Integrating Clinical and Economic Evaluation in Implementing AI Systems in Health Care

Bijan J. Borah, PhD

- Professor of Health Services Research, Mayo Clinic College of Medicine & Science
- Scientific Director, Clinical and Economic Evaluation Program, Kern Center for the Science of Health Care Delivery

International Health Economics Association 2025 Congress, Bali, Indonesia, 20 July 2025

Plan for the Talk

- Discuss primarily about clinical and economic evaluation in the context of learning health system
- Provide the general framework for AI implementation and evaluation
  - I will do so in the context of 2 ongoing cases studies

Acknowledgements & COIs

- All the Mayo and other colleagues with whom I am collaborating with the case studies
- No COI to report with respect to this work

Clinical Evaluation Framework for AI in Healthcare

- Technical Validation
  - Accuracy, sensitivity, specificity, AUC – performance metrics
  - Robustness and generalizability – diverse populations, geographies, time
  - Fairness and bias audits – demographic, other subgroups (e.g., race, insurance)
- Clinical Validation
  - Retrospective validation – typically happens in the model building
  - Prospective validation – assessment of outcomes of interests
    - RCTs
- Implementation Evaluation
  - Usability testing – stakeholder (clinicians, patients) input/experience
  - Adoption and fidelity – potential barriers to adoption
  - Human factor analysis – clinician-AI interaction, alert fatigue

Economic Evaluation Framework

- Cost-effectiveness Analysis
  - Incremental cost-effectiveness ratio or ICER (AI vs. standard of care)
  - Modeling approaches – decision trees, Markov models, microsimulation to project long-term impacts
  - Perspective – societal, health system, payer
- Budget Impact Analysis
  - Evaluate affordability and financial feasibility of the AI system
- Return on Investment (ROI)
  - Relevant for health care providers that adopt AI systems (potentially costly)
  - Evaluate efficiency gains (e.g., optimal outcomes) vs. implementation and maintenance costs

Integrated Frameworks and Standards (Dynamic)

- SPIRIT-AI / CONSORT-AI Guidelines[1]
  - Extensions of current reporting guidelines for clinical trials to ensure transparency in AI evaluations
- DECIDE-AI Framework[2]
  - Evaluate affordability and financial feasibility of the AI system
- Health Technology Assessment (HTA) Adaptations
  - Reimbursement agencies are adapting HTA methodologies to accommodate unique characteristics of AI (e.g., continuous learning models, black-box algorithms such as deep learning models)

[1]Liu, X., Cruz Rivera, S., Moher, D. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* **26**, 1364–1374 (2020). https://doi.org/10.1038/s41591-020-1034-x
[2]Vasey, B., Nagendran, M., Campbell, B. *et al.* Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med* **28**, 924–933 (2022). https://doi.org/10.1038/s41591-022-01772-9
SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence)
CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence)
DECIDE-AI (Developmental and Exploratory Clinical Investigation of Decision-support systems driven by Artificial Intelligence)

Integrated Frameworks and Standards (Continued)
FDA Guidance*

Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry and Other Interested Parties

