Original Paper

Human- Versus Machine Learning–Based Triage Using Digitalized Patient Histories in Primary Care: Comparative Study

Artin Entezarjou[1], MD; Anna-Karin Edstedt Bonamy[2,3], PhD, MD; Simon Benjaminsson[4], PhD; Pawel Herman[5*], PhD; Patrik Midlöv[1*], PhD, MD

[1]Center for Primary Health Care Research, Department of Clinical Sciences in Malmö/Family Medicine, Lund University, Malmö, Sweden

[2]Clinical Epidemiology Division, Department of Medicine Solna, Karolinska Institute, Stockholm, Sweden

[3]Doctrin AB, Stockholm, Sweden

[4]Smartera AB, Stockholm, Sweden

[5]Department of Computational Science and Technology, KTH Royal Institute of Technology, Stockholm, Sweden

[*]these authors contributed equally

Corresponding Author:
Artin Entezarjou, MD
Center for Primary Health Care Research
Department of Clinical Sciences in Malmö/Family Medicine
Lund University
Box 50332
Malmö, 202 13
Sweden
Phone: 46 40391400
Email: artin.entezarjou@med.lu.se

Abstract

Background: Smartphones have made it possible for patients to digitally report symptoms before physical primary care visits. Using machine learning (ML), these data offer an opportunity to support decisions about the appropriate level of care (triage).

Objective: The purpose of this study was to explore the interrater reliability between human physicians and an automated ML-based triage method.

Methods: After testing several models, a naïve Bayes triage model was created using data from digital medical histories, capable of classifying digital medical history reports as either in need of urgent physical examination or not in need of urgent physical examination. The model was tested on 300 digital medical history reports and classification was compared with the majority vote of an expert panel of 5 primary care physicians (PCPs). Reliability between raters was measured using both Cohen κ (adjusted for chance agreement) and percentage agreement (not adjusted for chance agreement).

Results: Interrater reliability as measured by Cohen κ was 0.17 when comparing the majority vote of the reference group with the model. Agreement was 74% (138/186) for cases judged not in need of urgent physical examination and 42% (38/90) for cases judged to be in need of urgent physical examination. No specific features linked to the model's triage decision could be identified. Between physicians within the panel, Cohen κ was 0.2. Intrarater reliability when 1 physician retriaged 50 reports resulted in Cohen κ of 0.55.

Conclusions: Low interrater and intrarater agreement in triage decisions among PCPs limits the possibility to use human decisions as a reference for ML to automate triage in primary care.

(JMIR Med Inform 2020;8(9):e18930) doi: 10.2196/18930

KEYWORDS

machine learning; artificial intelligence; decision support; primary care; triage

Introduction

Health care digitalization has the potential to mitigate increasing primary care workloads [1,2]. Time-constrained primary care physicians (PCPs) interrupt patient queries within the first 30 seconds of consultations [3], contributing to inadequate gathering of medical histories [4,5]. To reduce PCP workload and to ensure patients are directed to the appropriate level of care, nurse-led telephone triage is commonly used [6,7]. However, nurses face similar time constraints as physicians, which results in incomplete gathering of medical histories [8] and inappropriate levels of care recommended in up to 31% of cases [9,10].

Leveraging the wide use of smartphones, a large portion of patient history can today be acquired before the patient interacts with his/her health care provider. Automated patient interviewing software has been shown to gather reliable and relevant clinical information [11], and may thus save clinicians time and reduce workloads.

Existing "symptom checkers" can provide triage recommendations directly to patients. However, their accuracy is low, ranging from 33% to 78%, with higher accuracy reported only for more acute conditions [12]. Furthermore, patient adherence to symptom checker recommendations seems low at just 65% [13], compared with 81%-100% adherence to advice from triage nurses [7]. Thus, clinician decision-support software may be a better solution for optimizing triage.

With rapid developments in machine learning (ML), labeled automated patient interviewing software data offer a promising opportunity for enhancing triage software accuracy, providing appropriate access to primary care. Recent research shows promising utility of ML to aid in emergency department triage compared with commonly used algorithms [14]. However, the performance of such a system compared with human triage has, to the best of our knowledge, never been evaluated. Furthermore, ML research in the primary care setting is lacking, despite over 60% of health care visits being conducted in primary care [15].

Thus, this study sought to investigate interrater reliability between human physicians and an automated ML-based triage method, as well as evaluating interrater reliability of triage decisions between a panel of physicians assessing the same patient histories from an automated patient interviewing software.

