Text clustering model for defining cause of death among PLHIV registered in the National AIDS Program, Thailand, 2002-2022

1166

Apiratee Kanphukiew[1], Wiphawee Kiatchanon[1], Niramon Punsuwan[2], Khunkansasi Pimpakhan[3], Suchunya Aungkulanon[4], Cheewanan Lertpiriyasuwat[2], Poonchana Wareechai[4], Natdanai Khosanam[2], Anunthawip Chaimao[3], Sanny Chen Northbrook[1]

[1]Division of Global HIV/TB, U.S. Centers for Disease Control and Prevention, Nonthaburi, Thailand; [2]Division of AIDS and STI, Department of Disease Control, MOPH, Nonthaburi, Thailand; [3]Division of Digital Disease Control, Department of Disease Control, MOPH, Nonthaburi, Thailand; [4]National Health Security Office, Bangkok, Thailand

BACKGROUND

- In Thailand, causes of death are recorded in free-text format, creating challenges for analysis in the National AIDS Program (NAP) data system.
- The system struggles to extract and categorize causes of death from the Ministry of Interior's text-based death registry, hindering monitoring of HIV-related mortality.
- A text clustering model was developed to classify causes of death automatically and distinguish AIDS-related deaths from other causes.
- The model focuses on HIV-positive individuals who died between 2002 and 2022.

The text clustering model exhibited strong performance in classifying causes of death among PLHIV registered in Thailand's National AIDS Program into AIDS-related, non-AIDS-related, and ill-defined categories.

Enhancing data completeness and quality, along with leveraging future advancements in AI, is recommended to further improve the model's effectiveness.

METHODS

- Primary causes of death were classified using Thai HIV surveillance definitions and WHO clinical staging guidelines.
- A text clustering model was developed to categorize deaths into three groups: AIDS-related, non-AIDS-related, and ill-defined causes. (Figure 1)
- Clustering utilized similarity ratios, keywords, and >85% fuzzy string-matching scores in Python.
- Model performance was evaluated with a confusion matrix, measuring precision, recall, and F-1 score.

RESULTS

- 165,541 deaths recorded in NAP, 99% (163,251) had specific cause of death information, with 58% AIDS-related, 36% non-AIDS-related, and 6% ill-defined.
- Deduplication identified 26,913 unique causes of death and 31 major disease groups.
- The text clustering model achieved high performance with 95% precision, recall, and F-1 score overall. (Figure 2)
- Precision was lowest for immune deficiency/LTFU (79%), hematologic diseases (77%), and assault (52%). (Figure 2)
- Recall and F-1 scores were lowest for hematologic diseases (24% recall, 37% F-1) and assault (66% F-1). (Figure 2)
- Tuberculosis and cancer achieved perfect scores for precision, recall, and F-1 (100%). (Figure 2-3)

Figure 3: Sample Categories for Word Cloud



CONCLUSIONS

- The text clustering model demonstrated strong overall performance, with high precision, recall, and F1-scores.
- Challenges in certain categories, such as low scores for specific causes of death, highlight the impact of incomplete death certificate information.
- Leveraging advanced AI techniques in the future could further enhance the model's accuracy and system effectiveness.
- Improving data completeness and quality will be key to overcoming current limitations.

ACKNOWLEDGEMENT

- National Health Security Office, Thailand.
- Division of AIDS and STI, Department of Disease Control, Ministry of Public Health, Thailand.
- Division of Digital Disease Control, Department of Disease Control, Ministry of Public Health, Thailand.
- Division of Global HIV/TB, U.S. Centers for Disease Control and Prevention, Thailand Ministry of Public Health-U.S. Centers for Disease Control and Prevention Collaboration, Thailand.

The findings and conclusions in this presentation are those of the author(s) and do not necessarily represent the official position of the U.S. Centers for Disease Control and Prevention.

