ORIGINAL RESEARCH

Use of Patient-Reported Symptom Data in Clinical Decision Rules for Predicting Influenza in a Telemedicine Setting

W. Zane Billings, Annika Cleven, Jacqueline Dworaczyk, Ariella Perry Dale, PhD, MPH, Mark Ebell, PhD, Brian McKay, PhD, and Andreas Handel, PhD

Introduction: **Increased use of telemedicine could potentially streamline influenza diagnosis and reduce transmission. However, telemedicine diagnoses are dependent on accurate symptom reporting by patients. If patients disagree with clinicians on symptoms, previously derived diagnostic rules may be inaccurate.**

Methods: **We performed a secondary data analysis of a prospective, nonrandomized cohort study at a university student health center. Patients who reported an upper respiratory complaint were required to report symptoms, and their clinician was required to report the same list of symptoms. We examined the performance of 5 previously developed clinical decision rules (CDRs) for influenza on both symptom reports. These predictions were compared against PCR diagnoses. We analyzed the agreement between symptom reports, and we built new predictive models using both sets of data.**

Results: **CDR performance was always lower for the patient-reported symptom data, compared with clinician-reported symptom data. CDRs often resulted in different predictions for the same individual, driven by disagreement in symptom reporting. We were able to fit new models to the patient-reported data, which performed slightly worse than previously derived CDRs. These models and models built on clinician-reported data both suffered from calibration issues.**

Discussion: **Patients and clinicians frequently disagree about symptom presence, which leads to reduced accuracy when CDRs built with clinician data are applied to patient-reported symptoms. Predictive models using patient-reported symptom data performed worse than models using clinician-reported data and prior results in the literature. However, the differences are minor, and developing new models with more data may be possible. (J Am Board Fam Med 2023;00:000–000.)**

Keywords: **Clinical Decision Rules, Cohort Studies, Infectious Diseases, Influenza, Prospective Studies, Respiratory Tract Diseases, Students, Telemedicine, Triage**

Introduction

Influenza causes disease in millions of individuals, including hundreds of thousands of hospitalizations, every year in the United States alone.[1] Globally, seasonal influenza is estimated to cause hundreds of thousands of deaths each year, disproportionately affecting the elderly.[2]

Clinical decision rules (CDRs, also called clinical prediction rules) are tools used by physicians to diagnose patients based on observable evidence.[3–6] Since many of these CDRs are based on signs and symptoms which can be observed by patients, CDRs may be a useful tool for remote forward triage services. However, patients and clinicians can disagree on what symptoms are present.[7–14] Most

This article was externally peer reviewed.

Submitted 29 March 2023; revised 22 May 2023; accepted 25 May 2023.

This is the Ahead of Print version of the article.

From the Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA (WZB, APD, ME, AH); Department of Mathematics, St. Olaf College, Northfield, MN (AC); Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, AZ (JD); Department of Family and Consumer Sciences, University of Georgia, Athens, GA (BM).

Funding: WZB was funded by the University of Georgia Graduate School. AC and JD were funded by National Science Foundation grant #1659683 through the Population Biology of Infectious Diseases Research Experience for Undergraduates site. AH acknowledges partial support from NIH grants AI170116 and U01AI150747.

Conflict of interest: The authors have no conflicts of interest to declare.

Corresponding author: Andreas Handel, PhD, Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA 30602 (E-mail: ahandel@uga.edu).

CDRs based on signs and symptoms were designed using clinician-reported data. The usefulness of these rules for remote triage therefore depends on whether patients can accurately provide necessary information. Robust forward triage systems have the potential to reduce burden on the health care system, but to our knowledge, no one has studied whether these rules are valid in a remote health care context.

The recent rise in telemedicine may provide unique opportunities to reduce influenza transmission during epidemics,[15,16] as well as improve surveillance,[17,18] diagnosis,[19] and treatment.[20] Virtual visits are becoming more popular, and can improve the quality and equity of health care.[21] Implementing forward triage systems, which sort patients into risk groups before any in-person health care visits, through telemedicine can leverage these advantages, especially if automated systems are implemented. Patients who have low risk could be recommended to stay home, rather than seeking in-person health care services.[21–23] Screening out these low risk patients reduces the potential contacts for infected individuals receiving in-person health care, potentially reducing transmission during an epidemic.[24,25]

In our analysis, we evaluated several previously developed CDRs for the diagnosis of influenza to see how they performed for both clinician and patient-reported symptoms. We then examined differences between symptom reports by patients and by clinicians to determine if disagreement was a major factor in determining differences in CDR performance. Finally, we fit similar models to patient-reported symptom data to determine if updated CDRs would be beneficial for triage. More accurate CDRs for triage could reduce the burden of influenza by reducing transmission and improving treatment.

Methods

Collection and Preparation of Data

The data used in this secondary analysis were collected from a university health center from December 2016 through February 2017. Patients with an upper respiratory complaint filled out a questionnaire before their visit, and indicated the presence or absence of several symptoms. Patients were required to answer all questions on the survey. At the time of the visit, a clinician was required to mark the same symptoms as present or absent. Previous publications detail the study design and data collection methods.[26,27] Briefly, patients 18 years and older who presented with influenza-like illness (ILI) were recruited and given informed consent. ILI was defined as cough or at least two of the following symptoms: headache, fever, chills, fatigue, muscle pain, sore throat, or joint pain. Patients were excluded if English was not their preferred language for appointments, they did not provide consent, or they withdrew consent at any time. All data were deidentified before we received them.

A total of 19 symptoms and the duration of illness were assessed by both the clinician and patient. Duration of illness was collected as free text data, so we recoded this variable as a dichotomous indicator of whether the onset of disease was less than 48 hours before the clinic visit, which we called acute onset. Going forward, when we say "symptom," we include acute onset as well.

In our study sample, all patients received a diagnosis from the clinician, but some additionally received a PCR diagnosis. Clinicians in our study were not blinded to lab results before making a diagnosis, but still sometimes disagreed with PCR results (see Appendix). Since PCR is considered the "gold standard" of viral diagnoses,[28] we elected to use the PCR subset for our analyses. The PCR tested for both influenza A and influenza B, and we report the number of observed cases of each type. In all our following analyses we combined influenza A and B cases, which is consistent with the methodology of previous studies.[29,30]

We estimated the prevalence of each symptom as reported by clinicians and by patients in the overall group, as well as stratified by diagnosis. We also report descriptive statistics for age and sex, which were collected for the PCR subset of the study.

Evaluation of Clinical Decision Rules

We applied several CDRs to both patient-reported and clinician-reported symptom data. We chose to apply five CDRs in total that could be used by a clinician or implemented as part of a telemedicine screening service. We used three heuristic decision rules: presence of both cough and fever (CF); presence of cough and fever with acute onset of disease (CFA); and presence of cough, fever, and myalgia all simultaneously (CFM).[31,32] We also used a weighted score rule derived from a logistic regression model (WS), which included both

fever and cough simultaneously, acute onset, myalgia, chills or sweats;[29] and a decision tree model (TM), which included fever, acute onset, cough, and chills or sweats.[30]

The three heuristic rules all produce binary outcomes, assigning a patient to the high risk group if they display all indicated criteria, or the low risk group otherwise. The score and tree both produce numeric probabilities of predicted risk, which were converted into risk groups using predefined thresholds. Patients with risk below 10% (the testing threshold) were assigned to the low risk group, patients with risk below 50% (the treatment threshold) were assigned to the moderate risk group, and patients with risk at least 50% or greater were assigned to the high risk group, following a standard model of threshold diagnosis.[22,29] As a sensitivity analysis, we varied these thresholds (shown in the Appendix). We compared the performance in our data to previously reported performance metrics.[6] For the heuristic rules, AUROCC (equivalent to balanced accuracy in the case of binary predictions) values were derived from the sensitivity and specificity reported in the original article.[32] For the WS, AUROCC was taken from a previous external validation and was calculated on the entire set of patients.[6,33] For the TM, AUROCC was calculated from the validation set.[30]

We evaluated the agreement between patient and clinician symptom reporting using unweighted Cohen's kappa.[34] Qualitative assessment of agreement using the kappa estimates was based on previously published guidelines for use in medical settings.[35] As a sensitivity analysis, we calculated the percent agreement, the prevalence-and-bias-adjusted kappa (PABAK),[36] Gwet's AC1 statistic[37,38], and Krippendorff's α statistic[38,39] (shown in the Appendix). We calculated 95% confidence intervals for these statistics using the empirical percentiles of the statistic of interest calculated on 10,000 bootstrap resamples.[41]

Developing New Prediction Models

We assessed whether patient-reported symptom data could be used to build CDRs with better performance. We fit new models separately to the patient-reported and clinician-reported data. To better assess the performance of our new models, we divided our data into 70% derivation and 30% validation subgroups. Sampling for the data split was stratified by influenza diagnosis to ensure the prevalence of both groups was similar to the overall prevalence.

To develop a weighted score, we used several variable selection methods to fit models, and selected our final model based on AIC, *a priori* important symptoms, and parsimony. We fit a multivariable logistic regression model with diagnosis predicted by the selected variables, and rounded the coefficients to the nearest half (coefficients were doubled if rounding resulted in half-points). We fit a secondary logistic regression model with diagnosis predicted only by the score to estimate the risk associated with each score value.

We considered four different tree-building algorithms to construct a decision tree model: recursive partitioning (CART),[42,43] fast-and-frugal tree[44,45], conditional inference[46–48], and C5.0.[49–51] We then selected the best tree using Area Under the Receiver Operating Characteristic Curve (AUROCC) and parsimony. We did not manually prune or adjust trees.

Finally, we fit several machine learning models, which are less interpretable but often more powerful. We used 10-fold cross-validation repeated 100 times on the derivation set to train the models. We evaluated the performance of all models using AUROCC. All models were trained only on the derivation set, and performance was estimated on both the derivation set and the validation set separately. The Appendix contains more details on our methodology.

Implementation

Our study is a secondary data analysis of previously collected data, and the data were not collected with our research questions in mind. A formal hypothesis testing framework is inappropriate in this context, as tests would have limited power and inflated false discovery rates. Therefore, we elected not to conduct any formal hypothesis tests, and our results should be interpreted as exploratory.