DRAFT GUIDANCE

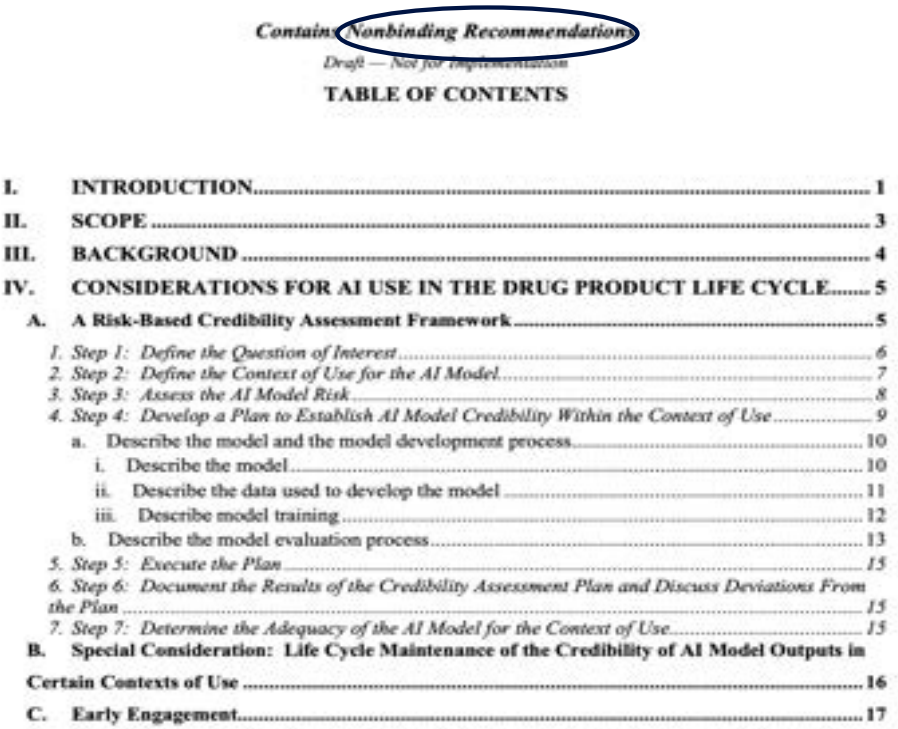
This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document, contact (CDER) Tala Fakhouri, 301-837-7407; (CBER) Office of Communication, Outreach and Development, 800-835-4709 or 240-402-8010; or (CDRH) Digital Health Center of Excellence, digitalhealth@fda.hhs.gov.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)
Center for Veterinary Medicine (CVM)
Oncology Center of Excellence (OCE)
Office of Combination Products (OCP)
Office of Inspections and Investigations (OII)

January 2025
Artificial Intelligence



Contains Nonbinding Recommendations
Draft — Not for Implementation

TABLE OF CONTENTS

I. INTRODUCTION 1
II. SCOPE 3
III. BACKGROUND 4
IV. CONSIDERATIONS FOR AI USE IN THE DRUG PRODUCT LIFE CYCLE 5
A. A Risk-Based Credibility Assessment Framework 5
1. Step 1: Define the Question of Interest 6
2. Step 2: Define the Context of Use for the AI Model 7
3. Step 3: Assess the AI Model Risk 8
4. Step 4: Develop a Plan to Establish AI Model Credibility Within the Context of Use 9
a. Describe the model and the model development process 10
i. Describe the model 10
ii. Describe the data used to develop the model 11
iii. Describe model training 12
b. Describe the model evaluation process 13
5. Step 5: Execute the Plan 15
6. Step 6: Document the Results of the Credibility Assessment Plan and Discuss Deviations From the Plan 15
7. Step 7: Determine the Adequacy of the AI Model for the Context of Use 15
B. Special Consideration: Life Cycle Maintenance of the Credibility of AI Model Outputs in Certain Contexts of Use 16
C. Early Engagement 17

*https://www.fda.gov/media/184830/download

Integrated Frameworks and Standards (Continued)
Other Regulatory Authorities

- National Institute for Health and Care Excellence or NICE
  - Evidence Standards Framework (ESF) for digital health technologies[1]
  - Stratifies evidence requirements based on the risk and function of digital tools, including AI
- European Medicines Agency's AI Initiatives[2]
  - Recently published 2024 AI Observatory Report

[1]https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies
[2]https://www.ema.europa.eu/en/about-us/how-we-work/data-regulation-big-data-other-sources/artificial-intelligence

Special Considerations in AI Evaluation

- Dynamic updating
  - Reassessment needed if models ingested with new data and/or evolve over time
  - Very fast-paced developments
    - AI to LLMs to Agentic AI
    - Multi-model AI systems that use different forms of data
      - structured data (e.g., patient demographics from EHR,
      - unstructured data (clinical notes),
      - images (e.g., radiology),
      - sensor data (e.g., from wearables)
- Transparency and explainability
  - Essential for clinician and patient trust and regulatory approval
- Regulatory landscape
  - Evolving guidance from FDA, EMA, European Union Medical Device Regulation (EU MDR) and other global regulatory bodies

I will talk about the implementation and evaluation of AI in the care delivery in a large health system in the U.S.

