Title

Evaluating the predictive ability of natural language processing in identifying tertiary/quaternary cases in prioritization workflows for interhospital transfer

Permalink

https://escholarship.org/uc/item/80j6n61g

Journal

JAMIA Open, 6(3)

ISSN

2574-2531

Authors

Lee, Timothy
Lukac, Paul J
Vangala, Sitaram
et al.

Publication Date

2023-07-04

DOI

10.1093/jamiaopen/ooad069

Peer reviewed

JAMIA Open, 2023, **6(3)**, ooad069
https://doi.org/10.1093/jamiaopen/ooad069
Research and Applications

Research and Applications

Evaluating the predictive ability of natural language processing in identifying tertiary/quaternary cases in prioritization workflows for interhospital transfer

Timothy Lee, MD, MS[1], Paul J. Lukac, MD, MBA, MS*[,2,3], Sitaram Vangala, MS[4], Kamran Kowsari, PhD[3], Vu Vu, BS[3], Spencer Fogelman, MS[5], Michael A. Pfeffer, MD[6], Douglas S. Bell, MD, PhD[7]

[1]Altamed Health Services, Commerce, CA, United States, [2]Department of Pediatrics, University of California, Los Angeles, Los Angeles, CA, United States, [3]Office of Health Informatics and Analytics, University of California, Los Angeles, Los Angeles, CA, United States, [4]Department of Medicine Statistics Core, University of California, Los Angeles, Los Angeles, CA, United States, [5]Nationwide Insurance, Scottsdale, AZ, United States, [6]Department of Medicine, Stanford University, Palo Alto, CA, United States, and [7]Department of Medicine, University of California, Los Angeles, Los Angeles, CA, United States

Author contributions: T. Lee and P.J. Lukac contributed equally and are considered co-senior authors of this work.

***Corresponding author:** Paul J. Lukac, MD, MBA, MS, Department of Pediatrics & Office of Health Informatics and Analytics, University of California, Los Angeles, 10833, Le Conte Ave, MDCC 22-432, Los Angeles, CA 90095 (plukac@mednet.ucla.edu)

ABSTRACT

Objectives: Tertiary and quaternary (TQ) care refers to complex cases requiring highly specialized health services. Our study aimed to compare the ability of a natural language processing (NLP) model to an existing human workflow in predictively identifying TQ cases for transfer requests to an academic health center.

Materials and methods: Data on interhospital transfers were queried from the electronic health record for the 6-month period from July 1, 2020 to December 31, 2020. The NLP model was allowed to generate predictions on the same cases as the human predictive workflow during the study period. These predictions were then retrospectively compared to the true TQ outcomes.

Results: There were 1895 transfer cases labeled by both the human predictive workflow and the NLP model, all of which had retrospective confirmation of the true TQ label. The NLP model receiver operating characteristic curve had an area under the curve of 0.91. Using a model probability threshold of ≥0.3 to be considered TQ positive, accuracy was 81.5% for the NLP model versus 80.3% for the human predictions ($P = .198$) while sensitivity was 83.6% versus 67.7% ($P < .001$).

Discussion: The NLP model was as accurate as the human workflow but significantly more sensitive. This translated to 15.9% more TQ cases identified by the NLP model.

Conclusion: Integrating an NLP model into existing workflows as automated decision support could translate to more TQ cases identified at the onset of the transfer process.

LAY SUMMARY

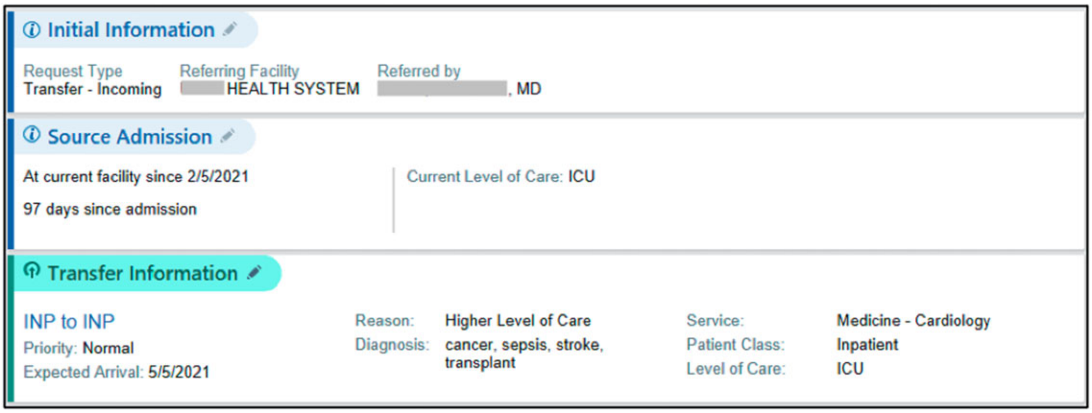
Selection and triaging of patients who are under consideration for interhospital transfer present a challenge. Transfer teams responsible for the intake of patients and determination of patient complexity rely on very little information to classify the level of care these patients may require. The accurate and timely identification of tertiary and quaternary (TQ) patients, who often require highly specialized services and procedures in intensive care units, is of vital importance for both patients, who benefit from receiving appropriate level of care and access to needed services faster; and for hospitals, where more accurate triage allows for improved resource allotment. At our institution, this process was previously dependent solely on human judgment, that of the transfer team. In this study, we utilized natural language processing and an ensemble neural network to predict TQ status of potential interhospital transfers. The model exhibited accuracy on par with human predictions and sensitivity (recall) statistically significantly better than human predictions. This should translate to improved workflow efficiency and an increase in identified TQ cases.

