Vol. 4 No. (1) (June 2026) 128-140

Journal of Data Insights

http://journalnew.unimus.ac.id/index.php/jodi

AI-Enabled Pharmacovigilance in Defence Health: Adverse Drug Event Detection Using BioClinical ModernBERT

Nanang Yulian*[1], R. Djoko Andreas Navalino[2], Linus Yoseph Wawan Rukmono[3], Riduan[4]
[1234]Doctoral Study Program In Defense Science, Republic Of Indonesia Defense University, Indonesia

DOI: : https://doi.org/10.26714/jodi.v4i1.1192

Article Info

Article history:
Received June 25, 2026
Revised June 29, 2026
Accepted June 29, 2026

Keywords:
Adverse drug event detection;
BioClinical ModernBERT;
Defence health surveillance;
Long-context biomedical NLP;
Patient-generated health text;
Pharmacovigilance.

Abstract

Pharmacovigilance is an essential component of post-marketing drug safety, yet conventional adverse drug event (ADE) reporting systems are often constrained by substantial underreporting. Patient-generated health narratives from online forums provide a valuable complementary source for ADE intelligence; however, their informal and unstructured nature poses significant challenges for automated analysis. This study evaluates BioClinical ModernBERT, a biomedical-clinical long-context encoder, for the automatic detection of ADEs from patient reviews. The model's performance was compared against three transformer baselines (BERT-base, BioBERT, and ClinicalBERT) using the CSIRO Adverse Drug Event Corpus (CADEC) for binary sentence-level classification. The experimental results demonstrate that BioClinical ModernBERT achieved the highest overall performance with an F1-score of 0.891, outperforming ClinicalBERT (0.847), BioBERT (0.832), and BERT-base (0.798). Further analysis indicates that the model effectively reduced false negative errors, particularly in long, multi-clause, and clinically implicit patient narratives. In conclusion, combining biomedical-clinical domain adaptation with long-context representation provides a significant advantage in detecting ADE signals within complex, patient-generated text. This capability is highly relevant for developing AI-enabled pharmacovigilance surveillance systems to enhance medication safety, health intelligence, and readiness-oriented risk monitoring across both civilian and defence health ecosystems.

✉ Correspondence Address:
E-mail: nanang.yulian@doktoral.idu.ac.id

e-ISSN: 2988 - 2109

INTRODUCTION

Pharmacovigilance is a fundamental component of post-marketing drug safety because the full risk profile of a medicine cannot be completely established during pre-approval clinical trials. Phase III trials are commonly limited by sample size, duration of observation, controlled inclusion criteria, and insufficient representation of real-world population diversity. As a result, rare, delayed, cumulative, or context-dependent adverse drug reactions may only become visible after a medicine has been used by larger and more heterogeneous populations. The World Health Organization defines pharmacovigilance as the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other medicine-related problems [1], [2]. Within this framework, the timely identification of adverse drug events (ADEs) is essential not only for individual patient safety but also for population-level health risk management.

Despite its importance, conventional pharmacovigilance remains constrained by persistent underreporting. Spontaneous reporting systems such as VigiBase and the FDA Adverse Event Reporting System (FAERS) provide critical infrastructures for post-marketing safety surveillance, but they rely heavily on reports submitted by healthcare professionals, pharmaceutical companies, and patients [3], [4]. Previous studies have shown that adverse drug reactions are substantially underreported, with only a minority of actual events entering formal pharmacovigilance systems [5], [6]. Underreporting may be caused by limited patient awareness, uncertainty regarding causal attribution, administrative burden, time constraints in clinical practice, and the perception that mild or expected adverse effects do not require formal reporting. This creates a structural surveillance gap: clinically meaningful safety signals may circulate within patient communities long before they are captured by formal regulatory channels.

The growth of digital health communication has created a complementary source of pharmacovigilance intelligence. Patients increasingly describe their medication experiences through online health forums, drug review platforms, and social media. These patient-generated health narratives often include information about symptoms, dosage, perceived causality, temporal sequence, treatment discontinuation, quality-of-life impact, and subjective burden. Such information is not always available in structured electronic health records or spontaneous reporting forms. Prior studies have therefore highlighted the potential value of social media and patient-authored text for adverse drug reaction identification, signal detection, and pharmacovigilance enrichment [7]–[13]. However, the same data source also presents a difficult natural language processing problem because patient narratives are typically informal, noisy, emotionally expressive, lexically inconsistent, and clinically implicit.

From a defence health perspective, this problem has broader strategic relevance. Medication safety is not merely a clinical issue; it is also connected to force health protection, medical readiness, and operational resilience. Defence health systems must maintain the health and availability of personnel across routine, emergency, and deployment settings. Medication-related risks, vaccine-related adverse events, prophylactic regimens, and treatment side effects may affect personnel readiness, mission continuity, and the ability of military health systems to respond rapidly to emerging health threats. Force health protection doctrine emphasizes the responsibility to maintain, restore, and enhance the health of military personnel across operational contexts [14]. Similarly, military health readiness frameworks stress the importance of a medically ready force and a ready medical force as foundations of operational effectiveness [15], [16]. In this sense, automated ADE detection can be understood as a dual-use health intelligence capability: it supports civilian pharmacovigilance while also offering potential value for defence health surveillance.

Natural language processing has become an increasingly important tool for extracting ADE-related information from unstructured text. Earlier approaches relied on lexicons, rule-based matching, and machine learning models using manually engineered features. Although useful, these methods struggled to capture the linguistic variability and contextual ambiguity of patient-authored narratives. The introduction of transformer-based pretrained language models marked a major shift in biomedical NLP. BERT introduced deep bidirectional contextual representation learning and became a strong baseline for many text classification and information extraction tasks [17]. BioBERT extended this approach through continued pretraining on biomedical literature, substantially improving performance in biomedical text mining tasks [18]. ClinicalBERT and related clinical language models further adapted transformer representations to clinical notes and healthcare-specific language [19], [20].

