communications medicine

Article

https://doi.org/10.1038/s43856-024-00631-7

A natural language processing approach to detect inconsistencies in death investigation notes attributing suicide circumstances

Check for updates

Song Wang[1], Yiliang Zhou[2], Ziqiang Han[3], Cui Tao[4], Yunyu Xiao[2], Ying Ding[5], Joydeep Ghosh[1] & Yifan Peng[2] ✉

Abstract

Background Data accuracy is essential for scientific research and policy development. The National Violent Death Reporting System (NVDRS) data is widely used for discovering the patterns and causing factors of death. Recent studies suggested the annotation inconsistencies within the NVDRS and the potential impact on erroneous suicide-circumstance attributions.

Methods We present an empirical Natural Language Processing (NLP) approach to detect annotation inconsistencies and adopt a cross-validation-like paradigm to identify possible label errors. We analyzed 267,804 suicide death incidents between 2003 and 2020 from the NVDRS. We measured annotation inconsistency by the degree of changes in the F-1 score.

Results Our results show that incorporating the target state's data into training the suicide-circumstance classifier brings an increase of 5.4% to the F-1 score on the target state's test set and a decrease of 1.1% on other states' test set.

Conclusions To conclude, we present an NLP framework to detect the annotation inconsistencies, show the effectiveness of identifying and rectifying possible label errors, and eventually propose an improvement solution to improve the coding consistency of human annotators.

Plain Language Summary

Data accuracy is essential for scientific research and policy development. The National Violent Death Reporting System (NVDRS) contains the recording of individual suicide incidents taking place in the United States, and the contributing suicide circumstances. We used a computational method to check the accuracy of NVDRS records. Our method identified and rectified possible errors in labeling within the database. This method could be used to improve the label accuracy in the NVDRS database, enabling more accurate recording and study of suicide circumstances. Improved data recording of suicide circumstances could potentially be used to develop improved approaches to prevent suicide in the future.

In recent years, the United States (U.S.) has experienced a concerning increase in suicide deaths, marked by an alarming 36% suicide rate rise between 2000 and 2021[1]. Understanding the suicide circumstances is critical and essential for effective interventions and suicide prevention policymaking.

The National Violent Death Reporting System (NVDRS) is a comprehensive surveillance initiative gathering violent fatality data from all 50 U.S. states, the District of Columbia, and Puerto Rico[2]. It meticulously documents information about suicide victims, including demographics and vital social determinants of health (SDoH). The database also contains death investigation notes for each incident, describing the circumstances potentially contributing to the suicide. The NVDRS coded a series of suicide circumstance variables[3], which were manually annotated by human abstractors utilizing the information contained in the death investigation notes[4]. These suicide circumstance variables indicated the presence status of suicide-related social factors (e.g., Family Relationship Crisis, Mental Health Crisis, and Physical Health Crisis). The NVDRS provides a standardized coding manual to maintain data quality and offer routine coding training for annotators (i.e., abstractors). However, it is noteworthy that only 5% of the incident annotations were verified by two independent annotators, leaving an overwhelming 95% of the data reliant on assessments by a single annotator[4]. This absence of a peer verification process increases the risk of annotation inconsistencies at both state and even intra-state levels. Moreover, despite annotators' compliance with the coding guidelines, there still

[1]Cockrell School of Engineering, The University of Texas at Austin, Austin, TX, USA. [2]Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [3]School of Political Science and Public Administration, Shandong University, Qingdao, Shandong, China. [4]Department of AI and Informatics, Mayo Clinic, Jacksonville, FL, USA. [5]School of Information, The University of Texas at Austin, Austin, TX, USA. ✉e-mail: yip4002@med.cornell.edu

exists a potential for annotation inconsistencies due to the possible gaps in expertise and human errors[5].

In our prior research, we developed Natural Language Processing (NLP) methods to extract suicide circumstances from the NVDRS narratives[3]. Our findings highlighted the performance disparities across states and suggested the concerns of inconsistencies in the NVDRS data annotations. Several studies have explored to address data annotation errors in NLP through various approaches[6–13], for example, utilizing conventional probabilistic approaches[14], training machine learning models (e.g., Support Vector Machines)[15–22], and developing generative models via active learning[23]. However, the conventional probabilistic approaches cannot handle infrequent events or compare events with similar probabilities. This is primarily because the probabilities cannot be calculated or compared with high confidence. At the same time, the conventional supervised training paradigm needs high-quality annotated data during the training process. This poses a limitation when applying these methods to the NVDRS dataset, where only 5% of the data were verified by two annotators. However, previous attempts mainly focused on NLP tasks in general domains, such as Part-of-Speech (POS) tagging and Named Entity Recognition (NER). Such approaches cannot be directly applied to identifying mis-labelings in free-text death investigation notes.

This study introduced an empirical NLP approach utilizing transformer-based models to detect potential data annotation inconsistencies in death investigation notes. In our evaluation, we measured the annotation discrepancies of suicide circumstances across all U.S. states. Here, we refer to the state under evaluation as the 'target state' and all other states as the 'other states'. For each suicide circumstance variable, we trained one transformer-based binary classifier[3] using the data sampled from the target state and other states. Then, we assessed the annotation inconsistencies by determining the change in the F-1 score after excluding the training data from the target state and re-training the classifier. We also designed a cross-validation-like framework to identify problematic data instances that were causing these inconsistencies. These problematic instances were manually rectified, after which we retrained the classifiers to evaluate the effectiveness of the corrections. In this work, we used F-1 scores as an underlying evaluation metric for comparison. The F-1 score is the harmonic mean of Precision (the ratio of true positive predictions to all positive predictions) and Recall (the ratio of true positive predictions to all actual positives). The F-1 score balances Precision and Recall into one single value. A higher F-1 score indicates better model performance. Finally, we analyzed the Odds Ratio (OR) computed for various demographic subgroups (age, sex, race) to better understand the risk of bias.

Our experiments show the efficacy of our approach in identifying potential annotation errors in NVDRS's death investigation notes. Moreover, correcting these errors yields an average F-1 score improvement of 3.85%. In summary, our work aims to enhance the understanding of annotation inconsistencies found in unstructured death investigation notes in the NVDRS. By addressing these inconsistencies, we hope to facilitate the use of NVDRS data in discovering suicide circumstances, enabling longitudinal change analysis and trend analysis, and helping develop targeted suicide prevention strategies at the national, state, and local levels.

Methods

Data source

This work utilizes data from the National Violent Death Reporting System (NVDRS) dataset, covering 267,804 recorded suicide death incidents from 2003 to 2020 across all 50 U.S. states, Puerto Rico, and the District of Columbia[2]. To access the NVDRS dataset, researchers must meet certain eligibility requirements and take steps to ensure confidentiality and data security. Our research was approved by the NVDRS Restricted Access Database (RAD) proposal, which gave us the required permissions to access the data and undertake the work described here. We have also obtained the approval from Weill Cornell Medicine's Institutional Review Board to undertake our study 23-12026810-01, titled "Use AI/ML to Address the Crisis of Suicide".

Each incident instance is accompanied by two death investigation notes, one from the Coroner or Medical Examiner (CME) perspective and the other from the Law Enforcement (LE) perspective. The NVDRS contains over 600 unique data elements for each incident, including the identification of suicide crises—precipitating events contributing to the occurrence of suicides, that occurred within 2 weeks before a suicide death[4]. Examples of suicide crises include Family Relationship, Physical Health, and Mental Health crises. Suicide crises are annotated based on the content of the CME and LE reports. The data annotator (i.e., abstractor) selects from a list of predefined crises and must code all known crises related to each incident. If either the CME report or LE report indicates the presence of a crisis, the abstractor must acknowledge and record this crisis in the database[4].

This study has three tasks: validating the inter-state annotation inconsistencies, identifying specific data instances that caused these inconsistencies, and verifying the improvement in annotation consistencies after removing the identified problematic data instances. We present our methods and conduct experiments using three crises as illustrative examples: Family Relationship Crisis, Mental Health Crisis, and Physical Health Crisis (the state-wise statistics are detailed in Table 1). These variables were selected for their higher prevalence of positive instances in the NVDRS dataset and their poor classification scores, as demonstrated in prior work[3]. Definitions and examples of these three crises can be found in Supplementary Table 1. We also addressed the positive/negative class imbalance in the NVDRS dataset through data pre-processing. First, states with fewer than 10 positive instances were excluded to ensure adequate training data. Next, for each crisis, we created a balanced class distribution for every state by keeping the positive instances intact and down-sampling the negative instances, ensuring an equal number of both.

