1. Title Page

Title: Utility of Predictive Systems in Diagnostic Errors (UPSIDE)

Principal Investigator and Team Members

Research Team

UCSF
Andrew Auerbach, MD, MPH (PI)
Tiffany Lee, BA
Gilmer Valdes, PhD, DABR
Colin Hubbard, PhD
Sandra Oreper, PhD
Mark Segal, PhD
John Boscardin, PhD
Sumant Ranji, MD

Brigham and Women's Hospital
Jeffrey Schnipper, MD, MPH
Anuj Dalal, MD

University of Colorado
Katie Raffel, MD

Vizient Inc.
Julie Cerese, RN, MBA, PhD
Ellen Flynn, RN, MBA, JD
Alyssa Harris, MPH
Sam Hohmann, PhD

Advisory Board:
Tejal Gandhi, MD, MPH
Gordon Schiff, MD
Urmimala Sarkar, MD
Kaveh Shojania, MD
David Bates, MD, MPH
Mark van der Laan, PhD

UPSIDE Research Group:

Beth Israel Deaconess Medical Center – David Feinbloom, MD, Bethany N. Roy, MD, Shoshana J. Herzig, MD, MPH
Brigham and Women's Hospital – Mohammed Wazir, MD, Esteban F. Gershanik, MD, MPH, MSc, Abhishek Goyal, MD, MPH, Pooja R. Chitneni, MD
Brigham and Women's Faulkner Hospital – Sharran Burney, NP, Janice Galinsky, NP, Sarah Rastegar, PA, Danielle Moore, PA
Cedars-Sinai Medical Center – Carl Berdahl ,MD, Edward G. Seferian, MD
ChristianaCare – Krithika Suri, MBBS, Tea Ramishvili, MD, Deepak Vedamurthy, MD, MSHS
Emory University Medical Center – Daniel P. Hunt, MD, Amisha S. Mehta, MD, Haritha Katakam, MD
Harborview Medical Center – Stephanie A. Field, MD, Barbara Karatasakis, MD, Katharina Beeler, MD, Allison M. Himmel, MD
Johns Hopkins Bayview Hopkins Medical Center – Shaker Eid, MBA, MD, Sonal Gandhi, MBBS, Ivonne M. Pena, MBBS
Massachusetts General Hospital – Zachary S. Ranta, MD, Samuel D. Lipten, MD, David J. Lucier, MD, MBA, MPH, Beth Walker-Corkery, MPH
Mayo Clinic Rochester – Jennifer Kleinman Sween, MD, Robert W. Kirchoff, MD, MS, Katie M. Rieck, MD, MHA, Gururaj J. Kolar, MD, Riddhi S. Parikh, MBBS, Caroline Burton, MD, Chandrasagar Dugani, MD, PhD
The Miriam Hospital – Kwame Dapaah-Afriyie, MD, MBA, Arkadiy Finn, MD
Medical College of Wisconsin – Sushma B. Raju, MD, Asif Surani, MD, Ankur Segon, MD, MPH, Sanjay Bhandari, MD
Northwestern Memorial Hospital and Northwestern Medicine Central DuPage Hospital – Gopi J. Astik, MD, Kevin J. O'Leary, MD
Oregon Health and Science University – A. Shams Helminski, MD, James Anstey, MD, Mengyu Zhou, MD, Angela E. Alday, MD, Stephanie A.C. Halvorson, MD
University of California, San Francisco – Armond M. Esmaili, MD, Peter Barish, MD, Cynthia Fenton, MD, Molly Kantor, MD
University of Chicago – Kwang Jin Choi, MD, Andrew W. Schram, MD, Gregory Ruhnke, MD
University of Colorado – Hemali Patel, MD, Anunta Virapongse, MD, MPH/MSPH, Marisha Burden, MD, Li-Kheng Ngov, MD, Angela Keniston, MSPH
University of Kentucky – Preetham Talari, MD, MBA, John B. Romond, MD, Sarah E. Vick, MD, Mark V. Williams, MD, MHM
University of Michigan – Ruby A. Marr, MD, Ashwin B. Gupta, MD, Jeffrey M. Rohde, MD
University of Missouri-Columbia – Catherine Jones, MD, Christine Schneider, DO, S. Hasan Naqvi, MD
University of Pennsylvania and Pennsylvania Presbyterian Hospitals – Frances Mao, MD, Michele M. Fang, MD, S. Ryan Greysen, MD, MHS, Pranav Shah, MD
University of Washington Montlake Campus – Christopher S. Kim, MD, MBA, Maya Narayanan, MD, MPH
University of Washington Northwest Campus – Benjamin Wolpaw, MD, Sonja M. Ellingson, MD

