Application of Machine Learning to Enhance e-Triggers to Detect and Learn from Diagnostic Safety Events

Principal Investigator: Hardeep Singh
Team Members: Andrew J. Zimolzak, MD, MMSc[1], Daniel Murphy, MD, MBA[1], Li Wei, MS[1], Usman Mir, MBBS, MPH[1], Ashish Gupta, MD, MBA[1], Viralkumar Vaghani, MBBS, MPH, MS[1], Devika Subramanian, PhD, MS[2], Dean F. Sittig, PhD[3]

Organizations: [1]Houston Veterans Affairs Center for Innovations in Quality, Effectiveness and Safety, Michael E. DeBakey VA Medical Center, Houston, TX, Section of Health Services Research, Department of Medicine, Baylor College of Medicine, Houston, TX

[2]Department of Computer Science, Rice University, Houston, TX

[3]University of Texas – Memorial Hermann Center for Healthcare Quality & Safety, School of Biomedical Informatics, University of Texas Health Science Center at Houston, TX

Dates of Project: 9/30/2019 – 9/29/2023

Federal Project Officer: Stephen Raab

Acknowledgement of Agency Support: Agency of Healthcare Research and Quality

Grant Award Number: R01 HS27363

1. Structured Abstract

Purpose: To develop and test algorithms to detect ED records with evidence of missed opportunities of diagnoses (MOD) and then enhance e-trigger predictive values by using machine learning (ML) techniques.

Scope: Emergency departments (ED) are particularly at risk for diagnostic errors, as critical decisions are often made in the absence of a complete work-up. There is a lack of reliable and valid measures for diagnostic safety.[3] However, healthcare organizations (HCOs) can now leverage their electronic data for quality improvement.[4] Such tools may be superior to voluntary reporting, as large data sets can be mined for diagnostic safety events, reducing the number of records requiring human review.

Methods: We developed, refined, tested, and applied Safer Dx e-triggers to detect potential records with diagnostic errors in the ED. Trained clinicians used standardized data collection instruments to review charts flagged by each e-trigger. We also tested whether machine learning (ML) can enhance e-trigger predictive values and emulate human chart reviewers at a larger scale.

Results:
Aim 1
The PPV varied from 10 % to 52.4 % in six triggers.

Aim 2
For Trigger 1 (treat-and-release ED visit for dizziness followed by hospitalization within 30 days), the best-performing ML algorithm (random forest) correctly identified 31 of the 33 true positives and 30 of the 35 true negatives (86% PPV). For Trigger 2 (hospitalizations within 10 days after treat-and-release ED visit for abdominal pain), ML correctly identified 26 of 31 true positives and 71 of 73 true negatives (93% PPV).

Key Words: Diagnostic error, machine learning

2. Purpose

In this study, we developed and tested algorithms to detect, or "trigger," review of ED records with evidence of missed opportunities of diagnoses (MOD) and then explored machine learning (ML) techniques to predict diagnostic errors using EHR-enriched data.

Aim 1 – To develop, refine, test, and apply Safer Dx e-triggers to enable detection, measurement, and learning from diagnostic errors in diverse emergency department (ED) settings. We will calculate the frequency of diagnostic errors in the ED based on these e-triggers and describe the burden of preventable diagnostic harm.

Aim 2 - To explore machine learning techniques that yield robust, accurate models to predict diagnostic errors using EHR-enriched data derived from expert-labeled patient records containing diagnostic errors (from Aim 1).

Hypothesis 1: Use of unsupervised deep learning will facilitate large-scale retrieval of patient records with high probability of containing diagnostic errors.

Hypothesis 2: Supervised machine-learned deep neural network models learned from manually labeled chart reviews (byproduct of Aim 1), enriched with similar records retrieved using unsupervised deep learning, can (a) identify diagnostic errors with higher sensitivity, specificity, and positive predictive value than manually derived e-triggers and (b) scale up Safer Dx e-trigger development and deployment processes of Aim 1 by reducing the number of refinement cycles and manual chart reviews needed.

