Natural Language Processing for Clinical Laboratory Data Repository Systems: Implementation and Evaluation for Respiratory Viruses

Elham Dolatabadi¹,⁴, Branson Chen², Sarah A. Buchan²,³,⁵, Alex Marchand-Austin², Mahmoud Azimaee²,⁴, Allison J. McGeer⁴,⁵,⁷,⁸, Samira Mubarek⁶,⁷, Jeffrey C. Kwong²,³,⁵,⁶,⁹,¹⁰

¹ Vector Institute, Toronto, Ontario, Canada
² ICES, Toronto, Ontario, Canada
³ Public Health Ontario, Toronto, Ontario, Canada
⁴ Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
⁵ Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
⁶ Sunnybrook Research Institute, Toronto, Ontario, Canada
⁷ Dept of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada
⁸ Sinai Health System, Toronto, Ontario, Canada
⁹ Department of Family and Community Medicine, University of Toronto, Toronto, Ontario, Canada
¹⁰ University Health Network, Toronto, Ontario, Canada

Abstract

Background: With the growing volume and complexity of laboratory repositories, it has become tedious to parse unstructured data into structured and tabulated formats for secondary uses such as decision support, quality assurance, and outcome analysis. However, advances in Natural Language Processing (NLP) approaches have enabled efficient and automated extraction of clinically meaningful medical concepts from unstructured reports.

Objective: In this study, we aimed to determine the feasibility of using the NLP model for information extraction as an alternative approach to a time-consuming and operationally resource-intensive handcrafted rule-based tool. Therefore, we sought to develop and evaluate a deep learning-based NLP model to derive knowledge and extract information from text-based laboratory reports sourced from a provincial laboratory repository system.

Methods: The NLP model, a hierarchical multi-label classifier, was trained on a corpus of laboratory reports covering testing for 14 different respiratory viruses and viral subtypes. The corpus included 85k unique laboratory reports annotated by eight Subject Matter Experts (SME). The model’s performance stability and variation were analyzed across fine-grained and coarse-grained classes. Moreover, the model’s generalizability was also evaluated internally and externally on various test sets.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Results: The NLP model was trained several times with random initialization on the development corpus, and the results of the top ten best-performing models are presented in this paper. Overall, the NLP model performed well on internal, out-of-time (pre-COVID-19), and external (different laboratories) test sets with micro-averaged F1 scores >94% across all classes. Higher Precision and Recall scores with less variability were observed for the internal and pre-COVID-19 test sets. As expected, the model’s performance varied across categories and virus types due to the imbalanced nature of the corpus and sample sizes per class. There were intrinsically fewer classes of viruses being detected than those tested; therefore, the model’s performance (lowest F1-score of 57%) was noticeably lower in the “detected” cases.

Conclusions: We demonstrated that deep learning-based NLP models are promising solutions for information extraction from text-based laboratory reports. These approaches enable scalable, timely, and practical access to high-quality and encoded laboratory data if integrated into laboratory information system repositories.

Keywords: Natural Language Processing; Clinical Laboratory Information Systems; Deep Learning; Classification; Respiratory Virus Reports; Infectious disease

Introduction

Clinical laboratory data account for a large proportion of data stored in electronic health record systems worldwide and present a wealth of information vital for evidence-based decision-making and public health improvement. Laboratory information systems record, manage, and store laboratory test data to facilitate reporting to clinicians and jurisdictional laboratory information repositories. These repositories often include test orders and results from various laboratory service providers, such as hospitals, public health agencies, and private companies, and are populated as part of clinical care.

Several factors limit the secondary use of laboratory data for other purposes. The most important are concerns about the quality of the data, lack of standardization, and difficulty in extracting the needed information. Laboratory data varies over time due to evolving standards of care and changing population demographics. Furthermore, specific categories of laboratory data are reported as free text in an unstructured format with no standard vocabulary in the actual contents, which adds more complexity for their secondary uses. Therefore, efforts are needed to eliminate redundancies, extract the necessary information, and derive accurate interpretations from laboratory data.

