ISARIC

International Severe Acute Respiratory and emerging Infections Consortium

A global federation of clinical research networks, providing a proficient, coordinated, and agile research response to outbreak-prone infectious disease

Analysis Plan for ISARIC International COVID-19 Data

This document details the plan for an analysis of COVID-19 patient data submitted to the ISARIC database. Details of this cohort can be found in the MedRxiv reports available here: https://www.medrxiv.org/content/10.1101/2020.07.17.20155218v15

All contributors to the ISARIC COVID-19 database are invited to participate in this analysis through review and input on the statistical analysis plan and resulting publication. The outputs of this work will be disseminated as widely as possible to inform patient care and public health policy, this will include submission for publication in an international, peer-reviewed journal. ISARIC aims to include the names of all those who contribute data used in this analysis as cited collaborators on the resulting publication, subject to the ISARIC publication policy.

Title of proposed research
Automatic Identification of Signals in Clinical Notes using Natural Language Processing
Working Group Lead(s) (name, ORCID ID, email, institution, country) [1]
Omid Rohanian, University of Oxford, UK, omid.rohanian@eng.ox.ac.uk
Working group members (name, ORCID ID, email, institution, country)
Additional researchers based at institutions in low- or middle-income countries are invited to join the working group. Hannah Jauncey, University of Oxford, UK, hannah.jauncey@ndm.ox.ac.uk Christiana Kartsonaki, University of Oxford, UK, christiana.kartsonaki@dph.ox.ac.uk Bronner Gonçalves, University of Oxford, UK, bronner.goncalves@ndm.ox.ac.uk Laura Merson, University of Oxford, UK, laura.merson@ndm.ox.ac.uk David Clifton, University of Oxford, UK, david.clifton@eng.ox.ac.uk Jose A. Calvache, Universidad del Cauca, Colombia, jacalvache@unicauca.edu.co
Data Analyst (name, ORCID ID, email, institution, country)

[1] Either chair and/or co-chair are based in an institution in an LMIC. If you would like to be connected with a potential partner researcher, please let us know at data@isaric.org.

Muhammad Nasir Khoso, 0000-0002-7135-2252
drmuhammadnasirkhoso@gmail.com, South City Hospital Pakistan
Omid Rohanian, University of Oxford, UK, omid.rohanian@eng.ox.ac.uk

Research Plan

Background, Research Objectives, Scientific Value

The ISARIC data include vast amounts of multi-lingual free text data entries that challenge our ability to perform meaningful analysis. These non-standard data use a range of terms and definitions to describe equivalent clinical events making harmonisation very challenging. Experience from the COVID-19 pandemic revealed that existing manual practices in data curation cannot scale to datasets of larger size and complexity than those considered to date. This process will be supported by automated ML algorithms that will permit human curation to scale to (i) datasets of larger size, and (ii) datasets of larger complexity, as is required for the next generation of the ISARIC platform. Working with partners across 60 countries has brought significant challenges in harmonising the international diversity of languages, units, data collection frameworks and clinical definitions so that they can be used in robust pooled analysis. Data undergo curation to restructure, convert, clean, and align them into equivalent formats that can be compared. Curation of disparate datasets is a time- and expertise-intensive process which has been a limiting factor in our ability to make use of the rich data shared by ISARIC partners, especially free text data where comorbidities, unusual symptoms and novel treatments are often recorded. Developing and validating approaches to identify and standardise clinically relevant terms recorded as free text would accelerate our curation process and enable access to these important variables.

We have experience in developing semi-autonomous approaches to data curation with data from 1.5 million patient admissions from electronic health records across the United Kingdom. This experience will be applied to innovate methods for the large volumes of multi-language, unstructured data in the ISARIC database to match controlled terminologies that can be efficiently analysed. Natural Language Processing (NLP) methods have advanced rapidly within recent years, allowing ML to identify structure in free-text data that may be represented in multiple different languages in ISARIC – this multi-lingual complexity is a key feature of the ISARIC platform, given its global reach, and therefore complex NLP methods based on modern ML approaches (“transformers”) are required to support curation of the data. Our exemplar will include the automated identification of cancer patients, across multiple languages, which will be compared directly with manually-obtained labels in this proof-of-principle study.

Proposed Target Population

To maximise the representation of all populations affected by COVID-19, we will include terms from the entirety of the ISARIC dataset, including all patients from all participating countries.

Outputs

We will construct tailored NLP models for direct use with ISARIC free-text fields, comparing (i) conventional, lightweight ML models (e.g. logistic regression and SVM) and (ii) state-of-the-art models based using deep learning ("transformers", etc.). A first proof-of-principle approach will be to construct models that automatically identify cancer patients from free-text data, across all ISARIC records annotated to date, with NLP metrics (precision, recall, and F-score) in addition to conventional metrics (AUC, etc.) compared with manually-curated labels obtained previously.

Outputs will include:

- Open-source models of types (i) and (ii) made available for use with ISARIC
- Labels for adding into the ISARIC record (e.g., "Cancer Patient?")
- An in-depth analysis of the strengths and weaknesses of the models developed and a quantitative assessment of where the state-of-the-art is and how future work can improve upon current methods

This approach will be extended to additional terms and signals of novel findings.