Methods

Context

The automated patient interviewing software technology used in this study (produced by Doctrin AB, Stockholm, Sweden) is being used by several primary care providers in Sweden since 2017. Patients access the platform using their smartphone, tablet, or computer, choosing their chief complaint from a prespecified list. An automated medical history is then taken, allowing patients to briefly formulate ideas, concerns, and expectations in free-form text, and subsequently answer a symptom-specific multiple-choice survey. The software selects suitable subsequent survey questions based on the patient's answers (Table 1).

Table 1. Examples of automated patient interviewing software survey questions. Chosen answers subsequently appear in reports used for triage.

Survey question	Answer format
"How long have you had a cough?"	Short answer: specify number of days, months or years
"How has your cough been since it started"	Multiple choice (one option allowed): "Not changing" "Getting worse" "Improving" "Gone away"
"Do you have any of the following symptoms?"	Multiple choice (multiple options allowed): "Runny nose" "Shortness of breath" "Chest pain" "Sore throat" "Swollen glands" "Fever"
If a patient reports fever: "What was the highest temperature you have had when you measured it?"	Multiple choice: "37°C" […] "Over 40 C"
"How many days in a row have you had fever?"	Short answer: specify number of days

Answers are presented to a PCP as a summarized report for review and further doctor–patient communication may occur asynchronously through a live text chat (eVisit). Physicians can prescribe medications, order laboratory samples, provide patient information, or remain available online for up to 72 hours for conservative management. Anonymized data from the automated

patient interviewing software report and subsequent chat are saved in a database used for this study. Clinical decisions regarding triage and treatment are, however, recorded separately in the patient medical record and were not accessible for study.

Data for Classification

Data used in this study were composed of 2 subsets. The first subset consisted of 300 automated patient interviewing software reports labeled by a selected expert PCP with over 10 years of clinical experience and a year of experience with online consultations. The reports represented the 10 most common chief complaints in the platform (common cold, cough, eye redness, genital problems, hay fever, rash, headache, sinus symptoms, sore throat, and urinary tract infections) with an equal marginal distribution between chief complaints. Automated patient interviewing software reports were triaged by the expert PCP to one of 4 levels: (1) Start a digital chat-based consultation; (2) Refer the patient to a primary care center for nonurgent care; (3) Refer the patient to a primary care center for urgent care; or (4) Refer the patient to the emergency department.

The second subset was 300 new automated patient interviewing software reports labeled by a panel of 5 PCPs (1 intern [AE], 2 residents, and 2 specialists). Sample sizes were chosen for feasibility reasons. Each PCP individually triaged automated patient interviewing software reports with an identical distribution of chief complaints as in the first subset. Each automated patient interviewing software report was labeled with a triage level as determined by a majority vote by the panel.

Triage categories in both subsets were then dichotomized into 2 triage levels used for further analyses: (1) No need for urgent physical examination (triage levels 1 and 2) or (2) Need of urgent physical examination (triage levels 3 and 4).

Exclusion Criteria

Because of incorrect formatting of one of the reports in the triage interface used by the panel, 299 automated patient interviewing software reports were triaged instead of 300.

Automated patient interviewing software reports describing cases with an ongoing medical contact or a different chief complaint from the one specified were classified as inappropriate for triage, which occurred in 37 reports classified by at least one panel member. These were manually reviewed by one of the authors (AB) for inclusion or exclusion by expert opinion, resulting in the exclusion of 17 cases from the analysis.

If the panel voting strategy did not result in a majority for 1 triage level, the automated patient interviewing software report was also excluded from the analysis, which occurred in 6 cases.

Initially, 22 automated patient interviewing software reports had missing triage data from some panel members. After applying the exclusion criteria, 16 automated patient interviewing software reports with missing triage data remained for analysis.

Model Analyses

To examine the potential of our ML-based approach for triage, we used the available data and corresponding dichotomized triage categories in a series of classification tests with 3 classifiers: (1) a simple linear naïve Bayes classifier, which assumes statistical independence of input features; (2) logistic regression, commonly used for binary classification problems; and (3) random forest, an ensemble decision tree approach, which is considered particularly suitable for high-dimensional problems.

Because of many questions from the automated patient interviewing software reports only appearing very rarely in the small-sized training data, feature space was reduced by only including those which were used in more than 5% of the training samples. This resulted in 243 features. As a few fields included brief free-form text, the classifiers were trained and tested both with and without information extracted from these text data. Text was handled by first removing common Swedish stop words. The remaining commonly used words appearing in more than 10% of the training samples were included as a bag-of-words model where each word was treated as an input feature to the classifier [16]. This resulted in a total of 53 features.