Our institute has developed a specific information extraction workflow to manage the interpretation of a large volume of provincial clinical laboratory results, as shown in Figure 1. The workflow, called a semi-rule-based workflow, relies on time-consuming and operationally resource-intensive approaches, including a library of rule-based and hand-crafted tools. These tools are explicitly programmed for various laboratory result categories and must be refined continually. To address
challenges with our existing semi-rule-based workflow and automate the exhaustive information retrieval task, we built a deep learning-based natural language processing (NLP) tool. The objective of this study was to assess the feasibility of our deep learning-based NLP model and evaluate its performance relative to the semi-rule-based workflow.

The development of NLP methods is essential to automatically transform laboratory reports into a structured representation that scales data usability for research, quality improvement, and clinical purposes. NLP enables automated extraction of information, and its use in the clinical domain is growing with increasing uptake in various applications such as biomedical named entity recognition, summarization, and clinical prediction tasks. More recently, deep learning approaches such as convolutional neural networks (CNN), recurrent neural networks (RNN), and RNN variants such as bidirectional Long Short-Term Memory (Bi-LSTM) have been successfully applied to clinical NLP tasks. They are now considered the baseline techniques for various information extraction tasks.

Figure 1. Semi-rule-based workflow vs. Fully automated Deep Learning NLP approach. Semi-rule-based relies on time-consuming and operationally resource-intensive approaches for the information extraction task. The corpus was derived from the Ontario Laboratories Information System (OLIS). OLIS has over 100 contributors, including hospital, commercial, and public health laboratories. Following basic text-cleaning steps, around 87k unique laboratory reports were collected and included in our corpus to be used in parallel by both approaches: Semi-rule-based and Deep Learning NLP. Semi-rule-based workflow is a multistep procedure where all the unique reports were grouped by the LOINCs, year, and location in the first step. In the second step, SMEs created a list of dictionaries for terms related to the different viruses and strains and a set of if-then-else rules to generate interpretations and extract information from each
In this study, we focused on automating the retrieval of information related to respiratory viruses from the laboratory repository of Canada's most populous province. Respiratory viruses account for a substantial burden of disease globally, causing respiratory and non-respiratory illnesses. It is impossible to distinguish which respiratory virus is causing infection based on clinical examination alone, necessitating laboratory testing for confirmation. We sought to (1) implement a deep-learning-based NLP predictive model to extract respiratory virus information from the laboratory repository and (2) evaluate the generalizability and robustness of predictions (extracted information) across different categories of respiratory viruses and test sets. This study's findings can inform public health practitioners and researchers about using NLP approaches to empower and facilitate access and retrieval of information from a collection of text-based laboratory reports without any previous time-consuming handcrafted rule-based approaches. Furthermore, it leads to a scalable and easily deployable automated information extraction tool.

Methods

Study Design

The dataset used in this study is a collection of laboratory reports that cover testing for 14 different respiratory viruses and viral subtypes (Table 1), most of which are in the form of texts. The reports are text-based and require cleaning, parsing, and encoding.

The dataset was derived from the Ontario Laboratories Information System (OLIS). OLIS has over 100 contributors, which comprise hospital, commercial, and public health laboratories, adding to the complexity and variability of the clinical data. These data were analyzed at ICES (formerly the Institute for Clinical Evaluative Sciences).

The automated encoding of laboratory testing reports into respiratory viruses is framed as a multi-label classification task according to a hierarchy. There are two levels of classification hierarchy; at each level, the classification is multi-label. Each input text sequence can be assigned to a non-empty subset of various labels, as shown in Figure 2. In the first level of the hierarchy, the classifier assigns outputs to mutually non-exclusive fine-grained labels. The fine-grained labels are reassigned to a coarse-grained set of labels in the second level of the classification hierarchy. In this work, “sequence” refers to the input laboratory reports to the NLP model, which may be a single or several sentences. A “label” also refers to a status of a specific virus or strain being tested or Detected (e.g., influenza A is detected). Tested and Detected are not mutually exclusive; it must first be determined if the specimen is
tested for any virus and then flagged as detected if the result is positive. **Detected** is a subset of the **Tested**.

