The Journal of Prevention of Alzheimer's Disease 12 (2025) 100002

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Original Article

Artificial intelligence-enabled safety monitoring in Alzheimer's disease clinical trials

Gustavo A. Jimenez-Maggiora*, Michael C. Donohue, Michael S. Rafii, Rema Raman, Paul S. Aisen

Alzheimer's Therapeutic Research Institute, University of Southern California, San Diego, CA, United States

ARTICLE INFO

Keywords:
Artificial intelligence
Alzheimer's disease
Clinical trial
Safety
Adverse event
Medical coding
Classification

ABSTRACT

Background: Investigators conducting clinical trials have an ethical, scientific, and regulatory obligation to protect the safety of trial participants. Traditionally, safety monitoring includes manual review and coding of adverse event data by expert clinicians.
Objectives: Our study explores the use of natural language processing (NLP) and artificial intelligence (AI) methods to streamline and standardize clinician coding of adverse event data in Alzheimer's disease (AD) clinical trials.
Design: Our quantitative retrospective study aimed to develop a gold standard AD adverse event data set, evaluate the predictive performance of NLP-based models to classify adverse events, and determine whether automated coding is more efficient, accurate, reliable, and consistent than clinician coding.
Setting: Our study was conducted at the University of Southern California's Alzheimer's Therapeutic Research Institute (ATRI). ATRI serves as the clinical and data coordinating center for the Alzheimer's Clinical Trial Consortium (ACTC).
Participants: We collected demographic and adverse event data from eight completed clinical trials in participants (n=1920) with symptomatic AD conducted between 2005 and 2020.
Measurements: Original expert clinician-confirmed codes were used for all model performance comparisons. F1 score was used as the primary model selection metric. Final classifier performance was evaluated using predictive accuracy. Clinician effort was measured in time to code, review, and confirm coded adverse events.
Results: In a sample of 1000 adverse events, AI-based AE coding achieved higher accuracy (~20% increase in accuracy) and was more cost-effective (~80% cost reduction) than traditional clinician coding.
Conclusions: Our study results demonstrate how approaches that effectively combine AI and human expertise can improve the efficiency and quality of adverse event coding and clinical trial safety monitoring.

1. Introduction

Despite recent therapeutic breakthroughs, Alzheimer's disease (AD) remains one of the most serious threats to human health and well-being in the new millennium [1]. As one of the leading causes of death in adults over the age of 65 in the United States and globally, the need for a redoubling of efforts to develop effective and safe therapies has never been more pressing. As the number of countries with aging populations continues to expand, this need becomes ever more urgent. In response to this need, and bolstered by recent therapeutic development successes, the number of clinical trials in the AD drug development pipeline across all phases of development (1, 2, and 3) increased from 172 to 187 between 2022 and 2023, representing an 8.7% increase year over year [2,3]. While encouraging, this growth also introduces new challenges. As new therapeutic trials are launched, the number of participants required to meet new and ongoing trial enrollment targets will rise by more than 57 thousand participants [3]. Modernizing trial processes and procedures to accommodate this growth will be critical to ensure success. Traditional approaches to ensure the safety of trial participants during and after exposure to these novel interventions are one example of processes ripe for transformation.

Adverse Events (AEs) are "any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have a causal relationship with this treatment" [4]. AEs are collected, assessed, classified, and scrutinized as part of a systematic approach to trial safety monitoring. In therapeutic trials, these activities contribute critical information to ensure study participant safety and establish the safety profile of an investigational product. While methodological issues related to the analysis of safety data remain an area of debate among clinical trialists [5,6], the

* Corresponding author.
E-mail address: gustavoj@usc.edu (G.A. Jimenez-Maggiora).

https://doi.org/10.1016/j.tjpad.2024.100002

Available online 1 January 2025

AE data used to support safety analyses and reporting, which combine structured and unstructured (free-text) data elements, must be manually coded and reviewed by expert clinicians using standardized coding dictionaries, such as the Medical Dictionary for Regulatory Activities (MedDRA) [7]. The high cost, time delays, and potential for bias inherent in these manual coding processes present an opportunity to explore the use of novel natural language processing and computational classification methods [8]. Effective use of these methods may help investigators address these challenges.

Text classification involves assigning categories to text, like a librarian sorting books by genre. With the rapid digitization of textual data, computational classification methods have emerged. These methods begin by preparing the text, often breaking it into discrete sentences or words (tokenization). Traditional methods use set rules to match keywords for classification. While effective and easy to understand, these methods don't generalize well to new data. To overcome this shortcoming, investigators proposed using supervised statistical learning methods to train predictive models on pre-classified text data. More advanced methods, such as neural networks, learn complex patterns from data and can outperform older techniques. Unlike traditional rule-based methods, however, these methods are sensitive to the size and distribution of the data on which they are trained.

A common approach in text classification involves dividing the problem into multiple subtasks, each building on the results of the previous one. One frequent subtask is named entity recognition (NER), which identifies and annotates segments of text with categories like person names, organization names, locations, dates and times, currency values, and percentages. In medical NLP applications, these categories might be diseases, symptoms, treatments, and medications.

A rich literature has developed in the fields of biomedical informatics and natural language processing (NLP) that assess the use of computational methods to classify medical text containing information about patient Adverse Events (AEs) and adverse drug reactions (ADRs). These studies evaluate the use of information extraction (IE) approaches, such as named entity recognition (NER) and relationship extraction (RE), to extract structured data from unstructured textual information [9].

