

Appendix E-1: Statistical Analysis

All statistical analyses were performed using a standard software package (Stata, version 12.0; StataCorp). Descriptive statistics were generated for the entire sample for all potential predictors and after stratification by surgical site infection (SSI) status. A hold-out sample of 1,000 random observations was then excluded from the following regression models for later use for validation of our final prediction model. Simple logistic regression models were created for all potential predictors to assess for an unadjusted association with SSI. A multiple logistic regression model to predict SSI was then created with a backward selection method. Specifically, all covariates were initially included in the multivariate model. We removed covariates in a stepwise fashion on the basis of the likelihood-ratio chi-square p value. All main effects were included in the model in addition to all biologically plausible interaction terms between predictors. Interaction terms were also included between all medical diagnoses and the duration of time between the diagnosis and the day of surgery to account for possible reductions in infection risk over time (for example, a reduction in SSI risk for surgery conducted 10 years compared with 1 year after hospitalization for sepsis). Interactions were eliminated first, followed by main effects, with an exit criterion of a likelihood ratio chi-square p value of >0.05 . However, there was forced inclusion of insulin dependence ($p = 0.07$), as it was found to be a strong confounder of the overall association of diabetes with infection risk. Regression diagnostics were then performed; the observations with the highest residuals and leverage were identified and reviewed for biological plausibility; no values required removal from the sample. The discriminatory capability of the final model (model 1) was assessed by receiver operating characteristic (ROC) curve analysis; goodness of fit was assessed by the Hosmer-Lemeshow (H-L) chi-square test. A point scoring system was then created on the basis of the final coefficients for model 1; points were assigned by dividing the respective coefficients for each predictor by the smallest coefficient in model 1 and rounding to the nearest half-point. No interaction terms were included, so the resulting risk score was determined as the sum of the individual point values. Regression diagnostics were then repeated for a model with the risk score as the single predictor (model 2). The influence of potentially modifiable, lifestyle-related risk factors on SSI risk was then estimated for the sample; we defined this as the difference between the estimated SSI risk at baseline and the estimated SSI risk after eliminating morbid obesity, tobacco use, and diabetes. Finally, the risk score was validated on the hold-out sample of 1,000 observations by repeated assessment of discriminatory capability via ROC curve analysis and goodness of fit by the H-L chi-square test.