Uncertainty-Aware Multi-Label Routing of Clinical Text to Surveillance Pathways

Agathe Zecevic[1,2], Angus Roberts[3], Sebastian S. Zeki[1,4]
[1]Faculty of Life Sciences & Medicine, King's College London, United Kingdom
[2]Clinical Scientific Computing, Guy's and St Thomas' NHS Foundation Trust, United Kingdom
[3]Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, United Kingdom
[4]Gastroenterology Department, Guy's and St Thomas' NHS Foundation Trust, United Kingdom
agathe.zecevic2@nhs.net

Abstract

Clinical decision support systems that operate across multiple downstream care pathways must first determine which pathway or pathways are relevant for a given patient. We study this routing problem in gastrointestinal surveillance, where paired endoscopy and histopathology text reports may indicate multiple concurrent conditions and therefore require multi-label routing. In this context, standard hard-label evaluation can be insufficient: a model may achieve reasonable overall performance while still excluding clinically important pathways when uncertain.

We formulate gastrointestinal report routing as a multi-label uncertainty-aware classification task over six pathway labels and compare lightweight lexical baselines, frozen embedding models and a fine-tuned transformer baseline under two complementary uncertainty mechanisms: threshold-based abstention and set-valued conformal prediction. Using 1,773 paired reports from a single NHS trust with disjoint train, calibration and test splits, we evaluate both hard-routing performance and the downstream review burden introduced by uncertainty-aware prediction.

The fine-tuned ClinicalBERT model achieved the strongest overall performance (0.811 subset accuracy, 0.861 macro-F1) and the lowest AURC of 0.084 under min-margin abstention. Threshold-based abstention consistently reduced exact-match routing error on accepted reports. For conformal routing at $\alpha = 0.10$, Mondrian calibration achieved high mean positive-label recall coverage across learned baselines (0.883-0.917). The fine-tuned model achieved 0.891 mean recall coverage with a mean prediction set size of 1.70, 0.642 candidate-label precision and 0.61 false-positive labels per report. Compared with a recall-tuned threshold baseline at similar recall, Mondrian CP produced smaller candidate sets, higher candidate-label precision and fewer false-positive pathway suggestions.

These results show that uncertainty-aware evaluation exposes clinically important failure modes missed by aggregate metrics. They also show that high-recall routing is not cost-free: set-valued prediction can reduce missed-pathway risk but must be interpreted as candidate generation for downstream review rather than automated pathway selection.

1 Introduction

Text-based clinical decision support systems use information recorded in free-text clinical documents to assist with screening, surveillance, triage and follow-up decisions (Sutton et al., 2020). In cases where a single system supports multiple possible downstream actions, an initial routing step is often required to determine which clinical pathway or pathways are relevant for a given patient. We refer to this as the task of *routing clinical text to clinical pathways*.

This task exposes an evaluation gap between NLP-in-the-lab and NLP-in-deployment. Standard multi-label models are typically assessed using aggregate point-prediction metrics such as accuracy or F1-score. However, point prediction alone can be insufficient in health-related use-cases: a model may achieve reasonable overall performance while still excluding a clinically important pathway on the cases where it is most uncertain. A hard routing decision returns a fixed label set, whereas a set-valued prediction can instead return a small candidate set of plausible pathways. An abstaining system can defer highly uncertain cases for manual review. These mechanisms address a question that aggregate point-prediction metrics do not fully capture: not only *what* the model predicts on average, but also *how reliably* it behaves when uncertainty is clinically important. This motivates uncertainty-aware evaluation frameworks for clinical NLP, particularly for multi-label tasks in which several pathways may be simultaneously relevant.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 181–190

We study this problem in gastrointestinal (GI) surveillance. Clinical decision support systems for GI surveillance (Zecevic et al., 2024; Peterson et al., 2021; Song et al., 2022) increasingly target individual conditions and ground themselves in national guidance such as Barrett's oesophagus, colorectal polyps (Fitzgerald et al., 2014; Rutter et al., 2020), or gastric atrophy (Banks et al., 2019). These conditions are clinically important because they are associated with increased risk of disease progression or gastrointestinal malignancy and therefore require pathway-specific follow-up. Decisions often depend on combining information from endoscopy reports describing macroscopic procedure findings, with histopathology reports providing microscopic confirmation from biopsies taken at procedure time. In practice, a single endoscopy-histopathology report pair can indicate multiple concurrent conditions. For instance, incidental colorectal polyps can be discovered during IBD surveillance, making routing inherently a multi-label problem. Misrouting carries significant clinical consequences: a Barrett's patient erroneously sent to the *Normal* pathway may miss surveillance intervals, whereas routing an incomplete or abandoned procedure to an inappropriate Gastric Atrophy pathway wastes downstream resources (Codipilly et al., 2018; White and Banks, 2022).

