**Lightweight transformers for clinical natural language processing**

Omid Rohanian\(^1\), Mohammadmahdi Nouriborji\(^2\), Hannah Jauncey\(^4\), Samaneh Kouchaki\(^5\), Farhad Nooralahzadeh\(^6\), ISARIC Clinical Characterisation Group\(^8\), Lei Clifton\(^9\), Laura Merson\(^8\) and David A. Clifton\(^1\)

\(^1\)Department of Engineering Science, University of Oxford, Oxford, UK, \(^2\)NLPie Research, Oxford, UK, \(^3\)Sharif University of Technology, Tehran, Iran, \(^4\)Infectious Diseases Data Observatory (IDDO), University of Oxford, Oxford, UK, \(^5\)Department of Electrical and Electronic Engineering, University of Surrey, Guildford, UK, \(^6\)University of Zürich, Zürich, Switzerland, \(^7\)University Hospital of Zürich, Zürich, Switzerland, \(^8\)ISARIC, Pandemic Sciences Institute, University of Oxford, Oxford, UK, and \(^9\)Oxford-Suzhou Centre for Advanced Research, Suzhou, China

**Corresponding author:** Omid Rohanian; Email: omid.rohanian@eng.ox.ac.uk

(Received 18 June 2023; revised 23 November 2023; accepted 26 November 2023)

**Abstract**

Specialised pre-trained language models are becoming more frequent in Natural language Processing (NLP) since they can potentially outperform models trained on generic texts. BioBERT (Sanh et al., Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv: 1910.01108*, 2019) and BioClinicalBERT (Alsentzer et al., Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019) are two examples of such models that have shown promise in medical NLP tasks. Many of these models are overparametrised and resource-intensive, but thanks to techniques like knowledge distillation, it is possible to create smaller versions that perform almost as well as their larger counterparts. In this work, we specifically focus on development of compact language models for processing clinical texts (i.e. progress notes, discharge summaries, etc). We developed a number of efficient lightweight clinical transformers using knowledge distillation and continual learning, with the number of parameters ranging from 15 million to 65 million. These models performed comparably to larger models such as BioBERT and ClinicalBioBERT and significantly outperformed other compact models trained on general or biomedical data. Our extensive evaluation was done across several standard datasets and covered a wide range of clinical text-mining tasks, including natural language inference, relation extraction, named entity recognition and sequence classification. To our knowledge, this is the first comprehensive study specifically focused on creating efficient and compact transformers for clinical NLP tasks. The models and code used in this study can be found on our Huggingface profile at https://huggingface.co/nlpie and Github page at https://github.com/nlpie-research/Lightweight-Clinical-Transformers, respectively, promoting reproducibility of our results.

**Keywords:** Machine learning; natural language processing for biomedical texts

1. **Introduction**

Large language models pre-trained on generic texts serve as the foundation upon which most state-of-the-art Natural language Processing (NLP) models are built. There is ample evidence that, for certain domains and downstream tasks, models that are pre-trained on specialised data outperform baselines that have only relied on generic texts (Sanh *et al.* 2019; Alsentzer *et al.* 2019; Beltagy, Lo, and Cohan 2019; Nguyen, Vu, and Nguyen 2020; Chalkidis *et al.* 2020).

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.
These models, however, are heavy in size and number of parameters, making them unsuitable for devices with limited memory and processing capacity. Furthermore, the rate at which leading technology corporations build these progressively larger and more resource-intensive models is a subject of debate in the Artificial Intelligence (AI) community (Bender et al. 2021), and there is interest in developing methods that would make these tools more accessible by creating smaller and faster versions of them that would run reasonably well on smaller devices (Li et al. 2020; Schick and Schütze 2021). This would allow independent researchers, particularly those from low-income nations, to contribute to the advancement of AI.

From the point of view of energy consumption and environmental impact, developing smaller transformer-based language models can be thought of as a step towards green AI (Schwartz et al. 2020), an approach to developing AI that prioritises sustainable use of computational resources and development of models with minimal carbon footprint (Strubell, Ganesh, and McCallum 2019).