Key words: natural language processing; machine learning; decision support systems; clinical; computer-assisted decision making

Introduction

Artificial intelligence (AI) has garnered attention for a wide range of potential applications in healthcare.[1–4] There is ample excitement for the future of AI, but there are relatively few examples of AI being implemented and studied in operational workflows for health systems.[5,6] Indeed, there is increasing interest in utilizing the predictive ability of AI for operational improvement in healthcare, an area ripe for exploration and study.[6–8] Furthermore, the promise of an AI intervention that offers automated decision support appeals to many health leaders who plan to implement AI in some way in the near future.[8–10] One such use case is the challenge

Received: 29 March 2023. **Revised:** 26 July 2023. **Editorial Decision:** 2 August 2023. **Accepted:** 6 August 2023



Initial Information

Request Type
Transfer - Incoming

Referring Facility
[redacted] HEALTH SYSTEM

Referred by
[redacted], MD

Source Admission

At current facility since 2/5/2021

97 days since admission

Current Level of Care: ICU

Transfer Information

INP to INP

Priority: Normal

Expected Arrival: 5/5/2021

Reason: Higher Level of Care

Diagnosis: cancer, sepsis, stroke, transplant

Service: Medicine - Cardiology

Patient Class: Inpatient

Level of Care: ICU

Figure 1. Transfer Center navigator in Epic. © 2023 Epic Systems Corporation.

of identifying TQ cases in need of transfer to academic health centers (AHCs).

The definition and the problem of identifying TQ

Tertiary care is a designation for specialized medical services not widely available in the community, while quaternary care represents an extension of tertiary care that includes highly specialized services, such as experimental medicine and/or uncommon diagnostic or surgical procedures.[11–13] As an AHC, it is within the scope of the organizational mission to provide and target resources for TQ services.[14] TQ cases are typically medically complex and require substantial institutional resources.[15] There is an intricate mix of quality, patient safety, financials costs, and potential reimbursement that incentivizes AHCs to focus on TQ cases.[16–19]

AHCs are typically regional hubs for escalating care in complex TQ cases, and the resource constraints associated with interhospital transfer to these limited capacity facilities is a high priority issue.[20,21] Increased wait times for transfer can result in canceled referrals and are associated with poorer outcomes, making early identification and prioritization vital.[22] The health system transfer center at the University of California, Los Angeles is a centralized department responsible for administering interhospital transfer referral requests from various regional hospitals to our AHC. With ever increasing demand for patient beds and finite capacity, there is an impetus to improve the identification and prioritization of TQ cases requested for transfer. However, there is no consistent, accurate way to predictively identify TQ patients when initially referred. Currently, a manual workflow involving the judgment of transfer center staff is used to predict TQ cases, and there is no requirement to record a prediction, which leaves some cases unlabeled. This potential for bias in user-dependent predictive TQ labeling was the primary reason that operational leaders at our institution sought an automated decision support tool to provide early TQ identification and prioritization as an adjunct to the existing human workflow.

Transfer center workflow: the opportunity for automation

The transfer process begins with an interhospital request to the transfer center. Intake is recorded by a transfer center staff member entering information about the case into the electronic health record (EHR—vendor Epic) via a transfer center specific navigator (Figure 1). This information includes a free text box filled in with the diagnosis (eg, transplant), among other things, as well as 4 discrete data fields of interest entered on the initial request.

The transfer center staff, composed of all registered nurses, are then tasked with prospectively evaluating incoming transfer cases and flagging a case as "TQ" (called an FYI flag) when appropriate, a designation that persists in the patient encounter record. However, entering this flag is not a mandatory component of the workflow, and staff may forget to complete this step. Of note, there is no "non-TQ" workflow in which a "non-TQ" label is entered, as this would add time burden while not ultimately impacting operational or clinical decisions. Clinical obligations of the transfer center staff beyond triaging include continued follow-up on patient stability for transfer.

When cases have been medically and financially cleared, the patient placement team prioritizes cases for admission and bed placement. The TQ flag is used at this point in the workflow and is available as a decision aid to help identify which cases to prioritize. The TQ flag is displayed on a central screen showing an EHR dashboard called the TC (transfer center) Log (Figure 2).

The project team performed workflow analysis to assess where the display of the natural language processing (NLP) model output would have maximum utility yet minimal disruption to current workflows. Two potential integration points were identified: (1) immediately after the initial intake workflow for transfer requests in the transfer center navigator and (2) on the TC Log in the transfer center central command for patient placement staff to see.