We therefore formulate GI report routing as a multi-label classification task with two complementary uncertainty-handling mechanisms. First, we study *abstention*, where the model defers cases that are too uncertain for automated routing. Second, we study *set-valued conformal prediction*, where the model returns a small candidate set of pathways designed to better retain clinically relevant labels under uncertainty. We make three contributions:

- We formulate GI surveillance pathway routing from paired endoscopy-histopathology reports as a multi-label uncertainty-aware classification problem.
- We show that threshold-based abstention substantially reduces exact-match routing error on accepted cases across lightweight and transformer-based baselines.
- We show that positive-only Mondrian conformal calibration provides a high-recall candidate-generation layer for clinically important pathways, while explicitly quantifying the associated false-positive review burden.

2 Related Work

Clinical text routing and triage. Prior work on clinical text routing and triage spans several adjacent settings. NLP has been used in referral workflows to predict target speciality from referral notes and triage referrals (Spasic and Button, 2020). Free-text triage has also been widely studied in emergency care (Stewart et al., 2023). However, these studies generally route a single text stream toward one operational endpoint rather than assigning paired reports to potentially multiple concurrent pathways.

Selective prediction in NLP. Selective prediction adds an abstention layer to traditional classifiers, producing risk-coverage trade-offs (El-Yaniv and Wiener, 2010). Wen et al. (2025) survey selective prediction in the context of LLMs, summarising common approaches such as threshold-based and learned selection functions. Fisch et al. (2022) address multi-label settings with conformal guarantees that limit false-positive predictions, which is critical for domains like healthcare.

Conformal prediction. Split conformal prediction provides distribution-free marginal coverage guarantees by calibrating nonconformity scores on a held-out set (Vladimir Vovk, 2005; Angelopoulos and Bates, 2022). The Mondrian extension refines this idea by yielding label-conditional guarantees that are especially relevant under class imbalance or asymmetric misrouting costs. Campos et al. (2024) survey conformal prediction for NLP, including both marginal and class-conditional variants. In clinical applications, conformal methods have recently been explored for medical multiple-choice question answering (Ke et al., 2025), medical text mining for epidemiological surveillance (Genari and Goedert, 2025), multi-label ICD coding from clinical narratives (Zhang et al., 2025a) and medical entity extraction from radiology text (Shrestha and Kim, 2026).

This study is distinct from our prior Barrett's oesophagus decision-support work, which focused on condition-specific surveillance scheduling. This work studies the upstream multi-label routing problem across six GI pathways using a newly annotated routing dataset.

3 Methods

3.1 Task Formulation

Given a merged endoscopy-histopathology report x, we predict a label set $\hat{Y} \subseteq \mathcal{L}$

where $\mathcal{L} = \{$*Normal*, *Barrett's*, *Colorectal Polyps*, *Gastric Atrophy*, *Other*, *Incomplete*$\}$. A selective predictor (f, g) consists of a base classifier f and a selection function $g : \mathcal{X} \to \{0, 1\}$ that determines whether to accept ($g(x) = 1$) or abstain ($g(x) = 0$). We decompose routing into six independent binary decisions, one per label, with associated probability scores $f_j(x) \in [0, 1]$. We evaluate via risk-coverage curves: *selective risk*, which is defined as the error rate on accepted instances versus *coverage*, defined as the fraction of instances accepted. For the point-prediction baselines, each label j is predicted positive when $f_j(x) \geq 0.5$.

Label logic. Gold labels are multi-label, but not all label combinations are clinically coherent. In our case, the label *Normal* denotes the absence of any pathway-relevant clinical finding and therefore does not co-occur with any other label in the annotated datasets.

3.2 Baselines

We compare four clinical pathway routing baselines with increased computational complexity.

B0: Keyword baseline This pipeline consists of regex patterns for each class matched against the merged report text. We incorporate negation cues (such as "no evidence of", "history of", "to rule out") to improve baseline performance. We derive heuristic scores for each label as the number of matched patterns divided by the total number of patterns defined for that label.

B1: TF-IDF + Logistic Regression. We split each report into affirmed and negated text spans using the same negation cues as B0 and fit separate TF-IDF vectorisers to each channel. For each label, the regularisation strength $C \in \{0.01, 0.1, 1.0, 10.0\}$ was selected by 3-fold cross-validation on the training split only, using macro-F1 as the selection criterion. Per-label probability calibration was performed on the training split only via Platt scaling with stratified K-fold cross-validation (K = 3). The positive and negative channels use feature budgets of 10,000 and 5,000 features respectively and are concatenated into a 15,000-dimensional representation. Both vectorisers use n-grams of order 1-2.

