

**Engineering and Validating Predictive
Infection Surveillance Strategies for
Methicillin-resistant *Staphylococcus aureus***

(2008–09)

Stephen Trusheim
Epidemiology
123 Ottawa Ave N
Golden Valley MN 55422

Abstract

Methicillin-resistant *Staphylococcus aureus* (MRSA) is endemic at nearly every hospital in America, infecting an estimated 94,000 patients every year and directly killing more than 19,000. To prevent MRSA outbreaks, the Institute for Healthcare Improvement recommends that hospitals employ “universal surveillance,” which involves testing every admitted patient for MRSA; however, nearly all hospitals employ less-effective MRSA surveillance strategies, due to cost.

To reduce costs of MRSA surveillance while ensuring that all cases of MRSA are identified, I developed a new surveillance strategy that I termed “predictive surveillance.” I engineered four software components based on predictive healthcare methods and validated the components against data from prior studies. I first engineered a LogitBoost algorithm to accurately predict MRSA-colonized patients, then used those predictions to recommend specific MRSA tests for each patient. I next engineered software that generated accurate cost-estimates for every available MRSA surveillance strategy and, finally, engineered an interactive website that provided this software to hospitals, clinics, and the public.

Results suggest hospital cost-savings of more than 25% per year associated with use of a predictive surveillance strategy. This cost-savings will enable more hospitals to employ effective means of MRSA surveillance, preventing further spread of this deadly infectious disease.

Introduction

Methicillin-resistant *Staphylococcus aureus* (MRSA) has been described as “a major public health problem” by the United States Centers for Disease Control (1). MRSA is endemic at nearly every hospital in the United States, infecting an estimated 94,000 patients each year and directly killing more than a quarter of those infected (1,2). To prevent outbreaks of this highly contagious and deadly disease, the Institute for Healthcare Improvement has recommended that hospitals actively test patients for MRSA upon admission — a process known as MRSA surveillance (3-5). However, the commonly recommended method for MRSA surveillance, “universal surveillance,” significantly increases costs for hospital medical care; for this reason, only three hospitals in the United States report using universal surveillance (6). The goal of my study was to engineer software that reduces the cost of MRSA surveillance by accurately targeting at-risk patients for MRSA testing — a new method that I termed “predictive surveillance.” I additionally sought to engineer an interactive website enabling the public to determine their own MRSA risks and see MRSA trends, giving smaller clinics and offices the same predictive capability as larger hospitals.

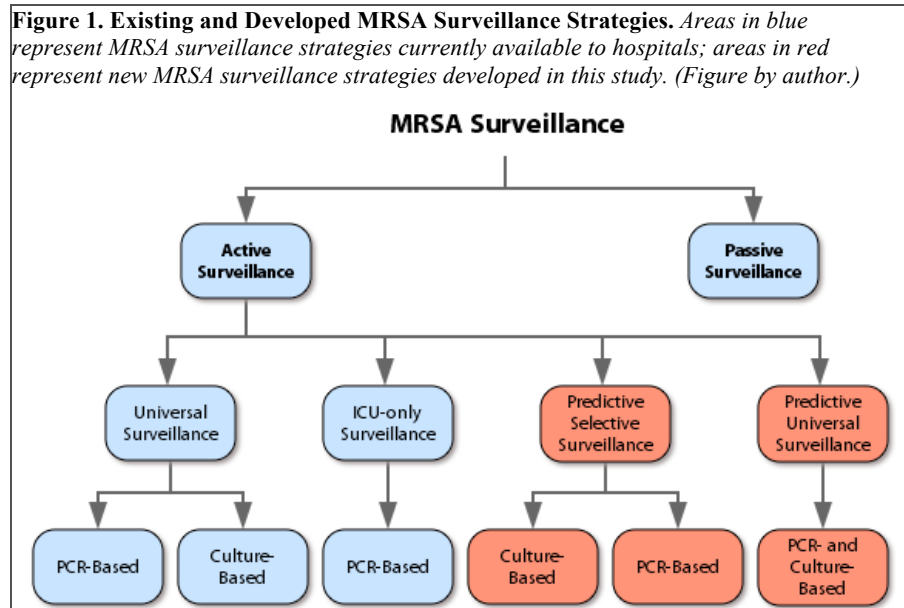
MRSA is a highly infectious, drug-resistant strain of the bacteria *Staphylococcus aureus* and is the leading cause of surgical-site infections and lower respiratory tract infections in the United States (7). In 2005 alone, the disease colonized more than an estimated 275,000 patients, causing a life-threatening infection in 94,000 and killing more than 19,000 (1,2,7). MRSA commonly colonizes nasal passages of otherwise healthy patients without causing significant infections (8,9). However, in the confines of a hospital, nasal MRSA infections are easily transmitted to open wounds and respiratory tracts of critically ill patients, whose immune systems cannot fight off infection (8,10). In fact, a recent study indicated that 85% of all MRSA cases are hospital-

acquired, and those commonly acquired strains are resistant to nearly all antibacterial medications (1,7).

The Institute for Healthcare Improvement has prioritized MRSA prevention in its “Five

Million Lives” campaign (5). Their 2006 plan agrees with other studies, recommending that hospitals implement universal surveillance as part of a five-step plan to decrease transmission of the disease in hospitals (3,5,11,12). While a recent study by Robicsek et al. (2008) reported that universal surveillance results in a 70% decrease in hospital-acquired MRSA infections, work by Peterson et al. (2007) estimated costs for universal surveillance at \$600,000 per year for a medium-sized hospital system (11,12). Many U.S. hospitals, for this reason, do not use any form of active MRSA surveillance, instead waiting to identify MRSA-colonized patients only after patients develop symptoms of the disease — a strategy known as “passive surveillance” (Figure 1).