**Table 1:** Details of the respiratory viruses embedded in text-based laboratory reports derived from the Ontario Laboratories Information System (OLIS). Specimens may be tested for one or more of the following viruses: *influenza, respiratory syncytial virus (RSV), adenoviruses, seasonal coronaviruses, enterovirus/rhinoviruses, parainfluenza viruses, human metapneumovirus (HMV), and bocavirus*. The testing modalities employed include single and multiplex polymerase chain reaction, direct fluorescent antibody, viral culture, and enzyme immunoassay rapid antigen tests. Repeated testing may involve multiple laboratories and testing modalities.

<table>
<thead>
<tr>
<th>Viruses</th>
<th>Mention Counts (#)*</th>
<th>Tested* (%)</th>
<th>Detected* (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenovirus</td>
<td>21614</td>
<td>45</td>
<td>2</td>
</tr>
<tr>
<td>Bocavirus</td>
<td>5112</td>
<td>96</td>
<td>5</td>
</tr>
<tr>
<td>Coronavirus (seasonal)</td>
<td>9128</td>
<td>95</td>
<td>9</td>
</tr>
<tr>
<td>Any Influenza</td>
<td>49282</td>
<td>78</td>
<td>35</td>
</tr>
<tr>
<td>Influenza A</td>
<td>44753</td>
<td>80</td>
<td>30</td>
</tr>
<tr>
<td>Influenza A H1</td>
<td>6797</td>
<td>N/A†</td>
<td>17</td>
</tr>
<tr>
<td>Influenza A H3</td>
<td>9929</td>
<td>N/A†</td>
<td>18</td>
</tr>
<tr>
<td>Influenza B</td>
<td>40840</td>
<td>78</td>
<td>12</td>
</tr>
<tr>
<td>Enterovirus/Rhinovirus</td>
<td>13262</td>
<td>92</td>
<td>19</td>
</tr>
<tr>
<td>Human Metapneumovirus (HMV)</td>
<td>21194</td>
<td>46</td>
<td>3</td>
</tr>
<tr>
<td>Parainfluenza</td>
<td>21584</td>
<td>46</td>
<td>4</td>
</tr>
<tr>
<td>Any Respiratory Syncytial Virus (RSV)</td>
<td>38080</td>
<td>68</td>
<td>11</td>
</tr>
<tr>
<td>RSV A</td>
<td>11227</td>
<td>N/A†</td>
<td>2</td>
</tr>
<tr>
<td>RSV B</td>
<td>11094</td>
<td>N/A†</td>
<td>3</td>
</tr>
</tbody>
</table>

*Mention Counts (#) represents the counts of specific virus terms from all the distinct laboratory reports (unique sequences). It does not provide any clinical information regarding the prevalence of the above-mentioned viruses in our region.

* **Tested** and **Detected** represent the proportion of mentions flagged as tested or positively detected by our parser, respectively. Note: **Tested** and **Detected** are not mutually exclusive; we first determine whether it was tested for (i.e., have a result) and then flag it as detected if the result is positive. **Detected** is a subset of the **Tested**.

†The subtypes of *Influenza A* and *RSV* were only analyzed for detection but not testing, as the scope of the planned analyses for using the respiratory virus data was primarily focused on the larger virus categories.
To summarize, the information extraction for an input text sequence involves retrieving virus types and identifying their status as being tested and/or detected. Figure 2 illustrates a running example of the input and output of the deep learning-based NLP model.

**Figure 2.** The fully automated deep learning-based NLP approach is a hierarchical-based multi-label classification task that retrieves virus (or strain) types and identifies their status as being tested and/or detected. Each input sequence was assigned to 24 mutually non-exclusive fine-grained labels in the first level of the hierarchy. In the second level of the classification hierarchy, the fine-grained labels were reassigned to a coarse-grained set of labels (k=6). An example laboratory report processed through the deep learning-based NLP approach for automated extraction and encoding of information is also shown.

Notes: A sequence refers to the input laboratory reports to the NLP approach, which may be a single or several sentences. A label also refers to a status of a specific virus or strain (tested or detected). *Influenza is tested* implies it was tested for any Influenza type; however, the total number of ‘Influenza is tested’ is greater than the total number of Influenza A tested + Influenza B tested since not all Influenza is mentioned for types. The same applies to “Influenza is detected” and “RSV is tested”.