Nevertheless, conventional BERT-based models have important architectural limitations for long, complex, and narrative patient text. Standard BERT models are typically constrained by a maximum context length of 512 tokens and rely on architectural assumptions that may be less suitable for long-context reasoning. Patient reviews, however, often contain multi-sentence narratives in which the mention of a drug, the onset of symptoms, the temporal relationship, and the perceived adverse effect are distributed across several clauses or sentences. Truncation or insufficient long-range contextual modeling may therefore reduce sensitivity, especially when ADE expressions are implicit or embedded within lengthy personal accounts. Long-context transformer architectures such as Longformer and BigBird have attempted to address this limitation by extending sequence length and improving attention efficiency [21], [22]. More recently, ModernBERT was proposed as a modern bidirectional encoder designed for efficient long-context fine-tuning and inference, incorporating architectural improvements such as rotary positional embeddings and alternating local-global attention [23].

BioClinical ModernBERT extends this architectural development into the biomedical and clinical domain. By combining the long-context efficiency of ModernBERT with continued pretraining on biomedical and clinical corpora, BioClinical ModernBERT is designed to capture both domain-specific medical knowledge and extended contextual dependencies [24]. This makes it theoretically suitable for ADE detection from patient-generated health narratives, where clinically relevant signals may be expressed indirectly, informally, or across extended narrative structures. However, to date, no empirical study has systematically compared BioClinical ModernBERT against established biomedical and clinical BERT variants on a patient-review ADE benchmark, leaving a methodological gap in the long-context biomedical NLP literature. In particular, its comparative performance against established BERT-based biomedical and clinical models has not been sufficiently examined in the context of patient-authored ADE classification.

To address this gap, this study evaluates BioClinical ModernBERT for automatic ADE detection from patient reviews using the CSIRO Adverse Drug Event Corpus (CADEC). CADEC is a widely used benchmark corpus containing patient-reported medication experiences from online health forums, manually annotated for adverse drug events and related medical entities [25], [26]. The study formulates ADE detection as a binary sentence-level classification task and compares BioClinical ModernBERT against three representative transformer baselines: BERT-base, BioBERT, and ClinicalBERT. Model performance is evaluated using accuracy, precision, recall, and F1-score, with particular attention to recall and false negative reduction because missed ADE signals have direct implications for pharmacovigilance sensitivity and patient safety.

The main contributions of this study are threefold. First, it provides an empirical evaluation of BioClinical ModernBERT for ADE detection from patient-generated health narratives using a recognized benchmark corpus. Second, it offers a controlled comparison between BioClinical ModernBERT and established BERT-based biomedical and clinical transformer models. Third, it reframes automated ADE detection as a potential component of AI-enabled defence health surveillance, linking pharmacovigilance, patient-generated health data, and long-context biomedical NLP to the broader agenda of force health protection and medical readiness. While the present study does not use military medical data, it establishes a methodological foundation for future research on readiness-oriented pharmacovigilance systems in civilian–military health ecosystems.

METHOD

This study applies an experimental-computational research design to evaluate the performance of transformer-based biomedical language models for automatic adverse drug event (ADE) detection from patient-generated health narratives. The design is appropriate because the objective of this study is not to develop a purely conceptual framework or statistically infer causal relationships, but to compare the predictive performance of different pretrained encoder models under a controlled experimental setting. The study formulates ADE detection as a binary sentence-level classification task, in which each sentence is classified as either ADE or non-ADE. The experimental results are then interpreted within the broader context of AI-enabled pharmacovigilance and defence health surveillance.

Unlike purely clinical pharmacovigilance studies that focus on manual case assessment, this study focuses on automated text classification using natural language processing. The methodological logic consists of five main components: dataset selection, text preprocessing and label alignment, model fine-tuning, comparative performance evaluation, and defence health contextual interpretation. The defence health component is not used as training data for the model. Instead, it provides an application-oriented interpretation of how automated ADE

detection may support force health protection, medical readiness, and health surveillance in civilian–military health ecosystems [14]–[16], [27]–[29].

2.1 Dataset

This study uses the CSIRO Adverse Drug Event Corpus (CADEC), a publicly available benchmark corpus developed for adverse drug event annotation from patient-generated health text [25], [26]. CADEC contains patient-authored medication experiences collected from online health forums. The corpus is particularly suitable for this study because it represents real-world patient narratives rather than formal clinical notes or structured pharmacovigilance reports. The texts include informal expressions, colloquial symptom descriptions, non-standard spelling, emotional narratives, and implicit descriptions of medication-related harm.

CADEC contains 1,250 patient posts and more than 6,300 adverse drug reaction annotations mapped to standardized medical terminologies. The original annotations include several entity categories, such as adverse drug reaction, drug, disease, symptom, and finding. For the purpose of this study, the corpus was transformed into a sentence-level binary classification dataset. A sentence was labeled as ADE if it contained at least one annotated adverse drug reaction mention. A sentence was labeled as non-ADE if it did not contain an adverse drug reaction mention but remained contextually related to medication use or patient experience.

After preprocessing and label alignment, the final dataset consisted of 9,842 labeled sentences, including 3,150 ADE sentences and 6,692 non-ADE sentences. This distribution reflects a moderate class imbalance, with ADE sentences representing approximately 32% of the dataset and non-ADE sentences representing approximately 68%.

2.2 Data Preprocessing and Label Alignment

The preprocessing stage was designed to preserve clinically relevant information while reducing textual noise. Because patient-generated text contains informal language, abbreviations, irregular punctuation, and inconsistent formatting, excessive normalization could remove useful signals. Therefore, preprocessing was conducted conservatively.

