

Article

Set Prediction for Outpatient Diagnosis Coding with Sparse Mahalanobis Conformal Scoring

Kamonrat Tangudomkit , Sawrawit Chairat  and Sitthichok Chaichulee * 

Department of Biomedical Sciences and Biomedical Engineering, Faculty of Medicine, Prince of Songkla University, Songkhla 90110, Thailand

* Correspondence: sitthichok.c@psu.ac.th

Abstract

Diagnosis coding is a large-scale multi-label task in which each clinical encounter may require one or more coding labels from a large label space. Conventional top- k and threshold-based classifiers provide practical coding suggestions but do not directly characterize uncertainty over alternative coding sets. This study proposes sparse Mahalanobis conformal scoring for set prediction in diagnosis coding under extreme multi-label classification (XMC), intended for coding-assist workflows that require compact and reviewable coding suggestions. A sparse XMC model first generates candidate coding labels for each encounter. Candidate label sets are then constructed from the sparse proposal space and scored using a diagonal Mahalanobis nonconformity function calibrated on held-out data. Empirical conformal p -values are assigned to candidate sets, and downstream decision rules are used to obtain a final coding output from the retained region. The framework was evaluated using outpatient EHR data from a tertiary-care hospital, comprising approximately 8.0 million visits from 2018 to 2025 and up to 12,829 diagnosis labels. The primary SMaCS output achieved Micro-F1 close to the strongest threshold-based comparator and the highest exact match ratio among flexible-size decision rules. Compared with the other nonconformity scores, the Mahalanobis score produced a smaller retained region with fewer distinct labels, while preserving the same point-prediction performance. Additional analyses examined conformal region validity, robustness to label-frequency thresholds, code-depth performance, label-frequency subgroups, sample cardinality, department-level variation, and confidence–credibility stratification. Our results suggest that sparse Mahalanobis conformal scoring provides a useful framework for uncertainty-informed outpatient coding set prediction, while also highlighting the importance of candidate-space adequacy in extreme multi-label diagnosis coding.

Keywords: diagnosis coding; electronic health records; set prediction; conformal scoring; extreme multi-label classification; sparse prediction



Academic Editor: Carson K. Leung

Received: 11 May 2026

Revised: 22 June 2026

Accepted: 7 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Diagnosis coding is the process of transforming clinical encounter data into standardized diagnostic codes. These codes are important not only for clinical documentation, but also for reimbursement, quality monitoring, epidemiological analysis, and health-system planning [1,2]. In routine outpatient care, a single visit may contain multiple symptoms, follow-up conditions, chronic comorbidities, and diagnostic impressions. As a result, the prediction target is often not one code, but a set of codes assigned to the same en-

counter. This makes outpatient coding more naturally a set-prediction problem than a simple single-label classification problem.

From a machine learning perspective, outpatient coding presents the main characteristics of extreme multi-label classification (XMC): a large output space, highly imbalanced label frequencies, a long tail of rare codes, dependencies among clinically related codes, and substantial variation in the number of codes assigned to individual encounters. A practical coding system must therefore determine not only which codes are relevant, but also how many codes should be included in the final prediction.

Most multi-label coding models produce per-label scores or ranked candidate lists that must subsequently be converted into discrete code sets. Common decision rules, such as fixed top- k prediction and global thresholding, are computationally efficient but impose restrictive assumptions. A fixed top- k rule assigns the same number of codes to every encounter, despite substantial variation in encounter complexity and code cardinality. Threshold-based rules can adapt the prediction size, but their behavior may differ considerably between frequent and rare labels. These rules produce a single final code set and therefore do not convey uncertainty about other clinically plausible coding combinations.

1.1. Research Gaps

Extending outpatient diagnosis coding from point prediction to set prediction raises a practical question: how can clinically plausible coding sets be represented at scale? Conformal prediction (CP) provides a framework for constructing uncertainty-aware prediction sets using a held-out calibration set and a nonconformity score [3–5]. However, applying CP directly to outpatient diagnosis coding presents two closely related challenges. First, set-valued prediction over a large multi-label space may require evaluating an enormous number of possible code combinations. Because the number of candidate sets grows combinatorially with the number of labels, direct conformal prediction over the full label power set is generally computationally impractical in an extreme multi-label setting. Second, conformal scoring can evaluate only the candidate sets considered during inference. If the candidate-generation mechanism excludes clinically relevant codes, no downstream nonconformity score or calibration procedure can recover the correct coding set. Uncertainty quantification therefore cannot be separated from the design and adequacy of the candidate space. These challenges indicate the need for a framework that restricts the candidate space sufficiently to remain computationally feasible while preserving enough clinically plausible alternatives for meaningful conformal evaluation.

1.2. The Proposed Framework

This study develops *sparse Mahalanobis conformal scoring* for set prediction in outpatient coding. The proposed framework combines a sparse extreme multi-label base model with conformal candidate-set scoring. First, the base model produces a sparse proposal set of likely labels for each encounter. Second, candidate coding sets are constructed from this restricted proposal space rather than from the full label space. Third, each candidate set is represented as a binary multi-label vector and scored against the base-model output using a diagonal Mahalanobis nonconformity function. This scoring rule allows label-wise residual variability to be incorporated into the candidate evaluation step. Finally, empirical conformal p -values are computed from held-out calibration data and used to form a conformal candidate region, from which practical decision rules can be derived. The proposed framework is not presented as a fully coverage-controlled point predictor. In this study, conformal scoring is used to evaluate the plausibility of candidate coding sets inside a computationally restricted search space generated by the base model. The practical value of the framework therefore depends on both calibration behavior and candidate-space

adequacy. The study evaluates not only the conformal candidate region itself, but also several downstream decision rules derived from it, including forced maximum p -value prediction (force), best- p selection within the retained region (best- p), and expected-F1 maximization within the retained region (Exp-F1). This design is more consistent with outpatient coding practice, where a system returns a compact and usable coding set rather than an abstract uncertainty object alone [6].

1.3. Contributions

The present study makes four main contributions.

1. It frames outpatient coding as an extreme multi-label set-prediction problem, where the output is a variable-sized code set rather than a ranked label list alone.
2. It develops a sparse Mahalanobis conformal scoring framework that assigns empirical conformal p -values to candidate coding sets generated from sparse model outputs.
3. It provides component-level ablations that separate the effects of proposal-support size, decision rule, and nonconformity score on coverage, region compactness, and point-prediction performance.
4. It evaluates several decision rules for converting conformal candidate regions into usable coding outputs and studies their behavior across label-frequency thresholds, code depth, sample cardinality, departments, and confidence–credibility strata.

2. Related Work

Previous research relevant to this study includes automated diagnosis coding, extreme multi-label settings, decision rules for constructing final multi-label outputs, and conformal prediction for multi-label problems.

2.1. Automated Diagnosis and ICD Coding

Automated diagnosis coding has been studied for many years, beginning with rule-based, dictionary-based, and feature-engineered systems, and later moving toward deep learning and pretrained language models [1,2,7–9]. In the broader medical-coding literature, most work has framed diagnosis code assignment as a large-scale multi-label task and has introduced neural architectures that learn from free-text clinical notes [10–15]. More recent studies have further improved code prediction by leveraging code hierarchies, label relations, and more efficient decoding strategies [16,17]. Recent Transformer-based work has further emphasized that automated ICD coding combines long-document modeling with an extreme multi-label output space [18]. However, many representative benchmarks and model developments have been centered on discharge summaries or inpatient-style documentation, whereas the outpatient setting has different documentation characteristics and workflow constraints [19,20]. Recent outpatient-specific work has shown that diagnosis coding in outpatient departments has distinct challenges related to document representation, case heterogeneity, and coder-assist workflows [20]. In outpatient practice, visits are usually shorter, more heterogeneous, and often assigned variable numbers of codes. Studies of ambulatory coding have also shown that clinicians may omit secondary diagnoses and that code selection can vary across users and EHR search interfaces, supporting the need for coding-assist methods that return concise and reviewable suggestions [21,22]. These characteristics make the task particularly challenging when prediction must remain both accurate and computationally feasible.

2.2. Extreme Multi-Label Classification (XMC)

Outpatient diagnosis coding exhibits the central characteristics of XMC. The output vocabulary can be very large, label frequencies are highly imbalanced, rare codes form

a substantial long tail, and clinically related diagnoses frequently co-occur [10,15,23,24]. Similar challenges have been reported in diagnosis coding studies that emphasize large code vocabularies, hierarchical code structure, label imbalance, and dependencies among co-occurring diagnosis codes [9,11,13,17]. Recent benchmark work using MIMIC-IV further supports the view that ICD coding is an extreme multi-label task requiring standardized preprocessing, large-scale evaluation, and reproducible comparison across models [24]. Moreover, the number of codes assigned to an encounter varies substantially from case to case. A practical coding model must therefore address at least three difficulties simultaneously: large output dimensionality, severe label imbalance, and uncertainty about the appropriate size of the final predicted code set. These issues are not fully resolved by simply improving the base classifier, because the final coding output still depends on how model scores are converted into a discrete set of predicted labels.

2.3. Decision Rules for Final Set Construction

In most multi-label coding pipelines, the base model produces either per-label scores or a ranked list of candidate labels, after which a decision rule is used to obtain the final prediction [9,24]. In practice, this often takes the form of ranked truncation, top- k prediction, or threshold-based selection [7,9]. Such rules are simple and efficient, but they impose strong assumptions on a problem in which case difficulty and code cardinality vary widely across visits [24,25]. A fixed top- k rule forces every encounter to receive the same number of predicted codes, whereas simple thresholding may behave inconsistently across frequent and infrequent labels. More importantly, these approaches usually return a single final prediction without explicitly quantifying how plausible alternative coding sets may be [26,27]. For outpatient coding, where several related code combinations may be clinically reasonable, this limitation becomes important.

2.4. Conformal Prediction for Multi-Label Outputs

CP offers a model-agnostic framework for transforming model outputs into uncertainty-aware prediction sets using a calibration dataset and a nonconformity score [3–5]. In recent years, conformal methods have been extended beyond standard single-label classification to multiclass and multi-label confidence sets, tree-based multi-label prediction, and Mahalanobis-based nonconformity measures for multi-label outputs [26,28,29]. Related work has also considered the practical usefulness of multi-label conformal sets when false positives and prediction-set size are important [27]. These studies demonstrate the potential of conformal methods for multi-label uncertainty quantification, but they also reveal a central computational difficulty: when the prediction target is itself a set of labels, the candidate output space can become prohibitively large.

The present study addresses this difficulty by applying conformal scoring within a sparse candidate space generated by an extreme multi-label base model. This approach shifts attention from unconstrained enumeration of the full multi-label output space to the joint design of candidate generation, candidate scoring, and practical final-set selection.