Research Questions, Challenges, and Methodology

Clinical notes contain important information about patients and given that they contain misspellings, abbreviations, non-standard language and sometimes broken grammar, it is a challenge to process them automatically. The following are our research questions:

1. Can representation learning help us automatically identify cancer-related notes in a multi-lingual dataset of a large magnitude?
2. To what degree can we utilise standard non-neural models to address the above task?
3. Recurrent Neural Networks (RNNS) have an inherent understanding of order. When equipped with pre-trained embeddings, can RNNs improve upon the performance of non-neural models?
4. Given the challenging nature of this task, can we produce outputs that would be helpful to human annotators? What are the strengths and weaknesses of such models?
5. Can feature engineering help with the performance of the models? For instance, can the inclusion of demographic attributes help in this task?

In our experiments, we employ methods and tools from representation learning to construct informative embeddings for our task. Embeddings are real-valued vector representations for each word or concept. Since language is discrete and machine learning models only understand numerical representations, embeddings are a way to convert a discrete entity into a vector of numbers that could be utilised by the learning model to learn different nuances about it. We will use a standard embedding technique to represent our data.

In the first round of the experiments, we use a simple logistic regression classifier to classify clinical notes as cancer+ or cancer- using an aggregated representation. To build that classifier we rely on the annotated portion of the dataset (approx. 3K rows) which were classed into different subcategories of ‘neoplasm’. In the initial phase of the project, it was agreed that we focus on developing a binary classifier that would identify rows that are potentially related to cancer, with the hope that it can help human annotators to focus on a subset of the dataset. Multi-class classification was left for the future phase when we have a good understanding of how the models perform in the simpler binary case.

Using the cancer pilot as the example, there are two main challenges at this stage of the work:

1. The annotations only targeted potential cases of cancer. However, to train a classifier, we also need a negative cohort so as to teach the model what should not be classed as cancer-related. We rely on keyword-based filtering to exclude rows that contain a potentially cancer-related row. This approach is inevitably noisy, but has proved practically viable.

2. When producing the final output on the larger unannotated file, we lack a numerical metric to assess the performance of the system and have to simply rely on impression of the clinicians or ask annotators whether they find the outputs practically useful.

To address the above shortcomings, a separate gold standard is developed, initially of size 1000, where rows are manually annotated for both positive and negative cases of cancer. These 1000 rows are then taken as the gold standard and the trained models can be first tested on this unseen data. In this way we can have a numerical metric to discuss effectiveness of the models. This gold standard has already been prepared. The obvious issue is the lopsidedness of the labels. Out of 1000 rows, 969 are negative and only 31 are cancer-related. Nonetheless, this reflects the distribution of the labels occurring naturally in the dataset as the rows were not pre-selected for a better representation of the minority class.

To specifically address issue 2, we can produce the outputs on the entirety of the unlabelled data, and then randomly select a subset of the rows and ask a human expert to label them. Subsequently, we can compare the human- and computer-generated labels. Using kappa inter-annotator agreement, we can assess the performance of the different baselines and the one that is closest to human judgement could be taken as the most accurate.

As for the methodology employed, we have a non-neural logistic regression baseline model that uses static weights. This is our strong baseline. We can add an SVM to have a better comparison between baselines. These models are ii) non-neural and ii) lack an inherent understanding of linearity and order in text.

For the neural models, we have developed three separate RNNs (Bi-LSTM, CNN-Bi-LSTM, and CNN-Bi-LSTM-Attention) to address the task and to compare with simpler methods. All these three use BioSentVec (Chen, Peng, & Lu, 2019) for the pretrained static vectors which seem to have a good coverage for tasks of this kind. Using a standardised list of keywords that could signal cancer, we construct a negative cohort based on simple keyword-based filtering which will be applied to the entire dataset. We train the three models on the positive and negative cohorts combined and test the trained models on the gold standard. We use f-score, accuracy, recall, and precision to report the results. The models at the moment are not hyper-parameter optimised and their training is not monitored on a separate DEV set and they are simply trained for a set number of epochs. At the moment, the simpler Bi-LSTM model that uses pretrained embeddings is performing best out of the three. A confusion matrix will be used to graphically show where the model is performing best and where it is being inaccurate.

For evaluation, we can consider training/testing this approach for datasets in different calendar time periods (e.g. 1st vs. 2nd pandemic year) as profile of patients might have changed. We can also stratify our test samples based on the degree of detailedness of notes.

References

Chen, Q., Peng, Y., & Lu, Z. (2019, jun). BioSentVec: creating sentence embeddings for biomedical texts. In 2019 IEEE international conference on healthcare informatics (ICHI).
IEEE. Retrieved from https://doi.org/10.1109%2Fichi.2019.8904728 doi: 10.1109/ichi.2019.8904728

Data Required

(DM) Demographics Domain
(SA) Clinical and Adverse Events Domain