Costs for MRSA surveillance depend mainly on two factors: the type of MRSA tests administered and the amount of time each patient spends in post-screening isolation (10). There are two types of FDA-approved MRSA screening tests: culture-based tests, such as Spectra MRSA and BD Diagnostics CHROMagar, and PCR-based tests, such as BD Diagnostics GeneOhm and Cepheid Xpert MRSA (13-17). Both types of tests are approved for clinical



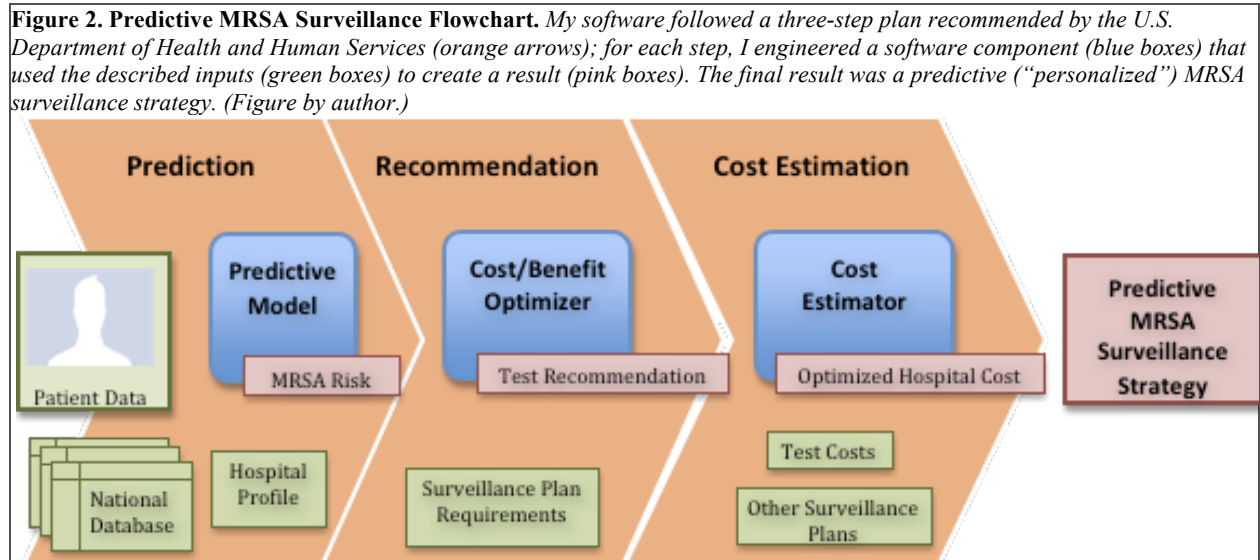
diagnosis but have varying costs and turn-around times. Culture-based tests cost between \$5 and \$15 per patient and give results within one to three days; PCR-based tests cost between \$50 and \$75 per patient and give results within three hours of screening (13-17). The slower culture-based tests are associated with higher isolation costs, because patients can spread the disease while waiting for test results. Costs for universal surveillance are particularly high because every admitted patient is typically screened with an expensive PCR-based test (5).

MRSA surveillance costs can be reduced by intelligently tailoring testing plans for each admitted patient — a practice known as “personalized healthcare.” To develop my software, I used three elements of personalized healthcare proposed in a 2007 report by the U.S. Department of Health and Human Services: prediction, recommendation, and cost reduction (18). I designed software to minimize hospital testing costs and reduce incorrect isolation time by enabling hospitals to personalize MRSA testing methods and isolation strategies for each patient during the admission process — a strategy I termed “predictive surveillance.” I designed and developed comprehensive web-based Java software enabling effective predictive surveillance, consisting of three components (Figure 2):

1. *Predictive Model*: This component generated a predictive model based on machine learning technology. The model analyzed patient risk profiles to predict patients who are likely to be colonized by MRSA.
2. *Testing Strategy Optimizer*: This component used the *Predictive Model* to determine the optimal MRSA test for each patient, optimizing current MRSA surveillance strategies.
3. *Hospital Cost Estimator*: This component estimated total hospital costs for MRSA surveillance to determine overall benefits of the surveillance strategies generated by the *Testing Strategy Optimizer*.

In addition, I engineered two website components to create a public MRSA website:

1. *Interactive Risk Analysis*: This component uses my *Predictive Model* to interactively predict a website user's risk for MRSA.
2. *Public MRSA Reports*: This component constantly generates up-to-date public reports about MRSA, using my *Predictive Model* and *Hospital Cost Estimator*.



Materials: Patient Data

To develop my initial model, I used patient data collected in a 2008 study by Robicsek et al., conducted at Evanston Northwestern Hospital in Chicago, IL (11). These data consisted of patient profiles that included the following information for a cohort of 48,203 patients: demographics, diagnosis codes, and care information (details are provided in Appendix A). Peterson et al. studied this same cohort to determine surveillance costs (12).