Figure 1: Hierarchical algorithm for coding AIDS deaths 3 Main Groups, 31 disease categories

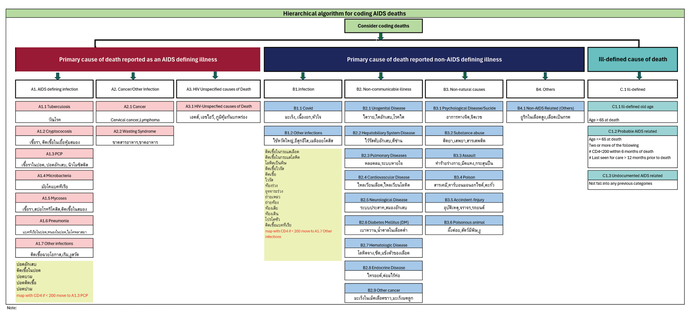
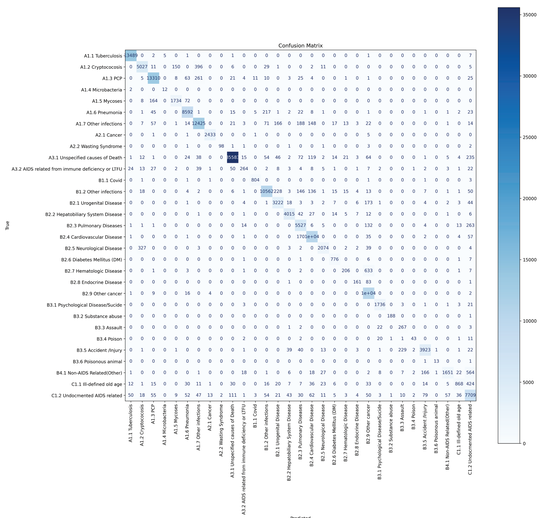


Figure 2: Model Performance and Prediction



Accuracy: 94.51%

	precision	recall	f1-score	support
A1.1 Tuberculosis	0.99	1.00	1.00	13506
A1.2 Cryptococosis	0.92	0.89	0.91	5638
A1.3 PCP	0.97	0.97	0.97	13752
A1.4 Microbacteria	0.71	0.86	0.77	14
A1.5 Mycoses	0.91	0.88	0.89	1978
A1.6 Pneumonia	0.97	0.96	0.96	8938
A1.7 Other infections	0.94	0.94	0.94	13171
A2.1 Cancer	0.99	1.00	0.99	2441
A2.2 Wasting Syndrome	0.98	0.91	0.94	108
A3.1 Unspecified causes of Death	0.99	0.98	0.99	36316
A3.2 AIDS related from immune deficiency or LTFU	0.79	0.54	0.64	490
B1.1 Covid	0.98	0.99	0.98	812
B1.2 Other infections	0.96	0.94	0.95	11213
B2.1 Urogenital Disease	0.87	0.92	0.89	3494
B2.2 Hepatobiliary System Disease	0.97	0.97	0.97	4131
B2.3 Pulmonary Diseases	0.88	0.93	0.90	5967
B2.4 Cardiovascular Disease	0.95	0.97	0.96	10584
B2.5 Neurological Disease	0.95	0.84	0.90	2456
B2.6 Diabetes Mellitus (DM)	0.91	0.98	0.94	792
B2.7 Hematologic Disease	0.77	0.24	0.37	854
B2.8 Endocrine Disease	0.83	0.65	0.73	246
B2.9 Other cancer	0.89	1.00	0.94	10216
B3.1 Psychological Disease/Sucide	0.97	0.98	0.98	1768
B3.2 Substance abuse	0.99	0.99	0.99	189
B3.3 Assault	0.52	0.91	0.66	295
B3.4 Poison	0.86	0.53	0.66	81
B3.5 Accident /Injury	0.93	0.92	0.92	4277
B3.6 Poisonous animal	0.87	0.87	0.87	15
B4.1 Non-AIDS Related(Other)	0.96	0.66	0.78	2495
C1.1 Ill-defined old age	0.90	0.56	0.69	1559
C1.2 Undocnented AIDS related	0.81	0.90	0.85	8541
accuracy			0.95	166337
macro avg	0.90	0.86	0.87	166337
weighted avg	0.95	0.95	0.94	166337