Research Setting

- Mayo Clinic is an academic hospital providing healthcare at different geographical locations including
  - Arizona
  - Florida
  - Minnesota
  - Wisconsin
- ~1.5 million patients from 150 countries each year
- It has an integrated EHR across all the locations (Epic, Verona, USA)

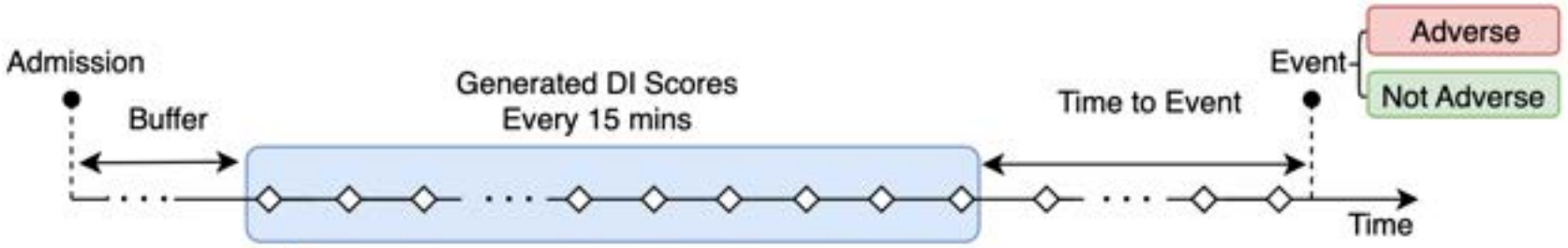
Mayo Clinic Care Network Partners (45 in total): Facilitating Clinical, Education and Research Collaborations

Case Study 1:
Deterioration Prediction for Hospitalized Patients

- Early warning systems (EWS) in hospitals:
  - Enhancing patient care
  - Reducing adverse events
  - Evolving from manual assessments to technologically advanced systems

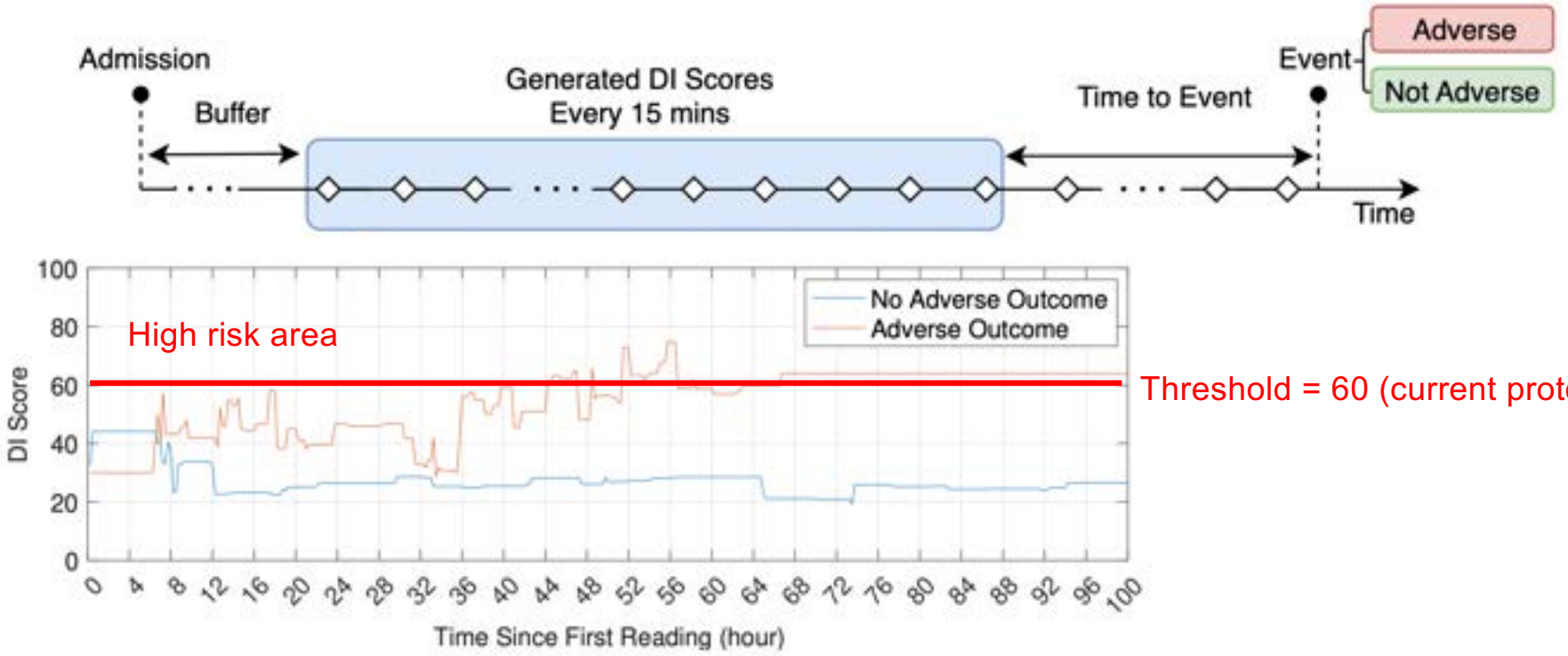
- Objective:
  - Accurate proactive identification of patient deterioration to prevent morbidity and mortality in the hospital setting