3. Materials and Methods

3.1. Study Design and Data Partitioning

This was a single-center retrospective secondary data analysis using outpatient electronic health records from Songkhlanagarind Hospital, Faculty of Medicine, Prince of Songkla University, a tertiary-care teaching hospital, covering encounters from January 2018 to December 2025. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Human Research Ethics Committee of the Faculty of Medicine, Prince of Songkla University (Approval no. REC.68-562-38-2). All data were

de-identified at the time of retrieval. Each observation corresponded to one outpatient encounter, and the prediction target was the set of ICD-10 diagnosis codes assigned to that encounter. Because a single visit could be associated with more than one code, the task was formulated as multi-label prediction under an extreme multi-label setting rather than ordinary single-label classification. Encounters were eligible when a physician note and at least one assigned diagnosis code were available, and encounters with no recorded note or administrative-only record were excluded.

To reflect the intended deployment setting, the dataset was partitioned chronologically into training, calibration, and test sets. The training set was used to fit the base extreme multi-label model and estimate the residual-variance terms. The held-out calibration set was used to form the empirical reference distribution of nonconformity scores and calculate conformal p -values. The test set was used only for final evaluation and was not used for model fitting or conformal calibration. Under the full label space, the partition contained 6,346,718 training, 819,354 calibration, and 804,574 test visits (approximately 8.0 million in total). In addition to the main analysis, sensitivity analyses were conducted across different minimum label-frequency thresholds in the training data in order to examine robustness to long-tail label pruning.

3.2. Prediction Task and Data Representation

Figure 1 summarizes the overall workflow used in this study. Outpatient encounter data were first transformed into sparse structured-EHR and clinical-note representations, which were then passed to the base XMC model to produce sparse label scores. These scores defined the proposal label set used for candidate-set construction, conformal scoring, and final decision-rule selection.

3.2.1. Label Space and Notation

Let $\mathcal{Y} = \{1, \dots, d\}$ denote the coding label space, where d is the number of labels. For encounter i , let x_i denote the encounter input representation and let $Y_i \subseteq \mathcal{Y}$ denote the observed set of assigned codes. The corresponding binary multi-label vector is written as

$$\mathbf{y}_i \in \{0, 1\}^d, \quad (1)$$

where $y_{ij} = 1$ if label j is assigned to encounter i and $y_{ij} = 0$ otherwise.

The base model produces a sparse score vector

$$\mathbf{o}_i = f(x_i) \in [0, 1]^d, \quad (2)$$

where o_{ij} denotes the score assigned to label j for encounter i . Only a sparse subset of labels is returned with nonzero scores. This subset defines the proposal label set

$$S_i = \{j \in \mathcal{Y} : o_{ij} > 0\}. \quad (3)$$

The aim of the proposed method is not only to rank labels, but to evaluate plausible sets of labels for each encounter. Accordingly, the conformal stage operates on candidate coding sets generated from the sparse proposal set rather than directly on individual labels.

3.2.2. Encounter Representation

Each outpatient encounter was represented by the concatenation of two sparse text-derived feature spaces: a structured visit representation and an unstructured clinical note representation (see Figure 1). This followed the PECOS-based outpatient XMC pipeline used in our coding experiments [23].

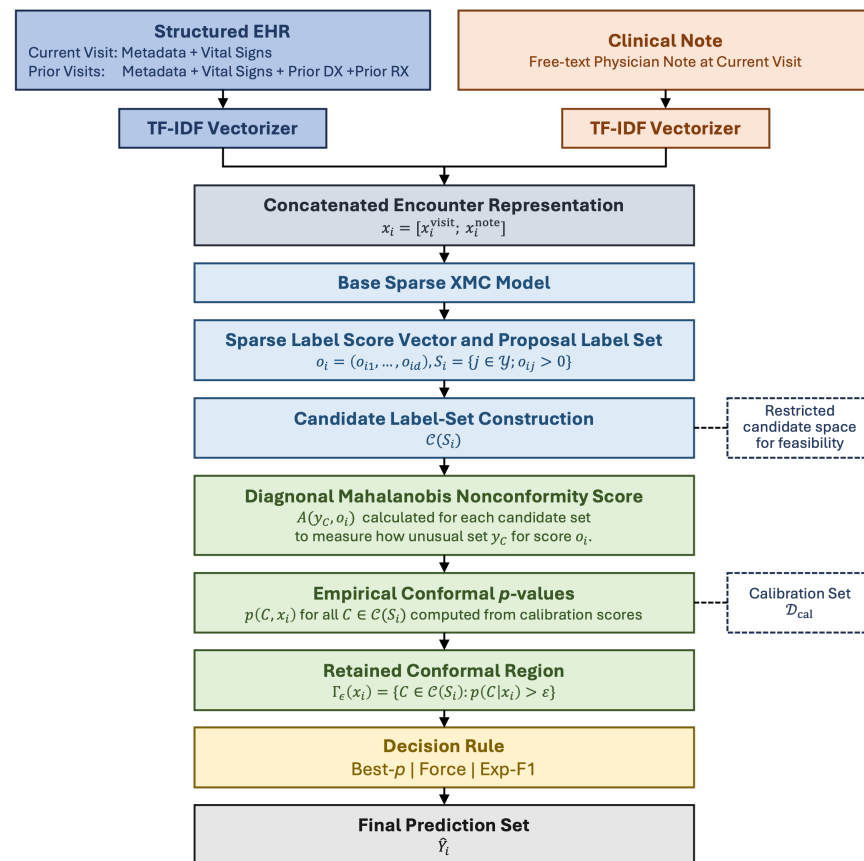


Figure 1. Overview of the proposed sparse Mahalanobis conformal scoring (SMaCS) framework for outpatient coding set prediction. Structured EHR data and the clinical note are transformed into sparse representations by independent TF-IDF models and passed to the base multi-label model to produce a label score vector. Candidate coding sets are then constructed from the sparse proposal space, evaluated using the nonconformity score, converted to empirical conformal p -values, and filtered into a retained conformal region. Final outputs are obtained by applying a decision rule, including forced maximum p -value prediction (force), best- p selection within the retained region (best- p), and expected-F1 maximization within the retained region (Exp-F1).

For the structured component, the current visit and recent visit history were serialized into token-based documents. The structured visit document (see Figure 2) included visit-level metadata and vital signs from the current encounter, while historical visits contributed prior diagnoses and prescriptions. Diagnoses and prescriptions from the current encounter were excluded from the structured representation in order to avoid label leakage. In the underlying XMC pipeline, the structured document was formed from the current visit together with the three most recent prior visits.

For the unstructured component, the physician-authored clinical note associated with the encounter was processed as a free-text document. Separate TF-IDF vectorizers were fitted to the structured visit documents and to the clinical note documents [30]. Each vectorizer used an n -gram range of 1–2 and was restricted to a maximum of 2^{18} features, with ℓ_2 normalization applied within each modality. The two modality-specific sparse vectors were then concatenated to form the encounter representation

$$x_i = [x_i^{\text{visit}}, x_i^{\text{note}}]. \quad (4)$$

This representation preserves the distinction between structured longitudinal context and the current clinical narrative while remaining suitable for large-scale sparse linear XMC modeling.

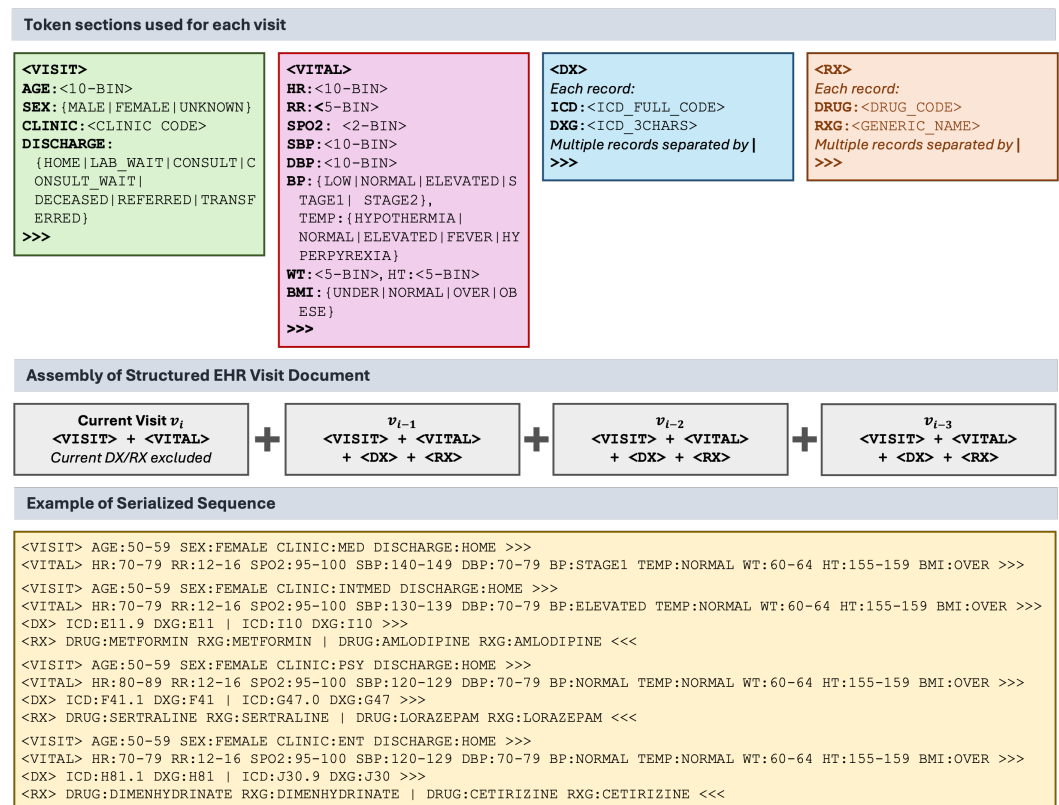


Figure 2. Structured EHR visit schema used to serialize outpatient EHR data into a uniform token sequence. Each visit is represented by four sections: <VISIT> for visit metadata, <VITAL> for vital signs, <DX> for diagnosis codes, and <RX> for prescriptions. Sections are terminated by the >>> token, and multiple diagnosis or prescription records are separated by the | token. For the current visit, <DX> and <RX> sections are excluded to avoid label leakage, whereas the three previous visits include <DX> and <RX> histories to provide longitudinal clinical context. The **lower panel** provides an example of the resulting serialized visit document.

3.3. PECOS-Based XMC Model

3.3.1. XR-Linear Training and Inference

In our study, the base prediction model was implemented using the PECOS framework for XMC, specifically the XR-Linear pipeline [23]. PECOS was selected because it is designed for large, sparse, and correlated output spaces and can produce sparse candidate labels efficiently, which is essential for the downstream conformal stage.

Model training followed the PECOS workflow. First, label embeddings were created from the training labels and training features. Second, the label tree was generated. Third, the recursive linear model was trained.

The base model served two roles in the overall framework. First, it generated the sparse proposal labels used to construct candidate coding sets. Second, it provided the score vector against which each candidate set was evaluated in the conformal stage. The conformal layer therefore operated on top of a trained PECOS XR-Linear model.