Objective

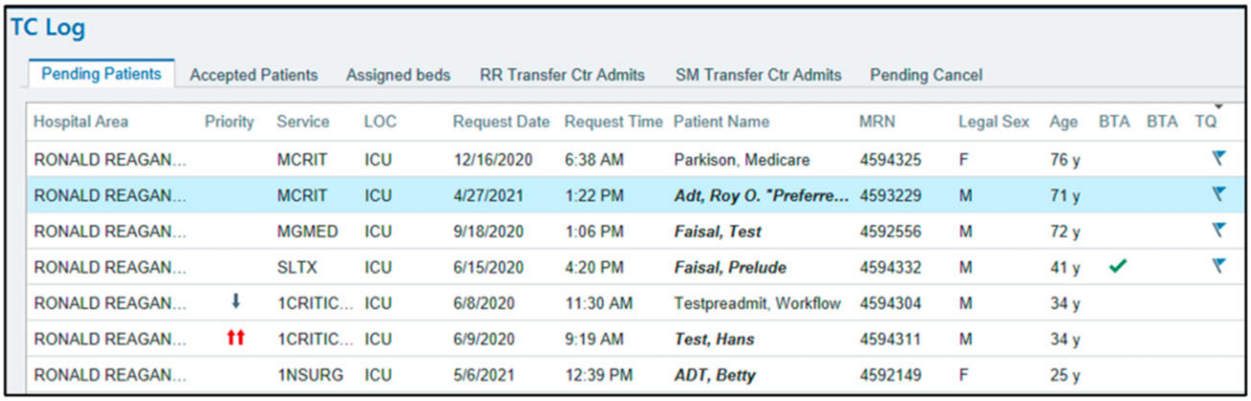
The organizational goal for this NLP project was to create a consistent, reliable way to predictively identify TQ cases early in the transfer process. For the transfer center this meant providing end users with a meaningful, automated output of the model-generated predictions of TQ integrated into the existing EHR workflow.

Methods

Model brief

Data characteristics

The 5 features in our model were chosen from the transfer center staff workflow and extracted from the EHR database



Hospital Area	Priority	Service	LOC	Request Date	Request Time	Patient Name	MRN	Legal Sex	Age	BTA	BTA	TQ
RONALD REAGAN...		MCRIT	ICU	12/16/2020	6:38 AM	Parkison, Medicare	4594325	F	76 y			⚑
RONALD REAGAN...		MCRIT	ICU	4/27/2021	1:22 PM	***Adt, Roy O. "Preferre...***	4593229	M	71 y			⚑
RONALD REAGAN...		MGMED	ICU	9/18/2020	1:06 PM	***Faisal, Test***	4592556	M	72 y			⚑
RONALD REAGAN...		SLTX	ICU	6/15/2020	4:20 PM	***Faisal, Prelude***	4594332	M	41 y	✓		⚑
RONALD REAGAN...	↓	1CRITIC...	ICU	6/8/2020	11:30 AM	Testpreadmit, Workflow	4594304	M	34 y			
RONALD REAGAN...	↑↑	1CRITIC...	ICU	6/9/2020	9:19 AM	***Test, Hans***	4594311	M	34 y			
RONALD REAGAN...		1NSURG	ICU	5/6/2021	12:39 PM	***ADT, Betty***	4592149	F	25 y			

Figure 2. Transfer Center log of transfer cases. © 2023 Epic Systems Corporation.

as follows: Diagnosis Free Text, Referring Facility, Transfer Reason, Requested Level of Care, and Requested Service. The Diagnosis Free Text box contains one or more sentences, making it amenable to the application of NLP, and is also central to determining TQ status.

The NLP model was trained on transfer center patient records queried from the EHR between the dates of March 1, 2019 and June 30, 2020. The inclusion criterion was adults (>18 years old) with referral to the AHC through the transfer center. During the training phase, model development data were divided into a 70% training/30% test dataset. The total number of patient records used in training and testing was 3491 and 1496. Of note, the model was later run in real-time on a separate data set for direct comparison of the NLP model to the existing human workflow. This is described in more detail in "Study Design."

The outcomes were labeled using a monthly financial report (here after called Post Hoc Report) that labels a case as TQ by using a proprietary classification method created by UCLA Health. There is no specific definition of TQ in the literature that identifies which conditions should be considered TQ. Thus, a workgroup of UCLA Health subject matter experts and organizational leaders convened to create a classification method to identify TQ cases and prioritize transfers. This classification method considers patient demographics, diagnosis codes, payor-generated data, and other financial data.

Our dataset included cases that were already financially and medically cleared for transfer to UCLA, were subsequently admitted, and then completed hospitalizations at UCLA. These completed cases were then reviewed against the Post Hoc Report to label cases as TQ in our data set; any case without a TQ label on the Post Hoc Report would be labeled as "non-TQ," even if the patient had been prospectively flagged as "TQ" by the transfer center.

Data preparation and transformation

The selected 5 features were concatenated for NLP analysis as if they were a sentence. Feature preparation and transformation processes consisted of the following steps:

1) Concatenated text fields were cleaned by removing numbers, punctuation, special characters, and extra whitespace.
2) Stop words—words with no semantic importance, such as "the"—were removed using the default stop words list provided in the Natural Language Toolkit (NLTK) library.[23]
3) Lemmatization—a process that replaces the suffix of a word with a different one or removes the suffix of a word completely to get the basic word form.
4) Medical abbreviations and jargon were expanded using a custom dictionary.
5) Concatenated sentences were tokenized into unigrams (individual words) and bigrams (2-word combinations).
6) Tokens were used to calculate term frequency-inverse document frequency (TF-IDF) scores. TF-IDF scores decrease the importance of words that show up in all text fields and increase the importance of words that are unique to either TQ transfer requests or non-TQ transfer requests.

The concatenation of the free text box, which can contain a sentence or more, with the additional 4 discrete features makes this information ripe for the application of NLP. While some of the features may appear short, NLP enables more sophisticated and flexible processing of textual information. For example, among other things, NLP allows for a deeper understanding of language by considering the context, rather than the rules, of the text; it can generalize patterns and rules learned from training data to new, unseen examples; it handles ambiguity, such as homonyms, well; and it is both scalable and adaptable.

