Crowdsourced, Point-of-Care Decision Support for Surgical Risk Prediction

William Yuan, PhD (1,2) Jayson S. Marwaha, MD (3) Bryce K. Allen, PhD (1) Shana K. Rakowsky, MD (3) Anthony F. Chen, MD (4) Brendin R. Beaulieu-Jones, MD, MBA (1,3) Nathan P. Palmer, PhD (1) Joshua H. Wolf, MD (5) Joseph D. Feuerstein, MD (6) Isaac S. Kohane, MD, PhD(1) Gabriel A. Brat, MD, MPH (1,3)

1) Department of Biomedical Informatics, Harvard Medical School, 2) Computational Health Informatics Program, Boston Children's Hospital, 3) Department of Surgery, Beth Israel Deaconess Medical Center, 4) Department of Radiology, University of California, Davis Medical Center, 5) Department of Surgery, Sinai Hospital/Lifebridge Health, 6) Division of Gastroenterology and Center for Inflammatory Bowel Diseases, Beth Israel Deaconess Medical Center

Abstract

Effectively managing uncertainty is one of the most common and challenging aspects of medicine. Algorithmic guidance is unavailable for many decisions without established or well-defined standards, resulting in unwanted variation in care. Development and validation of a methodology to guide decisions without established standards. This is done by i) learning existing patterns of practice, ii) using short-term outcomes and expert review to assign value, and iii) characterizing tradeoffs associated with particular clinical behaviors. Using a retrospective observational cohort (2008-2020) from a national administrative claims records (87,000 patients) patients with recurrent, medically managed ulcerative colitis were identified through administrative coding. Exclusion criteria included annotations of colon cancer, prior history of colectomy, and absence of prescription management. Deep-learning based risk-profiling using available patient history and inferred physician gestalt prior to a visit was conducted to predict 6-month colectomy status and etiology (emergency/non-emergency). Our model was able to **preemptively identify 78% of emergency surgeries**, with **an average lead time of 381 days**. Patients flagged for surgery by the model were, compared to unflagged patients, **23 times more likely to undergo surgery** (emergency and non-emergency) and **cost an additional $200,000 over six months**. Flagged patients **consumed twice the quantity of corticosteroids** and were **half as likely to experience steroid-free remission**. Surgeries overlooked by the model were found to be incidental to patient disease severity. Feature interpretation of the trained model was found to be consistent with degrees of physician concern. Finally, a set of case-based scenarios presented to a panel of expert clinicians found stronger agreement between panelist judgment and model predictions than with actual provider behavior. For clinical decisions without a gold standard, the collective behavior of a wide population of physicians is likely to represent a valuable source of guidance. Our approach has the potential to inform many clinical decisions currently limited by uncertainty by contextualizing how decisions are made by the wider community of physicians and clarifying the tradeoffs of different behaviors.

Introduction

Motivation:

- Existing applications of AI clinical decision support are limited in two ways: i) unrealistic claims of prediction value and ii) restriction to labeled tasks.

I) Prediction value:

- Existing models provide guidance by predicting elements of patient physiology in a reliable and consistent way.
- The ability for algorithms to make predictions that extend **beyond** what clinicians already suspect is dependent on the types of data available.
- Models trained over **clinician-initiated data** (diagnoses, procedures, prescriptions, etc.) that are generated through expressions of physician judgement are less likely to generate novel predictions.
- Clinician-initiated data is one of the most available modalities, in the form of administrative claims and EMR. Identifying robust use cases for these data is critical to improving the acceptance of AI.

II) Task selection:

- Existing models rely on gold-standard labels to align predictions against. This scheme is best suited for predictions of physiological fact about a patient (ex: presence/absence of bowel obstruction from x-ray image).
- A much wider set of clinical scenarios are those without labels of correctness: whether a patient **ought** to undergo a procedure or receive a drug is not a matter of fact.
- Models that expect providers to act based on predictions of future behavior alone implicitly assume optimal provider behavior in the training set: a very strong assumption.
- Clarifying the role of AI decision support can expand the types of clinical scenarios they are applied in to include those defined by differences in provider opinion.

Community Risk Intuition (CRI):

- We propose a methodology two address these challenges by contextualizing how a decision is made at the population level.
- In ulcerative colitis (UC) management, the decision of who to refer to surgery and when is a pivotal one.
- Providers must weigh the risks/benefits of intervention against continued management.
- The **nomination** of surgical candidates represents a policy question without labels of correctness/incorrectness.
- The **appropriateness** of model predictions is used instead to evaluate the quality of the algorithm.