Numerous works exist in the NLP literature with the aim to develop fast, efficient and lightweight versions of larger transformer-based models (Sanh et al. 2019; Jiao et al. 2020; Sun et al. 2020; Dehghani et al. 2021). However, there are comparatively fewer compact models developed for special domains like law, biology and medicine (Ozyurt 2020; Bambroo and Awasthi 2021; Rohanian et al. 2022). The present work focuses on development of efficient lightweight language models specifically developed for clinical NLP tasks. These models can be used to process a range of different clinical texts including patient history, discharge summaries and progress notes. The contributions of this work are as follows:

- We pre-trained five different compact clinical models using either distillation or continual learning on the MIMIC-III notes dataset.
- We used three different distillation techniques for training models in varying sizes and architectures.
- We evaluated our models on named entity recognition (NER), relation extraction (RE) and sequence classification (CLS) on four widely used clinical datasets plus an internal cancer identification dataset.
- We are the first to focus exclusively on developing compact clinical language models, and we make all of our models publicly available on Huggingface.a

2. Clinical notes in electronic health records (EHR)

Clinical notes are written documents generated by medical practitioners in order to communicate information about a patient treated at a health facility (Rethans, Martin, and Metsemakers 1994). These documents are regarded as ‘unstructured data’. This means that, unlike tabular data that are categorised and quantifiable, clinical notes are irregular, disorganised and not coded using predefined terms that domain experts would all understand (Boulton and Hammersley 2006; Rosenbloom et al. 2011). Clinical notes contain a wide range of information about a patient, ranging from medical history and response to medication to discharge summaries and even billing. Fast Healthcare Interoperability Resources (FHIR)b identifies eight different types of clinical notes, including consultations notes, imaging narratives, laboratory reports and procedure notes (Bender and Sartipi 2013).

Unstructured data constitutes 80% of all electronic health records (EHR) data (Kong 2019; Mahbub et al. 2022) and can potentially contain information that is otherwise not present elsewhere in the patient’s EHR (Zhang et al. 2022). They can, therefore, be exploited by computational

---

ahttps://huggingface.co/nlpie

models to infer more information about a patient or develop predictive models for patient monitoring. There are numerous examples in the literature where clinical notes have been used, sometimes in conjunction with structured data, to develop diagnostic or predictive models. Some examples include adverse drug effects (Dandala, Joopudi, and Devarakonda 2019), (Mahendran and MclNnes 2021), self-harm and drug abuse prediction (Obeid et al. 2020; Ridgway et al. 2021), hospitalisation and readmission risk (Huang, Altosaar, and Ranganath 2019; Song et al. 2022), mortality prediction (Si and Roberts 2019; Ye et al. 2020) and automatic phenotype annotation (Zhang et al. 2022).

Unless preprocessed and redacted, clinical notes might contain private information about patients, and models trained on this data are known to be prone to adversarial attacks (Lehman et al. 2021). For this reason, they are yet to be widely used for research and different de-identification methods have been developed to automatically remove personal identifiable information from text documents in order for them to be securely shared with other researchers (Melamud and Shivade 2019; Hartman et al. 2020).

2.1. Biomedical vs clinical texts

There is a distinction in the literature between clinical and biomedical texts and they are understood to be different in terms of their linguistic qualities (Alsentzer et al. 2019). Clinical notes are collected by healthcare professionals when the patient is seen or being treated. They are free-text, without a fixed structure, can contain spelling errors, abbreviations, non-standard grammar, differences in personal style and words and phrases from different languages. These characteristics contribute to the fact that they are still underutilised as a resource (Sanyal, Rubin, and Banerjee 2022).

What is referred to as biomedical texts, on the other hand, are often compilations of scientific texts in the biomedical and life sciences from resources such as PubMed. They are written in a more polished standard style, and while they do overlap with clinical texts, they are larger in size and easier to process using standard NLP methods.

2.2. Language models for clinical texts

Due to the differences between biomedical and clinical texts (Section 2.1), transformer-based language models that have been only trained on generic and biomedical texts are not always sufficient to capture all the complexities of clinical notes. For this reason, it is common to use pre-trained models as a starting point and either use fine-tuning to adapt them to clinical notes (van Aken et al. 2021; Agnikula Kshatriya et al. 2021) or use continual learning and further pre-train a model like BERT or BioBERT (Lee et al. 2020a) on clinical texts (Si et al. 2019; Alsentzer et al. 2019; Qiu et al. 2020).

2.3. Adapting to new domains via continual learning

Continual learning is a powerful alternative to the standard transfer learning approach which involves pre-training and fine-tuning on a target task. In this paradigm, models can adapt to new domains during the pre-training stage. This linearity resembles biological learning and also alleviates the need for excessive model re-training (Mehta et al. 2021).