B2: Frozen BGE-Large embeddings + Linear head. The frozen-embedding baseline extracts fixed representations from BGE-Large-EN-v1.5, a text encoder pre-trained with contrastive learning (Xiao et al., 2024). Each report is tokenised to a maximum of 512 subword tokens and encoded with the fixed model weights. A masked mean pool over the last hidden states of non-padding tokens produces a single 1024-dimensional vector per report. A separate one-vs-rest logistic regression classifier is then fitted per label on these frozen embeddings.

B3: Fine-tuned ClinicalBERT. We fine-tune ClinicalBERT (Alsentzer et al., 2019) for multi-label classification by adding a linear classification head on top of a masked mean-pool representation of the encoder's last hidden state. The head produces six independent logits, trained with binary cross-entropy. We optimise with AdamW (weight decay 0.01) using a linear warmup over the first 10% of training steps followed by linear decay to zero. Gradients are clipped to a maximum norm of 1.0. Learning rate and number of epochs are selected jointly by grid search over $10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}$ and up to 15 epochs, evaluated by macro F1 on a 15% hold-out subset of the training split kept strictly separate from the calibration data.

For each baseline, we compute results at a uniform threshold of 0.5 and at per-label thresholds selected on the calibration set to achieve at least 90% recall for each label independently.

3.3 Selective Prediction

Threshold abstention. We use a *min-margin* selection function: accept report x if

$$g_\tau(x) = \min_j \max\big(f_j(x),\, 1 - f_j(x)\big) \;\geq\; \tau, \quad (1)$$

where $\tau \in [0.5, 1]$. This score equals the confidence of the least certain label: it falls to 0.5 whenever any label score is close to the decision boundary, so a report is abstained as soon as any single label is uncertain. This aligns the selection function directly with exact-match risk, which fails whenever any label is wrong. By contrast, max-confidence ($\max_j f_j$) can accept a case with one highly confident label while several others remain uncertain, which is an undesirable behaviour in our setting. We vary τ to generate risk-coverage curves and report the Area Under the Risk-Coverage curve (AURC) (Geifman et al., 2019), where a lower AURC indicates better selective classification.

3.4 Conformal Prediction

Conformal prediction (Vladimir Vovk, 2005) is a distribution-free framework for constructing prediction sets with finite-sample coverage guarantees, requiring only data exchangeability. We apply it

independently per label j and evaluate whether the returned candidate set retains truly present labels. For each label, the main quantity of interest is the fraction of positive test examples for which label j is included in the prediction set. We report this as per-label recall coverage:

$$P\big(j \in \mathcal{C}(x) \mid Y_j = 1\big) \;\geq\; 1 - \alpha, \qquad (2)$$

where $\mathcal{C}(x) \subseteq \mathcal{L}$ is the prediction set returned for report x. We report empirical positive-class inclusion as the fraction of positive test examples for which label *j* is included in the prediction set, we do not make any joint coverage claim across labels or across the full multi-label set.

Split conformal prediction. We apply split conformal prediction independently per label j, using the LAC (Least Ambiguous Classifier) nonconformity score as implemented in MAPIE (Taquet et al., 2022):

$$s_j(x, y_j) = 1 - f(x)_{j,y_j}, \qquad (3)$$

where $f(x)_{j,y_j}$ is the model's predicted probability for the *true* class $y_j \in \{0, 1\}$ of label j. For each label j, we set the calibration threshold $\hat{q}_j$ to the $\lceil (n_j + 1)(1 - \alpha) \rceil / n_j$ empirical quantile of the calibration scores $\{s_j(x_i, y_{j,i})\}_{i=1}^{n_j}$, computed either over all calibration examples (marginal calibration) or over the positive calibration examples only (Mondrian calibration). Label j is included in the prediction set if

$$f_j(x) \geq 1 - \hat{q}_j, \qquad (4)$$

that is, if the model assigns sufficiently high probability to the positive class.

Marginal versus Mondrian calibration. Under *marginal* calibration, $\hat{q}_j$ is estimated from all calibration examples for label j, mixing positive and negative instances. This does not directly target the conditional quantity in Eq. (2); under class imbalance, it can therefore yield poor positive-class inclusion for rare labels. For a minority label such as GA in our case, most calibration examples are negative and the model typically assigns them small positive-class probabilities, which leads to calibration scores clustering near 0. In our experiments, this noticeably reduced GA inclusion under marginal calibration at $\alpha = 0.10$. This is mitigated in *Mondrian* calibration, where $\hat{q}_j$ is instead estimated only from the positive calibration examples for label j, matching the positive-class inclusion objective in Eq. (2).