Methodology and Results

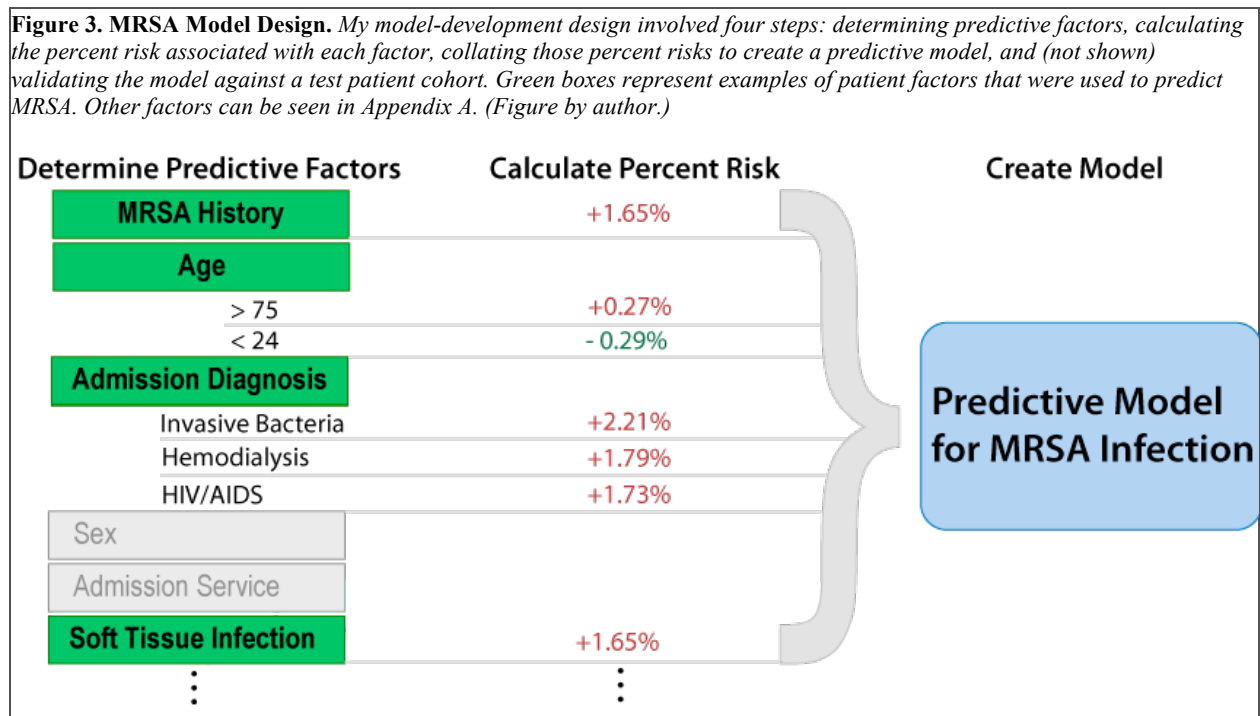
Prediction: Designing a Predictive Model

Data Collection: I began by designing a relational database to contain the Robicsek patient profiles and then engineered a data-transfer algorithm that mapped spreadsheet columns to database columns and converted textual spreadsheet data into binary formats. While I used the

Robicsek cohort to develop my software, I designed my software to analyze continuously uploaded data from multiple national databases.

Design: I next developed a design for my *Predictive Model*, involving four steps:

1. Determining risk factors that best predict MRSA infection (green boxes in Figure 3);
2. Calculating percent risk associated with each factor (percents in Figure 3);
3. Collating calculated percent risks to develop the *Predictive Model*; and
4. Validating accuracy of the *Predictive Model*.



Prototype 1: I engineered an algorithm that fit my design using iteratively reweighted least-squares logistic regression (IRLS regression), as described in papers by Derr et al., Bewick et al. (2005) and Movellan et al. (2006) (19-21). The IRLS logistic regression algorithm developed a predictive model for MRSA infection by:

- Repeatedly fitting a predictive matrix of classifiers to patient profiles, generating a matrix of predicted MRSA outcomes;
- Evaluating optimality of the predicted outcomes, using a least-squares error function;

- Iteratively re-calculating the predictive classifiers for maximum accuracy, using stepwise Newton-Raphson optimization; and
- Predicting patient outcomes by applying a logistic binary classification formula.

Full mathematical details of the IRLS regression algorithm are described in Appendix B.

Prototype 1 Results: When tested, the MRSA model that was generated by the IRLS logistic regression algorithm showed greater than 96% accuracy but less than 4% sensitivity (Table 1). The fitted model correctly predicted only 65 out of 1,682 patients actually infected with MRSA, making this prototype unsuitable for predicting MRSA infections.

Table 1. IRLS Algorithm Results				
	Correct Predictions	Expected Predictions	Percent Correct	
Total	46,506	48,203	96.5%	Accuracy
MRSA-Negative	46,441	46,521	99.8%	Specificity
Positive	65	1,682	3.9%	Sensitivity

Prototype 2: To develop a model with higher sensitivity, I used adaptive boosted tree-based logistic regression (LogitBoost), described in papers by Freund and Schapire (1995), Friedman et al. (1999), and Rennie (2003) (22-24). I modified an existing implementation of LogitBoost (25). This modified LogitBoost algorithm developed a predictive model, using my MRSA model design, by:

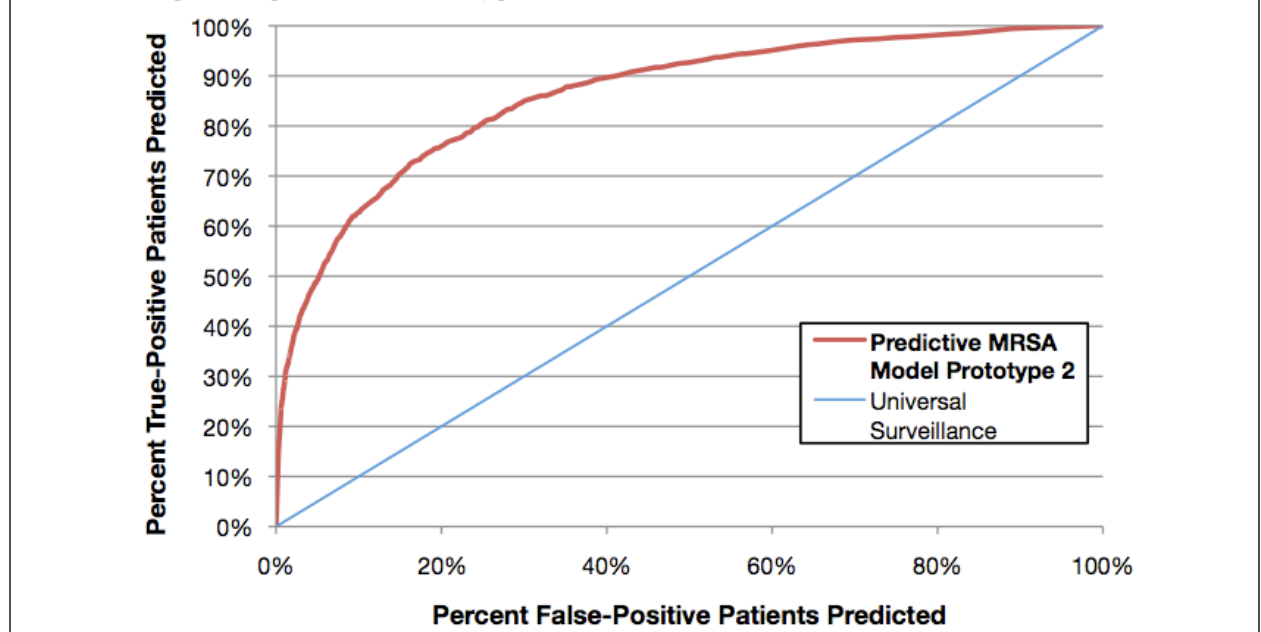
- Calculating an intermediate prediction for each patient by computing inverse probabilities of MRSA infection;
- Fitting a binary decision tree to patient risk factors by computing minimum variance between weighted intermediate predictions and true MRSA diagnoses;
- Adding the fitted binary decision tree to a list of predictive classifiers;
- Determining misclassified patients by calculating minimum variance from zero;
- Calculating accuracy of intermediate predictions by computing log-likelihood of accurate results; and

- Repeating all above steps until the difference in log-likelihood accuracy was less than 0.0005.

The predictive classifiers were used to generate probabilities of MRSA infection for each patient. The full mathematical details of this algorithm are described in Appendix C.

Prototype 2 Results: I calculated the accuracy of Prototype 2 by analyzing a receiver-operating characteristic curve (“ROC Curve”) generated by the model (Figure 4). A left-endpoint estimate showed that Prototype 2 gave an area under curve of 0.86 compared to an area under curve of 0.5 for universal surveillance methods.

Figure 4. Prototype 2 ROC Curve. Prototype 2 developed a model that is more sensitive than universal surveillance. The x-axis of the figure displays the percent of patients who were false-positive for MRSA (i.e. patients in a cohort who were predicted positive but were truly negative); the y-axis displays the percent of patients who were true-positive for MRSA (i.e. patients in a cohort who were predicted positive and were truly positive).



Recommendation: Developing a Testing Strategy Optimizer

I next designed a cost-benefit analysis algorithm that used my *Predictive Model* to optimize MRSA surveillance strategies for each patient in a hospital (Figure 5) by:

- Inputting true-positive and false-negative testing requirements defined by a hospital;
- Iterating all possible percent risk cutoffs to calculate the optimal cutoff, satisfying

hospital requirements;

- Converting predicted percent risks to positive or negative patient predictions for MRSA, using a binary test; and
- Recommending optimal testing methods, using rules-based decision trees and factoring in available testing options.

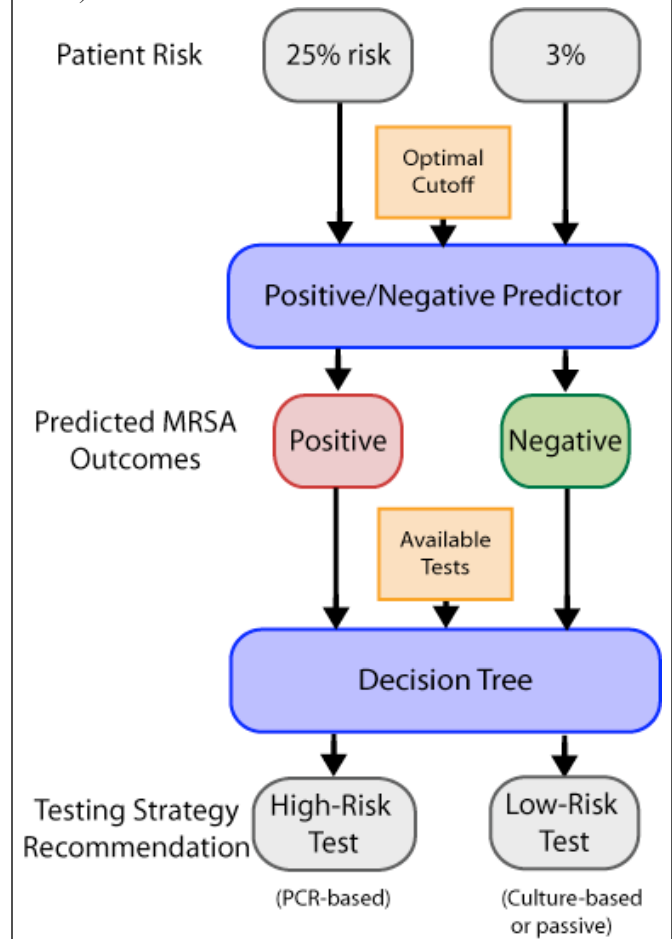
This algorithm recommended an optimal testing strategy for each patient based on the available testing options for each hospital.