3. Scope

Background:

Reducing diagnostic errors is a monumental challenge for patient safety.[1] The National Academies of Sciences, Engineering, and Medicine's (NASEM) report, Improving Diagnosis in Health Care,[2] recommends that accrediting organizations and Medicare "require that healthcare organizations have programs in place to monitor the diagnostic process and identify, learn from, and reduce diagnostic errors and near misses in a timely fashion."[2] Four years later, external incentives to implement these recommendations are absent, and guidance on how to develop these programs is limited. Improving diagnosis remains a rare organizational priority. Progress in reducing diagnostic errors is slow partly due to poorly defined methods to identify errors, high-risk situations, and adverse events. Emergency departments (ED) are particularly at risk for errors and preventable harm. The diagnostic process in the ED occurs in time- and information-constrained circumstances, and critical decisions are often made in the absence of a complete work-up.

There is a compelling need to create measurement methods that provide diagnostic safety data to clinicians and leaders who in turn act upon these data to reduce variations in care and prevent diagnostic harm. Though measurement has become foundational to quality and safety improvement, no reliable and valid measures for diagnostic safety currently exist.[3] However, healthcare organizations (HCOs) now have an opportunity to explore health information technology capabilities to use their own ever-increasing stores of electronic data for learning, research, and quality improvement.[4] Electronic trigger (e-trigger) tools, which mine vast amounts of clinical and administrative data to identify signals for likely adverse events,[4] offer a promising method to do so. Such tools are more efficient and effective than voluntary reporting and offer the ability to quickly mine large data sets for diagnostic safety events, reducing the number of records requiring human review to those at highest risk of harm.

Context:

The use of Machine Learning (ML) to estimate the likelihood of a diagnostic error from detailed patient-level information may allow for improved performance over existing rules-based systems and may stimulate the development and use of more sophisticated second-generation e-trigger tools. The goal of supervised ML is to learn a model that takes several situation-specific inputs (independent variables related to clinical data, such as vital signs, test results, exam findings, timings of notes or test results) and uses that to predict an output/outcome or probability of an outcome (dependent variable, such as potential diagnostic error).[5] More importantly, ML methods can derive value from the numerous unlabeled examples (or unreviewed charts) as well as the expert-labeled examples.[6] We will develop a system that can automatically retrieve suspected cases of diagnostic error from electronic data repositories, using similarity to expert-labeled charts as a criterion, thus serving as a set of second-generation e-triggers. The similarity-based retrieval process thus offers the potential for expanding the scope of discovery of diagnostic errors over and beyond previously developed e-triggers.

Settings:

The Department of Veterans Affairs (VA) provides care to 9 million veterans at 1321 healthcare facilities, including 172 medical centers and 1138 outpatients on a comprehensive, in-house--designed EHR that has been integrated into all facilities since 2000. It also caters to AHRQ priority populations.

Participants:

Veteran patients, Adult (18-64 years), Geriatric (65+ years)

Incidence:

Although certainty or timeliness of diagnosis is not always achievable in the ED given its challenging context, patients' health concerns need to be addressed.[7] Diagnostic errors appear frequently in ED malpractice claims.[8] There are about 141 million annual ED visits in the US. A conservative estimate of 5% of adults experiencing diagnostic errors translates to about 7 million cases of ED-based diagnostic errors.[2,9,10,11] About half of all diagnostic errors have potential for severe or permanent harm[12-14]---many by adding risk from delaying optimal treatment or providing suboptimal and even dangerous alternative actions. Diagnostic errors result from a complex interplay between various patient (e.g., health literacy, typicality of presenting symptoms, complexity, and behaviors), provider/care-team (e.g., cognitive load, information gathering and synthesis), and systems (e.g., health information technology, crowding, and interruptions) factors.

