S A E M Academic Emergency Medicine
Official Journal of the Society for Academic Emergency Medicine

ORIGINAL CONTRIBUTION

Are Mortality and Acute Morbidity in Patients Presenting With Nonspecific Complaints Predictable Using Routine Variables?

Mirjam A. Jenny, PhD, Ralph Hertwig, PhD, Selina Ackermann, PhD, Anna S. Messmer, MD, Julia Karakoumis, MD, Christian H. Nickel, MD, and Roland Bingisser, MD

Abstract

Objectives: Patients presenting to the emergency department (ED) with nonspecific complaints are difficult to accurately triage, risk stratify, and diagnose. This can delay appropriate treatment. The extent to which key medical outcomes are at all predictable in these patients, and which (if any) predictors are useful, has previously been unclear. To investigate these questions, we tested an array of statistical and machine learning models in a large group of patients and estimated the predictability of mortality (which occurred in 6.6% of our sample of patients), acute morbidity (58%), and presence of acute infectious disease (28.2%).

Methods: To investigate whether the best available tools can predict the three key outcomes, we fed data from a sample of 1,278 ED patients with nonspecific complaints into 17 state-of-the-art statistical and machine learning models. The patient sample stems from a diagnostic multicenter study with prospective 30-day follow-up conducted in Switzerland. Predictability of the three key medical outcomes was quantified by computing the area under the receiver operating characteristic curve (AUC) for each model.

Results: The models performed at different levels but, on average, the predictability of the target outcomes ranged between 0.71 and 0.82. The better models clearly outperformed physicians' intuitive judgments of how ill patients looked (AUC = 0.67 for mortality, 0.65 for morbidity, and 0.60 for infectious disease).

Conclusions: Modeling techniques can be used to derive formalized models that, on average, predict the outcomes of mortality, acute morbidity, and acute infectious disease in patients with nonspecific complaints with a level of accuracy far beyond chance. The models also predicted these outcomes more accurately than did physicians' intuitive judgments of how ill the patients look; however, the latter was among the small set of best predictors for mortality and acute morbidity. These results lay the groundwork for further refining triage and risk stratification tools for patients with nonspecific complaints. More research, informed by whether the goal of a model is high sensitivity or high specificity, is needed to develop readily applicable clinical decision support tools (e.g., decision trees) that could be supported by electronic health records.

ACADEMIC EMERGENCY MEDICINE 2015;22:1155–1163 © 2015 by the Society for Academic Emergency Medicine

Emergency physicians (EPs) frequently encounter patients with nonspecific complaints. These patients tend to report general feelings of weakness, discomfort, fatigue, or dizziness, but not more specific complaints.[1] They are often undertriaged (i.e., their initial risk assessments are too low), and the severity of the illness causing their nonspecific complaints is often misjudged.[2] These systematic misjudgments can result in delays in definitive treatment of the underlying cause of their nonspecific chief complaints.

Improved triage, rapid diagnosis, timely treatment, and appropriate disposition decisions to the properly

From the Center for Adaptive Rationality, Max Planck Institute for Human Development (MAJ, RH), Berlin-Brandenburg, Germany; the Department of Emergency Medicine, University Hospital Basel (SA, JK, CHN, RB), Basel-Stadt, Switzerland; and the Division of Trauma, Emergency, and Acute Medicine, King's College Hospital, Denmark Hill (ASM), London, UK.

Received December 17, 2014; revision received May 25, 2015; accepted May 27, 2015.

MAJ was supported in part by the German Academic Exchange Service. CHN has received honoraria from ThermoFisher. The rest of the authors have no conflicts or disclosures to report.

Supervising Editor: Gary Gaddis, MD, PhD.

Address for correspondence and reprints: Mirjam A. Jenny, PhD; e-mail: jenny@mpib-berlin.mpg.de.

doi: 10.1111/acem.12755
ISSN 1069-6563
PII ISSN 1069-6563583

resourced site are therefore imperative for patients presenting with nonspecific complaints. It is reasonable to expect that these patients' clinical outcomes could be improved by increased diagnostic and therapeutic accuracy and by better outcome prediction. However, it remains an open question whether and to what extent these patients' key medical outcomes are at all predictable. Given that patients with nonspecific chief complaints account for 5% to 20% of nontrauma emergency presentations in university hospitals, the relevance of this issue is clear.[1–4]

The capacity to predict key medical outcomes in patients with nonspecific complaints is desirable for two reasons. First, triage is designed to stratify patients according to their risk.[5] Patients who present with low risk of acute morbidity and mortality are more likely than those presenting with high risk to be able to wait safely until resources become available. Despite several validated triage tools, elderly patients and patients with nonspecific complaints tend to be undertriaged. Further, there is considerable overlap between triage categories with respect to outcomes, and patients categorized into the same category can have different medical outcomes.[2] Second, disposition after a concise work-up should be based on objective criteria, such as the resource needs of the patient, and the risk of short-term mortality and acute morbidity. The question of which readily available predictors can improve the risk stratification of these patients and inform subsequent disposition decisions remains to be resolved.

