UKRI Centre for Doctoral Training in Artificial Intelligence for Healthcare

Imperial College London

Paper Summary and Report

INSPECT: A Multimodal Dataset for Pulmonary Embolism Diagnosis and Prognosis

Huang, Shih-Cheng, et al.

Fiona Kekwick

March 27, 2024

Module: 70152 - Research Tutorial - AI and Machine Learning for Healthcare

1 Paper Details

Title: INSPECT: A Multimodal Dataset for Pulmonary Embolism Diagnosis and Prognosis

Authors: Huang, Shih-Cheng, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P. Lungren, Curtis P. Langlotz, Serena Yeung, Nigam H. Shah, and Jason A. Fries.

Publication details: Published in the thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track

2 Overview

This paper introduces a new dataset: INSPECT, which is a multi-modal, longitudinal dataset for the diagnosis and prognosis of pulmonary embolism and pulmonary hypertension. It contains electronic health record data combined with CT images and their associated radiology report impressions. All the data is from Stanford Health Care Centre. It also introduces a benchmark AI model that uses the dataset to create a mutli-modal model to diagnose pulmonary embolism and make future prognoses.

3 Background

One of the main problems faced by AI models in the healthcare domain is the lack of realistic, large-scale clinical datasets. Medical data is in its nature multi-modal, and patients' overall medical history can be used to inform diagnosis and prognosis. In fact, in a study performed by Leslie et al. [1], found that correct clinical information in addition to the CT images aided radiologists with improving radiology reports. This suggests that AI diagnosis based on radiology scans could be aided by a patients' health records. In this paper the medical dataset introduced contains electronic health record data, radiology report impressions and CTPA scans for the purpose of diagnosing pulmonary embolisms and pulmonary hypertension:

Pulmonary embolism is when blood clots blocking flow to an artery that feed to lungs. Pulmonary embolism (PE) is a serious medical condition, which was one of the causes for over 2600 deaths in the UK, 2021 [2].

Pulmonary Hypertension (PH) is high blood pressure in the blood vessels that supply the lungs (pulmonary arteries). Pulmonary emboli can lead to high blood pressure in the lunges and right side of the heart, a form of Pulmonary hypertension.

CTPA (computed tomography pulmonary angiogram) are 3D CT scans usd to diagnose pulmonary embolism. They are formed from many 2D CT slices taken at different planes, taken by rotating X-ray.

There have been a lot of works using multi-modal fusion methods in medical imaging [3], but few have focused on prognostic tasks due to the lack of longitudinal datasets [4]. Most medical imaging datasets are small, do not include mulit-modal data, or do not have diagnostic and prognostic labels [4]. Moreover, while there have been some large multi-modal medical datasets, for example UK Biobank [5] and MIMIC [6]. These lack either text data such as radiology reports (UK Biobank), or 3D imaging and prognostic labels (MIMIC).

4 Dataset

The RESPECT dataset contains CTPA image slices and their associated radiology report impression. The radiology report impression is described as 'a summary of the most important findings and possible causes'.

Also included is structured longitudinal electronic health record data (EHR) and clinically relevant labels, which include diagnostic and prognostic labels [4]. Every piece of data contains a timestamp and patient ID, giving full timelines for patients, the EHR data contains information in the form of medical codes that detail every interaction the patient has with the hospital and information, such as medications, lab results and vitals.

To maintain patient privacy the patient IDs are de-identified and the patient timelines are randomly shifted into the future, with shifts that preserve the intervals between data recorded for each patient. All unstructured text from the EHR data was removed to reduce the chance of protected health information (PHI) being released, and each CT slice was manually reviewed for PHI.

5 Benchmark Model

5.1 Benchmark model tasks

Diagnostic tasks: The model predicts the diagnosis of pulmonary embolism in each patient, as either positive or negative.
Prognostic tasks:

- A prognosis of the patient developing pulmonary hypertension is given, again as binary variable.
- A prognosis of in hospital mortality is given at 3 time intervals: 1 month, 6 months, 12 months
- A prognosis of re-admission to hospital is given for 3 time intervals: 1 month, 6 months, 12 months.

This paper uses a late fusion approach, where there are individual models for each modality, where each model returns a probability for the prognostic and diagnostic tasks. These probabilities are fused together using learned weights, w, which form a weighted mean to combine the different predictions from the individual models, giving an overall prediction from the model for each task.

5.2 Benchmark Model Methodology Overview

A graphical model overview is given in Fig. 1. The benchmark model consists of 3 models fused together, 2 models that take the EHR data and one model that takes the imaging data. The model looks at the snapshot in time at which each patient had a CTPA scan and a radiology report detailing if there is a diagnosis of pulmonary embolism. All the EHR data for each patient taken before the CTPA scan is fed into the two EHR models, and the CTPA images are fed into the imaging model. All the data for each patient taken after the radiology report is only used to give determine prognosis labels, an then is discarded. The radiology report impressions are only used to generate binary labels for the classification of pulmonary embolism, and then are discarded. The generated diagnostic and prognostic labels are fed into the supervised learning models, alongside the EHR and CTPA scans.

