

THE ECOSYSTEM OF EVIDENCE

Navigating the future
of the ecosystem of evidence

Artificial intelligence and digital transformation
in research and healthcare

11th EBHC International Conference
Taormina, 21st – 24th October 2026

#EBHC2026

Governing a deployed deep learning model

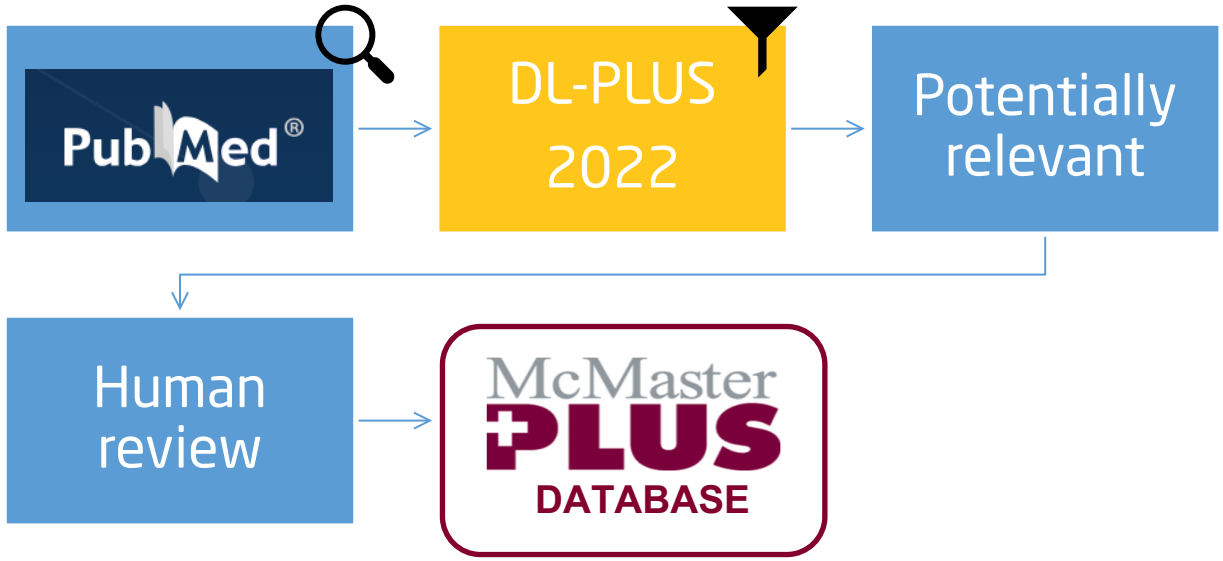
Practical use case determining process and decision points for ensuring optimal performance

Cynthia Lokker, PhD
Health Information Research Unit
McMaster University
Co-authors: Wael Abdelkader; Rick Parrish; Alfonso Iorio, Ashirbani Saha; Muhammad Afzal

30th GIMBE 1996-2026 EVIDENCE FOR HEALTH



Background



DL-PLUS has supported our PLUS process since 2022 (deep learning BIOBERT based model)

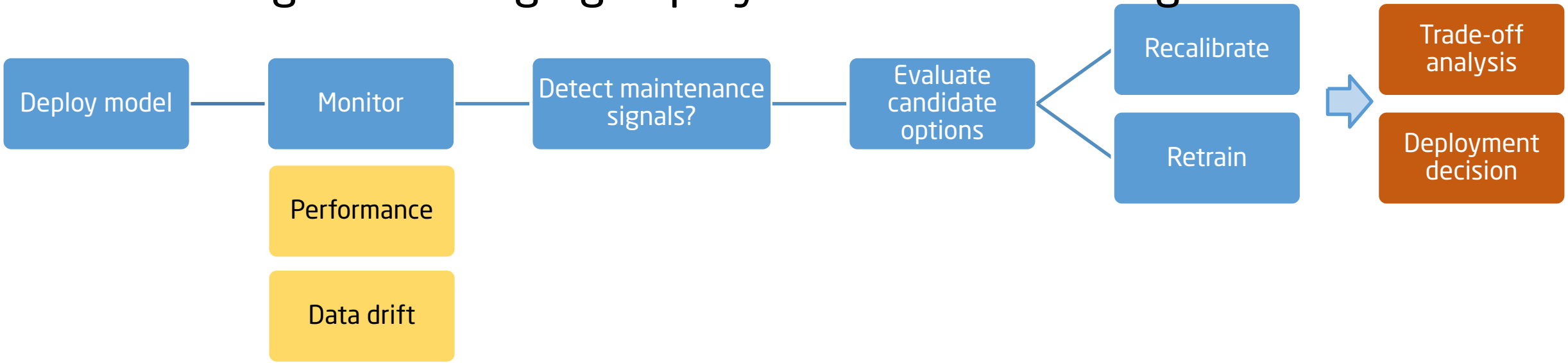
- @99% recall target
- reduced human article review workload by 53%