Our goal was to estimate the predictability of short-term mortality, acute morbidity, and presence of infectious disease in patients presenting to emergency departments (EDs) with nonspecific complaints by determining the accuracy of different predictive models, including machine learning models. The latter permit the use of nonlinear modeling techniques. Because it is impossible to know a priori which model will perform best in a given data set, we tested an array of models.[6] This approach offers the additional benefit of allowing us to estimate the average predictability of key medical outcomes in patients with nonspecific complaints, thus rendering our general conclusions independent of specific prediction methods. This analysis is clinically relevant in that it lays the groundwork for understanding whether it is possible to predict key medical outcomes in patients with nonspecific complaints, by identifying promising candidate variables and determining how their influences should be quantified and grouped.

METHODS

Study Design

This was a retrospective cohort analysis using data from patients prospectively recruited to the Basel Nonspecific Complaints (BANC) study. The study protocol was approved by the local ethics committee and preregistered (ClinicalTrials.gov identifier: NCT00920491). Each participating patient signed an informed consent form at the time of study enrollment.

Study Setting and Population

Identification, screening, and enrollment of the patients is described elsewhere.[1] Briefly, all patients with Emergency Severity Index assignments of 2 or 3 presenting with nonspecific complaints (e.g., general weakness, fatigue, dizziness, or other chief complaints as defined in the BANC study[3]) to three centers in Switzerland (the University Hospital in Basel, the Cantonal Hospital in Liestal, and the Cantonal Hospital in Aarau) were prospectively included. Nonspecific complaints are all complaints that are not part of the set of specific complaints for which evidence-based protocols exist for emergency care and are complaints for which a highly probable early working diagnosis cannot be clearly established. In the original publication,[1] the data set included 218 patients from the University Hospital in Basel. We have subsequently enrolled more patients: at the time of our analysis, the BANC database contained data from 1,278 patients in the three study centers. Figure 1 and Figure 2 provide more detail on how subjects were identified and selected.

Study Protocol

Data collection is described in detail elsewhere.[1] Numerous candidate variables for our modeling processes were assessed, complemented by baseline predictors such as age, sex, and comorbidity. In addition, we obtained a measure of the physician's first overall impression of each patient, the Gestalt-like impression of "how ill the patient looks." Details from the patient's history, physical examination, and standard laboratory examinations were also used as potential predictor variables.

The candidate variables were chosen by experienced clinicians (certified in internal medicine and emergency medicine and with at least 10 years' clinical experience) in a modified broadband Delphi process. Specifically, the clinicians selected predictors they considered potentially valuable in predicting three key medical outcomes: 30-day mortality, morbidity, and presence of an infectious disease. The leader of the project prepared a list of all predictors to be further considered by the team. This list was distributed among a group of experienced EPs. Each of them assessed the potential value of all predictors by giving a binary judgment (yes/no) in a spreadsheet. The results of these judgments were gathered by the leader and distributed again, individually to the experienced EPs. A second round of individual judgments by these experts was followed by a group meeting with the goal of finalizing the list of potential predictor variables. The ED resident physicians who participated in the subsequent data collection were trained (lecture as well as on-site training) in how to follow the BANC protocol. Their clinical experience was not recorded individually, but ranged between 1 and 4 years.

In addition to conventional predictors, such as vital signs, history, and laboratory values, one novel predictor, "looking ill," was constructed for the BANC cohort. Data for this predictor were collected at the very outset of each physician–patient interaction, with physicians rating how ill each patient looked on a scale from 0 (patient looks to be in excellent health) to 100 (patient looks