Epic Deterioration Index

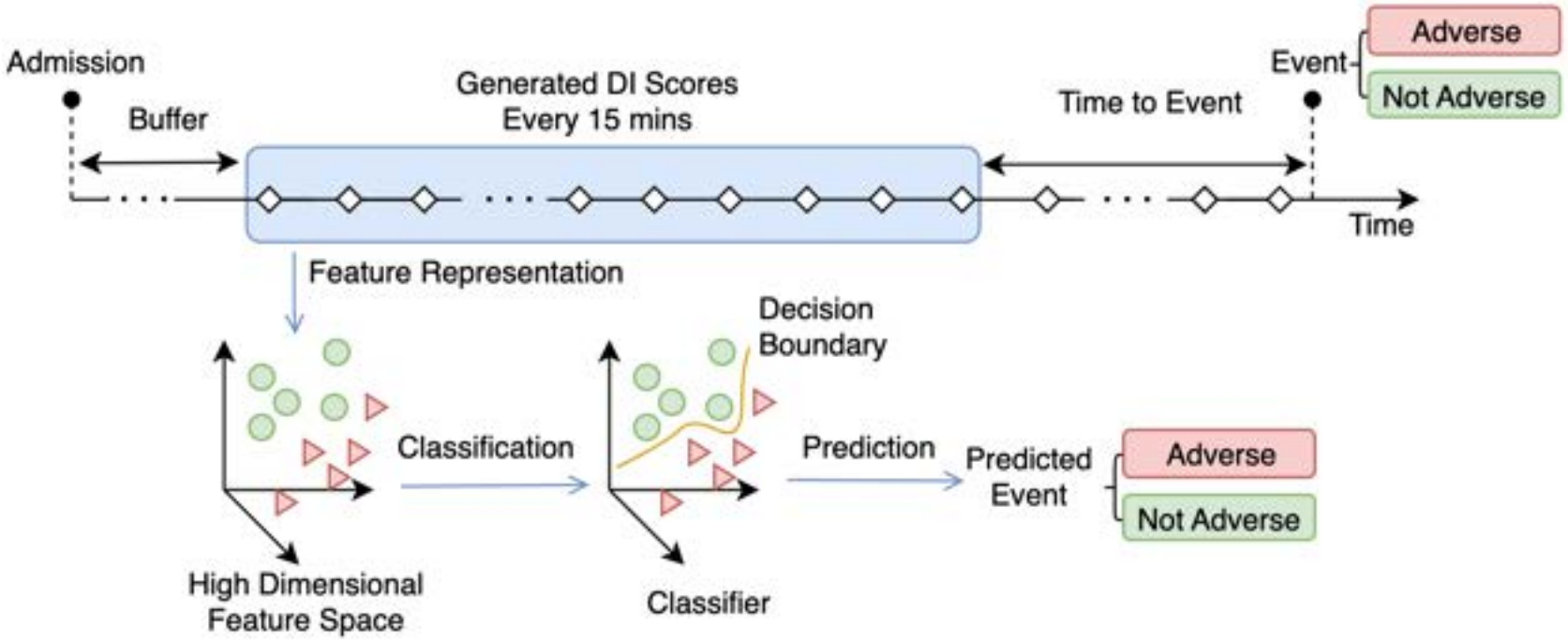


- Epic's Deterioration Index (EDI) is a proprietary algorithm and one of the most widely used EWS deployed in hundreds of hospitals across the U.S., including Mayo Clinic
- The risk score is determined based on 17 parameters
- Goal: Adverse Events Prediction: all-cause mortality, cardiac arrest, transfer to intensive care, and evaluation by the rapid response team

