Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hybrid Rule-Based and Machine Learning Classification of Military Working Dog Medical Problems Using SNOMED CT and the Veterinary Extension

Marc Knobbe

marc.g.knobbe.mil@army.mil

Defense Health Agency

Ore Fakiyesi

United States Military Academy

Joshua Frey

United States Military Academy

Sara Mullaney

United States Military Academy

Article

Keywords:

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10757940/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hybrid Rule-Based and Machine Learning Classification of Military Working Dog Medical Problems Using SNOMED CT and the Veterinary Extension

Marc Knobbe[1,*], Ore Fakiyesi[2], Joshua Frey[2], and Sara Mullaney[2]

[1]Defense Health Agency, Public Health Directorate, Falls Church, Virginia, USA
[2]Department of Chemical, Biological Science & Engineering, United States Military Academy, West Point, NY, USA
[*]marc.g.knobbe.mil@army.mil

ABSTRACT

The Master Problem List in the Department of War's veterinary electronic health record combines structured SNOMED CT selections with unstructured free text, and the record system provides no automated surveillance capability. We developed an automated pipeline that maps Master Problem List entries recorded between January 1, 2018, and December 31, 2025, to standardized SNOMED CT concepts, combining multi-tier rule-based matching, SapBERT biomedical concept linking, and supervised classification of entries no rule resolved. For 73,856 records (13,653 unique terms), the pipeline assigned 1 of 19 Level 1 clinical categories to 93.7% of records, matching 80.0% through concept-based tiers and assigning the remainder by anatomical keyword. The SNOMED CT Veterinary Extension supplied 22.8% of matched concepts. For fully unmatched terms, term frequency-inverse document frequency features with logistic regression outperformed a neural network on biomedical embeddings (80.0% and 76.4% top-1 accuracy) and a few-shot large language model (58.6%), and was deployed to categorize them. A blinded panel of 9 veterinary specialists rating 384 stratified terms served as the reference standard, and pipeline agreement with expert consensus was 82.7% across the dataset. The framework closes a clinical-informatics capability gap and supports population-level health surveillance of military working dogs.

Introduction

Military Working Dogs (MWDs) serve vital roles in detection, patrol, and security capabilities across the U.S. Department of War (DoW), and their health directly affects operational readiness, effectiveness, and working longevity. Illness and injury are the primary causes of morbidity, lost duty days, and early retirement from service across working dog populations, with musculoskeletal, dermatologic, and gastrointestinal problems cited most often[1–6]. Noncombat-related health problems account for a substantial proportion of veterinary encounters and lost duty days in both deployed and garrison settings[4,6], and recent reports document significant burdens from vector-borne and other infectious diseases[7]. Longitudinal studies of working farm dogs found acute injury, lameness, and musculoskeletal disease to be the leading causes of morbidity and end of working career[8–10], and among young, non-deployed MWDs medical problems were associated with breed, age, and installation-level factors[11]. These studies identified the health problems that shorten working careers, but each relied on manual record extraction rather than automated analysis, which limits the scale and frequency of epidemiologic investigation in working dog populations.

In the DoW, MWD population-level surveillance of medical problems is limited by the structure of the available clinical data and medical record system. The DoW MWD enterprise lacks a systematic health monitoring capability that could give veterinary and kennel leadership timely feedback, identify emerging health threats, and guide interventions. The Master Problem List (MPL) field contains a mix of structured entries from a drop-down list of Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and Veterinary Extension dictionary terms and unstructured free-text describing diagnoses, clinical signs, procedures, and findings[12]. The Veterinary Services Systems Management (VSSM), the electronic veterinary health record, has no automated capability for sorting MPL entries by body system or problem type, so researchers must export the raw data and manually review thousands of entries. The DoW has since directed an "AI-first" posture, accelerating the adoption of artificial intelligence across all components in support of operational requirements[13]. Health services that sustain force readiness fall within that scope, as does military veterinary medicine, where no such capability currently reaches the working-dog medical record.

This medical record challenge is not unique to military veterinary medicine. The absence of standardized diagnostic coding in veterinary records is recognized as a barrier to using electronic health records for research, surveillance, and process

improvement pathways[14,15]. Most veterinary records contain free-text notations without a systematic diagnostic coding structure[14]. AI and machine learning are emerging tools that can be employed to close this gap.

SNOMED CT is the most comprehensive international clinical reference terminology and biomedical ontology, maintained by SNOMED International and used in both human and veterinary medicine[16]. It organizes over 529,000 active clinical concepts into a poly-hierarchy connected through "IS-A" relationships, and its concept definitions are formalized in description logic, which makes the hierarchy computable and supports automated linking between a concept and its ancestors or descendants[17,18]. SNOMED CT is complemented by a SNOMED Veterinary Extension of approximately 37,000 veterinary-specific concepts maintained by the Veterinary Terminology Services Laboratory[19], which encodes common veterinary problems lacking direct equivalents in the International Edition and shares the same "IS-A" hierarchy and identifier scheme.

In computer science, a neural network is a model built from layers of simple numeric units whose connection weights are fitted to example data, so the model learns to map an input to an output without pre-programmed rules. A transformer is a neural network architecture that processes a span of text as a whole, weighing every word against every other word, and represents each term as a vector of numbers called an embedding, positioned so terms appearing in similar contexts sit close together. Natural language processing (NLP) in medicine relies heavily on transformers, most prominently Bidirectional Encoder Representations from Transformers (BERT)[20]. Domain-specific transformer variants are built by pretraining the same architecture on medical text. SapBERT (self-alignment pretraining for BERT) extends PubMedBERT, a transformer pretrained on biomedical literature[21], by adding contrastive training on synonym pairs from the Unified Medical Language System (UMLS)[22]. That training organizes the embedding space specifically for biomedical entity linking[23]. Model selection therefore depends on the downstream task, whether text classification, code prediction, or concept-level entity linking.

Applied systems for veterinary records have been built on these architectures. VetTag pretrained a transformer language model on a large corpus of veterinary clinical notes to predict standardized diagnosis codes from narrative text[14]. PetBERT adapted BERT to companion animal clinical narratives for the same coding task[15]. VetLLM showed that a large language model could predict veterinary diagnosis codes with minimal task-specific training data[24].

Classifying a clinical term involves two choices, how the text is represented as numbers and which classifier is fitted on top of that representation. Representations range from term frequency-inverse document frequency (TF-IDF) vectors, which count the words and character sequences a term contains, to the contextual embeddings a transformer produces. Classifiers range from logistic regression, which weighs the features and separates categories with a straight decision boundary, to a multi-layer perceptron (MLP), the simplest form of neural network. An MLP passes the representation through an intermediate hidden layer of units before predicting a category, so it can capture nonlinear relationships that a linear model cannot.

The purpose of this study was to develop an automated classification system that maps structured and free-text MPL entries to standardized SNOMED CT concepts and assigns each entry to a clinical category. The pipeline routes each entry through a sequence of matching tiers ordered from the most precise method to the most permissive, so every entry is resolved by the strictest rule that succeeds.

The study had four aims. Aim 1 was to quantify the proportion of records and unique terms resolved at each tier. Aim 2 was to compare four methods for assigning clinical categories to the fully unmatched Tier 5 terms on a common held-out split: logistic regression on TF-IDF features, logistic regression on SapBERT embeddings, a multi-layer perceptron on SapBERT embeddings, and a general-purpose open-weight large language model applied few-shot. Aim 3 was to develop and evaluate a SapBERT-based method for linking Tier 4 terms to specific SNOMED CT concepts, accepting or referring candidate matches by cosine-similarity threshold. Aim 4 was to quantify the contribution of the SNOMED CT Veterinary Extension and the depth of SNOMED CT matching, including the proportion of entries resolving to leaf concepts. Across the rule-based and machine learning aims, a blinded, fully crossed expert panel of veterinary specialists served as the reference standard against which pipeline category assignments were scored.

Methods

Study design and data sources

This was a retrospective analysis of clinical records, designed as the development and validation of an hybrid classification system rather than as an epidemiologic study. All MPL entries recorded in the VSSM electronic health record system for U.S. government-owned animals (GOA) working dogs, between January 1, 2018 and December 31, 2025 were included. The resulting MPL dataset contained 73,856 records with 13,653 unique terms or phrases. Two SNOMED CT terminology releases were used as the reference standard: the SNOMED CT International Edition (February 2026 release)[16] and the SNOMED CT Veterinary Extension (September 2025 release)[19].

Classification framework

Figure 1 summarizes the classification pipeline. Two numbering schemes are used throughout. Matching tiers (Tiers 0 to 5) describe how an MPL entry was matched, from direct identification (Tier 0) to no match (Tier 5), and are defined in Table 1.

SNOMED CT Levels 1 to 4 describe the granularity of the assigned category, with Level 1 the broadest clinical category and Levels 2 to 4 progressively more specific subcategories. A classification framework was developed using 19 Level 1 categories anchored directly to the SNOMED CT hierarchy. The 19 categories comprise 13 body system categories corresponding to direct children of "Disorder of body system" (Integumentary, Musculoskeletal, Digestive, Genitourinary, Respiratory, Cardiovascular, Neurological, Visual System, Auditory, Endocrine, Hematopoietic, Lymphatic, and Immune); three cross-area categories (Infectious Disease, Behavioral, and Neoplastic Disease); and three non-clinical-finding categories (Procedure, Event, and Clinical Finding Not Otherwise Specified (NOS)). Because SNOMED CT uses a poly-hierarchical structure, a priority system was used to resolve conflicts when concepts span multiple categories: body system categories take highest priority, followed by cross-area categories, non-finding categories, and Clinical Finding NOS as the catch-all.