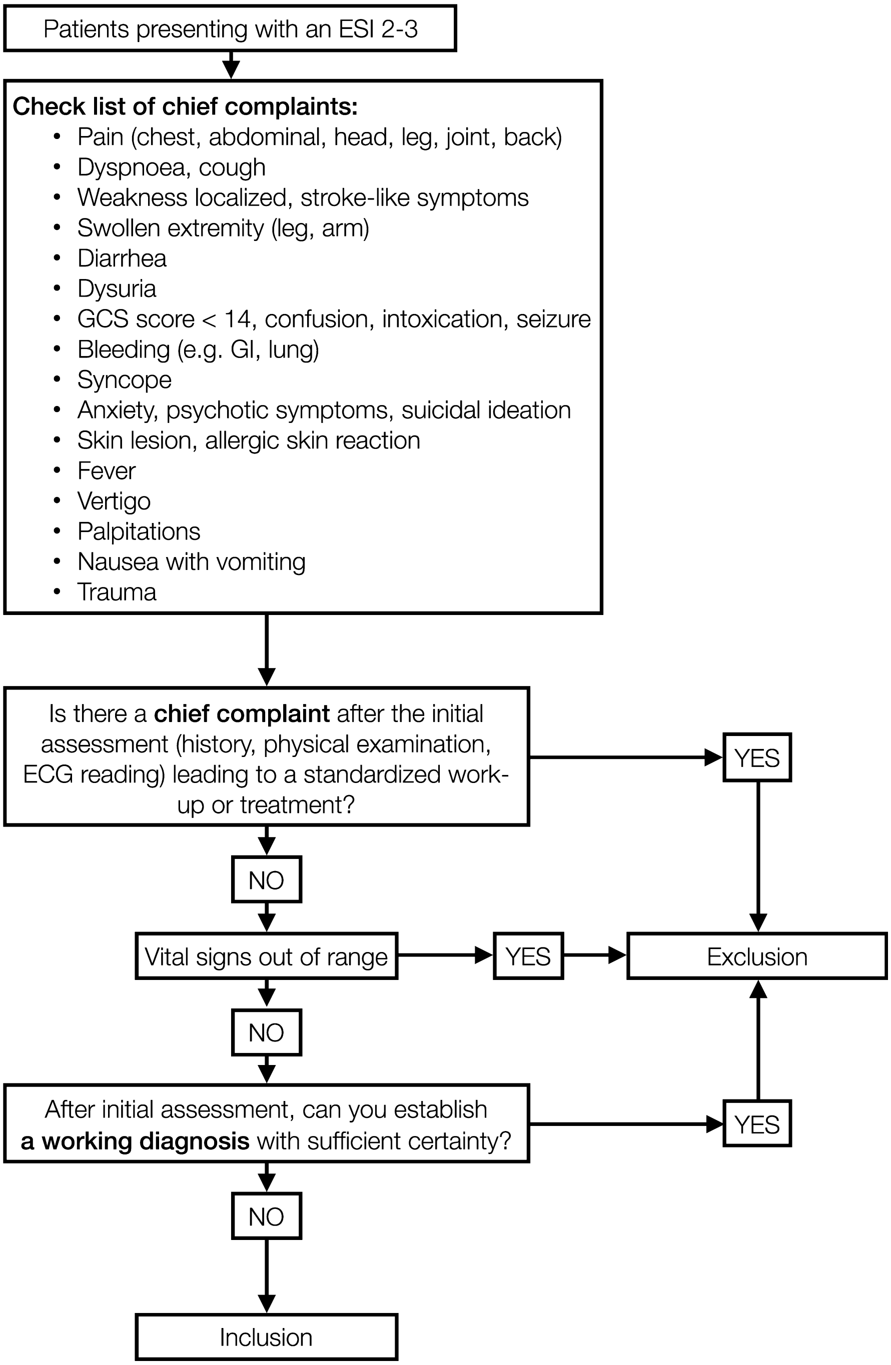
Patients presenting with an ESI 2-3
Check list of chief complaints:
• Pain (chest, abdominal, head, leg, joint, back)
• Dyspnoea, cough
• Weakness localized, stroke-like symptoms
• Swollen extremity (leg, arm)
• Diarrhea
• Dysuria
• GCS score < 14, confusion, intoxication, seizure
• Bleeding (e.g. GI, lung)
• Syncope
• Anxiety, psychotic symptoms, suicidal ideation
• Skin lesion, allergic skin reaction
• Fever
• Vertigo
• Palpitations
• Nausea with vomiting
• Trauma
Is there a chief complaint after the initial assessment (history, physical examination, ECG reading) leading to a standardized work-up or treatment?
YES
NO
Vital signs out of range
YES
Exclusion
NO
After initial assessment, can you establish a working diagnosis with sufficient certainty?
YES
NO
Inclusion

Figure 1. Identification of patients with nonspecific complaints in the BANC study.
ECG = electrocardiogram; ESI = Emergency Severity Index; GCS = Glasgow Coma Scale; GI = gastrointestinal. Figure adapted from Acad Emerg Med 2010;17:284-92.

extremely ill). The predictor's construct validity was tested; it showed a good linear relationship with acute morbidity (Spearman rank correlation coefficient $\rho = 0.96$) and a higher-than-chance median predictability for acute morbidity (area under the receiver operating characteristic [ROC] curve [AUC] = 0.65; interquartile range [IQR] = 0.64 to 0.67). External validation using a second sample (all patients presenting to the ED of the University Hospital of Basel) also indicated that the predictor showed a good linear relationship with acute morbidity ($\rho = 0.97$, $n = 1{,}196$), fair median predictability for acute morbidity (AUC = 0.72; IQR = 0.70 to 0.73), and good inter-rater reliability between physicians and nurses (intraclass correlation using linear mixed-effects models = 0.49).[7]

Outcomes

Our analysis focused on three medical outcomes as defined by the BANC framework[1]: short-term mortality (death occurring within 30 days of initial presentation to the ED), acute morbidity (any potentially life-threatening condition or any condition that requires immediate

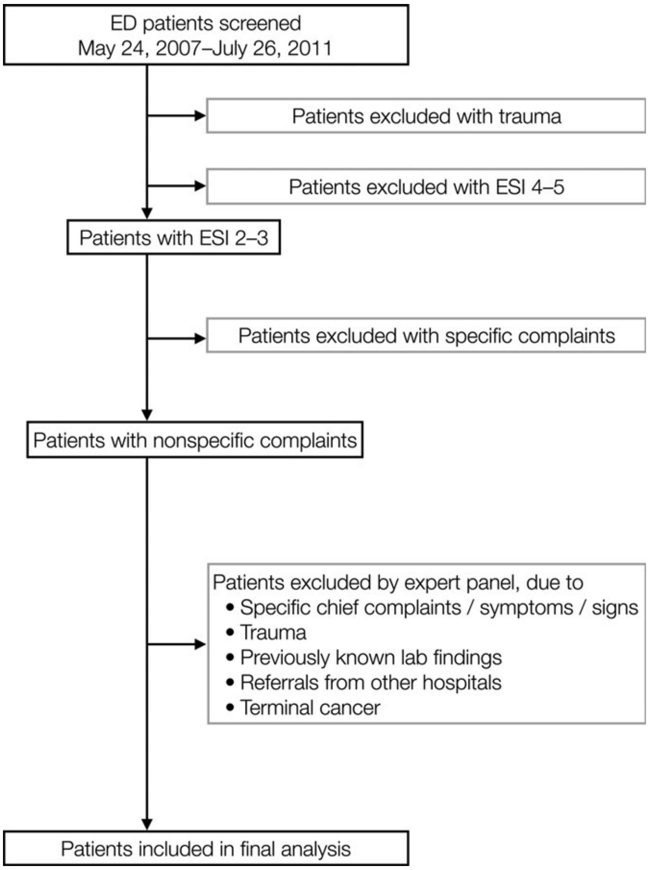
ED patients screened
May 24, 2007–July 26, 2011
Patients excluded with trauma
Patients excluded with ESI 4–5
Patients with ESI 2–3
Patients excluded with specific complaints
Patients with nonspecific complaints
Patients excluded by expert panel, due to
• Specific chief complaints / symptoms / signs
• Trauma
• Previously known lab findings
• Referrals from other hospitals
• Terminal cancer
Patients included in final analysis