First, we trained the models on the first subset and tested them in a single pass on the second subset with labels based on the majority vote of the 5 PCPs. We complemented this analysis with a cross-validation approach on the data without text information to better estimate generalization capabilities across the 2 subsets of data. We performed 10-fold cross-validation by dividing the union of the 2 subsets into 10 data clusters, where the mixture of the 2 subsets in 9 out of 10 clusters was used for training and the remaining cluster accounting for 10% (ie, 1/10) served as a test set. By applying this scheme 10 times with different 10% test folds, we could obtain an estimate of the second moment of the generalization classification performance. The cross-validation results were followed up with a nonparametric Friedman test.

We made an attempt at investigating the key input features that had a decisive role in classification. To this end, we ranked the coefficients in the regression models built using naïve Bayes and logistic regression methods as well as variable importance with a random forest approach [17]. We employed the correlation of rank, Kendall τ estimator, to examine the consistency of feature ranking produced by the 3 classifiers:

$$\tau = [(n_c - n_d)]/[n(n-1)/2]$$

where n is the number of features, n_c is the number of concordant feature pairs, and n_d is the number of discordant feature pairs. The pairwise relation between feature pairs (f_i, g_i) and (f_j, g_j) is considered as concordant if the ranking order between features f is the same as for features g, that is, rank (f_i,) > rank (f_j,) and rank (g_i,) > rank (g_j,), or rank (f_i,) < rank (f_j,), and rank (g_i,) < rank (g_j,). If neither of these relation pairs is preserved, feature pairs are referred to as discordant.

Finally, in order to exploit diagnostic evaluation made by each individual PCP in the second data subset, rather than directly considering the majority vote as the data sample label, we built 5 independent naïve Bayes classifiers. Each one of them was trained on labels from the second subset corresponding to 1 of

the 5 panel PCPs. We then evaluated the majority vote of the dichotomized responses of individual classifiers and employed a cross-validation scheme to estimate generalization properties.

Human Versus Model Analysis

To measure the agreement between the PCPs and a classification model, we chose a naïve Bayes approach (referred to as "the model"). Cohen κ [18] was calculated to evaluate interrater reliability of triage level within the panel, as well as interrater reliability between the model results and the panel:

$$\kappa = (p_o - p_e)/(1 - p_e)$$

where p_o is the observed ratio of agreement between 2 raters and p_e is the probability of chance agreement. Cohen κ provides a measure of agreement between raters while accounting for chance agreements. This is in contrast to percentage agreement, which merely quantifies the ratio of cases with the same classification in relation to different classifications made by 2 or more assessors, without accounting for chance agreements. A Cohen κ<0.20 is generally regarded as low, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1.00 as almost perfect agreement [18].

Additional Analyses

To explore how the brief free-form text influenced the classification, the classifier was retrained without features extracted from the brief free-form text. This analysis was conducted with a linear naïve Bayes approach.

To evaluate intrarater reliability of the training data, 50 of the 300 automated patient interviewing software reports available were chosen for retriage by the same expert PCP. These reports were chosen randomly from the full set but checked to include an even variation of all available symptoms. Cohen κ was used to assess agreement with prior triage.

Furthermore, to evaluate the impact of missing data on our results, we reran the analyses with automated patient interviewing software reports with missing triage data excluded.

Ethical Considerations

The study was approved by the Swedish Ethical Review Authority on April 24, 2019 (reference number 2019-01516).

Data Sharing Statement

Data on triage decisions made by panel members and our expert PCP are available to the Department of Clinical Sciences in Malmö at Lund university, to the Department of Computational Science and Technology at the Royal Institute of Technology, and to Doctrin AB, Stockholm Sweden 10 years following publication. Data can be accessed for a prespecified purpose after approval by all 3 parties above.

Results

Comparisons Between the Three Models

After exclusion, 276 automated patient interviewing software reports were usable as labeled test-set data (Figure 1). The single-pass test results as well as cross-validation outcomes are presented in Table 2. There was no evidence for rejecting the null hypothesis (P>.10), so the performance of all 3 classifiers is considered comparable even though one can observe a trend favorable for random forest.

Figure 1. Flowchart of automated patient interviewing software report exclusion criteria.