Validate annotation inconsistency

Inspired by Zeng et al.[13], our approach is grounded on the assumption that if the label annotations for two datasets are consistent, the models trained separately on these datasets should exhibit equivalent predictive capabilities when applied to each other. In practical terms, given a dataset D, if we train a model using one of its subsets to predict the remaining portion, we anticipate observing a comparable evaluation performance for both subsets.

Based on this assumption, we first explore whether the label annotations in the target state s are consistent with those in all other states (Step 1 in Fig. 1). Specifically, given the annotated data of target state $D_s \subset D$ (where D_s has a size of x), we sample m exclusive subsets (each with a size of x) from the annotated data of other states, denoted as D_{other}. It is worth noting that $D_s \cap D_{other} = \varnothing$.

We then split D_s and D_{other} into training, validation, and test sets, respectively, with a ratio of 8:1:1, and construct three different training sets of the same size: (1) PureOthers exclusively comprising samples from states other than the target state, (2) Others+Target combining samples of other states with samples of the target state, and (3) Target+Others similarly combining samples of the target and samples of other states in order. For each training set, we trained one transformer-based binary classifier, specifically using the Bidirectional Encoder Representations from Transformers (BERT) model[24]. Our goal is to compare the classification performances between different training set combinations. Specifically, we assess the inconsistencies between every state and other states in the annotations of Physical Health, Family Relationship, and Mental Health crises. To quantify the inconsistency, we compute the ΔF-1's on the test sets for both the target state and other states. The inconsistency is measured as the difference between the average F-1 score of models trained using mixed training data (Others+Target and Target+Others) and the F-1 score of the model trained solely on data from other states (PureOthers)

$$\triangle F - 1 = Difference\left(F - 1_{Mixed} - F - 1_{PureOthers}\right) \tag{1}$$

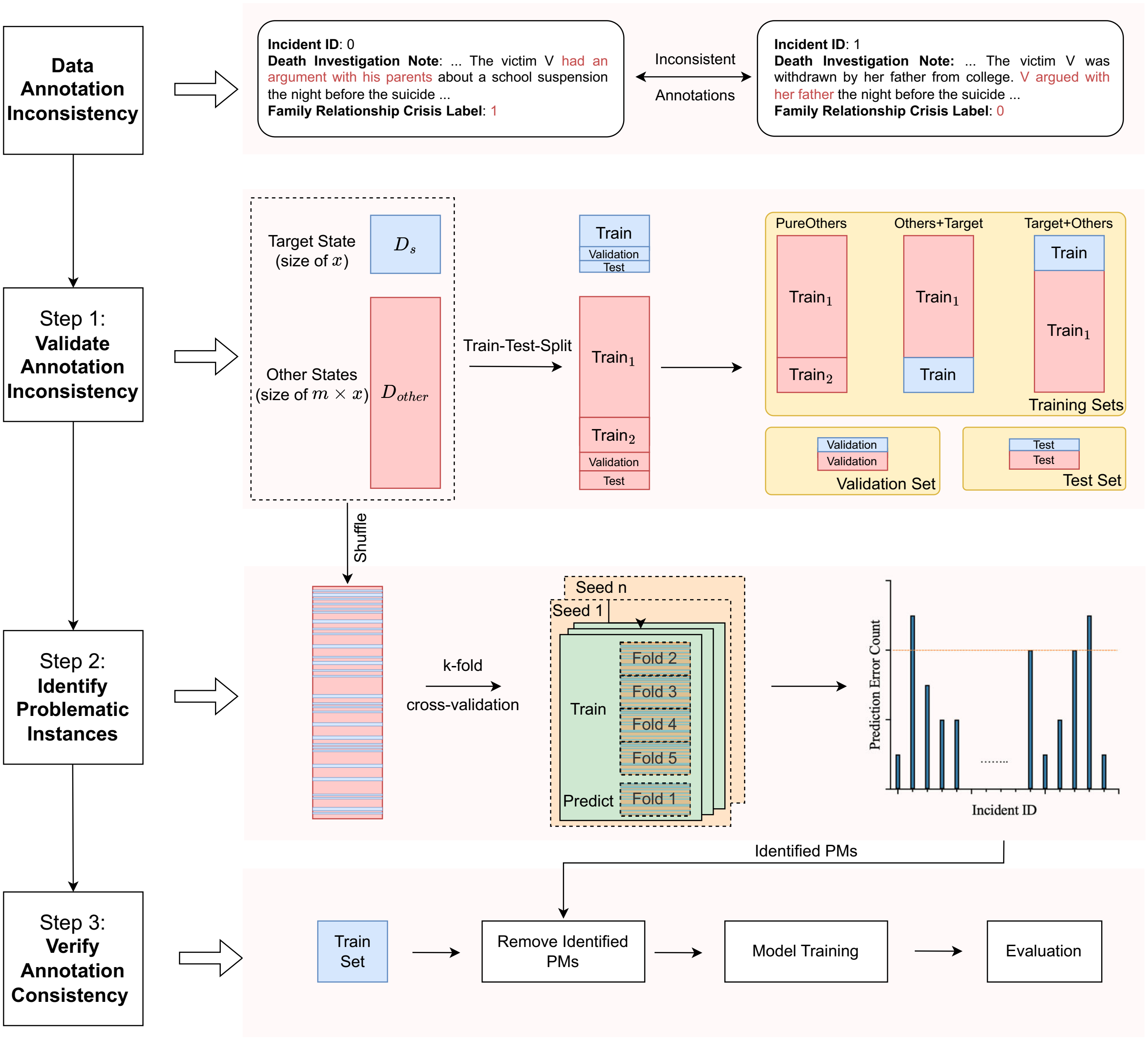
$$F - 1_{Mixed} = Mean\left(F - 1_{Others+Target}, F - 1_{Target+Others}\right) \tag{2}$$

Table 1 | State-wise data statistics

State	Physical Health			Mental Health			Family Relationship		
	Positive	Negative	Total	Positive	Negative	Total	Positive	Negative	Total
Alabama	4	327	331	1	491	492	8	159	167
Alaska	102	213	315	1	622	623	50	177	227
Arizona	198	1116	1314	69	2762	2831	117	552	669
Arkansas	1	50	51	1	107	108	0	34	34
California	179	884	1063	113	2964	3077	117	233	350
Colorado	536	2779	3315	204	8330	8534	324	5695	6019
Connecticut	1	627	628	1	1177	1178	1	152	153
Delaware	15	88	103	7	222	229	7	36	43
District of Columbia	10	13	23	9	160	169	14	17	31
Florida	0	0	0	0	0	0	0	1	1
Georgia	240	1113	1353	29	2845	2874	104	577	681
Hawaii	3	174	177	3	303	306	2	85	87
Idaho	50	36	86	43	126	169	25	31	56
Illinois	161	718	879	39	2506	2545	94	313	407
Indiana	255	495	750	68	1599	1667	103	205	308
Iowa	233	324	557	301	1283	1584	75	302	377
Kansas	67	761	828	19	1396	1415	87	233	320
Kentucky	142	1212	1354	84	1783	1867	89	358	447
Louisiana	86	288	374	14	894	908	54	250	304
Maine	76	257	333	38	612	650	50	101	151
Maryland	113	412	525	42	4284	4326	103	173	276
Massachusetts	283	598	881	44	3061	3105	149	239	388
Michigan	216	2849	3065	30	5279	5309	83	1268	1351
Minnesota	184	669	853	213	2618	2831	133	544	677
Mississippi	4	24	28	0	19	19	4	14	18
Missouri	121	487	608	41	1470	1511	83	171	254
Montana	2	154	156	3	186	189	2	53	55
Nebraska	39	137	176	59	245	304	20	66	86
Nevada	68	416	484	18	734	752	54	171	225
New Hampshire	38	270	308	15	993	1008	2	151	153
New Jersey	217	623	840	114	2320	2434	189	193	382
New Mexico	141	622	763	50	1041	1091	67	439	506
New York	83	815	898	60	3887	3947	54	380	434
North Carolina	640	1929	2569	104	5527	5631	675	280	955
North Dakota	17	42	59	10	152	162	14	46	60
Ohio	470	607	1077	95	8439	8534	300	2028	2328
Oklahoma	299	825	1124	59	2784	2843	104	793	897
Oregon	161	1108	1269	42	2729	2771	76	402	478
Pennsylvania	288	930	1218	13	2749	2762	65	346	411
Rhode Island	29	112	141	20	523	543	22	75	97
South Carolina	105	1224	1329	72	1964	2036	71	364	435
South Dakota	0	16	16	0	56	56	0	12	12
Tennessee	95	129	224	11	652	663	78	54	132
Texas	29	103	132	46	365	411	48	51	99
Utah	676	760	1436	191	2641	2832	550	348	898
Vermont	42	101	143	36	421	457	25	34	59
Virginia	493	1315	1808	352	4817	5169	621	444	1065
Washington	570	697	1267	166	2762	2928	348	560	908
West Virginia	12	204	216	1	350	351	1	100	101