Aim 1: Rule-based classification of MPL entries to SNOMED CT

The SNOMED CT International Edition and Veterinary Extension were combined into a single searchable reference dictionary linking standardized concept names to unique concept identifiers and hierarchy positions. The 19 Level 1 categories were mapped to their positions in the SNOMED CT IS-A tree, and all descendant concepts were identified by traversing parent-child relationships in the combined International and Veterinary terminology. Before matching, each MPL entry underwent text preprocessing[25]. That included case normalization, tokenization, removal of common non-clinical stop words while retaining clinically meaningful negation terms, expansion of veterinary abbreviations (e.g., "OE" to "otitis externa"), lemmatization ("fractured" and "fractures" both map to "fracture"), and standardization of punctuation and whitespace.

After preprocessing, each unique MPL entry was compared against the SNOMED CT dictionary using a tiered matching strategy (Table 1). Tiers were ordered by match strictness. Earlier tiers required the entry text to match a SNOMED CT term with little or no alteration, while later tiers progressively relaxed strictness, ending with keyword-only category assignment. Each entry was assigned to the first tier that produced a match, so every entry received the strictest match available. Tier 0 handled MWD-specific terms with hardcoded mappings to exact phrases; these mappings were validated against the active SNOMED CT concept set. Direct mappings were added by the authors in two situations, (1) when the SNOMED reference release retired a previously-mapped concept and a structurally-equivalent active replacement existed, and (2) when an MPL term had unambiguous synonym-level equivalence to a single active SNOMED concept (for example, the veterinary surgical term "caudectomy" mapping to SNOMED "amputation of tail"). Direct mappings assigned a Level 1 clinical category, and additionally carried a specific SNOMED CT concept identifier only where an active concept existed for the term. Several of the highest-frequency working-dog entries, such as specific preventive injections and "gastropexy, right paracostal", received a category without a concept identifier, so most Tier 0 records carried a category alone.

Tier 1 attempted exact case-insensitive matching against SNOMED CT synonyms. Tiers 2a and 2b applied normalized and lemmatized matching, and Tier 2c rematched after removing clinical modifiers such as laterality, severity, and grade. Tier 3 applied fuzzy string matching (RapidFuzz token sort ratio ≥ 90). Tier 4 assigned a clinical category based on the presence of anatomical keywords even when no specific SNOMED CT concept could be identified. For each matched SNOMED CT term, the system traced the term's position in the SNOMED CT hierarchy to determine the Level 1 category, applying the poly-hierarchy priority rules described above.

Aim 2: Comparison of machine learning methods for unmatched terms

For clinical-category classification of fully unmatched (Tier 5) terms, four classification configurations were evaluated. All four were trained and evaluated on the same labeled set, the 3,677 terms that Tiers 0, 1, 2a, 2b, 2c, and 3 resolved to a Level 1 category. That Level 1 category is the label the classifiers learn to predict, derived by hierarchy traversal from the matched concept for the concept-resolving tiers and taken from the hand-selected mapping for Tier 0 terms that carried a category without a concept identifier. Terms resolved by modifier stripping (Tier 2c) carry laterality, severity, and acuity qualifiers, the same phrasing pattern that characterizes entries no tier could resolve; they were therefore retained in the labeled set. A single stratified 80/20 train/test split, spanning the Level 1 categories present in the matched data, was used for all four methods, and hyperparameters were fixed a priori at standard values rather than tuned, so no method gained an advantage from a search on the held-out split.

First, TF-IDF vectors were generated, treating each unique MPL item name as a "document." Both word-level features (single words and two-word combinations) and character-level features (sequences of 2 to 5 characters) were included so that misspellings and morphological variants of clinically informative tokens still contribute to the representation. A logistic regression classifier was trained on these vectors with class weights.

Second, a multi-layer perceptron (MLP) was implemented in PyTorch using SapBERT embeddings (768-dimensional dense vectors) of each MPL term as input features, fitted on the same labeled terms and evaluated on the same held-out split as the other classifiers. The MLP architecture consisted of one hidden layer with 128 units, ReLU activation, 30% dropout regularization, and a 19-class linear output layer, one unit per Level 1 category, of which 18 were populated in the matched data. The model was trained for 30 epochs using the Adam optimizer (learning rate 1×10^{-3}, weight decay 1×10^{-4}) with

class-weighted cross-entropy loss. A stratified 80/20 train-test split was used for evaluation. Third, a scikit-learn logistic regression classifier was trained on the same SapBERT embeddings as a linear baseline, using the same class weighting as the TF-IDF model, isolating the effect of feature representation.

Fourth, a general-purpose open-weight large language model (`gemma4:12b`, a 12-billion-parameter Gemma model served locally through Ollama) was evaluated in a few-shot setting without any task-specific training or fine-tuning. For each fully unmatched term, the model was prompted with the 19 Level 1 category definitions and up to 10 labeled examples per category drawn only from the training split, and instructed to return the predicted category. The examples came only from the training split and the test terms were held out, so the large language model was evaluated on the same 80/20 split as the supervised classifiers, but with less task-specific supervision (up to 190 in-context examples rather than the full labeled training set). The classifier applied to the Tier 5 terms was TF-IDF with logistic regression, selected on this held-out comparison before the expert panel was scored. The multi-layer perceptron and the SapBERT logistic regression were applied to the same terms as second opinions, and terms on which the three models disagreed were flagged for expert review.

Aim 3: Concept linking of keyword-matched terms

Concept linking is the task of mapping a free-text clinical phrase to a specific SNOMED CT concept identifier rather than to a broad clinical category. A Tier 4 keyword match assigned a Level 1 category when the entry contained a recognizable anatomical word (for example, "tooth" $\rightarrow$ Digestive) but the surrounding text matched no specific SNOMED CT term well enough for the rule-based tiers to identify a unique concept. The body system was known while the underlying concept was not. SapBERT[23] was used to identify candidate SNOMED CT concept matches based on semantic similarity for these keyword-matched terms. All active SNOMED CT synonyms from the combined International and Veterinary editions (approximately 916,000 terms) were encoded into 768-dimensional vectors using the `cambridgeltl/SapBERT-from-PubMedBERT-fulltext` model. Because each SNOMED CT concept carries a Fully Specified Name together with one or more synonymous terms, this synonym count substantially exceeds the roughly 566,000 combined concepts in the two releases (529,000 International and 37,000 Veterinary). Each keyword-matched MPL term was encoded using the same model, and the five nearest SNOMED CT synonyms were retrieved by cosine similarity. Terms with a top-1 cosine similarity score of 0.85 or above were considered high-confidence matches, scores between 0.70 and 0.85 were considered borderline and flagged for review, and scores below 0.70 were considered unlikely to have a valid SNOMED CT match. Both thresholds were set a priori rather than tuned on the data.

Aim 4: Veterinary Extension attribution and hierarchy depth

Source attribution of matched concepts (International Edition vs. Veterinary Extension) was performed by joining the matched SNOMED CT concept identifiers against the active concept tables in each release using the `moduleId` field of the Release Format 2 distribution files; concepts present in both releases were attributed to the International Edition as the source. For each matched concept, two metrics were recorded to characterize specificity: hierarchy depth, defined as the shortest IS-A path length from the matched concept to its Level 1 root, and leaf concept status, defined as a concept with no descendants in the SNOMED CT hierarchy.

Expert panel review and sampling

The reference standard was a blinded, fully crossed expert panel review. Reviewers were recruited through an open call to veterinary specialists within the U.S. Army Veterinary Corps. Nine specialists across three specialties, the College of Veterinary Preventive Medicine, the College of Veterinary Internal Medicine, and the College of Veterinary Emergency and Critical Care, completed the review. Each analyzed term carried exactly nine independent ratings. A custom web application (Streamlit, hosted on Posit Connect Cloud) standardized the review. For each term, the reviewer searched the combined International and Veterinary Extension dictionary and selected the single best-fitting SNOMED CT concept, blinded to the pipeline's assignment and to other reviewers' decisions. Level 1 through Level 4 categories were derived automatically from the selected concept, so a single judgment yielded both concept-level accuracy and category-level agreement. When no adequate concept existed, the reviewer assigned a Level 1 clinical category directly. Entries could also be marked administrative (non-clinical) or unclassifiable. For fields listing two or more distinct problems, the reviewer flagged the field and coded the first-listed problem in the main search and the second in an additional search field, and the pipeline was scored correct if it matched any reviewer-coded concept. Reviewers worked from a written Expert Reviewer Guide describing the classification pipeline, code-selection rules, and decision flowchart.