Label	Train ($n = 789$)		Cal. ($n = 523$)		Test ($n = 461$)	
	n	%	*n*	%	*n*	%
Normal	77	9.8	53	10.1	45	9.8
Barr	185	23.4	65	12.4	50	10.8
CP	115	14.6	53	10.1	53	11.5
GA	63	8.0	30	5.7	42	9.1
Other	500	63.4	361	69.0	314	68.1
Incomplete	103	13.1	50	9.6	60	13.0
Multi-label (≥ 2)	249	31.6	89	17.0	101	21.9
Mean label cardinality	1.32		1.17		1.22	

Table 1: Per-label prevalence across datasets. *Multi-label* reports cases with two or more labels. *Mean label cardinality* is the average number of labels per report.

Abstention rule. We add a separate deployment heuristic on top of conformal prediction. A report is abstained when its prediction set size $|\mathcal{C}(x)| > K$, with $K = 3$. This size threshold is not part of the conformal guarantee but a pragmatic rule intended to defer highly ambiguous cases to manual review. We chose $K = 3$ a priori based on the observed degree of clinically plausible co-occurrence in the training and calibration sets. We evaluate at $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$ and report per-label empirical recall coverage (Eq. (2)), mean prediction set size and abstention rate.

4 Experiments

4.1 Data

We use six months of anonymised paired endoscopy and histopathology reports from an NHS hospital trust, merged into single documents per procedure. Data acquisition was authorised through institutional board review (IRAS ID = 257283). We remove patients who had opted out of their data being used (NHS Data Opt-Out). We construct three non-overlapping splits: a training set ($n = 789$), a calibration set ($n = 523$) and a test set ($n = 461$). The training set only is enriched for the minority classes (Barr, GA, Incomplete), the calibration and test set are not enriched.

Reports were manually annotated by a clinical researcher using labels from $\mathcal{L}$. Ambiguous cases (for instance "history of Barrett's" with no active findings) were resolved via predefined annotation rules. Table 1 reports the label distribution across datasets and the average number of labels per merged report. More details about the annotation rules and label co-occurrence can be found in Appendix A.

4.2 Results

Baseline comparison. Table 2 compares the four routing baselines on exact-match subset accuracy,

Model	Subset Acc. [95% CI]	$F1_{macro}$ [95% CI]	$Precision_{macro}$ [95% CI]	AURC [95% CI]
Keyword	0.538 [0.492, 0.584]	0.705 [0.670, 0.737]	0.646 [0.612, 0.682]	0.3718 [0.3183, 0.4264]
TF-IDF	0.646 [0.605, 0.692]	0.782 [0.743, 0.817]	0.840 [0.799, 0.880]	0.1683 [0.1318, 0.2119]
Embedding	0.705 [0.666, 0.746]	0.776 [0.733, 0.814]	**0.898** [0.861, 0.932]	0.1338 [0.1026, 0.1663]
Fine-tuned	**0.811** [0.777, 0.846]	**0.861** [0.829, 0.891]	0.873 [0.837, 0.910]	**0.0836** [0.0578, 0.1127]

Table 2: Baseline comparison on point-prediction and selective prediction performance. Reported metrics are subset accuracy, macro F1, macro precision and area under the risk-coverage curve (AURC). 95% confidence intervals were computed by non-parametric bootstrap over test reports with 1,000 resamples.

macro F1, and AURC. TF-IDF is a competitive lightweight baseline, achieving 0.646 subset accuracy and 0.782 macro-F1. The frozen BGE-Large embedding model improves subset accuracy to 0.705 and exhibits the highest macro precision (0.898). The fine-tuned ClinicalBERT model performs best overall, reaching 0.811 subset accuracy, 0.861 macro-F1 and the lowest AURC (0.0836). Paired bootstrap comparisons in Appendix B support the improvement of the fine-tuned model over the other learned baselines.

Selective prediction. Figure 1 shows risk–coverage curves under min-margin abstention. All models show decreasing selective risk as coverage decreases. This indicates that the confidence score is informative about cases where hard routing is more reliable. The fine-tuned model has the lowest AURC (0.084), followed by the frozen embedding model (0.134) and TF-IDF (0.168). The keyword baseline remains substantially weaker both at full coverage and under abstention, which can be explained by how its confidence scores are computed.

Conformal prediction. Table 3 summarises Mondrian conformal routing at $\alpha = 0.10$ for the learned baselines. Mean recall coverage remains high across models, ranging from 0.883 for the frozen embedding model to 0.917 for TF-IDF. The operational burden differs across baselines. The fine-tuned model provides the best trade-off, with the highest set precision (0.642), the smallest mean set size (1.70), the lowest false-positive burden (0.61 FP labels/report) and no abstentions under the $K = 3$ rule. These results show why recall coverage alone is insufficient: high-recall conformal routing must be interpreted together with set precision and false-positive pathway burden.