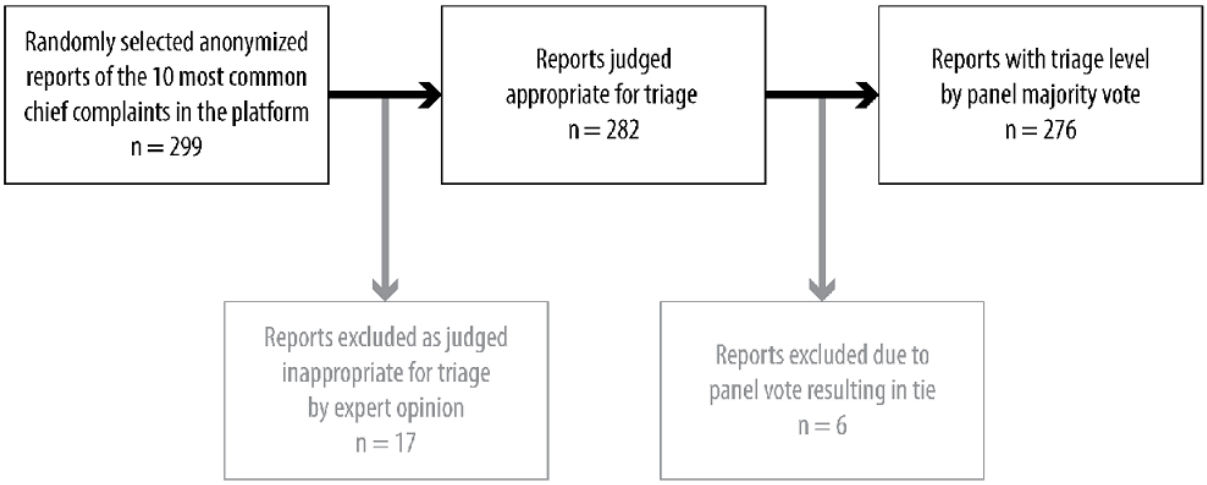

Table 2. Classification results obtained with naïve Bayes, logistic regression, and random forest in a single-pass test as well as in 10-fold cross-validation over the entire combined data set.

Classifier	Test results (training on the first and test on the second data subset), %	10-fold cross-validation (the first and second subsets combined), %[a]
Naïve Bayes	64.1	66.6 (7.6)
Logistic regression	60.1	64.5 (9.0)
Random forest	67.4	69.5 (7.7)

[a]The values for cross-validation are the mean and standard deviation of the classification accuracy obtained over 10 test folds.

Five Classifiers Versus One

Mean cross-validation accuracy calculated using the ensemble performance (majority vote) of the 5 naïve Bayes classifiers, each trained on the labels of one panel member, was 65.3% (SD 8.2%). Comparing this with the model, that is, the single naïve Bayes classifier (mean cross-validation accuracy 66.7% [SD 8.0%]), the null hypothesis could not be rejected (Wilcoxon signed-rank test, n=10, $P>.24$).

Decisive Features for Classification

Because the 3 classification approaches offer insights into feature weighing in the regression function that determines the classification boundary, we investigated more closely the distribution of such feature importance factors (see the "Methods" section). The results are inconclusive as the distribution is rather uniform and the pairwise correlations between feature rankings, Kendell τ (see the "Methods" section), produced by the classifiers are moderate (max 0.32 between naïve Bayes and random forest). This result implies that the given average level of accuracy can be achieved based on different sets of features.

Agreement Between Model and Human Triage

Because there was no statistically significant difference in the performance reported by the 3 classifiers, we decided to rely on the naïve Bayes approach in the next stages of our work due to its intuitive linear formulation. Cohen κ between the naïve Bayes model and the panel majority vote triage was 0.17 (Table 3), with 64% agreement. Excluding the information contained in brief free-form text resulted in the corresponding Cohen κ of 0.15. Within the reference group, average Cohen κ was 0.20, ranging from 0.10 to 0.30.

These results did not differ when analyses were rerun with missing cases excluded. No statistically significant difference in distribution of chief complaint symptoms could be found between reports with and without missing data (chi-square test, $P>.99$).

Using panel majority vote as the gold standard, the model correctly classified 74% (138/186) of nonurgent cases, but only 42% (38/90) of urgent cases. Adding free-form text data had a negligible effect on these numbers (Table 4).

When 50 automated patient interviewing software reports were selected for retriage by our selected expert PCP, Cohen κ was 0.55 with 78% agreement between retriage and previous triage.