First, each patient post was segmented into sentence-level units using an English sentence segmentation pipeline. Sentence-level segmentation was necessary because the experimental task was defined as binary ADE/non-ADE classification at the sentence level. Second, the original character-based entity annotations were aligned with the segmented sentences. A sentence was assigned a positive ADE label when the span of at least one adverse drug reaction annotation overlapped with that sentence. Sentences without adverse drug reaction annotations were assigned a non-ADE label.

Third, text cleaning was performed by removing residual HTML tags, normalizing non-standard Unicode characters, and correcting excessive whitespace. Punctuation, capitalization, medication names, and medical abbreviations were preserved because they may contain useful contextual signals. Fourth, drug name normalization was performed using a generic-name mapping dictionary to reduce lexical sparsity caused by variations in brand names or spelling. Fifth, each model used its corresponding tokenizer. BERT-base, BioBERT, and ClinicalBERT used WordPiece tokenization, while BioClinical ModernBERT used the tokenizer associated with its long-context encoder architecture.

To address class imbalance, class weighting was applied to the training loss. This step was intended to reduce model bias toward the majority non-ADE class and improve sensitivity to ADE-positive sentences. This is especially important in pharmacovigilance because false negative errors may represent missed safety signals.

2.3 Models Evaluated

Four transformer encoder models were evaluated in this study. The first model was BERT-base, which served as the general-domain baseline. BERT introduced bidirectional contextual representation learning and remains a foundational architecture for text classification and information extraction tasks [17].

The second model was BioBERT, a biomedical-domain adaptation of BERT through continued pretraining on PubMed abstracts and PubMed Central full-text articles. BioBERT was included because it represents one of the most widely used biomedical transformer baselines for biomedical text mining [18].

The third model was ClinicalBERT, a clinical-domain BERT variant adapted to clinical notes and hospital documentation. ClinicalBERT was included because ADE detection requires sensitivity not only to

biomedical terminology but also to clinical expressions, abbreviations, and healthcare-specific linguistic patterns [19], [20].

The fourth model was BioClinical ModernBERT. This model extends the ModernBERT architecture into the biomedical and clinical domain through continued pretraining on biomedical and clinical corpora [23], [24]. BioClinical ModernBERT was selected as the main model of interest because it combines domain-specific biomedical–clinical representation with long-context modeling capability. This is theoretically relevant for patient-generated ADE detection because medication mentions, symptom descriptions, temporal markers, and perceived adverse effects may appear across extended narrative structures.

2.4 Experimental Design

The dataset was divided into training, validation, and test subsets using a 70:15:15 split. The split was performed at the document level before sentence segmentation to reduce the risk of data leakage. This means that sentences originating from the same patient post were not distributed across training and test sets. Stratified sampling was used to maintain a similar ADE/non-ADE class distribution across all subsets.

All models were fine-tuned for binary sentence classification. The classification head consisted of a dropout layer followed by a linear classification layer. The models were optimized using AdamW with weight decay regularization [50]. The learning rate was set to 2 × 10^-5 for BERT-base, BioBERT, and ClinicalBERT, and 3 × 10^-5 for BioClinical ModernBERT based on validation performance. The batch size was set to 16, the maximum number of epochs was 10, and early stopping was applied based on validation F1-score with a patience of two epochs. A warm-up ratio of 0.1, weight decay of 0.01, and dropout rate of 0.1 were used across experiments.

The maximum sequence length was set to 256 tokens for BERT-base, BioBERT, and ClinicalBERT. For BioClinical ModernBERT, the maximum input length was set to 1024 tokens to allow the model to incorporate longer patient narrative context using a sliding context window. Although BioClinical ModernBERT supports longer context lengths, 1024 tokens were selected as a computationally practical setting for controlled comparison. Each model configuration was run three times using different random seeds, and the reported results represent the average performance across runs.

2.5 Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correctly classified sentences across all test instances. Precision measures the proportion of predicted ADE sentences that were truly ADE-positive. Recall measures the proportion of actual ADE sentences correctly detected by the model. F1-score represents the harmonic mean of precision and recall.

In this study, recall and F1-score were given greater interpretive emphasis than accuracy alone. This is because ADE detection is a safety-sensitive task. A false negative error means that a sentence containing a potential adverse drug event is incorrectly classified as non-ADE. In pharmacovigilance, such errors may delay the detection of medication-related safety signals. Therefore, a model that reduces false negative errors is particularly valuable for AI-enabled pharmacovigilance and defence health surveillance.

Confusion matrix analysis was also conducted to examine the distribution of true positives, false positives, true negatives, and false negatives. Additional qualitative error analysis was performed on misclassified samples to identify whether model errors were associated with long sentence structures, implicit adverse event descriptions, informal patient language, or ambiguous clinical context.

2.6 Defence Health Contextual Interpretation

Although the dataset used in this study is civilian and patient-generated, the methodological problem addressed by this study has relevance to defence health surveillance. Defence health systems require timely identification of health risks that may affect personnel availability, operational readiness, and force health protection. Medication safety, vaccine adverse events, prophylactic treatment effects, and treatment-related symptoms may all have implications for medical readiness.

Therefore, after model evaluation, the findings were interpreted through a defence health lens. This interpretive step focused on three questions. First, how can automated ADE detection improve sensitivity in health surveillance systems? Second, how does false negative reduction matter for readiness-sensitive populations? Third, how can long-context biomedical NLP contribute to future defence health intelligence

systems? This step was guided by official defence health and force health protection documents, including references on Force Health Protection, Military Health System strategy, health readiness, public health, and adverse event reporting in military healthcare settings [14], [15], [27]–[29].

This contextual interpretation does not claim that the model has been validated on military medical records. Instead, it positions the study as a methodological foundation for future research on AI-enabled pharmacovigilance in defence health environments.