Current Protocol



Machine Learning Model



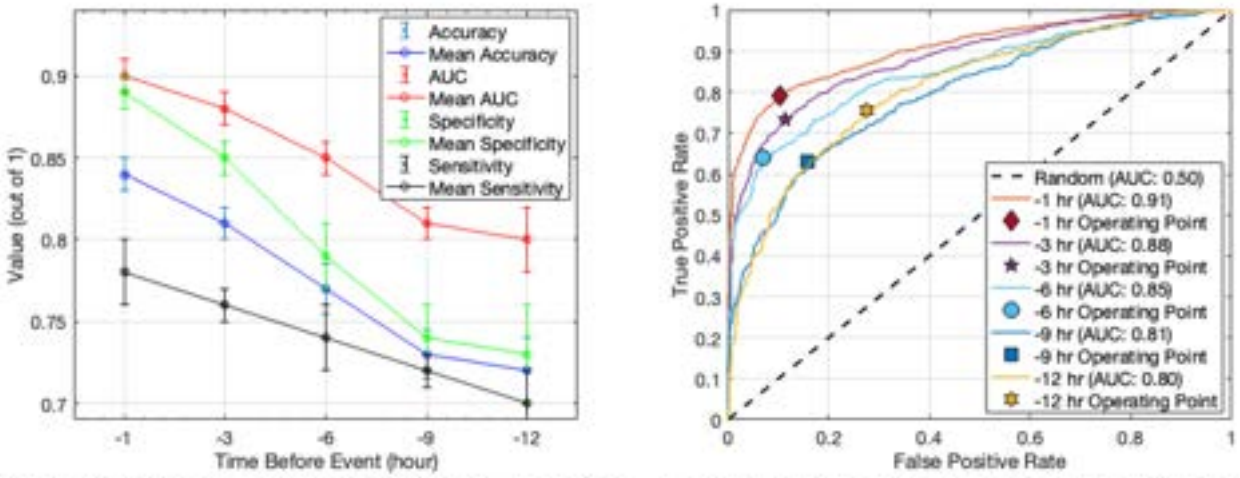
Data

- The DI scores were collected for adult patients (≥18 y-o) hospitalized on medical or surgical services during 8-23-2021 to 3-31-2022
- The DI score was calculated every 15 mins for a total of 59,617 encounters in four different locations across the U.S.:
  - Rochester, Minnesota (RST) 25,127 with 2,802 AEs
  - Mayo Clinic Health System (Minnesota and Wisconsin) (MCHS) 16,330 with 779 AEs
  - Jacksonville, Florida (FLA) 9,695 with 825 AEs
  - Scottsdale, Arizona (ARZ) 8,465 with 1,567 AEs

Results

Performance (Mean±SD) of models in predicting AE from DI scores.

Model	Sensitivity	Specificity	Accuracy	F1-Score	AUC
Thresholding (any)	0·25±0·01	0·89±0·01	0·57±0·01	0·53±0·01	0·57±0·01
Thresholding (last)	0·13±0·01	0·99±0·00	0·56±0·01	0·46±0·01	0·56±0·01
ML	**0·85±0·01**	**0·91±0·01**	**0·88±0·01**	**0·88±0·01**	**0·94±0·01**



(A) Average accuracy, AUC, specificity, sensitivity, and corresponding standard deviation of the best model in prediction of the correct event using retrospective DI scores for various hours before the event. Results are reported after averaging over 10-fold cross-validation. (B) ROC curve of the best model in prediction of the correct event using retrospective DI scores for various hours before the event. Results are reported after averaging over 10-fold cross-validation.

eClinicalMedicine
Volume 66, December 2023, 102312

Articles

Novel machine learning model to improve performance of an early warning system in hospitalized patients: a retrospective multisite cross-validation study

Hojjat Salehinejad [a b], Anne M. Meehan [c], Parvez A. Rahman [a], Marcia A. Core [d], Bijan J. Borah [a f], Pedro J. Caraballo [c e f]

[a] Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, USA
[b] Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA
[c] Department of Medicine, Mayo Clinic, Rochester, MN, USA
[d] Department of Information Technology, Mayo Clinic, Rochester, MN, USA
[e] Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA

Received 26 July 2023, Revised 26 October 2023, Accepted 26 October 2023, Available online 16 November 2023, Version of Record 15 December 2023.