University of Wisconsin-Madison – Farah A. Kaiksow, MD, MPP, Jordan S. Kenik, MD, MPH, David Sterken, MD
Vanderbilt University Medical Center – Michelle E. Lewis, MD, Bhavish R. Manwani, MD, Russell W. Ledford, MD, Chase J. Webber, DO, Eduard E. Vasilevskis, MD, MPH, Ryan J. Buckley, MD, Sunil B. Kripalani, MD, MSc

Yale New Haven Hospital – Christopher Sankey, MD, Sharon R. Ostfeld-Johns, MD, Katherine Gielissen, MD, MHS, Thilan Wijesekera, MD, MHS, Eric Jordan, MD
Zuckerberg San Francisco General Hospital – Abhishek Karwa, DO, Bethlehem Churnet, MD, David Chia, MD, MS, Katherine Brooks, MD
Organization: University of California San Francisco

Inclusive Dates of the Project: 9/30/2019-10/1/2023

Federal Project Officer: Stephen Raab, MD, MPH

Acknowledgment of Agency Support, and Grant Number: AHRQ R01HS027369

2. Structured Abstract

Purpose: To determine how often delayed or missed diagnosis occur among patients who die or are transferred to the ICU in the hospital, what causes these diagnostic errors, and the harms caused by diagnostic errors.

Scope: In total, 2,428 patients who died or were transferred to the ICU at 29 academic medical centers in the United States.

Methods: We conducted a retrospective study of randomly selected charts of hospitalized adult patients who died or went to the ICU after the second hospital day. Using a structured tool, two trained physicians reviewed each patient's medical record. We used statistical methods to determine factors associated with errors and which factors were potential targets for future improvement studies.

Results: Overall, 550 patients (23.0%, 95% CI 20.9-25.3%) experienced a diagnostic error. Errors contributed to temporary harm, permanent harm, or death in 436 (17.8% of patients, 95% CI 15.9%-19.8%). Diagnostic process faults in our cohort most highly associated with errors were problems with assessment (adjusted relative risk 2.89, 95% CI 2.23, 3.73), and testing (adjusted relative risk of 2.85, 95% CI 2.16, 3.76). These associations corresponded to adjusted proportion attributable fractions of 21.4% (95% CI 16.4-26.4%) and 19.9% (95% CI 14.7%-25.1%), respectively.

Key Words: Diagnostic error, Patient safety, diagnosis, quality, health services research

3. Purpose (Objectives of the study)

Our goal was to understand how often mistakes in diagnosis occur among patients in hospitals that are part of a national research group (HOMERuN) and share data with a benchmarking organization (Vizient).

This study did three things:

Aim 1: We looked at how many diagnostic errors happen among patients who either die in the hospital or are moved to the ICU 2 days or more after being admitted. We did this by carefully examining cases in medical centers connected to HOMERuN.

Aim 2: We combined the information from review of cases with administrative data from Vizient to find out what factors contribute to the risk of diagnostic errors in our group of patients. We used this data to calculate the adjusted incidence and impact of these contributing factors.

Aim 3: We used advanced computer methods to create models that can identify patients who likely experienced a diagnostic error based on our data.

By bringing together data from Vizient and HOMERuN and using advanced analysis techniques, we gained a comprehensive understanding of how often diagnostic errors happen in hospitalized patients who have suffered harm. We also created models to understand the factors that make a diagnostic error more or less likely. Our goal is to develop efficient tools that can help healthcare institutions improve and monitor diagnosis in the hospital setting.