We know how to build and deploy models.
We don't know how to manage them after deployment.

Lokker C, Bagheri E, Abdelkader W, et al. Deep learning to refine the identification of high-quality clinical research articles from the biomedical literature: Performance evaluation. J Biomed Inform. 2023;142:104384.

Aims

To develop and implement a practical framework for monitoring and managing deployed machine learning models.



If monitoring identifies opportunities for improvement, what do we consider when evaluating our options and making decisions about the next stage?

Methods

1 → 2 → 3

Monitoring

✓We have 42,748 articles processed between 2022 and Sep 2025

✓19,144 annotated by human reviewers

Performance since deployment

- Recall, review burden, number needed to read, work saved, AUROC

Data drift

- Model prediction scores over time
- Precision over time (COVID-19 effect)

Candidate update options

1. Calibrate the threshold score for a positive prediction and assess effect on sensitivity and NNR in the 42K articles

2. Adding the 19K newly tagged articles to our original training dataset, retrain the same model to see if performance is enhanced.

Trade-off analysis

Apply the candidate options to an unseen set of 8620 articles processed by DL-PLUS

Measure the trade-off:

reduction in human review burden for each additional relevant article that is missed (ie., false negative)

Results

- Monitoring identified opportunities to improve efficiency despite stable model performance.
- Threshold recalibration and retraining both improved efficiency, but with different operational trade-offs.

Comparison of candidate options using operational benefits and costs

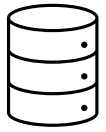
Strategy (external dataset, n = 8620)	Recall	Work saved over sampling	NNR	Operational BENEFIT Article reviews avoided	Operational COST Additional missed articles	Trade-off ratio (avoided/ missed)
Current deployment: DL-PLUS at 3.58%	99%	44%	2.01	0	0	0
Best recalibrated model: DL-PLUS at 10%	98.7%	54%	1.82	441	+7	63
Retrained model DL-PLUS-2 at 3.58%	98.3%	53%	1.67	842	+14	60
Retrained + recalibrated DL-PLUS at 10%	96.2%	56%	1.50	1279	+67	19

The optimal deployment strategy depends on tolerance for missed articles vs. review burden

Limits

Data Limitations

- Articles below the deployment threshold were not routinely reviewed
- Post-deployment sensitivity could not be directly estimated



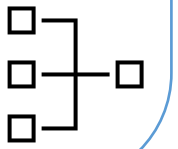
Evaluation Limitations

- Single literature surveillance use case
- Single evaluation cycle



Framework Limitations

- Evaluated only two update options
- Other update options remain unexplored



The framework was evaluated in a real-world deployment and should be tested across additional models, tasks, and domains

Conclusions

- Monitoring showed DL-PLUS remained effective
- **Monitoring identified two reasons to evaluate deployment changes:**
- accumulation of newly annotated articles and opportunities to further reduce reviewer burden
- **Threshold recalibration and retraining were evaluated as candidate update options.**
- **Trade-off analysis enabled transparent comparison of recalibration and retraining.**

Monitoring identifies opportunities for evaluation. Trade-off analysis enables evidence-based deployment decisions.