The idea here is to adapt the model to new streams of data while retaining knowledge of the previous domains (Parisi et al. 2019). Using this strategy, we may pre-train models that have previously been trained on biomedical texts and expose them to clinical notes and the vocabulary

---

It should be noted that some types of clinical notes resemble biomedical texts more than others. For instance, discharge summaries tend to be longer and more coherent than short and hastily written progress notes.
associated with them. While this method is not specifically related to model compression, we can develop lightweight clinical models by using already compressed biomedical models, such as BioMobileBERT, and through continual learning adapt them to the clinical domain. We explore this approach in this work in Section 3.2.

2.4. Language model compression

As discussed in Section 1, issues like overparameterisation, computational overhead and the negative environmental impact of large pre-trained language models have led researchers to develop strategies for compressing these models into smaller, faster, but almost equally performant versions (Sun et al. 2019).

Knowledge distillation (KD) (Hinton et al. 2015) is a well studied and powerful technique that is designed to create such models in a ‘teacher–student’ setup, where the smaller student model learns to mimic its teacher, either task-specifically by using the teacher’s outputs as soft labels or task-agnostically by looking at the outputs of a pre-training objective such as Masked Language Modelling (Devlin et al. 2019). The latter option allows for greater flexibility because the student may then independently be fine-tuned on the target task (Wang et al. 2020).

DistilBERT (Sanh et al. 2019) is a notable example of such an effort in NLP, inspiring a slew of alternative ‘distilled’ versions of commonly used models. DistilGPT2, DistilRoBERTa, DistilBART and DistilT5 are a few examples. More recently, other powerful approaches have also appeared in the literature, some of which will be covered in Section 3.

The efficacy of KD-based compression in specialised domains like biomedical and clinical texts is still understudied. Rohanian et al. (2022) is an example of a work that focuses on development of compact biomedical transformers. To the best of our knowledge, there is no work specifically targeting KD for models trained on clinical texts. As discussed in Section 2.1, these texts contain linguistic features and terminology that differ from generic biomedical texts, necessitating a separate treatment.

3. Methods

In this work, we utilise KD methods (Section 2.4) to train small-sized and efficient language models specialised for processing of clinical texts. First, we use KD approaches to directly extract compact models from the BioClinicalBERT model; second, we employ continual learning to pre-train existing compact biomedical models (e.g. BioDistilBERT and BioMobileBERT) using the MIMIC-III notes dataset (Johnson et al. 2016).

3.1. Compressing clinical models via distillation

In order to distil compact models from BioClinicalBERT, three different KD methods are explored in this work: DistilClinicalBERT, TinyClinicalBERT and ClinicalMiniALBERT (Fig. 3). These methods are described in detail below.

3.1.1. DistilClinicalBERT

This approach follows the distillation process outlined in DistilBERT Sanh et al. (2019) with the aim of aligning the output distributions of the student and teacher models based on the Masked Language Modelling (MLM) objective, as well as aligning their last hidden states. The loss used for this approach is defined as

\[
L(X, Y) = \lambda_1 L_{output}(f_s(X), f_t(X)) + \lambda_2 L_{align}(h_s(X), h_t(X)) + \lambda_3 L_{mlm}(f_s(X), Y)
\]

https://doi.org/10.1017/S1351324923000542 Published online by Cambridge University Press
where $X$ is the input to the model, $Y$ represents the MLM labels, $f_s(X)$ and $f_t(X)$ denote the outputs of the student and teacher models, respectively, $h_s(X)$ and $h_t(X)$ are the last hidden states of the student and teacher, $L_{output}$ is a KL-Divergence loss for aligning the output distributions of the student and teacher, $L_{align}$ is a cosine embedding loss for aligning the last hidden states of the student and teacher, $L_{mlm}$ represents the original MLM loss and $\lambda_1$ to $\lambda_3$ are hyperparameters controlling the weighting of each component in the loss function.

The student used in this approach uses six hidden layers, a hidden dimension size of 768 and an expansion rate of four for the MLP blocks, resulting in 65M parameters in total. For initialising the student model, we follow the method as described by Sanh et al. (2019) and Lee et al. (2020b). This involves using the same embedding size as the teacher and borrowing pre-trained weights from a subset of the teacher’s layers.

