Minimizing HAI-Prevention Cost, with Probabilistic Guarantees

Center for Financial and Business Analytics
DATA SCIENCE INSTITUTE

Elioth Sanabria and David D. Yao
IEOR Dept, Columbia Engineering

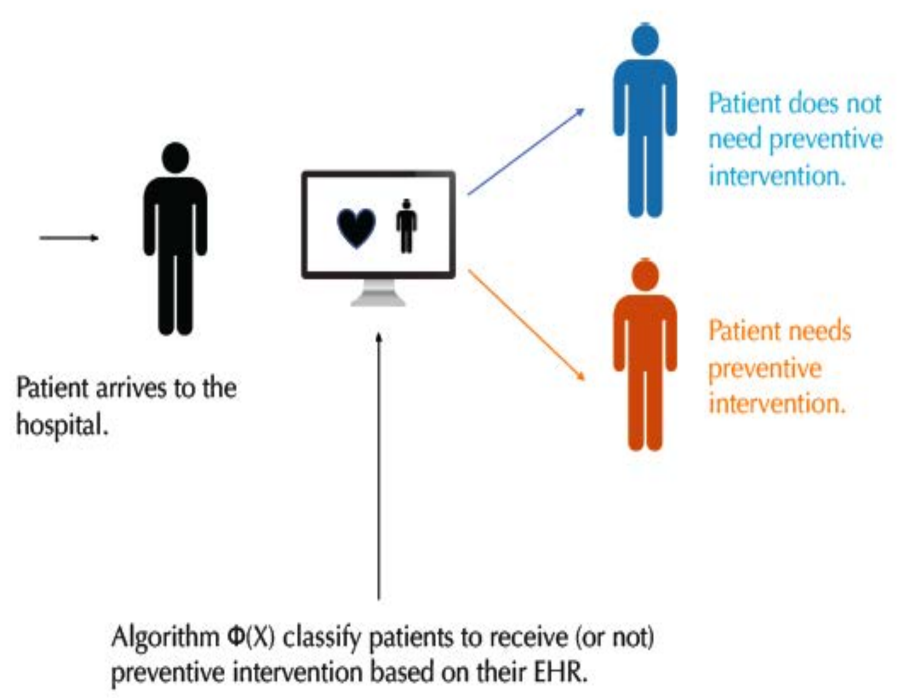
Data Science Institute
COLUMBIA UNIVERSITY

Healthcare-Associated Infection (HAI)

Healthcare associated infections (HAI) are estimated to cost US hospitals $9.8B per year. (Starting 2008, Medicare stopped reimbursing hospitals expenses due to HAI.) Costs for taking preventive measures are order-of-magnitude lower, but success rates are far from 100%; refer to the two tables below. We want to develop a machine learning (ML) classification scheme based on patient admission data only, as illustrated in the figure below.

sources of infection (non-surgical)	cost per case	infection rate
catheter-associated urinary tract (CU)	$896	1.15%
central line-associated bloodstream (CB)	$45,814	0.35%
clostridium difficile infection (CDiff)	$11,285	0.37%
ventilator-associated pneumonia (VP)	$40,144	0.25%

HAI	cost	succ rate	example preventive measures
CU	$16.5	69%	reduction in usage, timely removal
CB	$22.26	66%	sterile barrier precautions, chlorhexidine disinfection
CDiff	$115.62	80%	antibiotic stewardship, limit patient contact, extra care
VP	$215	46%	head of bed angle 130, daily interruption of sedation



A Jointly Optimized Classification-Prediction Scheme

The model involves two interlaced optimization problems. At the top is a cost minimization problem that explicitly accounts for the asymmetry between the cost of infection and the cost of prevention. The infection probabilities used in the cost model are solutions to a cross-entropy (CE) minimization problem that fits data with a suitable ML algorithm (e.g., logit regression, random forest, deep neutral network, etc). Here, the challenge is to deal with the intrinsic bias in data: infected cases are only around 1-2% of all patients. Our approach is to add a weighting (or "oversampling") coefficient to the CE objective and make it a decision variable too, in the same spirit as a Lagrangian multiplier.