Table 1 (continued) | State-wise data statistics

State	Physical Health			Mental Health			Family Relationship		
	Positive	Negative	Total	Positive	Negative	Total	Positive	Negative	Total
Wisconsin	297	978	1275	17	2811	2828	169	537	706
Wyoming	0	51	51	0	85	85	1	13	14
Puerto Rico	20	156	176	19	600	619	15	64	79



Incident ID: 0
Death Investigation Note: ... The victim V had an argument with his parents about a school suspension the night before the suicide ...
Family Relationship Crisis Label: 1

Incident ID: 1
Death Investigation Note: ... The victim V was withdrawn by her father from college. V argued with her father the night before the suicide ...
Family Relationship Crisis Label: 0

Fig. 1 | Annotation inconsistency example and our proposed framework. In Step 1, the size of other states' $Train_2$ set equals the size of the target state's *Train* set, ensuring the three new training sets are of the same size. In Step 2, the *k*-fold cross-validation procedure is repeated *n* times using different random seeds. For each data instance, we recorded its prediction error counts, and eventually identified the problematic instances by thresholding the prediction error counts. PMs Potential Mistakes.

When incorporating the data from the target state into training, a larger positive ΔF-1 on the test set of the target state, accompanied by a smaller negative ΔF-1 on the test set of other states, indicates a more pronounced annotation inconsistency between the target state and other states.

Identify problematic instances

To identify problematic data instances in the target state which might cause the label inconsistencies between D_s and D_{other}, we introduce a *k*-fold cross-validation approach (Step 2 in Fig. 1), inspired by the approach in Wang et al.[21]. Our method involves the following steps: we concatenate D_s and D_{other} into one set, we randomly shuffle the data to ensure it is well-mixed, and we divide the shuffled dataset into *k* folds. Each unique fold is treated as a hold-out set, while the remaining *k*-1 folds serve as the training set. We train independent suicide circumstance classifiers to identify problematic instances in each fold. Throughout this process, each individual data sample gets assigned to a specific fold where it remains for the duration

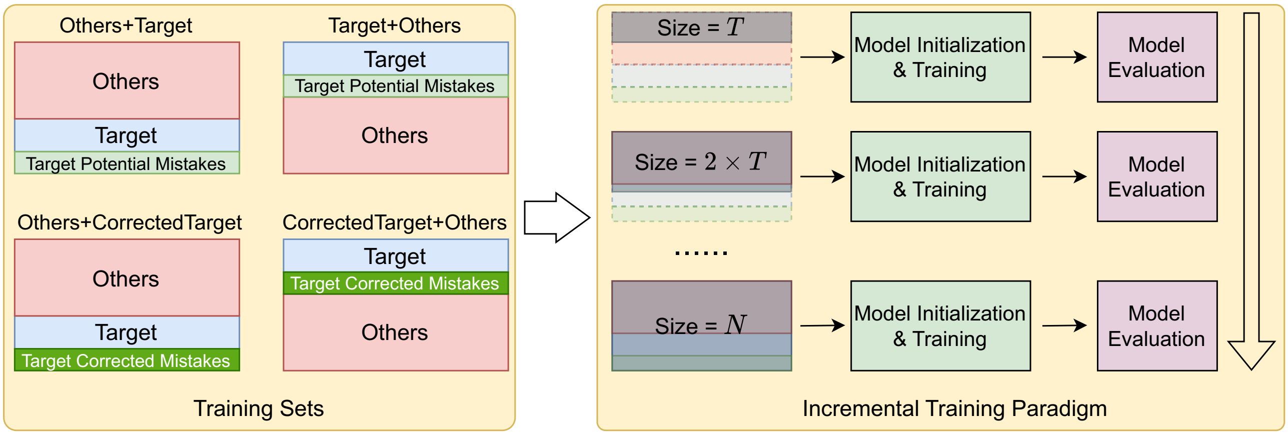


Fig. 2 | Illustration of the proposed incremental training paradigm. Four combinations of training data are shown. For each combination of training data, we feed it to the incremental training paradigm. T is the step size, and N is the total number of data instances in the target state s and other states.

of the cross-validation. This ensures that each data sample is utilized once in the hold-out set and contributes to training the model k-1 times. For each data sample in the hold-out set, we compare the model's prediction to the ground truth label and count the number of discrepancies.

To reduce randomness and enhance the reliability of our findings, we repeat the k-fold cross-validation procedure multiple times (i.e., n times), employing different random data partitions with distinct random seeds for each iteration. Then, for each data instance in the dataset, we get n estimations. We denote c_i ($0 \le c_i \le n$) as the number of times a data instance x_i is flagged for potential labeling errors across all n estimations. This count c_i indicates the confidence level that x_i might have labeling mistakes. Following this, we apply a thresholding mechanism to the counts of prediction errors for each data sample in D_s. This thresholding enables us to effectively identify and flag those data instances that repeatedly show inconsistencies.

Verify annotation consistency

Once we've identified the problematic data instances in D_s, our next step is to evaluate whether these potential mistakes negatively impact the model's performance. To this end, we systematically remove data instances identified as potential mistakes from the training dataset. By removing these instances and re-training the model, we can assess the impact of these potential mistakes on the model's performance (Step 3 in Fig. 1). To measure the effectiveness of these removals, we introduce a random baseline for comparison, which randomly removed the same number of instances from the training set as those identified as problematic.

On a separate front, our efforts extend to manual correction of potential mistakes. After identifying the potential mistakes, we recruited two annotators to manually identify and correct the actual mis-labelings. The actual mis-labelings are defined as instances where the two annotators identify ground truth annotations as incorrect. Two annotators received training on annotating labels following the NVDRS coding manual and resolved disagreements through discussion. Our objective is to show how consistent annotations can enhance the performance of classifiers. We employ an incremental training paradigm to demonstrate this with four training sets (Fig. 2): Others+Target, comprising the data from other states and the original data from target state; Others+CorrectedTarget, comprising the data from other states and the data from target state after correction; Target+Others, comprising the original data from the target state and the data from other states, and CorrectedTarget +Others with the data from the target state after correction and the data from other states.

For each training set, we progressively incorporate more training samples in an incremental manner using a step size of T, to have a finer-grained view of how the corrected data impact the model performance. We train the classification models and analyze the performances on the test set. This process helps validate the label consistency and the effectiveness of the corrected data. We repeat all experiments n=5 times using different random seeds.

Risk of bias analysis

To better understand the risk of bias in the data annotation, we employed logistic regression models to examine whether the relationship between the suicide circumstances and demographic variables (i.e., race, age, and sex) has changed as we removed the identified mistakes. NVDRS captures victim's sex at the time of the incident according to the Death Certificate (DC). NVDRS follows U.S. Department of Health and Human Services (HHS) and Office of Management and Budget (OMB) standards for race/ethnicity categorization, which defines standards for collecting and presenting data on race and ethnicity for all Federal reporting. In this work, we followed the HHS standard and used two categories for ethnicity (Hispanic or Latino, and Not Hispanic or Latino), and five categories for data on race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White) following the OMB and HHS standards. One distinct logistic regression model was developed for each suicide circumstance.