Table 3. Assessment of the triage performance: agreement between the naïve Bayes model and each panel member as well as their majority vote, and average interrater agreement among the panel members.[a]

Panel	Panel member versus naïve Bayes model (Cohen κ)	Panel member versus rest panel members (Cohen κ)
PCP1	0.09	0.21
PCP2	0.03	0.21
PCP3	0.24	0.18
PCP4	0.08	0.21
PCP5	0.13	0.17
Majority vote	0.17	N/A

[a]PCP1 had the least amount of clinical experience, whereas PCP4 and PCP5 had the most amount of clinical experience.

Table 4. Contingency table of model triage with panel majority vote as the gold standard.

	Truly urgent	Falsely nonurgent	Truly nonurgent	Falsely urgent
Naïve Bayes model trained on full information including brief free-form text	42% (38 out of 90 cases voted urgent)	58% (52 out of 90 cases voted urgent)	74% (138 out of 186 cases voted nonurgent)	26% (48 out of 186 cases voted nonurgent)
Naïve Bayes model trained with brief free-form text information excluded	42% (38 out of 90 cases voted urgent)	58% (52 out of 90 cases voted urgent)	73% (135 out of 186 cases voted nonurgent)	27% (51 out of 186 cases voted nonurgent)

Discussion

Principal Results

To our knowledge, this is the first study to evaluate human versus ML performance in primary care triage based on a digitalized patient history. The first principal finding of this investigation was that interrater reliability in human triage using automated patient interviewing software reports is low (Cohen κ 0.20). Consequently, our second principal finding was that interrater triage reliability between a statistical model trained on automated patient interviewing software reports and a human panel was low (Cohen κ 0.17).

Findings were robust when cases with missing triage data were excluded from the analysis. The performance of the model was mostly decided by the surveys as removing the free-form text had only marginal impact on Cohen κ (reduced to 0.15). Furthermore, the intrarater reliability was moderate, as seen by retriage of 50 automated patient interviewing software reports by the same PCP (Cohen κ 0.55).

Comparison With Prior Work

While we acknowledge that κ values seldom are comparable across studies [19], previous data have generally found high interrater reliability between triage nurses [20-22]. However, these studies were conducted in high-acuity emergency department settings, where indicators of urgency arguably are more clearly defined [23].

The primary care setting presents a particular challenge in that conditions are of low acuity, making the line between urgent and nonurgent care more difficult to draw. This is supported by the low intrarater agreement for our expert PCP as well as the low agreement between our panel members. Indeed, acquiring a true gold standard for triage is a well-known issue [24]. “Correct” triage is difficult to define, and thus difficult to label and automate using ML. We could not identify any particular features in the data that were linked to the model’s triage decision. As far as the clinicians are concerned, we did not study their clinical reasoning before reaching a triage decision, that is, we do not know on which features their decision was based.

Interpretation

A well-known bottleneck for the creation of reliable ML algorithms is the lack of large enough amounts of labeled training data but this study calls the reliability of labels themselves into question. Labeled data need to be consistent across different raters and over time. Consequently, while adding more automated patient interviewing software data to the training set exploited by the model could improve interrater reliability with humans, the interrater reliability between the humans themselves sets a limit on how useful an algorithm could be if labels are fully decided from human data. While the addition of free-form text did not offer any advantage to the performance of the model, as assessed by our gold standard, it is possible that larger amounts of free-text data would allow the model to leverage these data for improved performance.

Human clinical decision making is likely more prone to be affected by externalities such as stress and mental fatigue [25]. Such externalities may have been present to different extents among our panel, resulting in markedly variable triage decisions compared with each other and the model.

Furthermore, the low agreement between the panel and the model in our study may be due to the fact that variation in human interpretation of text-based cues from automated patient interviewing software data in a primary care setting [26] prevents PCPs from determining urgency as consistently as the model, given access to the same amount of data. It should be noted, however, that in the clinical setting, PCPs would acquire additional data through the eVisit chat before making a triage decision.

The model is trained on triage data from a senior expert PCP, but results show no trend toward higher agreement between more senior PCPs and the model. This suggests that triage decision making depends more on other factors such as PCP temperament and risk aversion than mere experience [27].

Accepting the panel majority vote as the gold standard, nonurgent cases were more often classified correctly compared with urgent cases (74% [138/186] vs 42% [38/90], respectively), even though higher triage accuracy would be expected for urgent conditions where red flags are more well-defined [12]. Selection bias through a disproportionately larger amount of training data on nonurgent automated patient interviewing software reports may explain part of this disparity. On the contrary, this disproportionality may still be representative of a primary care cohort which would utilize such a digital tool for mostly low acuity conditions. However, given the low agreement between panel members, one may also question the suitability of use of the panel majority vote as the gold standard.