Model architecture

In the design and development phases, we evaluated several different traditional and novel machine learning techniques such as support vector machine, multinomial naive Bayes, logistic regression, gradient boosting, long short-term memory (LSTM) using GloVe, and a convolutional recurrent neural network using GloVe.[24–28] All proved inferior to our final model in sensitivity and precision.

The final model was an ensemble of 6 deep (all 3 layers) and 3 shallow (all either 1 or 2 layers) neural networks, with TF-IDF vectors as input, that generated the probability of a transfer request being either TQ or non-TQ (Figure 3). Each individual neural network randomly consisted of 1, 2, and 3 hidden layers with 128, 192, and 256 nodes in each layer. Notably, the construction of an ensemble deep learning algorithm does not typically require tuning of the hyperparameters of the base models because it represents a combination of multiple deep learning models and produces superior predictive performance by reducing the errors of bias.

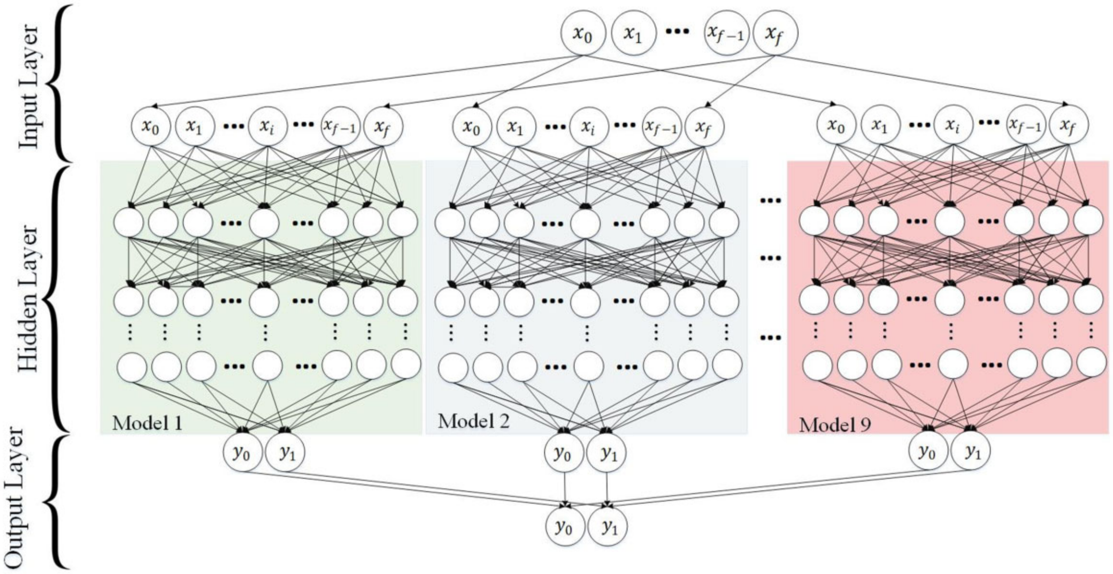

Input Layer
Hidden Layer
Output Layer
x_0 x_1 ... x_{f-1} x_f
x_0 x_1 ... x_i ... x_{f-1} x_f
Model 1
Model 2
Model 9
y_0 y_1

Figure 3. The implemented ensemble model consists of 9 neural networks; for simplicity, only 3 are shown here. The upper level depicts the "Input Layer," which utilizes TF-IDF; the middle level depicts the "Hidden Layer"; and the lower level, consisting of 2 nodes, depicts the "Output Layer."

The networks used a dropout of 0.5 and batch normalization to prevent overfitting.[29,30] The individual layers used the rectified linear unit (ReLU) activation function.[31] Neural network weights were trained using the sparse categorical cross entropy loss function and the Adam optimizer while using accuracy for validation.[32] The weights were trained for 500 epochs, and data were fed into the model as batches of 64 training examples. The final output probability was the average of the 9 individual probabilities, calculated in a process known as soft voting which is based on custom random multimodel deep learning.[33]

Model prediction

The predictive thresholds were determined during design phase. We classified probabilities into 3 likelihood categories: low, medium, and high. These categories were determined by a multidisciplinary team of business stakeholders, subject matter experts, and the data science team and were based on 2 probability thresholds: 0.3 and 0.7. Model probabilities <0.3 are considered strong negative for predicted non-TQ and placed in the *low* category. Model probabilities between 0.3 and 0.7 are considered possibly TQ and placed in the *medium* category. Model probabilities >0.7 are considered strong positive for TQ and placed in the *high* category.

Study design

The aim of our research was to retrospectively compare the ability of the NLP model to identify TQ cases compared to the predictive labeling of the existing human workflow. Both were compared to the Post Hoc Report for confirmation of actual TQ designation. The study period was between July 1, 2020 and December 31, 2020, which is after the model was trained and before the model was implemented into the workflow. The study inclusion criteria were as follows: Adult (>18) patient cases with referral to the AHC through the transfer center for which data was available from the EHR and was considered a completed case with data available from financial claims after hospitalization.

Statistical analysis

We report demographic summary statistics of the study population. We report the diagnostic performance of our NLP model to detect TQ cases using a receiver operating characteristic (ROC) curve as well as a calibration curve.

To compare the NLP model probabilistic output to the binary classification of the existing human workflow and the Post Hoc Report generated TQ label, the model probabilities were transformed into a binary classification of predicted TQ or non-TQ. This was done by taking predictive scores generated by the NLP model and using thresholds determined by the project team (0.7 and 0.3) to create 2 groupings: *NLP—High* ($P > .7$) and *NLP—Medium* ($P \geq .3$). These 2 groupings were used as proxies to compare the performance of the model to the human predictions, with the Post Hoc Report providing actual TQ classifications.