Specifically, the predictor variable represented the specific comparison group (i.e., Black, youth (age under 24), female) and was coded as 1. This was then contrasted with the reference group (i.e., white, adult, male), coded as 0. We calculated the ORs for each comparison group using the coefficient estimate affiliated with the predictor variable obtained from the corresponding logistic regression model. The OR quantifies the likelihood of the specific circumstance occurring in a comparison group versus the reference group. The OR is computed as follows: $OR = e^{Coefficient\ Estimate\ for\ the\ Comparison\ Group}$. ORs greater than 1 indicate that the comparison group had higher circumstance rates than the reference group. We further calculated a 95% confidence interval (CI) for each OR based on the standard error of the coefficient estimate and the Z-score as follows:

$$Lower\ CI\ Bound = e^{Coefficient\ Estimate} - Z \times Standard\ Error,$$

$$Higher\ CI\ Bound = e^{Coefficient\ Estimate} + Z \times Standard\ Error$$

For two illustrative states (Ohio and Colorado), we computed the ORs of each circumstance variable in three sets of annotations: the original annotations from the NVDRS, the annotations after removing the mistakes identified by our method, and the annotations after randomly dropping the same number of instances as the identified mistakes. By comparing the ORs for the same subgroup in different sets of annotations, we can examine whether the relationship between the suicide circumstances and demographic variables has changed.

Table 2 | Statistics of the identified problematic data instances

State	Physical Health			Family Relationship			Mental Health		
	Total	PMs	(%)	Total	PMs	(%)	Total	PMs	(%)
Ohio	1077	159	14.8	2328	324	13.9	9654	143	1.5
Colorado	3315	254	7.7	6019	294	4.9	8534	168	2.0

PMs Potential Mistakes.

Statistics and reproducibility

In this study, we used BioBERT[25] as our backbone model, known for its state-of-the-art performance, as demonstrated in our prior study[3]. BioBERT works with sequences of up to 512 tokens, producing 768-dimensional representations. About 5.1% of the NVDRS data have input length longer than 512 tokens, and they were truncated before being fed to BioBERT. We framed suicide crisis detection as a text classification problem by feeding concatenated CME and LE notes into BioBERT and training it to classify whether a suicide crisis of interest is mentioned in the text. We appended a fully connected layer on top of BioBERT for classification.

For each crisis, states with fewer than 10 positive instances were excluded to ensure adequate training data for validating annotation inconsistency. For experiments, we sampled $m = 4$ exclusive subsets from the annotated data of other states. We conducted the experiments five times ($n = 5$) to strike a balance between achieving a reliable evaluation and maintaining a reasonable running time. It also ensures that both training and testing sets contain sufficient variations. Each iteration used a different random seed, and we reported the range of micro F-1 scores along with the average. For problematic instance discovery, we chose $k = 5$ for k-fold cross-validation following common machine learning practices. A higher frequency of discrepancies between prediction results and ground truth labels increases the probability of an incorrect ground truth label. We set the threshold at 5, effectively minimizing the number of false potential mistakes.

In our prior work, we applied BERT-based models to classify crises in NVDRS narratives[3]. We selected Physical Health, Family Relationship and Mental Health crises in this study due to their higher frequency of positive instances and poor classification Area under the ROC Curve (AUC) scores compared to other crises (Table 2 and Fig. 2 in Wang et al.[3]). Similarly, we chose Ohio and Colorado as illustrative states for their higher frequency of positive instances and superior state-wise classification F-1 scores compared to other states (Tables A2 and A3 in Wang et al.[3]).

Binary Cross Entropy Loss and Adam optimizer were used during model training. We trained all the models for 30 epochs, and model selection was based on their performances on validation sets. The framework was implemented using PyTorch. We conducted our experiments using an Intel Xeon 6226 R 16-core processor and Nvidia RTX A6000 GPUs.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

Validating annotation inconsistency

Figure 3 (Supplementary Data 1) shows ΔF-1's on the test sets of the target state (↑) and other states (↓). In Fig. 3(a), outcomes for the Physical Health Crisis show that when the target state's data was added to the training set, approximately 83.7% (36 out of 43) of states improved their prediction performance on the target state's test set (indicated by a positive ΔF-1), while around 69.8% (30 out of 43) of states experienced a performance drop on other states' test sets (indicated by a negative ΔF-1).

In Fig. 3b, results for the Family Relationship Crisis reveal that when the target state's data was included in the training, 32.5% (13 out of 40) of states improved their prediction performance on the target state's test set. In comparison, 40% (16 out of 40) of states experienced a performance decrease on other states' test sets.

In Fig. 3c, findings for the Mental Health Crisis demonstrate that after including the target state's data in training, approximately 33.3% (13 out of 39) of states improved their prediction performance on the target state's test set. In comparison, around 43.6% (17 out of 39) of states saw a performance drop on other states' test sets.

Discovering problematic instances

Figure 4 (Supplementary Table 2) shows the prediction error count distributions (log scale for better readability) for two illustrative states, Ohio and Colorado. Table 2 offers a detailed statistical summary of these problematic data instances. For Ohio, our problematic instance discovery identified 159 potential mistakes out of 1077 Family Relationship Crisis annotations (14.8%), 324 out of 2328 Physical Health Crisis annotations (13.9%), and 143 out of 9654 Mental Health Crisis annotations (1.5%). For Colorado, our method detected 254 potential mistakes out of 3315 Family Relationship Crisis annotations (7.7%), 294 out of 6019 Physical Health Crisis annotations (4.9%), and 168 out of 8534 Mental Health Crisis annotations (2.0%).

Verifying annotation consistency

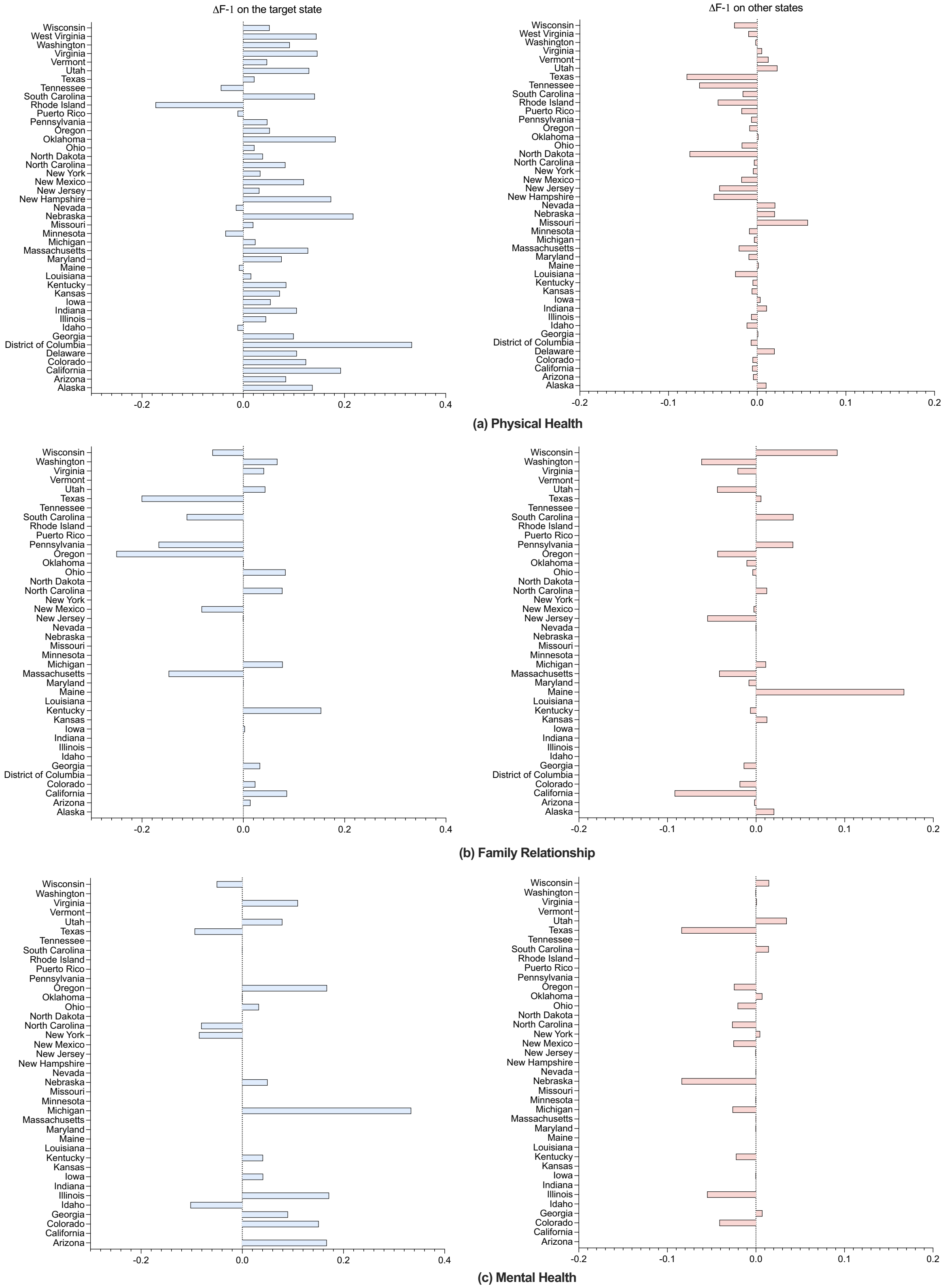
Figure 5 (Detailed F-1 scores are available in Table 3) visually represents our annotation consistency verification results. After removing potential mistakes from Ohio's data, we observed notable improvements in the average micro F-1 scores for each crisis on other states' test sets. In contrast, the random baseline resulted in smaller performance gains for all three crises. Specifically, for Family Relationship Crisis, the score increased from 0.695 to 0.713 after removing the potential mistakes, compared to an increase from 0.695 to 0.701 with random dropping. For Physical Health Crisis, it improved from 0.645 to 0.664 after removing the potential mistakes, as opposed to an increase from 0.645 to 0.654 with random dropping. For Mental Health Crisis, it rose from 0.571 to 0.600 when the potential mistakes were removed, in contrast to an increase from 0.571 to 0.585 with random dropping.