Recent studies detail NLP-based systems achieving high performance (measured in terms of classification precision, recall, and F1 score) in conducting ADR-related NER and RE tasks on medical text from multiple data sources. These sources include electronic health records (EHR) [10–22], EHR patient messages [23], social media [24], and spontaneous ADR reporting systems [19,25–29]. The classification approaches used in these studies encompass rule-based techniques, traditional supervised machine learning algorithms, and deep learning or neural network learning methods. The range of NER performance measures reported in these studies (precision: .51-.80; recall: .63-.83; F1 Score: .51-.94) provides evidence of the potential utility of these approaches as well as the need for further investigation.

Despite these advances, the empirical study of computational methods for AE classification in clinical trials has been limited. To our knowledge, the only studies of this type published recently have been our preliminary work focused on Alzheimer's disease (AD) clinical studies [30–32].

Given this gap in the extant literature and the challenges mentioned above, there is an urgent need to continue the exploration of these methods for use in clinical trials safety monitoring. The purpose of our study was to evaluate the use of NLP and artificial intelligence (AI) methods to streamline and standardize clinician coding of adverse event data in Alzheimer's disease clinical trials. More specifically, our study aimed to 1) develop a gold standard Alzheimer's disease adverse event data set to facilitate the assessment of coding models, 2) evaluate the predictive performance of a set of NLP-based models to classify adverse events into analyzable codes, and 3) determine whether automated coding improves efficiency, accuracy, reliability, and consistency relative to clinician coding.

2. Methods

2.1. Gold Standard Adverse Event Data Set

To build a gold standard AE data set, we collected demographic and adverse event data from eight completed mid-to-late phase (2, 2/3, and 3) clinical trials in participants with symptomatic AD conducted between 2005 and 2020 (protocol name [protocol initials], start date, ClinicalTrials.gov ID): A) Docosahexaenoic Acid (DHA), 2007, NCT00440050; B) FYN kinase inhibitor AZD0530 (FYN), 2014, NCT02167256; C) Homocysteine (HC), 2003, NCT00056225; D) Intranasal Insulin (INI), 2014, NCT01767909; E) Simvastatin (LL), 2002, NCT00053599; F) CERE-110 Nerve Growth Factor (NGF), 2008, NCT00876863; G) Resveratrol (RES), 2012, NCT01504854; H) Valproate Neuroprotection (VN), 2003, NCT00071721.

These data included participant age and sex, the textual descriptions of each AE, and the associated expert clinician-confirmed MedDRA lower-level term (LLT) codes (see Table 1). These codes were used for all subsequent comparisons. MedDRA version 26.0 (published March 2023) was used as the medical coding reference in our study.

Data (studies: A, C, E, G, H) used in the preparation of this manuscript/publication/article were obtained from the University of California, San Diego Alzheimer's Disease Cooperative Study Legacy database (www.adcs.org). Data for studies B, D, and F were obtained from the University of Southern California's Alzheimer's Therapeutic Institute (atri.usc.edu).

2.2. Exploratory Analysis

An exploratory analysis of the gold standard AE data set was conducted to inform model development. The results of this analysis guided the selection of predictors (feature selection), predictor transformations (feature engineering), modeling techniques, and parameter settings. The relevant details are described in the sections below.

2.3. Hold-out Test Data Set

Study A was held out from the model development process. This selection, which allowed us to maximize the size of the training set while ensuring a sufficiently large holdout data set, was informed by our exploratory analysis, which indicated that the distribution of LLT codes between studies was highly consistent (see supplementary information - Figure S1). A random sample (without replacement) of 1000 AEs from Study A was re-coded by a clinician and the best-performing model to support the final performance assessment of each approach. The original expert clinician-confirmed LLT codes were used for all comparisons. The performance metric selected for final performance evaluation was classifier accuracy relative to the original LLT codes.

An analysis of the intercoder reliability (ICR) between the clinician and the best-performing classification model was performed using Cohen's Kappa statistic [49].

2.4. Classification Model Development

A comprehensive approach was taken to classification model development that considered multiple modeling approaches: Rule-based (Levenshtein Distance Algorithm (LDA) [33], machine learning (Multinomial Logistic Regression [34], Light Gradient Boosting Machine (LightGBM) [35], Random Forests (RF) [36]), deep learning using transformer-based architectures (BERT [37], BioBERT [38], PubMedBERT [39], Bioformer [40]), and hybrid approaches that combine machine learning and Rule-based methods [41].

The model development data set (studies: B-H) (n=10208) was randomly split 80%/20% into training and testing sets stratified by LLT code.

Table 1
Trial cohorts are balanced by age and sex. A large number of unique codes are observed within and between studies. Most adverse events are described using short phrases (3-6 words). The most common codes are similar across trials and match those typically seen in an elderly population.