Performance characteristics were estimated for both NLP model thresholds compared with the human predictions, and included sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and Cohen's kappa; 95% confidence intervals were obtained for each parameter using the nonparametric bootstrap; 10 000 bootstrap samples were used, and *P*-values were reported based on interval inversion. An additional analysis was performed to identify the probability cut-point maximizing Youden's J statistic (the sum of sensitivity and specificity). All analyses used a 5% significance level and were performed using R v. 4.1.0.[34]

Results

Described in detail below are the demographics of the study population (Table 1); NLP model performance versus human performance in categorical prediction (Table 2); and performance characteristics of human predictions versus NLP predictions at 2 distinct thresholds (Table 3).

Demographics

Table 1 presents a summary of the main demographic variables of the study population.

NLP model performance

Figure 4 shows the ROC curve for the NLP model predicted probabilities with an area under the curve (AUC) of 0.91. We calculated a calibration curve seen in Figure 5, which shows that with the NLP model predictions the high probabilities are overestimated, and the low probabilities are underestimated.

Table 1. Demographic data for evaluation population.

	Total (*N* = 1895)
Age	
Mean (SD)	56.4 (18.8)
Median (Q1, Q3)	59 (42, 70)
Min, Max	17 109
Sex	
Male	1003 (53%)
Female	892 (47%)
Ethnicity	
Hispanic or Latino	573 (30.2%)
Not Hispanic or Latino	1305 (68.9%)
Unknown	17 (0.9%)
Race	
White or Caucasian	956 (50.4%)
Black or African American	244 (12.9%)
Asian or Pacific Islander	127 (6.7%)
Multiple races	29 (1.5%)
Other	513 (27%)
Unknown	26 (1.4%)

Table 2. Categorical predictions versus actual TQ outcome.

	TQ (*N* = 703)	Non-TQ (*N* = 1192)
Human workflow (prediction)		
TQ	476 (67.7%)	147 (12.3%)
Non-TQ	227 (32.3%)	1045 (87.7%)
NLP—High		
$P > .7$	487 (69.3%)	126 (10.6%)
$P \leq .7$	216 (30.7%)	1066 (89.4%)
NLP—Medium		
$P \geq .3$	588 (83.6%)	236 (19.8%)
$P < .3$	115 (16.4%)	956 (80.2%)

For columns "TQ" and "non-TQ," percentages are for each column.

Model predictions versus human predictions

Of the total sample size of 1895, TQ accounted for 703 cases and non-TQ for 1192 cases. Table 2 presents the summarized confusion matrices for the human predictions (*Human Workflow*) and the model performance at 2 thresholds (*NLP—High* and *NLP—Medium*), all compared to the actual outcome classification of TQ. Table 3 shows the sensitivity, specificity, PPV, NPV, accuracy, and Cohen's kappa of the *Human Workflow* and the 2 models, with *Human Workflow* as the reference category.

NLP—High represents an NLP model threshold of 0.7 (effectively grouping the "medium" category with the "low") compared to the actual TQ outcome. In this comparison, all of the test characteristics for the *NLP—High* were slightly better than the *Human Workflow*, though none achieved statistical significance. *NLP—Medium* represents an NLP model threshold of 0.3 (effectively grouping the medium category with high category) compared to the actual TQ outcome. The comparison between *Human Workflow* and *NLP—Medium* showed statistically significant differences between sensitivity, specificity, PPV, NPV, and kappa with *P*-values <.05.

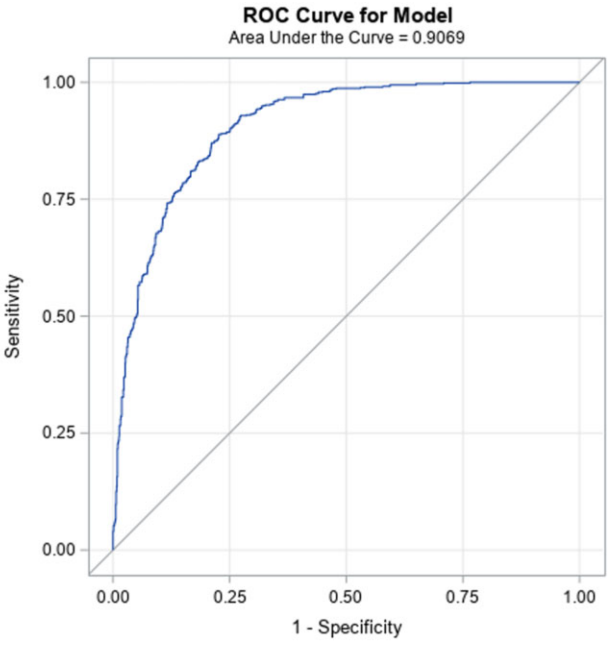


Figure 4. NLP model receiver operating characteristic (ROC) curve with area under the curve (AUC).

Table 3. Performance characteristics for human workflow predictions versus NLP—High and NLP—Medium models.