All analyses, figures, and tables were completed in R version 4.3.0 (2023-04-21 ucrt)[52] using the `boot package`,[40,41] and several packages from the `tidyverse suite`.[53–61] We fitted our models using the `tidymodels infrastructure`[62–71]. The manuscript was prepared using R markdown with the `bookdown package`.[72–75] Tables were generated with `gtsummary`[76] and `flextable`.[77] Figures were generated with `ggplot2`.[59,78]

In the Appendix, we provide detailed session information (including a list of packages and versions), all necessary code and data, and instructions for reproducing our analysis.

Results

Descriptive Analysis

In total, there were $n = 250$ patients in our study with symptom reports and a PCR diagnosis. The prevalence in our data was about 51% (127 out of 250 patients), with 118 cases of Influenza A and 9 cases of influenza B. There were slightly more females (148) than males (102) in the group, and most participants were young adults. Only 10% of participants were older than 22.

The prevalence of each symptom is shown in Table 1. Patients tended to report more symptoms than clinicians. Cough and fatigue were slightly more common in influenza positive patients, while chills/sweats and subjective fever were much more common in influenza positive patients. No symptoms were more common in influenza negative patients. Overall, clinicians reported several symptoms less commonly than patients: chest congestion, chest pain, ear pain, shortness of breath, and sneezing. Physicians were more likely to report fever, runny nose, and pharyngitis. Some symptoms also show interaction effects between the rater and the diagnosis. (ie, one rater was more likely to report a symptom, but only in one diagnosis group.) For example, clinicians more commonly reported eye pain in influenza positive patients, and less commonly reported headache in influenza negative patients.

Evaluation of Previous Influenza CDRs

Table 2 shows the five CDRs we applied (CF, CFA, CFM[32]; WS[29]; and TM[30]), the symptoms they use, and the previously reported AUROCC for each CDR. The table also shows the AUROCC when the rule was used to make predictions with the patient and clinician reported symptoms. A CDR that makes perfect predictions would have an

Table 1. Prevalence of Each Symptom as Reported by Clinicians and Patients

	Influenza + (n = 127)		Influenza – (n = 123)		Overall (n = 250)	
	Clinician	Patient	Clinician	Patient	Clinician	Patient
Total number of symptoms	10 (4, 17)	11 (6, 20)	8 (3, 15)	10 (4, 18)	10 (3, 17)	11 (4, 20)
Acute onset	70 (55%)	65 (51%)	53 (43%)	61 (50%)	123 (49%)	126 (50%)
Chest congestion	32 (25%)	80 (63%)	30 (24%)	47 (38%)	62 (25%)	127 (51%)
Chest pain	12 (9.4%)	44 (35%)	10 (8.1%)	24 (20%)	22 (8.8%)	68 (27%)
Chills sweats	116 (91%)	115 (91%)	76 (62%)	84 (68%)	192 (77%)	199 (80%)
Cough	126 (99%)	122 (96%)	111 (90%)	102 (83%)	237 (95%)	224 (90%)
Ear pain	7 (5.5%)	27 (21%)	12 (9.8%)	35 (28%)	19 (7.6%)	62 (25%)
Eye pain	64 (50%)	21 (17%)	20 (16%)	19 (15%)	84 (34%)	40 (16%)
Fatigue	113 (89%)	120 (94%)	75 (61%)	108 (88%)	188 (75%)	228 (91%)
Headache	112 (88%)	103 (81%)	76 (62%)	98 (80%)	188 (75%)	201 (80%)
Itchy eye	5 (3.9%)	25 (20%)	3 (2.4%)	27 (22%)	8 (3.2%)	52 (21%)
Myalgia	106 (83%)	111 (87%)	58 (47%)	98 (80%)	164 (66%)	209 (84%)
Nasal congestion	122 (96%)	99 (78%)	101 (82%)	90 (73%)	223 (89%)	189 (76%)
Pharyngitis	121 (95%)	106 (83%)	114 (93%)	110 (89%)	235 (94%)	216 (86%)
Runny nose	121 (95%)	93 (73%)	97 (79%)	78 (63%)	218 (87%)	171 (68%)
Shortness of breath	16 (13%)	55 (43%)	17 (14%)	36 (29%)	33 (13%)	91 (36%)
Sneeze	16 (13%)	68 (54%)	12 (9.8%)	57 (46%)	28 (11%)	125 (50%)
Subjective fever	113 (89%)	96 (76%)	71 (58%)	58 (47%)	184 (74%)	154 (62%)
Swollen lymph nodes	11 (8.7%)	55 (43%)	31 (25%)	62 (50%)	42 (17%)	117 (47%)
Tooth pain	0 (0%)	26 (20%)	2 (1.6%)	34 (28%)	2 (0.8%)	60 (24%)
Wheeze	15 (12%)	52 (41%)	16 (13%)	31 (25%)	31 (12%)	83 (33%)

Notes: We calculated the prevalence of each symptom in the overall subsample, as well as stratified by influenza diagnosis. The table shows the number of participants positive (Point Prevalence) for all symptoms, and the median (Range) for the total number of symptoms.

Table 2. Details on Previously Developed CDRs Along with Prior Reported AUROCC

CDR	Symptoms	Source	Previously Reported	Clinician-reported Symptoms	Patient-reported Symptoms
CF	Cough, fever	Monto 2000	0.66	0.70	0.69
CFA	Cough, fever, acute onset	Monto 2000	0.65	0.63	0.61
CFM	Cough, fever, myalgia	Monto 2000	0.65	0.73	0.68
WS	Fever and cough, acute onset, myalgia, chills/sweats	van Vugt 2015	0.71	0.77	0.69
TM	Fever, acute onset, cough, chills/sweats	Afonso 2012	0.80	0.71	0.69

Abbreviations: AUROCC, Area Under the Receiver Operating Characteristic Curve; CDR, Clinical decision rules; CF, presence of cough and fever; CFM, presence of cough, fever, and myalgia; CFA, presence of cough and fever with acute onset of disease; TM, decision tree model; WS, logistic regression model.
Notes: We show AUROCC values reported in previous studies, along with the AUROCC values when our clinician-reported data and patient-reported data are used in the CDRs and compared to the true PCR diagnoses.

AUROCC of 1, while random guessing would have an AUROCC of 0.5.

The CFA and TM rules performed worse on our data, while the CF, CFM, and WS rules performed slightly better. The WS rule was the best performing rule using the clinician-reported symptom data, while multiple rules (WS, TM, and CF) performed similarly on the patient data. Every score performed worse when the patient-reported symptoms were used, but any CDR that performed better than previously reported was still better when the patient-reported data were used. The drop in performance was small for most rules: CF, CFA, and the tree model were only slightly different from the clinician-reported symptom metrics. There was a substantive drop in performance for the CFM rule and the WS.

Analysis of CDR Agreement

To investigate the differences between patient-based and clinician-based CDR performance, we assessed the agreement between their predictions. For the three discrete heuristic CDRs, we obtained Cohen's kappa values of $\kappa = 0.52$; 95% *CI:* 0.41,0.62 for CF, $\kappa = 0.57$; 95% *CI:* 0.46,0.67 for CFA, and $\kappa = 0.50$; 95% *CI*: 0.39,0.60 for CFM. All the kappa values represent a moderate level of agreement.[35] Table 3 shows the contingency tables for each of the heuristic rules with the PCR diagnosis. Patients had a slightly lower accuracy for each of the three rules, despite a higher specificity (true negative rate). Clinicians had a higher sensitivity (true positive rate) for all three rules.

Rather than discretizing the predictions from the WS and TM, we visually assessed the correlation between the results from clinician-reported and patient-reported symptoms (Figure 1). Most of the scores tended to be large, and patients and clinicians tended to agree more on larger scores. For the TM, patients and clinicians were also likely to agree when the model predicted its minimum value for a patient.

Assessment of Interrater Agreement

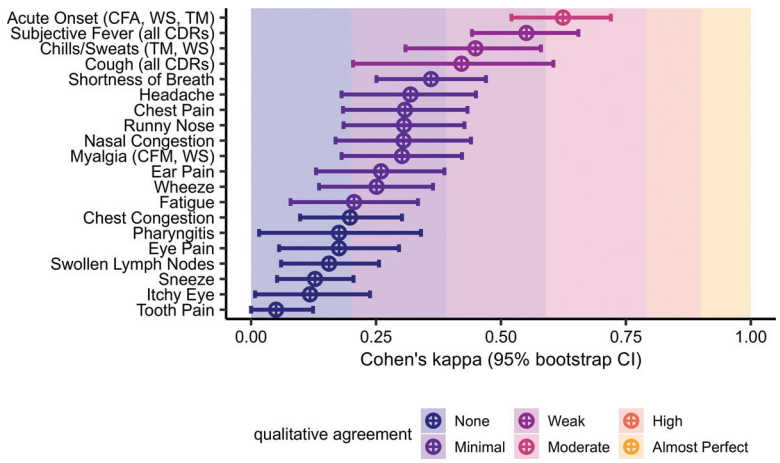
To understand the disagreement in CDR predictions between patient-reported and clinician-reported data, we examined the agreement between clinician and patient symptom reports. Figure 1 shows the calculated Cohen's kappa statistics and confidence intervals for each symptom. The only symptom which achieved moderate agreement was acute onset

Table 3. Number of Patients Who Were Predicted to Have Influenza by Each of the Three Heuristic CDRs, Which Produce Binary Outcomes

	Clinician (n = 250)		Patient (n = 250)	
	Influenza +	Influenza −	Influenza +	Influenza −
CF				
Positive	112 (88%)	60 (49%)	91 (72%)	42 (34%)
Negative	15 (12%)	63 (51%)	36 (28%)	81 (66%)
CFA				
Positive	66 (52%)	31 (25%)	50 (39%)	22 (18%)
Negative	61 (48%)	92 (75%)	77 (61%)	101 (82%)
CFM				
Positive	100 (79%)	40 (33%)	85 (67%)	38 (31%)
Negative	27 (21%)	83 (67%)	42 (33%)	85 (69%)

Abbreviations: CDR, Clinical decision rules; CF, presence of cough and fever; CFM, presence of cough, fever, and myalgia; CFA, presence of cough and fever with acute onset of disease.
Notes: The predictions are stratified by PCR influenza diagnosis.