4. Scope

Background

Currently, most research has looked into problems in the healthcare system that lead to diagnostic errors in clinics and emergency rooms. Less information exists about how often diagnostic errors happen in hospitals, what factors increase the risk of these errors, and how they affect patients (like leading to death or a transfer to the intensive care unit). This lack of data is a substantial problem, because most of the money spent on healthcare happens in hospitals, and hospitals, especially academic ones, are crucial for training future doctors.

Additionally, though there have been efforts to improve safety in healthcare, not enough attention has been given to the performance gaps that cause diagnostic errors and how to address them. Last, hospitals have a lot of data, and there's an opportunity to create electronic systems that can detect diagnostic errors. Closing this knowledge gap is important for making healthcare better and safer.

Context

Three decades of research have focused on the role of systems and policies on patient safety and have led to important insights as to how to reduce adverse outcomes such as infections due to urinary catheters, ventilator-associated pneumonia, or adverse drug events. Although the role of clinicians in these systems has been recognized as part of improving patient safety, less attention has been paid to how clinicians gather, interpret, and convey information about diagnoses and patients' clinical statuses. Our study tried to gain a broad perspective on this question by undertaking a national study using methods design to examine diagnostic processes and errors directly.

Settings

This study was undertaken as a collaboration among 29 academic centers participating in the Hospital Medicine ReEngineering Network (HOMERuN), a national collaborative of academic medical centers including university-based centers, community-based teaching hospitals, and safety-net hospitals.

Participants

We initially identified patients by reviewing administrative data obtained from participating sites, specifically using the Vizient® Clinical Data Base with permission from Vizient, Inc. This process resulted in an initial cohort of 487,532 patients admitted to the participating sites between January 1, 2019, and December 31, 2019, all of whom had a medical diagnosis according to the criteria set by the Centers for Medicare & Medicaid Services (CMS) in Baltimore, MD. Among these patients, 24,591 (5.0% either died during their hospitalization or were transferred to the Intensive Care Unit (ICU).

Given the variation in the size of the participating sites, we employed a random selection process within each site to ensure a balanced representation of cases for review. Reviewers then assessed cases in a randomized order, excluding patients whose cases were identified in error (e.g., lacking a medical diagnosis), those transferred to the ICU for policy reasons (e.g., medication desensitization), cases involving admissions solely for comfort or hospice care, those resulting from out-of-hospital cardiac arrests, or cases for which the medical records were unavailable. This screening led to 2,997 eligible cases, which underwent review until 100 charts were adjudicated at each site or until the data collection period concluded.

Following the completion of exclusions and reviews, our final cohort was composed of 2,428 patients.

Incidence

As a retrospective study, we were not able to determine incidence of diagnostic errors.

Prevalence:

Our study determined a prevalence of diagnostic errors of 23.0% (95% CI 20.9-25.3%).

5. Methods

Study Design:

We conducted a retrospective multicenter cohort study of adult patients who died or were transferred to the intensive care unit (ICU after the second hospital day. We excluded patients who were transferred to the ICU earlier in their course to eliminate cases due to mis-triage from the emergency department rather than inpatient diagnostic errors.

Data sources and collection:

We used data collected directly from the medical record via chart review as well as administrative data gathered from our hospitals.

Adjudication Methodology:

In the course of this study, a meticulous adjudication methodology was employed. All cases underwent thorough review by two physicians who were specifically trained in error adjudications. Stringent oversight and quality-checking processes were implemented to ensure the credibility of the assessments. This approach, involving two physician reviews, is a widely accepted practice in patient safety research, as observed in our previous investigations on readmissions and diagnostic errors.

For the adjudication to be finalized, both physician reviewers had to reach a unanimous agreement on the assessment. In instances when agreement proved elusive, a third trained reviewer was brought in to resolve any discrepancies. It's important to note that, in cases when two or three physicians conducted reviews, the focus was on achieving complete agreement rather than measuring inter-rater reliability. Previous research from our team has shown that adjudications performed by two trained physicians, independent of the case, followed by expert over-reads, consistently yield results with a Cohen's kappa value greater than 0.7 for identifying diagnostic errors.

