Published in final edited form as:

Anesth Analg. 2013 December ; 117(6): 1338–1346. doi:10.1213/ANE.0b013e3182a46d6d.

How to Improve the Performance of Intraoperative Risk Models: An Example with Vital Signs Using the Surgical Apgar Score

Joseph A. Hyder, MD, PhD^{*}, Daryl J. Kor, MD[†], Robert R. Cima, MD[‡], and Arun Subramanian, MBBS[†]

^{*}Department of Anesthesiology, Mayo Clinic, Rochester, MN

[†]Department of Anesthesiology, Division of Critical Care Medicine, Mayo Clinic, Rochester, MN

[‡]Department of Surgery, Mayo Clinic, Rochester, MN

Abstract

BACKGROUND—Computerized reviews of patient data promise to improve patient care through early and accurate identification of at-risk and well patients. The significance of sampling strategy for patient vital signs data is not known. In the instance of the surgical Apgar score (SAS), we hypothesized that larger sampling intervals would improve the specificity and overall predictive ability of this tool.

METHODS—We used electronic intraoperative data from general and vascular surgical patients in a single-institution registry of the American College of Surgeons National Surgical Quality Improvement Program. The SAS, consisting of lowest heart rate, lowest mean arterial blood pressure, and estimated blood loss between incision and skin closure, was calculated using 5 methods: instantaneously and using intervals of 5 and 10 minutes with and without interval overlap. Major complications including death were assessed at 30 days postoperatively.

Address correspondence to Joseph A. Hyder, MD, PhD, Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA. Address e-mail to joseph.a.hyder@gmail.com.

Joseph A. Hyder, MD, PhD, is currently affiliated with Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA.

Joseph A. Hyder, MD, PhD, is currently affiliated with the Center for Surgery and Public Health, Brigham and Women's Hospital, Boston, MA and Mayo Clinic, Department of Anesthesiology, Rochester MN.

The authors declare no conflicts of interest.

Reprints will not be available from the authors.

DISCLOSURES

Name: Joseph A. Hyder, MD, PhD.

Contribution: This author helped design the study, conduct the study, analyze the data, and write the manuscript.

Attestation: Joseph A. Hyder has seen the original study data, reviewed the analysis of the data, approved the final manuscript, and is the author responsible for archiving the study files.

Name: Daryl J. Kor, MD.

Contribution: This author helped design the study, conduct the study, and write the manuscript.

Attestation: Daryl J. Kor has seen the original study data, reviewed the analysis of the data, and approved the final manuscript.

Name: Robert R. Cima, MD.

Contribution: This author helped design the study and write the manuscript.

Attestation: Robert Cima reviewed the analysis of the data and approved the final manuscript.

Name: Arun Subramanian, MBBS.

Contribution: This author helped design the study, conduct the study, analyze the data, and write the manuscript.

Attestation: Arun Subramanian has seen the original study data, reviewed the analysis of the data, and approved the final manuscript.

This manuscript was handled by: Sorin J. Brull, MD, FCARCSI (Hon).

RESULTS—Among 3000 patients, 272 (9.1%) experienced major complications or death. As the sampling interval increased from instantaneous (shortest) to 10 minutes without overlap (largest), the sensitivity, positive predictive value, and negative predictive value did not change significantly, but significant improvements were noted for specificity (79.5% to 82.9% across methods, P for trend <0.001) and accuracy (76.0% to 79.3% across methods, P for trend <0.01). In multivariate modeling, the predictive utility of the SAS as measured by the c -statistic nearly increased from $c = +0.012$ ($P = 0.038$) to $c = +0.021$ ($P < 0.002$) between the shortest and largest sampling intervals, respectively. Compared with a preoperative risk model, the net reclassification improvement and integrated discrimination improvement for the shortest versus largest sampling intervals of the SAS were net reclassification improvement 0.01 ($P = 0.8$) vs 0.06 ($P = 0.02$), and for integrated discrimination improvement, they were 0.008 ($P < 0.01$) vs 0.015 ($P < 0.001$).

CONCLUSIONS—When vital signs data are recorded in compliance with American Society of Anesthesiologists' standards, the sampling strategy for vital signs significantly influences performance of the SAS. Computerized reviews of patient data are subject to the choice of sampling methods for vital signs and may have the potential to be optimized for safe, efficient patient care.

The widespread adoption of the electronic health record (EHR) promises to improve many aspects of patient care.¹ For inpatients and especially surgical patients, the EHR may allow early identification of at-risk patients and appropriate identification of patients at very low risk. Such information would allow for early and efficient distribution of hospital resources, whether to escalate care for the sick or to de-escalate care for the well. In addition, early identification of at-risk patients would aid enrollment into time-sensitive clinical trials.^{2–4}

Fully leveraging the EHR requires computerized reviews of clinical data including thousands of vital signs. Although a single set of vital signs might be useful during patient care,⁵ meaningful use of thousands of vital signs depends on sampling methods to correctly identify worrisome patterns.^{6–10} Although a few computerized reviews have demonstrated moderate success in identifying at-risk patients,^{11,12} it is unknown whether data sampling methods affect the success of any filtering strategies. In this study, we selected 1 robust data-filtering method, the surgical Apgar score (SAS), as a test case for the impact of sampling-based variability. The SAS, which uses lowest heart rate (HR), lowest mean arterial blood pressure (MAP), and estimated blood loss (EBL) between incision and skin closure, has been validated to predict 30-day major complications but was developed with hand-charted anesthesia records with 5-minute sampling intervals for vital sign data.^{13–16} The performance of the SAS may be, according to its originators, highly dependent on measurement variability, but this has never been formally investigated.¹⁷ We applied different data sampling strategies to vital signs data composing the SAS and then examined variation in the score and in its ability to predict postoperative risk. No single test can definitively compare utility of models, so we used multiple methods to describe changes in discrimination, calibration, reclassification, sensitivity, specificity, positive and negative predictive values, and accuracy of the SAS by sampling method. We hypothesized that larger sampling intervals would improve predictive ability through improved specificity.