Figure 1. Cohen's kappa values for each symptom. Cohen's kappa was used to measure agreement between clinician diagnoses and the lab test methods. Qualitative agreement categories were assigned based on previously published guidelines for clinical research.

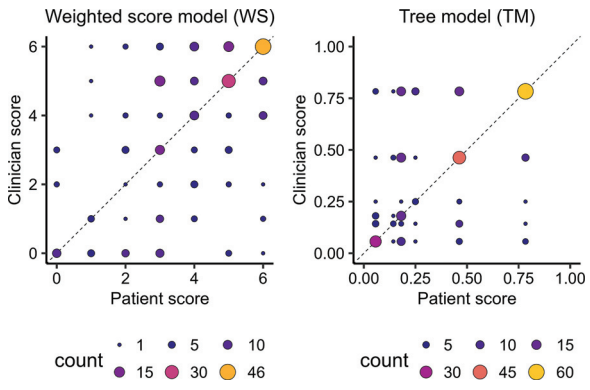


($\kappa = 0.62$; 95% *CI:* 0.52,0.72), according to the clinical guidelines. Symptoms with weak agreement were cough ($\kappa = 0.42$; 95% *CI:* 0.20,0.60), chills and sweats ($\kappa = 0.45$; 95% *CI:* 0.31,0.58) and subjective fever ($\kappa = 0.55$; 95% *CI:* 0.44,0.66), which were common across the CDRs we used. However myalgia (minimal agreement; $\kappa = 0.30$;95% *CI:* 0.18, 0.42) was also included in some of the CDRs.

Patients tended to report a higher number of symptoms overall (Figure 2), including symptoms which were rarely reported by physicians like tooth pain, and symptoms with specific clinical definitions like swollen lymph nodes and chest congestion (Table 1). Patients also were less likely to report certain symptoms, including pharyngitis, runny nose, and nasal congestion. These discrepancies occur for symptoms with lower Cohen's kappa values. However, patients and physicians were about equally likely to report acute onset, supported by a higher kappa value.

In our sensitivity analysis using other measurements of inter-rater agreement, there were no qualitative differences when using other kappa-based statistics. Krippendorff's α showed inconsistent trends.

Figure 2. Clinician versus patient scores for both of the continuous CDRs. The CDRs only have a discrete set of outputs, so the size and color of the points reflects the number of patients (overlapping observations) at each location. If the models agreed perfectly, all observations would fall on the dashed line.



Development of New Models

The differences between patient-reported and clinician-reported symptoms, and subsequent differences in CDR performance, suggest that CDRs developed using patient data might perform better than previous scores developed using clinician-reported data. We built new models using the patient-reported data by emulating the previously developed rules. We selected a point score, a decision tree, and a machine learning algorithm for further examination. We split the data into a derivation set of 176 patients, and a validation set of the remaining 74 patients. All models were trained only on the derivation set.

Based on our selection criteria, the best score model used symptoms selected via LASSO penalization.[79] The score model contained the symptoms chills or sweats (2 points), cough (5 points), and fever

Table 4. Derivation Set and Validation Set AUROCC for Each of the Three Selected Models, Trained and Evaluated on Either the Clinician or Patient Data

	Derivation group		Validation group	
	Clinician	Patient	Clinician	Patient
LASSO point score	0.86	0.78	0.71	0.60
Conditional inference tree	0.79	0.80	0.63	0.57
Naive Bayes classifier	0.83	0.79	0.74	0.68

Notes: The same individuals were used in the derivation and validation sets regardless of whether the clinician-reported symptom data or patient-reported symptom data were used for modeling.

(4 points). The tree we selected was a conditional inference tree containing the variables fever, shortness of breath, wheeze, and cough. Out of the machine learning models we fit, we selected a naive Bayes classification model, which performed competitively on both the clinician-data and patient-data models, and included all symptoms. For comparison, we applied the same modeling procedures to the clinician-reported symptom data. (See Appendix for modeling details.)

Table 4 shows the AUROCC of each of the selected models, using the clinician and the patient data. When trained on the clinician-reported data, the score and naive Bayes models performed better on both the derivation and validation sets than when trained on the patient-reported data. The conditional inference tree performed better on the validation group but worse on the derivation group when trained on the clinician data.

When trained to the patient-reported symptom data, all three models performed well on the derivation group, but their performance dropped substantially on the validation group. The validation group performance estimates the performance on new data, so all three models are likely overfit. The naive Bayes model appeared to overfit the least.

We examined the quantitative risk predictions made by the models, categorizing patients with risk ≤10% as low risk, patients with risk >10% but ≤50% as medium risk, and patients with risk > 50% as high risk. All three models assigned over half of the study participants to the high-risk group, and almost none to the low-risk group (Table 5). Patients in the high-risk group are recommended to seek in-person care in the context of a telemedicine forward triage system.

If we increase the thresholds for risk groups, a few more patients are classified as low or moderate risk. For the patient data models, the majority of patients remain in the high risk group. As a sensitivity analysis, we used the same procedures to fit models to the clinician-reported data. While models fit to

Table 5. Risk Group Statistics for the Models Built Using the Patient Data. The Models Were Trained Using the Derivation Set of Patient-Reported Symptom Data, and Evaluated on Both the Derivation and Validation Sets Separately

	Derivation group			Validation group		
	Flu/Total (%)	LR	In Group (%)	Flu/Total (%)	LR	In Group (%)
LASSO score						
Low	0/5 (0.0)	0.0	2.9	0/0 (NA)	NA	0.0
Moderate	20/77 (26.0)	0.3	44.3	16/34 (47.1)	0.8	45.3
High	68/92 (73.9)	2.8	52.9	23/41 (56.1)	1.2	54.7
Conditional inference tree (manual)						
Low	1/32 (3.1)	0.0	18.4	5/12 (41.7)	0.7	15.8
Moderate	6/25 (24.0)	0.3	14.4	6/14 (42.9)	0.7	18.4
High	81/117 (69.2)	2.2	67.2	28/50 (56.0)	1.2	65.8
Naive Bayes classifier						
Low	0/0 (NA)	NA	0.0	0/0 (NA)	NA	0.0
Moderate	0/7 (0.0)	0.0	4.0	0/4 (0.0)	0.0	5.3
High	88/167 (52.7)	1.1	96.0	39/72 (54.2)	1.1	94.7

Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; LR, stratum-specific likelihood ratio.
Notes: The models were trained using the derivation set of clinician-reported symptom data, and evaluated on both the derivation and validation sets separately. We obtained quantitative risk predictions for each individual from the models, and assigned individuals with a risk less than 10% to the low risk group, individuals with a risk between 10% and 50% to the moderate risk group, and individuals with a risk greater than 50% to the high risk group.

the clinician data were slightly better at identifying low- and medium-risk patients, the majority of patients were still placed in the high risk group by these models (see Appendix).

Discussion

We found that previously developed CDRs perform less well when used with patient-reported symptom data, as opposed to clinician-reported symptom data. Our analysis implies that patient-reported symptom data are likely to be less reliable for influenza triage than clinician-reported symptom data. We observed notable disagreement in many influenza-like illness symptoms, which may explain this discrepancy. Neither the previously developed CDRs, nor our new models fit to the patient-reported data could achieve the same performance with patient-reported symptom data as the best models using the clinician-reported data. However, evaluating the magnitude of these differences is difficult, and further evaluation (eg, a cost-benefit analysis) is necessary to determine whether the difference in predictive power of the models is meaningful in clinical practice.

As clinicians train for several years to identify signs and symptoms of illness, our results may not be surprising. Previous studies identified that patients and clinicians defined "chest congestion[7]", sinus-related symptoms,[11,12] and throat-related symptoms (among others).[13] Given the prior evidence for multiple symptoms, similar discrepancies likely exist with other symptoms. The design of the questionnaire could potentially be modified to better capture the information that would be gained by a clinician's assessment of the patient. The prior work suggests that patients may not understand what a given symptom means, so providing definitions or guides to self-assessing a symptom may be beneficial. Consistent with prior observations, patients in our study also tended to report more symptoms, which could point to issues with the questionnaire designed. All patients in our study were those who sought out health care and wanted to see a clinician, which may bias the reporting of symptoms. This bias might be present in a telemedicine triage context as well.

Our study was limited by a small sample size with accurate diagnoses, which makes fitting predictive models difficult, and a larger sample with accurate reference standards might provide more insight. Our study sample was also composed of young adults aged 18–25 living on a college campus. Our sample is likely unrepresentative of the general population, and our results may reflect a healthy worker bias. Young adults who are able to attend college are typically at low risk for influenza complications, and our study sample is biased toward less severe cases of influenza, which may be more difficult to distinguish from other nonsevere ILIs (eg, rhinovirus or RSV). This bias could explain our issues with model calibration in the low risk group – without any truly high risk patients in our sample, the risk predictions cannot be accurately calibrated. More demographic variation in future studies would also allow for known risk factors like age to be implemented in influenza risk models.

Analyzing the model goodness-of-fit using risk group predictions reveals further questions. Inclusion criteria for our study population included seeking health care and presenting with at least 2 symptoms, so potentially every member of our population is at high risk of influenza. The distribution of risk estimates in our population indicates that patient-reported CDRs might be viable in other populations which is more likely to feature diverse "true" risks of influenza across individuals.

Furthermore, combining patient-reported questionnaires with home rapid testing may provide a viable alternative to prediction methods based only on symptom data.[80] While rapid tests have a high false negative rate, they are cheap (compared with PCR testing), easy to use, and may provide more objective information. Combining rapid tests with symptom questionnaires and CDRs that are optimized for detection of low-risk cases may counterbalance the low sensitivity of the test.