Figure 2. Study flow of the BANC study. ESI = Emergency Severity Index. Figure adapted from Acad Emerg Med 2010;17:284-92.

intervention to prevent serious morbidity, permanent disability, or death), and presence of an acute infectious disease needing immediate treatment. This third outcome was chosen due to its high occurrence in the BANC cohort and the need to identify infection at an early stage to prevent poor outcomes due to delayed treatment of infectious sequelae (e.g., severe sepsis and septic shock). Written 30-day follow-up data were obtained from the patients' primary care physicians, together with hospital discharge reports for all hospitalized patients. When hospitalization outlasted 30 days, we obtained interim hospital records. Two physicians certified in internal medicine who were blinded to the patients' baseline data reviewed the final reports to determine the outcomes.

Data Analysis

We used R version 3.0.2 for data analysis, making use of several R-packages (e.g., caret[8] and party[9]). Of the 213 selected potential predictor variables, 87 were neither highly intercorrelated with other predictors, nor showed zero variance, nor had ≥10% missing values. The statistical and machine learning models we tested differ in the type of statistical relationship they assume between predictor and outcome and in their sensitivity to between-predictor correlations. The intercorrelations were therefore determined by a heuristic algorithm, available in the R-package caret, which defines a certain threshold (in our case, a Pearson correlation coefficient of 0.75) and ensures that no pairwise correlation exceeds this threshold.[8] Because they carry little or no information, continuous variables with zero or near-zero variance were omitted. To this end, we applied another statistical rule of thumb contained in the caret package, keeping the default settings.[8] The missing values for the predictors with < 10% missing values were imputed using random forests (we used the default settings of the missForest package).[10]

We tested whether each of the 87 potential predictors had a statistically significant relationship with the respective predefined outcomes using Bonferroni-corrected t-tests for continuous variables and Fisher's exact tests for categorical variables (both with a p-value cutoff of 0.05; see Figure 3). To estimate the predictability of the key medical outcomes, we fed the predictors into 17 statistical and machine learning models for each outcome (Figure 4). The models were fitted, cross-validated, and tested on the data. Note that the development of a single best model for each condition would depend on whether the aim was to maximize sensitivity or specificity. Our goal was not to develop a single model for each condition, but to identify a set of predictive variables and to maximize their combined predictive power (AUC) by testing multiple models.

The modeling scheme tested an array of structurally different models, making it possible to detect linear, nonlinear, and also nonmonotonic relationships (e.g., by random forests). Where necessary, variables were transformed using the R-package caret (centered and scaled for the regression models, the linear discriminant analyses, nearest shrunken centroids, k-nearest neighbors, and neural networks, for which the spatial sign transformation was additionally applied). To avoid overfitting, we used a state-of-the art cross-validation procedure: the models' tuning parameters were first determined using a 10-fold cross-validation on two-thirds of the data while applying the "one-standard-error" method.[8,11] The

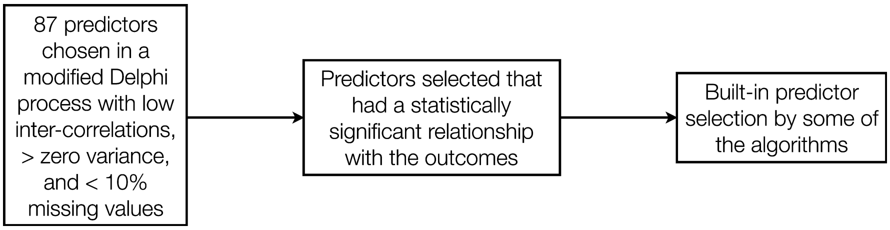
87 predictors chosen in a modified Delphi process with low inter-correlations, > zero variance, and < 10% missing values
Predictors selected that had a statistically significant relationship with the outcomes
Built-in predictor selection by some of the algorithms

Figure 3. Selection of the Basel Non-specific Complaints Study predictors for the predictability analysis.

Model	Implementation
Logistic regression	glm from stats and stepAIC from MASS
Penalized logistic regression	glmnet from glmnet
Linear discriminant analysis	lda from MASS
Partial least squares discriminant analysis	plsda from caret
Nearest shrunken centroids	pamr.train from pamr
Nonlinear discriminant analysis	mda from mda
Neural networks	nnet from nnet
Flexible discriminant analysis	fda from mda
Support vector machines	svm from e1071
K-nearest neighbors	knn3 from caret
Naïve Bayes	NaiveBayes from e1071
Classification and regression trees	rpart from rpart
J48 algorithm	J48 from RWeka
PART rule	PART from RWeka
Bagged trees	bagging from ipred
Random forests	cforest from party
Boosted trees	gbm from gbm

Figure 4. The 17 statistical and machine learning models with their specific functions and R-packages used to predict 30-day mortality, acute morbidity, and presence of acute infectious disease.

tuning parameters were selected to maximize the AUCs. Subsequently, we used the same two-thirds of the data to fit the remaining parameters. Finally, the models derived from the two-thirds of the data underwent validation using the last third of the data ("holdout" set). The mean (median) predictive power (AUC) of all models was used as a robust estimation of the predictability of each criterion.