**Corpus Development Description**

**Ontario Laboratories Information System (OLIS)**

To create the corpus for this study, over a million observations corresponding to 99 unique Logical Observation Identifiers Names and Codes (LOINC)s were pulled from OLIS, and the text-based laboratory results were extracted from the observations. OLIS was created and is managed by Ontario Health, from whom ICES receives an
ongoing data feed. At the time of writing, the OLIS data holding at ICES consists of >9,000 unique LOINCs and >5 billion laboratory observations across 150 laboratory test centers in Ontario. As such, the clinical laboratory data has considerable complexity and variability.

**Development of the Ground Truth**

In this study, we leveraged the semi-rule-based workflow, an information extraction workflow relying on a library of rule-based and hand-crafted tools, to create ground truth for the deep learning model. A group of eight Subject Matter Experts (SME), including two infectious disease epidemiologists, two infectious disease microbiologists, a genomic specialist, a research methodologist, a data analyst, and a machine learning scientist, were engaged in performing the required tasks in the workflow. These tasks include basic text cleaning, quality checking, and rule-based algorithm development for interpreting reports, as shown in Figure 1. In our institute, LOINCs are mainly used to filter OLIS observations into relevant groupings (e.g., respiratory viruses) and not for encoding and interpretation since they are not always used appropriately by those entering the data into OLIS. Consequently, the SMEs identified a list of 99 LOINCs related to respiratory viruses, and all the laboratory reports in OLIS corresponding to these LOINC codes were retrieved. The workflow consists of the following tasks:

1. The data analyst and data scientist first scanned the text strings and, after performing basic text cleaning (e.g., removing punctuations, stop words, case normalization, lemmatization, and stemming) and removing duplicates, they created a meaningful list of 87k unique laboratory reports.

2. The unique reports were grouped by laboratory and facility names, LOINCs, and year. Then, three SMEs, including two analysts and an infectious disease specialist, manually reviewed multiple samples per group and created a knowledge base and sets of if-then-else rules to generate interpretations for each laboratory report. Specifically, the knowledge base consisted of dictionaries for terms related to the different viruses and strains. The if-then-else rules provided instructions for grouping virus terms with respective results packaged as a python library, which we refer to in this study as the rule-based text parser.

3. Following the initial development of the rule-based text parser, it was improved based on inputs from three other SMEs in an iterative manner. The text parser was applied to the entire corpus to generate annotations at each iteration. Then the data analyst manually reviewed the interpretations and flagged unclear results to be reviewed by SMEs at another iteration. In addition, a small random sample of unflagged test results was provided to SMEs to be reviewed at this iteration. The SMEs subsequently reviewed the list and provided new rules to be added to the text parser. This procedure was repeated until there were no more flagged test results.
Model Development and Evaluation

NLP Model Description
The deep learning-based NLP model consisted of three components that were trained jointly: the Word Embedding layer, the Bidirectional long short-term memory (Bi-LSTM) layer, and the Output layer. The word embedding layer computed a vector representation of each word in the text as a combination of a character-based representation learning model, and word vectors initialized with pre-trained GloVe embedding. The embedding layer was coupled with a Bi-LSTM on top of it to generate conceptually and contextually meaningful representations of words. An output layer of size equal to the number of distinct labels was placed on top of Bi-LSTM, and the last hidden state of the Bi-LSTM was projected into the output layer.

Model Evaluation
The model's robustness and generalizability were evaluated internally and externally on various test sets, as in Table 2. The internal test set used for model training was a randomly sampled subset representing 10% of the laboratory reports from OLIS from 2007 to 2018. The performance of the model was also evaluated on two out-of-time test sets, including samples from an entirely different time period: (1) a large pre-COVID-19 (2019) sample and (2) a small post-COVID-19 (2020) sample. A separate test set, denoted as the external test set, included samples up to 2019 from two separate laboratories (testing sites not included in the development of the model) and was used to assess the external generalizability of the model. F1, Precision, and Recall scores were calculated for the model's predictions. The paired t-test was used to determine whether a statistically significant difference in the F1 scores between classes and test sets existed. In addition, 95% confidence intervals were calculated for the Precision and Recall scores to quantify the uncertainty of the model's estimates.