In conclusion, we find that patient-reported symptom data are less accurate than clinician-reported symptom data for predicting influenza cases using CDRs. Our results follow naturally from previous work showing discrepancies between clinician and patient reports of symptoms, and highlight critical issues with patient-based triage systems. However, clinical evaluation is needed to determine whether the difference in performance is meaningful in a real-world context. We conjecture that improved questionnaires or the possible addition of home test results could make patient reports more useful. Regardless, improving remote triage for telemedicine cases is critical to prepare public

health infrastructure for upcoming influenza pandemics. These CDRs may be a cost-effective tool for combating future influenza epidemics, but further development is needed.

We thank the Infectious Disease Epidemiology Research Group at the University of Georgia for feedback on our research.

To see this article online, please go to: http://jabfm.org/content/00/00/000.full.

References

1. Rolfes MA, Foppa IM, Garg S, et al. Annual estimates of the burden of seasonal influenza in the United States: A tool for strengthening influenza surveillance and preparedness. Influenza Other Respir Viruses 2018;12:132–7.
2. Iuliano AD, Roguski KM, Chang HH, Global Seasonal Influenza-associated Mortality Collaborator Network, et al. Estimates of global seasonal influenza-associated respiratory mortality: A modelling study. Lancet 2018;391:1285–300.
3. McIsaac WJ, Goel V, To T, Low DE. The validity of a sore throat score in family practice. CMAJ 2000;163:811–5.
4. Writing Group for the Christopher Study Investigators*. Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-Dimer testing, and computed tomography. JAMA 2006;295:172–9.
5. Wells PS, Owen C, Doucette S, Fergusson D, Tran H. Does this patient have deep vein thrombosis? JAMA 2006;295:199–207.
6. Ebell MH, Rahmatullah I, Cai X, et al. A systematic review of clinical prediction rules for the diagnosis of influenza. J Am Board Fam Med 2021;34:1123–40.
7. McCoul ED, Mohammed AE, Debbaneh PM, Carratola M, Patel AS. Differences in the intended meaning of congestion between patients and clinicians. JAMA Otolaryngol Head Neck Surg 2019;145:634–40.
8. Schwartz C, Winchester DE. Discrepancy between patient-reported and clinician-documented symptoms for myocardial perfusion imaging: Initial findings from a prospective registry. International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care 2021;33:mzab076.
9. Xu J, Schwartz K, Monsur J, Northrup J, Neale AV. Patient-clinician agreement on signs and symptoms of "strep throat" A MetroNet study. Fam Pract 2004;21:599–604.
10. Barbara AM, Loeb M, Dolovich L, Brazil K, Russell M. Agreement between self-report and medical records on signs and symptoms of respiratory illness. Prim Care Respir J 2012;21:145–52.
11. Riley CA, Soneru CP, Navarro A, et al. Layperson perception of symptoms caused by the sinuses. Otolaryngol Head Neck Surg 2023;168:1038–46.
12. Riley CA, Navarro AI, Trinh L, et al. What do we mean when we have a "sinus infection?" Int Forum Allergy Rhinol 2023;13:129–39.
13. Fischer JL, Tolisano AM, Navarro AI, et al. Layperson perception of reflux-related symptoms. OTO Open 2023;7:e51.
14. Sommerfeldt JM, Fischer JL, Morrison DA, McCoul ED, Riley CA, Tolisano AM. A Dizzying complaint: investigating the intended meaning of dizziness among patients and providers. Laryngoscope 2021;131:E1443–E1449.
15. Colbert GB, Venegas-Vera AV, Lerma EV. Utility of telemedicine in the COVID-19 era. Reviews in Cardiovascular Medicine 2020;21:583–7.
16. Gupta VS, Popp EC, Garcia EI, et al. Telemedicine as a component of forward triage in a pandemic. Healthc (Amst) 2021;9:100567.
17. Blozik E, Grandchamp C, von Overbeck J. Influenza surveillance using data from a telemedicine centre. Int J Public Health 2012;57:447–52.
18. Lucero-Obusan C, Winston CA, Schirmer PL, Oda G, Holodniy M. Enhanced Influenza Surveillance Using Telephone Triage and Electronic Syndromic Surveillance in the Department of Veterans Affairs, 2011-2015. Public Health Rep 2017;132:16S–22S. Jul/Aug.
19. Choo H, Kim M, Choi J, Shin J, Shin S-Y. Influenza screening via deep learning using a combination of epidemiological and patient-generated health data: development and validation study. J Med Internet Res 2020;22:e21369.
20. Xiao A, Zhao H, Xia J, et al. Triage modeling for differential diagnosis between COVID-19 and human influenza a pneumonia: classification and regression tree analysis. Front Med (Lausanne) 2021;8:673253.
21. Duffy S, Lee TH. In-person health care as option B. N Engl J Med 2018;378:104–6.
22. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980;302:1109–17.
23. Ebell MH, Locatelli I, Senn N. A novel approach to the determination of clinical decision thresholds. Evid Based Med 2015;20:41–7.
24. Rothberg MB, Martinez KA. Influenza management via direct to consumer telemedicine: an observational study. J Gen Intern Med 2020;35:3111–3.
25. Hautz WE, Exadaktylos A, Sauter TC. Online forward triage during the COVID-19 outbreak. Emerg Med J 2021;38:106–8.
26. Dale AP. Diagnosis, treatment, and impact on function of influenza in a college health population. Published online 2018.

27. Dale AP, Ebell M, McKay B, Handel A, Forehand R, Dobbin K. Impact of a rapid point of care test for influenza on guideline consistent care and antibiotic use. J Am Board Fam Med 2019;32:226–33.
28. Merckx J, Wali R, Schiller I, et al. Diagnostic accuracy of novel and traditional rapid tests for influenza infection compared with reverse transcriptase polymerase chain reaction: a systematic review and meta-analysis. Ann Intern Med 2017;167:394–409.
29. Ebell MH, Afonso AM, Gonzales R, Stein J, Genton B, Senn N. Development and validation of a clinical decision rule for the diagnosis of influenza. J Am Board Fam Med 2012;25:55–62.
30. Afonso AM, Ebell MH, Gonzales R, Stein J, Genton B, Senn N. The use of classification and regression trees to predict the likelihood of seasonal influenza. Fam Pract 2012;29:671–7.
31. Govaert TM, Dinant GJ, Aretz K, Knottnerus JA. The predictive value of influenza symptomatology in elderly people. Fam Pract 1998;15:16–22.
32. Monto AS, Gravenstein S, Elliott M, Colopy M, Schweinle J. Clinical signs and symptoms predicting influenza infection. Arch Intern Med 2000;160:3243–7.
33. van Vugt SF, Broekhuizen BD, Zuithoff NP, GRACE Consortium, et al. Validity of a clinical model to predict influenza in patients presenting with symptoms of lower respiratory tract infection in primary care. Fam Pract 2015;32:408–14.
34. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37–46.
35. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22:276–82.
36. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423–9.
37. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61:29–48.
38. Gwet KL. On Krippendorff's Alpha coefficient. Published online 2015:16.
39. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Med Res Methodol 2016;16:93.
40. Davison AC, Hinkley DV. *Bootstrap Methods and Their Application*. Cambridge University Press; 1997.
41. Canty A, Ripley B. *Boot: bootstrap functions (originally by Angelo Canty for s)*.; 2021. Available at: https://CRAN.R-project.org/package=boot.
42. Therneau T, Atkinson B. *Rpart: recursive partitioning and regression trees*.; 2022. Available at: https://CRAN.R-project.org/package=rpart.
43. Breiman L. *Classification and Regression Trees*. 1st ed. Routledge; 1984.
44. Phillips N, Neth H, Woike J, Gaissmaier W. *FFTrees: generate, visualise, and evaluate fast-and-frugal decision trees*.; 2022. Available at: https://CRAN.R-project.org/package=FFTrees.
45. Phillips ND, Neth H, Woike JK, Gaissmaier W. FFTrees: A toolbox to create, visualize, and evaluate fast-and-frugal decision trees. Judgm decis mak 2017;12:344–68.
46. Hothorn T, Zeileis A. *Partykit: A toolkit for recursive partytioning*.; 2022. Available at: http://partykit.r-forge.r-project.org/partykit/.
47. Hothorn T, Zeileis A. Partykit: A modular toolkit for recursive partytioning in R. Journal of Machine Learning Research 2015;16:3905–9.
48. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. Journal of Computational and Graphical Statistics 2006;15:651–74.
49. Kuhn M, Quinlan R. *C50: C5.0 decision trees and rule-based models*.; 2022. Available at: https://topepo.github.io/C5.0/.
50. Quinlan JR. *C4.5: programs for machine learning*. Morgan Kaufmann; 1993.
51. Kuhn M, Johnson K. *Applied predictive modeling*. 1st ed. 2013, Corr. 2nd printing 2018 edition. Springer; 2013.
52. Core Team R. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing 2022; Available at: https://www.R-project.org/.
53. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. JOSS 2019;4:1686.
54. Wickham H. *Ggplot2: elegant graphics for data analysis*. Springer-Verlag New York; 2016. Available at: https://ggplot2.tidyverse.org.
55. Wickham H. *Tidyverse: easily install and load the tidyverse*.; 2021. Available at: https://CRAN.R-project.org/package=tidyverse.
56. Wickham H, Girlich M. *Tidyr: tidy messy data*.; 2022. Available at: https://CRAN.R-project.org/package=tidyr.
57. Müller K, Wickham H. *Tibble: simple data frames*.; 2022. Available at: https://CRAN.R-project.org/package=tibble.
58. Henry L, Wickham H. *Purrr: functional programming tools*.; 2020. Available at: https://CRAN.R-project.org/package=purrr.
59. Wickham H, Chang W, Henry L, et al. *Ggplot2: create elegant data visualisations using the grammar of graphics*.; 2022. Available at: https://CRAN.R-project.org/package=ggplot2.
60. Wickham H. *Forcats: tools for working with categorical variables (factors)*.; 2021. Available at: https://CRAN.R-project.org/package=forcats.
61. Wickham H, François R, Henry L, Müller K. *Dplyr: a grammar of data manipulation*.; 2022. Available at: https://CRAN.R-project.org/package=dplyr.