Silent Validation

- Implementation of the model in silent mode in Epic (EHR)
  - Worked with Mayo IT team to deploy the model in practice
- Receiving DI scores in real-time and generating prediction (adverse event positive/negative)
- Storing prediction in a database without intervention
- Data will be collected for a period of time
- Model's performance is being evaluated against real-world outcomes

Prospective Testing

- Upon success and approval by the practice leadership, the model predictions will be displayed in the patient chart
- Prospectively, the DI Model will be rolled out at one department or one hospital at a time (e.g., step-wedge, adaptive)
- Clinical and economic outcomes will be evaluated
- Minimum clinical outcomes
  - All-cause mortality, cardiac arrest, transfer to intensive care, and evaluation by the rapid response team, length of stay
- Minimum economic outcomes
  - Hospitalization cost
  - Internal vs. standardized cost

Developing a standardized healthcare cost data warehouse

Sue L Visscher [1], James M Naessens [1 2], Barbara P Yawn [3], Megan S Reinalda [4], Stephanie S Anderson [4], Bijan J Borah [5 6]

Affiliations + expand

PMID: 28606088 PMCID: PMC5469019 DOI: 10.1186/s12913-017-2327-8

- Time-Driven Activity-Based Costing (TDABC)*
  - Estimate the cost per unit of time for different personnel
  - Estimate the time required of each type of personnel
  - Total cost = time spent * cost per unit of time

*Kaplan, R. S., & Anderson, S. R. (2004). *Time-driven activity-based costing.* Harvard Business Review, 82(11), 131–138.
*Kaplan, R. S., & Porter, M. E. (2011). *How to solve the cost crisis in health care.* Harvard Business Review, 89(9), 46–52.
*Kaplan, R. S., Witkowski, M. L., Abbott, M. D., et al. (2014). *Using time-driven activity-based costing to identify value improvement opportunities in healthcare.* Journal of Healthcare Management, 59(6), 399–412.

Case Study 2

Use of Artificial Intelligence in Preoperative Assessment for Surgical Triage

MAYO CLINIC

Introduction

- **Objective**: Use analytical decision support tools to identify patients with the highest likelihood of needing surgery at Mayo Clinic Rochester
- **Current State**: External spine surgery referrals are manually triaged by physicians, NP, PAs, yielding surgery in only 30–35% of new appointments
- **Initial Use**: Tool to be piloted in the Spine Center (Neurologic Surgery and Orthopedic Surgery) with future expansion across the House of Surgery
- **Burden on Clinicians**: Surgeons and NP/PAs spend significant time manually reviewing patient records for triaging daily

Introduction (continued)

- **Benefits of Analytics**
  - Reduce triage time
  - Identify new variables influencing surgical need
- **Pilot and Expansion**
  - Algorithm to be tested in Spine Center before broader diffusion
- **Institutional Goals Supported**:
  - Support Workforce: Reduce workload on clinical staff with triage decision support
  - Attract Serious/Complex Cases: Prioritize patients requiring specialized surgical care
  - Optimize Capacity: Improve surgical yield by focusing on high-probability cases

