SUPPORTING INFORMATION–Training the Model

When training a model for binary classification, the goal is typically to learn a function $f\ \cdot$ that can be used to predict the label for a new example. In our problem, the goal is to accurately predict a newly admitted patient's probability of testing positive for *C. difficile* during the current admission. Let $\mathbf{x}_{\mathrm{i}}$ represent the ith training example, (i.e., patient admission). $\mathbf{x}_{\mathrm{i}}$ is a d-dimensional feature vector that is in $\ 0,1\ ^{d}$. Let $\mathcal{X}$ represent the feature space that $\mathbf{x}_{\mathrm{i}}$ lies in. Let y_{i} represent a binary label indicating whether or not the ith patient tested positive for *C. difficile* during the current admission. Then our learning task is defined as follows: $\mathcal{D} = \quad \mathbf{x}_{\mathrm{i}}, \mathrm{y}_{\mathrm{i}} \quad \mathbf{x}_{\mathrm{i}} \in \mathcal{X}, \mathrm{y}_{\mathrm{i}} \in \ -1,1\ {}_{\mathrm{i}=1}^{n}$, where n is the number of unique patient admissions available for training. In general, with logistic regression, we seek a function $f\colon \mathbb{R}^{d} \rightarrow \ 0,1\ $ of the form:

$$f\ \boldsymbol{x}_{i}\ = \frac{1}{1 + e^{-\ b_{0} + \boldsymbol{w}^{T}\boldsymbol{x}_{i}}}$$

where $\boldsymbol{w} \in \mathbb{R}^{d}$ (and $\boldsymbol{x} \in \mathbb{R}^{d}$). Solving for the regression coefficients $\boldsymbol{w}$ and the offset b_{0} is a maximum likelihood estimation problem. When d is large, i.e., the data lie in a high-dimensional space, it is easy to overfit. Therefore, to improve generalizability to unseen future patient cases, and reduce the likelihood of overfitting to the training data, we employ L2-regularized logistic regression.[18] In L2-regularized logistic regression, a regularization term $\frac{1}{2}\quad \boldsymbol{w}\ ^{2}\ is$ included in the objective function.

$$\min_{\boldsymbol{w}} \frac{1}{2}\quad \boldsymbol{w}\ ^{2} +\ C\quad {}_{i=1}^{n} \log(1 + \exp^{-y_{i}\boldsymbol{w}^{T}\boldsymbol{x}_{i}}) \qquad \text{(Eq. 1)}$$

C is a scalar tuning parameter that controls the tradeoff between the number of errors on the training set and the complexity of the model. Note, we add an extra constant dimension to $\boldsymbol{w}$ and compute the offset b_0 implicitly. The solution to Eq. 1 depends on the $\{\ \mathbf{x}_{\mathrm{i}}, \mathrm{y}_{\mathrm{i}}\ \}_{\mathrm{i}=1}^{n}$ employed in the training. The training data is used in Eq. 1 to find the optimal setting of $\boldsymbol{w}$. The hyperparameter C in Eq. 1 was found using five-fold cross-validation on the training set, sweeping the value from 2^{-8} to 2^{-1}.

SUPPORTING INFORMATION–Calculating Colonization Pressure

In terms of *C. difficile*, colonization pressure (CP) aims to measure the number of infected patients, in a unit or hospital. In our analysis, the contribution a patient, *p*, makes to the CP on day, $CPP(t)$, depends on when the patient tested positive for the first and last time, t_f and t_l, and when the patient is discharged from the hospital t_d (where time is measured in days from the day of admission). While the patient continues to test positive he or she contributes a constant amount to the CP. After the last positive test result (which is often the first positive test result, since testing for a cure is not recommended) a patient contributes to the CP for no more than 14 days. During this time period, the patient is assumed to be treated or in isolation, and we assume a linearly decreasing relationship.

Equation 1 defines this function.

$$CPP\ \ p,t\ \ =\ \ \begin{matrix} 1 & t \in [t_f, t_l] \\ -\frac{t}{14}+\frac{(t_l+14)}{14} & t \in [t_l, \min\ \ t_d, t_l+14\ \ ] \\ 0 & otherwise \end{matrix} \qquad \text{Eq.1}$$

We have time-stamped locations for each patient, thus we calculate a colonization pressure for each unit, $CPU\ \ u,t$, as in Equation 2. The $CPU\ \ u,t$ depends on each patient’s contribution to the colonization pressure on that day and each patient’s length of stay in unit, u, on day t, $LOS(u,p,t)$

$$CPU\ \ u,t\ \ =\ \ {}_p CPP\ \ p,t\ \ *\frac{LOS(u,p,t)}{24} \qquad \text{Eq.}$$

2

When extracting the relevant **unit-wide colonization pressure** for a new patient on a given day we sum the $CPU\ \ u,t$ across all units in which that patient spent any time. As a result, the unit-wide colonization pressure varies across patients for a given day. The **hospital-wide colonization pressure** is calculated as $_{u}CPU(u,t)$, and is the same across all patients on a given day.