RESULTS

Baseline characteristics for 1,278 patients from the BANC cohort were analyzed and are reported in Table 1. Eighty-four (6.6%) patients died during the 30-day follow-up, 742 (58.0%) were classified as suffering from acute morbidity, and 360 (28.2%) from infectious conditions.

For mortality (6.6% of patients), the data were upsampled[8] to account for the low base rate of mortality and the resulting class imbalance. On the basis of the upsampled data, Bonferroni-corrected t-tests and Fisher's exact tests selected 41 of the original 87 predictors as potential predictors in the statistical and machine learning models (Table 2). Figure 5A shows the performances of all models for mortality in the holdout set. The median predictability (AUC) was 0.82 (IQR = 0.77 to 0.85). Because the penalized version of logistic regression (elastic net), which is less prone to overfitting, did not show better predictive power than the version without penalty, the former was excluded from the analysis to reduce redundancy. With an AUC of 0.86, flexible discriminant analysis[8,12] showed the best performance. This model allows for the detection of nonlinear relationships between a combination of continuous independent predictors and a criterion.

For acute morbidity (58.0% of patients), 14 predictors were selected as useful on the basis of the Bonferroni-corrected t-tests and Fisher's exact tests (Table 2). The median predictability (AUC) was 0.80 (IQR = 0.80–0.81; see Figure 5B). Again, penalized logistic regression did not clearly outperform its nonpenalized counterpart and was therefore omitted to reduce redundancy. With an AUC of 0.82, random forests[9,13] showed the best performance. This model is able to detect nonlinear and nonmonotonic relationships by producing a large number of decision trees and outputting the majority prediction of those trees—in other words, by producing a committee prediction.

Table 1
Baseline Characteristics of the Basel Non-specific Complaints Study Population (N = 1,278).

Characteristics	Summary Distribution	
Number of patients, N (%)		
All	1,278	
Male	496	(38.8)
Female	782	(61.2)
Age (yr), median (IQR)		
All	81	(74–87)
Male	79	(70–85)
Female	83	(76–88)
Emergency Severity Index, N (%)		
Not available, direct boarders	110	(8.6)
3	1,116	(87.3)
2	52	(4.1)
Charlson Comorbidity Index, not age adjusted, median (IQR)	2	(1–3)
Number of concomitant drugs, median (IQR)	5	(3–8)

Table 2
p-values of the predictors entering the models to predict key medical outcomes (derived using Bonferroni-corrected t-tests and Fisher's exact tests[8]).

Predictor	Mortality	Acute Morbidity	Infectious Disease
Katz activities of daily living (ADLs)			
Acute worsening of ADLs	0.00000^{+}		
Immobility	0.00000^{+}		
Assisted eating	0.00000^{+}		
Incontinence	0.00001^{+}		
Toilet hygiene	0.00000^{+}		
Personal hygiene and grooming	0.00000^{+}		
Laboratory			
C-reactive protein	0.00000^{+}	0.00000^{+}	0.00000^{+}
Serum urea (BUN)	0.00000^{+}	0.00000^{+}	0.00048^{+}
Creatinine	0.00000^{+}	0.00000^{+}	
Phosphate	0.00000^{+}	0.00000^{+}	
Leukocytosis in urinalysis		0.00000^{+}	0.00030^{+}
Albumin	0.00000^{-}	0.00000^{-}	0.00000^{-}
Hemoglobin	0.00000^{-}	0.00000^{-}	
Potassium	0.00000^{+}		
Calcium	0.00000^{-}		
Sodium	0.00022^{+}	0.00000^{-}	
Medication			
ACE inhibitors, AR blockers	0.00013^{+}		
Corticosteroids	0.00011^{+}		
Other antihypertensive drugs	0.00020^{+}		
Number of drugs	0.00042^{+}		
Comorbidities			
Charlson Comorbidity Index (CCI)	0.00000^{+}	0.00001^{+}	
Congestive heart failure	0.00000^{+}	0.00002^{+}	
Tumor (any solid, nonmetastatic)	0.00000^{+}		
Moderate to severe renal disease	0.00000^{+}	0.00000^{+}	
Heart disease (coronary, hypertensive, valvular)		0.00024^{+}	
Psychiatric conditions	0.00000^{-}		
Connective tissue disease	0.00011^{-}		
Physical examination			
Heart rate	0.00009^{+}		
Respiratory rate	0.00000^{+}	0.00012^{+}	
Central venous pressure (low)	0.00000^{+}		
Congested neck veins	0.00000^{+}		
Rales	0.00000^{+}		
Glasgow Coma Scale	0.00000^{-}		
Systolic blood pressure	0.00000^{-}		
Clinical signs			
Looking ill	0.00000^{+}	0.00000^{+}	0.00001^{+}
Sex (male)	0.00000^{+}		
Age	0.00002^{+}		
Clock drawing test (points)	0.00000^{+}		
Adequate quality of patient history	0.00000^{-}		
Body mass index	0.00000^{-}		
Alcohol consumption	0.00052^{-}		
Complaints			
Inappetence	0.00000^{+}		
Dizziness	0.00002^{-}		

+Positive/–negative association between predictor and outcome (higher/lower mean predictor value for patients with the outcome).