Similar trends were observed in Colorado. After removing potential mistakes, the average micro F-1 score for Family Relationship Crisis on the test set of other states increases from 0.705 to 0.726, compared to an increase from 0.705 to 0.714 with random dropping. For Physical Health Crisis, it increased from 0.684 to 0.694, in contrast to an increase from 0.684 to 0.690 with random dropping. For Mental Health Crisis, it rose from 0.574 to 0.607, compared to an increase from 0.574 to 0.587 with random dropping.

Rectifying problematic data

Two annotators achieved a high Inter-Annotator Agreement (IAA) of 0.893 (Kappa value). Among the 159 potential mistakes, 89 were confirmed as actual mis-labelings. These included 87 instances where the Family Relationship Crisis labels were incorrectly labeled as '0' in the ground truth annotations, and 2 instances where the labels were mistakenly labeled as '1' in the ground truth annotations.

Figure 6 (Supplementary Data 2) shows the changes in the average micro F-1 scores during the incremental training process. In Fig. 6a, the model performance on the test set of other states exhibits significant improvement when the corrected data is fed to the model at the beginning of the training process (to the left of the black vertical dashed line), and at the end of the training process (to the right of the red vertical dashed line). The label correction boosts the eventual average micro F-1 score on the test set of other states from 0.691 to 0.733. A substantial improvement can also be observed on Ohio's test set in Fig. 6b. The label correction enhances the eventual average micro F-1 score on Ohio's test set from 0.679 to 0.714. This result demonstrates that, after correction, the corrected data instances benefit the model performances on both the test sets of other states and the target state, regardless of whether they are fed to the model at the beginning or the end of the training process.



(a) Physical Health

(b) Family Relationship

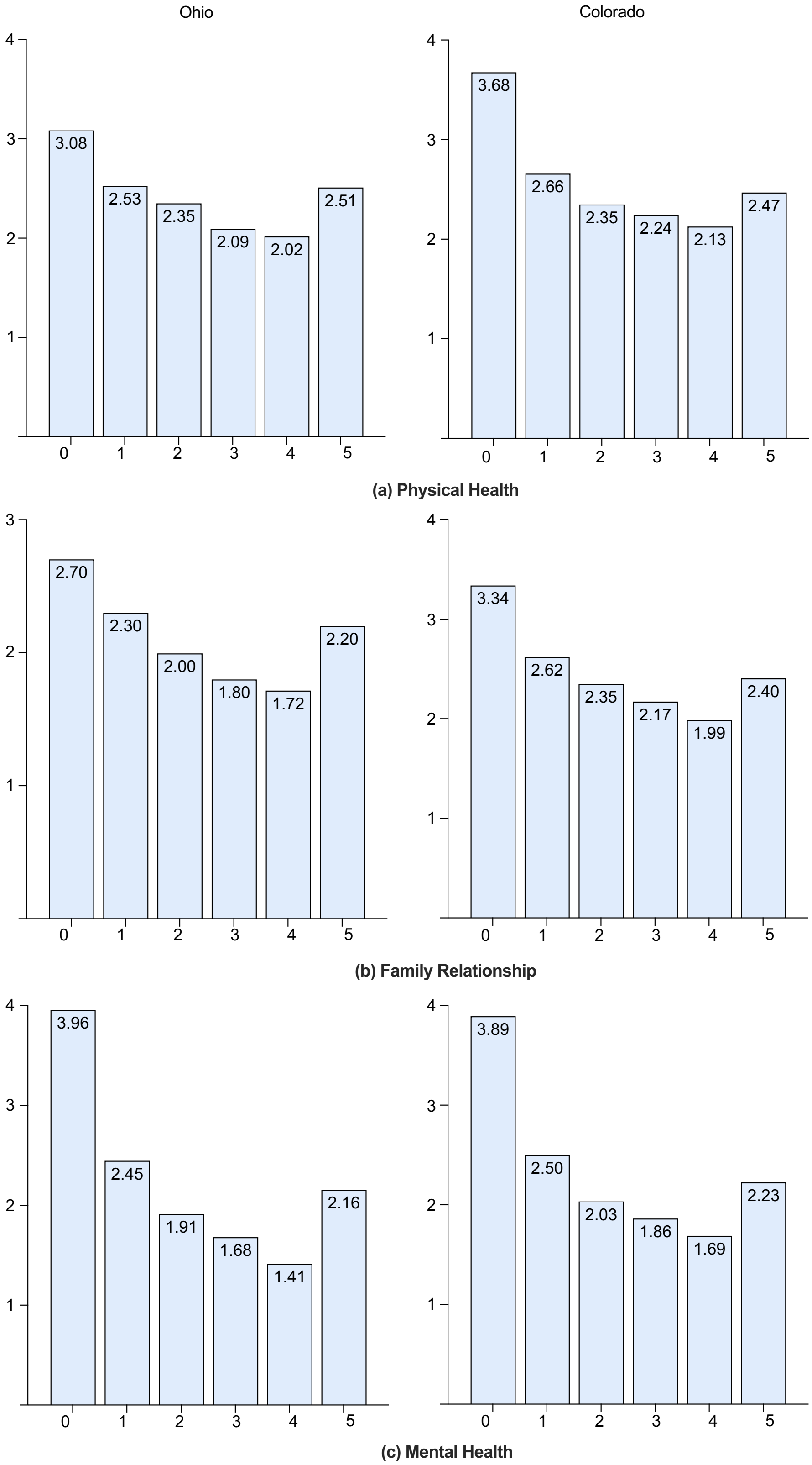
(c) Mental Health

Fig. 3 | ΔF-1's on the test sets of the target state (↑) and other states (↓). a ΔF-1's on the test sets for the Physical Health circumstance, (**b**) ΔF-1's on the test sets for the Family Relationship circumstance, (**c**) ΔF-1's on the test sets for the Mental Health circumstance.

Risk of bias analysis

Table 4 shows the Odds Ratio (OR) comparisons between youth vs adults, Blacks vs whites, and females vs males. Notably, in relation to the Mental Health Crisis in Colorado, the OR for youth in the original NVDRS annotations (OR = 0.89, 95%CI = 0.59–1.33) is similar to that in the annotations after random dropping (OR = 0.88, 95% CI = 0.58–1.34). However, it differs from the OR in the annotations after removing the mistakes identified by our method (OR = 0.65, 95% CI = 0.31–1.36).

Fig. 4 | Prediction error count distributions (log scale) of Ohio and Colorado. a Prediction error count distributions (log scale) for the Physical Health circumstance, (**b**) Prediction error count distributions (log scale) for the Family Relationship circumstance, (**c**) Prediction error count distributions (log scale) for the Mental Health circumstance. Data instances with a prediction error count equal to 5 will be identified as potential mistakes.



Similarly, the OR for the Black individuals in the original NVDRS annotations (OR = 0.68, 95% CI = 0.49–0.93) is similar to that in the annotations after random dropping (OR = 0.67, 95% CI = 0.48–0.93), but deviates from the OR in the annotations after removing the mistakes identified by our method (OR = 0.51, 95% CI = 0.07–3.7).