3.3.2. Sparse Proposal Labels

The conformal layer does not replace the need for a strong base model. If clinically relevant labels are not present in the sparse proposal set, they cannot be recovered later by conformal scoring. For this reason, the quality of the proposal mechanism is an essential part of the overall method, and candidate-space adequacy was examined explicitly in the evaluation.

The proposal label set for encounter i was defined as the set of labels with nonzero scores returned by the PECOS/XR-Linear inference pipeline:

$$S_i = \{j \in \mathcal{Y} : o_{ij} > 0\}. \quad (5)$$

For each encounter, nonzero labels were ranked by model score, and the 20 highest-scoring labels were retained as the proposal support. The conformal stage was then applied within this sparse proposal space.

3.4. Sparse Mahalanobis Conformal Scoring

3.4.1. Candidate-Space Construction

Direct conformal scoring over the full multi-label power set is infeasible in an extreme multi-label setting. Therefore, candidate coding sets were constructed only from the sparse proposal set S_i returned by the PECOS/XR-Linear inference pipeline. The candidate space for encounter i is written as

$$\mathcal{C}(S_i) = \{C \subseteq S_i : L_i \leq |C| \leq U_i\}, \quad (6)$$

where $|C|$ denotes the number of labels in the candidate set C .

Let $k_i = |S_i|$ denote the proposal-set size for encounter i . Candidate coding sets were generated over a restricted cardinality range centered around a target set size k_0 with window h :

$$L_i = \max(1, k_0 - h), \quad U_i = \min(k_i, k_0 + h). \quad (7)$$

In the main implementation, the candidate generator used $k_0 = 1$ and $h = 2$, determined from the training set. If the proposal set was empty, the highest-scoring label was used to define a singleton proposal set.

All label combinations with cardinality between L_i and U_i were enumerated when the total number of combinations was manageable. If the number of combinations exceeded a predefined cap M , candidate sets were sampled across cardinalities rather than fully enumerated. In the main implementation, the candidate cap was set to $M = 50,000$.

The restricted candidate-space design is central to the proposed method. It improves feasibility, but it also means that conformal scoring can only evaluate coding sets that are reachable through the sparse proposal mechanism. For that reason, candidate-space adequacy was treated as part of the method rather than only as a downstream diagnostic.

3.4.2. Diagonal Mahalanobis Nonconformity Score

The proposed method, termed *sparse Mahalanobis conformal scoring* (SMaCS), evaluates each candidate coding set using a diagonal Mahalanobis nonconformity score within a conformal framework [3,29]. Let $\mathbf{y}$ denote a candidate binary label-set vector and let $\mathbf{o}$ denote the corresponding base-model score vector for the same encounter. The nonconformity score is defined as

$$A(\mathbf{y}, \mathbf{o}) = \sqrt{(\mathbf{y} - \mathbf{o})^\top \mathbf{D}^{-1}(\mathbf{y} - \mathbf{o})}, \quad (8)$$

where $\mathbf{D}$ is a diagonal covariance matrix estimated from residuals between observed multi-label vectors and base-model outputs.

For label j , the residual is defined as $r_{ij} = y_{ij} - o_{ij}$. The diagonal variance term is estimated empirically as

$$\hat{\sigma}_j^2 = \mathbb{E}[(y_{ij} - o_{ij})^2] - (\mathbb{E}[y_{ij} - o_{ij}])^2. \quad (9)$$

To avoid unstable inverse weights, a ridge floor is imposed,

$$\hat{\sigma}_j^2 \leftarrow \max(\hat{\sigma}_j^2, \lambda_{\text{ridge}}), \quad (10)$$

and the inverse variance weight is defined as

$$w_j = \frac{1}{\hat{\sigma}_j^2}. \quad (11)$$

This yields a diagonal weighting matrix $\mathbf{D}^{-1} = \text{diag}(w_1, \dots, w_d)$. In the main implementation, the ridge term was set to $\lambda_{\text{ridge}} = 10^{-8}$.

3.4.3. Calibration Scores and Empirical Conformal p -Values

Let $\mathcal{D}_{\text{cal}} = \{(x_i, Y_i)\}_{i=1}^n$ denote the calibration set. For each calibration encounter, the observed multi-label vector $\mathbf{y}_i$ and the predicted score vector $\mathbf{o}_i$ are used to compute a calibration nonconformity score

$$a_i = A(\mathbf{y}_i, \mathbf{o}_i). \quad (12)$$

The calibration scores are sorted to form the empirical reference distribution used at inference.

For a test encounter with score vector $\mathbf{o}$ and a candidate coding set represented by binary vector $\mathbf{y}$, the candidate nonconformity score is computed as

$$a = A(\mathbf{y}, \mathbf{o}). \quad (13)$$

Its empirical conformal p -value is then defined as

$$p(\mathbf{y} \mid x) = \frac{1 + |\{i : a_i \geq a\}|}{n + 1}, \quad (14)$$

which follows the standard empirical conformal construction [3]. A candidate set is retained in the conformal region if

$$p(\mathbf{y} \mid x) > \epsilon, \quad (15)$$

where ϵ is the user-specified retention threshold. In the main implementation, empirical p -values were used without randomized smoothing.

For encounter i , the retained conformal region is written as

$$\Gamma_\epsilon(x_i) = \{C \in \mathcal{C}(S_i) : p(C \mid x_i) > \epsilon\}. \quad (16)$$

The conformal procedure is evaluated over the coding sets in the generated candidate space, $\mathcal{C}(S_i)$. The resulting region is a candidate-space conformal region whose overall coverage reflects both the conformal calibration within this space and the ability of the candidate-generation procedure to include the true coding set.

3.4.4. Decision Rules Derived from the Conformal Region

The retained conformal region may contain multiple candidate coding sets for the same encounter. Because outpatient coding systems must ultimately return a usable coding output, additional decision rules were applied to convert the retained conformal region into a final prediction. In this study, three decision rules were examined: force, best- p , and Exp-F1 selection:

- **Force:** The final prediction was always returned, regardless of whether it belonged to the retained conformal region. All candidate sets generated from the sparse proposal

space were scored, their empirical conformal p -values were computed, and the candidate with the largest p -value was selected. The force rule therefore preserved the candidate-space restriction but removed the region-retention requirement.

- **Best- p :** The final prediction was the retained candidate set with the largest empirical conformal p -value. This rule remained restricted to the conformal region and therefore selected among candidates satisfying the retention criterion $p(C \mid x) > \epsilon$. If no candidate satisfied the retention criterion, no final coding set was returned.
- **Exp-F1:** the final prediction was chosen by applying an Exp-F1 selection criterion over the retained conformal candidate sets. This rule was used to favor compact candidate sets that balanced precision and recall more effectively than the union rule while remaining more region-aware than simple best- p selection.

For each candidate set C , Exp-F1 was calculated as

$$\text{Exp-F1}(C) = \frac{2 \sum_{j \in C} \hat{p}_j}{|C| + \sum_{j=1}^L \hat{p}_j}, \quad (17)$$

where $\hat{p}_j$ is the predicted probability of label j , $|C|$ is the number of labels in the candidate set C , and L is the total number of labels.

The term $\sum_{j \in C} \hat{p}_j$ estimates the number of correctly selected labels in the candidate set C , while $\sum_{j=1}^L \hat{p}_j$ estimates the total number of positive labels. Exp-F1 therefore balances the expected number of correctly selected labels against the size of the candidate set. The retained candidate with the highest Exp-F1 value was selected as the final prediction. Exp-F1 is a downstream selection rule based on the predicted marginal label probabilities and is not part of the conformal calibration procedure.

If no candidate remained after conformal filtering, the method returned an empty prediction. Ties were resolved by first selecting the candidate with the higher conformal p -value, followed by the candidate with fewer labels. The operating significance level ϵ and the decision-rule settings were selected using the calibration set. These choices were fixed before evaluation on the test set.

In the main experiments, the Exp-F1 rule was treated as the primary SMaCS output for point-prediction analyses.

Unless otherwise stated, the main decision-rule analyses were reported at $\epsilon = 0.01$, with the full ϵ -sweep reported in the conformal-region analysis.

3.4.5. Alternative Nonconformity Scores

Three alternative nonconformity scores were compared with the Mahalanobis nonconformity score:

- **L_1 residual score:** The L_1 score measures the total absolute difference between the candidate label vector and the vector scores, with all labels receiving equal weight.

$$\alpha_{L_1}(\mathbf{y}, \mathbf{o}) = \sum_{j=1}^d |y_j - o_j|. \quad (18)$$

- **Euclidean L_2 score:** Unlike the Mahalanobis score, the L_2 score does not adjust the residuals according to label-specific residual variability.

$$\alpha_{L_2}(\mathbf{y}, \mathbf{o}) = \sqrt{\sum_{j=1}^d (y_j - o_j)^2}. \quad (19)$$

- **Rank-mismatch score:** This score used only the ordering of the vector scores. Let $\tilde{\mathbf{y}}^{(r)}$ be the binary vector representing the r highest-scoring labels in $\mathbf{o}$, where r is the number of labels in candidate $\mathbf{y}$. The score was defined as

$$\alpha_{\text{rank}}(\mathbf{y}, \mathbf{o}) = \sum_{j=1}^d |y_j - \tilde{y}_j^{(r)}|, \quad r = \sum_{j=1}^d y_j. \quad (20)$$

A score of zero means that the candidate exactly matches the r highest-scoring labels.

Each alternative score was calibrated separately and then applied within the same conformal procedure as the Mahalanobis score.

3.4.6. Confidence and Credibility Scores

Two additional quantities were recorded to characterize the strength of the candidate ranking: credibility and confidence. Credibility was defined as the largest empirical conformal p -value among all candidate sets generated for the encounter. Confidence was defined as one minus the second-largest empirical conformal p -value. If $p_{(1)}$ and $p_{(2)}$ denote the largest and second-largest candidate-set p -values, respectively, then

$$\text{credibility}(x) = p_{(1)} \quad \text{and} \quad \text{confidence}(x) = 1 - p_{(2)}. \quad (21)$$

These quantities were computed directly from the candidate-level conformal scores returned by the force rule and were later used for stratified performance analyses.

3.5. Evaluation Protocol

To place SmaCS in context, conventional point-prediction comparators were evaluated using the test set, including global max-F1 thresholding, local max-F1 thresholding, and top- k prediction with $k = 1, 2$, and 3. Component-level ablations examined proposal-support size, decision rules, and nonconformity scores. In the proposal-support ablation, label recall measured the proportion of true labels included in the proposal set, whereas proposal recall measured the proportion of encounters whose complete true label set was included.

3.5.1. Prediction Metrics

Prediction performance was assessed after applying each decision rule or comparator method to produce a final coding set for each encounter. The main point-prediction metrics were Micro-Precision, Micro-Recall, Micro-F1, Macro-F1, Weighted-F1, Weighted-Jaccard, exact match ratio, empty prediction rate, and average prediction size. These metrics were reported for the comparator methods and for the SmaCS decision rules. Empty prediction rate was included because some strategies could return no final label under stricter retention settings, whereas average prediction size summarized the compactness of the returned coding set.