For acute infectious disease (28.2% of patients), five predictors were identified as useful on the basis of Bonferroni-corrected t-tests and Fisher's exact tests (Table 2). The median predictability (AUC) was 0.71 (IQR = 0.69 to 0.72; see Figure 5C). Because the penalized version of logistic regression did not show clearly superior predictive power to the version without penalty, it was again excluded from the analysis. With an AUC of 0.73, naïve Bayes[8] showed the best performance. Naïve Bayes uses the Bayes rule to compute the probability that a case belongs to a specific class based on its predictor values, but assumes that the predictors are all independent.

Which were the best *predictors* of the three key medical outcomes? The random forests model,[9] which was among the best in predicting mortality and acute morbidity and very close to the best model in predicting infectious disease, yields an easily interpretable importance measure for each variable based on AUCs.[14,15] We were therefore able to determine the best predictors by identifying the predictors with the highest variable importance. Specifically, we ordered the predictors in descending order of importance and identified the best few predictors before their importance decreased sharply. For mortality, these were (in descending order) albumin, respiratory rate, Charlson Comorbidity Index (CCI), and "looking ill"; for acute morbidity, they were sodium, serum urea (BUN), C-reactive protein, leukocytosis in urinalysis, and "looking ill"; and for infectious disease, C-reactive protein. How well were each of these predictors able to independently predict the medical outcomes? For mortality, the AUCs were 0.78 for albumin, 0.69 for respiratory rate and 0.67 for CCI and "looking ill." For acute morbidity, the AUCs were 0.58 for sodium, 0.67 for BUN and C-reactive protein, 0.63 for leukocytosis in urinalysis, and 0.65 for "looking ill." Finally, for infectious disease, the AUC was 0.75 for C-reactive protein.

We found fair to good predictability of key medical outcomes (median AUC = 0.71 to 0.82) in patients with nonspecific complaints, complementing the finding that physicians' diagnostic accuracy for nonspecific complaints is better than chance.[16] Moreover, the better models clearly outperformed the physicians' Gestalt-like judgments (AUC = 0.65 for morbidity, 0.67 for mortality, and 0.60 for infectious disease) as well as all individual predictors. Key overall predictors were C-reactive protein and "looking ill," which were among the best predictors for two of the three outcomes. These findings confirm that key medical outcomes are indeed predictable in patients with nonspecific complaints. We can thus proceed to developing triage and risk stratification tools that are tailored to this population.

DISCUSSION

The major finding of our investigation is that modeling techniques can be used to derive formalized models that, on average, predict the outcomes of mortality, acute morbidity, and acute infectious disease in patients with nonspecific complaints with a level of accuracy that clearly exceeds both chance and physicians' intuitive judgment of how ill patients look. In clinical practice, it will be possible to capitalize on this better-than-chance