Selection and training of adjudicators were also meticulous. Reviewers were active clinicians caring for general medical inpatients and received training through a comprehensive, 2-day, live video conference. Subsequently, they reviewed standardized cases with expert reviewers until 100% agreement was observed on a minimum of 10 standard cases. Only then did the teams proceed to adjudicate the study cases.

To ensure the validity and consistency of reviews across sites, multiple quality assurance steps were taken. Each site presented at least one case quarterly to study team members, who provided feedback and corrections as necessary. Issues raised during these sessions were used to create a Frequently Asked Questions document, offering specific guidance on various aspects. Additionally, every tenth case from each site underwent independent expert over-read by the research team.

As a final validity check, the research team directly re-examined a minimum of 10 redacted patient charts and original case review forms from sites whose error rates deviated significantly from the group mean error rate. This process confirmed high concordance, with the exception of one site, for which data were retained only from cases over-read and confirmed by two additional team members.

The determination of errors and underlying causes involved a comprehensive examination of the entire electronic medical record for each hospitalization. Adjudicators focused on the reason for admission, events leading up to ICU transfer or death, and correlation of diagnostic decision-making documentation with objective data.

Interventions:

None

Measures:

Outcome Measures: Our main focus was on whether a diagnostic error was present, as determined by the SAFER-DX tool. To understand the root causes of these errors, we utilized the DEER Taxonomy Tool, which helped identify specific failure points during the episode of care. The DEER tool was applied consistently, regardless of the reviewer's belief in the presence of a diagnostic error. This allowed us to identify common failure points and those more frequently associated with diagnostic errors. The DEER Taxonomy covers eight categories: Access/Presentation, Patient History, Physical Exam, Diagnostic Test Ordering, Performance and Interpretation, Patient Follow-Up, Subspecialty Consultation/Referral, Healthcare Team Communication and Collaboration, and Patient Experience.

For each case with an error, we assessed the extent of the error's contribution to patient harm using the NCC-MERP scale, which provides explicit definitions of harm (e.g., considering an error to have led to death if it "contributed to or resulted in the patient’s death").

Confounding and adjustment measures: Administrative data from Vizient was used to categorize comorbidities based on the Elixhauser method and to define inpatient diagnoses associated with diagnostic errors using ICD-10 codes. All analyses, including the estimation of univariate proportions and their confidence intervals, involved weighted estimation. Each observation was weighted by the inverse of the sampling probability, defined as the ratio of cases reviewed in each hospital to the total number of eligible ICU transfers and deaths for review at each hospital during the study period.

Statistical methods: We employed multivariable Cox proportional hazard models with clustering effects, setting the time variable to unity and handling ties with the Breslow method. Robust variance estimators were used to construct confidence intervals of parameter estimates. Given the common occurrence of error outcomes in our data, we opted for a modified form of Cox regression to directly estimate the prevalence ratio instead of the odds ratio. Covariates for multivariable models were selected based on substantive knowledge and a priori hypotheses regarding the relationship between each variable and diagnostic error.

We calculated adjusted preventable attributable fractions, considering the sampling design, to offer insights into which features contributed the most to diagnostic errors in absolute terms.

Limitations

First and foremost, our data might be influenced by documentation and detection biases. To mitigate documentation biases, we encouraged chart reviewers to utilize all available medical record documentation and exercise sound judgment in interpreting patterns indicative of the diagnostic process. Additionally, we addressed detection biases by providing extensive training to all reviewers at the outset, employing methods known for producing high inter-rater reliability.

To enhance validity across different sites, cases underwent over-reads and reviews by members of the core research team, with extensive cross-checking of data. However, it's important to note that our data cannot directly measure whether a diagnostic error corresponds to a specific cognitive process, such as anchoring on a diagnosis to the exclusion of others. Communication gaps and issues with team dynamics may also be underdetected in the medical record, leading to a low prevalence of these issues in our study. Similarly, our data doesn't allow us to assess whether patients experienced various types of harm, such as emotional or financial, related to diagnostic errors.

Local reviewers' adjudications might have been influenced by local norms and professional standards, potentially shaping assessments of error likelihood and associated harms. We addressed these concerns through training and inter-site over-read of cases. The relationship between the most common fault associated with diagnostic errors, clinical assessment, and other faults cannot be disentangled. For instance, testing process faults might lead to problems in clinical assessment, and vice versa.