62. Kuhn M, Wickham H. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*.; 2020. Available at: https://www.tidymodels.org.
63. Kuhn M, Wickham H. *Tidymodels: easily install and load the tidymodels packages*.; 2022. Available at: https://CRAN.R-project.org/package=tidymodels.
64. Silge J, Chow F, Kuhn M, Wickham H. *Rsample: GENERAL RESAMPLING INFRASTRUCTURE*.; 2022. Available at: https://CRAN.R-project.org/package=rsample.
65. Kuhn M, Wickham H. *Recipes: preprocessing and feature engineering steps for modeling*.; 2022. Available at: https://CRAN.R-project.org/package=recipes.
66. Kuhn M, Vaughan D. *Parsnip: a common api to modeling and analysis functions*.; 2022. Available at: https://CRAN.R-project.org/package=parsnip.
67. Kuhn M. *Tune: Tidy Tuning Tools*.; 2022. Available at: https://CRAN.R-project.org/package=tune.
68. Kuhn M, Vaughan D. *Yardstick: Tidy Characterizations of Model Performance*.; 2022. Available at: https://CRAN.R-project.org/package=yardstick.
69. Vaughan D. *Workflows: Modeling Workflows*.; 2022. Available at: https://CRAN.R-project.org/package=workflows.
70. Kuhn M. *Workflowsets: Create a Collection of Tidymodels Workflows*.; 2022. Available at: https://CRAN.R-project.org/package=workflowsets.
71. Kuhn M, Frick H. *Dials: Tools for Creating Tuning Parameter Values*.; 2022. Available at: https://CRAN.R-project.org/package=dials.
72. Allaire J, Xie Y, McPherson J, et al. *Rmarkdown: Dynamic Documents for r*.; 2022. Available at: https://CRAN.R-project.org/package=rmarkdown.
73. Xie Y. *Bookdown: Authoring Books and Technical Documents with r Markdown*.; 2022. Available at: https://CRAN.R-project.org/package=bookdown.
74. Xie Y, Allaire JJ, Grolemund G. *R Markdown: The Definitive Guide*. Chapman; Hall/CRC; 2018. Available at: https://bookdown.org/yihui/rmarkdown.
75. Xie Y, Dervieux C, Riederer E. R Markdown Cookbook. Chapman; Hall/CRC; 2020. Available at: https://bookdown.org/yihui/rmarkdown-cookbook.
76. Sjoberg DD, Curry M, Larmarange J, Lavery J, Whiting K, Zabor EC. *Gtsummary: presentation-ready data summary and analytic result tables*; 2022.
77. Gohel D. *Flextable: functions for tabular reporting*; 2022. Available at: https://CRAN.R-project.org/package=flextable.
78. Wickham H. *Ggplot2: elegant graphics for data analysis*. 2nd ed. 2016. Springer; 2016.
79. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) 1996;58:267–88. Accessed November 30, 2022. Available at: https://www.jstor.org/stable/2346178
80. Cai X, Ebell MH, Geyer RE, Thompson M, Gentile NL, Lutz B. The impact of a rapid home test on telehealth decision-making for influenza: A clinical vignette study. BMC Prim Care 2022; 23:75.

Appendix.

Use of Patient-reported Symptom Data in Clinical Decision Rules for Predicting Influenza in a Telemedicine Setting

1. Instructions for Reproducing Analysis

1. Either clone the git repository, or download and unzip the folder.
2. Navigate to the "R" subdirectory and follow the directions there for the order to run code files.
3. When you run a code file, either "run all" or "source" the script from your IDE/GUI. (You could also run via command line if you prefer but it is unnecessary.)

2. Detailed Methods and Results

2.1 Sample Size and Data Cleaning

In total, we had records for 3117 unique visits to the clinic. Of these records, 7 were duplicate entries in the data set we received, which were removed as they were attributable to clerical issues with the electronic system. In addition, 635 were missing symptom data. These records were collected during the first few weeks of data collection, and missing values were due to issues with the collection protocol and database. These patients were excluded from the analysis, as the mechanism of missingness was known to be unrelated to any of the fields of interest. The final study sample included 2475 with complete data, not all these patients received a lab diagnosis.

All patients received a final diagnosis by their clinician. One subset of 250 patients received reverse transcription polymerase chain reaction (PCR) diagnoses, and a second, mutually exclusive subset of 420 patients received rapid influenza diagnostic test (RIDT) diagnoses. Patients were specifically recruited into the PCR group, and out of patients in the "usual care" (non-PCR) group, RIDT tests were administered at the clinician's discretion. Notably, the original study[1] reported 264 records in the PCR group, but we only had 250 nonmissing nonduplicate patients in this group.

2.2 CDR Assessment

We note that the TM utilizes the patient's measured temperature rather than subjective fever. However, patients were not asked to measure their own temperature at home during our study, so we assumed that any report of subjective fever corresponded with a fever greater than 37.3°C. This likely impacted the performance of the TM on our data.

2.3 Score Models

To develop a weighted score CDR, we followed the method used for the development of the FluScore CDR[2], with some minor deviations. We examined the differences in symptom prevalences between diagnostic groups, correlations between symptoms, univariate logistic regression models for each symptom, a full multivariable model, a multivariable model using bidirectional stepwise elimination for variable selection, and a multivariable model using LASSO penalization for variable selection to determine which predictors should be included in the score. We constructed several candidate scores and used information criteria (AIC/BIC), our knowledge of a priori important symptoms[3], and parsimony to choose the best score model. We fit a multivariable unpenalized logistic regression model including the identified predictors of interest and then rounded the coefficients (doubling to avoid half points) to create a score model. Appendix Table 1 shows the performance of the candidate models when

Appendix Table 1. Model Performance Metrics for the Score Models

Name	AIC	BIC	Tjur R^2	Brier Score
LASSO score	196.64	209.28	0.28	0.18
A priori symptom score	198.62	217.58	0.28	0.18
Re-fit FluScore model (Ebell 2012)	199.65	215.44	0.27	0.18
Cough/fever symptom score	199.66	209.14	0.25	0.19
Cough/fever heuristic	200.72	207.04	0.24	0.19
Cough/fever/acute onset symptom score	201.11	213.75	0.26	0.19
Cough/fever/myalgia symptom score	201.34	213.97	0.26	0.19
LASSO heuristic	204.82	211.14	0.22	0.19
Cough/fever/myalgia heuristic	208.96	215.28	0.20	0.20
Cough/fever/acute onset heuristic	232.31	238.63	0.07	0.23

Abbreviation: LASSO, Least Absolute Shrinkage and Selection Operator.
Notes: The models shown were fitted to the patient-reported data, and metrics were calculated using only the derivation set.

using the patient-reported symptom data. Since the names of each model were arbitrarily chosen by us, we show the coefficients with confidence intervals for each of the score models in Appendix Table 2. Coefficients and confidence intervals for each of the score models fit to the clinician-reported symptom data are shown in Table 3.

2.3.1. Tree Models

The best tree model was selected based on AUROCC, unnecessary shown in Appendix Table 5. A diagram of the conditional inference tree fitted to the patient data is shown in Figure 1, and the tree fitted to the clinician data is shown in Figure 2

2.3.2. Machine Learning Models

The candidate machine learning models were CART, conditional inference, and C5.0 decision trees with hyperparameter tuning; Bayesian Additive Regression Trees (BART); random forest; gradient-boosted tree using xgboost; logistic regression; logistic regression with LASSO penalization; logistic regression with elastic net penalization; *k*-Nearest Neighbors (knn); naive Bayes; and Support Vector Machine (SVM) models with linear, polynomial, and Radial Basis Function (RBF) kernels.

Hyperparameters were selected for these models via a grid search with 25 candidate levels for each hyperparameter chosen by latin hypercube search of the parameter space. Candidate models were evaluated using 10-fold cross validation repeated 100 times on the derivation set (for precision of out-of-sample error estimates), and the hyperparameter set maximizing the AUROCC for each model was selected as the best set for that model. We then evaluated the models by fitting the best model of each time to the derivation set, and examining the out-of-sample performance on the validation set. Several of these models had similar validation set performances (AUROCC within 0.01 units).

We selected the naive Bayes model as the model to present in the main text due to the competitive performance on both the clinician and patient data, and the relative simplicity of the classifier. While the naive Bayes model is difficult to interpret and difficult to compute by hand, the calculations are computationally efficient and simple. In a telemedicine setting where all calculations can be automated, these limitations matter much less than they would in a traditional health care setting.

Appendix Table 2. Estimated Logistic Regression Coefficients (b) for the Patient-Reported Symptom Data

Score Model	Symptom	b	Points	95% CI
A priori symptoms	Cough	2.85	6	1.41, 4.80
	Subjective_fever	1.99	4	1.21, 2.82
	Acute_onset	−0.27	−1	−1.02, 0.45
	Chills_sweats	1.24	2	0.26, 2.29
	Myalgia	−0.69	−1	−1.85, 0.44
LASSO	Chills_sweats	1.07	2	0.13, 2.07
	Cough	2.74	5	1.34, 4.66
	Subjective_fever	1.81	4	1.09, 2.56
Ebell flu score symptoms	Acute_onset	−0.39	−1	−1.13, 0.31
	Myalgia	−0.70	−1	−1.85, 0.43
	Chills_sweats	1.16	2	0.20, 2.21
	Cough:subjective_fever	2.19	4	1.46, 2.99
CF (unweighted)	Cough:subjective_fever	2.17	2	1.50, 2.88
CFA (unweighted)	Cough:subjective_fever:acute_onset	1.23	1	0.55, 1.96
CFM (unweighted)	Cough:subjective_fever:myalgia	1.93	2	1.28, 2.61
LASSO variables (unweighted)	Chills_sweats:cough:subjective_fever	2.05	2	1.39, 2.75
CF (weighted)	Cough	2.63	5	1.24, 4.55
	Subjective_fever	2.04	4	1.35, 2.77
CFA (weighted)	Cough	2.61	5	1.21, 4.53
	Subjective_fever	2.10	4	1.39, 2.87
	Acute_onset	−0.27	−1	−1.00, 0.44
CFM (weighted)	Cough	2.68	5	1.27, 4.60
	Subjective_fever	2.11	4	1.38, 2.89
	Myalgia	−0.30	−1	−1.34, 0.75

Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; CF, presence of cough and fever; CFM, presence of cough, fever, and myalgia; CFA, presence of cough and fever with acute onset of disease.