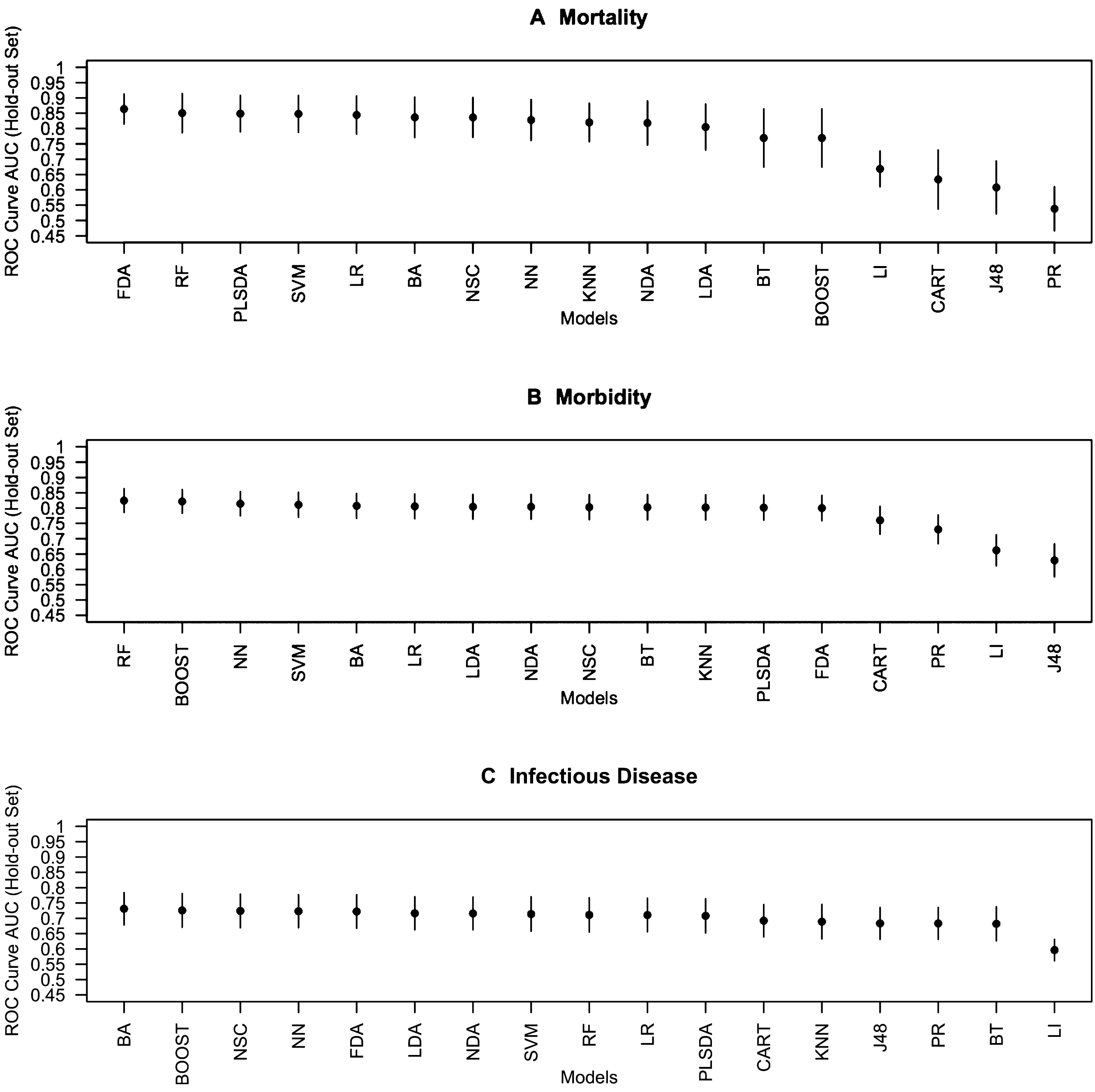

Figure 5. Comparing statistical and machine learning models' ability to predict 30-day mortality, acute morbidity, and presence of acute infectious disease with physicians' intuitive judgment of how ill a patient looked. All models were cross-validated and evaluated with respect to the area under the ROC curve (AUC; error bars are bootstrapped 95% confidence intervals) according to signal detection theory in the holdout set. BA = naïve Bayes; BOOST = boosted trees; BT = bagged trees; CART = classification and regression trees; FDA = flexible discriminant analysis; J48 = J48 algorithm; KNN = k-nearest neighbors; LDA = linear discriminant analysis; LI = looking ill; LR = logistic regression; NDA = nonlinear discriminant analysis; NN = neural networks; NSC = nearest shrunken centroids; PLSDA = partial least squares discriminant analysis; PR = PART rule; RF = random forests; SVM = support vector machines.

predictability of key medical outcomes by implementing these (and future) results of the modeling process in the form of decision tools. The models' performances varied, but averaged between 0.71 and 0.82, allowing for risk stratification in patients with nonspecific complaints. A second major finding concerns one of the of best predictor sets. Although most of the predictors identified may not surprise the experienced EP, the predictive utility of the physician's first overall impression (a Gestalt-like "looking ill" assessment) is perhaps less expected. This assessment can be obtained at presentation, does not take much time or other resources, and appears to be among the best predictors in two of the three key medical outcomes we investigated (AUC = 0.65 for morbidity, 0.67 for mortality, and 0.60 for infectious disease).

The group of patients with nonspecific complaints is exceptionally difficult in terms of triage, risk stratification, and disposition. Yet, our finding of clearly better-than-chance predictability of these patients' outcomes renders us optimistic that their triage, risk stratification, and disposition can be improved. Note that the AUC considers, separately for each model, all possible decision thresholds and thus all potential weightings of misses and false alarms and quantifies predictability over all of these. Developing specific models tailored to minimize either "false-negative" misses or "false-positive" alarms may result in different and potentially even higher predictability.

Analyzing the models' performances across all possible decision thresholds suffices to identify the most useful predictors for mortality, morbidity, and presence of infectious disease and to estimate the predictors' combined predictive power. However, it is not appropriate when the goal is to identify a single best model that supports clinical decisions. The reason is that some thresholds are irrelevant for clinical practice. Here, it would be necessary to develop and test different models depending on whether they aim to maximize sensitivity or specificity. We are in the process of refining the obtained models to design decision support tools for use in clinical practice. Depending on whether the objective is to improve triage or risk stratification in patients with nonspecific complaints, different candidate predictors can be recruited. Some of them can be obtained at triage (e.g., complaints, current medication, comorbidities, and vital signs) and thus used to develop decision support tools for triage. Others require more time (e.g., laboratory data) and are therefore more suitable for developing decision support tools for risk stratification.

Our results cannot yet be evaluated against similar analyses, as no comparable prospective studies have been conducted. However, serum urea (BUN), which predicted mortality, acute morbidity, and infectious disease in our analyses, has also been found to predict mortality in other conditions frequently presenting as emergencies, such as pneumonia or chronic obstructive pulmonary disease.[17] It has also been incorporated in scores guiding disposition decisions (e.g., CURB and CURB-65[18]).

Perhaps the most unexpected predictor identified in our analysis, the physician's first subjective Gestalt-like overall assessment ("looking ill"), should be analyzed in more detail. What is behind such a snap judgment? While patient histories indicating tumors or heart disease point to the possibility of a serious outcome, and paleness may point to anemia, "looking ill" may promptly and effortlessly integrate multiple such predictors that convey a "signal" to the physician. Should "looking ill" be interpreted as a dubious intuitive judgment that some physicians rightfully distrust?[19] Or should it be considered as a kind of meta-predictor synthesizing multiple qualitative and/or quantitative predictors? Which predictors does this judgment exploit? How much experience does a physician require before such a Gestalt-like impression reaches its maximum accuracy? These questions are all unanswered and can be addressed with the tools of cognitive data science. As a first step toward answering them, research needs to identify predictors (e.g., clinical signs, laboratory results, and comorbidities) that are highly predictive of "looking ill" and to develop hypotheses about which cues physicians may be spotting and using in the formation of their first impression.