Study	Participants	Female	Age		Adverse Events	Unique Codes	Words per Event Description		Most Frequent Codes
	N	*N* (%)	*M* + *SD*	*Range*	*N*	*N*	*M* + *SD*	*Range*	*Code* (%)
A	360	187 (51.9%)	76.49 ± 8.49	50–91	1,517	573	2.79 ± 2.24	1.00–20.00	- Fall (6.5%) - Urinary tract infection (2.6%) - Agitation (2.2%) - Diarrhea (1.6%) - Anxiety (1.5%)
B	128	62 (48.4%)	71.13 ± 7.65	54–85	497	231	2.63 ± 1.97	1.00–14.00	- Diarrhea NOS (5.0%) - Fall (4.8%) - Urinary tract infection NOS (3.8%) - Nausea (3.4%) - Headache (3.0%)
C	385	209 (54.3%)	76.32 ± 8.07	54–94	2,330	608	3.25 ± 3.23	1.00–31.00	- Fall (6.5%) - Depressed mood (3.4%) - Loss of energy (3.1%) - Drowsiness (3.0%) - Diarrhea (2.7%)
D	214	103 (48.1%)	70.54 ± 6.82	55–84	711	344	2.80 ± 1.97	1.00–14.00	- Headache (4.1%) - Urinary tract infection NOS (3.5%) - Fall (2.7%) - Upper respiratory tract infection NOS (2.3%) - Post lumbar puncture syndrome (1.7%)
E	370	222 (60.0%)	73.94 ± 9.34	50–93	2,353	539	2.84 ± 2.76	1.00–24.00	- Fall (6.8%) - Agitation (4.5%) - Loss of energy (3.8%) - Anxiety (3.7%) - Diarrhea (3.6%)
F	51	23 (45.1%)	68.47 ± 6.09	56–79	747	357	2.65 ± 2.16	1.00–20.00	- Headache (4.6%) - Urinary tract infection (3.7%) - Fall (3.2%) - Nausea (2.3%) - Agitation (1.9%)
G	116	65 (56.0%)	70.86 ± 7.98	51–89	362	173	2.52 ± 2.10	1.00–16.00	- Headache (9.7%) - Hypertension (6.6%) - Rash (3.6%) - Chills (2.8%) - Infusion site infiltration (2.8%)
H	296	174 (58.8%)	75.59 ± 7.86	53–89	3,208	570	2.62 ± 2.55	1.00–36.00	- Confusion (7.0%) - Fall (5.8%) - Agitation (3.8%) - Drowsiness (3.6%) - Loss of energy (2.7%)
Total	1,920	1,045 (54.4%)	72.92 ± 7.84	50–94	11,725	1,762	2.76 ± 2.41	1.00–36.00	- Fall (5.7%) - Agitation (2.8%) - Drowsiness (2.4%) - Loss of energy (2.3%) - Confusion (2.3%)

Rule-based, machine-learning, and hybrid models were implemented with R statistical software (v4.3.1; R Core Team 2023) [42]. Deep learning models were implemented with Python (v3.11.7) [43]. Transformer-based models were implemented with the HuggingFace framework [44]. Final analyses were performed with R statistical software (v4.3.1; R Core Team 2023) [42].

2.4.1. Rule-based Approach: Normalized Levenshtein Distance Algorithm

The rule-based approach was based on the Levenshtein Distance Algorithm (LDA) [33]. The LDA method, also known as the edit distance, measures the similarity between strings of text characters by calculating the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string to another. Lower values indicate a higher level of similarity, with an LDA score of zero indicating identical strings. The normalized LDA (NLDA) is a variation of the Levenshtein Distance Algorithm calculated by dividing the LDA score between two strings by the length of the longer string. The resulting score falls between 0 and 1, where 0 indicates identical strings. The NLDA is used to compare strings, regardless of individual length. In our study, LDA and NLDA scores were calculated using the stringdist (v0.9.1) R package [45].

For each AE textual description in the model development data set, our NLDA-based coding algorithm performed the following steps:

1. Punctuation, numbers, and irrelevant white spaces were removed.
2. The text was broken down into words (tokens) and common meaningless words (stopwords) such as prepositions, articles, and conjunctions were removed.
3. The remaining words (tokens) were combined into unique forward and backward n-grams (n=[1,3]) with n_max=3 and n_min=min(token count, 2).
4. Each n-gram was compared to every LLT code to generate an NLDA score.
5. The NLDA scores were sorted by rank and the LLT with the lowest score was assigned. In the event of a tie, the code corresponding to the LLT with the longest length was assigned.

2.4.2. Machine Learning Approach

The machine learning approach considered three algorithms: Multinomial Logistic Regression (MLR) [34], Light Gradient Boosting Machine (LightGBM) [35], and Random Forests (RF) [36]. Each of these algorithms has parameters that were tuned using grid search. 10-fold cross-validation was used to control for bias during parameter tuning. The following R packages were used for model implementation: MLR - glmnet (v4.1-8) [34], LightGBM - lightgbm (v3.3.5) [46], RF - ranger (v0.15.1) [47].

To address the large number of rare LLT codes present in the model development data set, rare LLT codes with a prevalence of less than 0.3% were aggregated into a pooled "other" class (1352/1411, 95.82%). The remaining 59 LLT codes (59/1411, 4.18%) represented less than 5% of the LLT codes but were associated with 59.50% of the total AEs in the training data set (4859/8166).

To improve computability, the following steps were performed on the raw AE textual descriptions in the model development data set:

1. Punctuation, numbers, and irrelevant white spaces were removed.
2. The text was broken down into words (tokens) and common meaningless words (stopwords) such as prepositions, articles, and conjunctions were removed.
3. Tokens that appeared frequently (more than 200 times) were excluded.
4. The remaining tokens were transformed into embeddings using the term frequency-inverse document frequency (TF-IDF) method.
5. Tokens with zero variance (i.e., appeared only once) were removed.
6. The remaining embeddings were standardized and normalized.

2.4.2.1. Participant Age and Sex as Additional Predictors. In addition to using the AE textual description as a predictor, we also examined models that included participant age (binned in 5-year intervals) and sex (dichotomized) as additional predictors. These variables were selected based on input from our clinician-scientist co-authors to improve the clinical validity and meaningfulness of the resulting predictions.