The panel reviewed a stratified random sample of 384 terms. The sample comprised 48 terms from each of the eight matching tiers (Tiers 0 through 5), drawn from the pool of 13,653 unique terms. Sampling was jointly stratified by tier and Level 1 clinical category. For Tier 5, the categories used for stratification were those predicted by the multi-layer perceptron, the classifier deployed when the sample was drawn. Equal allocation across tiers was chosen for estimation precision rather than pooled power. At 48 terms per tier, tier-level accuracy of 90% is estimated to within approximately $\pm 9\%$, and the pooled

sample estimates overall accuracy to within approximately ±3–5% (95% Wilson intervals)[26]. The fully crossed design supports multi-rater agreement estimation across the complete sample[27–29].

Because the review application's search index had been built from description-level activity flags, it included concepts retired from the reference SNOMED CT release, and 728 of 3,292 reviewer selected concepts carried retired identifiers. After the panel closed, selections were reconciled against the concept-level active set. 570 were remapped to active successors using SNOMED CT historical associations. The remaining 158 had no active successor and were assigned a Level 1 category by adjudication against the concept-category lookup, leaving the reviewer's concept identifier unchanged.

Statistical analysis

Coverage and clinical category distributions are reported as record counts and percentages of the total MPL dataset. Hierarchy depth is summarized as the mean and maximum across matched MPL items rather than unique concepts, because many item variants resolve to the same concept.

For the classifier comparison, top-1 accuracy is the proportion of test terms whose predicted Level 1 category matches the reference label, and top-5 accuracy the proportion whose reference label appears among the model's five highest-ranked categories. The reference labels are the rule-based, SNOMED CT-derived categories, so these figures quantify agreement with the rule-based standard rather than independent accuracy. Macro F1, macro precision, and macro recall are reported, each averaged across categories without weighting, because the 19 Level 1 categories are highly uneven in size. Each accuracy is reported with a 95% Wilson interval. All four methods predict the same held-out terms, so accuracies are correlated, and differences between models were tested with exact McNemar tests conditioned on the discordant terms[30]. The three supervised classifiers were evaluated by stratified five-fold cross-validation over the same labeled terms (shuffled, fixed seed), reporting the mean and the between-fold standard deviation. Claims about the distance separating the models depend on these cross-validated estimates, because a single split is sensitive to which terms fall into the held-out fifth[31]. The large language model was not cross-validated, as it was prompted rather than trained. The deployed Tier 5 classifier was selected by top-1 accuracy on the held-out split before the panel was scored against it, and the cross-validated ranking agreed.

The reference standard was the modal reviewer concept and clinical category per term. Terms without a category majority above 50% were adjudicated by a senior clinician blinded to the pipeline output, and scoring used the post-adjudication consensus. Accuracy was computed over sampled terms carrying both a panel consensus and a pipeline category. Some terms received no pipeline category because the matched concept reached no Level 1 root. These were excluded from the denominator rather than scored as errors, and the effect of the alternative is reported alongside the accuracy figures.

Accuracies are reported as proportions with 95% Wilson score confidence intervals. Overall accuracy weights each tier by that tier's share of MPL records, because the panel allocated terms equally across tiers rather than in proportion to tier size, and an unweighted mean would over-represent the smaller tiers. Per-tier accuracies are reported separately. Wilson intervals do not combine across strata, so the interval for the weighted figure came from a bootstrap resampling terms within each tier over 10,000 replicates, with the record-share weights and the seed held fixed. Inter-rater reliability at the Level 1 category was quantified with Fleiss' κ[32], generalized for a varying number of category votes per term. Confidence intervals came from bootstrap resampling of terms[33,34]. Exact concept-identifier agreement is reported descriptively.

Software and AI disclosure

All text preprocessing, SNOMED CT hierarchy traversal, and rule-based matching were implemented in Python 3.11.14 in Jupyter notebooks, using the pandas, NLTK, and rapidfuzz libraries. Machine learning classification used scikit-learn (TF-IDF, logistic regression, and nearest-neighbor search), PyTorch (multi-layer perceptron), and the HuggingFace transformers library (SapBERT embeddings). The large language model classification used `gemma4:12b` (Ollama model digest `4eb23ef187e2`) served locally through Ollama. Inter-rater reliability statistics were computed using statsmodels, significance tests and sparse feature assembly using SciPy, and bootstrap confidence intervals using NumPy. Figures were produced with matplotlib. The expert review applications were built in Streamlit.

A large language model (Claude Opus 4.7 and 4.8, Anthropic) was used as a code debug and development assistant during pipeline implementation, Streamlit application development, and analysis. All scientific content, study design, data analysis decisions, manuscript text, and final code were drafted, reviewed, and edited by authors, who accept full accountability for the work.

Results

Aim 1: Rule-based classification of MPL entries to SNOMED CT

The MPL dataset contained 73,856 entries representing 13,653 unique terms. The rule-based NLP pipeline assigned a Level 1 clinical category to 93.7% of records and 76.1% of unique terms (Table 2). Of the categorized records, 80.0% were matched through the concept tiers (Tiers 0–3), which resolve an entry against SNOMED CT concept text although not every match

carries a concept identifier, and 13.7% received a category by anatomical keyword match (Tier 4) without a specific concept ID (Table 1). The 6.1% of records that matched no tier were assigned to Tier 5 and referred to the machine learning classifiers. A residual 92 records (0.1%) matched a concept whose IS-A path reached no Level 1 root and received no category.

Table 1 reports the records, unique terms, and records per unique term resolved at each tier. Two tiers carried most of the volume: exact SNOMED CT synonym matching (Tier 1) classified 51.3% of all records, and direct identification of MWD-specific terms (Tier 0) a further 25.2%, the latter consisting primarily of high-frequency preventive entries such as heartworm prevention, prophylactic gastropexy, and dental prophylaxis. Record coverage and term coverage diverge sharply. Tiers 0 and 1 averaged 45.8 and 15.4 records per unique term, whereas Tier 4 averaged 1.5. This shows that most of the free-text vocabulary consists of low-frequency variants that no concept tier resolved. Fuzzy matching was the exception among the looser tiers at 19.3 records per term, indicating that spelling and word-order variants recur rather than being idiosyncratic.

Table 2 shows the distribution of categorized records by Level 1 category. The two largest categories were non-disorder categories rather than body-system diagnoses. Clinical Finding NOS captured nonspecific findings such as overweight and underweight status, lameness, and blood typing, and Procedure covered surgical interventions and medication administrations. Digestive, Integumentary, and Musculoskeletal were the highest-volume body system categories. Table 3 lists their most frequent SNOMED CT concepts, which captured clinically important working-dog problems including dental fracture, giardiasis, dermatitis, skin and footpad lacerations, degenerative joint disease, and cranial cruciate ligament tears. All remaining categories fell below 4% of records. Heat injury, a priority concern in working dogs, did not resolve to a single category. Of 898 records whose entry text referenced heat injury (excluding estrus), 65.4% resolved to Clinical Finding NOS concepts such as hyperthermia and heat injury and 26.3% to Neurological concepts such as heat exhaustion and heat stroke, while 67 matched no tier at all. No records resolved to the Immune category, a structural consequence of the poly-hierarchy priority rules rather than an absence of immune-mediated problems from the MPL.

Aim 2: Comparison of machine learning methods for unmatched terms

Four methods were compared on a common stratified 80/20 split of the 3,677 terms that the concept-resolving tiers (Tiers 0 through 3, including the modifier-stripped Tier 2c) matched to a specific SNOMED CT concept, which served as labeled training data (Table 4). TF-IDF with logistic regression ranked first on top-1 accuracy (80.0%, 95% CI 77.0–82.8), on macro F1, and under five-fold cross-validation. On the shared split, the two leading models disagreed on 135 terms, 81 of them resolved correctly by TF-IDF alone, a difference favoring TF-IDF (exact McNemar test, $p = 0.025$). Holding the feature representation constant, the multi-layer perceptron's nonlinear hidden layer improved single-split top-1 accuracy over the linear model by 4.3 percentage points. TF-IDF was the only model with comparable macro precision and recall; both SapBERT-based models paired high recall with low precision, recovering small categories by over-assigning them. The general-purpose large language model, prompted few-shot without task-specific training, trailed all three supervised classifiers. The features weighted most heavily by the TF-IDF logistic regression in the eight largest clinical categories are shown in Figure 2.

TF-IDF with logistic regression, the highest-scoring classifier in Table 4, was applied to the 3,262 Tier 5 terms (covering 4,535 MPL entries), the complete set of terms no rule-based tier resolved. Clinical Finding NOS dominated the predicted distribution (52.2%), followed by Digestive (13.8%) and Musculoskeletal (9.9%). The deployed model assigns the catch-all category more often than the expert reviewers did: Clinical Finding NOS was the panel's consensus for 31.2% of the sampled Tier 5 terms, compared with 52.2% of predictions. The multi-layer perceptron and the SapBERT logistic regression were applied to the same terms as independent second opinions. All three models agreed on 53.2% of terms (58.1% by record volume); the deployed prediction matched the multi-layer perceptron on 60.6% of terms and the SapBERT logistic regression on 57.4%. The remaining terms, on which the three models did not agree unanimously, were flagged as candidates for expert review.