3.5.2. Conformal Region Evaluation

The retained conformal region was also evaluated as a set-valued object in order to describe its size, inclusion behavior, and empirical retention properties across different values of ϵ . The conformal region was summarized using set-wise coverage, label-wise coverage, average region size, empty-region rate, average set size, and average label size across a sequence of ϵ values. This analysis was used to show how the retained region changed as the retention threshold became more or less stringent.

3.5.3. Robustness and Subgroup Analyses

Additional analyses were performed to examine how the method behaved under different label-frequency thresholds and across clinically relevant subgroups. Robustness to long-tail label pruning was assessed by repeating the main experiment across minimum training-set label-count thresholds of 0, 200, 400, 600, 800, and 1000.

Subgroup analyses were then conducted at several levels. First, performance was summarized by code depth, including chapter, block, and full-code levels. Second, labels were grouped by frequency bin in order to examine the relation between label prevalence and predictive performance. Third, encounters were stratified by the number of assigned labels to study the effect of sample cardinality. Fourth, performance was summarized across departments to examine heterogeneity by service setting. Finally, confidence- and credibility-stratified analyses were used to assess how these conformal ranking quantities were related to point-prediction quality.

4. Results

4.1. Dataset Characteristics

Table 1 summarizes the numbers of visits and labels in the training, calibration, and test sets. Under the full label space, the training, calibration, and test sets contained 6,346,718, 819,354, and 804,574 visits, respectively, with 12,829 retained labels. Increasing the minimum label-count threshold progressively reduced both the number of labels and the number of visits, yielding 1101 retained labels at the most restrictive setting. Across all settings, the average number of assigned labels per encounter remained stable at approximately 1.38–1.40. This indicates that label-frequency filtering changed the size of the retained coding space more than the typical per-visit coding cardinality.

Table 1. Dataset characteristics by experimental setting.

Minimal Label Count in Training	Number of Labels	Number of Visits			Number of Patients				Average Assigned Labels	
		Training	Calibration	Test	Training	Calibration		Test		
		01/18–07/24	07/24–03/25	03/25–12/25		Seen	Unseen	Seen		Unseen
0	12,829	6,346,718	819,354	804,574	604,899	177,235	34,607	165,845	49,452	1.39
200	2685	5,879,346	784,527	768,646	540,029	171,997	34,637	160,328	49,388	1.40
400	1883	5,683,852	759,640	744,639	530,712	168,934	34,085	157,500	48,425	1.39
600	1529	5,535,098	741,382	726,761	522,924	166,467	33,640	155,196	47,704	1.39
800	1270	5,384,932	721,392	703,055	515,512	163,868	33,148	152,039	46,111	1.39
1000	1101	5,256,556	704,069	686,609	508,972	161,463	32,820	149,871	45,400	1.38

4.2. Overall Point-Prediction Performance

Table 2 compares the point-prediction performance of the conventional baselines and SMaCS with Exp-F1. SMaCS Exp-F1 achieved a Micro-F1 of 0.6112 with a negligible empty rate of 0.0001, an exact match ratio of 0.4754, and an average prediction size of 1.2514. It also attained a Weighted-F1 of 0.5950 and a Weighted-Jaccard of 0.4469. Global Max-F1 attained a slightly higher Micro-F1 of 0.6229, but produced empty predictions in 19.3% of encounters and a lower exact match ratio of 0.4083, despite a comparable average prediction size of 1.1783. Local Max-F1 returned empty predictions in 15.4% of encounters. SMaCS Exp-F1 therefore provided a more favorable operating point when output availability and exact set recovery were considered alongside Micro-F1.

The threshold-based and top-*k* comparators showed different operating points. Local Max-F1 yielded slightly higher Micro-Recall than Global Max-F1 (0.6062 vs. 0.5907), but at the cost of lower Micro-Precision and Micro-F1. Compared with Top-1, SMaCS Exp-F1 maintained a similar exact match ratio while increasing Micro-Recall from 0.4883 to 0.5967.

Top-2 and Top-3 increased recall substantially, but precision and exact match declined sharply as the output size increased.

Table 2. Point-prediction performance of the conventional baselines and SMaCS decision rules.

Decision Rule	Micro-Precision	Micro-Recall	Micro-F1	Macro-F1	Weighted-F1	Weighted-Jaccard	Exact Match Ratio	Empty Rate	Average Prediction
Global Max-F1	0.6587	0.5907	0.6229	0.1483	0.6013	0.4550	0.4083	0.1930	1.1783
Local Max-F1	0.6000	0.6062	0.6031	0.1605	0.6126	0.4642	0.3823	0.1539	1.3275
Top-1	0.6416	0.4883	0.5545	0.1462	0.5326	0.3864	0.4766	0.0000	1.0000
Top-2	0.4348	0.6618	0.5248	0.1482	0.5353	0.3828	0.0570	0.0000	2.0000
Top-3	0.3257	0.7437	0.4531	0.1288	0.4920	0.3413	0.0217	0.0000	3.0000
SMaCS Exp-F1	0.6265	0.5967	0.6112	0.1453	0.5950	0.4469	0.4754	0.0001	1.2514

4.3. Performance as a Function of Prediction Size

Figure 3 compares Micro-F1, Micro-Precision, and Micro-Recall as functions of average prediction size. The figure shows a clear trade-off between compactness and recall. Top- k methods moved steadily toward higher recall as k increased, but this was accompanied by a marked decrease in precision and Micro-F1. In contrast, the threshold-based methods and SMaCS decision rules occupied a much more compact part of the operating space, with average prediction sizes close to one to one-and-a-half labels per encounter.

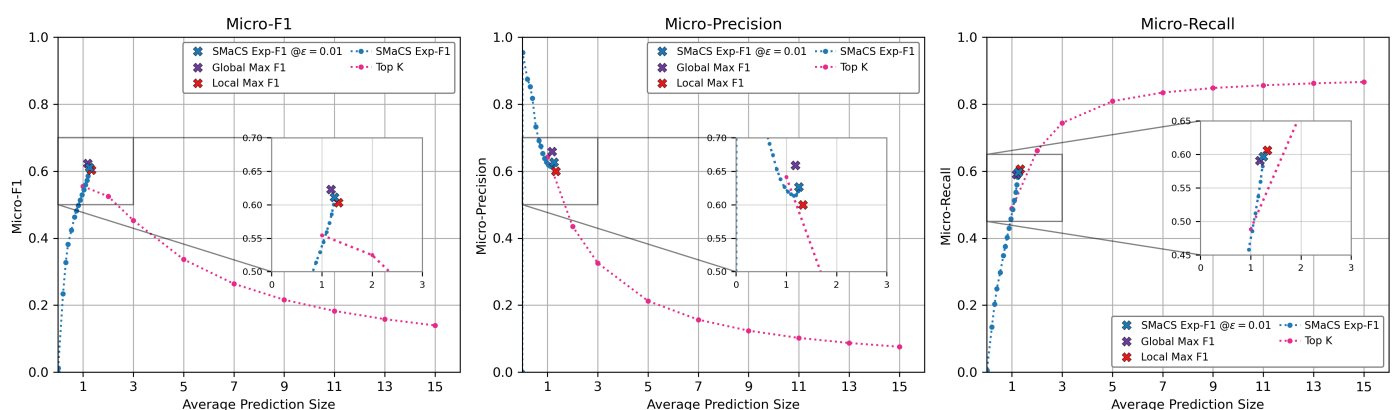


Figure 3. Comparison of Micro-F1, Micro-Precision, and Micro-Recall as functions of average prediction size for SMaCS Exp-F1 at $\epsilon = 0.01$, Global Max-F1, Local Max-F1, and Top- k prediction, illustrating the trade-off between prediction compactness and point-prediction performance across output sizes.

Within this compact region, SMaCS Exp-F1 provided the strongest overall balance. Its Micro-F1 was higher than that of SMaCS force and SMaCS best- p , while remaining close to the best threshold-based result. Global Max-F1 remained the strongest comparator in terms of Micro-F1, but SMaCS Exp-F1 provided a more favorable balance than Local Max-F1 and substantially improved upon the other conformal decision rules used as direct point predictors. These results support the use of Exp-F1 selection as the main decision rule for converting the retained conformal region into a practical coding output.

4.4. Conformal Region Validity

Table 3 and Figure 4 summarize the behavior of the retained conformal region across values of ϵ . Both set-wise coverage and label-wise coverage decreased monotonically as ϵ increased. At $\epsilon = 0.01$, set-wise coverage was 0.8236 and label-wise coverage was 0.8685. These values declined to 0.3324 and 0.3714 at $\epsilon = 0.50$, and to 0.1379 and 0.1349 at $\epsilon = 0.90$.

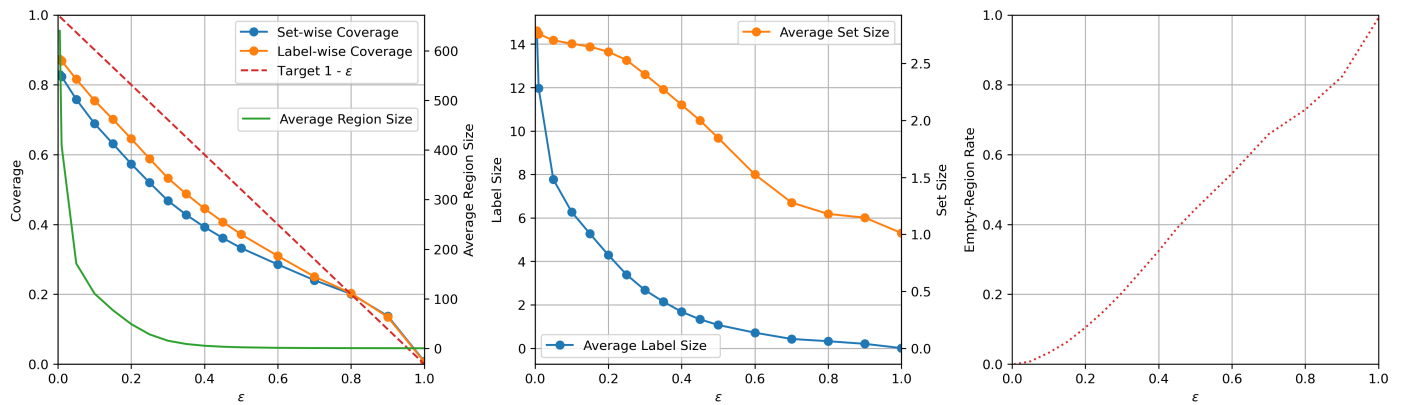


Figure 4. Behavior of the retained conformal region across different values of ϵ . The **left panel** shows set-wise coverage, label-wise coverage, the nominal target line ($1 - \epsilon$), and average region size. The **middle panel** shows average label size and average set size. The **right panel** shows the empty-region rate. Together, these panels summarize how the retained conformal region becomes smaller and more selective as ϵ increases.