Appendix Table 3. Estimated Logistic Regression Coefficients (b) for the Clinician-Reported Symptom Data

Score Model	Symptom	b	Points	95% CI
A priori symptoms	Cough	2.23	4	0.40, 5.20
	Subjective_fever	1.37	3	0.39, 2.40
	Acute_onset	0.20	0	−0.51, 0.91
	Chills_sweats	0.91	2	−0.14, 1.99
	Myalgia	0.67	1	−0.20, 1.52
LASSO	Chills_sweats	1.39	3	0.20, 2.65
	Subjective_fever	1.46	3	0.38, 2.61
	Myalgia	−0.39	−1	−1.53, 0.66
	Runny_nose	1.49	3	0.14, 3.02
	Eye_pain	1.37	3	0.49, 2.30
	Swollen_lymph_nodes	−2.20	−4	−3.49, −1.08
Ebell flu score symptoms	Acute_onset	0.16	0	−0.55, 0.85
	Myalgia	0.72	1	−0.12, 1.56
	Chills_sweats	0.81	2	−0.22, 1.87
	Cough:subjective_fever	1.54	3	0.63, 2.51
CF (unweighted)	Cough:subjective_fever	2.27	5	1.50, 3.13
CFA (unweighted)	Cough:subjective_fever:acute_onset	1.27	3	0.63, 1.93
CFM (unweighted)	Cough:subjective_fever:myalgia	1.95	2	1.30, 2.64
LASSO variables (unweighted)	Chills_sweats:subjective_fever:myalgia:runny_nose: eye_pain:swollen_lymph_nodes	−0.44	−1	−2.49, 1.38
CF (weighted)	Cough	2.62	5	0.86, 5.56
	Subjective_fever	2.19	4	1.38, 3.10
CFA (weighted)	Cough	2.66	5	0.89, 5.60
	Subjective_fever	2.10	4	1.27, 3.03
	Acute_onset	0.33	1	−0.36, 1.01
CFM (weighted)	Cough	2.18	4	0.35, 5.15
	Subjective_fever	1.71	3	0.81, 2.69
	Myalgia	0.96	2	0.17, 1.75

Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; CFM, presence of cough, fever, and myalgia; CFA, presence of cough and fever.
Notes: All models were fit only to the derivation set. Confidence intervals for the coefficients were calculated using the wald method.

3. Clinician and PCR Agreement

We had many more patients included in our study with clinician diagnoses ($n = 2475$) than PCR tests ($n = 250$). Using a larger sample size would likely help with model fitting. However, the clinicians in our study saw the PCR results before they made their final diagnosis, so we cannot directly assess the accuracy of the clinicians at predicting influenza.

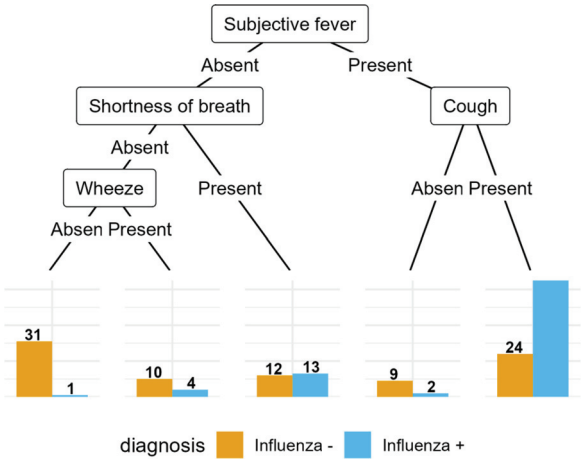
Despite having access to the PCR diagnoses, however, clinicians only agreed with the PCR results 86.0% (95% CI: 81.6%, 90.0%) of the time. Appendix Table 4 shows the contingency table of diagnoses by the clinicians versus the PCR results.

4. Additional IRR Statistics

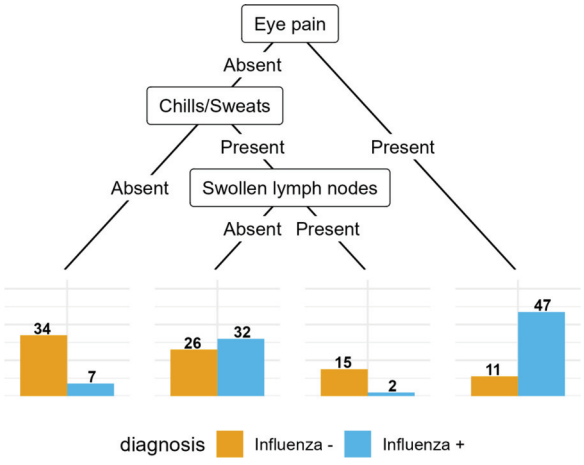
There are known problems with the interpretation of Cohen's kappa statistic. Cohen's kappa depends on the prevalence and variance of the data. That is, the percentage of yes/no answers affects Cohen's kappa, even if the actual percent agreement stays the same. Cohen's kappa is maximized when half of the cases are true 'yes' answers and half are true 'no' answers, which can lead to low kappa values when prevalence is high or low, regardless of the actual percentage agreement. This property is sometimes called "the paradox of kappa"[4,5].

Alternative statistics to Cohen's kappa have been proposed, including the prevalence-and-bias-adjusted kappa (PABAK)[6], Gwet's AC1 statistic[7,8], and Krippendorff's α statistic[8,9]. In addition to calculating Cohen's kappa, we calculated the percent agreement along with these three additional statistics (Figure 3). The percent agreement is not corrected for chance agreement. PABAK and AC1 are corrected for chance agreement and were developed to limit the so-called "paradox of kappa." Finally, Krippendorff's α is based on correcting chance disagreement rather than chance agreement, and whether it is similar or different from kappa-based statistics is inconsistent.

Appendix Figure 1. The conditional inference tree, fitted to the patient data.



Appendix Figure 2. The conditional inference tree, fitted to the clinician data.

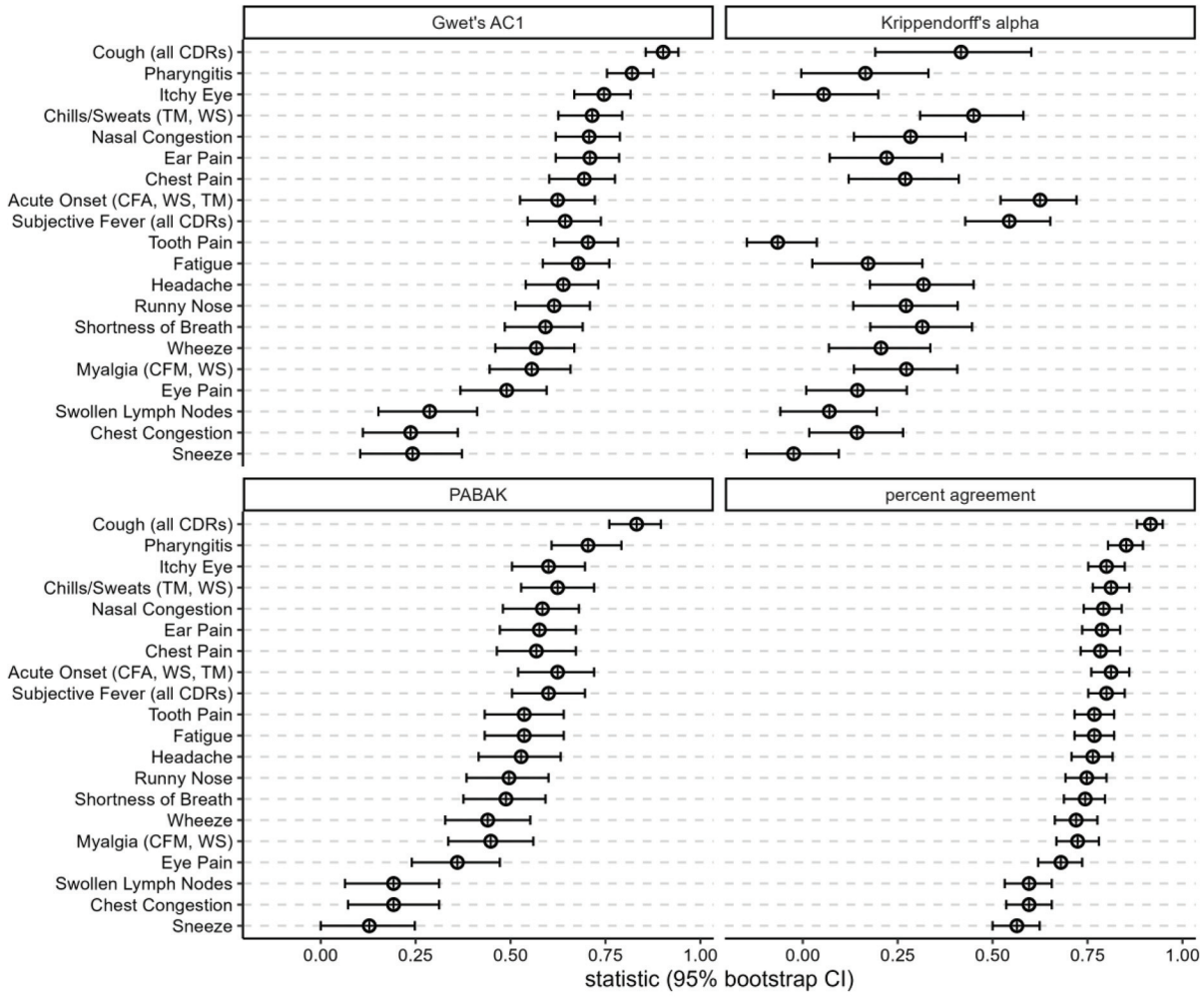


Appendix Table 4. Contigency Table for PCR versus Unblinded Clinician Diagnoses for the Same Patients

	PCR		
	Positive	Negative	Total
Clinician			
positive	116	24	140
negative	11	99	110
Total	127	123	250

Abbreviation: PCR, Polymerase chain reaction.
Notes: Most of the time, clinicians agreed with the PCR results, but rarely the diagnoses differed. Justifications for clinician diagnoses were not collected as part of the study.