Metrics %, (95% CI)	Human workflow (reference)	NLP—High	NLP—Medium
Sensitivity	67.7 (64.3-71.2)	69.3 (65.9-72.7)	***83.6 (80.9-86.4)***[a]
Specificity	87.7 (85.8-89.5)	89.4 (87.7-91.2)	***80.2 (77.9-82.5)***[a]
PPV	76.4 (73.1-79.7)	79.4 (76.2-82.6)	***71.4 (68.3-74.4)***[a]
NPV	82.2 (80.0-84.3)	83.2 (81.1-85.2)	***89.3 (87.4-91.1)***[a]
Accuracy	80.3 (78.5-82.1)	82.0 (80.2-83.7)	81.5 (79.7-83.2)
Kappa	56.7 (52.8-60.6)	60.3 (56.5-64.1)	***61.7 (58.1-65.2)***[b]

Significant results are in bold/italics.

[a] $P < .001$.

[b] $P = .010$.

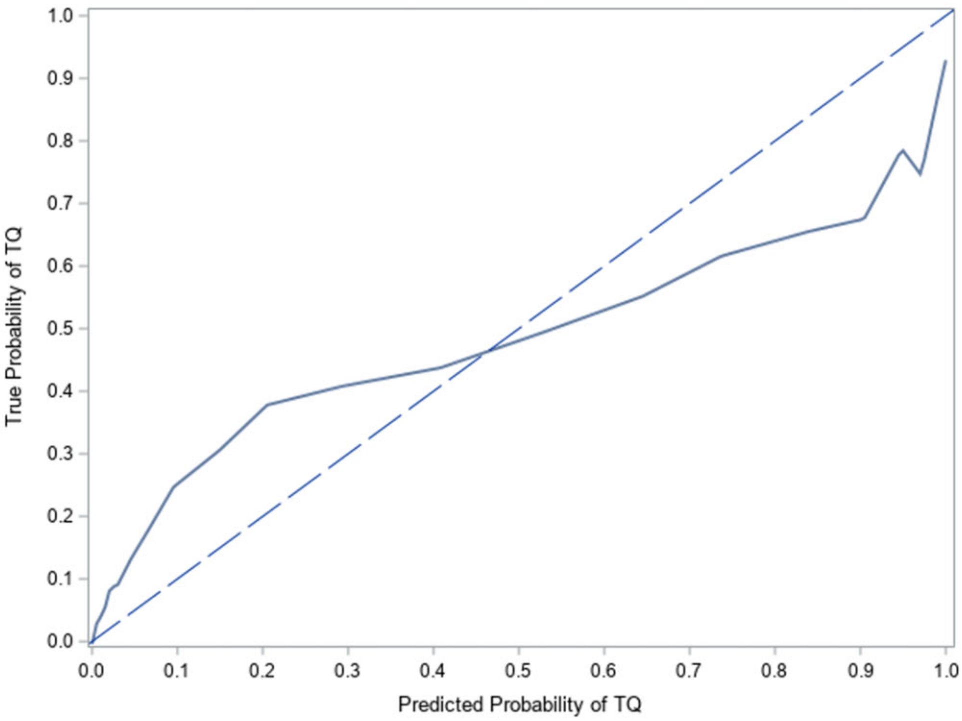


Figure 5. NLP model calibration curve.

Optimal threshold

We used the evaluation dataset to identify a threshold that maximizes sensitivity and specificity (also known as Youden's J). This generated a cutoff of roughly 0.2. In the dataset, this had a sensitivity of 88.9%, a specificity of 77.3%, and accuracy of 81.6%.

Discussion

Common applications of AI include deep learning using NLP to mine EHR text to identify and extract meaningful clinical information, like diagnoses.[35] For example, Kaur et al[36] applied an NLP model to identify patients who met criteria for a validated predictive index for asthma and compared the output to manual chart review, showing high concordance. Clapp et al[37] found similar predictive performances when comparing an NLP-based model to a validated risk stratification tool to identify patients at high risk for maternal mortality. These studies, which demonstrate similar performance between the NLP-based AI models and existing non-AI models, are consistent with our findings.

Model performance

The model performs very well at discriminating between TQ and non-TQ cases, with an AUC of 0.91. The calibration curve suggests that the model tends to overestimate higher probabilities and underestimate lower probabilities. We thus focused our evaluation on the classification performance of the dichotomized model predictions.

Interpreting the results of performance metrics comparisons

Overall, the NLP model modestly outperforms the human workflow at predicting TQ cases when compared to the Post Hoc Report actual TQ outcome. Test performance parameters of sensitivity, specificity, PPV, NPV, accuracy, and to a lesser degree the kappa statistic act as discrete performance metrics to compare the 2 NLP thresholds against the human predictions. For these comparisons to be meaningful outside of statistical significance, evaluating clinical interpretation and significance to operational workflow is vital to identifying the value of each test metric. Operational leaders want to identify TQ cases quickly, consistently, and as early as possible in the transfer process to flag them for prioritization. For this reason, some test parameters are more important to leadership than others. Besides accuracy, operational leaders want to maximize sensitivity to capture as many true TQ cases as possible, even at the expense of increased false positives. Thus, the operational imperative was to create a model with performance at least similar to the human predictions while allowing for more autonomous and consistent identification of comparable accuracy and higher sensitivity.

The *NLP—High* model threshold estimates for all of the test metrics were slightly better than the human predictions, though there were not any statistically significant differences. It can be inferred by the statistical testing that transfer center staff choosing to flag a case as TQ based solely on the "high" category to identify TQ cases would result in performance akin to human judgment while likely also saving time.