Aim 3: Concept linking of keyword-matched terms

SapBERT concept linking was applied to the 6,661 unique terms (10,121 records) assigned a Level 1 category by keyword matching (Tier 4) but lacking a specific SNOMED CT concept identifier. Figure 3 shows the distribution of top-1 cosine similarity scores against the accept (0.85) and review (0.70) thresholds. High-confidence matches were recovered for 13.9% of terms, covering 26.1% of records; 39.3% of terms fell in the borderline band and were flagged as candidates for expert review, and 46.8% scored below the review threshold and were left unlinked. Coverage ran consistently higher by record than by term, so the terms that linked successfully were the more frequently recorded ones, the same pattern seen across the matching tiers.

Aim 4: Veterinary Extension attribution and hierarchy depth

The 3,427 unique MPL items whose matched concept traced to a Level 1 root mapped to 1,928 unique SNOMED CT concepts, since multiple item variants frequently resolved to the same concept. Their mean hierarchy depth from the matched concept to its Level 1 root was 3.84 levels (maximum 9 levels). 27.4% of matched MPL items (938 of 3,427) resolved to leaf concepts. A leaf is a concept with no descendants in the SNOMED CT hierarchy. These metrics indicate that the matching pipeline typically

resolved MPL items to moderately specific SNOMED CT concepts rather than to broad categorical parents. Tier 4 keyword matches are excluded from the hierarchy-depth analysis because they receive only a Level 1 clinical category and not a specific SNOMED CT concept identifier.

Source attribution of the matched SNOMED CT concepts revealed reliance on the Veterinary Extension. Of 1,928 unique SNOMED CT concepts assigned to MPL terms by the rule-based pipeline, 439 (22.8%) originated in the SNOMED CT Veterinary Extension and the remaining 1,489 (77.2%) in the International Edition. These Veterinary Extension concepts accounted for 9,353 records, or 21.7% of the 43,065 records carrying a concept identifier, a denominator that excludes Tier 4 keyword matches and the Tier 0 mappings that assign a category without one. The contribution was concentrated in the categories where dog-specific terminology dominates and was negligible elsewhere (Table 2), ranging from 43.7% of concept-linked Integumentary records to 0.5% of Auditory records. The highest-volume Veterinary Extension concepts represented common working-dog problems with no direct equivalent in the International Edition, including tail wound (n = 837 records), hindlimb lameness (n = 767), acute small bowel diarrhea (n = 700), pododermatitis (n = 197), and elbow hygroma (n = 186). These results indicate that the International Edition alone would not have provided sufficient concept-level coverage for MWD problem-list classification and that the Veterinary Extension is essential for the clinical categories most relevant to working-dog surveillance.

Expert panel validation

The expert panel reviewed 384 terms, 48 from each of the eight matching tiers, with all 9 reviewers rating every term. Inter-rater agreement at the Level 1 clinical category was substantial (Fleiss' $\kappa = 0.796$, 95% CI 0.770–0.822). Reviewers selected the identical SNOMED CT concept for 71.9% of terms. 13 terms lacked a category majority above 50% and were referred to a senior clinician; 2 remained without a majority after adjudication.

Pipeline accuracy against the panel consensus was 82.7% (78.7–86.6) across the dataset, weighting each tier's accuracy by that tier's share of MPL records. Accuracy was scored on the 369 terms carrying both a consensus and a pipeline category (42 to 48 per tier) and varied sharply by matching tier (Figure 4). The remaining 15 terms received no pipeline category because the matched concept reached no Level 1 root, though all carried a clean panel consensus; excluding rather than scoring them as errors raises unweighted accuracy from 76.8% to 79.9%. The direct-mapping tier (Tier 0) reached 64.6% (50.4–76.6), below every automated string-matching tier despite being pre-specified. The deployed classifier for fully unmatched terms reached 66.7% (52.5–78.3) on the Tier 5 sample, the direct estimate of its performance on the terms it was built to handle, and consistent with its selection on the held-out comparison in Table 4.

Discussion

This study developed a hybrid NLP and machine learning system that maps structured and free-text MWD medical problem entries to standardized clinical categories using the SNOMED CT concept terminology and poly-hierarchy structure. The high proportion of records resolved by exact matching (Tiers 0 and 1, collectively 76.6%) may reflect end-user reliance on the structured SNOMED dictionary drop-down within the MPL field. A small number of unique terms accounted for a large share of records: the 407 directly identified terms in Tier 0 alone covered nearly one quarter of all records, dominated by heartworm prophylaxis, prophylactic gastropexy, and standardized dental procedures. This concentration suggests that a dictionary-based approach anchored to SNOMED CT can capture the majority of clinical documentation in a working dog population without complex machine learning for the bulk of the data.

The distribution of medical problems across clinical categories was consistent with previous reports of morbidity in military and other working dog populations. Procedures, however, accounted for 23.2% of records but only 8.9% of unique terms, rising to 26.3% and 10.7% once preventive, screening, and status entries resolving elsewhere in the hierarchy are included. A small, heavily repeated vocabulary of procedural and preventive entries generates roughly a quarter of all MPL volume, while the disorder vocabulary is long-tailed and individually infrequent.

13.7% of records matched only by anatomical keyword (Tier 4), identifying a body system but not a specific clinical concept (for example, benign prostatic hypertrophy). The MPL field is therefore functioning as a catch-all for diagnoses, clinical findings, procedures, preventive care, and partial descriptions, with both the level of specificity and the type of content determined by individual user practice rather than by a documentation standard. Meaningful population-level surveillance from the MPL will require clinical practice guidelines that standardize how the field is used, particularly with respect to (1) when an entry should record a diagnosis versus a finding, (2) the minimum level of clinical specificity expected, (3) preferential use of the structured drop-down list of SNOMED dictionary terms over free-text entry where possible, and (4) whether procedures and preventive treatments belong in the MPL field at all, given that they are already captured elsewhere in the electronic record. Until such guidelines exist, automated classification systems like the one described here are necessary to extract surveillance signals from MPL data, but they cannot account for inconsistent inputs.

The hierarchy-depth output has practical value for syndromic, disease, and injury surveillance, supporting aggregation at Level 1 for cross-category comparison and descent to Level 3 or Level 4 for focused analysis, although no surveillance analysis was undertaken here. Because each record retains a dog identifier alongside its SNOMED CT concept and clinical category, the classified dataset also supports individual-level and cohort analyses, such as whether dogs with a recorded lameness problem also carry a cranial cruciate ligament rupture, a direction for future work.

The mean hierarchy depth of 3.84 levels for matched MPL items indicates that the rule-based pipeline routinely resolved entries to moderately specific SNOMED CT concepts rather than to broad parent categories. The 27.4% leaf-concept rate is a potential marker of end-user behavior: among entries the pipeline could resolve to a concept, more than one in four reached the most granular term available in the hierarchy, which is consistent with users supplying clinically specific terminology. Neither of the MPL field's two input options, the structured dictionary or free text, enforces a minimum level of clinical specificity, and use of the dictionary is not required in VSSM.

The Veterinary Extension proved essential rather than supplementary. It supplied 22.8% of the concepts the pipeline assigned and dominated the categories where dog-specific terminology is densest, including common working-dog problems such as tail wound, pododermatitis, and elbow hygroma that have no direct International Edition equivalent. A comparable system built on the International Edition alone would leave those entries at the category level.

TF-IDF features combined with logistic regression outperformed a PyTorch multi-layer perceptron on SapBERT embeddings, and both exceeded a logistic regression on those same embeddings (Table 4). TF-IDF led the multi-layer perceptron on the shared held-out split (exact McNemar test, $p = 0.025$), although five-fold cross-validation separated the two by a smaller margin. Holding the representation constant, the multi-layer perceptron outperformed the linear model by 4.3 percentage points, showing that a nonlinear hidden layer captures meaningful patterns in the embedding space, yet the simpler TF-IDF representation still exceeded the best embedding-based model. TF-IDF with logistic regression was therefore deployed as the Tier 5 classifier, selected on the held-out comparison rather than on the expert panel, which scored it afterwards at 66.7% on the Tier 5 sample. The classifier assigns Clinical Finding NOS to 52.2% of unmatched terms, whereas the panel's consensus placed only 31.2% of the Tier 5 sample in that category. Tier 5 output is therefore best treated as a screening layer that routes entries for review rather than a substitute for terminology curation.

A general-purpose large language model (`gemma4:12b`) trailed all three supervised methods (58.6% top-1); with only a handful of in-context examples and no task-specific training, it could not match classifiers fit to the full labeled set on these short entries, which favors the compact supervised models for the batch-surveillance use case. On these short problem-list entries, short word patterns proved highly informative, and how a term was represented as numbers mattered at least as much as which classifier was fitted on top of that representation. This is consistent with broader observations in biomedical NLP that specialized domain features can outperform general-purpose embeddings when input regularities align strongly with the prediction target.

The tiered matching strategy was designed to maximize precision at each step before attempting less certain and more complex methods. SapBERT concept linking subsequently proposed a specific concept for 13.9% of the keyword-matched terms at high confidence, indicating that biomedical language models can narrow, but not close, the gap between a broad category and a specific concept. A residual 92 records matched a SNOMED CT concept whose IS-A path reached no Level 1 root. Most arose from lemmatization over-reach, where a term such as "extracted tooth" resolved to the Substance concept "extract" rather than to a dental procedure, and the remainder from gaps in the Level 1 anchor set for concepts such as regurgitation and vomiting. These failures cost coverage rather than accuracy. Hierarchy traversal returns no category when the matched concept sits on an unrelated branch, so the error surfaces as an unclassified record rather than a confident misclassification.