METHODS

Study Participants

For quality improvement purposes, Mayo Clinic, Rochester, MN, has maintained a database of surgical patients as part of institutional participation in the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP).¹⁸ All Mayo NSQIP patients ($N = 13,260$) who underwent general or vascular surgery between April 26, 2006 and January 28, 2010 were eligible for study inclusion. Of those, 12,987 had complete data describing intraoperative vital signs. A random sample of 3000 general and vascular patients of ASA physical status classification $<V$ and who underwent general anesthesia was included for investigation. Sample size was determined by approximate monthly clinical surgical volume at Mayo Clinic Rochester with the rationale that differences detectable from month to month may be of clinical value, but differences requiring much larger samples and larger periods of time to detect would potentially be of less clinical value. This study was approved by the IRB of Mayo Clinic, and the requirement for written informed consent was waived by the IRB.

The SAS

The SAS has been described elsewhere.¹⁷ Its construction is based on slowest HR, lowest MAP, and EBL between surgical incision and skin closure (Table 1). At Mayo Clinic, Rochester, MN, intraoperative records are kept in a proprietary SQL database within the Anesthesia Information Management System (PICIS ChartPlus, Wakefield, MA). This system records vital signs at 2-minute intervals. For recording patient HR, values from a pulse oximeter were used only if the $SaO_2 > 80\%$ (surrogate for quality of the data); otherwise, HR values from electrocardiogram data were used. If source data were missing for both, then no HR was recorded for a given time stamp. To calculate the lowest MAP, patient data were filtered to exclude any individual values < 25 mm Hg or > 180 mm Hg, consistent with others' publications on the SAS.¹⁴ When multiple measurements were available from a specific measurement modality, (e.g., data from 2 arterial catheters available), the higher of the values was recorded as the summary value from that modality for that time stamp. If no arterial catheter MAP was available, then noninvasive cuff pressure was used. For MAP data, 14.8% of cases had at least 1 gap > 5 minutes, and 1.1% of cases had at least 1 gap > 10 minutes. For HR, 1.1% of cases had at least 1 gap > 5 minutes, and 0.4% of cases had at least 1 gap > 10 minutes.

Sampling Methods

Five sampling methods for slowest HR and lowest MAP were established before initiating data analyses. These methods were based on "windows" or intervals of data and were established as follows: instantaneous (each data point is its own window), 5- and 10-minute nonoverlapping windows with windows beginning at the time of incision (0–5 minutes, 6–10 minutes, 11–15 minutes, etc., or 0–10 minutes, 11–20 minutes, 21–30 minutes, etc.) and overlapping windows of 5- and 10-minutes (0–5 minutes, 1–6 minutes, 2–7 minutes, etc., or 0–10 minutes, 1–11 minutes, 2–12 minutes, etc.). For overlapping intervals, a new sampling window was established with the next available vital signs or after 2 minutes if no new vital sign was recorded. Within each window, a median value was determined. Median values for

HR and MAP were the basis for the original SAS investigations, and median values were chosen for this investigation. EBL as recorded by the in-room anesthesia provider was calculated for the entire case.

Preoperative Variables

Preoperative health states and postoperative clinical data were classified using ACS NSQIP definitions.¹⁹ Model covariates were selected a priori based on review of the literature, presumed explanatory ability, and clinical applicability.²⁰ Efforts were made to exclude specific variables, such as clinical laboratory data, that are not commonly available for surgical patients and to include easily acquired covariates. The following ACS NSQIP preoperative variables were recorded from 3-level variables to binary as normal versus not: functional status before surgery, dyspnea, and congestive heart failure. Age was recorded as younger than 65 years vs 65 years or older. Surgical magnitude was classified either as minor (e.g., breast, endocrine, groin and umbilical herniorrhaphy, appendectomy, laparoscopic cholecystectomy, perianal procedures, and skin or soft-tissue operations) or major or extensive (all other operations) as in previous studies.^{15,21,22} ASA physical status classification was included as an unordered categorical variable.

Outcomes

The primary end point was major complications and/or death within 30 days of index surgery as recorded in the Mayo Clinic Rochester, MN, ACS NSQIP database. Consistent with previous investigations, the following ACS NSQIP-defined events²³ were defined as major complications: “acute renal failure, bleeding that required a transfusion of 4 units or more of red blood cells within 72 hours after surgery, cardiac arrest requiring cardiopulmonary resuscitation, coma of 24 hours or longer, deep venous thrombosis, myocardial infarction, unplanned intubation, ventilator use for 48 hours or more, pneumonia, pulmonary embolism, stroke, wound disruption, deep or organ-space surgical site infection, sepsis, septic shock, systemic inflammatory response syndrome, and vascular graft failure.”¹⁴ Additional complications coded were systematically recorded and individually reviewed and classified as major complications using Centers for Medicare & Medicaid Services listing of major complication based on the International Classification of Diseases, Ninth Revision.²⁴ Examples include acute pancreatitis and acute heart failure syndromes. Additional complications in the ACS NSQIP database were individually reviewed by 3 of the investigators (JAH, AS, DJK) and adjudicated by consensus to be major or not based on the need for intensive care unit admission or reoperation. Examples included anastomotic leak, cardiac tamponade, and gangrene.

Statistics

Differences in the distributions of clinical characteristics were tested with analysis of variance for linear variables and χ^2 tests for categorical variables. The SAS was categorized by values 0 to 4; 5 to 6; 7 to 8; 9 to 10 before model construction according to published precedent with the exception of the lowest score categories (0–2 and 3–4) that were collapsed due to small cell sizes in the lowest categories.¹⁵ Sampling-based differences in the SAS and its components (slowest HR, lowest HR category, lowest MAP, and lowest

MAP category) were assessed as within-subject differences in vital sign components by sampling method and tested by paired *t* tests.