Our reviewers had access to the entire medical record, reviewing information after the fact, which might not fully capture the rapid evolution of clinical scenarios or the absence of a gold standard diagnostic test for many inpatient conditions when reviewed in real time. External pressures on teams or clinicians, such as hospital census or physician workload, were not directly measured. It's important to note that our results do not represent the prevalence and severity of diagnostic errors across all hospitalized patients; rather, this study focused on a select sample of patients experiencing clinical deterioration. Moreover, the generalizability of our results to all U.S. hospitals is limited, given the predominantly academic medical center selection for this study.

Last, though our data are detailed and complex, they fall short of being sufficient for creating well-calibrated machine learning models to identify cases with diagnostic errors. We anticipate that more detailed electronic health record data, including audit logs and finer measures of physiology, will be necessary to achieve this objective.

6. Results

Principal Findings:

- Diagnostic Error Rate: Of 2,428 patient records at 29 hospitals that underwent review, 550 (23.0%, 95% CI 20.9-25.3%) experienced a diagnostic error.
- Harms due to diagnostic errors: Errors were judged to have contributed to temporary harm, permanent harm, or death in 436 (17.8% of patients, 95% CI 15.9%-19.8%); among 550 patients with DE, the error was judged to have contributed to temporary harm, permanent harm, or death in 77.1% (95% CI 72.3%-81.9%). Among all 1,863 patients who died, the DE was judged to have contributed to death in 121 (6.6%, 95% CI 5.3%-8.2%); within the group of patients who died and had a DE, the error contributed to the death in 29.4% (95% CI 24.0%-35.3%).
- Features associated with diagnostic errors: The most prevalent diagnostic process faults in our cohort were problems with assessment (e.g., delay in considering diagnosis, failure to recognize complications, or suboptimal prioritizing of potential diagnoses), access and presentation faults (e.g., incorrect triage, failure or delay in seeking care), and problems with testing (e.g., delay in ordering or performing needed tests, erroneous clinician interpretation of test). In multivariable models adjusting for patient sociodemographic factors, comorbidities, and all process faults, the two diagnostic processes most highly associated with DE were problems with assessment (adjusted relative risk 2.89, 95% CI 2.23, 3.73) and with testing (adjusted relative risk of 2.85, 95% CI 2.16, 3.76), corresponding to adjusted proportion attributable fractions of 21.4% (95% CI 16.4-26.4%) and 19.9% (95% CI 14.7%-25.1%), respectively.
- Machine learning models: We constructed and tested a series of algorithms appropriate for tabular data including Support Vector Machine (With Kernel), Random Forests, Gradient Boosting algorithms, Classification and Regression Trees (CART), logistic regression with elastic net (GLMNET), Ridge and Lasso regularization, and EAML within our adjudicated cases and including data observable in the medical record (such as functional status). We were not able to generate models with c-statistic greater than 0.57, so we did not proceed with subsequent testing of predictive models or implementation of EAML methods.

Discussion

In this retrospective, multicenter study involving medical patients who either died in the hospital or were transferred to an Intensive Care Unit (ICU), diagnostic errors were prevalent and frequently linked to patient harm. Though various patient and system factors played a role in the likelihood of diagnostic errors, issues related to the diagnostic processes were identified as central to the propagation of errors. Notably, problems related to testing (such as selecting the appropriate test, timely ordering, and accurate interpretation of results) and assessment (such as avoiding biases in decision making like anchoring or confirmation biases) emerged as crucial areas for improvement in safety programs.

Our study builds upon foundational research in patient safety, which initially highlighted the overall impact of errors. Although earlier works like the Harvard Medical Practice Study rightly focused on procedural complications and medication errors, they also underscored the independent significance of diagnostic problems in safety gaps. Over the years, strides have been made to enhance medication and procedural safety, but progress in diagnostic safety, particularly in the hospital setting, has been relatively limited. Our findings contribute to previous studies in several key ways. First, we employed a standardized approach to identify our patient diagnostic pool, concentrating on patients who experienced severe harm in the hospital. Second, we applied a standardized adjudication process to all cases, enabling us to comprehend the pathophysiology of inpatient diagnostic errors reliably and validly. Third, we examined medical records from diverse health systems, providers, and patient populations, offering one of the most extensive and broadly applicable insights into diagnostic processes in the hospital to date.