Results
The development corpus, including training and test sets, included 87,500 sequences involving ~5 million tokens (summary statistics for the datasets are shown in Table 2). The NLP model was implemented in Tensorflow on an NVidia Tesla GPU, and Adam was used as the optimization algorithm. The maximum sequence length was fixed to 400 words. The model was trained several times with random initialization on the development corpus, and the results of the top ten best-performing models on the test sets are presented here. The results for the fine-grained classification in the first level of the hierarchy are aggregated by micro-averaging across fine-grained 24 labels and are shown in Table 3. The detailed performance for each label is also shown in Supplementary Table 1. The F1-score performance of the model in the second level of the hierarchy, coarse-grained multi-label classification, for 'Any Influenza', 'Any RSV', and 'Any Virus' are also shown in
Table 3. In addition, the variation of the model’s Precision and Recall scores using barplots and 95% confidence intervals are shown in Figure 3.

Table 2 Dataset statistics for laboratory descriptions of the development and test sets.

<table>
<thead>
<tr>
<th>Cohorts</th>
<th>seq(#)</th>
<th>Any Influenza virus* detected (#)</th>
<th>Any Influenza virus* tested (#)</th>
<th>Any RSV virus detected (#)</th>
<th>Any RSV virus tested (#)</th>
<th>Any virus detected (#)</th>
<th>Any virus tested (#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(years 2009-2018)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training set</td>
<td>60471</td>
<td>13792</td>
<td>35292</td>
<td>3959</td>
<td>27196</td>
<td>22284</td>
<td>40652</td>
</tr>
<tr>
<td>Internal test set</td>
<td>6719</td>
<td>1604</td>
<td>3941</td>
<td>428</td>
<td>3009</td>
<td>2541</td>
<td>4534</td>
</tr>
<tr>
<td>Out-of-time test sets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>post-COVID-19 (2020)</td>
<td>100</td>
<td>n/a</td>
<td>11</td>
<td>&lt;6</td>
<td>11</td>
<td>&lt;6</td>
<td>27</td>
</tr>
<tr>
<td>External test set</td>
<td>4213</td>
<td>964</td>
<td>3020</td>
<td>261</td>
<td>2546</td>
<td>1431</td>
<td>3237</td>
</tr>
</tbody>
</table>

Table 3 The prediction results (F1-score) of the top 10 best-performing models on the in-time, out-of-time, and external test sets. The fine-grained results are aggregated by micro-averaging across 24 fine-grained labels. The fine-grained results with the best-performing model are shown in Supplementary Table 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine-grained micro-averaged% (mean±std)</td>
<td>97.3±0.25</td>
<td>94.3±0.59</td>
</tr>
<tr>
<td>Coarse-grained Any Influenza virus Detected% (mean±std)</td>
<td>97.6±0.28</td>
<td>94.4±1.04</td>
</tr>
<tr>
<td>Coarse-grained Any Influenza virus Tested% (mean±std)</td>
<td>98.7±0.15</td>
<td>97.2±0.45</td>
</tr>
<tr>
<td>Coarse-grained Any RSV virus Detected% (mean±std)</td>
<td>90.9±1.7</td>
<td>81.5±3.63</td>
</tr>
<tr>
<td>Coarse-grained Any RSV virus Tested% (mean±std)</td>
<td>98.1±0.34</td>
<td>96.1±0.95</td>
</tr>
<tr>
<td>Coarse-grained Any virus Detected% (mean±std)</td>
<td>95.0±1.0</td>
<td>92.3±1.59</td>
</tr>
<tr>
<td>Coarse-grained Any virus Tested% (mean±std)</td>
<td>98.4±0.17</td>
<td>96.3±0.35</td>
</tr>
</tbody>
</table>