Point-of-care (POC) Intervention:

- The optimal window for algorithmic guidance is at the point-of-care.
  - Before: changes in patient state are unknown.
  - After: ability to change intervention is reduced.
- Administrative data used to train models is often only available after an encounter is over.
- In developing a CRI tool for UC **(crit-UC)**, we further consider model performance under these circumstances.
- POC data collection from the provider is simulated to bridge gap of data availability.

Methods

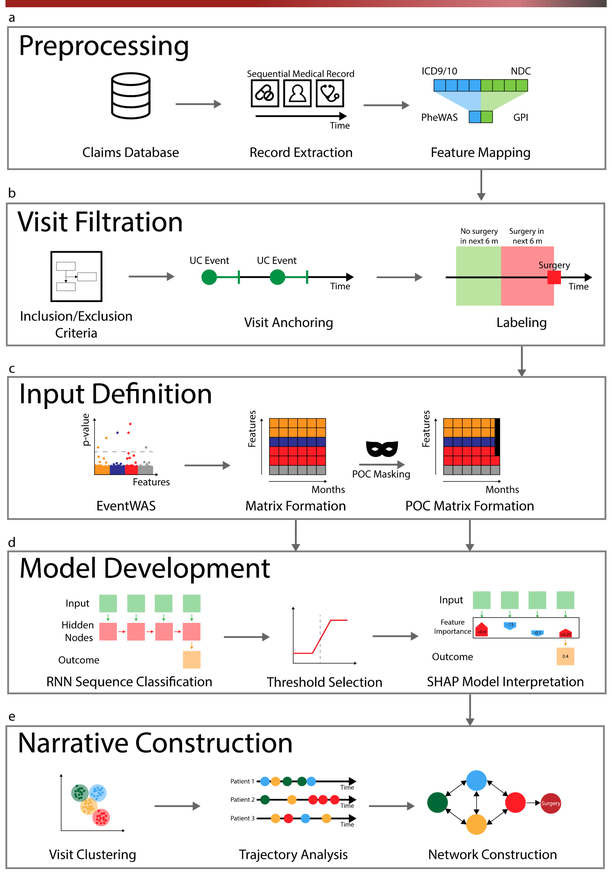

Figure 1: Study Overview

- Administrative coding was used to define a cohort of recurrently managed ulcerative colitis patients.
- A deep learning model was trained over 3M+ patient encounters to predict 6m risk of surgical intervention.
- Decision threshold was selected from short-term outcome analysis.
- Point-of-care deployment was simulated by masking model input.
- Model predictions were validated in an external hospital dataset.
- Predictions evaluated based on if they were "reasonable": if patient presentation was consistent with intervention at time of flag.
- Visit clustering was used to characterize patient trajectories.

Results

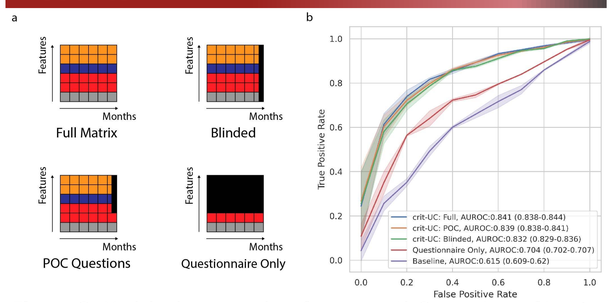

Figure 2: Model schemes and performance. A) Comparison of input blinding methods. B) AUROC of model implementations

- Single-element point-of-care questionnaire ("what drugs, if any, do you intend on prescribing at this encounter?") was necessary and sufficient for model performance.

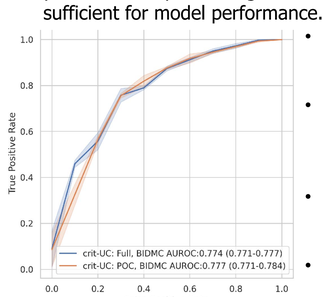

Figure 3: External model validation

- POC model performance was externally validated using EHR data.
- Operationalization of model requires selection of action threshold: the risk score above which intervention is considered.
- Patient risk scores for those who underwent surgery diverged 12+ months prior to surgery.
- Increasing threshold between 0.2 – 0.6 led to limited improvements in precision and F1 score.