Table 3. Empirical coverage and compactness of the retained conformal region.

ϵ	Set-Wise Coverage	Label-Wise Coverage	Average Region Size	Empty Rate	Average Set Size	Average Label Size
0.01	0.8236	0.8685	409.12	0.0001	2.76	11.97
0.05	0.7584	0.8161	171.04	0.0081	2.70	7.77
0.10	0.6885	0.7547	110.21	0.0326	2.67	6.28
0.15	0.6311	0.7012	76.77	0.0634	2.65	5.27
0.20	0.5735	0.6454	48.80	0.1047	2.60	4.29
0.25	0.5198	0.5885	28.04	0.1515	2.53	3.39
0.30	0.4682	0.5329	15.23	0.2045	2.41	2.67
0.35	0.4277	0.4874	8.66	0.2643	2.27	2.15
0.40	0.3926	0.4454	5.05	0.3243	2.14	1.68
0.50	0.3324	0.3714	1.96	0.4440	1.85	1.07
0.60	0.2855	0.3099	0.87	0.5456	1.53	0.72
0.70	0.2403	0.2506	0.42	0.6584	1.28	0.43
0.80	0.2011	0.2031	0.29	0.7289	1.18	0.33
0.90	0.1379	0.1349	0.18	0.8229	1.15	0.20

Average region size denotes the average number of retained candidate sets. Average set size denotes the average number of labels within each retained candidate set. Average label size denotes the average number of unique labels appearing across all retained candidate sets for an encounter.

This reduction in coverage was accompanied by a corresponding contraction of the retained region. Average region size decreased from 409.12 at $\epsilon = 0.01$ to 1.96 at $\epsilon = 0.50$ and 0.18 at $\epsilon = 0.90$. Average label size decreased from 11.97 at $\epsilon = 0.01$ to 1.07 at $\epsilon = 0.50$ and 0.20 at $\epsilon = 0.90$. At the same time, empty rate increased steadily from 0.0001 at $\epsilon = 0.01$ to 0.4440 at $\epsilon = 0.50$ and 0.8229 at $\epsilon = 0.90$. Figure 4 shows the same pattern graphically. As the retention threshold became more stringent, the conformal region became smaller and more selective, but also more likely to be empty.

The empirical region behavior did not follow the nominal ($1 - \epsilon$) line. This is consistent with the fact that retained candidates were evaluated over a restricted proposal space rather than over the full multi-label power set. In this setting, the conformal region should be interpreted as an empirical candidate-set filter within the sparse proposal mechanism, rather than as a full-space coverage guarantee.

In the present study, $\epsilon = 0.01$ was used as the main operating point for decision-rule evaluation because it preserved a broad retained candidate region with negligible empty-region rate, allowing Exp-F1 selection to operate over a sufficiently rich candidate set. It

should therefore be interpreted as a practical retention setting within the restricted proposal space, rather than as a nominal full-space coverage target.

4.5. Ablation Studies

Three ablations were performed to isolate the effects of proposal-support size, decision rule, and nonconformity scoring.

Table 4 shows the trade-off associated with proposal-support size. Increasing the support from 5 to 20 labels increased set-wise coverage from 0.7615 to 0.8236 and proposal recall from 0.7662 to 0.8303. At the same time, the number of generated candidate sets increased from 25 to 1350, and the average region size increased from 19.55 to 409.12. The 20-label support was retained for the main analysis because it provided the highest proposal recall and coverage while remaining within the available computational budget.

Table 4. Proposal-support ablation.

Proposal Support	Set-Wise Coverage	Label-Wise Coverage	Average Region Size	Average Label Size	Candidate Sets	Average Set Size	Label Recall	Proposal Recall
5 labels	0.7615	0.8050	19.55	4.40	25	2.15	0.8088	0.7662
10 labels	0.8021	0.8480	91.45	7.37	175	2.55	0.8527	0.8080
20 labels	0.8236	0.8685	409.12	11.97	1350	2.76	0.8739	0.8303

The decision-rule ablation in Table 5 held the proposal support at 20 labels and used the Mahalanobis nonconformity score. Force and best- p achieved Micro-F1 scores of 0.5453 and 0.5343, respectively. Exp-F1 increased Micro-F1 to 0.6112 and the exact match ratio to 0.4754, while increasing the average prediction size only from approximately 1.15 to 1.25 labels. Among the decision rules, Exp-F1 achieved the strongest point-prediction performance. This gain reflects the objective of Exp-F1, which uses the model probabilities to select the retained candidate with the highest estimated F1.

Table 5. Decision-rule ablation.

Decision Rule	Micro-F1	Macro-F1	Exact Match Ratio	Empty Rate	Average Prediction
Force	0.5453	0.1056	0.4279	0.0000	1.1536
Best- p	0.5343	0.1035	0.4158	0.0001	1.1505
Exp-F1	0.6112	0.1453	0.4754	0.0001	1.2514

With the proposal support fixed at 20 labels and Exp-F1, the nonconformity scores produced similar point-prediction performance (see Table 6). Mahalanobis produced an average region size of 409.12, compared with 439.91 for Euclidean L2, 559.06 for L1 residual, and 624.58 for rank mismatch. Its average union size was 11.97 labels, compared with approximately 20 labels for the other scores. This reduction was accompanied by a small percentage-point decrease in set-wise coverage. The main advantage of the Mahalanobis score was clear. It reduced the retained region to fewer candidate sets and fewer distinct labels, while maintaining similar Exp-F1 performance.

Overall, these ablations show that the main components contribute in different ways. Proposal-support size primarily determined proposal recall and coverage, Exp-F1 primarily improved point-prediction performance, and Mahalanobis scoring produced a more compact uncertainty region while maintaining similar predictive performance.

Table 6. Nonconformity-score ablation.

Nonconformity Score	Set-Wise Coverage	Label-Wise Coverage	Average Region Size	Average Label Size	Micro-F1	Exact Match Ratio	Average Prediction Size
Mahalanobis	0.8236	0.8685	409.12	11.97	0.6112	0.4754	1.2514
L_1 Residual	0.8283	0.8717	559.06	19.97	0.6120	0.4764	1.2512
Euclidean L_2	0.8284	0.8724	439.91	19.98	0.6120	0.4764	1.2514
Rank Mismatch	0.8299	0.8739	624.58	20.00	0.6122	0.4764	1.2520

4.6. Robustness to Label-Frequency Thresholding

Table 7 and Figure 5 show the effect of progressively restricting the retained label space. As the minimum training-set label count increased from 0 to 1000, Micro-F1 for the SMaCS Exp-F1 output increased from 0.6112 to 0.6601. Macro-F1 also increased substantially, from 0.1453 to 0.5760. Exact match ratio rose from 0.4754 to 0.5293. Average prediction size remained remarkably stable throughout, varying only between 1.2437 and 1.2514.

Table 7. Performance sensitivity to minimum label-count thresholds.

Metric	Minimal Label Count in Training Set					
	0	200	400	600	800	1000
Micro-F1	0.6112	0.6281	0.6370	0.6443	0.6537	0.6601
Macro-F1	0.1453	0.4597	0.5093	0.5350	0.5570	0.5760
Jaccard	0.4469	0.4663	0.4766	0.4844	0.4952	0.5018
Exact Match	0.4754	0.4943	0.5038	0.5126	0.5224	0.5293
Empty Rate	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
Average Prediction	1.2514	1.2463	1.2500	1.2437	1.2463	1.2473

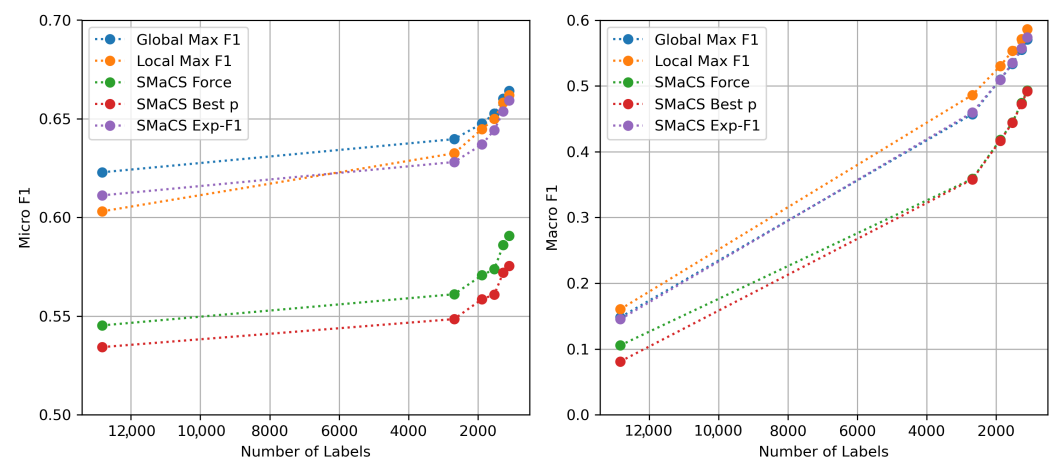


Figure 5. Performance under varying retained label-space sizes. The **left panel** shows Micro-F1 and the **right panel** shows Macro-F1 as functions of the number of retained labels after minimum-frequency filtering. These reflect how performance changes as the coding label space becomes more restricted.

These results indicate that long-tail label pruning improved point-prediction performance without materially changing the size of the returned coding set. Figure 5 shows that the gap between the different decision rules narrowed as the label space became smaller. SMaCS Exp-F1 remained competitive across the full range of retained label sets, while SMaCS force and best- p remained consistently below the leading threshold-based and Exp-F1 outputs. Jaccard also improved overall across the frequency thresholds.

4.7. Subgroup Analyses

Table 8 shows that performance depended strongly on code granularity. At the chapter level, Micro-F1 reached 0.7742 and label-wise coverage reached 0.9355. At the block level, these values declined to 0.6616 and 0.7001, respectively. At the full-code level, Micro-F1 was 0.6110 and label-wise coverage was 0.4893. This pattern indicates that the model performed better when labels were aggregated at coarser coding levels and became less accurate as diagnostic specificity increased.

Table 8. ICD-10 subgroup analysis.

Code Depth	Micro-Precision	Micro-Recall	Micro-F1	Recovered Label Fraction
Chapter	0.7881	0.7607	0.7742	0.9355
Block	0.6807	0.6436	0.6616	0.7001
Full Code	0.6265	0.5963	0.6110	0.4893

Table 9 shows the expected gradient across label-prevalence strata. For labels with frequency below 100, Micro-Recall was only 0.0997 and Micro-F1 was 0.1703. Performance improved markedly with label frequency, reaching Micro-F1 0.6333 and label-wise coverage 0.9952 in the ≥ 1000 group. These results indicate that the remaining performance gap is concentrated in the rarest labels.

Table 9. Label-frequency subgroup analysis.