Figure 2 shows the conformal operating trade-off. At the set level (Panel A), we show that decreasing α increases mean recall coverage but reduces candidate-label precision. The learned models follow similar trade-off curves, with the fine-tuned model providing the best operating point at $\alpha = 0.10$ in Table 3. Panel B uses GA to illustrate the clinical cost of high-recall routing: Mondrian CP substantially increases GA recall relative to the hard classifier but with a marked reduction in precision.

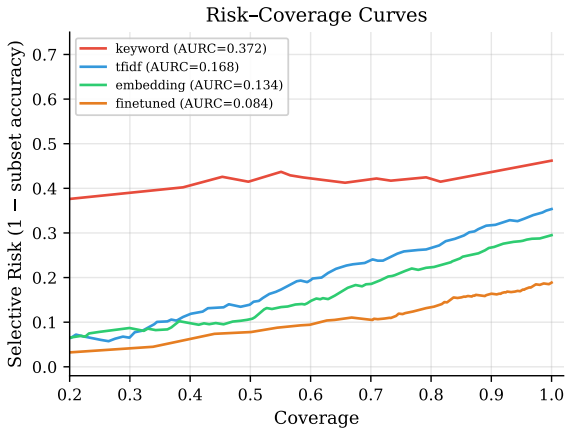

Figure 1: Selective risk-coverage curves under min-margin abstention. The low-coverage region (<0.2) is omitted as estimates become unstable with few accepted examples. AURC is computed over the full coverage range.

Model	Recall cov. ↑	Set prec. ↑	Mean set ↓	FP/report ↓	Abstain % ↓
TF-IDF	**0.917**	0.589	1.89	0.77	1.3
Embedding	0.883	0.607	1.79	0.70	1.1
Fine-tuned	0.891	**0.642**	**1.70**	**0.61**	**0.0**

Table 3: Operational summary of Mondrian conformal routing at $\alpha = 0.10$. Recall cov. is mean per-label positive-class inclusion coverage over test positives. Set precision is the candidate-label precision, defined as $\sum_i |\mathcal{C}_i \cap Y_i| / \sum_i |\mathcal{C}_i|$. FP/report is the average number of false-positive pathway labels included per report.

Per-class analysis. Table 4 compares hard prediction, a recall-tuned threshold baseline, marginal CP and Mondrian CP for the fine-tuned model at $\alpha = 0.10$. The recall-tuned threshold baseline achieves high mean recall (0.895), but candidate-label precision falls to 0.540, mean set size increases to 2.03 and the false-positive burden rises to 0.93 labels/report. Mondrian CP achieves a similar mean recall (0.891) with higher candidate-label precision (0.642), smaller sets (1.70) and fewer false-positive labels per report (0.61).

The GA row shows the main minority-label trade-off. Hard prediction has relatively high GA precision (0.80) but lower recall (0.67). Mondrian

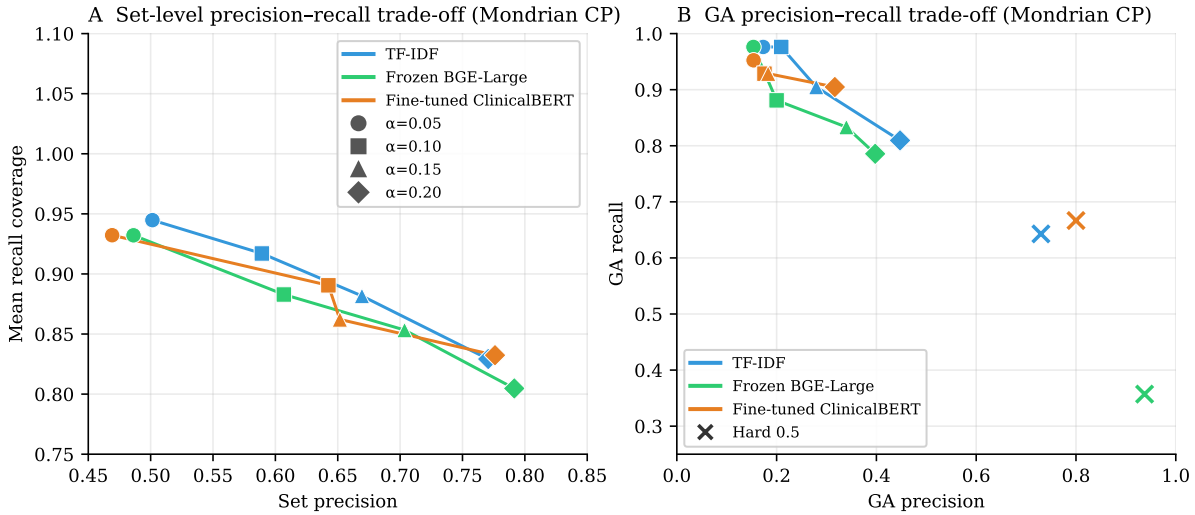

Figure 2: Precision-recall trade-offs for Mondrian conformal routing. **A:** Set-level trade-off across $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$ for the learned baselines, where precision is candidate-label precision and recall is mean per-label positive-class recall coverage. **B:** GA-specific precision–recall trade-off under Mondrian CP. The hard 0.5-threshold classifiers are shown as reference points.