Prevalence:

The frequency of diagnostic errors in the ED is largely unknown but likely is higher than comparable estimates for outpatient care.[9, 15, 16, 17.]

4. Methods

Study Design:

Aim 1

a. Trigger identification and prioritization: Our advisory committee representing diverse perspectives (practicing ED physicians and clinical operations stakeholders, safety/ED researchers, and informaticians) helped identify which of our prior pilot triggers were ready for further development and application and identified high-priority targets for development of additional triggers.
b. Defining trigger criteria: The advisory panel helped operationalize triggers and identify details of the context to focus on for further development based on potential impact on care and feasibility for trigger development.
c. Trigger algorithm construction and data retrieval: Queries were designed to run on VA data warehouse platforms. Each trigger algorithm used Structured Query Language (SQL) to automatically extract structured data fields, evaluate for certain pre-determined criteria, and output a list of patients who met the trigger criteria.
d. Trigger testing and refinement: We selected six triggers and extracted a sample data set to perform chart reviews to determine what additional information must be added or removed to improve the PPV of the trigger algorithm. Clinician reviewers iteratively evaluated samples of at least 25 trigger-positive cases for each trigger to identify clinical clues that were inappropriately captured or incorrectly ignored by the trigger algorithm. This review was repeated until no further improvements were identified.
e. Evaluation of trigger performance: After the algorithm for each of the six triggers was finalized, the revised trigger criteria was validated by conducting record reviews on a second unique cohort (validation cohort) on retrospective data.
f. Descriptive Analysis of MODs and Knowledge Synthesis: We performed structured review of selective medical records, using the Safer Dx Instrument[18] developed by our team, to determine missed opportunities and then used a taxonomy developed by our team to determine process breakdowns and patient harms related to MODs. We used descriptive statistics to depict characteristics, including clinical conditions involved, and the associated process breakdowns and level of harm from diagnostic errors.

Aim 2

Test Hypothesis 1 (unsupervised methods for large-scale retrieval):

a. Data Preparation: We performed pre-processing and standardization of the various data types in the EHR. We prepared vector representations of EHR data—both text notes and structured clinical data. This resulted in 181 candidate predictors, including demographics, lab values, vital signs, medications, orders in the emergency room and subsequent hospital admission (tests and consults), visit times, and risk factors (past diagnoses).
b. Unsupervised Learning: We attempted unsupervised ML and clustering in several ways.

First, discharge summaries of patients with a primary diagnosis of stroke had concepts extracted by natural language processing (NLP), and this high-dimensional data was embedded in a space that allows us to measure the distance between two notes. This is similar to the word2vec approach but with one vector for an entire discharge summary. This mainly served as proof of concept of NLP and clustering on our hardware platform, although the goal was also to search for stroke subtypes. Second, we generated embeddings of each patient's structured EHR data and applied clustering to investigate subtypes of trigger-positive patients. Third, we applied an encoder-decoder to ER notes and short case summaries generated by reviewers in Aim 1. The goal was for the algorithm to produce a case summary in the style of a human chart reviewer.

c. Model Validation: Natural language processing of text notes was performed in several ways (embedding stroke discharge summaries, and an encoder-decoder applied to ER notes and trained to produce a case summary in the style of a human chart reviewer). However, these methods were not developed enough to incorporate their output into the main analysis (which used supervised ML and structured EHR data only—see Hypothesis 2).