Frequency Bin	Micro-Precision	Micro-Recall	Micro-F1	Label-Wise Coverage
<100	0.5831	0.0997	0.1703	0.1946
100–499	0.5777	0.3085	0.4023	0.8721
500–999	0.5696	0.4187	0.4826	0.9732
≥ 1000	0.6305	0.6361	0.6333	0.9952

Table 10 shows that performance also varied with the number of assigned codes per encounter. Micro-F1 increased from 0.5894 in single-label encounters to 0.6897 in three-label encounters, then declined for four-label and five-or-more-label cases. Exact match ratio was highest in single-label encounters at 0.5387 and dropped to zero in the four-label and five-or-more groups. Average prediction size increased steadily with sample cardinality, from 1.1395 in single-label cases to 2.3777 in the five-or-more group. This indicates that increasingly complex coding sets remained difficult to recover exactly, even when Micro-F1 stayed moderate.

Table 10. Sample cardinality subgroup analysis.

Sample Group	Micro-F1	Macro-F1	Exact Match Ratio	Empty Rate	Average Prediction
1 label	0.5894	0.1408	0.5387	0.0001	1.1395
2 labels	0.6566	0.1063	0.2683	0.0001	1.4739
3 labels	0.6897	0.0704	0.2227	0.0002	1.9642
4 labels	0.6328	0.0425	0.0000	0.0002	2.1930
5+ labels	0.5543	0.0265	0.0000	0.0003	2.3777

Department-level heterogeneity is summarized in Table 11. The strongest Micro-F1 values were observed in Radiology (0.7386), Dentistry (0.7138), Psychiatry (0.6961), and Family and Preventive Medicine (0.6592). Lower values were observed in Emergency Medicine (0.3656) and Pathology (0.2506). Exact match ratio was especially high in Dentistry (0.6819) and Psychiatry (0.5784). Overall, the department results suggest that performance

was better in settings with more standardized documentation patterns and narrower diagnostic ranges, and less stable in services with broader or more heterogeneous case mix.

Table 11. Department subgroup analysis.

Department	Micro-F1	Macro-F1	Exact Match Ratio	Empty Rate	Average Prediction
Radiology	0.7386	0.0233	0.4869	0.0000	1.5640
Dentistry	0.7138	0.0017	0.6819	0.0001	1.0824
Psychiatry	0.6961	0.0137	0.5784	0.0019	1.2054
Family and Preventive Medicine	0.6592	0.0276	0.5029	0.0000	1.3085
Internal Medicine	0.6348	0.0800	0.4433	0.0001	1.3788
Otolaryngology	0.6335	0.0221	0.5518	0.0001	1.1438
Surgery	0.6113	0.0531	0.5313	0.0001	1.1618
Ophthalmology	0.5876	0.0153	0.4578	0.0000	1.2094
Obstetrics and Gynecology	0.5849	0.0118	0.5131	0.0003	1.1161
Orthopedics	0.5595	0.0314	0.4906	0.0005	1.1242
Pediatrics	0.5463	0.0361	0.4214	0.0002	1.2205
Anesthesiology	0.5427	0.0224	0.4929	0.0000	1.0988
Rehabilitation Medicine	0.5161	0.0178	0.3263	0.0000	1.3454
Emergency Medicine	0.3656	0.0183	0.3243	0.0000	1.0998
Pathology	0.2506	0.0074	0.1721	0.0000	1.1367

4.8. Confidence and Credibility Stratification

Table 12 shows that credibility was strongly associated with downstream prediction quality. Across the credibility strata, Micro-F1 increased monotonically from 0.3659 in the lowest bin to 0.7597 in the highest bin. Jaccard increased from 0.3173 to 0.7466, and exact match ratio increased from 0.2489 to 0.6776. Average prediction size changed only modestly across these strata, from 1.2036 to 1.1194. This indicates that credibility carried meaningful information about the reliability of the final coding output.

Table 12. Confidence and credibility stratified performance.

Stratum	Micro-F1	Jaccard	Exact Match Ratio	Average Prediction
Credibility				
[0.00, 0.20]	0.3659	0.3173	0.2489	1.2036
(0.20, 0.40]	0.4098	0.3429	0.2660	1.1669
(0.40, 0.60]	0.4300	0.3674	0.2931	1.1538
(0.60, 0.80]	0.5298	0.4761	0.3988	1.1688
(0.80, 1.00]	0.7597	0.7466	0.6776	1.1194
Confidence				
[0.00, 0.20]	0.6115	0.5488	0.3845	1.3032
(0.20, 0.40]	0.5734	0.5120	0.4105	1.2201
(0.40, 0.60]	0.5144	0.4694	0.4011	1.1359
(0.60, 0.80]	0.5215	0.4856	0.4247	1.1229
(0.80, 1.00]	0.6149	0.6171	0.5807	1.0805

Confidence was less strongly ordered. Micro-F1 was 0.6115 in the lowest confidence stratum, decreased through the middle bins, and returned to 0.6149 in the highest confidence stratum. Jaccard and exact match ratio showed a similar non-monotone pattern, although the highest confidence bin still achieved the best overall exact match ratio at 0.5807. Compared with confidence, credibility provided a more interpretable ranking of prediction quality in this setting.

4.9. Performance Across Credibility Deciles

To further assess whether the credibility score provides useful information about prediction reliability, test visits were divided into ten equal-frequency groups, from the lowest-credibility decile (D1) to the highest-credibility decile (D10). Table 13 and Figure 6 show that Micro-F1, exact match ratio, Jaccard, and set-wise coverage generally increased with credibility. Exact match ratio rose from approximately 0.3586 in D1 to 0.7649 in D10, and Jaccard increased from approximately 0.4766 to 0.8145. Although the increase was not strictly monotonic, higher credibility corresponded to more reliable predictions and a greater probability that the true coding set was retained in the conformal region.

Table 13. Performance and set-wise coverage across credibility deciles.

Decile	Credibility Range	Micro-F1	Jaccard	Exact Match Ratio	Set-Wise Coverage
D1	0.00–0.34	0.5324	0.4766	0.3586	0.7522
D2	0.34–0.46	0.5395	0.4811	0.3663	0.7845
D3	0.46–0.55	0.5452	0.4852	0.3724	0.7958
D4	0.55–0.62	0.5443	0.4948	0.3955	0.8004
D5	0.62–0.67	0.5719	0.5273	0.4310	0.8207
D6	0.67–0.73	0.5954	0.5414	0.4365	0.8306
D7	0.73–0.78	0.5971	0.5494	0.4480	0.8306
D8	0.78–0.84	0.6591	0.6163	0.5226	0.8531
D9	0.84–0.91	0.7539	0.7398	0.6581	0.8915
D10	0.91–1.00	0.8176	0.8145	0.7649	0.9082

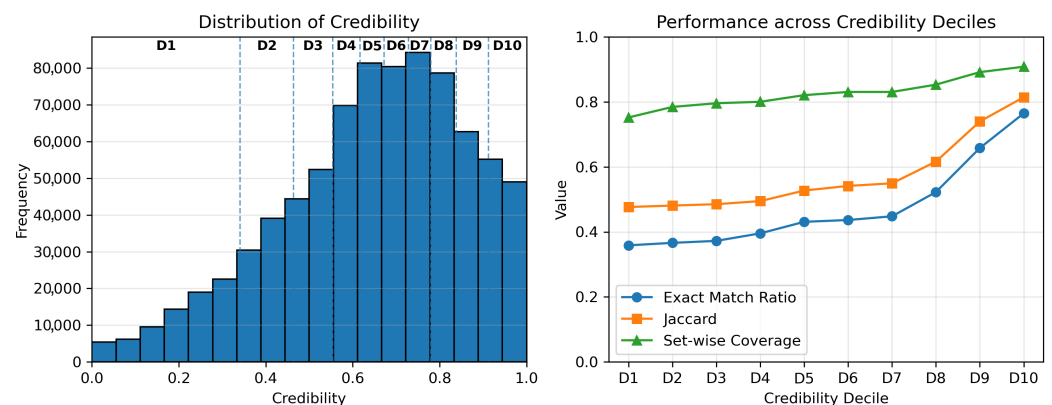


Figure 6. Distribution of credibility and predictive performance across credibility deciles. The **left panel** shows the distribution of credibility scores, with dashed vertical lines indicating the boundaries of ten equal-frequency credibility deciles. The **right panel** reports the exact match ratio, Jaccard score, and set-wise coverage within each credibility decile. Higher credibility is associated with improved performance and higher set-wise coverage.

The subgroup analysis in Figure 7 showed that this association varied across clinical departments. Exact match ratio generally improved across credibility deciles, although some departments showed flatter or non-monotonic patterns. Performance also differed across label-frequency strata. For visits whose true label occurred at least 1000 times in the training data, exact match ratio increased across credibility deciles. In contrast, visits containing labels with fewer than 100 training occurrences had consistently low exact match ratios and did not show the same positive gradient.

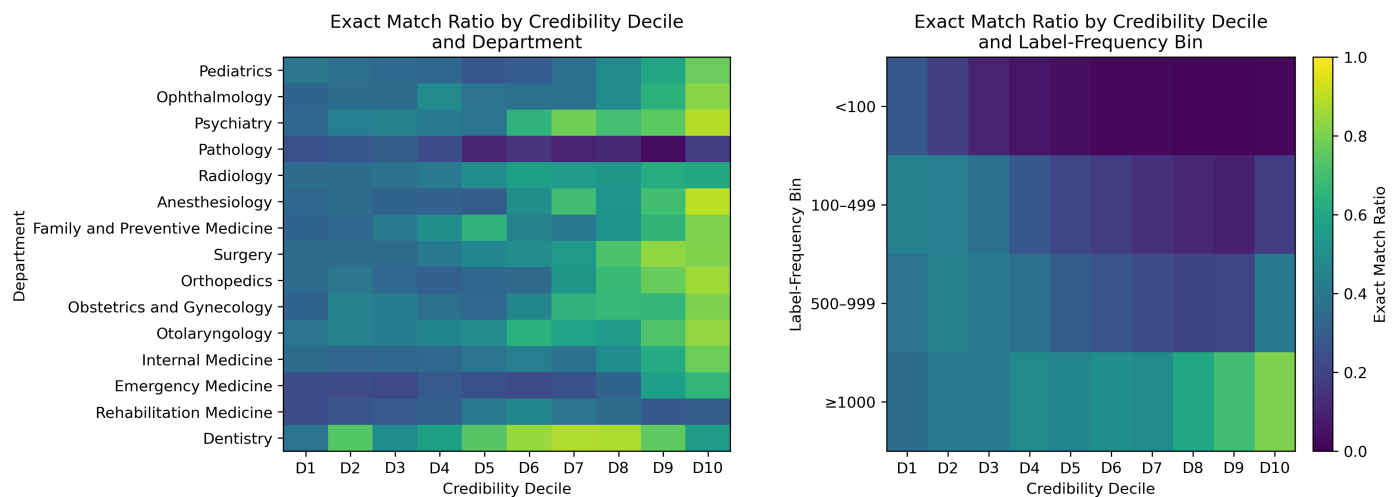


Figure 7. Exact match ratio across credibility deciles by department and label-frequency bin. The **left panel** shows the exact match ratio across credibility deciles within each clinical department. The **right panel** shows the exact match ratio across the label-frequency bins (<100, 100–499, 500–999, and ≥ 1000). Darker-to-lighter colors indicate lower-to-higher exact match ratios.