Cost Estimation: Developing a Hospital Cost Analyzer

I engineered a cost-analysis algorithm that determined total hospital costs for available MRSA testing strategies by analyzing the costs for MRSA tests as well as costs for patient isolation time. My software estimated costs for MRSA surveillance by:

- Inputting isolation cost estimates and applicable patient demographics;
- Determining an appropriate testing method for each patient, using the optimized MRSA testing strategy, using the *Testing Strategy Optimizer*;
- Estimating per-patient cost of each available MRSA test by adding costs for test kits, laboratory overhead, labor, and disposable equipment;

Figure 5. Testing Strategy Optimizer Procedure. *The Testing Strategy Optimizer inputted patient risk percentages created by the Predictive Model, converted those into predicted MRSA outcomes, then used a decision tree to determine the appropriate appropriate testing strategy (i.e. which test was most appropriate for the patient.) Patients at high risk were typically recommended for PCR-based testing, while patients at low risk were typically recommended for culture-based or passive testing. (Figure by author.)*



- Calculating time spent in isolation by subtracting pre-test isolation time and post-test isolation time from total length of stay;
- Calculating costs for patient isolation time, using estimates provided by a hospital; and
- Repeating the above steps for every available MRSA surveillance strategy, including predictive surveillance strategies and universal surveillance strategies.

Validation of Results: I validated that the results provided by my *Hospital Cost Estimator* accurately represented total hospital costs by estimating costs for universal PCR-based surveillance for the average-size hospital system described in the Robicsek study. I inputted cost estimates provided by Peterson et al., who used MRSA tests provided at manufacturer's volume discount pricing (26). Table 2, provided later, shows that my *Hospital Cost Estimator* predicted total hospital costs for universal PCR-based surveillance of \$934,000, with an overall test cost of \$665,000. While my overall cost was higher than Peterson estimates of \$600,000, the Peterson study only factored in costs for MRSA tests; when patient isolation costs were disregarded, my results were within a 11% margin of error of the Peterson study (12).

Public Education: Developing an Interactive MRSA Website

Interactive MRSA Risk Analysis: I engineered a website based on a Java Spring interactive-website framework that:

- Inputs a patient profile, using a web-based form;
- Uses my *Predictive Model* to determine MRSA-colonization risk for the inputted patient profile; and
- Computes the variance between the patient's colonization risk and national-average risk, converting MRSA colonization risk into risk quartiles (high/normal/low).

Public MRSA Reports: Additionally, I engineered software that displays graphs and charts

that I made available to the public on my *Interactive MRSA Website*. I used open-source JFreeChart chart-generation software to graph the incidence of MRSA among patient groups in the United States through single-variable incidence analysis of the patient database.

Designing Predictive MRSA Surveillance Strategies

I applied my software, which included the *Predictive Model*, the *Testing Strategy Optimizer*, and the *Hospital Cost Estimator*, to create two new predictive MRSA surveillance strategies for the Chicago hospital system described in the Robicsek data (11):

1. The first strategy recommended PCR-based testing for high-risk patients and passive testing for low-risk patients — a strategy I termed “**predictive specific surveillance.**”
2. The second strategy recommended PCR-based testing for high-risk patients and culture-based testing for low-risk patients — a strategy I termed “**predictive universal surveillance.**”

Cost Estimation and Results: Combining hospital-specific costs and national estimates, I estimated costs for true-positive isolation (i.e. isolating a MRSA-positive patient) at \$30/day, costs for false-positive isolation (i.e. isolating a MRSA-negative patient) at \$30/patient plus \$45/day, and costs for false-negative isolation (i.e. not isolating a MRSA-positive patient) at \$250/patient plus \$250/day (10,27-29). I used manufacturer volume-discounted prices to estimate MRSA test costs: PCR-based tests at \$16/test, and culture-based tests at \$5.50/test (13-17).

My *Hospital Cost Estimator* estimated total costs of predictive universal surveillance at \$719,000/year, and costs of predictive specific surveillance at \$688,000/year (Table 2; highlighted in blue). Costs of universal surveillance were higher; costs for PCR-based universal surveillance were estimated at \$934,000/year, costs for culture-based universal surveillance were

estimated at \$1,129,000/year, and costs for passive surveillance were estimated at \$1,946,000/year (Table 2).

Table 2. Cost Estimates for Available MRSA Surveillance Strategies					
	True-Positive Isolation Cost (USD)	False-Negative Isolation Cost (USD)	False-Positive Isolation Cost (USD)	Test Cost (USD)	Total Cost (USD)
Predictive Universal Surveillance	\$157,592.00	\$223,625.00	\$20,890.00	\$317,000.00	\$719,000.00
Predictive Specific Surveillance	\$141,723.00	\$399,573.00	\$12,046.00	\$135,000.00	\$688,000.00
Universal PCR-Based Surveillance	\$171,621.00	\$42,122.00	\$55,237.00	\$665,000.00	\$934,000.00
Universal Culture-Based Surveillance	\$98,679.00	\$791,432.00	\$10,832.00	\$228,000.00	\$1,129,000.00
Passive Surveillance	\$0.00	\$1,946,250.00	\$0.00	\$0.00	\$1,946,000.00

Discussion

I successfully engineered four software modules based on predictive healthcare methods to predict high-risk patients, recommend optimal testing methods, analyze the overall cost of MRSA at a hospital, and deliver public reports about MRSA infections. My *Predictive Model* employs a LogitBoost model-development algorithm to accurately predict MRSA-colonized patients more efficiently than current surveillance models; my *Testing Strategy Optimizer* uses results from the *Predictive Model* to recommend a PCR-based, culture-based, or passive MRSA test for each patient admitted to a hospital; my *Hospital Cost Estimator* generates accurate cost estimates, giving hospitals tools for ensuring patient safety while selecting the most cost-effective MRSA surveillance strategy; and my *Interactive MRSA Website* provides interactive individual predictions of MRSA risk and public reports about MRSA infections in the United States.