Appendix Figure 3. Additional IRR statistics for agreement between symptom reports. Abbreviations: IRR, Incidence rate ratio; PABAK, Prevalence-adjusted kappa; CI, Confidence interval.



Notes: Gwet's AC1 statistic, PABAK, and percent agreement show the same overall trends as the Cohen's kappa statistic reported in the main text. However, Krippendorff's a is quite different, and shows no systematic pattern in the differences from the other 4 statistics.

Our observed Krippendorff's α values vary widely, and do not show a general trend along with the kappa-type statistics we computed. In general, the AC1 and PABAK values follow the same trend as the reported Cohen's kappa values in the main text. Notably, Gwet's AC1, when interpreted with the same guidelines used for Cohen's kappa, is larger and assigns some symptoms to a higher agreement level. Cough and pharyngitis are marked as high agreement using AC1, which may indicate that pharyngitis should be considered in the development of influenza CPRs. Since pharyngitis was not included in the CPRs we tested, and cough already had one of the highest agreement ratings in our main analysis, these findings do not substantially change our conclusions.

5. Performance of All Models

We evaluated the performance of all the candidate models. Appendix Table 5 shows the derivation and validation set AUROCC values on both the clinician-reported and patient-reported data for all the models we fit.

6. Risk Groups for Clinician Data Models

We used the same 10% and 50% thresholds to place patients into risk groups using models fit to the clinician-reported symptom data. We used the same modeling procedures as for the patient-reported data, but model tuning was performed using the clinician-reported data instead.

Appendix Table 5. Estimated AUROCC for All Candidate Models

	Derivation group		Validation group	
	Clinician	Patient	Clinician	Patient
A priori symptom score	0.77	0.79	0.75	0.56
Cough/fever heuristic	0.71	0.75	0.67	0.55
Cough/fever symptom score	0.71	0.76	0.67	0.57
Cough/fever/acute onset heuristic	0.64	0.62	0.61	0.57
Cough/fever/acute onset symptom score	0.73	0.77	0.67	0.55
Cough/fever/myalgia heuristic	0.72	0.72	0.75	0.58
Cough/fever/myalgia symptom score	0.75	0.76	0.74	0.55
Re-fit FluScore model (Ebell 2012)	0.77	0.79	0.74	0.57
LASSO score	0.86	0.78	0.71	0.60
LASSO heuristic	0.50	0.74		0.57
CART (manual)	0.81	0.82	0.67	0.55
FFT	0.77	0.73	0.70	0.53
C5.0 tree (manual)	0.73	0.79	0.65	0.55
Conditional inference tree (manual)	0.79	0.80	0.63	0.57
Bayesian Additive Regression Trees (BART)	0.86	0.81	0.70	0.64
C5.0 tree (tuned)	0.85	0.75	0.60	0.57
CART (tuned)	0.81	0.82	0.67	0.55
Conditional inference tree (tuned)	0.79	0.79	0.63	0.55
Elastic net logistic regression	0.87	0.83	0.70	0.65
Unpenalized logistic regression	0.88	0.84	0.67	0.62
k-Nearest Neighbors classifier	0.89	0.92	0.66	0.65
LASSO logistic regression	0.87	0.83	0.70	0.65
Naive Bayes classifier	0.83	0.79	0.74	0.68
Random forest	0.89	0.87	0.73	0.59
SVM (linear kernel)	0.85	0.82	0.75	0.65
SVM (polynomial kernel)	0.83	0.79	0.74	0.67
SVM (RBF kernel)	0.83	0.82	0.74	0.69
Gradient-boosted tree	0.85	0.87	0.71	0.61

Abbreviations: AUROCC, Area Under the Receiver Operating Characteristic Curve; LASSO, Least Absolute Shrinkage and Selection Operator; SVM, Support vector machines; CART, Classification and Regression Tree Algorithm.
Notes: The AUROCC was not estimable for the LASSO heuristic model on the validation set of clinician-reported symptom data, as all patients were assigned the same score in this set.

The models trained to the clinician data, with the exception of the tree model, performed slightly better at placing patients in the low and moderate risk groups (Appendix Table 6). However, the majority of patients were still placed in the high risk group for all 3 of the best-performing models, with no patients being identified as low risk by the conditional inference tree model.

7. Risk Group Threshold Analysis

While the 10% and 50% thresholds are based on the expert knowledge of practicing physicians,[2,10] a recent study suggested increased thresholds of 25% and 60% in the context of telehealth visits for influenza-like illness.[11]

7.1 25%/60% Thresholds

We recomputed the risk groups and stratum-specific statistics for both the patient (Appendix Table 7) and clinician (Appendix Table 8) reported data using the 25% and 60% thresholds.

For the patient models, while more patients were classified as low or moderate risk, the majority of patients remained in the high risk group (as compared with the risk groups using the 10% and 50% thresholds). For the clinician data models, the Naive Bayes and LASSO score models showed similar trends. Slightly more patients were categorized as low or moderate risk overall, but the majority of patients remained in the high risk group. However, for the conditional inference tree model, there was an even distribution of patients across the three risk groups.

Appendix Table 6. Risk Group Statistics for the Models Built Using the Clinician Data

	Derivation group			Validation group		
	Flu/Total (%)	LR	In Group (%)	Flu/Total (%)	LR	In Group (%)
LASSO score						
Low	0/18 (0.0)	0.0	10.3	3/9 (33.3)	0.5	11.8
Moderate	19/67 (28.4)	0.4	38.5	7/25 (28.0)	0.4	32.9
High	69/89 (77.5)	3.4	51.1	29/42 (69.0)	2.1	55.3
Conditional inference tree (manual)						
Low	0/0 (NA)	NA	0.0	0/0 (NA)	NA	0.0
Moderate	9/58 (15.5)	0.2	33.3	8/24 (33.3)	0.5	31.6
High	79/116 (68.1)	2.1	66.7	31/52 (59.6)	1.4	68.4
Naive Bayes classifier						
Low	4/36 (11.1)	0.1	20.7	1/9 (11.1)	0.1	11.8
Moderate	3/14 (21.4)	0.3	8.0	2/7 (28.6)	0.4	9.2
High	81/124 (65.3)	1.8	71.3	36/60 (60.0)	1.4	78.9

Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; LR, Stratum-specific likelihood ratio.
Notes: The models were trained using the derivation set of clinician-reported symptom data, and evaluated on both the derivation and validation sets separately. We assigned risk groups using a 25% testing threshold and a 60% treatment threshold. We obtained quantitative risk predictions for each individual from the models, and assigned individuals with a risk less than 10% to the low risk group, individuals with a risk between 10% and 50% to the moderate risk group, and individuals with a risk greater than 50% to the high risk group.

7.2 30%/70% Thresholds

We additionally recomputed the risk groups and stratum-specific statistics using thresholds of 30% and 70% for both the patients (Appendix Table 9) and clinicians (Appendix Table 10). Increasing the thresholds to be even higher should increase the number of patients in the low risk group, but may be difficult to justify clinically.

The patient data models continued to exhibit the same problem: even with these high thresholds, the majority of patients were classified as high risk, across all models and both samples. However, the differences from the 25% and 60% threshold analysis are minor. For the clinician data models, most models remained exactly the same, with the exception of the Naive Bayes model on the derivation group. Each of the models only predicts a discrete set of risk estimates, so if a change in the threshold does not reach the next discrete risk estimate, none of the stratum-specific statistics will change.

Appendix Table 7. Risk Group Statistics for the Models Built Using the Patient Data

	Derivation group			Validation group		
	Flu/Total (%)	LR	In Group (%)	Flu/Total (%)	LR	In Group (%)
Conditional inference tree (manual)						
Low	3/43 (7.0)	0.1	24.7	8/22 (36.4)	0.5	28.9
Moderate	17/39 (43.6)	0.8	22.4	8/13 (61.5)	1.5	17.1
High	68/92 (73.9)	2.8	52.9	23/41 (56.1)	1.2	53.9
Naive Bayes classifier						
Low	0/2 (0.0)	0.0	1.1	0/0 (NA)	NA	0.0
Moderate	0/10 (0.0)	0.0	5.7	0/5 (0.0)	0.0	6.6
High	88/162 (54.3)	1.2	93.1	39/71 (54.9)	1.2	93.4
LASSO score						
Low	5/41 (12.2)	0.1	23.6	7/23 (30.4)	0.4	30.7
Moderate	20/49 (40.8)	0.7	28.2	9/12 (75.0)	2.8	16.0
High	63/84 (75.0)	2.9	48.3	23/40 (57.5)	1.2	53.3

Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; LR, stratum-specific likelihood ratio.
Notes: We assigned risk groups using a 25% testing threshold and a 60% treatment threshold.

Appendix Table 8. Risk Group Statistics for the Models Built Using the Clinician Data

	Derivation group			Validation group		
	Flu/Total (%)	LR	In Group (%)	Flu/Total (%)	LR	In Group (%)
Conditional inference tree (manual)						
Low	9/58 (15.5)	0.2	33.3	8/24 (33.3)	0.5	31.6
Moderate	32/58 (55.2)	1.2	33.3	14/26 (53.8)	1.1	34.2
High	47/58 (81.0)	4.2	33.3	17/26 (65.4)	1.8	34.2
Naive Bayes classifier						
Low	5/42 (11.9)	0.1	24.1	1/13 (7.7)	0.1	17.1
Moderate	8/15 (53.3)	1.1	8.6	4/8 (50.0)	0.9	10.5
High	75/117 (64.1)	1.7	67.2	34/55 (61.8)	1.5	72.4
LASSO score						
Low	7/58 (12.1)	0.1	33.3	8/24 (33.3)	0.5	31.6
Moderate	12/28 (42.9)	0.7	16.1	2/11 (18.2)	0.2	14.5
High	69/88 (78.4)	3.5	50.6	29/41 (70.7)	2.3	53.9

Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; LR, Stratum-specific likelihood ratio.
Notes: We assigned risk groups using a 25% testing threshold and a 60% treatment threshold.