Overall accuracy of the *NLP—Medium* threshold was slightly better than the human predictions but not significantly so (81.5% vs 80.3%, $P = .198$). However, the *NLP—Medium* threshold achieved a notable improvement in sensitivity over the human predictions (83.6% vs 67.7%). Since the model generates a probability score as soon as the data fields are entered, this translates to 15.9% ($n = 112$) more positive TQ cases that could have been caught at the onset of the transfer process. Of course, this would be at the expense of more false positives, but the operational team is willing to tolerate these false positives given the opportunity to increase true positive TQ cases and subsequent earlier bed placement. Thus, though the *NLP—Medium* threshold generated a lower specificity and PPV compared to human predictions, it was not outside the bounds of expectations or tolerated differences of the operational team. Furthermore, the *NLP—Medium* threshold NPV was significantly higher than the human predictions. If the model predicts that a case is in the "low" category, it could reasonably be trusted that the case truly was not TQ compared to a human prediction alone.

Given the operational team's desire to retain accuracy but maximize sensitivity, the statistical team identified the optimal cutoff of 0.2. This cutoff has comparable accuracy and is more sensitive but less specific than all 3 groups: the *Human Predictions*, *NLP—High* threshold, and *NLP—Medium* threshold. In practice, this could translate to lowering the medium category from 0.3-0.7 to 0.2-0.7. This particular analysis could inform the thresholds used in the deployment of the model into the EHR.

The implications for automation in clinical workflows

Overall, this study shows that there is potential for improving TQ identification at the beginning of the transfer workflow using an NLP model. The *NLP—Medium* threshold performed well against the existing human workflow with a similar accuracy but significantly improved sensitivity. Assuming the "medium" and "high" categories (0.3-0.7 and 0.7-1.0, respectively) displayed to the end-user translate to a user entered flag of TQ, the NLP model would allow for a higher rate of true positive TQ identification. Even if the medium cases do not end up being TQ, the *NLP—High* threshold comparison reveals that should only the "high" category lead to a TQ designation, the model would perform at least as well as existing human workflow.

The NLP model is being implemented into the transfer center workflow as a form of automated decision support in order to improve upon the current human predictive identification of TQ cases and to prioritize TQ cases that are already medically and financially cleared at the beginning of interhospital transfer request intake. We plan to display the model-generated scores to transfer center staff, who can then use their best judgment on final TQ designation. Importantly, this will not be full automation (ie, model output automatically flags a case as TQ). We believe that the human interaction with the model output serves as a vital processing point to allow human judgment to continue to play a role in decisions that could impact patient care.

Our multidisciplinary team, including stakeholders from the transfer center, prototyped a build in the EHR for the model to display. The build is intended to minimize changes to existing workflows but maximally impact behavior. This will be done by displaying the model output as a percent probability TQ and probability category (low, medium, and high) with a correlated coded color (red, yellow, and green). The model generates scores almost instantaneously after the data are entered and will be displayed at 2 workflow points:

1) Transfer Center Navigator—the model output will be displayed as a new section in the existing workflow deployed in the same module as the current transfer center process.
2) TC Log—the model output will be visible as a new column in the existing transfer center caseload log that patient placement staff review to make decisions.

Limitations

Though we have tested the performance of the model on a limited dataset and compared performance metrics against the existing human predictions, we do not yet know how deploying and integrating the model directly into operational workflows will affect outcomes of the transfer process as well as other important key performance metrics (eg, turnaround time for TQ cases compared to regular cases). Safety and efficacy evaluations will be the subject of a future study examining the impact of the deployed model on the human decision-making process for TQ identification. Furthermore, the Post Hoc Report only has data for cases that have been completed in the hospital. Those cases that were initially referred to the transfer center but whose transfer request was subsequently cancelled, either for medical or financial reasons, were excluded from the training data and the final comparative analysis. These patients could not be used because their TQ outcomes were never determined. We do not believe that this exclusion created any bias. Additionally, as with all language models, performance can degrade over time due to the evolution of medical terminology. We continually monitor model performance via a customized and automated dashboard. Lastly, this model was trained exclusively at one large AHC and thus may not be directly generalizable to other institutions given that workflows and vernacular may differ.

Conclusion

NLP has the potential to automate and improve upon the human ability to predictively identify TQ cases referred to regional transfer centers. Our results suggest that integrating the NLP model into the existing workflow as an automated decision support tool could translate to more TQ cases identified and prioritized at the onset of the transfer process while saving hospital staff time. Furthermore, NLP and various AI techniques can be applied to other clinical and operational workflows, potentially improving performance and optimizing outcomes. Source code has been shared via GitHub (TQ model), and further studies to assess generalizability of this methodology could be informative.

Acknowledgments

We would like to thank the CTSI for help with statistical analysis. We would also like to thank Deb Chandler for assistance with acquiring the financial claims data.

Author contributions

T.L. and V.V. conceived this study. T.L., S.V., P.J.L., K.K., V.V., S.F., and D.S.B. collected and/or analyzed data, while S.V. performed formal statistical analyses. All co-authors participated in the writing, editing, and review of the manuscript.

Funding

This work was supported by NIH National Center for Advancing Translational Science (NCATS) UCLA CTSI (grant number TL1TR001883).

Conflicts of interest

None declared.

Data availability

The data underlying this article cannot be shared publicly due to patient healthcare data privacy protection requirements.