5. Discussion

This study evaluated outpatient coding as an extreme multi-label set-prediction problem and examined whether sparse Mahalanobis conformal scoring could improve the conversion of sparse model outputs into final coding sets. Three findings stand out. First, SMaCS Exp-F1 at $\epsilon = 0.01$ achieved competitive point-prediction performance, with a Micro-F1 of 0.6112, an exact match ratio of 0.4754, a negligible empty rate of 0.0001, and an average prediction size of 1.25. Although Global Max-F1 achieved a slightly higher Micro-F1 of 0.6229, it returned empty predictions in 19.3% of encounters and had a lower exact match ratio of 0.4083. Second, the retained conformal region behaved as a tunable filter: lower ϵ values retained broader candidate regions with higher empirical inclusion, whereas higher ϵ values produced smaller, more selective regions with higher empty rates. Third, the ablations clarified the roles of the main components. Proposal-support size controlled target-set reachability and computational cost, Exp-F1 provided the strongest point-prediction performance among the SMaCS decision rules, and Mahalanobis scoring produced a smaller retained region and label union than the alternative nonconformity scores while maintaining similar Exp-F1 performance.

5.1. Comparison to Previous Diagnosis Coding Studies

These findings fit the broader literature on automated medical coding, which has focused mainly on improving base prediction models through stronger encoders, label-aware architectures, and more efficient decoding strategies [1,2,8,9,11,13,15,16]. That work has established the importance of representation learning and large-scale multi-label modeling. The present results point to a related issue that has received less attention in outpatient coding: how sparse model scores should be converted into a final, variable-sized coding set. In this study, the same sparse XMC backbone produced substantially different outputs depending on whether the final coding set was selected by best- p , force, union, or Exp-F1. This indicates that post-score set construction is part of the predictive problem, not only a downstream implementation detail.

Compared with earlier diagnosis coding studies, the present study focuses on a different part of the coding pipeline. Prior work such as CAML, hierarchical label-wise attention, and pretrained language model approaches has mainly improved how clinical text is represented and how label-specific evidence is extracted from notes [11,13,15]. Other recent

studies have improved decoding efficiency or modeled relationships among diagnosis codes [16,17]. These studies are important because they improve the base prediction model. Additional model-centered ICD coding studies further show this direction: MultiResCNN improved clinical text representation using multi-filter and residual convolutional layers, HiLAT used hierarchical label-wise Transformer attention to improve explainability, and long-text Transformer models such as XR-LAT focused on handling long documents and extreme label spaces [12,14,18]. In contrast, the present study does not introduce a new text encoder or decoder. It starts from sparse XMC outputs and examines how candidate coding sets should be formed, scored, and reduced to a compact final variable-sized output. The contribution is therefore complementary to encoder- or decoder-focused coding models rather than a replacement for them.

The numerical performance of SmaCS Exp-F1 falls within the range reported by strong multi-label ICD coding models, although direct head-to-head comparison is limited by differences in datasets, coding systems, document types, label spaces, and evaluation protocols. On the MIMIC-III full-code discharge-summary benchmark, CAML reported Micro-F1 around 0.539 and MultiResCNN reported Micro-F1 around 0.552, while PLM-ICD later reported Micro-F1 around 0.598 using pretrained language models [11,12,15]. In the present outpatient setting, SmaCS Exp-F1 achieved Micro-F1 0.6112 with an average prediction size of 1.2514 and exact match ratio 0.4754. This comparison should not be interpreted as model superiority over previous studies because the data sources and label distributions differ. Instead, it shows that the proposed set-selection layer produced competitive performance at outpatient scale while addressing a different part of the coding pipeline: converting sparse candidate scores into compact variable-sized coding sets.

The study also differs from most inpatient or discharge-summary coding benchmarks. Those settings often involve longer documents and a different coding workflow, whereas outpatient coding uses shorter and more heterogeneous encounter records and may require rapid coder review across high visit volume [8,9,19,20]. Recent benchmark and workflow studies have emphasized that automated coding systems should be evaluated not only by aggregate predictive performance, but also by reliability, verification burden, and fit with coder workflows [6,24,31]. The present results support that view. SmaCS Exp-F1 did not simply maximize recall or return all plausible labels, it selected a compact coding set from the retained candidate region. This makes the method closer to a coding-assist setting, where the output must be short enough to review while still preserving useful prediction quality.

5.2. Decision-Rule Behavior and Compact Set Prediction

Regarding the comparison with threshold-based comparators, Global Max-F1 remained the strongest method in Micro-F1, which indicates that conventional thresholding remains a strong benchmark for final point prediction. At the same time, SmaCS Exp-F1 outperformed Local Max-F1 and clearly outperformed the other conformal decision rules. It also achieved the highest exact match ratio. This pattern suggests that the main contribution of SmaCS is not to replace thresholding on every metric, but to provide a more structured route from sparse label scores to a compact final coding set. That interpretation is consistent with recent work showing that multi-label prediction sets must be judged not only by inclusion, but also by false positives and set size [27]. In outpatient coding, where the output must remain short enough to be reviewed and used in routine care, compactness is part of model utility rather than a secondary preference [6].

5.3. Interpretation of the Conformal Region

The conformal results should be interpreted in line with the structure of the method. CP has been extended from standard single-label settings to multiclass and multi-label confidence sets, tree-based multi-label conformal methods, and Mahalanobis-based nonconformity scores [3–5,26,28,29]. The present study follows that line of work, but in a restricted candidate-space setting. Candidate coding sets were evaluated only within the sparse proposal pool returned by the XMC model. As a result, the retained conformal region should be interpreted as an empirical filter over plausible candidate sets inside that proposal space. The observed region behavior supports that interpretation. Here, ϵ determines the tolerated degree of atypicality relative to the calibration distribution. Candidate sets are retained when their nonconformity scores fall below the corresponding $1 - \epsilon$ empirical quantile, while empirical p -values and credibility quantify conformity continuously. The similar Exp-F1 performance despite large differences in region size suggests that many candidates retained by L_1 , Euclidean L_2 , and rank-mismatch scoring did not affect the final decision. The Mahalanobis score removed many of these unnecessary alternatives, leaving a smaller and more focused set of candidate codes for review without changing the final prediction, thereby concentrating the uncertainty representation. In extreme multi-label coding, this reduces both the number of candidate sets and the range of labels requiring further consideration without materially changing the downstream point prediction. As ϵ increased, set-wise coverage, label-wise coverage, candidate union size, and average region size all decreased, while the empty-region rate increased. This is consistent with a more selective retained region. At the same time, the empirical region behavior did not follow the nominal $(1 - \epsilon)$ line. In the present setting, that reflects the combination of conformal calibration with sparse proposal restriction rather than a full-space marginal coverage guarantee. The final coding quality therefore depended not only on the retained region, but also on how that region was summarized. Among the evaluated decision rules, Exp-F1 provided the strongest balance between inclusion and compactness.

5.4. Robustness, Subgroup Performance, and Reliability Signals

The robustness analyses reinforce the same interpretation. Performance improved steadily as the retained label space became smaller under higher minimum-frequency thresholds. Micro-F1 increased from 0.6112 to 0.6601, Macro-F1 increased from 0.1453 to 0.5760, and exact match ratio increased from 0.4754 to 0.5293, while average prediction size remained almost unchanged. This shows that long-tail label pruning changed the difficulty of the task more than it changed the size of the final coding set. The label-frequency subgroup analysis further showed that the remaining error was concentrated in rare labels. This pattern is consistent with prior work on extreme multi-label coding, where long-tail labels remain a major source of performance loss even when overall ranking quality is strong [10,15,23].

The subgroup analyses also show where the model worked well and where the task remained difficult. Performance was highest at the chapter level and declined at the block and full-code levels, which indicates that coding specificity remained an important source of difficulty. By sample cardinality, Micro-F1 improved from one-label to three-label encounters and then declined for more complex cases, while exact match ratio dropped sharply in the highest-cardinality groups. This suggests that increasingly complex coding sets remained difficult to recover exactly even when partial overlap was still captured. Department-level variation showed that performance was better in some specialties than others, with stronger results in Radiology, Dentistry, and Psychiatry and lower results in Emergency Medicine and Pathology. These patterns are consistent with differences in documentation regularity, coding diversity, and case mix across outpatient services [2,6,19].

The confidence and credibility analyses provide a useful practical result. Credibility showed a clear monotone relation with downstream quality: higher credibility strata had higher Micro-F1, Jaccard, and exact match ratio, with only modest changes in average prediction size. Confidence did not show the same ordered pattern. In this setting, the largest candidate-set conformal p -value carried more useful ranking information than one minus the second-largest p -value. This suggests that credibility can serve as an operational reliability signal for case prioritization or manual review. That may be more useful in practice than a large retained set or a nominal region interpretation alone.

Because the residual-variance matrix is diagonal, labels with small residual variance receive the largest inverse-variance weights and dominate the nonconformity score, whereas rare or noisy labels contribute little. The calibration distribution could also flag atypical encounters, supporting anomaly detection in outpatient coding-assist systems.

5.5. Limitations and Future Work

Several limitations should be noted. First, this was a single-center study. Coding practice, documentation style, and service mix may differ across institutions. Second, the conformal stage depended on candidate-space adequacy. If the sparse proposal pool omitted the correct code, no downstream scoring rule could recover it. In addition, the main candidate-generation setting restricted candidate-set cardinality to one to three labels. This improved computational feasibility but imposed a hard upper bound on exact set recovery for encounters with four or more assigned codes. Third, performance gains under stricter frequency thresholds partly reflect a smaller and less difficult retained label space. Fourth, the present study evaluated retrospective coding performance and did not include prospective workflow testing, coder interaction, or time-to-code outcomes. Those aspects are important for judging operational value [6]. Fifth, the diagonal covariance approximation improves computational stability but does not capture correlations between label residuals. Dependencies between clinically related codes may therefore be missed. Sparse or structured covariance models that include selected off-diagonal terms should be explored in future work. Finally, the present framework was built on sparse TF-IDF representations and a PECOS-based XR-Linear model. This design was computationally efficient and appropriate for large-scale XMC, but it does not determine whether the same conformal decision layer would behave similarly with stronger encoder-based backbones [15,16,31].

Future work should therefore focus on three areas. The first is candidate-space adequacy, including direct measurement of how often the true coding set remains reachable under sparse proposal generation and how that changes across specialties and label-frequency strata. The second is calibration design, including department-aware or service-aware calibration strategies for outpatient settings with heterogeneous documentation. The third is prospective evaluation in coding-assist workflows, especially for triage based on credibility and for compact code-set recommendation during routine care. In that setting, the most relevant question is likely not whether the model produces the largest retained set, but whether it returns a short and reliable coding set that reduces manual effort while preserving coding quality.