The curated direct-identification tier (Tier 0) was the weakest tier against the expert panel (64.6%, 50.4–76.6), below every automated string-matching tier despite being specified by hand. Its disagreements were not matching failures but category judgments. The expert reviewers placed "missing tooth" entries in Clinical Finding NOS or Digestive, "scrotal dermatitis" in Integumentary or Genitourinary, and exertional hyperthermia in Neurological rather than Clinical Finding NOS. Each reflects a defensible but contestable choice. Curated mappings will need periodic revalidation against expert consensus, not only the programmatic validation against the active concept.

The framework described here differs from previous NLP systems for veterinary records in several ways. VetTag and PetBERT were trained on large corpora of companion animal clinical notes and used deep learning architectures to predict diagnostic codes directly from narrative text[14,15]. VetLLM demonstrated that large language models could achieve comparable performance with minimal fine-tuning[24]. The system outlined in this project classifies short, semi-structured problem list entries from a specific data entry field rather than full clinical narratives, and uses a hybrid approach combining rule-based SNOMED CT matching with machine learning to handle entries that resist direct terminology mapping.

This study has several limitations. The reported machine-learning accuracies were obtained on a held-out split against the rule-based labels rather than against independent expert-labeled test data, so they reflect how well the classifiers reproduce the pipeline's own categorizations rather than absolute accuracy. Because only terms resolving to a SNOMED CT concept (Tiers

0–3) carry labels, the classifiers were trained and evaluated on terms the rule-based pipeline could already handle, whereas the deployment target (Tier 5) consists of terms that failed every tier, with misspellings, dense abbreviations, compound phrasing, and novel wording. This covariate shift, similar to spectrum bias in diagnostic test evaluation, means the held-out figures should be read as an upper bound on Tier 5 performance; the expert panel provides the direct deployment estimate, which was accordingly lower. That estimate also rests on 48 Tier 5 terms, so its interval is wide, and the sample was drawn before the deployed classifier changed, stratified by the categories the multi-layer perceptron predicted rather than TF-IDF's own. Finally, the SapBERT concept links accepted at high confidence were not validated against expert judgment at scale: the panel sampled Tier 4 terms without regard to similarity score, so only 9 of the 48 sampled Tier 4 terms fell in the accepted band, too few to estimate their accuracy. The thresholds therefore govern which links enter the dataset without an estimate of how often those links are correct, and accepted concept identifiers should be treated as candidates for confirmation rather than validated assignments.

Anchoring the classification framework to the SNOMED CT hierarchy introduces constraints. The poly-hierarchical structure means some findings could be assigned to more than one category, and the priority rules resolving those assignments are pragmatic rather than clinically definitive. The empty Immune category illustrates the effect. Immune-mediated problems are present in the MPL (pannus, immune-mediated polyarthritis, lupus, and pemphigus), but under the body-system-first ordering each was claimed by the affected body system or by Clinical Finding NOS, so a surveillance query organized by etiology rather than anatomy cannot be answered from a single Level 1 category. Free-text wording imposes a related constraint. Heat injury, a priority concern in working dogs, resolved to no single category: 65.4% of the 898 heat-related records fell in Clinical Finding NOS and 26.3% in Neurological, and 67 matched no tier. One clinical entity can therefore map to several SNOMED CT concepts depending on how the entry was worded, so surveillance for such an entity may require a curated concept set rather than a single category filter.

A hybrid classification system combining rule-based NLP with machine learning mapped the majority of free-text MWD medical problem entries to standardized SNOMED CT concepts and clinical categories. The tiered strategy assigned a clinical category to 93.7% of records, 80.0% of them through terminology matching and the remainder by anatomical keyword, with machine learning extending coverage to the entries no rule resolved. The framework provides a foundation for population-level health surveillance across the MWD enterprise, replacing the manual review that has limited epidemiological use of MPL data, and its standardized hierarchical output suits trend monitoring, comparison across installations and time periods, and integration with other SNOMED CT-based health information systems.

References

1. Worth, A. *et al.* Causes of loss or retirement from active duty for new zealand police german shepherd dogs. *Animal Welf.* **22**, 167–174, DOI: 10.7120/09627286.22.2.167 (2013).
2. Caron-Lormier, G., Harvey, N. D., England, G. C. W. & Asher, L. A new metric for quantifying the relative impact of risk factors on loss of working life illustrated in a population of working dogs. *PLoS ONE* **11**, e0165414, DOI: 10.1371/journal.pone.0165414 (2016).
3. Caron-Lormier, G., England, G., Green, M. & Asher, L. Using the incidence and impact of health conditions in guide dogs to investigate healthy ageing in working dogs. *The Vet. J.* **207**, 124–130, DOI: 10.1016/j.tvjl.2015.10.046 (2016).
4. Takara, M. S. & Harrell, K. Noncombat-related injuries or illnesses incurred by military working dogs in a combat zone. *J. Am. Vet. Med. Assoc.* **245**, 1124–1128, DOI: 10.2460/javma.245.10.1124 (2014).
5. WORTH, A. J. & CAVE, N. J. A veterinary perspective on preventing injuries and other problems that shorten the life of working dogs. *Revue Sci. et Tech. de l'OIE* **37**, 161–169, DOI: 10.20506/rst.37.1.2749 (2018).
6. Harroun-White, H. E., Facciolla, C. A. & Knobbe, M. G. Noncombat injury and illness prevalence and working score percentage quantify the impact on duty availability in us army special operations military working dogs. *J. Am. Vet. Med. Assoc.* **262**, 1209–1214, DOI: 10.2460/javma.24.02.0106 (2024).
7. Meyers, A. C. *et al.* Epidemiology of vector-borne pathogens among u.s. government working dogs. *Vector-Borne Zoonotic Dis.* **21**, 358–368, DOI: 10.1089/vbz.2020.2725 (2021).
8. Isaksen, K. E. *et al.* Teammate: a longitudinal study of new zealand working farm dogs. i. methods, population characteristics and health on enrolment. *BMC Vet. Res.* **16**, 59, DOI: 10.1186/s12917-020-2273-2 (2020).
9. Isaksen, K. E. *et al.* Teammate: A longitudinal study of new zealand working farm dogs. ii. occurrence of musculoskeletal abnormalities. *Front. Vet. Sci.* **7**, 624, DOI: 10.3389/fvets.2020.00624 (2020).
10. Isaksen, K. E. *et al.* Teammate: A longitudinal study of new zealand working farm dogs. iii. factors affecting the risk of dogs being lost from the workforce. *Animals* **11**, 1602, DOI: 10.3390/ani11061602 (2021).

11. Schuh-Renner, A., Rappole, C., Mullaney, S. B., Venn, E. & Grier, T. L. Factors associated with medical problems among young non-deployed u.s. military working dogs. *Prev. Vet. Medicine* **193**, 105390, DOI: 10.1016/j.prevetmed.2021.105390 (2021).

12. Mullaney, S. B., Rozier, L., Moore, G. D. & Ajufoh, C. Factors associated with military working dog global electronic veterinary health record master problem list entries, 2018-2022. *The Med. J.* (2023).

13. U.S. Department of War. Artificial intelligence strategy for the Department of War: Accelerating America's military AI dominance. Memorandum for Senior Pentagon Leadership, Commanders of the Combatant Commands, and Defense Agency and DoW Field Activity Directors. Office of the Secretary of War, Washington, DC (2026). Accessed August 2026.

14. Zhang, Y., Nie, A., Zehnder, A., Page, R. L. & Zou, J. Vettag: improving automated veterinary diagnosis coding via large-scale language modeling. *npj Digit. Medicine* **2**, 35, DOI: 10.1038/s41746-019-0113-1 (2019).

15. Farrell, S., Appleton, C., Noble, P.-J. M. & Al Moubayed, N. Petbert: automated icd-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. *Sci. Reports* **13**, 18015, DOI: 10.1038/s41598-023-45155-7 (2023).

16. International, S. Snomed ct – systematized nomenclature of medicine – clinical terms. Accessed February 2026.

17. Baader, F., Brandt, S. & Lutz, C. Pushing the EL envelope. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, 364–369 (Morgan Kaufmann, Edinburgh, Scotland, 2005).

18. W3C OWL Working Group. OWL 2 web ontology language document overview (second edition). W3C Recommendation (2012). Accessed April 2026.

19. Veterinary Terminology Services Laboratory. Veterinary extension to SNOMED CT. Virginia Maryland Regional College of Veterinary Medicine, Blacksburg, VA (2025). Accessed January 2026.

20. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186, DOI: 10.18653/v1/N19-1423 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).

21. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Comput. for Healthc.* **3**, 1–23, DOI: 10.1145/3458754 (2021).

22. Bodenreider, O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270, DOI: 10.1093/nar/gkh061 (2004).

23. Liu, F., Shareghi, E., Meng, Z., Basaldella, M. & Collier, N. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4228–4238, DOI: 10.18653/v1/2021.naacl-main.334 (Association for Computational Linguistics, 2021).