* seq(#) represents the counts of unique sequences; sequence refers to the input laboratory reports to the NLP model, which may be a single sentence or several sentences.
* Detected (%) and Tested (%) represent the aggregation of the proportion of any mentions of the virus terms from the total unique sequences in the dataset.
* The out-of-time (post-COVID-19) is a very small and imbalanced sample, including only 100 sequences with no mentions of any virus detected.
As expected, the performance on the internal test set was better than the out-of-time (pre-COVID-19) and external test sets. In this regard, the F1 score results of the test sets were compared, and noticeable differences were observed between the pairs of internal and out-of-time (pre-COVID-19) test sets, internal and out-of-time (post-COVID-19) test sets, and internal and external test sets. The out-of-time (post-COVID-19) test set was a small and imbalanced sample, including 100 sequences with <6 mentions of any virus as being detected. The sample included 12 sequences labelled as being tested for coronavirus, and our model correctly classified them with an F1 score of 0.67. Regarding the degree of uncertainty in the estimates, fewer variations in Precision and Recall scores are observed for the internal and out-of-time test sets (Pre-COVID-19). On the contrary, the estimates on the out-of-time (post-COVID-19) and external test sets have larger confidence intervals.

![Figure 3](image_url). The Precision and Recall scores of the predictions of the top 10 best-performing models with 95% confidence intervals. The fine-grained results are aggregated by micro-averaging across 24 fine-grained labels.

In general, the models’ estimates on any test sets were variable across classes with varying degrees of uncertainty. The averaged F1 scores of the estimates for both “fine-grained (micro-averaged)” and “coarse-grained Any virus” classes were above 90% on the internal test set. The F1 score for the coarse-grained ‘Any Influenza detected’ on all test sets was above 91%. Overall, the performance for “Coarse-grained detected” classes was lower than for “Coarse-grained tested” classes. Among the detected classes, the performance for ‘Any Influenza virus’ was evidently higher than ‘Any RSV virus’. The same result was observed between ‘Any Influenza virus’
and 'Any RSV virus'. Comparably, larger confidence intervals are evidenced for the 'Coarse-grained Any RSV detected' estimates.

Discussion

Principal Results

In this study, we demonstrated an implementation and evaluation of an NLP model for an automated and reductive information extraction task in a province-wide laboratory data repository. Our results suggest that the NLP model is a promising approach for information extraction from text-based laboratory reports as an alternative method to address the time-consuming and operationally resource-intensive nature of handcrafted rule-based models.

Model's Generalization across Various Test Sets

Overall, the NLP solution, which was a hierarchical multi-label classifier, performed well on internal, out-of-time (pre-COVID-19), and external (different laboratories) test sets; except for the internal test sets, the other test sets were sourced from either a more recent time period or other laboratory sites, but the model was able to generalize well with micro-averaged F1 score >94% across all classes. The performance of the model on the other out-of-time (post-COVID-19) test set was satisfactory, however, due to its small sample size with many underrepresented classes, it was not possible to draw any conclusion. The out-of-time (post-COVID-19) test set was pulled from the 2020 cohort to simulate a non-stationary production environment for observation.

Models' Stability and Performance Variation between Classes

In general, the model performances on any test sets were variable across classes and virus types due to the imbalanced nature of the corpus and sample sizes per class. There were intrinsically fewer classes of viruses detected compared with those tested. Therefore, the model's performance was noticeably lower in the "detected" cases. Among detected cases, the lowest performance was observed for RSV A, and the highest performance among the tested cases was observed for Influenza. Moreover, more considerable variations were observed for the positive predictive and sensitivity values of the "detected" classes and, in particular, for the "Any RSV virus detected" class.

Limitations

Although deep learning models promise great potential for digitized health data, putting the model into production and prospectively validating operational data is as crucial as model building and are a critical step in assessing and ensuring its operational effectiveness. However, we expect the model's performance to deteriorate as it goes into production, potentially impacting data quality. Moving forward, we plan to run a silent-period production validation to further prospectively explore the model's performance. During the silent period, our model will be integrated into the data quality and management workflow for the
laboratory data repository, and the outputs will be internally validated in a fashion that would avoid exposure to data users. We also plan to run rigorous evaluation and continuous refinement of the model in the silent period to assess its performance better before it enters production.