Our study relied on retrospective medical record reviews, a method employed in other foundational patient safety studies, and it has both strengths and limitations when applied to diagnostic processes. Similar to errors resulting from procedural complications or certain medication errors, chart reviews allow for sophisticated assessments requiring expert clinical guidance. For instance, determining the potential for patient harm based on the timing of test results and the initiation of definitive treatment involves clinical expertise. Although rule-based criteria might be more objective, they lack the flexibility to capture the spectrum of diagnostic errors in general medical patients, given the diverse medical problems they present and the current state of medical science. Conversely, because many aspects of the diagnostic process occur in a provider's mind or during team discussions, not all are reflected in electronic records. Future studies may need to incorporate surveys and interviews of clinicians regarding recent cases to capture the full depth of reasons for diagnostic errors. However, as a foundational exploration into this topic, our multicenter retrospective medical review strikes a fair balance between breadth and depth.

Conclusions

Diagnostic errors in this cohort of patients, all of whom were hospitalized and very ill, were unsettlingly common and harmful. Though prevalence and harms in our study are dismaying, our methods are important because they provide tools with which hospitals can begin to examine their own error rates and begin to correct them. Importantly, our adjudication methods can also be used to aid development of new artificial intelligence systems that detect errors or identify when additional diagnostic thinking might be needed (before an error takes place). Finally, the priority areas we identified — testing and assessment — provide an initial roadmap for intervention development and focus areas for research.

Significance

The prevalence of diagnostic errors in this seriously ill patient cohort is surprisingly high, with errors being associated with substantial harm. Some of our findings are likely influenced by how sick our patient cohort was in that severity of illness significantly compresses the time to make a diagnosis, but the challenge to healthcare to make improvements is substantial.

Implications

To prevent future diagnostic errors (DE), upcoming programs must address deficiencies in assessment, testing, and follow-up. Possible interventions include identifying patients at high risk for diagnostic errors, implementing triggers for a "diagnostic time-out" to re-evaluate the diagnostic process, establishing a diagnostic peer-consult service, offering decision support on pre-test and post-test probabilities, providing guidance on test selection for common diagnoses, and implementing alarms for patients exhibiting deterioration or unexpected lack of improvement.

Similar to programs designed to prevent other adverse events, these interventions aim to enhance diagnostic accuracy. However, the unique nature of DE, which remains predominantly clinician-centric even within complex health systems, necessitates tailored strategies. Additionally, individual clinician audit and feedback could be valuable, but it requires careful implementation to avoid the "second victim effect" — the potential re-traumatization of clinicians after a patient experiences an adverse event. This underscores the importance of approaching DE prevention with a nuanced understanding of its impact on healthcare providers within intricate healthcare systems.

Finally, the role of artificial intelligence in identifying diagnostic opportunities, screening for diagnostic errors, guiding speedier diagnoses, and training physicians and patients in how to make better and faster diagnoses is only now coming into focus. These tools and methods will provide transformational opportunities for patient safety, particularly if integrated into care with consideration for the complex and intricate sociotechnical challenges posed by all technology adoption in healthcare.

7. List of Publications and Products (Bibliography of Outputs) from the study.

Auerbach AD, Lee TM, Hubbard CC, Ranji SR, Raffel K, Valdes G, Boscardin J, Dalal AK, Harris A, Flynn E, Schnipper JL; for the UPSIDE Research Group. Diagnostic errors in hospitalized adults who died or were transferred to intensive care. JAMA Intern Med. 2024 Jan 8. doi:10.1001/jamainternmed.2023.7347

Dalal AK, Schnipper JL, Raffel K, Ranji S, Lee T, Auerbach A. Identifying and classifying diagnostic errors in acute care across hospitals: Early lessons from the Utility of Predictive Systems in Diagnostic Errors (UPSIDE) study. J Hosp Med. 2023 May 21. doi: 10.1002/jhm.13136