24. Jiang, Y., Irvin, J. A., Ng, A. Y. & Zou, J. Vetllm: Large language model for predicting diagnosis from veterinary notes. *Pac. Symp. on Biocomput.* **29**, 120–133, DOI: 10.1142/9789811286421_0010 (2024).

25. Nesca, M., Katz, A., Leung, C. K. & Lix, L. M. A scoping review of preprocessing methods for unstructured text data to assess data quality. *Int. J. Popul. Data Sci.* **7**, 1757, DOI: 10.23889/ijpds.v7i1.1757 (2022).

26. Cochran, W. G. *Sampling Techniques* (John Wiley & Sons, New York, NY, 1977), 3rd edn.

27. Rothschild, A. S., Lehmann, H. P. & Hripcsak, G. Inter-rater agreement in physician-coded problem lists. In *AMIA Annual Symposium Proceedings*, 644–648 (2005).

28. Kim, T. Y., Hardiker, N. & Coenen, A. Inter-terminology mapping of nursing problems. *J. Biomed. Informatics* **49**, 213–220, DOI: 10.1016/j.jbi.2014.03.001 (2014).

29. Sim, J. & Wright, C. C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys. Ther.* **85**, 257–268, DOI: 10.1093/ptj/85.3.257 (2005).

30. Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **10**, 1895–1923, DOI: 10.1162/089976698300017197 (1998).

31. Molinaro, A. M., Simon, R. & Pfeiffer, R. M. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–3307, DOI: 10.1093/bioinformatics/bti499 (2005).

32. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378–382, DOI: 10.1037/h0031619 (1971).

33. Zapf, A., Castell, S., Morawietz, L. & Karch, A. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Med. Res. Methodol.* **16**, 93, DOI: 10.1186/s12874-016-0200-9 (2016).

34. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–74, DOI: 10.2307/2529310 (1977).

Acknowledgements

The authors gratefully acknowledge the institutional support of the United States Military Academy at West Point. Additionaly the authors want to thank the veterinarians from the U.S. Army Veterinary Corps that volunteered to serve as expert reviewers. This work received no external funding. The SNOMED CT International Edition is used under license from SNOMED International. The SNOMED CT Veterinary Extension is used under license from the Veterinary Terminology Services Laboratory (VTSL) at the Virginia-Maryland College of Veterinary Medicine, Virginia Polytechnic Institute and State University. Access to SNOMED CT and the Unified Medical Language System (UMLS) Metathesaurus, used by the SapBERT pretraining corpus, is provided through the U.S. National Library of Medicine.

Author contributions statement

M.K. and S.M. devised the project concept. M.K. designed the multi-tier classification methodology and led software implementation, validation, and formal analysis, with O.F., and J.F. contributing to methodology development, software implementation, and visualization of results. O.F., and J.F. contributed time resources and curated data. S.M. additionally contributed to data curation. M.K. wrote the original manuscript draft; O.F., J.F., and S.M. critically reviewed and edited the manuscript. S.M. provided supervision and project administration. All authors reviewed and approved the final manuscript.

Additional information

Code availability. The Jupyter notebooks containing the classification pipeline, the classifier comparison and Tier 5 prediction scripts, the SapBERT concept-linking code, the expert review application, and the scripts that generate every reported number and figure are available from the corresponding author on reasonable request and approval from the relevant DoW authorities.

Data availability. The MPL data analyzed in this study are derived from the U.S. Defense Health Agency Veterinary Services Systems Management (VSSM) electronic health record system. These data are not publicly available. De-identified aggregate statistics, the SNOMED CT classification framework definitions, and all analysis code are available from the corresponding author upon reasonable request and approval from the relevant DoW authorities. The SNOMED CT International Edition used as the reference terminology is available under license from SNOMED International (with U.S. access provided through the National Library of Medicine UMLS), and the SNOMED CT Veterinary Extension is available under license from the Veterinary Terminology Services Laboratory at the Virginia-Maryland College of Veterinary Medicine.

Competing interests. The authors declare no competing interests.

Ethical approval and study protocol. The United States Army Animal Resource and Review Committee approved this project proposal, and the Defense Health Agency approved the use of Veterinary Services Systems Management electronic veterinary record data. The Department of Homeland Security Office of Science and Technology approved the use of working dog health records included in the DoW system The study was not registered; no formal protocol was published, and the approved project proposal served as the planning document.

Tier	Method	Description	Unique terms	Records, n (%)	Records per term
T0	Direct identification	MWD-specific terms with known category assignments (e.g., ProHeart 12, gastropexy, COHAT)	407	18,644 (25.2%)	45.8
T1	Exact match	Case-insensitive exact match to a SNOMED CT term or synonym	2,468	37,908 (51.3%)	15.4
T2a	Normalized match	Match after text normalization (punctuation, whitespace, abbreviation expansion)	327	565 (0.8%)	1.7
T2b	Lemmatized match	Match after reducing words to base dictionary forms	107	232 (0.3%)	2.2
T2c	Modifier-stripped match	Match after removing clinical modifiers such as laterality, severity, and grade	360	674 (0.9%)	1.9
T3	Fuzzy match	Approximate string similarity above 90%, capturing misspellings and word-order variations	61	1,177 (1.6%)	19.3
T4	Keyword match	Category assignment based on anatomical keywords when no SNOMED CT term match was found	6,661	10,121 (13.7%)	1.5
T5	No match	Term could not be matched by any rule-based tier; referred to the machine learning classifier for clinical-category classification	3,262	4,535 (6.1%)	1.4
	Total	All MPL entries	13,653	73,856 (100%)	5.4

Table 1. Tiered matching strategy for mapping MPL entries to SNOMED CT concepts. Tiers are applied in sequence, and each entry is assigned to the first tier that produces a match. Percentages are of all 73,856 MPL records; the eight tiers partition the dataset completely. Tiers 1 through 3 resolve an entry to a specific SNOMED CT concept identifier and Tier 0 assigns a curated Level 1 category, carrying a concept identifier only where an active concept exists for the term; Tier 4 assigns a Level 1 clinical category from an anatomical keyword without identifying a concept, and Tier 5 entries carry no match and were passed to the machine learning classifier. Records per term is records divided by unique terms and shows how far each tier's vocabulary stretches: the direct and exact tiers resolve a small, high-frequency vocabulary, whereas Tiers 4 and 5 consist mostly of low-frequency free-text variants.

Category	Records, n (%)	Veterinary Extension, % of concept-linked records
Clinical Finding NOS	18,916 (25.6%)	31.2
Procedure	17,110 (23.2%)	0.6
Digestive	9,312 (12.6%)	4.1
Integumentary	7,583 (10.3%)	43.7
Musculoskeletal	4,188 (5.7%)	18.7
Auditory	2,515 (3.4%)	0.5
Genitourinary	2,340 (3.2%)	0.5
Behavioral	1,860 (2.5%)	40.8
Infectious Disease	1,667 (2.3%)	0.7
Visual System	1,033 (1.4%)	7.4
Respiratory	794 (1.1%)	33.0
Cardiovascular	663 (0.9%)	13.0
Neurological	465 (0.6%)	3.1
Neoplastic Disease	418 (0.6%)	7.9
Hematopoietic	227 (0.3%)	9.7
Endocrine	97 (0.1%)	1.1
Lymphatic	30 (<0.1%)	3.3
Event	11 (<0.1%)	—
Immune	0 (0%)	—
Total categorized	69,229 (93.7%)	—
No Level 1 root	92 (0.1%)	—
Unmatched (Tier 5)	4,535 (6.1%)	—
Total records	73,856 (100%)	—

Table 2. Distribution of classified MPL records across the 19 Level 1 clinical categories, ordered by record volume. Percentages are of all 73,856 MPL records. The 19 categories together account for the 69,229 records (93.7%) assigned a Level 1 category; the remainder comprises 4,535 fully unmatched Tier 5 records and 92 records that matched a concept whose hierarchy path reached no Level 1 root. The Veterinary Extension column gives the percentage of each category's concept-linked records whose SNOMED CT concept originated in the Veterinary Extension rather than the International Edition (Aim 4). Immune contains no records because the poly-hierarchy priority rules assign immune-mediated problems to the affected body system rather than to a cross-area category.

Body system	SNOMED CT concept	Records	% of category
Digestive (9,312 records)			
	Fracture of tooth	1,661	17.8
	Giardiasis	1,442	15.5
	Dental calculus	626	6.7
	Gastroenteritis	503	5.4
	Periodontal disease	390	4.2
	Inflammation of intestine	315	3.4
Integumentary (7,583 records)			
	Dermatitis	888	11.7
	Laceration of skin	516	6.8
	Hot spots (acute moist dermatitis)	283	3.7
	Abrasion of skin	272	3.6
	Superficial pyoderma	269	3.5
	Footpad laceration	243	3.2
Musculoskeletal (4,188 records)			
	Degenerative joint disease	513	12.2
	Spondylosis	242	5.8
	Spondylosis deformans of lumbosacral joint	162	3.9
	Cranial cruciate ligament tear	116	2.8
	Fracture	101	2.4
	Arthritis	97	2.3

Table 3. Six most frequent SNOMED CT concepts within the three highest-volume body system categories, aggregated at the concept-identifier level, and listed under their clinical name rather than the first-encountered synonym. The larger Clinical Finding NOS and Procedure categories are omitted. Percentages are of all categorized records in each category; counts include only records matched to a specific SNOMED CT concept (Tiers 0–3), so keyword-only (Tier 4) records contribute to denominators but not to counts. The six concepts shown account for roughly a third to a half of the records in each category.