2.4.3. Deep Learning with Transformers

The deep learning approach considered five Transformer-based large language models: Bidirectional Encoder Representations from Transformers (BERT) (cased and uncased) [37], as well as BERT-derived models adapted to the biomedical domain: BioBERT [38], PubMedBERT [39], and Bioformer [40]. These models were fine-tuned on our training data using manual optimization based on the hyperparameters recommended by the BERT authors [37] to achieve the best results (batch size = 8, learning rate = 2e-5, number of epochs = 4).

To address the large number of rare LLT codes present in the model development data set, rare LLT codes with a prevalence of less than 0.3% were aggregated into a pooled "other" class (see prior section for details). The AE textual descriptions were processed using a model-specific workflow that included tokenization, text encoding, and padding/truncation steps. This workflow was implemented using the transformers (v4.36.1) Python package.

2.4.4. Hybrid Approach

A hybrid approach that combines multiple modeling approaches can achieve performance improvements relative to the performance of the individual component models [41]. Given the limited size of the model development data set, we hypothesized that a hybrid approach that combined the rule-based NLDA and the best-performing predictive model would outperform its underlying component models.

By making use of the fact that predictive models can generate both class labels and probabilities, we developed a hybrid approach with a decision rule that considered the level of confidence that the model assigns to each class label. The decision rule is defined below.

To classify each AE textual description in the model development data set, our hybrid algorithm performed the following steps:

1. Predicted the LLT code and probability (confidence) score using the best-performing predictive model.
2. Predicted the LLT code and NLDA score using the NLDA-based model.
3. Decision Rule: If the predicted probability score was low (≤ 0.1; i.e., low confidence in predicted LLT code) and the NLDA score was low (≤ 0.1; i.e., high similarity between strings), the predicted LLT code was replaced with the NLDA-based LLT code.

2.4.5. Coding Effort and Costs

We estimated the total cost of coding by assuming the following component costs (Equation 1):

$$TotalCodingCost = CodingCost + ReviewCost + ErrorResolutionCost \quad (1)$$

Each of the component costs was calculated as follows:

$$CodingCost = NumberOfAdverseEvents * AverageCodingTime(min) * WageClinician(min) \quad (2)$$

$$CodingCostPerAdverseEvent = \frac{CodingCost}{NumberOfAdverseEvents} \quad (3)$$

$$ReviewCost = NumberOfAdverseEvents * AverageReviewTime(min) * WageSeniorClinician(min) \quad (4)$$

$$ErrorResolutionCost = NumberOfAdverseEvents * ErrorRate * CodingCostPerAdverseEvent \quad (5)$$

2.4.5.1. Hourly Wages

To estimate the hourly wages for clinicians and senior clinicians, we referenced the most recent wage survey data from the U.S. Department of Labor [48]. Based on this survey, the estimated national median hourly wage for a physician (OC: 29-1210) in 2022 was $109.22. Informed by this data, we estimated the hourly wage for a senior clinical in the same period was 50% higher, or $163.83.

Similarly, the median hourly wage for computer and information research scientists (OC: 15-1221) in 2022 was estimated to be $65.69 [48].

2.4.5.2. Coding and Review Time. The holdout test data set (n=1000) was re-coded by a clinician in five batches of 200 events. The time to re-code each batch was averaged and divided by 1000 to estimate the average coding time per AE (21.24 seconds). Based on input from our clinician co-authors, senior clinician review time was estimated to be 10% of the time required by a clinician to code an AE.

2.5. Model Selection Criteria

To mitigate the effects of class imbalance inherent in AE data, we selected the F1 score (Equation 6), which is the harmonic mean of a classifier's precision (positive predictive value) and recall (sensitivity), as the primary model selection metric.

$$F1\ Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (6)$$

The advantage of the F1 Score is that it provides a balanced evaluation of a classifier's performance by considering both precision and recall, which focus on the accuracy and completeness of positive predictions, respectively. This property makes it a useful measure for performance comparisons across multiple classifiers.

In addition to raw F1 Score performance, we also considered the clinical meaningfulness of the model as an additional selection criterion.

Table 2
Machine Learning Models Outperform Other Approaches (model performance ranked by F1 Score). Model types: Rule-based: lev_distance; Machine Learning: lgbm, mlr, rf; Neural Networks/Deep Learning with Transformers: bert-base-cased, bert-base-uncased, biobert, bioformer, pubmedbert.

Model	Accuracy	F1 Score	Precision	Recall
lgbm	0.917	0.885	0.909	0.823
lgbm_age	0.914	0.879	0.899	0.818
rf_age	0.915	0.874	0.872	0.837
rf_age_sex	0.913	0.873	0.872	0.835
lgbm_sex	0.909	0.867	0.869	0.867
rf_sex	0.919	0.863	0.867	0.853
bert-base-uncased	0.933	0.849	0.845	0.868
rf	0.917	0.848	0.855	0.848
bert-base-cased	0.929	0.846	0.848	0.858
lgbm_age_sex	0.912	0.846	0.849	0.848
biobert	0.934	0.843	0.837	0.866
mlr_sex	0.902	0.842	0.865	0.779
mlr_age	0.902	0.838	0.868	0.789
lev_distance	0.649	0.837	0.369	0.445
pubmedbert	0.933	0.836	0.828	0.860
mlr_age_sex	0.903	0.831	0.869	0.793
mlr	0.899	0.812	0.832	0.789
bioformer	0.901	0.681	0.702	0.701

3. Results

3.1. Exploratory Analysis

Descriptive statistics of the overall and trial-level cohorts are presented in Table 1. The gold standard AE data set contains information on the AEs observed in 1920 trial participants across eight trials, roughly half of whom were female (54.4%) with an average age of 72.92 (SD=7.84; Range=[50, 94]) at the time of consent. Overall, 11725 adverse events were reported and assigned to 1762 unique LLT codes. The average number of words per AE description was 2.8 ± 2.4 (range: 1-36).