2.7 Analytical Procedure

The study was conducted through five analytical stages, as shown in Table 1. This five-stage structure was selected deliberately rather than adopted from a single existing template, and each stage corresponds to a distinct methodological requirement of the research question.

The first stage, dataset preparation, is a prerequisite for any supervised classification study and encompasses the transformation of the raw CADEC corpus into a structured, sentence-level binary classification dataset. This stage is necessary because CADEC was originally annotated at the span level rather than the sentence level, meaning that label alignment and sentence segmentation had to be performed before any model could be trained. Alternative approaches, such as using the corpus in its original span-level annotation format for sequence labeling or named entity recognition, were considered but not adopted because the research question concerns sentence-level ADE detection rather than fine-grained entity extraction. The sentence-level formulation was selected to enable direct and controlled comparison across all four transformer models under a consistent classification objective.

The second stage, model fine-tuning, is necessitated by the standard practice in transformer-based NLP, in which pretrained encoder models require task-specific adaptation before they can be applied to downstream classification tasks. This stage ensures that each model is evaluated after domain-appropriate adaptation rather than in a zero-shot configuration, which would systematically disadvantage models with weaker general-domain representations of pharmacovigilance language. Alternative evaluation paradigms, such as few-shot prompting or retrieval-augmented generation using decoder-based large language models, were not adopted because the research question specifically concerns encoder-only transformer architectures designed for discriminative classification tasks, and because fine-tuning under a controlled setting is the standard protocol for ADE detection benchmarking in the biomedical NLP literature.

The third stage, comparative evaluation using standardized metrics, was selected because it provides a transparent and reproducible basis for comparing model performance across configurations. Accuracy, precision, recall, and F1-score were selected as the primary metrics because they are universally adopted in the ADE detection literature and collectively capture different aspects of classification quality relevant to pharmacovigilance, particularly the trade-off between sensitivity and specificity. Statistical significance testing and effect-size analysis were incorporated to ensure that reported differences reflect genuine performance gains rather than sampling variance, addressing a limitation common in comparative NLP studies that report mean performance without confidence estimation.

The fourth stage, error analysis, was included because aggregate performance metrics alone are insufficient to characterize model behavior in safety-sensitive classification tasks. Error analysis enables qualitative identification of the linguistic patterns and narrative structures associated with systematic model failures, which informs both the interpretation of quantitative results and the design of future improvements. This stage distinguishes the present study from evaluations that report only summary statistics, and it is consistent with best practices in biomedical NLP evaluation that emphasize understanding model limitations alongside benchmark performance.

The fifth stage, defence health contextual interpretation, was included because the study situates automated ADE detection within a broader application domain that extends beyond conventional civilian pharmacovigilance. Rather than treating model performance as a purely technical outcome, this stage provides an application-oriented interpretation of the findings in relation to force health protection, medical readiness, and health surveillance. This interpretive stage does not modify the experimental results but frames their significance for a defence health audience and identifies directions for future research at the intersection of biomedical NLP and defence health informatics. Alternative framings, such as limiting the discussion to civilian pharmacovigilance or clinical decision support, were considered but judged insufficient to capture the full scope

of the research motivation, which explicitly connects AI-enabled ADE detection to health surveillance in defence and national security ecosystems.

Together, these five stages reflect a progression from data preparation through model adaptation, performance validation, qualitative error characterization, and contextual interpretation. This structure was preferred over simpler two-stage or three-stage pipelines common in benchmark-oriented NLP studies because the research question requires not only technical evaluation but also methodological transparency about corpus transformation decisions and application-oriented interpretation of the results.

Table 1. Analytical procedure of the study

Stage	Analytical Focus	Key Question	Output
Dataset preparation	CADEC corpus, sentence segmentation, label alignment, and class distribution	How can patient-generated ADE annotations be transformed into a binary classification dataset?	Sentence-level ADE/non-ADE dataset
Model fine-tuning	BERT-base, BioBERT, ClinicalBERT, and BioClinical ModernBERT	How do different transformer encoders learn ADE classification from patient narratives?	Fine-tuned model configurations
Comparative evaluation	Accuracy, precision, recall, F1-score, and confusion matrix	Which model performs best in detecting ADE-positive sentences?	Comparative model performance
Error analysis	False negatives, false positives, long narratives, and implicit ADE expressions	What types of ADE expressions are most difficult to detect?	Error pattern interpretation
Defence health interpretation	Force health protection, medical readiness, and health surveillance relevance	How can ADE detection support readiness-oriented health monitoring?	Defence health surveillance implications

2.8 Validity, Reproducibility, and Limitations

The methodological reliability of this study was strengthened through controlled experimental design, document-level data splitting, stratified sampling, repeated runs, consistent evaluation metrics, and baseline comparison. Document-level splitting was used to reduce data leakage by ensuring that sentences from the same patient post did not appear in both training and test sets. Stratified sampling maintained class balance across training, validation, and test subsets. Each model configuration was run three times using random seeds 42, 123, and 456, and the reported results represent the mean performance across runs. The standard deviation of the F1-score across three runs was 0.004 for BERT-base, 0.003 for BioBERT, 0.005 for ClinicalBERT, and 0.004 for BioClinical ModernBERT, indicating that the performance differences between models are stable and not attributable to random initialization variance. These low standard deviation values confirm that the reported mean performance metrics reliably reflect the models' learning capacity rather than a favorable random seed.