Assessment of the clinical utility of a novel risk factor using only a simple comparison of odds ratios or *P*-values is fraught with limitations.²⁵ Therefore, model performance across constructions of the SAS was assessed with multiple methods, detailed below, to inform conclusions from this study. The fundamental approach was to compare the incremental change in a base model's performance on the addition of sampling-based constructions of the SAS.

Because the independent predictive ability of the SAS has been demonstrated in prior studies,¹⁵ differences in model performance with the SAS were evaluated primarily for discrimination and reclassification rather than simple tests of statistical significance of the SAS, which was presumed. Discrimination was assessed using area under the curve, or c-statistic, as well as differences in c-statistic (χ^2) with the addition of each of the 5 SAS constructions. Differences in c-statistic between models were tested with 95% Wald confidence limits and χ^2 test for *P*-values using the ROCCONTRAST statement. Differences in net reclassification improvement (NRI) and integrated discrimination improvement (IDI) were calculated and tested using a z-test for illustrative purposes despite known shortcomings of the z-test for differences in IDI.^{26–28} Model performance was additionally assessed for NRI with table-based reclassification of events and nonevents based on the method of Pepe and Janes.²⁹ Multivariate logistic regression models were applied to each participant to calculate the estimated risk of an event for each participant. These values were used to classify each patient into risk strata. Cut-points for reclassification tables were established empirically based on risk of event according to ASA physical status classification (I vs II vs III or IV).

To facilitate lay interpretation of the reclassification data, 3 comparative methods were used. First, sensitivity and specificity of the SAS were calculated with an arbitrary cut-point of 6, with 95% confidence intervals determined with the exact method. This method additionally served as a sensitivity analysis on the grouping of different SAS risk scores. Second, for multivariate finding, the results of adding a theoretical “perfect new variable” were included to demonstrate perfect reclassification and perfect discrimination. Finally, we compared the SAS head-to-head with the ASA physical status classification to gauge the magnitude of the SAS on risk refinement in comparison to the ASA, which is widely considered a simple, common, and powerfully predictive clinical scoring system. All models were assessed for goodness of fit using the Hosmer–Lemeshow goodness-of-fit statistic. One model violated the goodness-of-fit assumption, and performance was investigated graphically across deciles of risk, suggesting modest heterogeneity of fit due to arbitrary decile construction, a known shortcoming of the Hosmer–Lemeshow goodness-of-fit test.³⁰ The model was included for comparison with other models. Models demonstrated no evidence of multicollinearity as assessed by tolerance values. All analyses were performed using SAS v9.13, 9.2 and 9.3 (Cary, North Carolina). Two-sided *P*-values <0.05 were considered significant. Figures were made with Microsoft Office Excel Edition 2003 (Redmond, WA) and GraphPad Prism6 (La Jolla, CA).

RESULTS

Table 1 presents the construction of the SAS by component. Points assigned to vital signs are analogous to β coefficients of risk, although, as with the Apgar score in pediatrics, lower scores are associated with worse outcomes. For the SAS components, a lower number for slowest HR is associated with a higher HR score, or lower risk of postoperative complication. In contrast, a lower number for lowest MAP is associated with a lower MAP score, or higher risk of postoperative complication.

Table 2 presents demographic characteristics of the 3000 patients and their primary surgeries. In total, 1909 (63.6%) patients underwent major surgery, and 273 (9.1%) experienced a major complication or death within 30 days of surgery. Thirty-seven (1.2%) patients died within 30 days of surgery. Patients were predominantly younger than 65 years of age (1875, 62.5%), female (1678, 55.9%), and ASA physical status II (50.3%), or ASA physical status III and IV (41.3%).

Figure 1 presents within-subject differences by paired t test with standard error of the mean for the SAS components of slowest HR score and lowest MAP score according to the sampling method used for vital signs. The shortest sampling interval (instantaneous) identified the most extreme values for slowest HR, resulting in a higher score for slowest HR compared with the coarsest sampling interval (10-minute median nonoverlapping). For lowest MAP, the shortest sampling interval (instantaneous) again identified the most extreme values compared with other methods. All differences across sampling method for HR score, MAP score, and SAS are statistically significant ($P < 0.05$).

Table 3 demonstrates the distribution of patients into the SAS category by sampling method. Data from this table illustrate a number of patterns relevant to the study hypothesis. For all sampling strategies, the lowest category of the SAS (0–4) included the fewest numbers of patients and the greatest risk. Associations between the SAS and major complication were statistically significant (all P -values < 0.01) in univariate models and multivariate models adjusted for ASA physical status classification (I, II, III, or IV), major surgery, emergency surgery, age, sex, wound classification, functional status before surgery, ascites, dialysis, dyspnea, congestive heart failure, and bleeding. Differences in the distribution of patients into SAS categories and the magnitudes of odds ratios by sampling method are not testable with statistical methods.

Of note, the adjusted model including 10-minute non-overlapping was the only model to have a P -value < 0.1 for the Hosmer–Lemeshow goodness-of-fit test ($P = 0.0296$). This result was graphically inspected and determined to be a result of risk binning, a known source of unpredictability with the Hosmer–Lemeshow test.³⁰ Because the aim of analyses is for clinical comparison, the model is included without alteration.

Table 4 presents incremental changes in model performance with the addition of different constructions of the SAS.

Multivariate Prediction—Overall and by Sampling Method for the SAS

The preoperative risk model demonstrated acceptable predictive ability ($c = 0.775$). Among nonevents, the majority (62%) were assigned to the lowest risk group (risk scores of $<5\%$). Among events, the majority (48%) were assigned to the highest risk group (risk scores of $>13\%$). The addition of the SAS resulted in significantly increased c-statistic in all cases ($P < 0.05$ for all).

Estimates for the incremental predictive ability of the SAS increased as the sampling method was altered from instantaneous to 10-minute nonoverlapping (c-statistic $+ 0.012$, $P = 0.0379$ vs $+ 0.021$, $P = 0.0015$ for 10-minute nonoverlapping sampling strategy); however, it was not possible to statistically test whether these changes were significant.