Another significant limitation of this study is that the model was only trained on respiratory virus laboratory reports. Even within that collection, some categories were naturally underrepresented, which impacted the model’s generalizability. So, during the silent period, more records from a diverse set of laboratory reports from various categories will be annotated and made available to the model, and the model will be updated accordingly.

**Comparison with Prior Work**

Deep learning-based NLP approaches have shown their efficiency in many clinical NLP tasks and have thoroughly permeated the informatics community. The existing body of literature has mainly focused on using deep learning models to extract and interpret cancer-related clinical concepts from free text or other clinically meaningful entities from radiology reports or hospital notes. Only one study at the time of writing explored the use of an NLP system, Topaz, for the automated extraction and classification of influenza-related terms from text emergency reports. To our knowledge, our study is the first to explore using deep learning models for efficient processing and extraction of clinically meaningful knowledge pertaining to respiratory viruses from a laboratory repository.

One strength of the NLP approach used in this study is its scalability for various text-based laboratory scenarios. As the size and complexity of laboratory data grow, so does the need for scalable and reusable tools for automated extraction of knowledge from vast amounts of clinical notes and quick generalization from one task to another. Manual processing of laboratory reports severely limits the utilization of rich information embedded in the data repositories and makes the process of data cleaning and quality improvement prohibitively expensive. Usually, the rules learned from cleaning one collection of laboratory reports show little generalizability toward other collections. On the other hand, deep learning-based NLP algorithms are well-poised to scale the information extraction process. Although building deep learning-based NLP models is computationally intensive and memory demanding, the benefit-to-cost ratio of these models in clinical settings will continue to increase.

**Conclusions**

The health industry is rapidly becoming digitized, and information extraction is a promising method for researchers and clinicians seeking quick retrieval of information embedded in texts. This study described developing and validating a deep learning-based NLP approach to extract respiratory virus testing information from laboratory reports. We demonstrated that our system can classify and encode large volumes of text-based laboratory reports with high performance without any
of the previous time-consuming handcrafted feature engineering approaches. Taken together, the findings in this study provide encouraging support that NLP-based information extraction should become an important component of laboratory information repositories to assist researchers, clinicians, and healthcare providers with their information and knowledge management tasks.

Acknowledgements
This study was a collaborative effort supported by the Vector Institute, an independent, not-for-profit corporation dedicated to research in the field of artificial intelligence (AI), and ICES, an independent, non-profit research organization that uses population-based health and social data to produce knowledge on a broad range of health care issues. Resources used in preparing this work were funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). Parts of this material are based on data and information compiled and provided by the Ontario Ministry of Health. This work was also supported by a SickKids-Canadian Institutes of Health Research New Investigator Grant in Child and Youth Health (NI19-1065). The analyses, conclusions, opinions, and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred.

Conflicts of Interest
The authors have no competing interests to declare.

Abbreviations
NLP: Natural Language Processing
LIS: Laboratory Information System
OLIS: Ontario Laboratory Information System
Bi-LSTM: bidirectional Long Short-Term Memory
EHR: Electronic Health Record
RSV: Respiratory syncytial virus
HMV: Human Metapneumovirus
SME: subject matter experts
CI: Confidence Interval

ETHICS STATEMENT
Projects that use data collected by ICES under section 45 of Ontario’s Personal Health Information Protection Act (PHIPA), and use no other data, are exempt from research ethics board review. The use of the data in this project is authorized under section 45 and approved by ICES’ Privacy and Legal Office. ICES is a prescribed entity under PHIPA. Section 45 of PHIPA authorizes ICES to collect personal health information, without consent, for the purpose of analysis or compiling statistical information concerning the management of, evaluation or monitoring of, the allocation of resources to or planning for all or part of the health system.

DATA AVAILABILITY STATEMENT
The data underlying this work is held securely in the coded form at ICES and, therefore, cannot be shared publicly due to data privacy concerns and legal data sharing agreements between ICES and data providers (e.g., healthcare organizations and government).
However, data access might be granted to those who meet specified criteria for confidential access, available at www.ices.on.ca/DAS (email das@ices.on.ca).

References