My *Predictive Model* was accurate when validated against cohort data from Robicsek et al.

and hospital costs from Peterson et al. Validation showed that my model is more sensitive than universal surveillance, giving an area under curve of 0.86 compared to 0.5 for universal surveillance (Figure 4). Testing 6% of the patient population identified over 50% of MRSA-colonized patients, compared to 6% of patients using universal surveillance. Additionally, testing 50% of the patient population with my model identified 92% of MRSA colonized patients, compared with only 50% using universal surveillance.

My *Hospital Cost Estimator* was accurate when validated against the Robicsek and Peterson data using manufacturer volume discount pricing. While my cost estimator gave higher costs for universal PCR-based surveillance than actual costs determined by the Peterson study, the Peterson study did not consider increased costs for patient isolation. When patient isolation costs were not considered, my predicted costs were within a margin of error of 11% compared to the Peterson costs, which were rounded to the nearest hundred-thousand (Table 2). This result shows that my *Hospital Cost Estimator* is a valid means of estimating total costs for all MRSA surveillance strategies.

My data supported conclusions drawn by Robicsek, Peterson, and the Institute for Healthcare Improvement: active surveillance for MRSA decreases overall hospital costs for treating and monitoring this highly infectious disease. My cost-estimation showed that total costs for passive surveillance were nearly \$817,000 more than active surveillance strategies, providing a strong argument for implementation of active surveillance strategies in hospitals across America.

Most importantly, my study shows the viability of using a predictive universal surveillance strategy. At the average-size hospital modeled, costs for MRSA surveillance were reduced by > \$214,000/year (23%) using *Predictive Universal Surveillance*, and costs were reduced by > \$245,000/year (26%) using *Predictive Specific Surveillance*. Data suggest that using *Predictive*

Universal Surveillance instead of traditional universal surveillance or passive surveillance can allow hospitals across the United States to prevent MRSA infections effectively while substantially reducing costs. The *Interactive MRSA Website* developed in this study is an important new analysis tool that provides predictive capability to hospitals, clinics, and the public to identify MRSA infections.

Conclusion

My results have several limitations. The dataset of patients from which I developed my model was based in a single geographical area. Although the algorithm I used in model development has been shown by previous studies to be resistant to over-fitting a single dataset, MRSA risk factors may vary across national populations (23,30). Future work will include cross-validating my model with other national datasets to ensure maximum accuracy.

My hospital cost modeling may not be accurate for the entire nation, for three reasons: first, modeling was based on a single hospital that successfully employed all five aspects of the 2006 Institute for Healthcare Improvement plan; cost estimates do not account for training, administration, doctor/nurse salaries, or other factors not directly related to testing and isolation. Second, modeling was based on costs for two commercially available tests, provided at manufacturer volume discount pricing and average labor estimates; a hospital's real testing costs will vary, depending on true test costs. Third, my estimations for patient-care costs are based on national studies, which may not accurately reflect true hospital costs. The results of my study, therefore, do not represent minimum or maximum cost savings for a given hospital, but do indicate that predictive surveillance strategies decrease overall costs for MRSA surveillance.

Future iterations of my software will factor in costs for training, administration, doctor/nurse salaries, and other factors indirectly related to MRSA surveillance. Future iterations will also include cross-validating my *Predictive Model* with databases of MRSA patients sponsored by the

United States Centers for Disease Control and Prevention.

My project represents the first application of predictive healthcare techniques to reduce the cost of MRSA surveillance across the nation as well as the first public website allowing the public to access real-time statistics of MRSA infection. The results of my study suggest large cost-savings associated with the use of predictive surveillance; in addition, the software and the website I engineered provides tools to realize more effective MRSA surveillance and MRSA analysis in hospitals. In all, my project allows hospitals and the public to achieve greater control of this deadly endemic disease.

Acknowledgements

Dr. Ronald McGlennen, my research advisor, provided feedback on the medical applicability of my software and reviewed my final paper for accuracy. Dr. Ari Robicsek provided an initial dataset from his 2008 MRSA study. Dr. Clifford McDonald and Dr. John Jernigan of the U.S. Centers for Disease Control and Prevention, Division of Healthcare Quality and Promotion, were instrumental in determining the nationwide applicability of this research and providing insight into medical techniques for MRSA prevention. My classmates and my teacher, Lois Fruen, peer-reviewed my paper for clarity and flow.

Sources Cited

- (1) Klevins, Monina R., et al. "Invasive Methicillin-Resistant *Staphylococcus aureus* Infections in the United States." *Journal of the American Medical Association* 298 (2007): 1763-1771. 28 May 2008
<http://www.cdc.gov/ncidod/dhqp/pdf/ar/InvasiveMRSA_JAMA2007.pdf>.
- (2) Blot, Stijn I., et al. "Outcome and Attributable Mortality in Critically Ill Patients With Bacteremia Involving Methicillin-Susceptible and Methicillin-Resistant *Staphylococcus aureus*." *Archives of Internal Medicine* 162 (2002): 2229-2235. 28 May 2008
<<http://archinte.ama-assn.org/cgi/content/abstract/162/19/2229>>.