Discussion

This study touches on a previously unexplored area of uncovering annotation inconsistencies in unstructured death investigation notes and resolving likely misattributed suicide circumstances. To bridge this gap, we proposed an empirical NLP approach. To the best of our knowledge, no

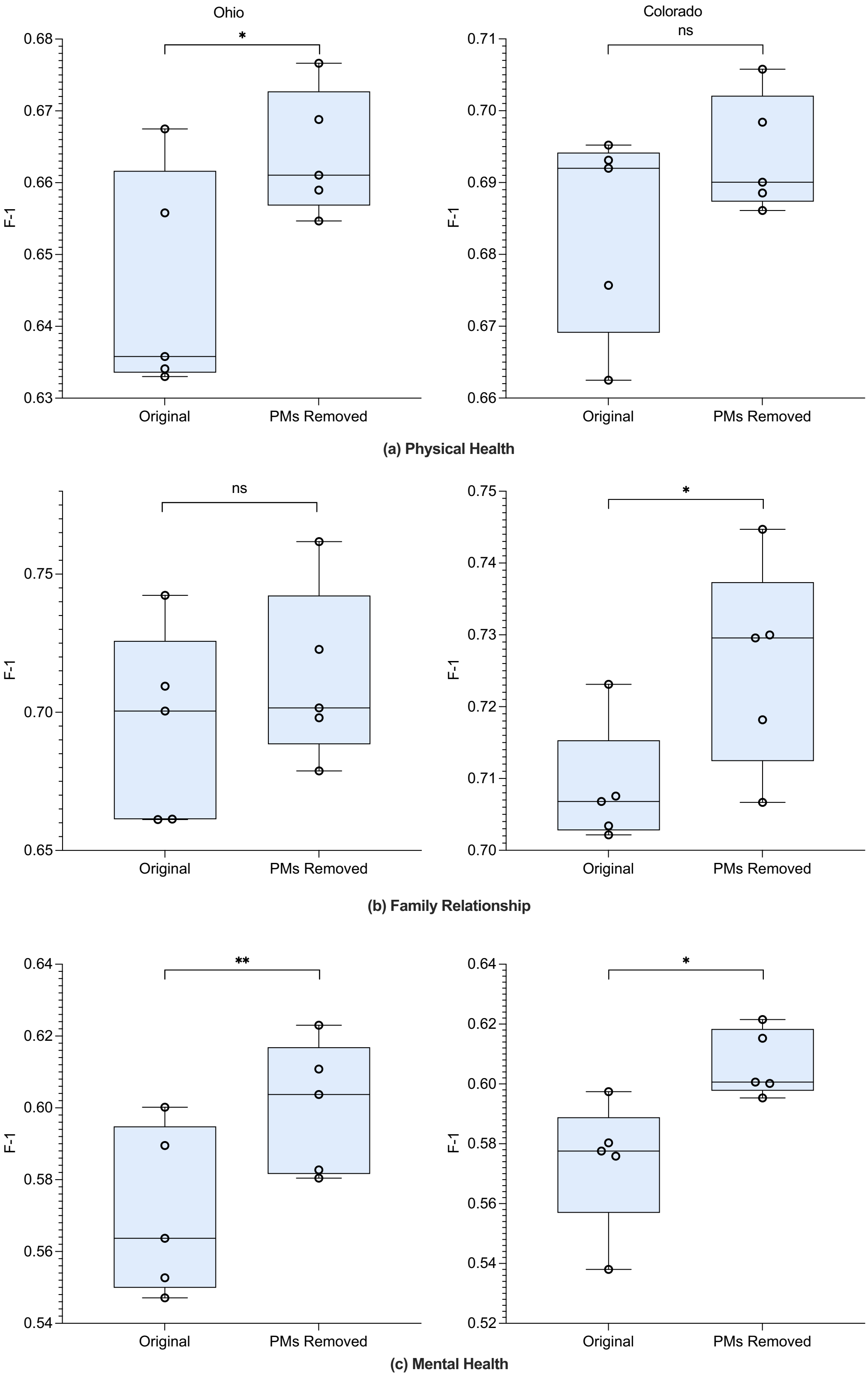
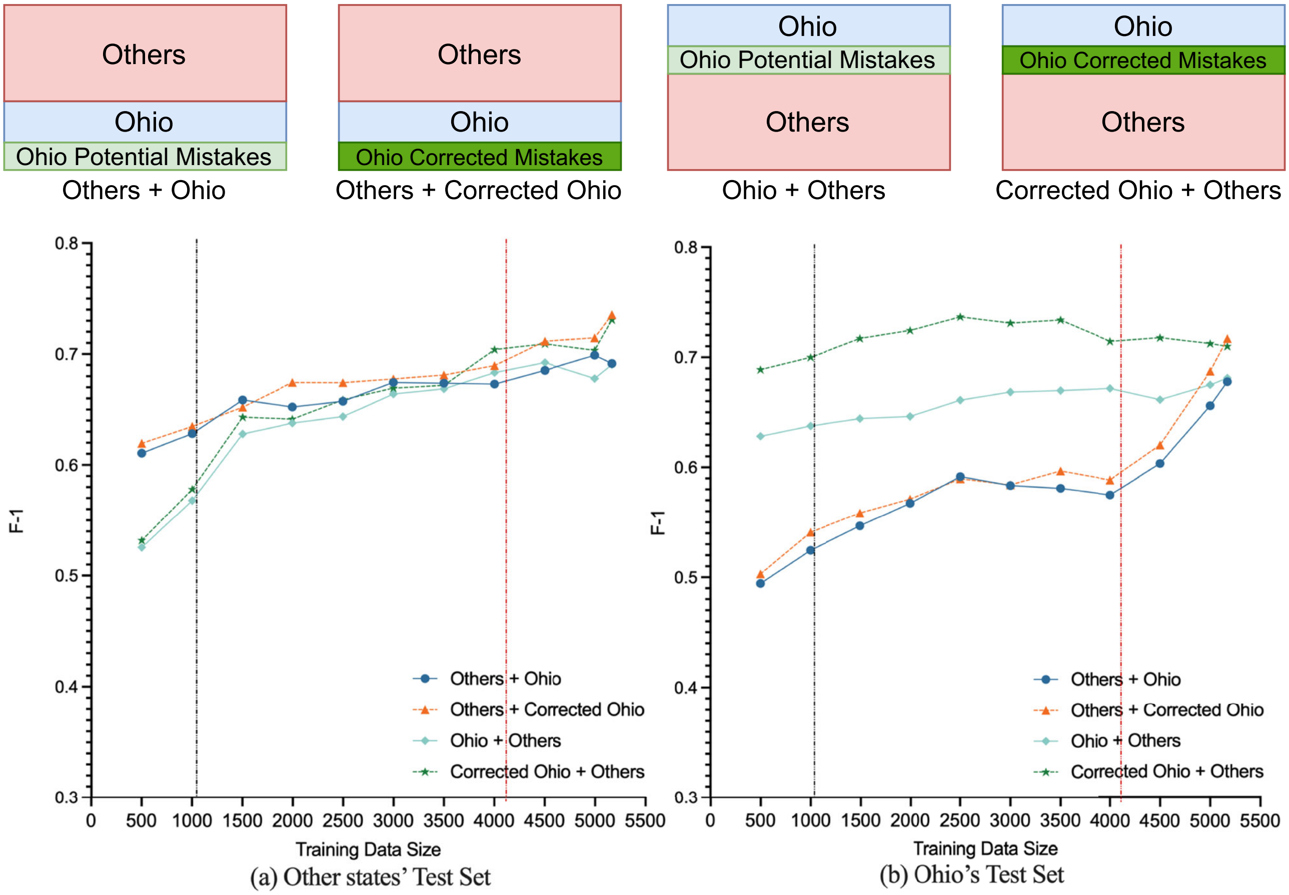


Fig. 5 | Comparison of F-1 scores. a Comparison of F-1 scores between models trained using 'Original' (before removing the identified potential mistakes), and 'PMs Removed' (after removing the identified potential mistakes) for the Physical Health circumstance, (**b**) Comparison of F-1 scores between models trained using 'Original' (before removing the identified potential mistakes), and 'PMs Removed' (after removing the identified potential mistakes) for the Family Relationship circumstance, (**c**) Comparison of F-1 scores between models trained using 'Original' (before removing the identified potential mistakes), and 'PMs Removed' (after removing the identified potential mistakes) for the Mental Health circumstance. PMs Potential Mistakes. The asterisk indicates statistical significance. There were 5 independent experiments conducted to derive the statistics.