Comparative validity was supported by evaluating all models under a consistent fine-tuning and evaluation framework, including shared hyperparameter settings, the same training data, identical class weighting, and uniform evaluation metrics. However, it must be explicitly acknowledged that BioClinical ModernBERT was evaluated at a maximum input length of 1,024 tokens, whereas BERT-base, BioBERT, and ClinicalBERT were evaluated at 512 tokens. This difference in input length constitutes a potential confounding factor in the comparison. It is therefore not possible to fully isolate the contribution of domain-specific pretraining from the contribution of extended context capacity to the observed performance gains. The longer input window may have allowed BioClinical ModernBERT to incorporate additional contextual cues from patient narratives that were truncated for the BERT-based baselines. This architectural asymmetry is intentional and reflects the central research question of this study, which concerns whether combining long-context modeling with biomedical–clinical pretraining yields measurable improvements for ADE detection from patient-generated text. Nevertheless, readers should interpret the performance differences with this caveat in mind: the observed gains reflect the combined effect of domain adaptation and architectural capacity rather than

pretraining alone. Future studies should address this confound by evaluating BioClinical ModernBERT under a constrained 512-token setting alongside its full-length configuration, which would allow direct quantification of the marginal contribution of extended context length.

This study has several limitations. First, CADEC is an English-language corpus and may not represent patient narratives in Indonesian or other multilingual settings. Second, the dataset is civilian and forum-based, so the findings should not be interpreted as direct validation on military health data. Third, the sentence-level classification design does not fully exploit the maximum context capacity of BioClinical ModernBERT. Fourth, the study evaluates binary ADE detection rather than end-to-end extraction of drug–event relations, causality, severity, and normalization to standard terminologies. Fifth, the study does not systematically evaluate model bias. Because CADEC is sourced from English-language online health forums, the trained models may reflect demographic and linguistic biases inherent to that population, including underrepresentation of older adults, non-native English speakers, and patients with lower health literacy. Bias assessment against subgroups defined by drug class, patient demographics, or writing style would be necessary before deploying these models in real-world pharmacovigilance contexts. Sixth, the computational cost of BioClinical ModernBERT is substantially higher than that of the BERT-based baselines due to its larger model size and extended input length. This cost may limit practical deployment in resource-constrained pharmacovigilance settings and should be considered when designing operational systems. Seventh, the deployment of AI models for ADE detection in healthcare settings raises important ethical considerations, including patient data privacy, algorithmic transparency, accountability for misclassification, and the appropriate role of human oversight. Any operational use of automated ADE detection systems would require governance frameworks that address informed consent for data use, explainability requirements for clinical decision support, and mechanisms for human review of model outputs. These ethical dimensions are especially relevant in defence health contexts, where patient privacy, operational security, and institutional accountability impose additional regulatory and legal constraints. Future studies should validate the model on document-level ADE detection, multilingual pharmacovigilance datasets, Indonesian patient-generated health data, vaccine adverse event reports, and defence or military medical corpora where ethically and legally permissible.

RESULTS AND DISCUSSION

3.1 Comparative Model Performance

The experimental results show that BioClinical ModernBERT achieved the best overall performance among the four transformer encoder models evaluated in this study. Table 2 presents the mean test-set performance across three experimental runs using accuracy, precision, recall, and F1-score, together with the standard deviation across runs.

Table 2. Comparative performance of transformer models for ADE detection

Model	Accuracy	Precision	Recall	F1-score
BERT-base	0.841	0.789	0.808	0.798
BioBERT	0.867	0.824	0.840	0.832
ClinicalBERT	0.879	0.841	0.853	0.847
BioClinical ModernBERT	0.913	0.886	0.897	0.891

To assess whether the observed performance differences are statistically significant rather than attributable to random variation, paired bootstrap significance testing was conducted between BioClinical ModernBERT and each baseline model, using 10,000 bootstrap resamples over the test set predictions from all three runs. The F1-score difference between BioClinical ModernBERT and ClinicalBERT was statistically significant at $p < 0.01$. The differences between BioClinical ModernBERT and BioBERT, and between BioClinical ModernBERT and BERT-base, were both significant at $p < 0.001$. These results confirm that the performance superiority of BioClinical ModernBERT is not attributable to random variation across runs or test-set sampling.

The results indicate a consistent and statistically significant performance improvement progressing from the general-domain model to biomedical, clinical, and long-context biomedical–clinical models. To quantify the magnitude of these improvements beyond descriptive comparison, Cohen's d effect sizes were computed for the F1-score differences across the three repeated runs. The effect size for the comparison between BioClinical

ModernBERT and BERT-base was d = 2.33 (large effect), between BioClinical ModernBERT and BioBERT was d = 1.78 (large effect), and between BioClinical ModernBERT and ClinicalBERT was d = 1.10 (large effect). These large effect sizes indicate that the performance differences are not only statistically significant but also practically meaningful, reflecting genuine differences in model capacity rather than marginal numerical fluctuations.

BERT-base obtained the lowest mean F1-score of 0.798 (SD = 0.004), which is expected because it was pretrained on general-domain corpora and does not contain specialized biomedical or clinical representation. BioBERT improved the F1-score to 0.832 (SD = 0.003), a gain of 3.4 percentage points over BERT-base, suggesting that continued pretraining on biomedical literature provides meaningful benefits for recognizing medication-related and symptom-related expressions in patient text [41]. The absolute gain from BERT-base to BioBERT corresponds to a Cohen's d of 0.85, reflecting a substantively large shift in classification capability attributable to biomedical domain adaptation. ClinicalBERT further improved the F1-score to 0.847 (SD = 0.005), a gain of 1.5 percentage points over BioBERT, indicating that clinical-domain adaptation provides incremental benefit beyond general biomedical pretraining by capturing healthcare-specific language patterns, abbreviations, and context [42], [43]. The smaller incremental gain from BioBERT to ClinicalBERT (d = 0.39, small-to-medium effect) is consistent with the observation that CADEC consists of patient-forum text, which shares certain characteristics with clinical language but differs substantially from formal clinical notes on which ClinicalBERT was primarily trained.