Notation:

$Y = \{0,1\}$ with $\mathrm{P}(Y=1) := \pi$, proportion of infected patient;

$X := (X_1, \dots, X_\ell)$, basic admission data;

misclassification costs: K_0, K_1; $K_1 >> K_0$;

decision variables: ρ, K, and $w = (w_i)_{i=1}^{\ell}$ (more below);

ML/classification scheme: $\phi := (\rho, K, w)$.

$$\min_{\phi_n} \hat{\mathcal{C}}_n(\phi_n) := \frac{1}{n}\sum_{i=1}^{n}\left[K_0(1-y_i)\mathbf{1}\{p(x_i)\ge\rho\} + K_1 y_i \mathbf{1}\{p(x_i)<\rho\}\right],$$

$$\min_{(w,K)} \frac{1}{n}\sum_{i=1}^{n}\left[-(1-y_i)\log(1-p(x_i)) - K y_i \log p(x_i)\right];$$

where $p(x_i)$ can be from logit regression: $p(x_i) := 1/(1+e^{-wx_i})$,

or from DNN,

$$x^{(1)} = \psi(W^{(1)}x_i),\ x^{(d)} = \psi(W^{(d)}x^{(d-1)}),\ x^{(D)} = \psi(W^{(D)}x^{(D-1)}) := p(x_i).$$

Convergence and Rate of Convergence

ϕ^*: optimal solution to the original problem;

ϕ_n^*: optimal solution to the "data-driven" version;

$\mathrm{E}\mathcal{C}(\phi^*)$, $\hat{\mathcal{C}}_n(\phi_n^*)$: corresponding objective values.

Applying the Dvoretzky-Kiefer-Wolfowitz/Massart bound, we can derive

- $\hat{\mathcal{C}}_n(\phi_n^*) \to \mathrm{E}\mathcal{C}(\phi^*)$ a.s.; and $\mathrm{P}\left(|\hat{\mathcal{C}}_n(\phi_n^*) - \mathrm{E}\mathcal{C}(\phi^*)| > \epsilon\right) \le 4\exp\left(-\frac{n\epsilon^2}{2K_1^2}\right)$.
- $\mathcal{C}(\phi_n^*) \to \mathrm{E}\mathcal{C}(\phi^*)$ a.s.; and $\mathrm{P}\left(|\mathcal{C}(\phi_n^*) - \mathrm{E}\mathcal{C}(\phi^*)| > \epsilon\right) \le 4\exp\left(-\frac{n\epsilon^2}{8K_1^2}\right)$.

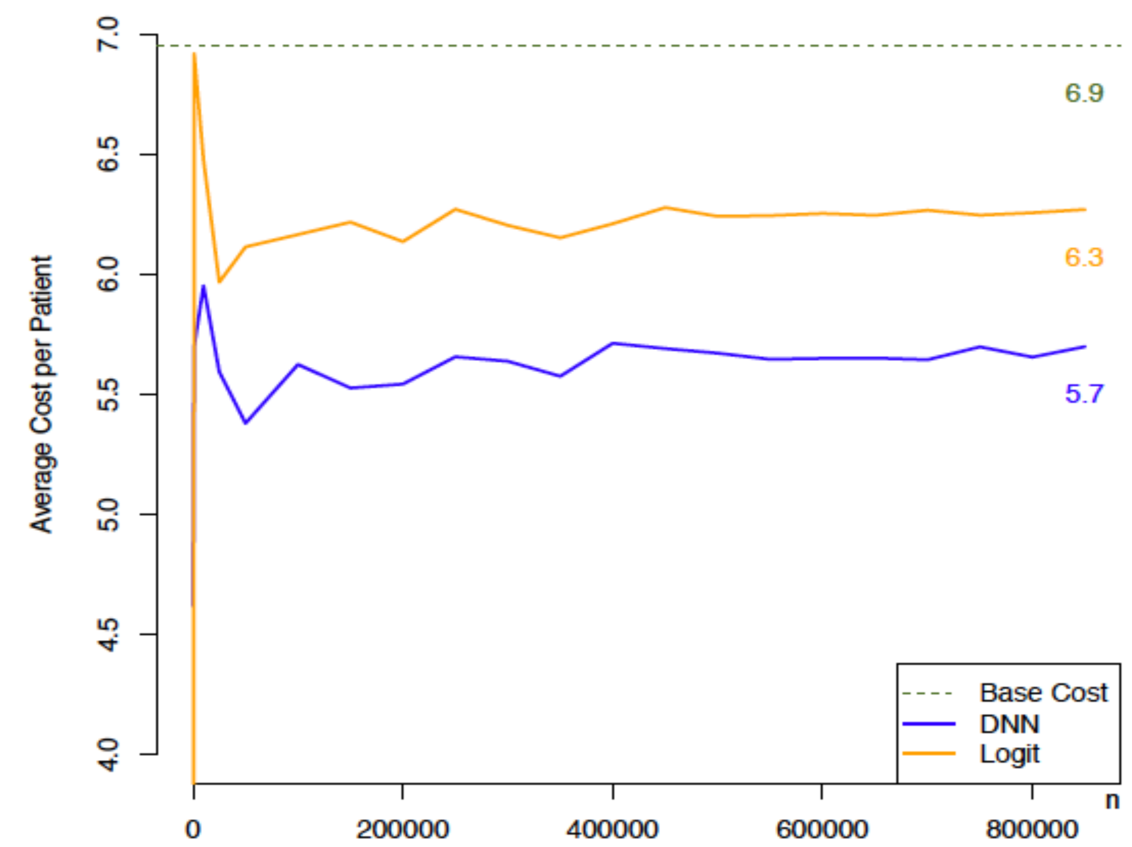
Suppose ϕ_n^* (for a given n, sufficiently large) is applied to another data set of size N, with the data being i.i.d. and following the same distribution as the original set, and denoting the corresponding cost as $\hat{\mathcal{C}}_N(\phi_n^*)$. Then,