Test Hypothesis 2 (supervised ML improves performance of diagnostic error e-triggers):

a. We used rules-based e-triggers from Aim 1 (based on expert input and existing frameworks) to find possible missed opportunities in diagnosis (MODs) in emergency care. Using Veterans Affairs national EHR data that contains records on >20 million unique individuals, we used two high-risk e-triggers: (a) patients with stroke risk factors who were discharged from an emergency department (ED) after presenting with dizziness or vertigo and subsequently hospitalized for stroke or TIA within 30 days and (b) patients discharged from ED with abdominal pain and abnormal temperature who were subsequently hospitalized within 10 days. We used labels from Aim 1 to train ML (clinicians reviewed a random sample of charts flagged by each e-trigger and labeled each chart as MOD or no MOD).
b. Clinician-labeled charts were divided into training and test sets. ML methods were regularized logistic regression and random forests (with limited maximum tree depth to mitigate overfitting). From the set of 181 candidate predictors, a smaller set was pre-selected based on bivariate association with MOD by t-test or chi-squared test as appropriate, with a threshold of $P = 0.1$. After pre-selection, 42 predictors remained for training ML algorithms. We calculated positive predictive values (PPV) of the rules-based trigger and compared these to the PPV of the ML-enhanced trigger.

Data Sources/Collection:

Veteran medical records were selected by the electronic trigger from the VA data warehouses. E-triggers were applied to >9 million patient records in the VA's corporate data warehouse. We used records from the year 2019 for Trigger 2, Trigger 3, Trigger 5, and Trigger 6. We used year 2016-2019 for Trigger 1 and year 2018-2019 for Trigger 4 to get 100 trigger-positive cases.

Interventions: N/A

Measures:

Trigger 1	Stroke hospitalization within 30 days of treat-and-release ED visit for dizziness in patients with two or more stroke risk factors
Trigger 2	Hospitalizations within 10 days after treat-and-release ED visit for abdominal pain with high (T>99.5 F or >37.5 C) or low body temperature (T<96.8 F or <36.3 C)
Trigger 3	3A: Treat-and-release ED visit followed by unscheduled return to ED 3B: Treat-and-release ED visit followed by unexpected hospitalization within 10 days
Trigger 4	Treat-and-release ED visit with benign or symptomatic diagnosis followed by a return ED visit for a serious diagnosis within 7 days for chest pain-myocardial infarction (MI) or pulmonary edema (PE) dyad, 10 days for abdominal pain-appendicitis or perforated diverticulitis dyad, 14 days for headache-subarachnoid hemorrhage (SAH) or meningitis dyads, 30 days for vertigo/dizziness-stroke dyad, and syncope-cardiac arrest, ventricular tachycardia (VT), ventricular fibrillation (VF), sick sinus syndrome, or atrioventricular block.
Trigger 5	Abnormal tests ordered during treat and release ED visits that were not followed up within 14 days (tests included positive urine culture, positive blood culture, and elevated thyroid-stimulating hormone).
Trigger 6	Treat-and-release ED visit followed by a primary care visit within 7 days with a different, more serious diagnosis.
Trigger Negative	Patients with treat-and-release ED visit who did not meet triggers 1, 2, 3, or 4.

Limitations:

We recognize several methodological and logistical challenges. We were not able to capture every missed diagnostic opportunity. Still, using e-triggers is far more efficient than the current standard of care, as there are currently no other comprehensive and widely adopted methods to identify, measure, and monitor diagnostic errors. Therefore, this approach represents a significant improvement over current status.

First, although chart reviews are one of the best available methods to detect MODs, the analyses are dependent on documentation, which may be suboptimal. To address this, we oversampled the identified charts to ensure that charts with adequate information are chosen for determination of MODs. Second, retrospective methods may introduce hindsight bias in judgments of error determination.[19,20] We did not interview providers as part of the analysis, which limited our ability to determine the breakdown point in the diagnostic process. Another limitation is that agreement for diagnostic errors tends to be much lower than for other types of errors.[21] Our Safer Dx instrument was a potentially objective way to detect MODs and partially overcome limitations of lower reviewer agreement for diagnostic errors. Third, we recognize that triggers did not identify all the errors that occur in diverse ED settings or across all types of providers (e.g., physicians vs. advanced practitioners), populations, or types of diseases (acute or chronic). Triggers such as those used in this study inevitably missed some errors, especially errors related to missed diagnosis of chronic conditions that are only diagnosed over a prolonged period of time (such as cancer) and errors related to follow-up of patients.