References

1. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books; 2019.
2. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. *J Glob Health*. 2018;8(2):020303.
3. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318.
4. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: a report from the national academy of medicine: a report from the national academy of medicine. *JAMA*. 2020;323(6):509-510.
5. Obermeyer Z, Weinstein JN. Adoption of artificial intelligence and machine learning is increasing, but irrational exuberance remains. *NEJM Catalyst*. 2020;1(1):2-18.
6. Lyell D, Coiera E, Chen J, *et al.* How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. *BMJ Health Care Inform*. 2021;28(1):e100301.
7. He J, Baxter SL, Xu J, *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25(1):30-36.
8. Pianykh OS, Guitron S, Parke D, *et al.* Improving healthcare operations management with machine learning. *Nat Mach Intell*. 2020;2(5):266-273.
9. Petersen C, Smith J, Freimuth RR, *et al.* Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. *J Am Med Inform Assoc*. 2021;28(4):677-684.
10. Healthitanalytics.com. Accessed May 26, 2021. https://healthitanalytics.com/news/90-of-hospitals-have-artificial-intelligence-strategies-in-place
11. Torrey T. How the 4 levels of medical care differ. Verywellhealth.-com. Accessed May 26, 2021. https://www.verywellhealth.com/primary-secondary-tertiary-and-quaternary-care-2615354
12. Tertiary Healthcare – MeSH – NCBI. Nih.gov. Accessed May 26, 2021. https://www.ncbi.nlm.nih.gov/mesh/68063128
13. Project Gutenberg. Quaternary care. Gutenberg.org. Accessed May 26, 2021. http://www.self.gutenberg.org/articles/Quaternary_care?View=embedded%27
14. Hochman M, Robinson J, Dhanireddy K. Implications of medicare's value-based payment initiative for specialty health systems. *Am J Med*. 2018;131(2):117-118.
15. Naessens JM, Van Such MB, Nesse RE, *et al.* Looking under the streetlight? A framework for differentiating performance measures by level of care in a value-based payment environment. *Acad Med*. 2017;92(7):943-950.
16. Cologne KG, Hwang GS, Senagore AJ. Cost of practice in a tertiary/quaternary referral center: is it sustainable? *Tech Coloproctol*. 2014;18(11):1035-1039.
17. Mehaffey JH, Hawkins RB, Mullen MG, *et al.* Access to quaternary care surgery: implications for accountable care organizations. *J Am Coll Surg*. 2017;224(4):525-529.
18. DiSesa VJ, Kaiser LR. What's in a name? The necessary transformation of the academic medical center in the era of population health and accountable care. *Acad Med*. 2015;90(7):842-845.
19. Zuckerman AM, Golden RN. The scale imperative for academic medical centers: part 1 – approach. *J Healthc Manag*. 2015;60(1):8-13.
20. Herrigel DJ, Carroll M, Fanning C, *et al.* Interhospital transfer handoff practices among US tertiary care centers: a descriptive survey. *J Hosp Med*. 2016;11(6):413-417.
21. Southard PA, Hedges JR, Hunter JG, *et al.* Impact of a transfer center on interhospital referrals and transfers to a tertiary care center. *Acad Emerg Med*. 2005;12(7):653-657.
22. Hanane T, Wiles S, Senussi MH, *et al.* Interhospital transfers of the critically ill: time spent at referring institutions influences survival. *J Crit Care*. 2017;39:1-5.
23. Natural Language Toolkit – NLTK 3.6.2 documentation. Nltk.org. Accessed May 28, 2021. https://www.nltk.org/
24. Colas F, Brazdil P. 2006. Comparison of SVM and some older classification algorithms in text classification tasks. In: *Conference Proceedings of the IFIP 19th World Computer Congress: Artificial Intelligence in Theory and Practice*. August 21-24, 2006:169-178. Santiago, Chile.
25. Kibriya AM, Frank E, Pfahringer B, *et al.* Multinomial naive Bayes for text categorization revisited. In: *Conference Proceedings of the Australasian Joint Conference on Artificial Intelligence*. 2004:488-499. Berlin, Heidelberg.
26. Kowsari K, Meimandi KJ, Heidarysafa M, *et al.* Text classification algorithms: a survey. *Information*. 2019;10(4):150.
27. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. October, 2014:1532-1543. Doha, Qatar.
28. Wang R, Li Z, Cao J, *et al.* Convolutional recurrent neural networks for text classification. In: *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*. July, 2019:1-6. Budapest, Hungary.
29. Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: a simple way to prevent neural networks from overfitting. Jmlr.org. Accessed May 28, 2021. https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_campaign=buffer&utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com
30. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D, eds. arXiv [csLG]. http://proceedings.mlr.press/v37/ioffe15.html, 2015:448-456, preprint: not peer reviewed.
31. Agarap AF. Deep learning using rectified linear units (ReLU). https://arxiv.org/abs/1803.08375, 2018, preprint: not peer reviewed.
32. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv [csLG]. http://arxiv.org/abs/1412.6980, 2014, preprint: not peer reviewed.
33. Kowsari K, Heidarysafa M, Brown DE, *et al.* RMDL: random multimodel deep learning for classification. In: *Proceedings of the 2nd International Conference on Information System and Data Mining*. April 2018:19-28. Lakeland, FL.
34. Ripley BD. The R project in statistical computing. *MSOR Connect*. 2001;1(1):23-25.
35. Liang H, Tsui BY, Ni H, *et al.* Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*. 2019;25(3):433-438.
36. Kaur H, Sohn S, Wi CI, *et al.* Automated chart review utilizing natural language processing algorithm for asthma predictive index. *BMC Pulm Med*. 2018;18(1):34.
37. Clapp MA, Kim E, James KE, *et al.* Comparison of natural language processing of clinical notes with a validated risk-stratification tool to predict severe maternal morbidity. *JAMA Netw Open*. 2022;5(10):e2234924.