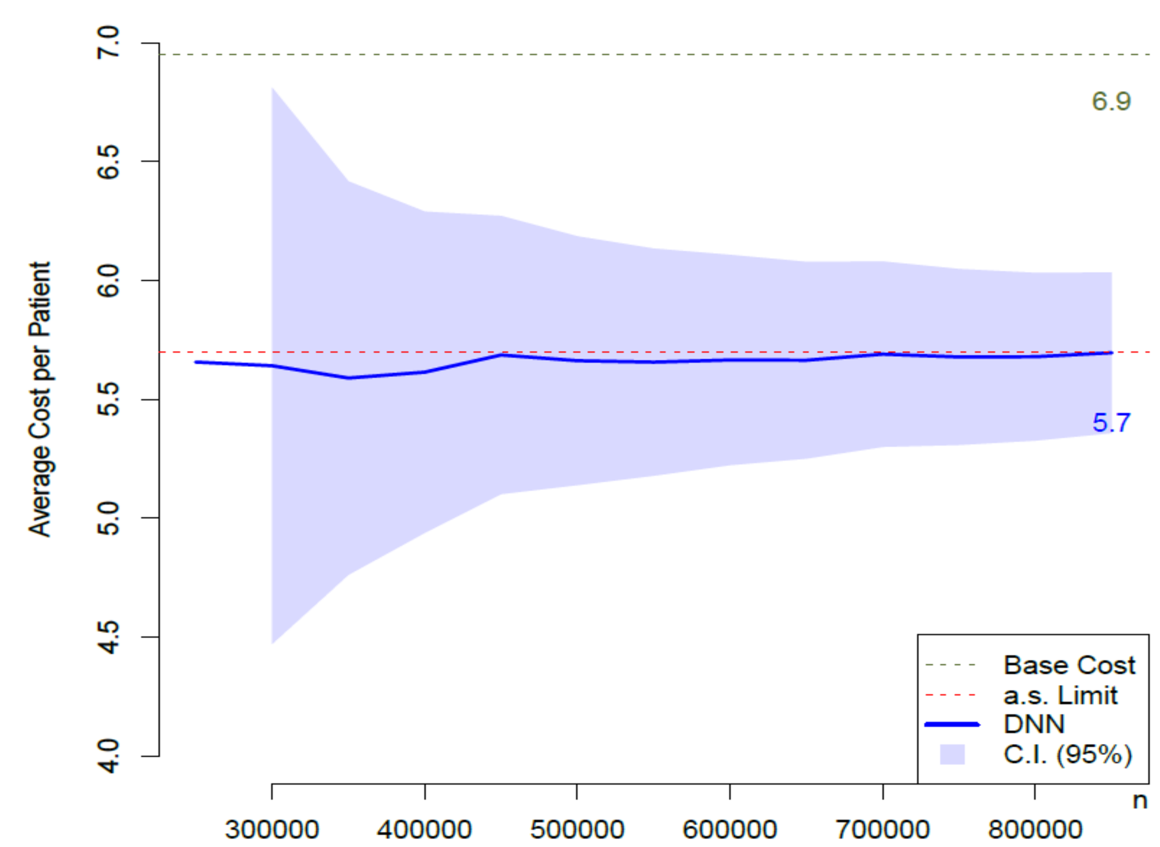
- $\hat{\mathcal{C}}_N(\phi_n^*) \xrightarrow{N\to\infty} \mathrm{E}\mathcal{C}(\phi^*)$ a.s.; and $\mathrm{P}\left(|\hat{\mathcal{C}}_N(\phi_n^*) - \mathrm{E}\mathcal{C}(\phi^*)| > \epsilon\right) \le \mathrm{P}\left(|Z| > \frac{\epsilon\sqrt{N}}{2\hat{\sigma}}\right)$,

where Z: standard normal, $\hat{\sigma}^2 := K_0^2(1-\pi) + K_1^2\pi$.

Cost Savings Achieved

As shown in the left figure below, logit regression achieves about 10% reduction, DNN does about 20%, from the base case (no ML). The right figure shows the confidence interval associated with the DNN performance.





Research Support and External Collaborations

Agency for Healthcare Research and Quality, AHRQ-R01-HS024915-01 (PI: Elaine Larson)
Columbia University Medical Center, School of Nursing