Fourth, detection of different types of errors required different methods, some of which were outside the scope of the study.

5. Results

Principal Findings:

Aim 1

E-Triggers	Reviewed Charts	Included for Analysis	MOD	No MOD	PPV
Trigger 1	100	88	47	41	47.00%
Trigger 2	120	103	31	72	25.80%
Trigger 3	200	186	34	152	17.00%
Trigger 3A	100	98	11	87	11.00%
Trigger 3B	100	88	23	65	23.00%
Trigger 4	100	81	18	63	18.00%
Trigger 5	105	105	55	50	52.40%
Trigger 6	100	92	10	82	10.00%
Trigger Negative	100	100	1	99	N/A

Abbreviations:
MOD= Missed Opportunities in Diagnosis
PPV=Positive Predictive Value
N/A=Not applicable

Aim 2

For Trigger 1 (dizziness e-trigger), reviewers identified MODs in 47 of 88 flagged records (47% PPV). The best-performing ML algorithm (random forest) correctly identified 31 of the 33 true positives and 30 of the 35 true negatives (86% PPV). Findings on chart review included lower documentation quality and lower rate of certain neurological examination components, as described previously.[3] For Trigger 2 (abdominal pain e-trigger), reviewers identified 31 MODs in 103 flagged records (30% PPV). ML correctly identified 26 of 31 true positives and 71 of 73 true negatives (93% PPV). Examples of diagnostic errors included missed diagnoses of cholangitis, cholecystitis, and infectious colitis.

Outcomes:

Aim 1

Our primary outcome was missed opportunities in diagnosis (MODs).[22] We determined whether each trigger-identified chart meets all the trigger criteria and either had a MOD ("true trigger positive") or not ("false trigger positive") by examining all relevant sections of the EHR (e.g., progress notes, consultations, laboratory, radiology, referral menus) for details. For these reviews, we relied on our validated Safer Dx instrument.[23] We also collected information on management errors in addition to diagnostic errors. Certain provider (e.g., years of experience, role [MD, NP, PA]), patient (e.g., age, sex, comorbidities), and available ED (e.g., location) characteristics also were collected as potential predictors of error to be used in exploratory subanalyses.

Aim 2

For Trigger 1 (dizziness e-trigger), PPV increased from 47% to 86% with the use of ML methods. Likewise, for Trigger 2 (abdominal pain e-trigger), PPV increased from 25.8% to 93% with the use of ML.

Discussion:

Aim 1

Our Safer Dx framework of e-triggers offers an efficient method to detect missed opportunities in diagnosis for the patients who presented to the ED. Currently, health systems are not using any sophisticated detection methods for diagnostic error and are finding them occasionally and passively through rudimentary incident reporting systems. Although our trigger performed modestly, PPVs to identify events of interest have been traditionally lower in the area of patient safety.[24-28] This approach could be applied to other EHR data warehouses to retrospectively identify diagnostic errors for learning and improvement purposes. E-trigger enhanced review procedures overcome several limitations of other safety measurement methods.[29] E-triggers could strengthen patient safety improvements efforts in health systems with limited resources and competing demands on quality measurement.

Aim 2

Machine learning enhanced the accuracy of electronic triggers to identify missed opportunities in diagnosis. Limitations include the time needed to prepare the variables used by ML, although, once this is done, the algorithm can run at a large scale. The relatively small number of expert-labeled records may impede the ability of ML to use all structured data and to estimate test set performance.

Conclusions:

A portfolio of e-triggers achieves reasonable accuracy to identify multiple types of diagnostic errors in emergency care. Implementing this portfolio in routine ED care nationally could accelerate quality improvement efforts to reduce diagnostic errors in ED settings.