LIMITATIONS

Although this is a multicenter study, all study sites are in Switzerland. Patients with nonspecific complaints are also common in other populations.[1–4] To what extent cultural or genetic differences need to be considered when generalizing our results to other regions or countries remains an open question.

Physicians were instructed to give their first overall impressions of how ill patients looked before taking note of any other signs, symptoms, details of history, or laboratory results. However, it is impossible to exclude the possibility that, in some cases, certain information had already been conveyed to the admitting physicians (e.g., by nurses not aware of the study goals).

The predictive performance, averaged across our diverse and large set of models, represents an approximate estimate of the average predictability of mortality, acute morbidity, and presence of acute infectious disease in patients with nonspecific complaints. Although the models' performances are reassuring, in that they show that these patients' key outcomes can be predicted at a rate much better than chance, the models do not constitute user-friendly clinical decision supports that can be readily employed in the ED. As discussed above, moreover, they were not designed to maximize sensitivity or specificity. We are currently developing models that are both tailored to maximize specific performance criteria and simple enough to be used in clinical practice.

Although we have shown that statistical and machine learning models predict three medical outcomes better than physicians' intuitive judgment alone, we have not yet compared the models' performance against physicians' ultimate clinical decisions. What degree of clinical accuracy modeling-based decision support tools can add to current clinical practice remains an open but crucial question that we can investigate with tools provided by cognitive data science. Once we have developed such decision support tools, we will be able to compare their accuracy to that of physicians' actual decisions.

As mentioned above, we used $p < 0.05$ as a cutoff to select the candidate predictors for our modeling procedure. This arguably stringent cutoff was intended to limit the large number of possible predictors. For the sake of comparability, the same cutoff was used for all three medical outcomes. Using a more lenient cutoff (e.g., 0.10 or 0.15), especially in the prediction of presence of infectious disease, might have resulted in more predictors being included in the models, potentially boosting predictive performance. In our planned analysis of models that aim to enhance sensitivity and/or specificity, we will also investigate the effect of different cutoff p-values.

CONCLUSIONS

We have demonstrated that mortality, acute morbidity, and presence of an acute infectious disease can be predicted using available standard predictors. Most importantly, for risk stratification in terms of short-term mortality and acute morbidity, the best predictors are easily obtainable from a short patient history combined with the physician's overall Gestalt-like first impression of "how ill a patient looks." The models we derived outperformed physicians' first impressions alone and were found to have substantial predictive accuracy.

The authors thank Susannah Goss for editing this manuscript and Max Kuhn for his support in implementing the modeling techniques.

References

1. Nemec M, Koller MT, Nickel CH, et al. Patients presenting to the emergency department with non-specific complaints: the Basel Non-specific Complaints (BANC) Study. Acad Emerg Med 2010;17:284–92.
2. Mockel M, Searle J, Muller R, et al. Chief complaints in medical emergencies: do they relate to underlying disease and outcome? The Charité Emergency Medicine Study (CHARITEM). Eur J Emerg Med 2013; 20:103–8.
3. Vanpee D, Swine CH, Vandenbossche P, Gillet JB. Epidemiological profile of geriatric patients admitted to the emergency department of a university hospital localized in a rural area. Eur J Emerg Med 2001;8:301–4.
4. Safwenberg U, Terént A, Lind L. The emergency department presenting complaint as predictor of in-hospital fatality. Eur J Emerg Med 2007;14:324–31.
5. Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Acad Emerg Med 2000;7:236–42.
6. Domingos P. A few useful things to know about machine learning. Commun ACM 2012;55:78.
7. Rohacek M, Nickel CH, Dietrich M, Bingisser R. Clinical intuition ratings are associated with morbidity and hospitalisation. Int J Clin Pract 2015;69:710–7.
8. Kuhn M, Johnson K. Applied Predictive Modeling. New York City, NY: Springer, 2013.
9. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods 2009;14:323–48.
10. Stekhoven DJ, Buhlmann P. MissForest: nonparametric missing value imputation for mixed-type data. Bioinformatics 2011;28:112–8.
11. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. New York, NY: Chapman and Hall, 1984.
12. Hastie T, Buja A, Tibshirani R. Penalized discriminant analysis. Ann Stat 1995;23:73–102.
13. Breiman L. Random forests. Mach Learn 2001;45:5–32.
14. Janitza S, Strobl C, Boulesteix AL. An AUC-based permutation variable importance measure for random forests. BMC Bioinformatics 2013;14:1–11.
15. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics 2008;9:307.
16. Hertwig R, Meier N, Nickel C, et al. Correlates of diagnostic accuracy in patients with nonspecific complaints. Med Decis Making 2013;33:533–43.
17. Ray P, Birolleau S, Lefort Y, et al. Acute respiratory failure in the elderly: etiology, emergency diagnosis and prognosis. Crit Care 2006;10(3):R82.
18. Lim WS, van der Eerden MM, Laing R, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax 2003;58:377–82.
19. Croskerry P. From mindless to mindful practice: cognitive bias and clinical decision making. N Engl J Med 2013;368:2445–8.