Consistent with a symptomatic AD trial population, the most common LLT codes were (in descending order): Fall (5.7%), Agitation (2.8%), Drowsiness (2.4%), Loss of energy (2.3%), Confusion (2.3%), Diarrhea (2.2%), Headache (2.0%), Depressed mood (2.0%), Anxiety (2.0%), and Dizziness (1.7%). The overall distribution of LLT codes was heavily skewed toward rare codes (n<10). The distributions of LLT codes across trials were consistently bimodal with similar peaks (see supplementary information - Figure S1). These findings suggested that this imbalance was representative of our population distribution. Therefore, rather than employing resampling methods, we opted to select a modeling approach that was resilient to a large number of infrequent classes and imbalanced data.

3.2. Model Development

Model performance metrics (accuracy, F1 score, precision, and recall) on the model development test set are presented in Table 2. The NLDA performed relatively poorly (F1 score=0.837) compared with the machine learning and deep learning approaches. Among the machine learning models, the LightLGBM model (lgbm) had the highest F1 score (0.885). The best-performing deep learning model was the base BERT uncased model (bert-base-uncased) (F1 score=0.849). The classification accuracy for these models was 0.649, 0.917, and 0.933, respectively. Upon further discussion with our clinician co-authors regarding the importance of clinical meaningfulness, we selected the best-performing model that included participant age and sex as additional predictors. This model was the Random Forest model with participant age and sex as additional predictors (rf_age_sex), which achieved an F1 score and accuracy of 0.873 and 0.913, respectively.

Table 3
AI Models Outperform Clinician Coding (model performance ranked by Accuracy).

Approach	Accuracy	F1 Score	Precision	Recall
Hybrid AI	0.884	0.954	0.931	0.780
AI	0.830	0.934	0.934	0.675
Clinician	0.706	0.889	0.606	0.600

3.3. Re-coding Process and Performance

The holdout data set (n=1000) was re-coded by a clinician and the selected model (rf_age_sex). During the clinician re-coding, several data quality issues were identified and referred to the study team for review. These issues included uncorrected typos, compound AEs, and suspected but unconfirmed AEs. Upon further discussion within the study team, twenty of these problematic AEs were removed from the final evaluation data set (n=980).

We also re-coded the holdout data set using a hybrid algorithm (Hybrid AI) that combined the results of the selected (rf_age_sex) and NLDA models using the approach described in the Methods section.

The performance of each of the three approaches is summarized in Table 3. The Hybrid AI model outperformed the AI model and clinician-based coding decisively, achieving the highest accuracy (0.887). This performance represents a 26.3% improvement in accuracy relative to the clinician coder (accuracy=0.706).

Intercoder reliability (ICR) between the Hybrid AI and AI models was almost perfect (Cohen's Kappa=0.937), while the agreement between the clinician and Hybrid AI models (Cohen's Kappa=0.431) and the clinician and AI models (Cohen's Kappa=0.393) was minimal.

Grouping the LLT codes by frequency provides further evidence of Hybrid AI and AI model outperformance relative to clinician coding. Figure 1 presents the accuracy of each approach across the LLT code frequency range (from rare to common) and shows that the AI and Hybrid AI models outperform the clinician coder performance across the entire range. At the high end of the range (i.e., the most frequent LLT codes), both AI models consistently achieve 100% accuracy while the clinician struggles to surpass 60% accuracy (see Figure 1). At the low end of the range (i.e., the infrequent LTT codes), Hybrid AI and AI model performance remain above 60%, while clinician performance falls below 50%. Comparing the AI and Hybrid AI models, the Hybrid AI model matches or surpasses the AI model's performance across the full range with marked outperformance in the low end of the range (see Figure 1).

A possible explanation for the Hybrid AI model's outperformance relative to the AI model may be found in the results of an analysis of intercoder reliability (ICR) between the AI and NLDA models. The high accuracy of each approach (NLDA accuracy=0.649, rf_age_sex accuracy=0.913) combined with the poor level of reliability between approaches (Cohen's Kappa=0.393) suggests that their performance may be complementary. Further analysis will be required to fully decompose the sources of performance.

3.4. Re-coding Cost Analysis

The costs of AI and clinician-based re-coding of the holdout data set (n=1000) are presented in Table 4. The estimated total costs of each approach are $161.10 and $902.16, respectively, yielding cost savings of more than 80%. The main sources of cost savings are 1) the AI approach has negligible initial coding costs, and 2) the AI approach's estimated 50% reduction in coding error rate (estimated by averaging the AI and Hybrid AI error rates) relative to the clinician error rate (15% vs. 30%) yields lower costs related to error resolution.