Table 3 | Comparison of F-1 scores between models trained using 'Original' (before removing the identified potential mistakes), 'PMs Randomly Dropped' (after randomly dropping the identified potential mistakes), and 'PMs Removed' (after removing the identified potential mistakes)

Crisis	Model	Ohio					Colorado				
		Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5
Physical Health	Original	0.656	0.636	0.633	0.668	0.634	0.692	0.676	0.693	0.695	0.663
	PMs Randomly Dropped	0.697	0.677	0.742	0.697	0.692	0.689	0.680	0.687	0.699	0.695
	PMs Removed	0.659	0.661	0.669	0.677	0.655	0.706	0.690	0.686	0.689	0.698
Family Relationship	Original	0.661	0.661	0.700	0.742	0.709	0.707	0.723	0.703	0.702	0.707
	PMs Randomly Dropped	0.632	0.649	0.657	0.665	0.665	0.718	0.719	0.717	0.709	0.707
	PMs Removed	0.723	0.698	0.762	0.679	0.701	0.729	0.745	0.718	0.730	0.707
Mental Health	Original	0.553	0.547	0.600	0.564	0.589	0.597	0.576	0.578	0.538	0.580
	PMs Randomly Dropped	0.563	0.590	0.598	0.580	0.593	0.556	0.602	0.581	0.590	0.604
	PMs Removed	0.623	0.604	0.580	0.611	0.583	0.601	0.615	0.600	0.621	0.595

PMs Potential Mistakes.



(a) Other states' Test Set

(b) Ohio's Test Set

Fig. 6 | Comparisons of average micro F-1 scores for Family Relationship Crisis when we gradually feed more training data to the model. a Comparisons of average micro F-1 scores in Other states' test set for the Family Relationship circumstance when we gradually feed more training data in an incremental manner to the model, (**b**) Comparisons of average micro F-1 scores in Ohio's test set for the Family Relationship circumstance when we gradually feed more training data in an incremental manner to the model. In each subplot, the black vertical dashed line on the left denotes when Ohio's data have all been fed to the model for training data Ohio+Others and CorrectedOhio+Others, while the red vertical dashed line on the right denotes when we start to feed Ohio's data to the model for Others+Ohio and Others+CorrectedOhio.

previous work applies a similar approach to any large-scale healthcare or mortality dataset.

Our analysis of data annotation inconsistencies across different states showed that adding the data from the target state into the training set led to improved performance in 49.8% of the target states when tested against the target state's test set. Meanwhile, testing on the test set of other states showed an average performance decrease in 51.1% of the target states. The performance variation, depending on the training data combinations, underscores the presence of annotation inconsistencies in the NVDRS dataset. These findings highlight the

Table 4 | Odds Ratio for circumstances between youth vs. adults, Blacks vs. whites, and females vs. males

Ohio									
Circumstance	**Youth**	**Adult**	**OR [95% CI]**	**Black**	**White**	**OR [95% CI]**	**Female**	**Male**	**OR [95% CI]**
Family Relationship Crisis									
Original	250	827	1.01 [0.75; 1.36]	103	947	1.53 [1.01; 2.32]	310	767	1.12 [0.85; 1.48]
Random Drop	216	702	1.04 [0.75; 1.44]	84	810	1.56 [0.99; 2.47]	268	650	1.16 [0.86; 1.57]
Our Method	216	702	1.05 [0.76; 1.45]	90	805	1.51 [0.97; 2.36]	256	662	1.06 [0.78; 1.44]
Mental Health Crisis									
Original	1218	8436	0.79 [0.43; 1.44]	859	8638	0.99 [0.51; 1.91]	2814	6840	0.95 [0.63; 1.43]
Random Drop	1201	8310	0.79 [0.43; 1.46]	847	8510	1.00 [0.52; 1.92]	2767	6744	0.96 [0.64; 1.45]
Our Method	1204	8307	0.81 [0.44; 1.48]	846	8510	1.00 [0.52; 1.94]	2780	6731	0.98 [0.65; 1.47]
Physical Health Crisis									
Original	30	2298	0.52 [0.18; 1.51]	84	2221	1.01 [0.60; 1.70]	442	1886	0.63 [0.48; 0.82]
Random Drop	26	1978	0.64 [0.22; 1.88]	69	1914	0.64 [0.33; 1.24]	369	1635	0.60 [0.44; 0.80]
Our Method	28	1976	0.34 [0.08; 1.45]	73	1910	0.79 [0.41; 1.52]	387	1617	0.49 [0.35; 0.68]
Colorado									
Family Relationship Crisis									
Original	684	2631	1.38 [1.07; 1.80]	112	3063	1.15 [0.64; 2.08]	946	2369	0.87 [0.67; 1.13]
Random Drop	626	2435	1.34 [1.02; 1.76]	100	2831	1.08 [0.57; 2.05]	873	2188	0.96 [0.74; 1.25]
Our Method	613	2448	1.43 [1.01; 2.01]	102	2827	1.35 [0.64; 2.83]	879	2182	0.88 [0.63; 1.24]
Mental Health Crisis									
Original	1237	7297	0.89 [0.59; 1.33]	199	8034	0.68 [0.49; 0.93]	2743	5791	0.20 [0.03; 1.42]
Random Drop	1205	7161	0.88 [0.58; 1.34]	195	7874	0.67 [0.48; 0.93]	2688	5678	0.20 [0.03; 1.45]
Our Method	1210	7156	0.65 [0.31; 1.36]	197	7874	0.51 [0.07; 3.70]	2695	5671	0.48 [0.27; 0.84]
Physical Health Crisis									
Original	251	5768	0.49 [0.24; 1.01]	120	5736	0.50 [0.38; 0.66]	1720	4299	0.12 [0.02; 0.86]
Random Drop	235	5490	0.67 [0.39; 1.14]	115	5454	0.65 [0.53; 0.81]	1640	4085	0.08 [0.01; 0.60]
Our Method	241	5484	0.64 [0.52; 0.79]	120	5454	0.67 [0.40; 1.12]	1641	4084	0.08 [0.01; 0.58]

OR Odds Ratio, *CI* Confidence Interval.

need to rectify label inconsistencies in death investigation notes to enhance data quality.

After employing a cross-validation-like paradigm and identifying likely inconsistent instances, we illustrated that removing these instances from the training set significantly improved model performance and generalizability. The consistent improvement observed across two states and within three distinct crisis scenarios suggests that removing potential mistakes can help align the label annotations of the target state with those of other states, underscoring the effectiveness of our approach in identifying annotation mistakes. We further corrected these data and observed performance enhancements in suicide circumstance detection.

Last but not least, our OR analysis measures the strength of association between suicide circumstances (e.g., Mental Health Crisis) and specific groups of suicide decedents (e.g., Youth) compared to a reference group (e.g., Adults). This helps in identifying whether certain subpopulations are more likely to have experienced specific suicide circumstances compared to their counterpart. We found that the OR results, after the correction of inconsistency data instances, are different from those obtained before data corrections were made. This observation indicates the importance of data accuracy in conducting data-driven suicide analysis.

While our study offers promising insights, it does have certain limitations. First, our problematic instance discovery method uses a cross-validation-like method, which can become computationally demanding as the dataset size increases. Secondly, although our proposed framework can work with various models, we only demonstrated the results utilizing BERT-based models. Several NLP tasks have recently showcased the effectiveness of Large Language Models (LLMs)[26,27], which could be a potential area of exploration for future studies. Moreover, for each incident, we concatenated the medical examiner report and the law enforcement report, while NVDRS acknowledges that information from the two data sources can occasionally conflict. At the same time, due to the input token limitations of BERT-based models, 5.1% of the records were truncated before being fed to the BERT model. If information about a certain circumstance were present at the end of one of these records, they would not be fed to the NLP model. The choice of parameters, such as the number of folds and the threshold for error identification, can be further tuned using Grid Search for better results. Future research could aim to identify and mitigate potential biases among subgroups. These biases may stem from how incidents are reported, how data is curated, and how conclusions are drawn, potentially leading to unjustified or skewed outcomes. Addressing these biases will enhance the reliability and fairness of research findings and their subsequent applications in suicide prevention. Lastly, although we demonstrated the effectiveness of manual label correction, automatic methods should be explored for scalability. This annotation inconsistency detection framework might also be applied to other state-based reporting systems, such as the Fatality Analysis Reporting System (FARS)[28]. Such an approach would provide a practical means for improving annotation consistency across large datasets and diverse sources.