- (3) "Mentor Hospital Registry: MRSA." 31 Mar. 2007. Institute for Healthcare Improvement. 23 Sept. 2008 <http://www.ihl.org/ihl/programs/campaign/mentor_registry_mrsa.htm>.
- (4) "Highmark Blue Cross Blue Shield-Changing Incentives to Promote Better Care." 2008. Blue Cross Blue Shield Association. 23 Sept. 2008 <<http://www.bcbs.com/issues/uninsured/highmark-blue-cross-blue-shiel.html>>.
- (5) Institute for Healthcare Improvement. *Getting Started Kit: Reduce Methicillin-Resistant Staphylococcus aureus (MRSA) Infection*. 1st ed. 12 Dec. 2006. 23 Sept. 2008 <<http://www.ihatoday.org/issues/quality/ihireductionguide.pdf>>.
- (6) Health-Care-Associated Infections In Hospitals: An Overview of State Reporting Programs and Individual Hospital Initiatives to Reduce Certain Infections. United States. Government Accountability Office. 2008. Government Accountability Office. Sept. 2008. 3 Dec. 2008 <<http://www.gao.gov/new.items/d08808.pdf>>.
- (7) Klein, Eili, et al. "Hospitalizations and Deaths Caused by Methicillin-Resistant *Staphylococcus aureus*, United States, 1999-2005." *Emerging Infectious Diseases* 13 (2007): 1840-1846. 28 May 2008 <<http://www.cdc.gov/eid>>.
- (8) "MRSA in Healthcare Settings." *Centers for Disease Control*. 17 Oct. 2007. CDC. 28 May 2008 <<http://www.cdc.gov/Features/MRSA>>.
- (9) Diederer, B.M.W., and J.A.J.W. Kluytmans. "The emergence of infections with community-associated methicillin resistant *Staphylococcus aureus*." *Journal of Infection* 10 (2006): 157-68. 28 May 2008 <<http://www.ncbi.nlm.nih.gov/pubmed/16289303>>.
- (10) Robicsek, Ari. Personal interview. 30 May 2008.
- (11) Robicsek, Ari, et al. "Universal Surveillance for Methicillin-Resistant *Staphylococcus aureus* in 3 Affiliated Hospitals." *Annals of Internal Medicine* 148 (2008): 409-418.
- (12) Peterson, Lance R., et al. "Case Study: an MRSA Intervention At Evanston Northwestern Healthcare." *The Joint Commission Journal on Quality and Patient Safety* 33 (2007): 732-738. 28 May 2008 <<http://www.jcrinc.com/26813/newsletters/32/>>.
- (13) Remel, Inc. *Spectra MRSA Product Insert*. 3 Mar. 2008.
- (14) BD Diagnostics. *BBL CHROMagar MRSA Product Insert*. May 2005.
- (15) BD Diagnostics. *BD GeneOhm™ MRSA Assay Product Insert*. May 2005.
- (16) BD Diagnostics. *BD GeneOhm™ MRSA Test Procedure*. 2006.
- (17) Cepheid. *Xpert™ MRSA*. Brochure. 2007.
- (18) *Personalized Health Care: Opportunities, Pathways, Resources*. United States Department of Health and Human Services. Sept. 2007. 24 Sept. 2008 <<http://www.hhs.gov/myhealthcare/>>.
- (19) Bewick, Viv., et al. "Statistics Review 14: Logistic Regression." *Critical Care* 9 (2005):

112-118.

- (20) Derr, Robert E. "Performing Exact Logistic Regression with the SAS System." 24 Sept. 2008 <<http://www.ats.ucla.edu/stat/sas/library/exactlogistic.pdf>>.
- (21) Movellan, Javier R. "Tutorial on Multivariate Logistic Regression." 23 Jul. 2006. 24 Sept. 2008 <<http://mplab.ucsd.edu/wordpress/tutorials/MultivariateLogisticRegression.pdf>>.
- (22) Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." 20 Sept. 1995. 24 Sept. 2008 <<http://citeseer.ist.psu.edu/freund95decisiontheoretic.html>>.
- (23) Friedman, Jerome, et al. "Additive Logistic Regression: a Statistical View of Boosting." 20 Aug. 1998. 24 Sept. 2008 <<http://citeseer.ist.psu.edu/friedman98additive.html>>.
- (24) Rennie, Jason. "Boosting with decision stumps and binary features." 10 Apr. 2003. 24 Sept. 2008 <<http://people.csail.mit.edu/jrennie/writing/stumps.pdf>>.
- (25) Witten, Ian H., and Eibe Frank. "Data Mining: Practical machine learning tools and techniques." Morgan Kaufmann: San Francisco, 2005.
- (26) Reese, Kathryn. Telephone interview. 18 Sept. 2008.
- (27) McGlennen, Ron. Personal interview. 29 May 2008.
- (28) Noskin, Gary A., et al. "The Burden of *Staphylococcus aureus* Infections on Hospitals in the United States." *Archives of Internal Medicine* 165 (2005): 1756-1761. 28 May 2008 <<http://archinte.ama-assn.org/cgi/content/abstract/165/15/1756>>.
- (29) Cosgrove, Sara E., et al. "The Impact of Methicillin Resistance in *Staphylococcus aureus* Bacteremia on Patient Outcomes: Mortality, Length of Stay, and Hospital Charges." *Infection Control and Hospital Epidemiology* 26 (2005): 166-174. 28 May 2008 <<http://www.ncbi.nlm.nih.gov/pubmed/15756888>>.
- (30) McDonald, Clifford, and John Jernigan. Telephone interview. 31 Oct 2008.