Strengths

This study has several strengths. First, it is one of few studies comparing human with ML performance using the same test data set for both groups. It is uniquely conducted in an eVisit primary care setting, where the need for reduced workload is high and where the ML algorithm has access to the same data as the clinician in the eVisit setting would. This contrasts with clinical or electronic health record–based ML tools which may not have access to key clinical data not recorded in the electronic health record [28]. Our data set was largely complete with only 1.4% missing data points. We also used training set data independent of validation test-set data, which is not always the case in other published research in the field [29]. Finally, the findings add nuance to the existing literature of ML versus human physicians [30].

Limitations

The results should be interpreted with consideration to several limitations. Our sample is not representative of a physical primary care population, as reports were acquired from an online consultation service database of self-selected patients being less likely to have life-threatening conditions [31]. Our data did not allow for out-of-sample external validation, as we do not know how these automated patient interviewing software reports ended up being triaged in their clinical setting. Lack of external validation also means that our low interrater reliability was likely overestimated [29]. However, even if externally valid

endpoint data could aid in defining a decision as "correct" retrospectively [32], defining "correct" triage prospectively may not be possible as some clinical outcomes cannot be predicted. In addition, the lack of consensus and use of a voting strategy in our panel are unconventional methods of defining a gold standard to compare ML-based performance and make comparison with other studies difficult. Future studies may use consensus techniques such as Delphi [33], incorporating PCP and emergency physician expertise, to mitigate lack of panel triage consensus.

Given the lack of agreement between our panel PCPs, using 1 expert PCP to provide training data may not be optimal. However, we did not observe any significant differences in cross-validation accuracy in this model compared with the ensemble performance of 5 models separately trained by each panel member.

Finally, our data set did not allow us to evaluate how the temporal provision of data affects the triage process in a way that would mimic the iterative clinical decision-making process. Thus, training data sets which make this possible may open up new opportunities for devising ML approaches that better mimic the human decision-making process.

Practical Implications

This study refutes implementation of the current ML model to fully automate binary triage in primary care, despite naïve Bayes being a reasonable ML algorithm to approach this problem. However, in the clinical setting, these reports are used as decision support in the interaction with patients, implying that uncertainties may be addressed by further interaction with the patient. Further development of the model with the suggestions made above may allow for fully automated triage in the future.

Conclusions

While digitalized patient histories have the potential to mitigate primary care workloads, leveraging patient history data to automate triage with ML methods is challenging given the difficulty for human physicians to triage consistently in a primary care setting. Future research should evaluate if external validation and temporal provision of training data may improve automated triage performance, as well as attempt to better identify which features drive triage decisions in a primary care setting.

Acknowledgments

The authors thank Johan Ekegren Gunnarsson and Sonja Petrovic Lundberg for preparing anonymized data sets and creating a graphic user interface for report review and triage. The authors also thank Markus Sandén (Doctrin AB, Stockholm, Sweden) for providing training data as our expert PCP. The authors thank residents Robin Back (Närhälsan Online and Närhälsan Ekmanska vårdcentral, Gothenburg, Sweden) and Lina Nyhlén (Vårdcentralen Lomma, Sweden), and specialists Miriam Pikkemaat (Department of Clinical Sciences in Malmö/Family Medicine, Lund university) and Annika Pahlmblad (Department of Clinical Sciences in Malmö/Family Medicine, Lund university) for being part of our expert panel. This study was partly funded by Vinnova, Sweden's Innovation Agency (Grant No. 2017-02348).

Authors' Contributions

AB, PH, SB, and PM were responsible for study concept and design; AB and AE were responsible for data acquisition; PH and SB performed analysis; AE was responsible for manuscript drafting; all authors were responsible for data interpretation, critical revision of the manuscript for important intellectual content, and final approval of the version to be published.

Conflicts of Interest

AB is the Chief Medical Officer of Doctrin AB, one of the project parties in this Vinnova-financed project. Other authors have no conflicts of interest to declare.