5.3 Benchmark Model Methodology

5.3.1 Label Generation

The diagnostic labels are generated using a clinical-longformer from [7], which is trained on CTPA reports from a dataset with 4351 reports described in Banerjee et al. 2019 [8]. The data is obtained from the same hospital as the INSPECT dataset and was manually with ground truth labels. After training, this NLP model was used to generate the pulmonary embolism diagnostic labels from the INSPECT dataset. The prognostic labels are generated using a new structured data-based phenotyping algorithm which uses a dictionary of relevant codes ICD and internal Stanford codes along with EHR input data to determine

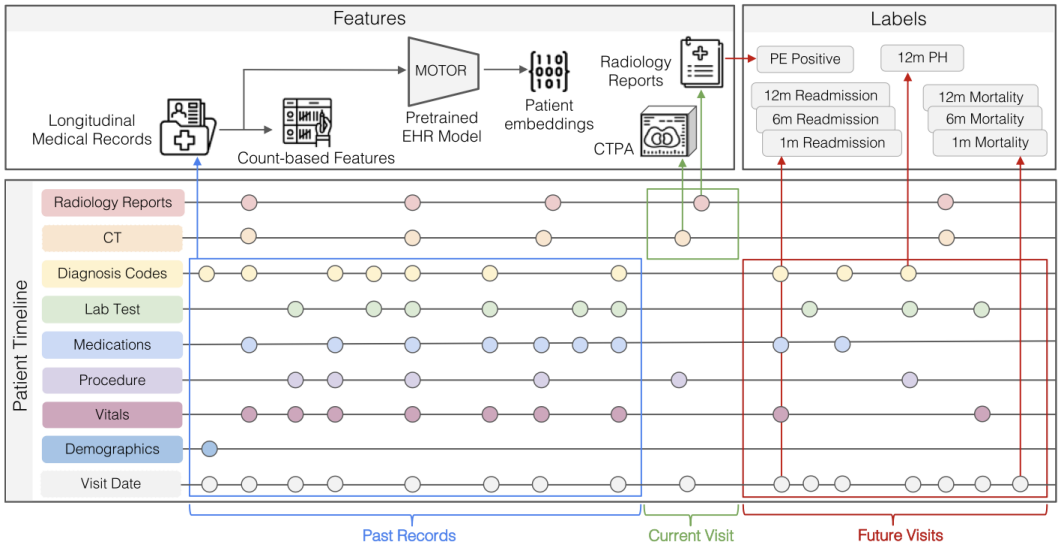

Figure 1: A diagram depicting the INSPECT dataset timeline and the INSPECT benchmark model architecture. [4]

a if there was prognosis of pulmonary hypertension for each patient. It was trained using the subset of the INSPECT dataset of 120 patients which have manually labelled diagnoses of PE.

5.3.2 Fusion model

Two independent models are used for the EHR data modality. The first approach encodes the EHR data by using a count featurisation method. The number of times each code appears in the patient EHR data is encoded into a column in a sparse matrix, and this matrix is input into a LightGBM model which is a gradient boosted decision tree model developed by Ke et al. [9]. The individual patient timelines from the EHR data are encoded into a MOTOR model, which is a self-supervised, time-to-event foundation model pre-trained on Stanford structured EHR data. The MOTOR backbone was frozen after pre-training and fitted with a linear probe.

The imaging model used takes an input of a set of pre-processed 2D CT slices for each 3D CTPA scan. The model takes a fixed number of slices, so if the number of slices for a scan is larger, then excess number of slices are selected randomly and discarded. These slices are fed into a slice encoder which is a ResNetV2 pre-trained using BigTransfer [10] and then finetuned. These embeddings are fed into a sequential model which is either a vision transformer model, a long short-term memory (RNN) or a gated recurrent unit (RNN). The paper does not say explicitly which model is used or if a combination of approaches are used.

A late fusion method is used; each model outputs a probability for each of the tasks and the probabilities for each task are combined using weights, w, which are learned on the validation dataset. The late fusion approach means that we can take predictions from the individual models or different combinations of models.

5.4 Results

5.4.1 Label generation results

The structured data-based pulmonary hypertension labeller had a recall of 0.91 and an accuracy of 0.80. The paper mentions that the low accuracy is likely because of the very small dataset used for training and testing (120 labels). The pulmonary embolism diagnostic labels generated from the radiology report impressions had a higher accuracy of 0.98, with only 12 out of 682 misclassifications.