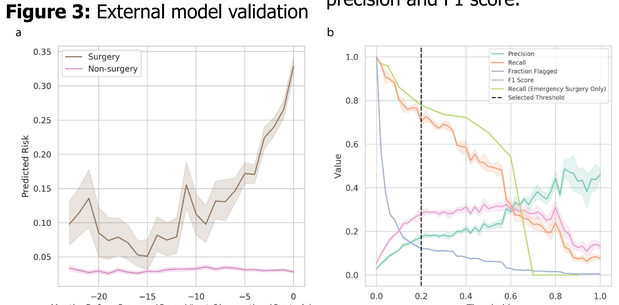

Figure 4: A) Comparison of surgery/non-surgery predicted risk. B) Performance metrics vs. model decision threshold.

Results

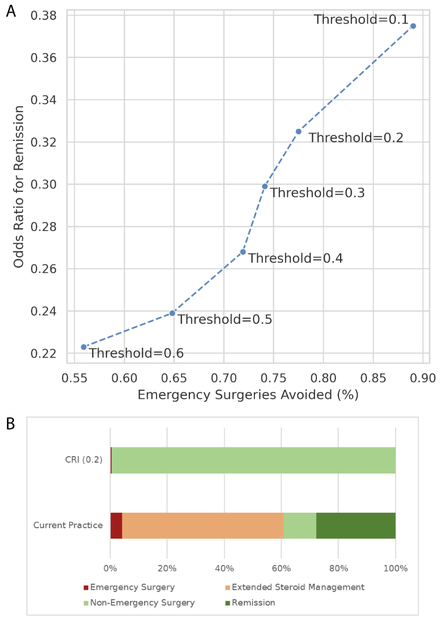

Figure 4: A) Tradeoff landscape between emergency surgery and remission selectivity. B) Outcome comparison between model management and current practice.

- Variation in decision thresholds represents differences in values
- Early intervention minimizes emergency surgery and unnecessary toxic medical exposure.
- Late intervention maximizes rates of remission and minimizes unnecessary surgery.
- Based on these findings, a threshold of 0.2 was selected and is characteristic of substantial variation in observed surgeries.

- Comparison of algorithm-managed care to current practice highlights acuity of patients at time of identification by model.
- A small number of patients deteriorated so quickly that emergency intervention was required before the model was able to flag them.
- 70+% of algorithmically identified patients went on to emergency intervention, delayed non-emergency intervention, or extended medical management.
- Roughly one quarter of flagged patients eventually achieved remission on their own, but rates of "unnecessary" surgery in current practice are unknown due to the censoring nature of the intervention.

	All Test Set Patients (8702)		Non-Surgery Test Set Patients (8256)	
All-time Outcomes (95 CI)	Flagged vs. Unflagged	p-value	Flagged vs. Unflagged	p-value
Surgery (all cause) Odds Ratio	23.1 (16.8-31.9)	<0.005	N/A	N/A
Emergency Surgery Odds Ratio	34.0 (16.7-69.2)	<0.005	N/A	N/A
Steroid-free Remission Odds Ratio	0.325 (0.286-0.368)	<0.005	0.382 (0.336-0.435)	<0.005
Days (all-cause) Surgery Accelerated	244 (185-303 days)	N/A	N/A	N/A
6-month Post-flag Outcomes (95 CI)				
Surgery Odds Ratio	7.20 (5.72-9.07)	<0.005	N/A	N/A
Emergency Surgery Odds Ratio	40.0 (14.4-110.4)	<0.005	N/A	N/A
Steroid-free Remission Odds Ratio	0.55 (0.53-0.57)	<0.005	0.61 (0.57-0.65)	<0.005
Mean Additional mg-equivalent Corticosteroid Consumption	801 (707 - 894)	<0.001	738 (642 - 834)	<0.001
Mean Additional Days with Corticosteroid	23.3 (20.9-25.6)	<0.001	23.6 (21.0-26.2)	<0.001
Mean Additional Costs over ($)	198701 (97800-299603)	<0.005	14542 (8178-20906)	<0.005

Table 1) Outcome comparison between flagged and unflagged patients.

