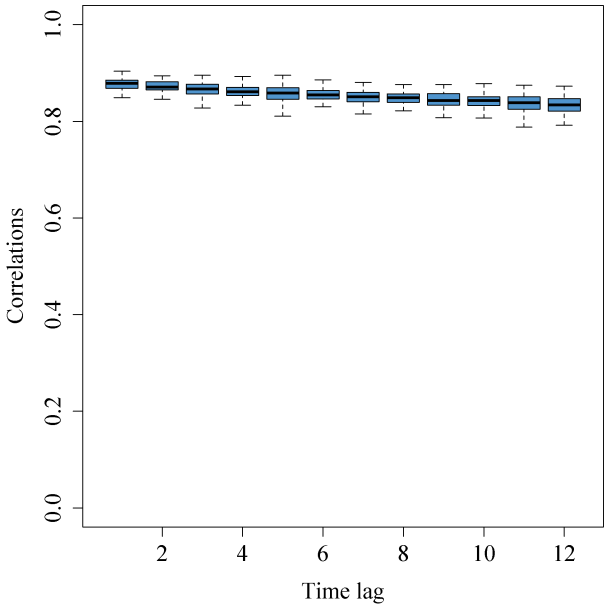


Figure 2: Temporal correlations between transfer networks at different lags.

# References

Butts, C. T. (2016). *sna: Tools for Social Network Analysis*. R package version 2.4.

Newcomb, T. M. (1951). Social psychological theory: Integrating individual and social approaches. In J. H. Rohrer and M. Sherif (Eds.), *Social Psychology at the Crossroads*. New York: Harper.

Polage, C. R., C. E. Gyorke, M. A. Kennedy, J. L. Leslie, D. L. Chin, S. Wang, H. H. Nguyen, B. Huang, Y.-W. Tang, L. W. Lee, K. Kim, S. Taylor, P. S. Romano, E. A. Panacek, P. B. Goodell, J. V. Solnick, and S. H. Cohen (2015, 11). Overdiagnosis of clostridium difficile infection in the molecular test era. *JAMA Internal Medicine 175*(11), 1792–1801.