5.4.2 Benchmark model results

Input Modality			**Diagnostic**	**Prognostic**						
Image	**EHR**		**PE**	**In-Hospital Mortality**			**Re-admission**			**PH**
CT	M	G	(+)	1 m	6 m	12 m	1 m	6 m	12 m	12 m
✓			<u>0.721</u>	0.794	0.755	0.748	0.549	0.515	0.525	0.661
	✓		0.677	<u>0.923</u>	<u>0.901</u>	<u>0.892</u>	<u>0.773</u>	<u>0.779</u>	<u>0.767</u>	0.824
		✓	0.681	0.848	0.865	0.855	0.737	0.740	0.728	<u>0.828</u>
✓	✓		0.761	**0.924**	0.903	0.895	0.774	0.777	0.764	0.820
✓		✓	0.765	0.867	0.875	0.866	0.740	0.736	0.722	0.830
	✓	✓	0.699	0.922	0.903	0.892	0.782	**0.786**	**0.774**	**0.849**
✓	✓	✓	**0.771**	**0.924**	**0.904**	**0.895**	**0.782**	0.784	0.771	0.843

Table 1: Results table taken from [4] giving the area under the receiver operating characteristic AUROC performance for each of the individual models in the top 3 rows, and for the different combinations of fused models in the bottom 4 rows. M is the structured EHR based MOTOR model, and G is the structured EHR based gradient-boosted trees model. The best overall models are bolded and the best individual models are underlined.

As expected, the best individual model at the diagnostic tasks is the imaging model, as the CTPA scans are typically taken with the purpose of diagnosing PE. In addition, we see the EHR models are best at the prognostic tasks, which again the paper mentions would be expected by clinicians. For the combined modality model, a significant improvement in the AUROC score of 7% is seen when the EHR models are combined with the imaging model. However, this paper did not find any significant improvements in the prognostic tasks when the modalities were combined. Indeed, the best AUROC score in the PH diagnostic task was the two combined EHR models, and it was made worse when the models were fused with the imaging model. The paper mentions that this is unexpected as clinicians have reported to using a past CTPA scan of pulmonary embolism when making a later pulmonary hypertension diagnosis.

6 Summary & Discussion

6.1 Discussion

We see that the MOTOR based model had a better individual AUROC score than the LightGBM model, but we see the best performance on the EHR data when the two models are fused. The paper does not explain why two models were selected for the EHR data. As both models are completely independent until a later fusion. The paper also did not comment or compare upon the performance of the two different EHR models.

The successful improved model performance when the EHR data models and imaging models are combined shows that AI imaging diagnostic models can be improved when given a patients past electronic health record data. This shows that these multi-modal medical datasets are important in developing diagnostic AI tools as it shows a case where a AI model can provide more accurate classification of disease from a CT scan by having access to patients' medical history. To improve upon AI methods of combining text and vision based medical data medical datasets like this one are required to test different multi-modal fusion methods. Most vision-language datasets from the computer vision field lack the detail and nuance seen in medical datasets where entire radiology reports are written based on one image compared to a non medical image dataset where images would typically be described with a few words or one sentence. For this reason, generic datasets combining text and language data likely are insufficient for finding the best multi-modal AI models for diagnostic tasks. Consequently, the impact of the dataset presented in this paper will likely be useful in testing and improving multi-model classification methods for medical diagnoses. [3]

However, the prognostic tasks were not improved by combining the imaging and EHR data, which goes against expectations from clinicians. The paper states this model is put forward just as a benchmark and not as a state of the art multi-modal model. Perhaps the method of label generation needs to be improved as the reported accuracy of just 0.80 is low. With incorrect labels the models may be learning some features that show signs of PH incorrectly. It may be that with a different model and different fusion method we will be able to see the prognostic tasks improved when the CTPA and EHR data are combined. The paper presents this dataset with the expectation that future researchers will build upon and improve their benchmark model methods. Indeed, their late fusion approach is just one way of combining multiple data modalities. Other vision-language models use earlier fusion methods where the text and image embeddings are seen by a fusion module encoder [3]. This method perhaps could offer an improvement as it may be that certain combinations of features in the EHR and CTPA data would offer information relevant to prognoses. There is a potential chance treating the data types entirely independently loses some of the information that requires knowing both imaging and text features in combination.