The presence of data annotation inconsistencies in NVDRS's death investigation notes not only hampers our understanding of suicide circumstances but also impedes the development, implementation, and evaluation of effective strategies, programs, and policies aimed at preventing suicide. In this work, we proposed an empirical NLP approach to detect the data annotation inconsistencies in the NVDRS, and verify the effectiveness of identifying and rectifying likely problematic instances. Experiment results showcase the capabilities and generalizability of our approach and suggest the limitations of this work. We intend to refine and expand our methodology to address data annotation inconsistencies across diverse data sources. Additionally, we advocate for establishing more stringent annotation guidelines and quality control measures to ensure the consistent and reliable annotation of datasets. By enhancing the accuracy and consistency of annotations within these datasets, we can elevate the performance and reliability of NLP models. This, in turn, equips scientists and policymakers with the means to improve the annotation accuracy for the NVDRS data, and then fundamentally supports discovering the true suicide circumstances, and eventually contributes to suicide prevention.

Code availability

We have made our code publicly available[29].

Data availability

The dataset analyzed in this study, NVDRS RAD, is accessible upon request to researchers who meet specific eligibility criteria and take steps to ensure confidentiality and data security. Our research was approved by the NVDRS Restricted Access Database (RAD) proposal, which gave us the required permissions to access the data and undertake the work described here. This restricted access is in place due to the confidential nature of the NVDRS data, which includes sensitive information that could potentially lead to the unintended disclosure of the identities of victims. To safeguard these data, the CDC protects these data by requiring users to fulfill certain eligibility requirements and implement the necessary to ensure the security of data, preserve confidentiality, and prevent unauthorized access. Researchers interested in accessing the NVDRS data can apply per the instructions provided at https://www.cdc.gov/nvdrs. The source data for Fig. 3 is in Supplementary Data 1, the source data for Fig. 5 is in Table 3, the source data for Fig. 6 is in Supplementary Data 2.

Received: 29 April 2024; Accepted: 7 October 2024;
Published online: 14 October 2024

References

1. CDC. Suicide Prevention. https://www.cdc.gov/suicide/.
2. CDC. The National Violent Death Reporting System. https://www.cdc.gov/nvdrs.
3. Wang, S. et al. An NLP approach to identify SDoH-related circumstance and suicide crisis from death investigation narratives. *J. Am. Med. Inform. Assoc.* **30**, 1408–1417 (2023).
4. Liu, G. S. et al. Surveillance for violent deaths - national violent death reporting system, 48 states, the District of Columbia, and Puerto Rico, 2020. *MMWR Surveill. Summ.* **72**, 1–38 (2023).
5. Hollenstein, N., Schneider, N. & Webber, B. Inconsistency detection in semantic annotation. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3986–3990. European Language Resources Association (ELRA), Portorož, Slovenia, 2016).
6. Květoň, P. & Oliva, K. (Semi-)Automatic Detection of Errors in PoS-Tagged Corpora. in *Proceedings of COLING 2002: The 19th International Conference on Computational Linguistics* (2002).
7. Ma, Q., Lu, B.-L., Murata, M., Ichikawa, M. & Isahara, H. On-line error detection of annotated corpus using modular neural networks. *Proceedings of the International Conference on Artificial Neural Networks 1185–1192*. (Springer-Verlag, Berlin, Heidelberg, 2001).
8. Ule, T. & Simov, K. Unexpected productions may well be errors. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. (European Language Resources Association (ELRA), Lisbon, Portugal, 2004).
9. Loftsson, H. Correcting a PoS-tagged corpus using three complementary methods. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 523–531. Association for Computational Linguistics, USA, 2009).
10. Kato, Y. & Matsubara, S. Correcting errors in a treebank based on synchronous tree substitution grammar. *Proceedings of the ACL 2010 Conference Short Papers* (pp. 74–79. Association for Computational Linguistics, USA, 2010).
11. Manning, C. D. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? in *Computational Linguistics and Intelligent Text Processing* 171–189 (Springer, 2011).
12. Nguyen, P.-T., Le, A.-C., Ho, T.-B. & Nguyen, V.-H. Vietnamese treebank construction and entropy-based error detection. *Lang. Resour. Eval.* **49**, 487–519 (2015).
13. Zeng, Q., Yu, M., Yu, W., Jiang, T. & Jiang, M. Validating label consistency in NER data annotation. *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems* (pp. 11–15. Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021).
14. Chong, D., Hong, J. & Manning, C. D. Detecting label errors by using pre-trained language models. *arXiv [cs.CL]* (2022).
15. Eskin, E. Detecting errors within a corpus using anomaly detection. in *Proceedings of 1st Meeting of the North American Chapter of the Association for Computational Linguistics* (2000).
16. Nakagawa, T. & Matsumoto, Y. Detecting errors in corpora using support vector machines. in *Proceedings of COLING 2002: The 19th International Conference on Computational Linguistics* (2002).
17. Dligach, D. & Palmer, M. Reducing the need for double annotation. *Proceedings of the 5th Linguistic Annotation Workshop* (pp. 65–73. Association for Computational Linguistics, Portland, Oregon, USA, 2011).
18. Amiri, H., Miller, T. & Savova, G. Spotting spurious data with neural networks. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2006–2016. Association for Computational Linguistics, New Orleans, Louisiana, 2018).
19. Swayamdipta, S. et al. Dataset cartography: mapping and diagnosing datasets with training dynamics. *arXiv [cs.CL]* (2020).
20. Yaghoub-Zadeh-Fard, M.-A., Benatallah, B., Chai Barukh, M. & Zamanirad, S. A study of incorrect paraphrases in crowdsourced user utterances. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 295–306. Association for Computational Linguistics, Minneapolis, Minnesota, 2019).
21. Wang, Z. et al. CrossWeigh: Training named entity tagger from imperfect annotations. in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Association for Computational Linguistics, 2019). https://doi.org/10.18653/v1/d19-1519.
22. Northcutt, C., Jiang, L. & Chuang, I. Confident learning: estimating uncertainty in dataset labels. *J. Artif. Intell. Res.* **70**, 1373–1411 (2021).
23. Rehbein, I. & Ruppenhofer, J. Detecting annotation noise in automatically labelled data. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1160–1170. Association for Computational Linguistics, Vancouver, Canada, 2017).
24. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv [cs.CL]* (2018).
25. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

26. Guevara, M. et al. Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med*. **7**, 6 (2024).
27. Keloth, V. K. et al. Large language models for social determinants of health information extraction from clinical notes - a generalizable approach across institutions. *medRxiv* https://doi.org/10.1101/2024.05.21.24307726 (2024).
28. Fatality Analysis Reporting System (FARS). *NHTSA* https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars.
29. *bionlplab/2024_Inconsistency_Detection: v1.0.0*. https://doi.org/10.5281/zenodo.13047596.

Acknowledgements

This study was supported by the AIM-AHEAD Consortium Development Program of NIH under grant number OT2OD032581, National Institute on Aging (NIA) of NIH under grant number RF1AG072799, National Institute of Allergy and Infectious Diseases (NIAID) of NIH under grant number 1R01AI130460, National Science Foundation under grant number 2145640, the Center for Health Economics of Treatment Interventions for Substance Use Disorder, HCV, and HIV (NIDA P30DA040500) and the National Institute for Health Care Management Research and Educational Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and NSF.

Author contributions

S.W., Y.X., and Y.P. contributed to the conception of the study and study design; S.W., Y.X., Y.P. contributed to the acquisition of the data; S.W., Y.Z., Y.X., and Y.P. contributed to the analysis and interpretation of the data; Z.H., C.T., Y.X., Y.D., J.G., and Y.P. provided strategic guidance; S.W. and Y.P. contributed to the paper organization and team logistics; S.W., Y.Z., Z.H., C.T., Y.X., Y.D., J.G., and Y.P. contributed to drafting and revising the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-024-00631-7.

Correspondence and requests for materials should be addressed to Yifan Peng.

Peer review information *Communications Medicine* thanks Amir Eliassaf and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.