CP increases GA recall to 0.93, but GA precision falls to 0.18, indicating that the conformal layer acts as a high-recall safety net rather than a final automatic GA decision. Marginal CP under-recovers positives for several labels. These results support interpreting conformal routing as candidate generation for downstream review.

5 Discussion and Conclusion

We evaluated a range of routing baselines with different computational requirements, motivated by the practical constraint that not all hospitals can deploy the same class of model. Even relatively lightweight approaches such as TF-IDF achieved reasonable overall performance on our annotated dataset. We highlight the fact that standard point-prediction metrics alone are not sufficient for this task as they do not characterise the risk of excluding clinically relevant pathways from a set-valued prediction. Some misclassifications have important clinical consequences, particularly false negatives on pathways where missing the correct route could prevent the relevant surveillance logic from being applied at all.

This motivated two separate uncertainty-handling strategies at inference time. First, we studied threshold-based abstention to reduce hard-routing errors on accepted cases. Second, we studied conformal set prediction to reduce the risk that clinically important pathways are excluded from the returned candidate set. Recall on clinically important pathways is more crucial than forcing a single-label decision in every case. Returning a small set of plausible pathways may be preferable to excluding the true one, because an irrelevant downstream pathway may simply yield no action, whereas omission of the correct pathway risks missing the appropriate follow-up entirely.

Our results support three conclusions. First, threshold-based abstention reduces exact-match routing error on accepted cases, with the strongest selective performance obtained by the fine-tuned model. Second, set-valued conformal prediction changes the evaluation question from whether a single hard label set is correct to whether clinically important pathways are retained for review. Third, high-recall routing is not cost-free. While Mondrian CP can recover minority pathways such as GA, this comes at the cost of increasing false-positives and therefore downstream review burden.

This distinction is important for deployment. A conformal set should not be interpreted as a final diagnosis or automatic pathway activation. Instead, it is a candidate-generation layer: labels included in $\mathcal{C}(x)$ are pathways that should not be safely excluded without further pathway-specific logic or manual review. Under this interpretation, Mondrian CP provides a useful safety mechanism because it reduces missed-pathway risk while making the review burden explicit. The comparison with a recall-tuned threshold baseline further showed that in our case, Mondrian CP achieved a better recall-

Method	Overall operating point					Per-label recall / precision					
	Mean recall	Set precision	Mean set	FP/report	Abstain %	Normal	Barrett's	CP	GA	Other	Incomplete
Hard 0.5	0.852	0.903	1.20	0.12	—	0.91 / 0.84	0.78 / 0.87	0.96 / 0.89	0.67 / 0.80	0.92 / 0.93	0.87 / 0.91
Tuned on cal. (≥90% recall)	0.895	0.540	2.03	0.93	—	0.93 / 0.40	0.74 / 0.86	0.96 / 0.80	0.95 / 0.11	0.90 / 0.93	0.88 / 0.88
Marginal CP	0.726	0.894	1.15	0.12	0.0	0.80 / 0.88	0.54 / 0.84	0.85 / 1.00	0.43 / 1.00	0.96 / 0.87	0.78 / 0.96
Mondrian CP	0.891	0.642	1.70	0.61	0.0	0.91 / 0.44	0.74 / 0.86	0.96 / 0.80	0.93 / 0.18	0.90 / 0.93	0.90 / 0.89

Table 4: Operating-point and per-label performance for the fine-tuned model at $\alpha = 0.10$

burden trade-off than simple recall-thresholding for the fine-tuned model.

Limitations. Our study uses reports from a single NHS site, in English, with a six-class routing schema, labelled by a single annotator. One class (*Other*) is heavily represented as it is a deliberate catch-all for conditions outside the current surveillance scope (for instance IBD, strictures, malignancy). The calibration set remains small for the rarest classes. With few positive calibration examples per minority class, the Mondrian threshold $\hat{q}_j$ can only take a small number of distinct values, making coverage insensitive to small changes in α.

Future work. We plan to add a local LLM zero-shot/few-shot baseline expanding on already existing work for single conditions (Zhang et al., 2025b), extend the routing schema to additional clinical pathways currently labelled as *Other* and validate the system prospectively on several temporally distant test sets.