6.2 Critical Analysis

The figure 1 shown in the Inspect paper gave a very clear and detailed overview of the dataset and benchmark model. The paper was great with its use of figures to aid explanations. Nonetheless, it did lack some clarity when explaining the reasoning behind the methods used, as it did not offer any explanation or analysis for the use of two independent EHR models and also the methods used in the sequential model used in the CTPA model were not well explained. Moreover, the background section of the paper perhaps could be improved by a more detailed medical information around pulmonary hypertension and its link to pulmonary embolism. It could be concluded from reading this paper that pulmonary hypertension is directly caused and related to pulmonary embolisms. However, pulmonary emboli (the blood clots) are just one of multiple types of high blood pressure in the pulmonary arteries. Other causes include lung disease, hypoxia, left-sided heart disease. This paper was submitted to the Neurips 2023 conference, so the expected audience were unlikely to have the relevant medical background knowledge on pulmonary hypertension. Additional information about how the PH is related to the pulmonary embolism cases seen in this dataset would be very helpful.

In addition, a small error is made in two of the statistics presented in this paper. It stated that pulmonary embolism is responsible for 'nearly 300,000 hospital admissions and approximately 180,000 fatalities each year in the United States'. However, the study they cite, Horlander et al. 2003, found that deaths per year dropped from 35 750 in 1979 to 24,947 in 1998. It also did not give an exact estimate hospital admissions, but estimated a figure of 250,000. Perhaps, a better, more recent study to have used would be from Virani et al. 2021. It found PE mortality at 8809, and 'any-mention mortality' at 36 494

in 2021 [11].

6.3 Conclusion

In conclusion, this paper introduces a novel longitudinal dataset for the diagnosis of pulmonary embolism and pulmonary hypertension. INSPECT is the largest multi-modal dataset which integrates 3D medical imaging with EHR data [4]. This dataset enables the development of AI multi-modal fusion methods for the purpose of making medical diagnoses and prognoses. The benchmark method was able to show an improvement in the diagnose accuracy of pulmonary hypertension when imaging data was combined with textual data, however was unable to provide the expected improvement for the prognosis of pulmonary hypertension when combing imaging and EHR modalities. Moreover, while the dataset does have a small number of prognostic and diagnostic labels, there is not a full set of ground truth labels so label generation methods are required for training, potentially reducing training accuracy if the generated labels are incorrect.

This benchmark model was presenting as a starting point for future work. It seems likely this dataset will be used by future medical AI researchers, fine-tuning multi-modal AI methods for medical specific data. Moreover pushing AI diagnostic models to use multiple data types and also make predictions about the disease progression, allowing models to get closer to clinical reality by considering patient's medical background.

7 Funding and Acknowledgements

This work was funded by the Ai4Health centre for doctoral training programme at Imperial college. The Ai4Health cdt is funded by the UKRI, grant number EP/S023283/1.

References

[1] Adones Leslie, AJ Jones, and PR Goddard. The influence of clinical information on the reporting of ct by radiologists. *The British journal of radiology*, 73(874):1052–1055, 2000.

[2] Deaths in 2019, 2021, and 2022 due to blod clots, office for national statistics 2022, 2022.

[3] Prashant Shrestha, Sanskar Amgain, Bidur Khanal, Cristian A Linte, and Binod Bhattarai. Medical vision language pretraining: A survey. *arXiv preprint arXiv:2312.06224*, 2023.

[4] Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Curtis Langlotz, Matthew Lungren, Serena Yeung, Nigam Shah, and Jason Fries. Inspect: A multimodal dataset for patient outcome prediction of pulmonary embolisms. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[6] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[7] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022.

[8] Imon Banerjee, Yuan Ling, Matthew C Chen, Sadid A Hasan, Curtis P Langlotz, Nathaniel Moradzadeh, Brian Chapman, Timothy Amrhein, David Mong, Daniel L Rubin, et al. Comparative effectiveness of convolutional neural network (cnn) and recurrent neural network (rnn) architectures for radiology text report classification. *Artificial intelligence in medicine*, 97:79–88, 2019.

[9] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[10] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[11] Salim S. Virani, Alvaro Alonso, Hugo J. Aparicio, Emelia J. Benjamin, Marcio S. Bittencourt, Clifton W. Callaway, April P. Carson, Alanna M. Chamberlain, Susan Cheng, Francesca N. Delling, Mitchell S.V. Elkind, Kelly R. Evenson, Jane F. Ferguson, Deepak K. Gupta, Sadiya S. Khan, Brett M. Kissela, Kristen L. Knutson, Chong D. Lee, Tené T. Lewis, Junxiu Liu, Matthew Shane Loop, Pamela L. Lutsey, Jun Ma, Jason Mackey, Seth S. Martin, David B. Matchar, Michael E. Mussolino, Sankar D. Navaneethan, Amanda Marma Perak, Gregory A. Roth, Zainab Samad, Gary M. Satou, Emily B. Schroeder, Svati H. Shah, Christina M. Shay, Andrew Stokes, Lisa B. VanWagner, Nae-Yuh Wang, Connie W. Tsao, and null null. Heart disease and stroke statistics—2021 update. *Circulation*, 143(8):e254–e743, 2021.