Reclassification—Overall and by Sampling Method for the SAS

Differences in reclassification by SAS sampling intervals were notable, although not testable statistically. In reclassification tables, the fraction of healthy patients (nonevents) reclassified appropriately to the lowest risk group ($<5\%$ absolute risk) was 4.4% for the instantaneous sampling method and improved to 10% with the 10-minute non-overlapping strategy. Of nonevents that were reclassified, most were redistributed from the intermediate risk group to the lowest risk group. Point estimates for NRI and IDI increased with larger sampling intervals, consistent with improved model performance, but these differences could not be tested statistically.

Clinical Context for the Reclassification Tables

To place these results in clinical context, consider the example of reclassification of patients who did not have a major complication (healthy). Without using the SAS variable, 1036 healthy patients were appropriately assigned to the lowest risk strata. With the addition of the SAS variable made using an instantaneous sampling method, up to an additional 121 patients were appropriately assigned to this lowest risk group. This represents up to 4% of all 3000 surgical patients who might avoid costly and unnecessary intensive postoperative monitoring. By contrast, with the addition of the SAS variable made using a 10-minute nonoverlapping sampling method, up to 185 patients were appropriately assigned to the lowest risk group (as many as 64 more patients than with the instantaneous method, as high as a 50% improvement over the instantaneous method, and 6% rather than 4% of all patients). This difference in specificity between the sampling methods may be large enough to affect approximately one of 20 surgical patients (no P -value is currently applicable to this estimate).

Comparison of the SAS to the ASA Physical Status Classification

To gauge the relative importance of the c-statistic, NRI, and IDI for the SAS, we compared the SAS to the ASA physical status classification, which is an established preoperative and potent predictor of postoperative complication. We compared model performance after adding either SAS or ASA physical status classification with a preoperative risk model (Table 4). These models were adjusted for major surgery, emergency surgery, age, sex, wound classification, functional status before surgery, ascites, dialysis, dyspnea, congestive heart failure, and bleeding. For both variables, the δ -c-statistic increase was statistically

significant ($P=0.0119$ for ASA physical status classification and $P=0.0043$ for SAS). The addition of the ASA physical status classification to the model resulted in improved sensitivity (as assessed by appropriate reclassification of events into the highest risk group with 9.2% of all events reassigned to this group) and minimal change in specificity (reclassification on nonevents). By contrast, the addition of the SAS with 10-minute overlap resulted in appropriate reclassification of nonevents (6.7% of non-events reclassified to the lowest risk group) and appropriate reclassification of events into the highest risk group (4.0% of events). The above differences, however, cannot be tested by conventional statistical methods. However, the net reclassification improvement and integrated discrimination improvement were positive and statistically significant for both variables. Sensitivity analyses varying cut-points for all reclassification tables did not result in material differences from the results presented with persistent, recognizable patterns of difference, directions, and levels of statistical significance.

Table 5 presents univariate changes in sensitivity, specificity, negative predictive value, positive predictive value, and accuracy for each sampling method for the SAS. An SAS of 6 or lower was considered a positive test. As sampling interval increased, the values for positive predictive value, negative predictive value, sensitivity, specificity, and accuracy increased, and the increases in specificity and accuracy were statistically significant (Cochran-Armitage trend 2-sided $P<0.01$ for each).

DISCUSSION

This study investigated the effect of different sampling methods for extracting vital signs data on the performance of the SAS, an established and robust data-filtering method to predict perioperative risk. In 3000 patients undergoing general and vascular surgery, increasing the sampling interval for vital signs contributed to systematic differences in the SAS itself and systematic differences in the ability of the SAS to predict postoperative complications and reclassify patients appropriately. Larger sampling intervals for the SAS resulted in better model discrimination and improved reclassification. These effects due to sampling method alone were large enough to affect approximately 1 in 20 patients and large in comparison to the overall effect of the SAS that was itself a strong predictor of complications (Table 4).

We are unaware of other investigations that examined the effects of sampling strategy on the performance of a computerized review of patient data. In univariate analyses, the SAS demonstrated good sensitivity and specificity, and the specificity and accuracy were significantly improved by increasing the sampling interval. In contrast to nearly all other tests that are used to determine risk of disease, these improvements in specificity and accuracy were achieved with no additional costs in time, personnel, equipment, patient comfort, laboratory resources, or new monitoring equipment. All that was required was a trivial modification of the sampling algorithm for vital signs.

In multivariate analyses, comparisons required more advanced methods but were analogous. The improvement in model prediction by adding the SAS was greater for the larger sampling intervals compared with the shortest interval, although formal statistical tests were not

available in every instance. The estimate for incremental increase in the c-statistic was greater when the largest (+ 0.021; 95% CI, 0.008–0.034) rather than the shortest sampling window (+ 0.012; 95% CI, 0.001–0.022) was used, although statistical significance could not be tested. In reclassification analyses, the utility of the SAS was driven by improved specificity, a finding consistent with univariate methods and in contrast to the ASA physical status classification. The ASA physical status classification variable did improve predictive performance to a similar extent to the SAS (+ 0.019 for the ASA compared with + 0.023 for the SAS) but appeared to do so through improved sensitivity.

The present analyses include investigations of individual SAS components for HR and MAP and are the first to do so since its derivation.¹⁷ In contrast to the original publication of the score, the present investigation demonstrated no statistically significant differences in lowest MAP pressure between those with and without subsequent major complications.¹⁷ This null finding persisted across sampling methods. The noncontributory effect of MAP on the utility of the SAS in the present study may reflect a previously unknown difference between computerized versus hand-charted anesthesia records, the marginal physiologic relevance of MAP within some range of values, the relative unimportance of MAP in a heterogeneous surgical sample when other variables are measured, or may reflect true differences in intraoperative performance among institutions.¹⁵ Further work is underway to address these possibilities. For the SAS and its improved utility with larger sampling intervals, the mechanism for improved specificity is most likely filtering out of uninformative extreme values for HR and less frequent assignment of patients to the highest risk SAS category.