		Single split		5-fold CV	Macro (single split)			
Features	Model	Top-1 (95% CI)	Top-5	Acc. (SD)	F1	Prec.	Rec.	Parameters
TF-IDF (word + char n-grams)	Logistic Regression	80.0% (77.0–82.8)	97.8%	79.7% (1.1)	0.72	0.72	0.76	∼14,000
SapBERT embeddings (768-dim)	Logistic Regression	72.1% (68.8–75.3)	98.6%	72.6% (0.9)	0.69	0.63	0.84	13,842
SapBERT embeddings (768-dim)	PyTorch MLP (128 hidden, dropout 0.3)	76.4% (73.2–79.3)	99.5%	77.5% (1.5)	0.69	0.63	0.81	100,754
Few-shot prompt (no training)	LLM (`gemma4:12b`)	58.6% (55.0–62.1)	—	—	0.53	0.48	0.72	∼12B

Table 4. Comparison of classification methods for Level 1 clinical-category classification of fully unmatched (Tier 5) MPL terms. All four methods were evaluated on the same stratified 80/20 train/test split of the concept-matched (Tiers 0–3) terms, so the top-1 accuracies are directly comparable and quantify agreement with the rule-based SNOMED CT-derived labels. Macro F1 is the unweighted mean of per-category F1 across the 18 Level 1 categories (Immune had 0 entries) represented in the labeled data, so every category contributes equally regardless of size; it is reported alongside accuracy because the category distribution is highly uneven and accuracy alone is dominated by the largest categories. Macro precision and recall are the unweighted means across the same categories, and a model with recall far above precision is assigning small categories more often than they occur. Five-fold cross-validation over the same terms is reported as mean accuracy with the standard deviation across folds. The language model was run few-shot with no fine-tuning, so it reports no top-5 or cross-validated accuracy, and its parameter count reflects pretrained model size.

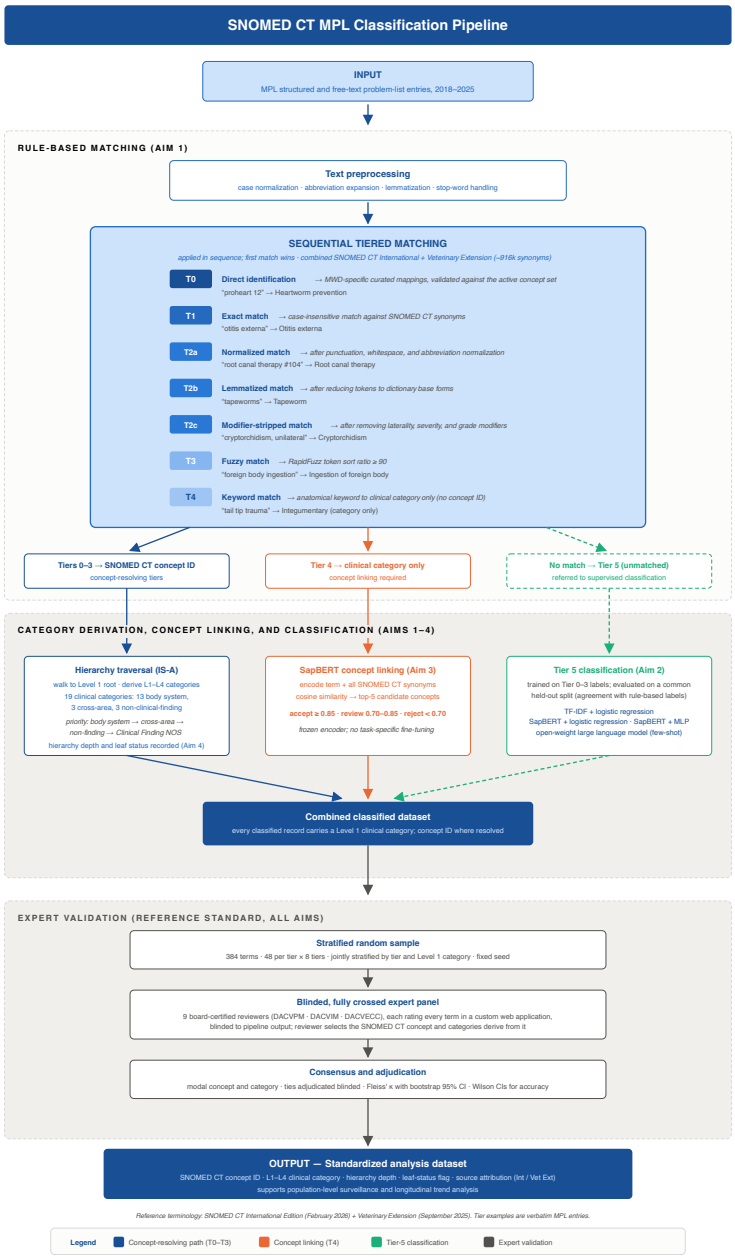


Figure 1. Overview of the SNOMED CT classification pipeline: rule-based sequential tiered matching (Tiers 0–4), hierarchy-based category derivation, SapBERT concept linking, supervised classification of unmatched (Tier 5) terms, and blinded expert-panel validation forming the reference standard.

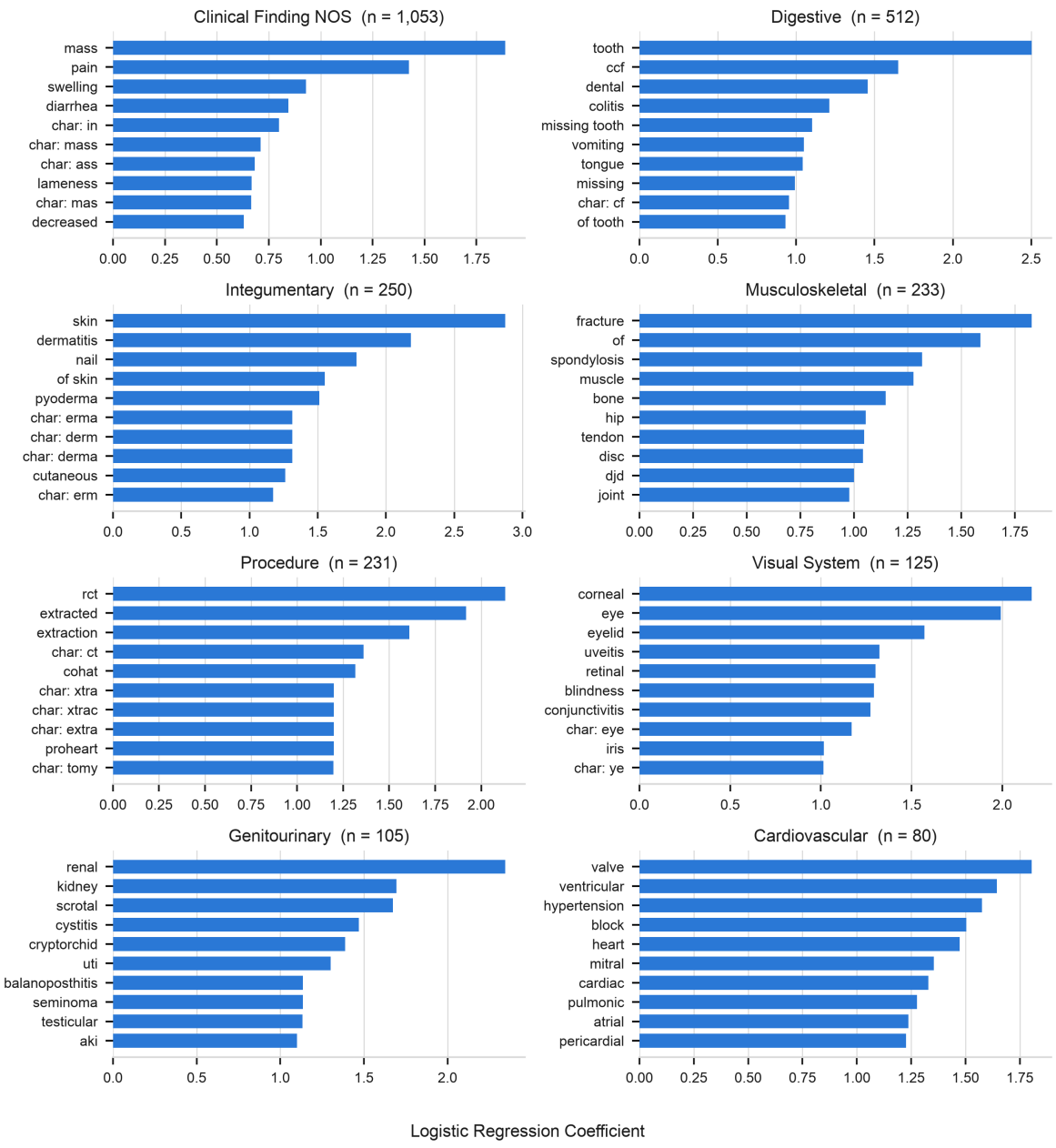


Figure 2. Highest-weighted TF-IDF features for the eight largest Level 1 categories, from the logistic regression reported in Table 4. The model was trained on the 3,677 concept-matched terms (Tiers 0–3), whose labels derive from SNOMED CT hierarchy traversal rather than from a token in the entry text, so a high weight reflects a learned association rather than a restatement of the anatomical-keyword rule used by Tier 4. Only the eight categories with the most labelled training terms are shown, with category size given in each panel heading. The weighted features are clinically interpretable and include working-dog shorthand (“rct” for root canal therapy, “cohat”, “proheart”), while the character n-grams capture morphological variants of informative tokens (“extracted”, “extraction”, “char: xtrac”).