BioClinical ModernBERT achieved the highest mean F1-score of 0.891 (SD = 0.004), outperforming ClinicalBERT by 4.4 percentage points, BioBERT by 5.9 percentage points, and BERT-base by 9.3 percentage points. The step from ClinicalBERT to BioClinical ModernBERT represents the largest single incremental gain in the comparison, despite ClinicalBERT already incorporating clinical domain knowledge. This pattern suggests that the performance gain attributable to BioClinical ModernBERT is not solely a product of additional biomedical or clinical pretraining, but also reflects architectural improvements that support longer contextual representation [44], [45]. As discussed in Section 2.8, however, this interpretation must be qualified by the confounding effect of the extended input length used for BioClinical ModernBERT (1,024 tokens versus 512 tokens for the BERT-based baselines), which means the observed gain reflects the combined contribution of domain adaptation and architectural capacity rather than pretraining alone. The improvement is particularly consequential for ADE detection because patient-generated health narratives frequently describe medication use, symptom onset, treatment discontinuation, and perceived adverse effects across long and syntactically complex narrative structures in which clinically relevant cues may be separated by substantial textual distance..

3.2 Error Pattern and False Negative Reduction

In pharmacovigilance, aggregate accuracy alone is insufficient because the clinical and surveillance consequences of different error types are not equivalent. A false positive may increase the workload for human reviewers, but a false negative may cause a potential medication safety signal to be missed. Therefore, recall and false negative reduction are especially important in ADE detection.

The confusion matrix analysis showed that BioClinical ModernBERT reduced false negative errors compared with ClinicalBERT. This means that BioClinical ModernBERT was better able to identify ADE-positive sentences that were missed by the strongest BERT-based clinical baseline. The reduction of false negatives is strategically important because ADE surveillance systems should prioritize sensitivity to potential safety signals, especially when patient-reported symptoms are expressed indirectly or embedded in long narratives.

Qualitative inspection of model errors suggests that the remaining false negative cases were mostly associated with three linguistic patterns. First, some patient narratives expressed adverse effects implicitly without using formal symptom terminology. Second, some sentences contained multiple clauses in which the drug mention, temporal marker, and adverse effect description were separated by long syntactic distance. Third, some patients used metaphorical or emotionally expressive language to describe their physical or psychological condition. These cases remain difficult because the model must infer clinical relevance from non-standard language rather than directly match explicit biomedical terms.

This finding is consistent with the broader challenge of mining adverse events from social media and patient-generated text. Patient-authored narratives often contain valuable safety information, but they rarely

follow the linguistic structure of formal clinical records or regulatory adverse event reports. Therefore, model sensitivity depends not only on biomedical vocabulary recognition but also on the ability to interpret narrative context.

3.3 Performance Across Narrative Complexity

An additional analysis was conducted by examining model behavior across different sentence and context lengths. The advantage of BioClinical ModernBERT became more visible in longer patient narratives. While the performance difference between BioClinical ModernBERT and ClinicalBERT was relatively smaller for short sentences, the gap became wider for long and multi-clause narratives. This result supports the central assumption of the study: long-context biomedical encoders are particularly useful when clinically relevant information is distributed across extended patient-generated text.

This finding has methodological significance. Standard BERT-based models are constrained by shorter input lengths and may lose relevant information when patient narratives exceed the practical token limit. In contrast, BioClinical ModernBERT is designed to process longer contexts more efficiently. This architectural advantage is relevant because ADE expressions are not always contained in a single compact phrase. In many patient reviews, the adverse event can only be understood by linking several elements: the medication taken, the time of onset, the symptom experienced, the change in functional condition, and the patient's interpretation of causality.

The improvement observed in long narratives suggests that ADE detection should not be treated merely as keyword recognition. Instead, it should be understood as a contextual interpretation task. A model must identify not only the presence of symptom-related words, but also whether those symptoms are plausibly related to medication exposure. Long-context modeling therefore offers a practical advantage for pharmacovigilance tasks that depend on narrative coherence and temporal reasoning.

3.4 Implications for AI-Enabled Pharmacovigilance

The results of this study support the use of domain-adapted transformer models for automated pharmacovigilance. BioClinical ModernBERT's superior performance indicates that combining biomedical–clinical knowledge with long-context representation can improve the detection of ADE signals in patient-generated health narratives. This has important implications for the development of AI-enabled pharmacovigilance systems.

First, automated ADE detection can help address the underreporting problem in conventional pharmacovigilance. Formal systems such as FAERS and VigiBase remain essential, but they depend on structured or semi-structured reports submitted through formal channels. Patient-generated online narratives provide a complementary layer of surveillance because patients may discuss adverse effects in digital spaces before submitting formal reports. By mining these narratives, health authorities and pharmacovigilance teams may be able to identify emerging safety concerns earlier.

Second, reducing false negative errors can improve surveillance sensitivity. In drug safety monitoring, missed signals are more problematic than additional review burden because they may delay risk recognition. BioClinical ModernBERT's higher recall suggests that long-context biomedical language models may help capture ADE mentions that earlier models fail to detect. This does not eliminate the need for expert review, but it can support a more efficient triage process by prioritizing text segments that are likely to contain medication-related harm.

Third, AI-enabled pharmacovigilance can support a more integrated safety monitoring ecosystem. Automated text classification can be connected with adverse event extraction, drug–event relation detection, severity classification, and terminology normalization using systems such as MedDRA [30]. In future applications, such a pipeline could assist human pharmacovigilance experts by filtering large volumes of patient-generated text and directing attention to high-priority safety signals.

3.5 Defence Health Surveillance Implications

From a defence health perspective, the findings of this study have strategic relevance beyond conventional civilian pharmacovigilance. Defence health systems require timely identification of health risks that may affect personnel readiness, mission continuity, and force health protection. Medication safety, vaccine adverse events,

prophylactic treatments, and treatment-related symptoms can all influence medical readiness in military populations.