### 3.1.2. TinyClinicalBERT

This is a layer-to-layer distillation approach based on TinyBERT (Jiao et al. 2020), which is intended to align the hidden states and attention maps of each layer of the student with a specific layer of the teacher. Because the student network typically uses a smaller hidden dimension size compared to its teacher, an embedding alignment loss is also included. The combined loss in this approach is defined below:

\[
L(X) = \lambda_0 L_{embed}(e_s(X), e_t(X)) \\
+ \sum_{l=1}^{N} \lambda_l L_{att}(a^{l}_s(X), a^{l}_t(X)) \\
+ \sum_{l=1}^{N} \lambda_l L_{hid}(h^{l}_s(X), h^{l}_t(X)) \\
+ \lambda_{(N+1)} L_{out}(f_s(X), f_t(X))
\]

where $N$ represents the number of layers in the student model. The embedding vectors for the student and teacher models before passing to the transformer encoder are represented by $e_s(X)$ and $e_t(X)$, respectively. For the $i$th layer of both the student and teacher models, the attention maps and hidden states are represented by $a^{i}_s(X)$, $a^{i}_t(X)$, $h^{l}_s(X)$ and $h^{l}_t(X)$, respectively. The mapping function $g(.)$ is used to determine the corresponding teacher layer index for each student layer and is the same mapping function used in TinyBERT. The mean squared error (MSE) loss $L_{embed}$ is used to align the embeddings of the student and teacher models, while the MSE losses $L_{att}$ and $L_{hid}$ align their attention maps and hidden states, respectively. The cross-entropy loss $L_{out}$ aligns their output distributions. Finally, hyperparameters $\lambda_0$ to $\lambda_{(N+1)}$ control the significance of each loss component.

The student model in this approach has four hidden layers, with a hidden dimension of 312 and an MLP expansion rate of 4, totalling 15 million parameters. Due to the difference in hidden dimensions between the student and teacher models, the student model is initialised with random weights.

### 3.1.3. ClinicalMiniALBERT

This is another layer-to-layer distillation approach with the difference that the student is not a fully-parameterised transformer, but a recursive one (e.g. ALBERT (Lan et al. 2019)). We follow the same distillation procedure introduced in MiniALBERT (Nouriborji et al. 2022), which is similar to Equation (2). The recursive student model in this method uses cross-layer parameter sharing and embedding factorisation in order to reduce the number of parameters and employs bottleneck adapters for layer-wise adaptation. Its architecture features a hidden dimension of 768, an MLP expansion rate of 4 and an embedding size of 312, which add up to a total of 18
3.2. Adapting compressed models to clinical tasks via continual learning

We investigate an alternative method for compressing clinical models through continual learning. In this approach, already compressed biomedical models are further specialised by pre-training on a collection of clinical texts using the MLM objective. Two different models, namely, ClinicalDistilBERT and ClinicalMobileBERT, are developed in this fashion.

To obtain ClinicalDistilBERT, we use the pre-trained BioDistilBERT (Rohanian et al. 2022) model and train it further for three epochs on MIMIC-III. This model has the same architecture as DistilClinicalBERT (as described in Section 3.1.1). ClinicalMobileBERT, on the other hand, is based on the pre-trained BioMobileBERT (Rohanian et al. 2022) model and is also further trained on MIMIC-III for three epochs. This model has a unique architecture that allows it to have a depth of up to 24 hidden layers while having only 25 million parameters.

4. Datasets and tasks

This section discusses the NLP tasks and datasets used in this work. We briefly explain the goals and the nature of each task, followed by information on the datasets used to evaluate the proposed models.

4.1. Tasks

We explored four prominent tasks in clinical NLP: NER, Relation Extraction (RE), natural language inference (NLI) and classification (CLS). Below, we provide definitions and concrete examples from each task to illustrate their objectives and characteristics.

4.1.1. Named entity recognition (NER)

NER is the task of automatically processing a text and identifying named entities, such as persons, organisations, locations and medical terms. For example, in the sentence ‘The patient was diagnosed with heart disease by Dr. Johnson at JR Hospital’, NER would identify ‘patient’, ‘heart disease’, ‘Dr. Johnson’ and ‘JR Hospital’ as named entities and classify them as ‘Person’, ‘Disease’, ‘Person’ and ‘Hospital’, respectively.

4.1.2. Relation extraction (RE)

RE is the task of recognising and extracting links between entities such as genes, proteins, diseases, treatments, tests and medical conditions. For example, in the sentence ‘The EGFR gene has been associated with increased risk of lung cancer’, the RE system may recognise the association between the EGFR gene and lung cancer as ‘connected to increased risk’.