Triage for Referral Patients
Current State vs. Expected Future State

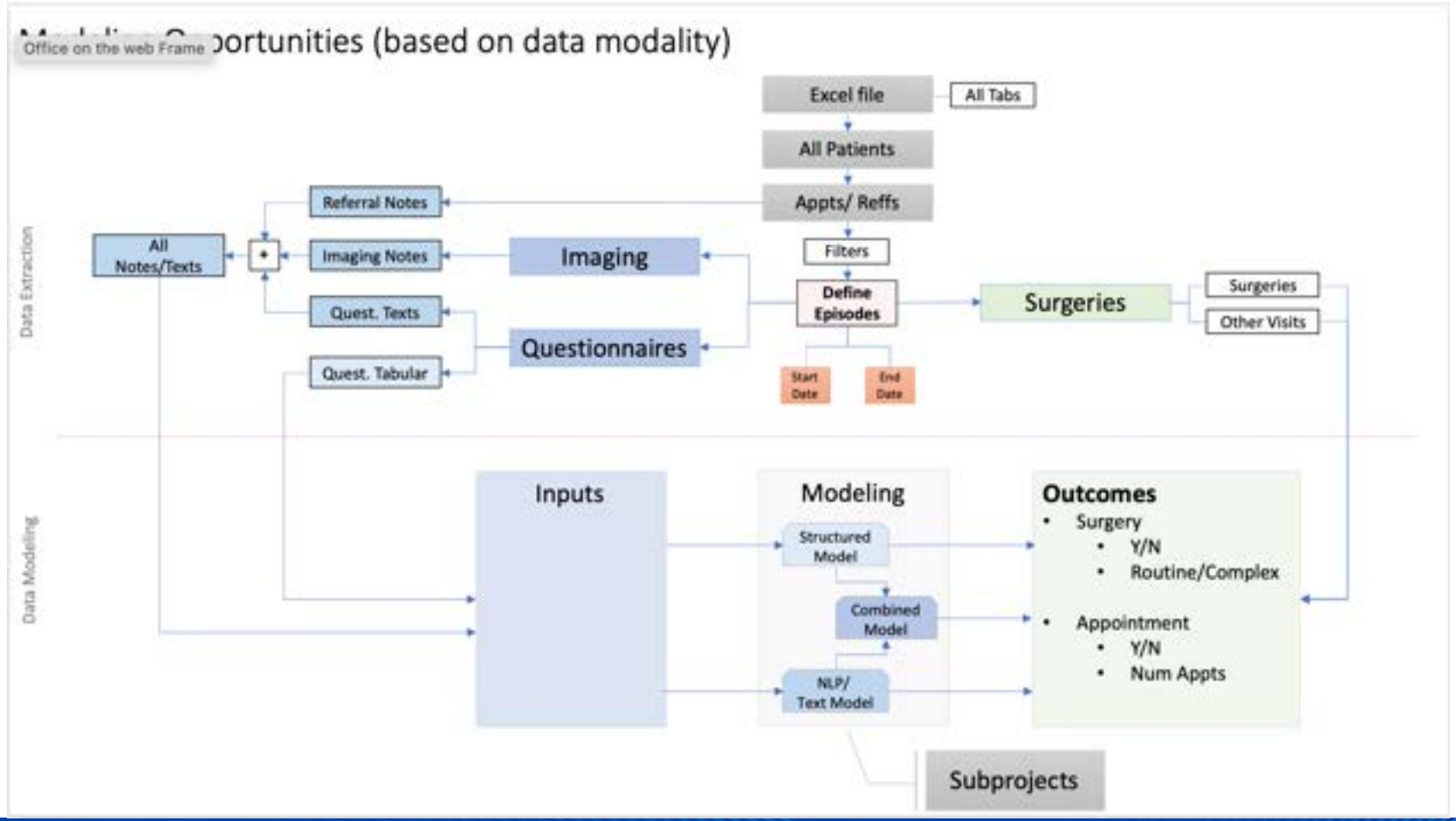
- **Current Practice**: Triage decisions are primarily based on patients' 1-year medical history (e.g., imaging reports, diagnoses) and triage questionnaires
- **Referral Selection**: Only referrals aligning with the expertise of Orthopedics and Neurosurgery are accepted for consultation or surgery
- **Proposed Hypothesis**: AI models can be trained to predict triage and surgical outcomes for each referral
- **Goal**: Use AI to optimize clinic appointment prioritization by identifying patients with complex needs requiring specialized surgical care