Appendices

Appendix A: Patient Information Modeled

Characteristic Name	Description	Possible Values
uniqueid	Internal model value - A unique ID for a this patient	Numeric
datasetid	Internal model value - The Study Dataset ID	Numeric
hospital	Hospital to which the patient patient was admitted	Hospital Name
length_of_stay	Length that the patient has stayed at the hospital	Days
insurer	Patient's insurer	Insurer Name
patient_type	Type of patient	Observation, Inpatient
admit_type	Type of patient admission	Emergency, Elective
admit_service	Service into which the patient was admitted	Medicine, Surgery, Gynecology, Psychology, Pediatrics, Rehab
icu_admit	Was this patient an ICU admission?	Yes or No

admission_date	Date the patient was admitted	Date (YYYY-MM-DD)
discharge_type	The location to which a patient was discharged	Home, Hospital, Long Term Care Facility, or Expired
age	Patient age	Numeric
sex	Patient sex	Male or Female
race	Patient race	White, Hispanic, Black, Asian, American Indian/Alaska Native, and Other.
previous_mrsa	Has the patient had MRSA in the past 90 days?	Yes or No
previous_admit	Has the patient been admitted to the hospital in the last 90 days?	Yes or No
pressure_ulcer	Does the patient have pressure ulcer?	Yes or No
high_temperature	Did the patient have a fever upon admission?	Yes or No
has_diabetes	Does the patient have diabetes?	Yes or No
has_hemodialysis	Does the patient have hemodialysis?	Yes or No
soft_tissue_infection	Does the patient have a soft tissue infection?	Yes or No
diagnosis_codes	The patient's admission diagnosis	ICD-9 Diagnosis Codes
admission_tested	Was this patient tested for MRSA upon admission?	Yes or No
mrsa_on_admit	Did this patient have MRSA when admitted? (Modeling Value)	Yes or No
previous_positives	Did this patient previously have MRSA?	Yes or No
previous_dates	The dates this patient previously had MRSA	Dates
renal_dialysis	Does this patient have renal dialysis?	Yes or No
cancer	Does this patient have cancer?	Yes or No
congestive_heart_failure	Does this patient have congestive heart failure?	Yes or No
chronic_lung	Does this patient have a chronic lung infection?	Yes or No

Appendix B: IRLS Regression Algorithm

This appendix describes the mathematical details of the IRLS regression algorithm that I implemented, based on formulas described by Derr et al., Bewick et al. (2005) and Movellan et al. (2006) (19-21).

All m patients and their n risk factors are considered a $m \times n$ mathematical matrix X ; MRSA outcomes are considered an $c \times m$ matrix Y where $c = 1$. The IRLS algorithm then repeatedly fits an $n \times c$ predictive matrix of classifier parameters B to X to generate a matrix of predicted outcomes \hat{y} . At each iteration, the algorithm evaluates the optimality of the predicted outcomes,

using the formula $\Phi(B) = -\sum_{j=1}^m \left[\sum_{k=1}^c (y_{j,k} - \hat{y}_{j,k})^2 \right] + \frac{\alpha}{2} \sum_{k=1}^c B^T B$, where α is a positive constant. The

gradient between y and \hat{y} is calculated, using the equation $\nabla_B \Phi = x^T (y - xB) + \alpha B$. Typically, B is updated to minimize the error with a Newton-Raphson stepwise optimization procedure, using the recursive formula $B_{j+1} = B_j + (\nabla_B \nabla_B \Phi_j)^{-1} \nabla_B \Phi_j$, until $\nabla_B \nabla_B \Phi < r$ where r is defined as a

cutoff. However, noting that the traditional Newton-Raphson method requires computing $\Phi(B)^*$, it is more efficient to solve the formula analytically and skip directly to the optimum value for B , \hat{B} using $\hat{B} = (x^T x + \alpha I_n)^{-1} + \alpha B$. The resulting matrix of optimal predictive classifiers \hat{B} is used to predict patient outcomes by evaluating $\frac{1}{1 + e^{-x\hat{B}}} > 0.5$, resulting in a patient prediction.

Appendix C: LogitBoost Algorithm

This appendix describes the LogitBoost algorithm that I used in my project, based on work by Freund and Schapire (1995), Friedman et al. (1999), and Rennie (2003) (22-24).

To fit a model to patient risk factors, each of the N patients is considered a member x_i of a patient vector x . A vector of weights w of length N is initialized, such that $w_i = \frac{1}{N}$. The classifier function $F(x)$ is initialized, such that $F(x) = 0$, and the estimate of patient probabilities $P(x)$ is initialized, such that $P(x_i) = \frac{1}{2}$. A limited number of boosting rounds M and a weight percent threshold T are defined.

The following steps are repeated for every boosting round $m = 1, 2, \dots, M$:

- The working response z_i is calculated, using the equation $z_i = \begin{cases} \frac{1}{p}, y^* = 1 \\ \frac{-1}{(1-p)}, y^* = 0 \end{cases}$, where y^* is the response output by $F(x)$. The value is thresholded, such that $-z_{\max} \leq z_i \leq z_{\max}$ for $z_{\max} \in [2, 4]$, an empirically-determined value.
- The working weight vector w is re-calculated, such that $w_i = \max(p(x_i)(1 - p(x_i)), \frac{T}{2})$.
- A binary decision tree $f(x)$ is calculated by computing least possible variance between current estimations z_i and true answers x_i , which are weighted using w_i for each risk factor in the dataset.

- $F(x)$ is updated using the equation $F(x) \leftarrow F(x) + \frac{1}{2} f_m(x)$, and $P(x)$ is updated using the equation $P(x) \leftarrow \frac{e^{F(x)}}{e^{F(x)} + e^{-F(x)}}$.
- Patients are eliminated where $w_i < \left(\sum_{j=1}^N w_j \right) \times T, T \in [0,1]$.
- The log-likelihood $L_m(x)$ is calculated, using $L_m(x) = \sum_{i=1}^N -2 \log(p(x_i))$, and the steps are repeated if $L_m - L_{m-1} > H$, where $H = 0.0005$, and if $m < M$.

$F(x)$ is used to predict the probability of each patient being infected with MRSA.