7.3 Continuous Risk Estimates

Overall, while varying the thresholds did assign more patients to the low and moderate risk groups, with both of our trials, the majority of patients were still assigned to the high risk group. This can be explained by examining the quantitative risk predictions made by the models without binning the estimates into groups.

Appendix Figure 4 shows histograms of the predicted risk for each model. The point score and tree models both produce a sparse set of discrete risk outcomes, so varying the threshold does not affect categorizations until the next measurement is crossed. While the naive bayes model has a larger set of possible outcomes, most of the predictions were close to a risk of 1.

We could arbitrarily choose even higher thresholds to attempt to improve the model metrics, or we could computationally optimize the stratum-specific likelihood ratios by choosing threshold values. But it is unlikely that such data-driven threshold choices would be contextually meaningful or robust across multiple studies.

Appendix Table 9. Risk Group Statistics for the Models Built Using the Patient Data

	Derivation group			Validation group		
	Flu/Total (%)	LR	In Group (%)	Flu/Total (%)	LR	In Group (%)
Conditional inference tree (manual)						
Low	7/57 (12.3)	0.1	32.8	11/26 (42.3)	0.7	34.2
Moderate	13/25 (52.0)	1.1	14.4	5/9 (55.6)	1.2	11.8
High	68/92 (73.9)	2.8	52.9	23/41 (56.1)	1.2	53.9
Naive Bayes classifier						
Low	0/3 (0.0)	0.0	1.7	0/1 (0.0)	0.0	1.3
Moderate	0/17 (0.0)	0.0	9.8	1/6 (16.7)	0.2	7.9
High	88/154 (57.1)	1.3	88.5	38/69 (55.1)	1.2	90.8
LASSO score						
Low	5/41 (12.2)	0.1	23.6	7/23 (30.4)	0.4	30.7
Moderate	20/49 (40.8)	0.7	28.2	9/12 (75.0)	2.8	16.0
High	63/84 (75.0)	2.9	48.3	23/40 (57.5)	1.2	53.3

Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; LR, Stratum-specific likelihood ratio.
Notes: We assigned risk groups using a 30% testing threshold and a 70% treatment threshold.

Appendix Table 10. Risk Group Statistics for the Models Built Using the Clinician Data

	Derivation group			Validation group		
	Flu/Total (%)	LR	In Group (%)	Flu/Total (%)	LR	In Group (%)
Conditional inference tree (manual)						
Low	9/58 (15.5)	0.2	33.3	8/24 (33.3)	0.5	31.6
Moderate	32/58 (55.2)	1.2	33.3	14/26 (53.8)	1.1	34.2
High	47/58 (81.0)	4.2	33.3	17/26 (65.4)	1.8	34.2
Naive Bayes classifier						
Low	5/45 (11.1)	0.1	25.9	1/13 (7.7)	0.1	17.1
Moderate	8/18 (44.4)	0.8	10.3	4/8 (50.0)	0.9	10.5
High	75/111 (67.6)	2.0	63.8	34/55 (61.8)	1.5	72.4
LASSO score						
Low	7/58 (12.1)	0.1	33.3	8/24 (33.3)	0.5	31.6
Moderate	12/28 (42.9)	0.7	16.1	2/11 (18.2)	0.2	14.5
High	69/88 (78.4)	3.5	50.6	29/41 (70.7)	2.3	53.9

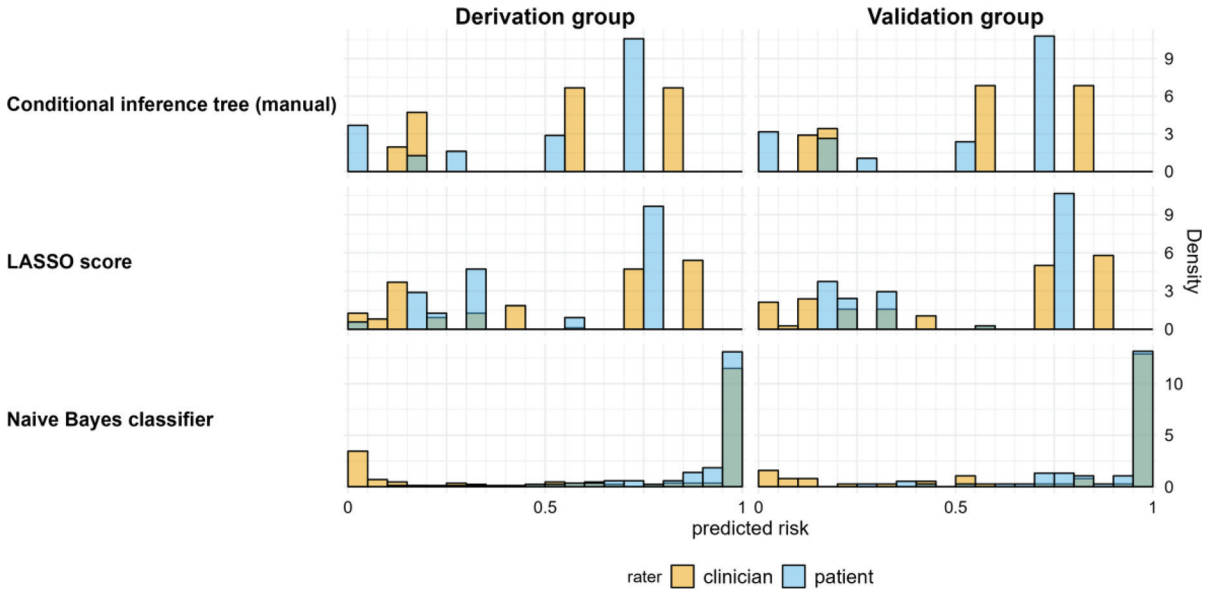
Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; LR, Stratum-specific likelihood ratio.
Notes: We assigned risk groups using a 30% testing threshold and a 70% treatment threshold.

Examining model calibration on the continuous risk estimates would be more revealing than optimizing thresholds for categorizing a continuous variable.

8. R Session and Package Information

```
## R version 4.2.2 (2022-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 × 64 (build 19045)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English United States.utf8
## [2] LC_CTYPE=English United States.utf8
```

Appendix Figure 4. Histograms of individual risks predicted by the models (shown on the left side). Bins represent a width of 5%. Across all models, patients were more often assigned a high risk, and most patients who were at high risk were assigned the same or very close risk estimates.



```
## [3] LC_MONETARY=English United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English United States.utf8
##
## attached base packages:
## [1] stats graphics grDevices utils datasets methods base
##
## other attached packages:
## [1] zlib_0.0.1 renv_0.16.0 gtsummary_1.6.2 tidyselect_1.2.0
## [5] dplyr_1.0.10 readr_2.1.3 here_1.0.1 flextable_0.8.3
## [9] knitr_1.41 bookdown_0.31 rmarkdown_2.19
##
## loaded via a namespace (and not attached):
## [1] xfun_0.36 purrr_1.0.1 colorspace_2.1-0
## [4] vctrs_0.5.1 generics_0.1.3 htmltools_0.5.4
## [7] yaml_2.3.6 base64enc_0.1 to 3 utf8_1.2.2
## [10] rlang_1.0.6 pillar_1.8.1 glue_1.6.2
## [13] withr_2.5.0 DBI_1.1.3 gdtools_0.2.4
## [16] uuid_1.1-0 lifecycle_1.0.3 stringr_1.5.0
## [19] munsell_0.5.0 gtable_0.3.1 zip_2.2.2
## [22] evaluate_0.19 tzdb_0.3.0 fastmap_1.1.0
## [25] fansi_1.0.3 Rcpp_1.0.9 scales_1.2.1
## [28] openssl_2.0.5 systemfonts_1.0.4 ggplot2_3.4.0
## [31] hms_1.1.2 askpass_1.1 digest_0.6.31
## [34] stringi_1.7.12 grid_4.2.2 rprojroot_2.0.3
## [37] cli_3.6.0 tools_4.2.2 magrittr_2.0.3
## [40] tibble_3.1.8 tidyr_1.2.1 pkgconfig_2.0.3
## [43] ellipsis_0.3.2 broom.helpers_1.11.0 data.table_1.14.6
## [46] xml2_1.3.3 assertthat_0.2.1 gt_0.8.0.9000
## [49] officer_0.5.1 rstudioapi_0.14 R6_2.5.1
## [52] compiler_4.2.2
```

References

1. Dale AP, Ebell M, McKay B, Handel A, Forehand R, Dobbin K. Impact of a Rapid Point of Care Test for Influenza on Guideline Consistent Care and Antibiotic Use. J Am Board Fam Med 2019;32:226–33.
2. Ebell MH, Afonso AM, Gonzales R, Stein J, Genton B, Senn N. Development and Validation of a Clinical Decision Rule for the Diagnosis of Influenza. J Am Board Fam Med 2012;25:55–62.
3. Monto AS, Gravenstein S, Elliott M, Colopy M, Schweinle J. Clinical signs and symptoms predicting influenza infection. Arch Intern Med 2000;160:3243–7.
4. Zec S, Soriani N, Comoretto R, Baldi I. High Agreement and High Prevalence: The Paradox of Cohen's Kappa. Open Nurs J 2017;11:211–8.
5. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. Kappa and AC1/2 statistics: Beyond the paradox. J Clin Epidemiol 2022;142:328–9.
6. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423–9.
7. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61:29–48.
8. Gwet KL. On Krippendorff's Alpha Coefficient. Published online 2015:16.
9. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Med Res Methodol 2016;16:93.
10. Sintchenko V, Gilbert GL, Coiera E, Dwyer D. Treat or test first? Decision analysis of empirical antiviral treatment of influenza virus infection versus treatment based on rapid test results. J Clin Virol 2002;25:15–21.
11. Cai X, Ebell MH, Geyer RE, Thompson M, Gentile NL, Lutz B. The impact of a rapid home test on telehealth decision-making for influenza: A clinical vignette study. BMC Prim Care 2022;23:75.