Overall, the results show that sparse Mahalanobis conformal scoring is best understood as a candidate-set scoring and selection layer within a sparse outpatient coding pipeline. Its main contribution is not a formal guarantee for the final point prediction, but a structured way to rank plausible coding sets and derive compact final outputs from them. Within that role, Exp-F1 selection gave the strongest overall balance between point-prediction quality and practical output size.

6. Conclusions

This study presented the SMaCS framework for set prediction in diagnosis coding under an extreme multi-label setting. The framework combined a sparse PECOS-based XMC model with candidate-set construction, conformal scoring, and decision-rule selection over the retained conformal region. In the main comparison, SMaCS Exp-F1 achieved competitive point-prediction performance and the highest exact match ratio while maintaining a compact average prediction size. The retained conformal region behaved as a tunable candidate filter. Mahalanobis scoring made this filter more useful by narrowing both the candidate-set space and the number of distinct labels presented for review, without sacrificing predictive performance. Lower values of ϵ preserved broader candidate regions with higher empirical inclusion, whereas larger values produced smaller and more selective regions. Among the evaluated decision rules, Exp-F1 provided the strongest overall balance between coding quality and output compactness. The subgroup and robustness analyses showed that performance depended on label frequency, code granularity, encounter cardinality, and department-level heterogeneity. Performance improved when the retained label space was restricted by minimum-frequency filtering, and credibility showed a clear relation with downstream prediction quality.

Overall, SMaCS is useful as a candidate-set scoring and selection layer within a sparse outpatient coding pipeline. Its main contribution is a structured way to convert sparse multi-label scores into compact final coding outputs while preserving information about candidate-set plausibility. Future work should focus on improving candidate-space adequacy and evaluating the framework in prospective coding-assist workflows.

Author Contributions: Conceptualization, K.T. and S.C. (Sitthichok Chaichulee); methodology, K.T. and S.C. (Sitthichok Chaichulee); software, K.T. and S.C. (Sawrawit Chairat); validation, K.T., S.C. (Sawrawit Chairat), and S.C. (Sitthichok Chaichulee); formal analysis, K.T.; investigation, K.T. and S.C. (Sawrawit Chairat); resources, S.C. (Sitthichok Chaichulee); data curation, K.T. and S.C. (Sawrawit Chairat); writing—original draft preparation, K.T.; writing—review and editing, K.T., S.C. (Sawrawit Chairat), and S.C. (Sitthichok Chaichulee); visualization, K.T. and S.C. (Sawrawit Chairat); supervision, S.C. (Sitthichok Chaichulee); project administration, S.C. (Sitthichok Chaichulee). All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Science, Research, and Innovation Fund (NSRF) and Prince of Songkla University (Grant No. MED6801219S). This research was supported by the Postdoctoral Fellowship Program from the Faculty of Medicine, Prince of Songkla University.

Institutional Review Board Statement: This secondary data analysis was conducted in accordance with the Declaration of Helsinki. The protocol was reviewed by the Human Research Ethics Committee of the Faculty of Medicine, Prince of Songkla University (Approval no. REC.68-562-38-2). All data were de-identified at the time of retrieval.

Informed Consent Statement: Informed consent for participation was not required because this study used deidentified secondary data, as approved by the Human Research Ethics Committee of the Faculty of Medicine, Prince of Songkla University.

Data Availability Statement: The data supporting the findings of this study are not publicly available due to institutional policy, ethical restrictions, and patient privacy considerations.

Acknowledgments: During the preparation of the manuscript, the authors used ChatGPT 5.4 for the purposes of proofreading, table formatting, and internal consistency review. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Stanfill, M.H.; Williams, M.; Fenton, S.H.; Jenders, R.A.; Hersh, W.R. A Systematic Literature Review of Automated Clinical Coding and Classification Systems. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 646–651. [\[CrossRef\]](#)
2. Dong, H.; Falis, M.; Whiteley, W.N.; Alex, B.; Matterson, J.; Ji, S.; Chen, J.; Wu, H. Automated Clinical Coding: What, Why, and Where We Are? *npj Digit. Med.* **2022**, *5*, 159. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Shafer, G.; Vovk, V. A tutorial on conformal prediction. *J. Mach. Learn. Res.* **2008**, *9*, 371–421.
4. Angelopoulos, A.N.; Bates, S. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [\[CrossRef\]](#)
5. Fontana, M.; Zeni, G.; Vantini, S. Conformal Prediction: A Unified Review of Theory and New Challenges. *Bernoulli* **2023**, *29*, 1–23. [\[CrossRef\]](#)
6. Gan, Y.; Rybinski, M.; Hachey, B.; Kummerfeld, J.K. Aligning AI Research with the Needs of Clinical Coding Workflows: Eight Recommendations Based on US Data Analysis and Critical Review. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vienna, Austria, 27 July–1 August 2025; pp. 909–922. [\[CrossRef\]](#)
7. Yan, C.; Fu, X.; Liu, X.; Zhang, Y.; Gao, Y.; Wu, J.; Li, Q. A Survey of Automated International Classification of Diseases Coding: Development, Challenges, and Applications. *Intell. Med.* **2022**, *2*, 161–173. [\[CrossRef\]](#)
8. Kaur, R.; Ginige, J.A.; Obst, O. AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Syst. Appl.* **2023**, *213*, 118997. [\[CrossRef\]](#)
9. Ji, S.; Li, X.; Sun, W.; Dong, H.; Taalas, A.; Zhang, Y.; Wu, H.; Pitkänen, E.; Marttinen, P. A Unified Review of Deep Learning for Automated Medical Coding. *ACM Comput. Surv.* **2024**, *56*, 306. [\[CrossRef\]](#)
10. Perotte, A.; Pivovarov, R.; Natarajan, K.; Weiskopf, N.; Wood, F.; Elhadad, N. Diagnosis Code Assignment: Models and Evaluation Metrics. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 231–237. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Mullenbach, J.; Wiegrefe, S.; Duke, J.; Sun, J.; Eisenstein, J. Explainable Prediction of Medical Codes from Clinical Text. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1 January 2018; pp. 1101–1111.
12. Li, F.; Yu, H. ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional Neural Network. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 8180–8187. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Dong, H.; Suárez-Paniagua, V.; Whiteley, W.; Wu, H. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *J. Biomed. Inform.* **2021**, *116*, 103728. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Liu, L.; Perez-Concha, O.; Nguyen, A.; Bennett, V.; Jorm, L. Hierarchical Label-wise Attention Transformer Model for Explainable ICD Coding. *J. Biomed. Inform.* **2022**, *133*, 104161. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Huang, C.W.; Tsai, S.C.; Chen, Y.N. PLM-ICD: Automatic ICD Coding with Pretrained Language Models. In Proceedings of the 4th Clinical Natural Language Processing Workshop, Seattle, WA, USA, 14 July 2022; pp. 10–20. [\[CrossRef\]](#)
16. Nguyen, T.T.; Schlegel, V.; Kashyap, A.R.; Winkler, S. A Two-Stage Decoder for Efficient ICD Coding. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, 9–14 July 2023; pp. 4658–4665. [\[CrossRef\]](#)
17. Luo, J.; Wang, X.; Wang, J.; Chang, A.; Wang, Y.; Ma, F. CoRelation: Boosting Automatic ICD Coding through Contextualized Code Relation Learning. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, 20–25 May 2024; pp. 3997–4007.
18. Liu, L.; Perez-Concha, O.; Nguyen, A.; Bennett, V.; Jorm, L. Automated ICD Coding Using Extreme Multi-label Long Text Transformer-based Models. *Artif. Intell. Med.* **2023**, *144*, 102662. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Masud, J.H.B.; Kuo, C.C.; Yeh, C.Y.; Yang, H.C.; Lin, M.C. Applying Deep Learning Model to Predict Diagnosis Code of Medical Records. *Diagnostics* **2023**, *13*, 2297. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Schlegel, V.; Kashyap, A.R.; Nguyen, T.T.; Yang, T.H.; Dwivedi, V.P.; Yin, W.H.; Wei, J.; Winkler, S. Automated Clinical Coding for Outpatient Departments. *arXiv* **2023**. [\[CrossRef\]](#)
21. Horsky, J.; Drucker, E.A.; Ramelson, H.Z. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. In Proceedings of the AMIA Annual Symposium Proceedings, Washington, DC, USA, 4–8 November 2017; Volume 2017, pp. 912–920.
22. Nguyen, A.N.; Truran, D.; Kemp, M.; Koopman, B.; Conlan, D.; O'Dwyer, J.; Zhang, M.; Karimi, S.; Hassanzadeh, H.; Lawley, M.J.; et al. Computer-Assisted Diagnostic Coding: Effectiveness of an NLP-based Approach Using SNOMED CT to ICD-10 Mappings. In Proceedings of the AMIA Annual Symposium Proceedings, San Francisco, CA, USA, 3–7 November 2018; Volume 2018, pp. 807–816.
23. Yu, H.F.; Zhong, K.; Zhang, J.; Chang, W.C.; Dhillon, I.S. PECOS: Prediction for Enormous and Correlated Output Spaces. *J. Mach. Learn. Res.* **2022**, *23*, 1–32. [\[CrossRef\]](#)
24. Nguyen, T.T.; Schlegel, V.; Kashyap, A.R.; Winkler, S.; Huang, S.S.; Liu, J.J.; Lin, C.J. MIMIC-IV-ICD: A New Benchmark for eXtreme MultiLabel Classification. *arXiv* **2023**. [\[CrossRef\]](#)

25. Babbar, R.; Schölkopf, B. Data Scarcity, Robustness and Extreme Multi-label Classification. *Mach. Learn.* **2019**, *108*, 1329–1351. [[CrossRef](#)]
26. Cauchois, M.; Gupta, S.; Duchi, J.C. Knowing What You Know: Valid and Validated Confidence Sets in Multiclass and Multilabel Prediction. *J. Mach. Learn. Res.* **2021**, *22*, 1–42.
27. Fisch, A.; Schuster, T.; Jaakkola, T.; Barzilay, R. Conformal Prediction Sets with Limited False Positives. In Proceedings of the 39th International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; Volume 162, pp. 6514–6532.
28. Tyagi, C.; Guo, W. Multi-label Classification under Uncertainty: A Tree-based Conformal Prediction Approach. In Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications, PMLR, Limassol, Cyprus, 13–15 September 2023; Volume 204, pp. 488–512.
29. Katsios, K.; Papadopoulos, H. Multi-label Conformal Prediction with a Mahalanobis Distance Nonconformity Measure. In Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications, PMLR, Milano, Italy, 9–11 September 2024; Volume 230, pp. 522–535.
30. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [[CrossRef](#)]
31. Yuan, M.; Shing, H.C.; Strong, M.; Shivade, C. Toward Reliable Clinical Coding with Language Models: Verification and Lightweight Adaptation. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Suzhou, China, 4–9 November 2025; pp. 173–184. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.