3.4.1. Model Development Costs

Model development required 50 hours of effort by a computer scientist ($65.69 hourly wage), yielding an estimated total model develop-

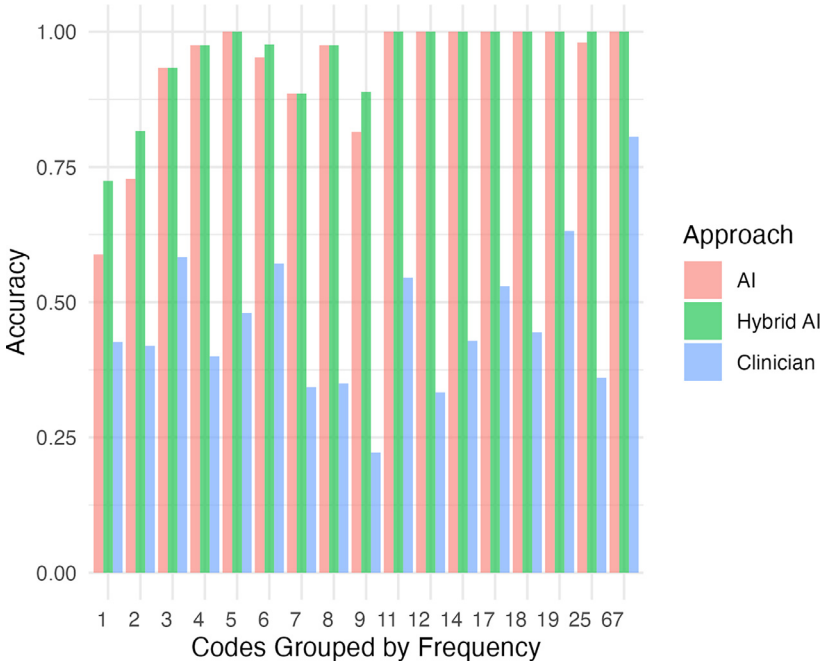

Figure 1. AI Models Consistently Outperform Clinician Coding Across the Code Frequency Range (from rare to common).

Table 4
AI Coding Significantly Reduces Costs Relative to Clinician Coding. AI cost savings were driven by the reduction of initial coding effort and the increased accuracy of the initial coding process.

	Clinician	AI
Recoded Adverse Events (N)	1000	1000
Coding Cost		
Average Coding Time (sec)	21.24	0
Clinician Wage ($)	109.22	0
Clinician Recoding Cost ($)	644.40	0
Review Cost		
Average Review Time (sec)	2.124	2.124
Senior Clinician Wage ($)	163.83	163.83
Senior Clinician Review Cost ($)	64.44	64.44
Error Resolution Cost		
Error Rate (%)	30	15
Average Coding Cost per Adverse Event ($)	0.64	0.64
Error Resolution Cost ($)	193.32	96.66
Total Cost		
Total Cost ($)	902.16	161.10

ment cost of $3,284.50. This cost was incurred before re-coding, so we consider it a sunk cost excluded from our cost model.

4. Discussion

Our study results provide supportive evidence that AI-based adverse event coding approaches can improve coding efficiency, accuracy, reliability, and consistency relative to clinician coding, thus reducing costs and improving the availability and quality of data for clinical trial safety monitoring and reporting. In our study, AI-based AE coding achieved higher accuracy (~20% increase in accuracy) and was more cost-effective (~80% cost reduction) than traditional clinician coding. Furthermore, the accuracy of AI-based AE coding consistently outperformed clinician coding across the LLT code frequency range (from rare to common codes), with AI accuracy outperformance consistently surpassing 20%, and in some cases exceeding 50%. Additional performance gains were observed by combining classic rule-based and AI-based methods to implement a Hybrid AI approach (+5% increase in accuracy). Increased access to AE data and methodological improvements via the use of novel natural language processing and modeling methods (e.g., Transformers, Large Language Models) may further improve artificial intelligence coding accuracy.

Our results do not negate or undermine the continued importance of expert clinician coding review and issue resolution, which remain critical steps in the AE coding process. Rather, AI technologies may serve as tools to produce accurate coding that, when integrated into clinical trial data management systems [49], can be rapidly reviewed by expert clinicians. Focusing resources and expertise on this phase of the safety monitoring process by utilizing AI methods to automate initial AI coding represents an important efficiency that can improve trial participant safety monitoring and reporting. Future work using community-engaged mixed methods exploratory research designs will assess the acceptability, feasibility, usability, efficacy, and trustworthiness of integrating AI-enabled coding workflows into trial safety monitoring programs.

4.1. Limitations

Our study has several limitations that we will address in future work. First, our data set is limited to academic clinical trials in symptomatic AD. In future work, we plan to expand our data set to include novel trial cohorts focused on the pre-symptomatic stages of Alzheimer's disease (e.g., A4/LEARN studies) and industry-sponsored trials. This work should allow us to assess the external validity of our results. Preliminary efforts toward this objective are underway.

Second, the high number of infrequent LLT codes present in our model development data set necessitated the use of a class pooling method that collapsed 95.8% (1352/1411) of the LLT codes into an "other" class. The remaining 59 LLT codes (59/1411, 4.2%) were associated with 59.5% of the total AEs in the training data set (4859/8166). In future work, we plan to expand our data set to increase the number of LLT codes that are well-represented and maximize our use of all available data.

Third, but related, the size of our data set was insufficient to properly assess the performance of the deep learning class of models, which in some cases, achieved high accuracy but middling F1 Scores. In future work, we plan to train models on larger data sets that may yield classification performance metrics that align with the state-of-the-art results reported in the literature.

Finally, while our study explored the effects of age and sex as predictors, the lack of participants from underrepresented sociodemographic

groups in the trial data sets used in our study limited our ability to explore the predictive effects of these factors. Other potentially important predictors include measures of cognitive impairment, genetic risk, biomarkers, and concomitant medications. In future work, we hope to gain access to AD trial cohorts that have more diverse and deeply characterized participants to examine these questions.