Application of the findings here with real-time data monitoring and real-time risk refinement in a clinical setting might improve the efficiency of postoperative triage across a variety of surgeries, but this has not been tested prospectively. The early identification of both at-risk and low-risk surgical patients is of great utility for patient safety and cost-efficient care.^{4,31} A risk prediction algorithm, including sampling methods for vital signs, might be altered in real time to emphasize sensitivity or specificity, depending on clinical or research needs.^{3,32,33} Given that the majority of surgical patients do not have major postoperative complications and that intensive postoperative monitoring is costly, particularly when unnecessary, any improved specificity without a change in sensitivity (as was the case here with increasing sampling interval) could result in appropriate de-escalation of care for postoperative patients, decreased hospital costs, and no change in outcomes. Future prospective application of the SAS would be required to test the effectiveness of this approach.

Conclusions from this investigation are limited by specific features of this study and the availability of methods to assess clinical applications of risk models. The specific features of this study include the composition of the patient sample, the example of the SAS, and the use of a composite end point. This study was limited to adult patients who received general anesthesia. The study benefited from a large sample size, diversity within noncardiac procedures, and detailed, consistent classification of preoperative and postoperative morbidity within the ACS NSQIP. The SAS is 1 example of a data-filtering method and is subject to idiosyncrasies. Nonetheless, this score is a formal intraoperative risk prediction tool that is well studied and is a strong predictor of postoperative complications or

wellness.^{13,34} It is simple and resistant to large fluctuations due to its 3-part composition and reliance on collapsed categories of risk.¹⁷ First, the score incorporates information from 3 variables, each of which is scored based on categories of values. Therefore, small differences in values do not always correspond to changes in score. In addition, the 10-level score encourages categorization into groups that minimizes spurious variation due to poorly populated risk strata. Finally, sampling strategy does not affect EBL, one of the 3 score components. The inherent robustness of the score and the variability demonstrated here add gravity to our findings that performance varies by sampling method. The differences in SAS by sampling methods illustrated here suggest that the performance of other methods to filter vital signs data may be more susceptible to sampling methodology. This study does not relate to the American Society of Anesthesiologists Standards for Basic Anesthetic Monitoring, which remain the standard for the practice of anesthesiology. The results here are dependent on the regular measurement and recording of vital signs as advocated by the American Society of Anesthesiologists.

In the present study, the outcome of interest was composite and heterogeneous, including major complications or death within 30 days of the index surgical procedure. The number of deaths ($N=37$, 1.2%) was insufficient to apply the investigation to mortality. Heterogeneity among outcomes may explain the limited success of the SAS in terms of sensitivity or increasing risk estimates for those with disease. Despite this limitation, composite outcomes comparable with the one used here are established tools for quality improvement,³⁵ and the present findings demonstrate the utility of the SAS for prediction generally, and more specifically, for identifying patients with the lowest risk of postoperative complication across a variety of sampling strategies.

Any conclusions about optimal sampling strategies or clinical application are limited by the maturity of statistical methods available for determining the clinical utility of a new risk factor.^{25,29,36} In some cases, conclusions about reclassification and discrimination by sampling strategy in multivariate models were limited by a lack of any statistical tests. However, the present study benefits from explicit reporting of multiple methods to assess model utility including univariate and multivariate assessments of strength and significance of association, goodness-of-fit, model calibration, discrimination, and reclassification. These approaches were internally consistent and were consistent with reporting recommendations.²⁹ The methods of net reclassification, although increasingly common in the literature, have not been previously applied to the SAS or to the ASA physical classification status.³⁷ The present findings demonstrated a new appreciation for the mechanisms of sensitivity and specificity that result in improved model discrimination.

In conclusion, concurrent automated reviews of patient data, such as the SAS, might improve detection of at-risk patients as well as identification of patients with very low risk. However, the present findings suggest that all computer-driven data reviews may be susceptible to sampling-based performance variation as well as optimization by sampling methodology. Future work to develop and apply concurrent automated reviews of patient vital signs may benefit from the explicit understanding of the potential influence of varied sampling strategies.

ACKNOWLEDGMENTS

The authors wish to acknowledge Roxanne Hyke and Sharon Nehring for their support with ACS NSQIP data management at Mayo Clinic as well as Andrew Hanson for programming support.

Funding: This work was supported, in part, by a grant from the Division of Critical Care, Mayo Clinic, Rochester, MN.