Acknowledgments

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [Grant reference number EP/Y035216/1] Centre for Doctoral Training in Data-Driven Health (DRIVE-Health) at King's College London, with additional support from the South East London Cancer Alliance (SELCA).

This research was supported by the Clinical Research Data and Analytics platform within Research and Development, Guy's and St Thomas' NHS Foundation Trust. The views expressed are those of the author(s) and not necessarily those of the NHS.

Code and Data availability

The datasets generated or analysed, or both during this study are not publicly available owing to ethical restrictions. Guy's and St Thomas' Electronic Research Records Interface (GERRI) is governed by the overarching Clinical Research Analytics Governance Group (CRAG).

The code used to run the experiments and reproduce the analyses can be made available upon reasonable request.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings.

Anastasios N. Angelopoulos and Stephen Bates. 2022. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification.

Matthew Banks, David Graham, Marnix Jansen, Takuji Gotoda, Sergio Coda, Massimiliano di Pietro, Noriya Uedo, Pradeep Bhandari, D Mark Pritchard, Ernst J Kuipers, Manuel Rodriguez-Justo, Marco R Novelli, Krish Ragunath, Neil Shepherd, and Mario Dinis-Ribeiro. 2019. British Society of Gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. *Gut*, 68(9):1545–1575.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. 2024. Conformal Prediction for Natural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

Don Chamil Codipilly, Apoorva Krishna Chandar, Siddharth Singh, Sachin Wani, Nicholas J. Shaheen, John M. Inadomi, Amitabh Chak, and Prasad G. Iyer. 2018. The Effect of Endoscopic Surveillance in Patients With Barrett's Esophagus: A Systematic Review and Meta-analysis. *Gastroenterology*, 154(8):2068–2086.

Ran El-Yaniv and Yair Wiener. 2010. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research*, 11(53):1605–1641.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. 2022. Conformal Prediction Sets with Limited False Positives.

Rebecca C Fitzgerald, Massimiliano di Pietro, Krish Ragunath, Yeng Ang, Jin-Yong Kang, Peter Watson, Nigel Trudgill, Praful Patel, Philip V Kaye, Scott Sanders, Maria O'Donovan, Elizabeth Bird-Lieberman, Pradeep Bhandari, Janusz A Jankowski, Stephen Attwood, Simon L Parsons, Duncan Loft, Jesper Lagergren, Paul Moayyedi, and 2 others. 2014. British Society of Gastroenterology guidelines on the

diagnosis and management of Barrett's oesophagus. *Gut*, 63(1):7–42.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. 2019. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers.

Juliano Genari and Guilherme Tegoni Goedert. 2025. Mining Unstructured Medical Texts With Conformal Active Learning.

Yusong Ke, Hongru Lin, Yuting Ruan, Junya Tang, and Li Li. 2025. Correctness Coverage Evaluation for Medical Multiple-Choice Question Answering Based on the Enhanced Conformal Prediction Framework.

Emma Peterson, Folasade P. May, Odet Kachikian, Camille Soroudi, Bita Naini, Yuna Kang, Anthony Myint, Gordon Guyant, Joann Elmore, Roshan Bastani, Cleo Maehara, and William Hsu. 2021. Automated identification and assignment of colonoscopy surveillance recommendations for individuals with colorectal polyps. *Gastrointestinal Endoscopy*, 94(5):978–987.

Matthew D Rutter, James East, Colin J Rees, Neil Cripps, James Docherty, Sunil Dolwani, Philip V Kaye, Kevin J Monahan, Marco R Novelli, Andrew Plumb, Brian P Saunders, Siwan Thomas-Gibson, Damian J M Tolan, Sophie Whyte, Stewart Bonnington, Alison Scope, Ruth Wong, Barbara Hibbert, John Marsh, and 3 others. 2020. British Society of Gastroenterology/Association of Coloproctology of Great Britain and Ireland/Public Health England post-polypectomy and post-colorectal cancer resection surveillance guidelines. *Gut*, 69(2):201.

Manil Shrestha and Edward Kim. 2026. Conformal Prediction for Risk-Controlled Medical Entity Extraction Across Clinical Domains.

Gyuseon Song, Su Jin Chung, Ji Yeon Seo, Sun Young Yang, Eun Hyo Jin, Goh Eun Chung, Sung Ryul Shim, Soonok Sa, Moongi Simon Hong, Kang Hyun Kim, Eunchan Jang, Chae Won Lee, Jung Ho Bae, and Hyun Wook Han. 2022. Natural Language Processing for Information Extraction of Gastric Diseases and Its Application in Large-Scale Clinical Research. *Journal of Clinical Medicine*, 11(11):2967.

Irena Spasic and Kate Button. 2020. Patient Triage by Topic Modeling of Referral Letters: Feasibility Study. *JMIR Medical Informatics*, 8(11):e21252.