Next steps for ML algorithms to identify diagnostic errors include incorporating clinical note text as a source of missed opportunity prediction and increasing the expert-labeled records on which the approach is tested. Furthermore, ML can be tested on additional emergency setting e-triggers (apart from only Triggers 1 and 2), and on e-triggers from other care settings. Machine learning shows promise as a tool to efficiently identify diagnostic errors for research and quality improvement purposes.

Significance:

ML-enhanced e-triggers could advance an organization's ability to monitor diagnostic errors for research, learning, and improvement.

Implications:

Rules-based e-triggers retrieve many cases of missed opportunities in diagnosis, mixed with substantial cases with no missed opportunity. If an initial set of manually reviewed charts are used to train an ML algorithm to separate "miss" from "no miss" cases, this would substantially reduce the burden of clinician-dependent manual chart review traditionally used for case analysis.

6. List of Publications and Products

Presentations:

Society to Improve Diagnoses in Medicine (SIDM) 2019

- Vaghani, V., Mushtaq, U., Zimolzak, A., Sittig, D., & Singh, H. Performance of an e-Trigger to Detect Missed Stroke Diagnosis in Patients with Headache or Dizziness Symptoms in Emergency Department [abstract]. In proceedings of The Diagnostic Error in Medicine 12th Annual International Conference [Internet]. SIDM: 2019 Nov 10-13; Washington, DC. Diagnosis. 2019;6(4): eA1-eA96. https://doi.org/10.1515/dx-2019-0075

SIDM 2020

- Vaghani, V., Murphy, D., Memon, S., Zimolzak, A., Subramanian, D., Upadhyay, D., Singh, H. Portfolio of e-Triggers to Identify Diagnostic Errors in Emergency Departments: A Prioritization Exercise [abstract]. In proceedings of The Diagnostic Error in Medicine 13th Annual International Conference [Internet]. SIDM: 2020 Oct 19-21 Diagnosis, vol. 8, no. 2, 2021, pp. eA1-eA74. https://doi.org/10.1515/dx-2021-0012
- Vaghani, V., Kapadia, P., Zimolzak, A., Singh, H., Subramanian, D. Development of a Machine Learning Enhanced Trigger to Detect Diagnostic Error [abstract]. In proceedings of The Diagnostic Error in Medicine 13th Annual International Conference [Internet]. SIDM: 2020 Oct 19-21 Diagnosis, vol. 8, no. 2, 2021, pp. eA1-eA74. https://doi.org/10.1515/dx-2021-0012

SIDM 2021

- Vaghani, V., Gupta, A., Kapadia, P., Wei, L., Sittig D. F., Singh, H. Developing and Testing a Bundle of E-Triggers to Identify Diagnostic Errors in Emergency Departments [abstract]. In proceedings of The Diagnostic Error in Medicine 14th Annual International Conference [Internet]. Diagnosis. 2022;9(2): 294-386. https://doi.org/10.1515/dx-2022-0024

SIDM 2022

- Vaghani, V., Gupta, A., Mir, U., Murphy, D., Wei, L., Sittig D. F., Singh, H. Performance of E-Triggers to Identify Diagnostic Errors in High-risk Clinical Presentation in ED [abstract]. In proceedings of The Diagnostic Error in Medicine 15th Annual International Conference [Internet]. SIDM: 2022 October 16-18; Minneapolis, Minnesota Diagnosis. 2023;10(2): A1-A77. https://doi.org/10.1515/dx-2023-0006
- Vaghani, V., Mushtaq, U., Murphy, D., Li, W., Mir, U., Sittig, D., Singh, H. Measuring Missed Opportunity in Diagnosis in Abnormal Test Result Follow up In Post ED Visit [abstract]. In proceedings of The Diagnostic Error in Medicine 15th Annual International Conference [Internet]. SIDM: 2022 October 16-18; Minneapolis, Minnesota. Diagnosis. 2023;10(2): A1-A77. https://doi.org/10.1515/dx-2023-0006

SIDM 2023

- Vaghani, V., Gupta, A., Mir, U., Murphy, D., Wei, L., Sittig D. F., Singh, H. A Portfolio of E-Triggers to Measure Multiple Types of Diagnostic Errors in Emergency Care. Poster session presented at Society to Improve Diagnosis in Medicine (SIDM) 16th Annual International Conference; 2023 October 8-11; Cleveland, OH.