References

1. Colwill JM, Cultice JM, Kruse RL. Will generalist physician supply meet demands of an increasing and aging population? Health Aff (Millwood) 2008 Jan;27(3):w232-w241. [doi: 10.1377/hlthaff.27.3.w232] [Medline: 18445642]
2. van den Berg MJ, van Loenen T, Westert GP. Accessible and continuous primary care may help reduce rates of emergency department use. An international survey in 34 countries. Fam Pract 2016 Feb 28;33(1):42-50. [doi: 10.1093/fampra/cmv082] [Medline: 26511726]
3. Rhoades DR, McFarland KF, Finch WH, Johnson AO. Speaking and interruptions during primary care office visits. Fam Med 2001;33(7):528-532. [Medline: 11456245]
4. Berner ES, Kasiraman RK, Yu F, Ray MN, Houston TK. Data quality in the outpatient setting: impact on clinical decision support systems. AMIA Annu Symp Proc 2005:41-45 [FREE Full text] [Medline: 16778998]
5. Burnett SJ, Deelchand V, Franklin BD, Moorthy K, Vincent C. Missing clinical information in NHS hospital outpatient clinics: prevalence, causes and effects on patient care. BMC Health Serv Res 2011 May 23;11(1):114 [FREE Full text] [doi: 10.1186/1472-6963-11-114] [Medline: 21605359]
6. Campbell JL, Fletcher E, Britten N, Green C, Holt TA, Lattimer V, et al. Telephone triage for management of same-day consultation requests in general practice (the ESTEEM trial): a cluster-randomised controlled trial and cost-consequence analysis. The Lancet 2014 Nov;384(9957):1859-1868. [doi: 10.1016/s0140-6736(14)61058-8] [Medline: 25098487]

7. Marklund B, Ström M, Månsson J, Borgquist L, Baigi A, Fridlund B. Computer-supported telephone nurse triage: an evaluation of medical quality and costs. J Nurs Manag 2007 Mar;15(2):180-187. [doi: 10.1111/j.1365-2834.2007.00659.x] [Medline: 17352701]
8. Shepherd G, Schwartz R. Frequency of incomplete medication histories obtained at triage. Am J Health Syst Pharm 2009 Jan 01;66(1):65-69. [doi: 10.2146/ajhp080171] [Medline: 19106346]
9. Ernesäter A, Engström M, Holmström I, Winblad U. Incident reporting in nurse-led national telephone triage in Sweden: the reported errors reveal a pattern that needs to be broken. J Telemed Telecare 2010 May 10;16(5):243-247. [doi: 10.1258/jtt.2009.090813] [Medline: 20457800]
10. Giesen P, Ferwerda R, Tijssen R, Mokkink H, Drijver R, van den Bosch W, et al. Safety of telephone triage in general practitioner cooperatives: do triage nurses correctly estimate urgency? Qual Saf Health Care 2007 Jun 01;16(3):181-184 [FREE Full text] [doi: 10.1136/qshc.2006.018846] [Medline: 17545343]
11. Zakim D. Development and significance of automated history-taking software for clinical medicine, clinical research and basic medical science. J Intern Med 2016 Sep 12;280(3):287-299 [FREE Full text] [doi: 10.1111/joim.12509] [Medline: 27071980]
12. Semigran H, Linder J, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 08;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
13. Verzantvoort NCM, Teunis T, Verheij TJM, van der Velden AW. Self-triage for acute primary care via a smartphone application: Practical, safe and efficient? PLoS One 2018 Jun 26;13(6):e0199284 [FREE Full text] [doi: 10.1371/journal.pone.0199284] [Medline: 29944708]
14. Shafaf N, Malek H. Applications of Machine Learning Approaches in Emergency Medicine; a Review Article. Arch Acad Emerg Med 2019;7(1):34 [FREE Full text] [Medline: 31555764]
15. Swedish Association of Local Authorities and Regions (SALAR). Statistics About Health Care and Regional Development 2015: Operations and Economy in Counties and Regions. Stockholm: Swedish Association of Local Authorities and Regions (SALAR); 2016:978-991.
16. Lebanon G, Mao Y, Dillon J. The locally weighted bag of words framework for document representation. J Mach Learn Res 2007;8(12/1/2007):2405-2441. [doi: 10.5555/1314498.1314576]
17. Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ, Population H. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. Popul Health Metr 2011 Aug 04;9:29 [FREE Full text] [doi: 10.1186/1478-7954-9-29] [Medline: 21816105]
18. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 2016 Jul 02;20(1):37-46. [doi: 10.1177/001316446002000104]
19. Feinstein AR, Cicchetti DV. High agreement but low Kappa: I. the problems of two paradoxes. Journal of Clinical Epidemiology 1990 Jan;43(6):543-549. [doi: 10.1016/0895-4356(90)90158-l] [Medline: 2348207]
20. Elias P, Damle A, Casale M, Branson K, Churi C, Komatireddy R, et al. A Web-Based Tool for Patient Triage in Emergency Department Settings: Validation Using the Emergency Severity Index. JMIR Med Inform 2015 Jun 10;3(2):e23 [FREE Full text] [doi: 10.2196/medinform.3508] [Medline: 26063343]
21. Grouse AI, Bishop RO, Bannon AM. The Manchester Triage System provides good reliability in an Australian emergency department. Emerg Med J 2009 Jul;26(7):484-486. [doi: 10.1136/emj.2008.065508] [Medline: 19546267]
22. Gerdtz MF, Collins M, Chu M, Grant A, Tchernomoroff R, Pollard C, et al. Optimizing triage consistency in Australian emergency departments: the Emergency Triage Education Kit. Emerg Med Australas 2008 Jun;20(3):250-259. [doi: 10.1111/j.1742-6723.2008.01089.x] [Medline: 18462405]
23. Widgren BR, Jourak M. Medical Emergency Triage and Treatment System (METTS): a new protocol in primary triage and secondary priority decision in emergency medicine. J Emerg Med 2011 Jun;40(6):623-628. [doi: 10.1016/j.jemermed.2008.04.003] [Medline: 18930373]
24. FitzGerald G, Jelinek GA, Scott D, Gerdtz MF. Emergency department triage revisited. Emerg Med J 2010 Feb;27(2):86-92. [doi: 10.1136/emj.2009.077081] [Medline: 20156855]
25. Allan JL, Johnston DW, Powell DJH, Farquharson B, Jones MC, Leckie G, et al. Clinical decisions and time since rest break: An analysis of decision fatigue in nurses. Health Psychol 2019 Apr;38(4):318-324. [doi: 10.1037/hea0000725] [Medline: 30896218]
26. Entezarjou A, Bolmsjö BB, Calling S, Midlöv P, Milos Nymberg V. Experiences of digital communication with automated patient interviews and asynchronous chat in Swedish primary care: a qualitative study. BMJ Open 2020 Jul 23;10(7):e036585 [FREE Full text] [doi: 10.1136/bmjopen-2019-036585] [Medline: 32709650]
27. Considine J, Botti M, Thomas S. Do knowledge and experience have specific roles in triage decision-making? Acad Emerg Med 2007 Aug;14(8):722-726 [FREE Full text] [doi: 10.1197/j.aem.2007.04.015] [Medline: 17656608]
28. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep 2016 Dec 17;6:26094 [FREE Full text] [doi: 10.1038/srep26094] [Medline: 27185194]
29. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/s2589-7500(19)30123-2]