REFERENCES

1. Jha AK, Classen DC. Getting moving on patient safety—harnessing electronic data for safer care. *N Engl J Med*. 2011; 365:1756–8. [PubMed: 22070474]
2. Kenzaka T, Okayama M, Kuroki S, Fukui M, Yahata S, Hayashi H, Kitao A, Sugiyama D, Kajii E, Hashimoto M. Importance of vital signs to the early diagnosis and severity of sepsis: association between vital signs and sequential organ failure assessment score in patients with sepsis. *Intern Med*. 2012; 51:871–6. [PubMed: 22504241]
3. Herasevich V, Pieper MS, Pulido J, Gajic O. Enrollment into a time sensitive clinical study in the critical care setting: results from computerized septic shock sniffer implementation. *J Am Med Inform Assoc*. 2011; 18:639–44. [PubMed: 21508415]
4. Sobol JB, Wunsch H. Triage of high-risk surgical patients for intensive care. *Crit Care*. 2011; 15:217. [PubMed: 21457500]
5. Sternbach G. Claude Beck: cardiac compression triads. *J Emerg Med*. 1988; 6:417–9. [PubMed: 3066820]
6. Bijker JB, van Klei WA, Kappen TH, van Wolfswinkel L, Moons KG, Kalkman CJ. Incidence of intraoperative hypotension as a function of the chosen definition: literature definitions applied to a retrospective cohort using automated data collection. *Anesthesiology*. 2007; 107:213–20. [PubMed: 17667564]
7. Hermeneit S, Müller M, Terzic A, Rodehorst A, Schamberger M, Böttger T. Influenceable surgical and anesthesiological risk factors for the development of cardiac and pulmonary complications in laparoscopic surgery of the colon. *Zentralbl Chir*. 2008; 133:156–63. [PubMed: 18415903]
8. Pratt W, Callery MP, Vollmer CM Jr. Optimal surgical performance attenuates physiologic risk in high-acuity operations. *J Am Coll Surg*. 2008; 207:717–30. [PubMed: 18954785]
9. Chau A, Ehrenfeld JM. Using real-time clinical decision support to improve performance on perioperative quality and process measures. *Anesthesiol Clin*. 2011; 29:57–69. [PubMed: 21295752]
10. Moore LE, Sharifpour M, Shanks A, Kheterpal S, Tremper KK, Mashour GA. Cerebral perfusion pressure below 60 mm Hg is common in the intraoperative setting. *J Neurosurg Anesthesiol*. 2012; 24:58–62. [PubMed: 21862931]
11. Kheterpal S, O'Reilly M, Englesbe MJ, Rosenberg AL, Shanks AM, Zhang L, Rothman ED, Campbell DA, Tremper KK. Preoperative and intraoperative predictors of cardiac adverse events after general, vascular, and urological surgery. *Anesthesiology*. 2009; 110:58–66. [PubMed: 19104171]
12. Sessler DI, Sigl JC, Kelley SD, Chamoun NG, Manberg PJ, Saager L, Kurz A, Greenwald S. Hospital stay and mortality are increased in patients having a “triple low” of low blood pressure, low bispectral index, and low minimum alveolar concentration of volatile anesthesia. *Anesthesiology*. 2012; 116:1195–203. [PubMed: 22546967]
13. Regenbogen SE, Bordeianou L, Hutter MM, Gawande AA. The intra-operative Surgical Apgar Score predicts postdis-charge complications after colon and rectal resection. *Surgery*. 2010; 148:559–66. [PubMed: 20227100]
14. Regenbogen SE, Ehrenfeld JM, Lipsitz SR, Greenberg CC, Hutter MM, Gawande AA. Utility of the surgical apgar score: validation in 4119 patients. *Arch Surg*. 2009; 144:30–6. [PubMed: 19153322]
15. Regenbogen SE, Lancaster RT, Lipsitz SR, Greenberg CC, Hutter MM, Gawande AA. Does the Surgical Apgar Score measure intraoperative performance? *Ann Surg*. 2008; 248:320–8. [PubMed: 18650644]

16. Regenbogen SE, Bordeianou L, Hutter MM, Gawande AA. The intraoperative Surgical Apgar Score predicts postdis-charge complications after colon and rectal resection. *Surgery*. 2010; 148:559–66. [PubMed: 20227100]
17. Gawande AA, Kwaan MR, Regenbogen SE, Lipsitz SA, Zinner MJ. An Apgar score for surgery. *J Am Coll Surg*. 2007; 204:201–8. [PubMed: 17254923]
18. Fink AS, Campbell DA Jr, Mentzer RM Jr, Henderson WG, Daley J, Bannister J, Hur K, Khuri SF. The National Surgical Quality Improvement Program in non-veterans administration hospitals: initial demonstration of feasibility. *Ann Surg*. 2002; 236:344–53. [PubMed: 12192321]
19. Khuri SF, Daley J, Henderson W, Hur K, Demakis J, Aust JB, Chong V, Fabri PJ, Gibbs JO, Grover F, Hammermeister K, Irvin G 3rd, McDonald G, Passaro E Jr, Phillips L, Scamman F, Spencer J, Strempel JF. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. *Ann Surg*. 1998; 228:491–507. [PubMed: 9790339]
20. Dimick JB, Osborne NH, Hall BL, Ko CY, Birkmeyer JD. Risk adjustment for comparing hospital quality with surgery: how many variables are needed? *J Am Coll Surg*. 2010; 210:503–8. [PubMed: 20347744]
21. Arvidsson S, Ouchterlony J, Nilsson S, Sjöstedt L, Svärdsudd K. The Gothenburg study of perioperative risk. I. Preoperative findings, postoperative complications. *Acta Anaesthesiol Scand*. 1994; 38:679–90. [PubMed: 7839777]
22. Arvidsson S, Ouchterlony J, Sjöstedt L, Svärdsudd K. Predicting postoperative adverse events. Clinical efficiency of four general classification systems. The project perioperative risk. *Acta Anaesthesiol Scand*. 1996; 40:783–91. [PubMed: 8874563]
23. Khuri SF, Daley J, Henderson W, Barbour G, Lowry P, Irvin G, Gibbs J, Grover F, Hammermeister K, Strempel JF. The National Veterans Administration Surgical Risk Study: risk adjustment for the comparative assessment of the quality of surgical care. *J Am Coll Surg*. 1995; 180:519–31. [PubMed: 7749526]
24. Centers for Medicare and Medicaid Services. Available at: <https://www.cms.gov>. Accessed October 21, 2010
25. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010; 21:128–38. [PubMed: 20010215]
26. Kerr KF, McClelland RL, Brown ER, Lumley T. Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *Am J Epidemiol*. 2011; 174:364–74. [PubMed: 21673124]
27. Pencina MJ, D'Agostino RB Sr, Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med*. 2012; 31:101–13. [PubMed: 22147389]
28. Bergstralh, EJ. ROCPLUS macro. 2009. Available at: <http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm>
29. Pepe MS, Janes H. Commentary: reporting standards are needed for evaluations of risk reclassification. *Int J Epidemiol*. 2011; 40:1106–8. [PubMed: 21571811]
30. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med*. 1997; 16:965–80. [PubMed: 9160492]
31. Smith T, Den Hartog D, Moerman T, Patka P, Van Lieshout EM, Schep NW. Accuracy of an expanded early warning score for patients in general and trauma surgery wards. *Br J Surg*. 2012; 99:192–7. [PubMed: 22183685]
32. Cannon CM, Braxton CC, Kling-Smith M, Mahnken JD, Carlton E, Moncure M. Utility of the shock index in predicting mortality in traumatically injured patients. *J Trauma*. 2009; 67:1426–30. [PubMed: 20009697]
33. Cao H, Eshelman L, Chbat N, Nielsen L, Gross B, Saeed M. Predicting ICU hemodynamic instability using continuous multiparameter trends. *Conf Proc IEEE Eng Med Biol Soc*. 2008; 2008:3803–6. [PubMed: 19163540]