Specific Aims

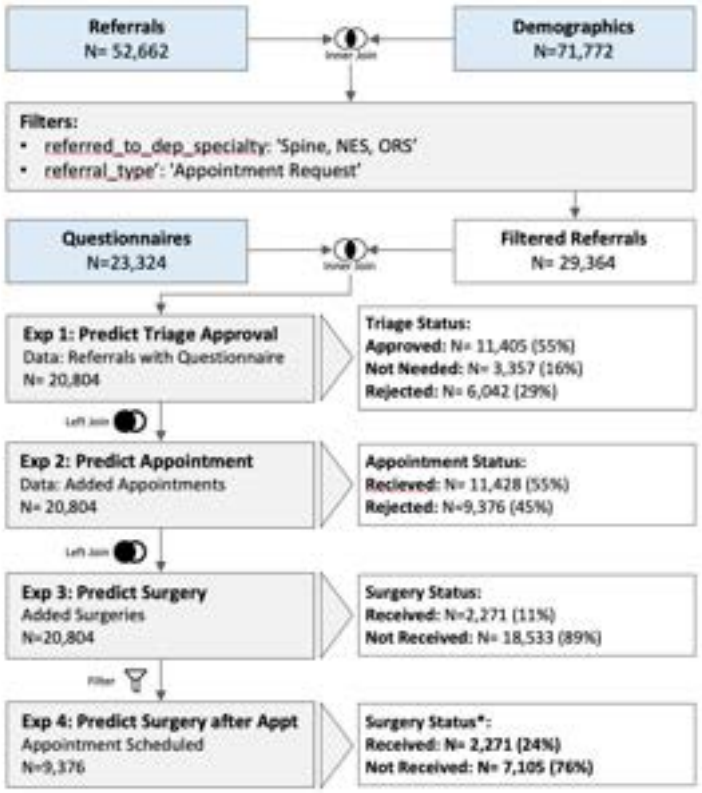
- Complete a retrospective analysis of spine patient charts to abstract medical and radiology records
- Model variables that are highly influential in determining the need for spine surgery at MCR
- Build and apply an algorithm to future patients to calculate a probability score for surgical need
- Prioritize clinical appointments for highest probability cases

Target Patient Data

- **Cohort**
  - 73,688 patients (52% female) referred to Orthopedics/Neurosurgery (2019–2023)
- **Outcomes**
  - Referral status (accepted vs denied) and encounter type (consult vs surgery) recorded
- **Data Sources (1-year prior to encounter)**:
  - Structured: EHR, demographics, ZIP code, insurance
  - Semi-structured: Triage questionnaires
  - Unstructured: Clinical/referral/triage notes, radiology reports, patient messages



Candidacy Using Questionnaires



Accuracy (macro)	Precision	Recall	F1-Score	AUC	
0.66 (95% CI:0.65-0.68)	0.67 (95% CI: 0.65-0.69)	0.75 (95% CI: 0.73-0.76)	0.71 (95% CI: 0.69-0.72)	0.71 (95% CI: 0.70-0.73)	
0.64 (95% CI: 0.62-0.65)	0.59 (95% CI: 0.57-0.61)	0.66 (95% CI: 0.64-0.68)	0.62 (95% CI: 0.60-0.64)	0.68 (95% CI: 0.66-0.69)	
0.77 (95% CI: 0.75-0.79)	0.30 (95% CI: 0.23-0.39)	0.09 (95% CI: 0.06-0.12)	0.14 (95% CI: 0.10-0.18)	0.54 (95% CI: 0.51-0.57)	
0.87	0.92	0.69	0.79	0.89	Needs further validation

Key Findings

- **4 ML Experiments**
  - Predicting triage approval
  - Predicting appointment scheduling
  - Predicting surgery for all triaged patients
  - Predicting surgery among patients with completed appointments
- **Key Findings**
  - Highest accuracy (AUC 0.89) in Experiment 4
  - Top predictors: pain duration, prior treatments, socioeconomic factors
- **Impact**
  - Demonstrates value of routine referral-stage data for early decision support across spine care stages

How these models benefit the practice

- **Triage Optimization**: Supports development of intelligent triage tools to manage rising spine referral volumes
- **Early Identification**: Predictive models can identify likely surgical candidates early to prioritize appointments and reduce wait times
- **Workflow Integration**: Embedding models in the referral process may enhance access equity and reveal biases in current triage practices
- **Resource Efficiency**: Helps allocate clinical resources more effectively and reduces unnecessary chart reviews and low-yield appointments
- **Provider Support**: Lightens the workload for surgeons by filtering out non-surgical cases early
- **Next Steps**: Embarked on the next phase of the project for silent testing in practice

Questions/Discussion

Borah.Bijan@mayo.edu