- Zimolzak AJ, Yu M, Wu Y, Wei L, Mir U, Gupta A, Vaghani V, Hassan A, Mower J, Subramanian D, Singh H. Machine Learning for Enhanced Electronic Trigger Detection of Diagnostic Errors. Poster session presented at Society to Improve Diagnosis in Medicine (SIDM) 16th Annual International Conference; 2023 October 8-11; Cleveland, OH. *Awarded best oral presentation of the conference as a whole.*

HSR&D/QUERI National Conference 2023

- Vaghani, V., Gupta, A., Mir, U., Mustaq, U., Wei, L., Sittig D. F., Singh, H. Performance of E-Triggers to Identify Diagnostic Errors in High-risk Clinical Presentation in ED. Poster session presented at: Health Services Research and Development Service (HSR&D) and Quality Enhancement Research Initiative (QUERI) National Conference; 2023 February 8-10; Baltimore, MD.

<u>Publications</u>

- Vaghani V, Wei L, Mushtaq U, Sittig DF, Bradford A, Singh H. Validation of an electronic trigger to measure missed diagnosis of stroke in emergency departments. Journal of the American Medical Informatics Association. 2021 Oct 1;28(10):2202-11.

<u>Publications (Under Review)</u>

- Zimolzak AJ, Wei L, Mir U, Gupta A, Vaghani V, Subramanian D, Singh H. Machine Learning to Enhance Electronic Detection of Diagnostic Errors.

<u>Publications (In Preparation)</u>

- Vaghani V, Mir U, Gupta A, Zimolzak AJ, Murphy. D, Khan. S, Wei L, Sittig DF, Singh H. Application of e-Triggers to Detect and Learn from Diagnostic Errors in Emergency Departments.

References

1. Singh H, Graber ML. Improving Diagnosis in Health Care — The Next Imperative for Patient Safety. New England Journal of Medicine. 2015 Dec 24;373(26):2493–5.
2. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care.
3. Singh H, Graber ML, Hofer TP. Measures to improve diagnostic safety in clinical practice. Journal of patient safety. 2019 Dec 1;15(4):311-6.
4. Murphy DR, Meyer AN, Sittig DF, Meeks DW, Thomas EJ, Singh H. Application of electronic trigger tools to identify targets for improving diagnostic safety. BMJ Quality & Safety. 2019 Feb 1;28(2):151-9.
5. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: springer; 2009 Aug.
6. Chapelle O, Schölkopf B, Zien A. A discussion of semi-supervised learning and transduction in Semi-supervised learning 2006 (pp. 473-478). MIT Press.
7. Gerolamo AM, Jutel A, Kovalsky D, Gentsch A, Doty AM, Rising KL. Patient-identified needs related to seeking a diagnosis in the emergency department. Annals of Emergency Medicine. 2018 Sep 1;72(3):282-8.
8. Gurley KL, Grossman SA, Janes M, Yu-Moe CW, Song E, Tibbles CD, Shapiro NI, Rosen CL. Comparison of emergency medicine malpractice cases involving residents to nonresident cases. Academic Emergency Medicine. 2018 Sep;25(9):980-6.
9. Singh H, Meyer AN, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ quality & safety. 2014 Sep 1;23(9):727-31.
10. Croskerry P, Sinclair D. Emergency medicine: a practice prone to error?. Canadian Journal of Emergency Medicine. 2001 Oct;3(4):271-6.
11. Graber ML. The incidence of diagnostic error in medicine. BMJ quality & safety. 2013 Oct 1;22(Suppl 2):ii21-7.
12. Newman-Toker DE, Pronovost PJ. Diagnostic errors—the next frontier for patient safety. Jama. 2009 Mar 11;301(10):1060-2.
13. Zwaan L, Thijs A, Wagner C, van der Wal G, Timmermans DR. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. Academic Medicine. 2012 Feb 1;87(2):149-56.
14. Singh H, Giardina TD, Meyer AN, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA internal medicine. 2013 Mar 25;173(6):418-25.
15. Medford-Davis L, Park E, Shlamovitz G, Suliburk J, Meyer AN, Singh H. Diagnostic errors related to acute abdominal pain in the emergency department. Emergency Medicine Journal. 2016 Apr 1;33(4):253-9.
16. Newman-Toker DE, Moy E, Valente E, Coffey R, Hines AL. Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample. Diagnosis. 2014 Jun 1;1(2):155-66.
17. Okafor N, Payne VL, Chathampally Y, Miller S, Doshi P, Singh H. Using voluntary reports from physicians to learn from diagnostic errors in emergency medicine. Emergency Medicine Journal. 2016 Apr 1;33(4):245-52.
18. Singh H, Khanna A, Spitzmueller C, Meyer AN. Recommendations for using the Revised Safer Dx Instrument to help measure and improve diagnostic safety. Diagnosis. 2019 Nov 26;6(4):315-23.
19. Fischhoff B. Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human perception and performance. 1975 Aug;1(3):288.
20. McNutt RA, Abrams MR, Hasler MS. Diagnosing diagnostic mistakes. AHRQ Web M&M (published online May 2005); http: www.webmm.ahrq. gov/printview.aspx, 2005.