30. Cook TS. Human versus machine in medicine: can scientific literature answer the question? The Lancet Digital Health 2019 Oct;1(6):e246-e247. [doi: 10.1016/s2589-7500(19)30124-4]
31. North F, Crane SJ, Stroebel RJ, Cha SS, Edell ES, Tulledge-Scheitel SM. Patient-generated secure messages and eVisits on a patient portal: are patients at risk? J Am Med Inform Assoc 2013 Nov 01;20(6):1143-1149 [FREE Full text] [doi: 10.1136/amiajnl-2012-001208] [Medline: 23703826]
32. Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Acad Emerg Med 2000 Mar;7(3):236-242 [FREE Full text] [doi: 10.1111/j.1553-2712.2000.tb01066.x] [Medline: 10730830]
33. Fry M, Burr G. Using the Delphi technique to design a self-reporting triage survey tool. Accid Emerg Nurs 2001 Oct;9(4):235-241 [FREE Full text] [doi: 10.1054/aaen.2001.0245] [Medline: 11855763]

Abbreviations

ML: machine learning
PCPs: primary care physicians

Edited by G Eysenbach; submitted 27.03.20; peer-reviewed by R Miotto, A Benis, D Gunasekeran; comments to author 27.04.20; revised version received 22.06.20; accepted 24.06.20; published 03.09.20

Please cite as:
Entezarjou A, Bonamy AKE, Benjaminsson S, Herman P, Midlöv P
Human- Versus Machine Learning–Based Triage Using Digitalized Patient Histories in Primary Care: Comparative Study
JMIR Med Inform 2020;8(9):e18930
URL: https://medinform.jmir.org/2020/9/e18930
doi: 10.2196/18930
PMID: 32880578