Jonathon Stewart, Juan Lu, Adrian Goudie, Glenn Arendts, Shiv Akarsh Meka, Sam Freeman, Katie Walker, Peter Sprivulis, Frank Sanfilippo, Mohammed Bennamoun, and Girish Dwivedi. 2023. Applications of natural language processing at emergency department triage: A narrative review. *PLOS ONE*, 18(12):e0279953.

Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1):17.

Vianney Taquet, Vincent Blot, Thomas Morzadec, Louis Lacombe, and Nicolas Brunel. 2022. MAPIE: an open-source library for distribution-free uncertainty quantification.

Glenn Shafer Vladimir Vovk, Alexander Gammerman. 2005. *Algorithmic Learning in a Random World*. Springer-Verlag, New York.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know Your Limits: A Survey of Abstention in Large Language Models.

Jonathan R. White and Matthew Banks. 2022. Identifying the pre-malignant stomach: from guidelines to practice. *Translational Gastroenterology and Hepatology*, 7:8–8.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed Resources For General Chinese Embeddings.

Agathe Zecevic, Laurence Jackson, Xinyue Zhang, Polychronis Pavlidis, Jason Dunn, Nigel Trudgill, Shahd Ahmed, Pierfrancesco Visaggi, Zanil YoonusNizar, Angus Roberts, and Sebastian S. Zeki. 2024. Automated decision making in Barrett's oesophagus: development and deployment of a natural language processing tool. *npj Digital Medicine*, 7(1):312.

Xin Zhang, Qiyu Wei, Yingjie Zhu, Fanyi Wu, and Sophia Ananiadou. 2025a. THCM-CAL: Temporal-Hierarchical Causal Modelling with Conformal Calibration for Clinical Risk Prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 916–928, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xinyue Zhang, Agathe Zecevic, Sebastian Zeki, and Angus Roberts. 2025b. Improving Barrett's Oesophagus Surveillance Scheduling with Large Language Models: A Structured Extraction Approach. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 176–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

A Label Schema and Annotation Rules

Table 5 summarises the six routing categories used for manual annotation of the multi-label task.

Annotation principles. Non-gastrointestinal background conditions were ignored unless they directly changed the interpretation of the procedure. Minor incidental findings that do not typically define a surveillance pathway on their own such as hiatus hernia or inlet patch were not sufficient to trigger *Other* in isolation.

Label	Description
Normal	No abnormal clinical finding identified, procedure is normal
Barrett's	Active Barrett's oesophagus finding
CP	Colorectal polyp finding of any relevant subtype
GA	Gastric atrophy and/or gastric intestinal metaplasia
Other	Active non-target pathology outside the predefined surveillance pathways (for instance: stricture, ulcer, tumour, coeliac-related abnormality)
Incomplete	Incomplete, failed, or abandoned procedure requiring repeat assessment or rebooking

Table 5: Routing label schema used for manual annotation.

Operational rules for recurrent ambiguous cases. The following predefined rules were applied throughout annotation:

- Historical-only mentions were ignored for routing. For example, "history of Barrett's" without evidence of active Barrett's in the current report was not annotated as *Barr*.
- *Normal* was reserved for reports with no pathway-relevant clinical finding and therefore did not co-occur with other labels in the annotations.
- Gastric atrophy and gastric intestinal metaplasia were annotated as *GA*. Co-occurring gastritis alone did not trigger an additional *Other* label.
- Atrophy described outside the stomach was not annotated as *GA*.
- *Incomplete* was assigned for failed, abandoned or incomplete procedures and could co-occur with other active pathway labels when relevant findings were also present.
- *Other* captured active non-target abnormalities outside the predefined surveillance pathways.

B Significance tests

Table 6: Pairwise paired bootstrap significance tests ($n = 1{,}000$ resamples, two-tailed). $\Delta = \text{score}(A) - \text{score}(B)$. $^{*}p < 0.05$, $^{**}p < 0.01$

Model A	Model B	Δ Subset Acc $\uparrow$	Δ Macro F1 $\uparrow$	Δ AURC $\downarrow$
Keyword	TF-IDF	-0.108^{**}	-0.077^{**}	$+0.203^{**}$
Keyword	Embedding	-0.167^{**}	-0.071^{**}	$+0.238^{**}$
Keyword	Fine-tuned	-0.273^{**}	-0.156^{**}	$+0.288^{**}$
TF-IDF	Embedding	-0.059^{**}	$+0.006$	$+0.034^{*}$
TF-IDF	Fine-tuned	-0.165^{**}	-0.079^{**}	$+0.085^{**}$
Embedding	Fine-tuned	-0.106^{**}	-0.085^{**}	$+0.050^{**}$