21. Thomas EJ, Lipsitz SR, Studdert DM, Brennan TA. The reliability of medical record review for estimating adverse event rates. Annals of internal medicine. 2002 Jun 4;136(11):812-6.
22. Singh H. Helping health care organizations to define diagnostic errors as missed opportunities in diagnosis. Joint Commission journal on quality and patient safety. 2014 Mar 1;40(3):AP1.
23. Al-Mutairi A, Meyer AN, Thomas EJ, Etchegaray JM, Roy KM, Davalos MC, Sheikh S, Singh H. Accuracy of the Safer Dx instrument to identify diagnostic errors in primary care. Journal of general internal medicine. 2016 Jun;31:602-8.
24. Silva MD, Martins MA, Viana LD, Passaglia LG, de Menezes RR, Oliveira JA, da Silva JL, Ribeiro AL. Evaluation of accuracy of IHI Trigger Tool in identifying adverse drug events: a prospective observational study. British journal of clinical pharmacology. 2018 Oct;84(10):2252-9.
25. Lipitz-Snyderman A, Classen D, Pfister D, Killen A, Atoria CL, Fortier E, Epstein AS, Anderson C, Weingart SN. Performance of a trigger tool for identifying adverse events in oncology. Journal of oncology practice. 2017 Mar;13(3):e223-30.
26. Klein DO, Rennenberg RJ, Koopmans RP, Prins MH. The ability of triggers to retrospectively predict potentially preventable adverse events in a sample of deceased patients. Preventive medicine reports. 2017 Dec 1;8:250-5.
27. Davis J, Harrington N, Fagan HB, Henry B, Savoy M. The accuracy of trigger tools to detect preventable adverse events in primary care: a systematic review. The Journal of the American Board of Family Medicine. 2018 Jan 1;31(1):113-25.
28. Musy SN, Ausserhofer D, Schwendimann R, Rothen HU, Jeitziner MM, Rutjes AW, Simon M. Trigger tool–based automated adverse event detection in electronic health records: systematic review. Journal of medical internet research. 2018 May 30;20(5):e198.
29. Liberman AL, Newman-Toker DE. Symptom-Disease Pair Analysis of Diagnostic Error (SPADE): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data. BMJ quality & safety. 2018 Jul 1;27(7):557-66.