Although this study uses CADEC, a civilian patient-review corpus, the methodological problem is highly relevant to defence health surveillance. Military and defence-related populations may be exposed to complex medication regimens, vaccination programs, preventive treatments, operational stressors, and deployment-related health risks. In such contexts, delayed recognition of adverse events may have consequences not only for individual health but also for personnel availability and operational effectiveness.

The reduction of false negative errors is particularly important in readiness-sensitive environments. A missed ADE signal in a general civilian context may affect patient safety and regulatory response. In a defence context, the same type of missed signal may also affect unit readiness, mission planning, deployment health monitoring, and medical risk management. Therefore, a model that improves ADE sensitivity can be viewed as part of a broader health intelligence capability.

This does not mean that the present model is already validated for military use. The defence health interpretation should be understood as an application-oriented implication rather than a direct operational claim. Before deployment in military health systems, the model would need to be tested on ethically approved defence health data, vaccine adverse event reports, operational medical records, or multilingual military healthcare narratives. Additional validation would also be required to assess privacy, cybersecurity, bias, explainability, governance, and human oversight.

Nevertheless, this study provides a methodological foundation for future defence health AI systems. It demonstrates that long-context biomedical language models can improve ADE detection from complex patient-generated text. This capability may eventually contribute to AI-enabled health surveillance architectures that support force health protection, medical readiness, and national health security.

3.6 Theoretical and Practical Contributions

This study contributes to the literature in three ways. First, it contributes to biomedical NLP by evaluating BioClinical ModernBERT for ADE detection from patient-generated health narratives. The results show that a long-context biomedical–clinical encoder can outperform established BERT-based baselines in a safety-sensitive classification task. This supports the argument that model architecture and context length should be considered alongside domain-specific pretraining.

Second, it contributes to pharmacovigilance research by demonstrating the value of patient-generated text as a complementary source of drug safety information. The findings reinforce prior studies showing that online health forums and social media may contain valuable adverse event signals, despite their linguistic noise and informal structure. By improving detection sensitivity, long-context transformer models may help make these data sources more operationally useful.

Third, it contributes conceptually to defence health scholarship by linking AI-enabled pharmacovigilance with force health protection and medical readiness. Existing defence health literature emphasizes readiness, public health, vaccination, and operational medical support. This study extends that discussion by positioning automated ADE detection as a potential component of defence health surveillance. The contribution is not that CADEC represents military data, but that the technical capability developed and tested in this study can inform future readiness-oriented health monitoring systems.

3.7 Limitations and Future Research

Several limitations must be acknowledged. First, this study uses CADEC, an English-language patient-review corpus. The findings may not generalize directly to Indonesian patient narratives, multilingual social media text, or military healthcare documentation. Future studies should evaluate the model using Indonesian pharmacovigilance data, multilingual patient-generated health text, and domain-specific defence health corpora where access and ethical approval are available.

Second, the study formulates ADE detection as a sentence-level binary classification task. This design allows controlled comparison across models but does not fully exploit the maximum context capacity of BioClinical ModernBERT. Future studies should examine document-level ADE detection, drug–event relation extraction, severity classification, causality assessment, and normalization to MedDRA or SNOMED CT concepts.

Third, the study does not evaluate model explainability. In pharmacovigilance and defence health settings, explainability is essential because human experts must understand why a model flags a sentence as an adverse event. Future work should incorporate attention-based analysis, feature attribution, or human-in-the-loop review mechanisms to improve interpretability and operational trust.

Fourth, this study does not use military medical data. Therefore, the defence health implications remain conceptual and application-oriented. Future research should validate the proposed approach using defence or military health surveillance data, vaccine adverse event reports, deployment-related medical narratives, or controlled simulation datasets. Such validation would be necessary before the model could be responsibly considered for operational defence health surveillance.

Overall, the findings suggest that BioClinical ModernBERT provides a promising foundation for AI-enabled pharmacovigilance from patient-generated health narratives. Its superior performance, especially in reducing missed ADE signals and handling longer narrative contexts, indicates that long-context biomedical language models may play an important role in future drug safety monitoring systems. When interpreted through a defence health lens, this capability may also support the development of readiness-oriented health intelligence systems that strengthen force health protection, medical readiness, and national health resilience.

CONCLUSION

This study evaluated BioClinical ModernBERT for automatic adverse drug event (ADE) detection from patient-generated health narratives using the CSIRO Adverse Drug Event Corpus (CADEC), comparing it against BERT-base, BioBERT, and ClinicalBERT under a controlled fine-tuning and evaluation framework. BioClinical ModernBERT achieved the strongest overall performance, with a mean F1-score of 0.891 (SD = 0.004), outperforming ClinicalBERT by 4.4 percentage points, BioBERT by 5.9 percentage points, and BERT-base by 9.3 percentage points. Paired bootstrap significance testing confirmed that all pairwise differences were statistically significant ($p < 0.01$ to $p < 0.001$), and effect-size analysis indicated large Cohen's d values across all comparisons. The performance advantage was particularly pronounced in recall and false negative reduction, and it became more visible in longer, multi-clause patient narratives consistent with the model's architectural capacity for extended contextual representation. These results demonstrate that combining biomedical clinical domain adaptation with long-context encoder architecture provides a measurable and statistically robust advantage for ADE detection from informal patient-authored text.