4.2. Conclusions

The recent growth in AD therapeutic development programs gives us renewed confidence that effective and safe interventions that can bring relief to the millions of individuals and families around the world suffering from AD are on the horizon. Our study results demonstrate how approaches that effectively combine AI and human expertise can positively contribute to these efforts.

Declaration of competing interest

GJM has received research support from the National Institutes of Health (NIH), the Alzheimer's Association, the American Heart Association, Gates Ventures, Eli Lilly, and Eisai.

MD has research grants from Eisai and Lilly, consults with Roche, and owns stock in Janssen. His wife is employed by Janssen.

MR has research grants from Eisai and Lilly, consults with AC Immune and Ionis, and serves on an advisory board for Alzheon, Aptah Bio, Biohaven, Keystone Bio, Prescient Imaging, Positrigo, and Embic.

RR has research grants from Eisai, Lilly, the American Heart Association, and the Alzheimer's Association.

PA has research grants from NIH, Lilly, and Eisai, and consults with Merck, Roche, BMS, Genentech, Abbvie, Biogen, ImmunoBrain Checkpoint, Arrowhead, AltPep, and Neurimmune.

Acknowledgments

MedDRA® the Medical Dictionary for Regulatory Activities terminology is the international medical terminology developed under the auspices of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). MedDRA® trademark is registered by IFPMA on behalf of ICH.

Data used in preparation of this manuscript/publication/article were obtained from the University of California, San Diego Alzheimer's Disease Cooperative Study (www.adcs.org)

Funding

Data collection and sharing for this project (studies: A, C, E, G, H) was funded by the University of California, San Diego Alzheimer's Disease Cooperative Study (ADCS) (National Institute on Aging Grant Number U01AG010483).

Data collection and sharing for this project (studies: B, D, F) was funded by the National Institute on Aging (Grant Numbers UH3TR000967, RF1AG041845, R01AG030048) and the University of Southern California's Alzheimer's Therapeutic Institute (ATRI).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tjpad.2024.100002.

References

[1] 2022 Alzheimer's disease facts and figures. Alzheimer's & Dementia 2022;18(4):700–89. doi:10.1002/alz.12638.
[2] Cummings J, Lee G, Nahed P, et al. Alzheimer's disease drug development pipeline: 2022. Alzheimer's & Dementia: Translational Research & Clinical Interventions 2022;8(1):e12295. doi:10.1002/trc2.12295.
[3] Cummings J, Zhou Y, Lee G, Zhong K, Fonseca J, Cheng F. Alzheimer's disease drug development pipeline: 2023. Alzheimer's & Dementia: Translational Research & Clinical Interventions 2023;9(2):e12385. doi:10.1002/trc2.12385.
[4] European Medicines Agency. Guideline for good clinical practice E6(R2). Published online 2016. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-6-r2-guideline-good-clinical-practice-step-5_en.pdf
[5] Phillips R, Hazell L, Sauzet O, Cornelius V. Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open 2019;9(2):e024537. doi:10.1136/bmjopen-2018-024537.
[6] Singh S, Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. Trials 2012;13(1):138. doi:10.1186/1745-6215-13-138.
[7] Brown EG, Wood L, Wood S. The Medical Dictionary for Regulatory Activities (MedDRA). Drug Safety 1999;20(2):109–17. doi:10.2165/00002018-199920020-00002.
[8] Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. International Journal of Medical Informatics 2019;132:103971. doi:10.1016/j.ijmedinf.2019.103971.
[9] Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing. Computational Linguistics, and Speech Recognition. 2nd ed. Pearson Prentice Hall; 2009.
[10] Alfattni G, Belousov M, Peek N, Nenadic G. Extracting drug names and associated attributes from discharge summaries: Text mining study. JMIR Medical Informatics 2021;9(5). doi:10.2196/24678.
[11] Borjali A, Magnéli M, Shin D, Malchau H, Muratoglu OK, Varadarajan KM. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation. Computers in Biology and Medicine 2021;129. doi:10.1016/j.compbiomed.2020.104140.
[12] Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting adverse drug events with rapidly trained classification models. Drug Safety 2019;42(1):147–56. doi:10.1007/s40264-018-0763-y.
[13] Christopoulou F, Tran TT, Sahu SK, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. Journal of the American Medical Informatics Association 2020;27(1):39–46. doi:10.1093/jamia/ocz101.
[14] Gupta S, Belouali A, Shah NJ, Atkins MB, Madhavan S. Automated identification of patients with immune-related adverse events from clinical notes using word embedding and machine learning. JCO clinical cancer informatics 2021;5:541–9. doi:10.1200/CCI.20.00109.
[15] Iqbal E, Mallah R, Rhodes D, et al. ADEPt, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health records. PLoS ONE 2017;12(11). doi:10.1371/journal.pone.0187121.
[16] Karhade AV, Bongers MER, Groot OQ, et al. Can natural language processing provide accurate, automated reporting of wound infection requiring reoperation after lumbar discectomy? Spine Journal 2020;20(10):1602–9. doi:10.1016/j.spinee.2020.02.021.
[17] Santiso S, Pérez A, Casillas A. Smoothing dense spaces for improved relation extraction between drugs and adverse reactions. International Journal of Medical Informatics 2019;128:39–45. doi:10.1016/j.ijmedinf.2019.05.009.
[18] Taggart M, Chapman WW, Steinberg BA, et al. Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients. JAMA network open 2018;1(6):e183451. doi:10.1001/jamanetworkopen.2018.3451.
[19] Wang L, Rastegar-Mojarad M, Ji Z, et al. Detecting pharmacovigilance signals combining electronic medical records with spontaneous reports: A case study of conventional disease-modifying antirheumatic drugs for rheumatoid arthritis. Frontiers in Pharmacology 2018;9(AUG). doi:10.3389/fphar.2018.00875.
[20] Wei Q, Ji Z, Li Z, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. Journal of the American Medical Informatics Association 2020;27(1):13–21. doi:10.1093/jamia/ocz063.
[21] Xu D, Yadav V, Bethard S. UArizona at the MADE1.0 NLP Challenge. Proceedings of Machine Learning Research 2018;90:57–65.
[22] Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. Journal of the American Medical Informatics Association 2020;27(1):65–72. doi:10.1093/jamia/ocz144.
[23] Chen J, Lalor J, Liu W, et al. Detecting hypoglycemia incidents reported in patients' secure messages: Using cost-sensitive learning and oversampling to reduce data imbalance. Journal of Medical Internet Research 2019;21(3). doi:10.2196/11990.
[24] Magge A, Tutubalina E, Miftahutdinov Z, et al. DeepADEMiner: A deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on twitter. Journal of the American Medical Informatics Association 2021;28(10):2184–92. doi:10.1093/jamia/ocab114.
[25] Kreimeyer K, Dang O, Spiker J, et al. Feature engineering and machine learning for causality assessment in pharmacovigilance: Lessons learned from application to the FDA adverse event reporting system. Computers in Biology and Medicine 2021;135. doi:10.1016/j.compbiomed.2021.104517.
[26] Kreimeyer K, Menschik D, Winiecki S, et al. Using probabilistic record linkage of structured and unstructured data to identify duplicate cases in spontaneous adverse event reporting systems. Drug Safety 2017;40(7):571–82. doi:10.1007/s40264-017-0523-4.
[27] Sutphin C, Lee K, Yepes AJ, Uzuner Ö, McInnes BT. Adverse drug event detection using reason assignments in FDA drug labels. Journal of Biomedical Informatics 2020;110. doi:10.1016/j.jbi.2020.103552.
[28] Ujiie S, Yada S, Wakamiya S, Aramaki E. Identification of adverse drug eventrelated japanese articles: Natural language processing analysis. JMIR Medical Informatics 2020;8(11). doi:10.2196/22661.