34. Chandra A, Mangam S, Marzouk D. A review of risk scoring systems utilised in patients undergoing gastrointestinal surgery. *J Gastrointest Surg.* 2009; 13:1529–38. [PubMed: 19319612]
35. Hall BL, Hamilton BH, Richards K, Bilimoria KY, Cohen ME, Ko CY. Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals. *Ann Surg.* 2009; 250:363–76. [PubMed: 19644350]
36. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. *Am J Epidemiol.* 2011; 173:1327–35. [PubMed: 21555714]
37. Ioannidis JP, Tzoulaki I. What makes a good predictor?: the evidence applied to coronary artery calcium score. *JAMA.* 2010; 303:1646–7. [PubMed: 20424257]

Figure 1. Within-patient averages for surgical Apgar score and its components, displayed by sampling method. Error bars display the SEM. For all scores, values for each sampling method are significantly different from the value for instantaneous sampling.

Table 1

The Surgical Apgar Score

Components	Surgical Apgar score, no. of points				
	0	1	2	3	4
Estimated blood loss, mL	>1000	601–1000	101–600	100	
Lowest mean arterial blood pressure, mm Hg	<40	40–54	55–69	70	
Lowest heart rate, beats per minute	>85	76–85	66–75	56–65	55

Table 2

Description of Patients and Procedures by Outcome

	Total, N (%)	No Event, N (%)	Major complication, N (%)	P
Total surgical characteristics	3000	2727 (90.9)	273 (9.1)	
Procedure type				<0.0001
Minor or intermediate	1091 (36.4)	1069 (39.2)	22 (8.1)	
Simple alimentary	293	284	9	
Breast	318	313	5	
Groin or umbilical hernia repair	121	119	2	
Skin, soft tissue, or bone	311	311	0	
Endovascular	48	45	1	
Major	1909 (63.6)	1658 (60.8)	251 (91.9)	
Complex alimentary, prostatectomy	1063	944	119	
Open vascular or any AAA	365	288	77	
Hepatobiliary or pancreas	250	208	42	
Bariatric	13	13	0	
Complex hernia, VATS, thoracotomy	178	167	11	
Skin, soft tissue or bone	40	38	2	
Wound classification (Contaminated or dirty)	328 (10.9)	266 (9.76)	62 (22.7)	<0.0001
Priority/emergency case	241 (8.0)	188 (6.9)	53 (19.4)	<0.0001
Patient characteristics				
Age, years (mean, SD)	58.1 (16.3)	57.7 (16.2)	63.9 (15.8)	<0.0001
Body mass index, kg/m ² (mean, SD)	29.2 (8.7)	29.1 (8.7)	29.5 (9.5)	0.5989
Age older than 65 y	1125 (37.5%)	982 (36.0)	143 (52.4)	<0.0001
Female sex	1678 (55.9%)	1533 (56.2)	145 (53.1)	0.5722
ASA physical status classification				<0.0001
I	164 (5.5)	160 (5.9)	4 (1.5)	
II	1508 (50.3)	1439 (52.8)	69 (25.3)	
III	1238 (41.3)	1077 (39.5)	161 (59.0)	
IV	90 (3.0)	51 (1.9)	39 (14.3)	
Functional status (not independent)	148 (4.9)	85 (3.1)	63 (23.2)	<0.0001
Weight loss	91 (3.0)	80 (2.9)	11 (4.0)	0.3142
Ascites	22 (0.7)	14 (0.5)	8 (2.9)	<0.0001
Dialysis	26 (0.9)	18 (0.7)	8 (2.9)	<0.0001
Dyspnea (any)	437 (14.7)	419 (13.6)	18 (23.8)	<0.0001
Congestive heart failure	20 (0.7)	13 (0.5)	7 (2.6)	<0.0001
Bleeding disorder	106 (3.5)	82 (3.0)	24 (8.8)	<0.0001

AAA = abdominal aortic aneurysm; VATS = video-assisted thoracoscopic surgery.

Table 3

Distributions and Univariate Associations of Surgical Apgar Score with Major Postoperative Complication by Sampling Method