- 80% of flagged patients did not undergo surgery: outcome analysis was used to evaluate if flagged presentations were reasonable.
- Relative to unflagged patients, presentations flagged by the model were substantially more likely to undergo surgery, emergency surgery, and were substantially less likely to achieve remission.
- To examine if flag timing was appropriate, 6-month post-flag outcomes were considered as well.
- In the six months following a flag, flagged patients experienced higher costs of care and consumed substantially more steroids than unflagged patients.
- The all-time odds ratio for surgery was lower than 6-month because the model flagged individuals more than six months in advance on average.

Results

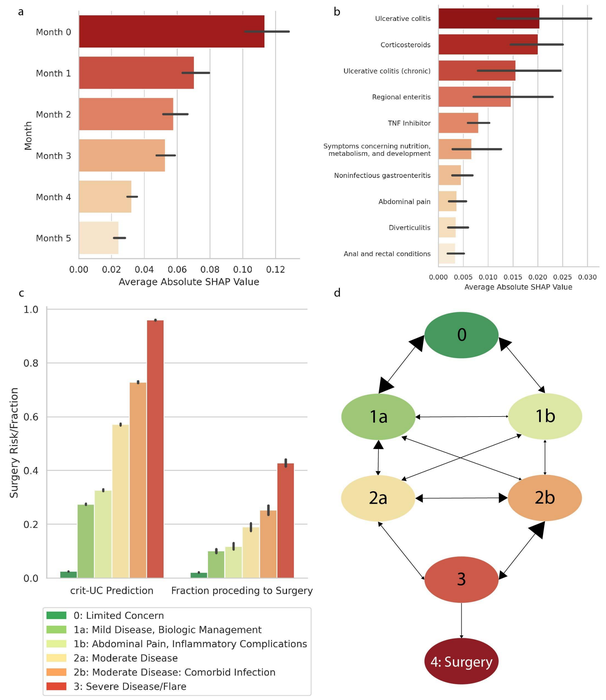

Figure 5) A) Feature interpretation by month. B) Feature interpretation by clinical event. C) Visit cluster property comparison. D) Markov transition diagram between visit clusters inferred through patient trajectories.

- To identify the primary signal utilized by the model, SHAP was used for feature identification.
- Features close to the visit and features relating to UC symptomology were found to be more important, as expected.
- Model predictions tracked with rate of observed surgery.
- Cluster identity alone was found to be sufficient for risk prediction, satisfying Markovian criteria.
- Model predictions were compared to the opinions of a focus group of gastroenterologists considering case vignettes. Panelists preferred behaviors recommended by the model over the behavior of the actual attendings (p = 0.038).

Conclusions

- CRI represents a novel methodology for AI-based decision support in situations without a "correct" choice.
- Model predictions were found to represent reasonable guidance based on analysis of short-term outcomes and manual case review.
- Substantial inter-provider variation in rates of surgery were observed among patients with comparable presentations
- Higher rates of referral may be warranted with the aim of standardizing care and reducing emergency intervention.
- This variation is addressed by leveraging the collective behavior of providers to define clinician-perceived risk levels.
- Individual providers may have their own algorithm for risk assessment but may not know how others would behave in the same situation.

Comparisons to existing AI methods:

- Clinical scenarios with well-defined labels enabling classical AI analysis often lack substantial clinical uncertainty or ambiguity:
  - Detection of diabetic retinopathy: physiological fact
  - Prediction of intubation in intensive care unit: observed behavior so strongly linked with physiology that no alternative behaviors are observed
- CRI does not purport to know patient physiology more than the provider, instead providing insight into how other providers behave and assess risks and benefits associated with an intervention.

Value of CRI to providers:

- CRI has substantial value in the situations that it is "**wrong**" in: presentations where predicted and actual behavior do not align represent opportunities for change.
- For scenarios defined by uncertainty, accuracy is an inappropriate metric of model performance or value.

Limitations

- Patient preferences towards surgical interventions, reimbursement policies, and socioeconomic factors represent important unmeasured potential confounders.
- Unmeasured healthcare interactions could influence model predictions.

Acknowledgements

We would like to thank Daniel Kahneman and Olivier Sibony for their insightful comments on a draft of this work. We would also like to thank the case-study reviewers for their skill and expertise in evaluating patient vignettes. WY is supported by T32HD040128 from the NICHD/NIH. JM is supported by T15LM007092 from the NIH/NLM and the Biomedical Informatics and Data Science Research Training (BIRT) Program of Harvard University. GB is supported by a Blavatnik Pilot Grant at HMS.