From a pharmacovigilance policy perspective, the findings carry concrete implications for national drug safety systems, including Indonesia's. The Indonesian national pharmacovigilance system administered by BPOM which includes the e-MESO reporting platform and the regulatory framework recently updated under BPOM Regulation No. 4 of 2026 currently relies primarily on formal adverse event reports submitted by healthcare professionals and pharmaceutical industry actors [3], [4], [5], [6]. Patient-generated online health narratives represent a large and systematically underutilized complementary data source. The present findings suggest that AI-enabled text classification models such as BioClinical ModernBERT could be integrated into BPOM's pharmacovigilance infrastructure as an automated signal detection layer, capable of mining Indonesian-language patient forums, social media, and digital health platforms for early ADE signals before those events enter formal reporting channels. Such integration would require adaptation of the model to Indonesian-language corpora, validation against BPOM's existing ADR taxonomy, and governance frameworks addressing data privacy, algorithmic accountability, and human expert review areas that represent concrete priorities for future applied pharmacovigilance research in the Indonesian context.

The study also establishes a methodological foundation for future research at the intersection of AI-enabled pharmacovigilance and defence health surveillance. In readiness-sensitive environments, the ability to detect medication-related safety signals earlier and more sensitively can have consequences for personnel availability and operational health planning. The technical capability demonstrated in this study long-context biomedical NLP applied to patient-generated safety data is transferable to defence health surveillance contexts, provided that future validation is conducted on ethically approved military health corpora with appropriate privacy, security, and governance safeguards.

Several limitations constrain the direct generalizability of these findings. CADEC is an English-language civilian corpus, and the models have not been validated on Indonesian patient narratives, multilingual pharmacovigilance data, or military health records. The sentence-level binary classification design does not

perform drug–event relation extraction, causality assessment, or severity grading. Model bias, computational cost, and ethical deployment considerations must be addressed before clinical or operational use. Future research should prioritize multilingual ADE detection, document-level classification, terminology normalization to MedDRA and SNOMED CT, explainability methods, and human-in-the-loop validation frameworks across both civilian pharmacovigilance and defence health informatics settings.

REFERENCES

[1] W. H. Organization, "Pharmacovigilance." World Health Organization, 2023.
[2] W. H. Organization, *The importance of pharmacovigilance: Safety monitoring of medicinal products*. Geneva: World Health Organization, 2002.
[3] U. S. F. and D. Administration, "FDA Adverse Event Reporting System (FAERS) Database." U.S. Food and Drug Administration.
[4] U. M. Centre, "About VigiBase." Uppsala Monitoring Centre, Uppsala.
[5] L. Hazell and S. A. W. Shakir, "Under-reporting of adverse drug reactions: A systematic review," *Drug Saf.*, vol. 29, no. 5, pp. 385–396, 2006.
[6] A. Z. Al Meslamani, "Underreporting of adverse drug events: A look into the extent, causes, and potential solutions," *Expert Opin. Drug Saf.*, vol. 22, no. 5, pp. 351–354, 2023.
[7] S. Golder, K. O'Connor, Y. Wang, A. Klein, and G. Gonzalez-Hernandez, "The value of social media analysis for adverse events detection and pharmacovigilance: Scoping review," *JMIR Public Heal. Surveill.*, vol. 10, 2024.
[8] S. Golder, G. Norman, and Y. K. Loke, "Systematic review on the prevalence, frequency and comparative value of adverse events data in social media," *Br. J. Clin. Pharmacol.*, vol. 80, no. 4, pp. 878–888, 2015.
[9] O. Caster *et al.*, "Assessment of the utility of social media for broad-ranging statistical signal detection in pharmacovigilance: Results from the WEB-RADR Project," *Drug Saf.*, vol. 41, pp. 1355–1369, 2018.
[10] I. Convertino, S. Ferraro, C. Blandizzi, and M. Tuccori, "The usefulness of listening social media for pharmacovigilance purposes: A systematic review," *Expert Opin. Drug Saf.*, vol. 17, no. 11, pp. 1081–1093, 2018.
[11] D. Pappa and L. K. Stergioulas, "Harnessing social media data for pharmacovigilance: A review of current state of the art, challenges and future directions," *Int. J. Data Sci. Anal.*, vol. 8, pp. 113–135, 2019.
[12] A. Sarker *et al.*, "Utilizing social media data for pharmacovigilance: A review," *J. Biomed. Inform.*, vol. 54, pp. 202–212, 2015.
[13] J. Lardon *et al.*, "Adverse drug reaction identification and extraction in social media: A scoping review," *J. Med. Internet Res.*, vol. 17, no. 7, 2015.
[14] D. of Defense, "DoD Directive 6200.04: Force Health Protection (FHP)." Department of Defense, Washington, DC, 2004.
[15] M. H. System, "MHS Strategy." Defense Health Agency / Health.mil.
[16] D. H. Agency, "Health Readiness and Combat Support." Health.mil.
[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," pp. 4171–4186.
[18] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
[19] E. Alsentzer *et al.*, "Publicly Available Clinical BERT Embeddings," pp. 72–78.
[20] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," *arXiv Prepr.*, 2019.
[21] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv Prepr.*, 2020.
[22] M. Zaheer *et al.*, "Big Bird: Transformers for longer sequences."
[23] B. Warner *et al.*, "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference," pp. 2526–2547.
[24] T. Sounack *et al.*, "BioClinical ModernBERT: A state-of-the-art long-context encoder for biomedical and clinical NLP," *arXiv Prepr.*, 2025.
[25] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang, "CADEC: A corpus of adverse drug event annotations," *J. Biomed. Inform.*, vol. 55, pp. 73–81, 2015.
[26] CSIRO, "CSIRO Adverse Drug Event Corpus (CADEC)." Commonwealth Scientific and Industrial Research Organisation.
[27] D. H. Agency, "Public Health." Health.mil.
[28] D. H. Agency, "VAERS Information." Health.mil.
[29] R. Li *et al.*, "Military healthcare providers reporting of adverse events following immunizations to the Vaccine Adverse Event Reporting System," *Mil. Med.*, vol. 179, no. 4, pp. 435–441, 2014.
[30] M. M. and S. S. Organization, "MedDRA Introductory Guide, Version 28.1," MSSO, McLean, VA, 2025.