Sampling method	Surgical Apgar score			
	0–4	5–6	7–8	9–10
Event/total (%)				
Instantaneous	33/119 (27.7)	80/554 (14.4)	125/1625 (7.7)	35/702 (5.0)
5-min overlap	34/112 (30.4)	82/519 (15.8)	121/1597 (7.6)	36/772 (4.7)
5-min step	34/107 (31.8)	83/515 (16.1)	120/1577 (7.6)	36/801 (4.5)
10-min overlap	34/107 (31.8)	80/489 (16.4)	124/1576 (7.9)	35/828 (4.2)
10-min step	32/97 (33.0)	87/488 (17.8)	114/1559 (7.3)	40/856 (4.7)
Unadjusted odds ratios (95% confidence interval)				
Instantaneous $c = 0.631$	7.3 (4.3–12.4)	3.2 (2.3–4.9)	1.6 (1.1–2.3)	1 (reference)
5-min overlap $c = 0.648$	8.9 (5.3–15.0)	3.8 (2.5–5.8)	1.7 (1.1–2.5)	1 (reference)
5-min step $c = 0.654$	9.9 (5.8–16.8)	4.1 (2.7–6.1)	1.8 (1.2–2.6)	1 (reference)
10-min overlap $c = 0.658$	10.6 (6.2–18.0)	4.4 (2.9–6.7)	1.9 (1.3–2.8)	1 (reference)
10-min step $c = 0.661$	10.0 (5.9–17.0)	4.4 (3.0–6.6)	1.6 (1.1–2.3)	1 (reference)
Adjusted odds ratios (95% confidence interval) ^a				
Instantaneous	3.1 (1.7–5.6)	2.0 (1.3–3.2)	1.5 (1.0–2.2)	1 (reference)
5-min overlap	3.9 (2.2–6.9)	2.3 (1.5–3.5)	1.5 (1.0–2.2)	1 (reference)
5-min step	4.3 (2.4–7.7)	2.4 (2.4–7.7)	1.5 (1.0–2.3)	1 (reference)
10-min overlap	4.5 (2.5–8.1)	2.5 (1.6–3.9)	1.7 (1.1–2.5)	1 (reference)
10-min step ^b	4.2 (2.3–7.5)	2.5 (1.6–3.8)	1.4 (1.0–2.1)	1 (reference)

C is the c-statistic. Surgical Apgar score is significantly associated with event for all models ($P < 0.0001$).

^a Odds ratio adjusted for ASA physical status classification (I, II, III, or IV), major surgery, emergency surgery, age, sex, wound classification, functional status before surgery, ascites, dialysis, dyspnea, congestive heart failure, and bleeding.

^b Hosmer–Lemeshow goodness-of-fit test $P < 0.10$ for this model.

Table 4

Incremental Performance of Different Constructions of the Surgical Apgar Score for Predicting Major Complication

		NONEVENTS			EVENTS				
		% reclassified by risk			% reclassified by risk				
	c-statistic (95% CI)	<0.05	0.5–0.13	>0.13	<0.05	0.5–0.13	>0.13	NRI ^a (P)	IDI ^a (P)
Perfect new variable	1.000	+62.0	−45.8	−16.2	−5.5	−42.5	+48.0	Positive	Positive
Preoperative risk model ^b	c = 0.775								
+Instantaneous	+ 0.012 (0.001-0.022)	+4.4	−4.5	+0.1	+1.5	+0.7	−2.2	0.0070 (0.7856)	0.0083 (0.0044)
+5-min rolling	+ 0.017 (0.005-0.029)	+5.4	−4.3	−1.1	+2.6	−0.4	−2.2	0.0173 (0.5600)	0.0123 (0.0006)
+5-min step	+ 0.019 (0.006-0.032)	+5.9	−6.4	+0.5	+2.6	−3.7	+1.1	0.0393 (0.1748)	0.0138 (0.0003)
+10-min rolling	+0.020 (0.006-0.032)	+6.8	−8.2	+1.4	+1.8	−4.4	+2.6	0.0609 (0.0216)	0.0138 (0.0002)
+10-min step	+0.021 (0.008-0.034)	+10.0	−8.3	−1.7	+3.3	−1.1	−2.2	0.0617 (0.0616)	0.0145 (0.0002)
Null risk model ^c	c = 0.756								
+ASA physical status classification	+ 0.019 (0.004-0.033)	−0.4	−2.6	+3.0	−1.1	−8.1	+9.2	0.0688 (0.0111)	0.0121 (0.0013)
+10-min rolling	+ 0.019 (0.007-0.038)	+6.7	−8.2	+1.4	+2.6	−6.6	+4.0	0.0678 (0.0175)	0.0165 (<0.0001)

The “perfect new variable” would result in perfect model discrimination (c = 1.000) and reclassification of all nonevents to the lowest risk group as well as reclassification of all events to the highest risk group. All differences in c-statistic from baseline models were statistically significant ($P < 0.05$).

^aNRI is net reclassification improvement; and IDI is integrated discrimination improvement.

^bPreoperative model Includes ASA physical status classification (I, II, III, or IV), major surgery, emergency surgery, age, sex, wound classification, functional status before surgery, ascites, dialysis, dyspnea, congestive heart failure, and bleeding.

^cincludes major surgery, emergency surgery, age, sex, wound classification, functional status before surgery, ascites, dialysis, dyspnea, congestive heart failure and bleeding. In these models, ASA physical status classification was replaced by surgical Apgar score using the 10-minute median rolling sampling method.

Table 5

Conventional Performance and 95% Confidence Intervals for the Surgical Apgar Score as a Single Test by Sampling Method

Sampling method	PPV%, (95% CI)	NPV%, (95% CI)	Sensitivity%, (95% CI)	Specificity%, (95% CI)	Accuracy%
Instantaneous	16.8 (14.0–19.8)	93.1 (92.0–94.1)	41.4 (35.5–47.5)	79.5 (78.0–81.0)	76.0 (74.5–77.5)
5-min overlap	18.4 (15.4–21.6)	93.4 (92.3–94.3)	42.5 (36.6–48.6)	81.1 (79.6–82.6)	77.6 (76.1–79.1)
5-min step	18.8 (15.8–22.1)	93.4 (92.4–94.4)	42.9 (36.9–49.0)	81.5 (80.0–82.9)	78.0 (76.4–79.4)
10-min overlap	19.1 (16.1–22.5)	93.4 (92.3–94.4)	41.8 (35.8–47.9)	82.3 (80.8–83.7)	77.9 (76.3–79.3)
10-min step ^a	20.3 (17.2–23.8)	93.6 (92.6–94.6)	43.4 (37.5–49.5)	82.9 (81.5–84.3)	79.3 (77.8–80.8)
<i>P</i> -value for trend	0.1090	0.5316	0.6985	0.0006	0.0013

A positive value for the surgical Apgar score is a score of 6 or less. 95% confidence intervals are by exact method. Cochran–Armitage trend test is reported with an asymptotic 2-sided value.

PPV = positive predictive value; NPV = negative predictive value.