[29] Usui M, Aramaki E, Iwao T, Wakamiya S, Sakamoto T, Mochizuki M. Extraction and standardization of patient complaints from electronic medication histories for pharmacovigilance: Natural language processing analysis in japanese. JMIR Medical Informatics 2018;20(9). doi:10.2196/11021.
[30] Jimenez-Maggiora G, Raman R, Ernstrom K, Rafii MS, Aisen PS. Automated classification of adverse events in clinical studies of alzheimer's disease. Published online December 8, 2016. https://www.ctad-alzheimer.com/files/files/PROGRAMfinal_CTAD2016_1dec.pdf
[31] Ravindranath P, Bruschi S, Ernstrom K, et al. Machine learning in automated classification of adverse events in clinical studies of alzheimer's disease. Published online July 16, 2017.
[32] Ravindranath P, Raman R, Chow T, Rafii M, Aisen P, Jimenez-Maggiora G. Deep learning in automated classification of adverse events in clinical studies of alzheimer's disease. Published online July 22, 2018.
[33] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Union 1966;10:707–10.
[34] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 2010;33(1):1–22.
[35] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In:; 2017. https://api.semanticscholar.org/CorpusID:3815895
[36] Breiman L. Random forests. Machine Learning 2001;45(1):5–32. doi:10.1023/A:1010933404324.
[37] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Published online 2018. doi:10.48550/ARXIV.1810.04805
[38] Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40. doi:10.1093/bioinformatics/btz682.
[39] Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. Published online 2020. doi:10.48550/ARXIV.2007.15779
[40] Fang L, Chen Q, Wei CH, Lu Z, Wang K. Bioformer: An efficient transformer language model for biomedical text mining.
[41] Berge GT, Granmo OC, Tveit TO, Ruthjersen AL, Sharma J. Combining unsupervised, supervised and rule-based learning: the case of detecting patient allergies in electronic health records. BMC Medical Informatics and Decision Making 2023;23(1):188. doi:10.1186/s12911-023-02271-8.
[42] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2023. https://www.R-project.org/
[43] Van Rossum G, Drake FL. Python 3 Reference Manual. CreateSpace 2009.
[44] Wolf T, Debut L, Sanh V, et al. HuggingFace's transformers: State-of-the-art natural language processing.
[45] der Loo MPJ van. The stringdist package for approximate string matching. The R Journal 2014;6(1):111–22. https://CRAN.R-project.org/package=stringdist.
[46] Shi Y, Ke G, Soukhavong D, et al. Lightgbm: Light Gradient Boosting Machine.; 2023. https://CRAN.R-project.org/package=lightgbm
[47] Wright MN, Ranger Ziegler A. A fast implementation of random forests for high dimensional data in c++ and r. Journal of Statistical Software 2017;77(1):117. doi:10.18637/jss.v077.i01.
[48] U. S. Bureau of Labor Statistics. May 2022 National Occupational Employment and Wage Estimates. Published online 2022. https://www.bls.gov/oes/2022/may/oes_nat.htm
[49] Jimenez-Maggiora GA, Bruschi S, Qiu H, So JS, Aisen PS. ATRI EDC: a novel cloud-native remote data capture system for large multicenter Alzheimer's disease and Alzheimer's disease-related dementias clinical trials. JAMIA Open 2022;5(1):ooab119. doi:10.1093/jamiaopen/ooab119.