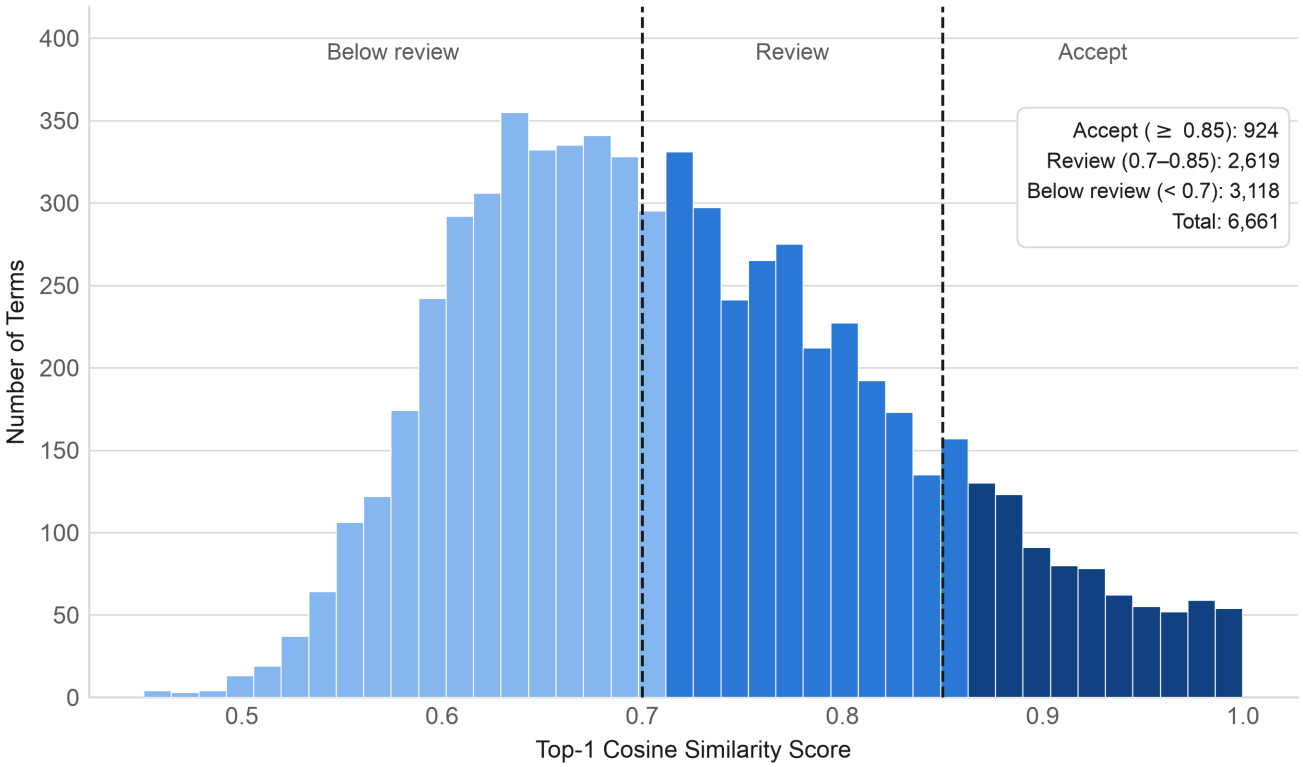


Figure 3. Distribution of top-1 SapBERT cosine similarity scores for the 6,661 Tier 4 keyword-matched terms, each scored against the full SNOMED CT vocabulary. Terms at or above the accept threshold (0.85) received the matched concept identifier; terms in the review band (0.70–0.85) were flagged as candidates for expert review; terms below 0.70 were left unlinked and retain a Level 1 category only. Band counts are given in the inset. The distribution is centred well below the accept threshold, with the mode near 0.65, so most keyword-matched entries could not be resolved to a specific concept at high confidence.

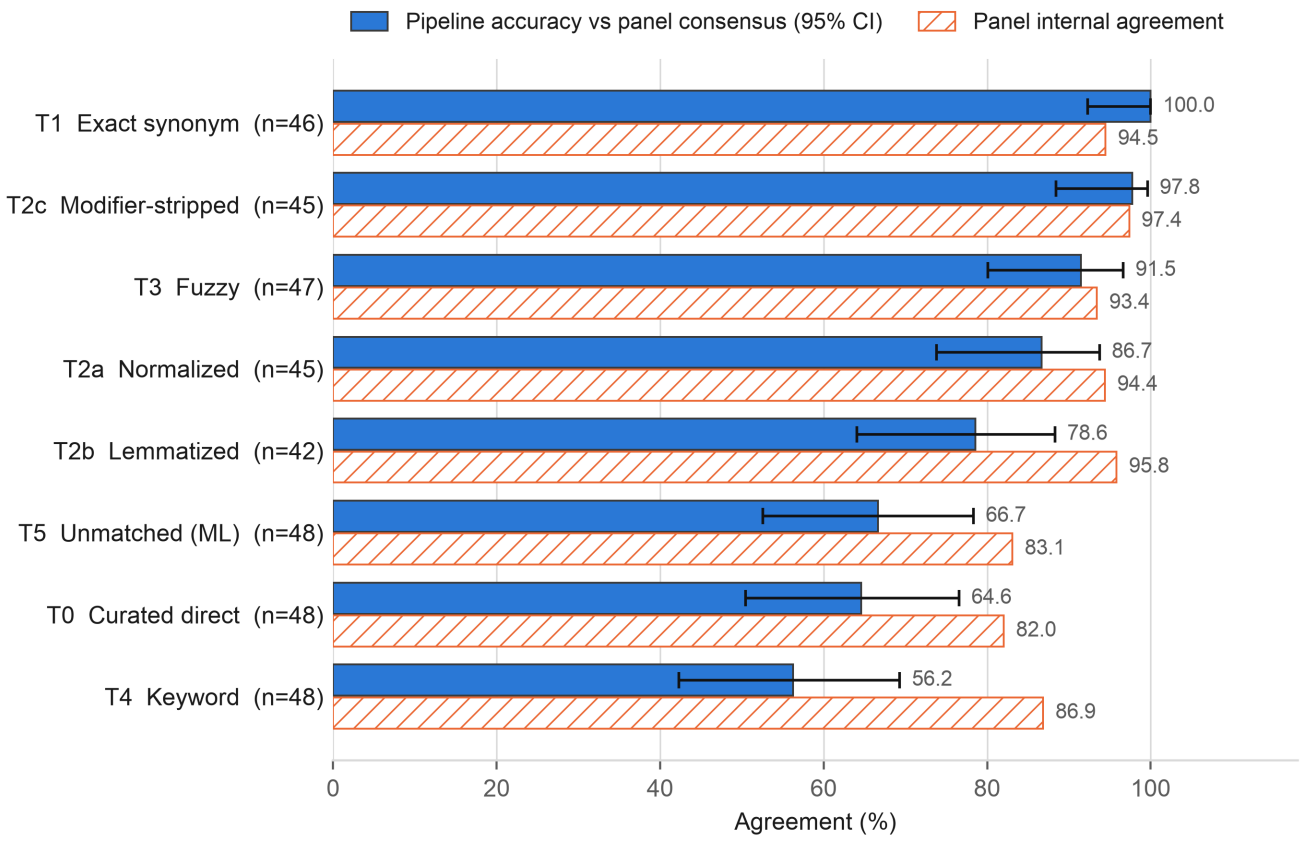


Figure 4. Pipeline performance by matching tier against the expert panel. Solid bars: percentage of terms for which the pipeline's Level 1 category matched the panel's modal consensus, with 95% Wilson score confidence intervals. Hatched bars: mean internal agreement among the nine reviewers on the same terms, indicating the reliability of the reference standard itself. Tiers are ordered by pipeline accuracy. The curated direct-mapping tier (T0) sits below every automated string-matching tier. The intervals are wide and overlap for most adjacent tiers, so the ordering is descriptive rather than a test of individual differences. Panel agreement remains high across all tiers, including those where pipeline accuracy is lowest, so the weak performance at T0, T4, and T5 reflects pipeline error rather than reviewer disagreement. Per-tier *n* is the number of the 48 terms drawn per tier that received a pipeline category; all sampled terms had a panel consensus. The 15 exclusions fall entirely in the concept-matching tiers (T1–T3), where a matched concept whose IS-A path reaches no Level 1 root yields no category.