4.1.3. Natural language inference (NLI)

In natural language inference (NLI), the goal is to determine the connection between two texts, such as a premise and a hypothesis. The connection may be defined as entailment, contradiction or neutral. For example, if given the premise ‘The patient is diagnosed with influenza’ and the hypothesis ‘The patient is being treated for bacterial infection’, the NLI system would find that the connection is ‘contradiction’. This task helps improve understanding of the connections between biomedical concepts in language.
Table 1. Training samples from the MedNLI dataset

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>Sentence 2</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemi-sternotomy and lead extraction on [<strong>3090-11-5</strong>]</td>
<td>Patient has no cardiac history</td>
<td>Contradiction</td>
</tr>
<tr>
<td>He was sent for Stat Head CT, neurology was</td>
<td>The patient has had a stroke</td>
<td>Entailment</td>
</tr>
<tr>
<td>called</td>
<td></td>
<td></td>
</tr>
<tr>
<td>His O2Sat remains above 90% on room air but</td>
<td>Patient has OND</td>
<td>Neutral</td>
</tr>
<tr>
<td>appears to drop when patient falls asleep</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.1.4. Sequence classification (CLS)

CLS is the task of assigning class labels to word sequences in a biomedical text, such as sentences or paragraphs. The aim is to correctly predict the sequence’s class label based on the contextual information provided in the text. For example, given the text ‘Patient has high fever and severe headache’, the system may predict the class label ‘symptoms of an illness’.

When the class labels contain assertion modifiers (negation, uncertainty, hypothetical, conditional, etc.) and reflect degrees of certainty, the task is referred to as assertion detection (AD) (Chen 2019), which can be regarded as a subtask of CLS. For example, in the statement ‘Patient has heightened sensitivity to light and mild headache, which may indicate migraine’, the AD system would predict the class label ‘uncertain’ or ‘possible’ based on the context.

4.2. Datasets

We evaluate all of our models on four publicly available datasets, namely, MedNLI, i2b2-2010, i2b2-2012, i2b2-2014, and one internal dataset named ISARIC Clinical Notes (ICN).

4.2.1. MedNLI

MedNLI Romanov and Shivade (2018) is a NLI task designed for medical texts, in which two sentences are given to the model and the model should predict one of the entailments, contradiction or neutral labels as the relation of the two given sentences, as shown in Table 1.

4.2.2. i2b2 datasets

i2b2-2010 (Uzuner et al. 2011) is a medical relation extraction dataset, in which the model is required to output the relation between two concepts in a given text. The relations are between ‘medical problems and treatments’, ‘medical problems and tests’ and ‘medical problems and other medical problems’. In total, this dataset uses nine labels which are as follows:

1. Treatment improves medical problem (TrIP)
2. Treatment worsens medical problem (TrWP)
3. Treatment causes medical problem (TrCP)
4. Treatment is administered for medical problem (TrAP)
5. Treatment is not administered because of medical problem (TrNAP)
6. Test reveals medical problem (TeRP)
7. Test conducted to investigate medical problem (TeCP)
8. Medical problem indicates medical problem (PIP)
9. No relations

This dataset is also used for NER, with the same entity labels as i2b2-2012.
Table 2. i2b2-2010 samples taken from the dataset’s guideline Uzuner et al. (2011). The concept pairs for which a relationship should be predicted are displayed in boldface. Following the pre-processing used in the BLUE benchmark, the concepts are replaced with tags and then passed to the model as shown in the second column.

<table>
<thead>
<tr>
<th>Raw</th>
<th>Pre-Processed</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was discharged to home to be followed for her <strong>coronary artery disease</strong> following <strong>two failed bypass graft procedure</strong></td>
<td>Was discharged to home to be followed medically for <strong>@treatment$</strong> following <strong>@problem$</strong></td>
<td>TrWP</td>
</tr>
<tr>
<td>She has an <strong>elevated cholesterol</strong> controlled with <strong>Zocor</strong></td>
<td>She has an <strong>@problem$</strong> controlled with <strong>@treatment$</strong></td>
<td>TrIP</td>
</tr>
<tr>
<td><strong>Bactrim</strong> could be a cause of <strong>these abnormalities</strong></td>
<td><strong>@treatment$</strong> could be a cause of <strong>@problem$</strong></td>
<td>TrCP</td>
</tr>
<tr>
<td>A <strong>lung biopsy</strong> was performed, revealing <strong>chorio carcinoma</strong></td>
<td>A <strong>@test$</strong> was performed, revealing <strong>@problem$</strong></td>
<td>TeRP</td>
</tr>
</tbody>
</table>

Figure 1. Samples from the i2b2-2012 dataset.

For fine-tuning our models on this dataset, we follow the same pre-processing used in the BLUE benchmark, which models the relation extraction task as a sentence classification by replacing the concepts with certain tags, as shown in Table 2.

i2b2-2012 (Sun, Rumshisky, and Uzuner 2013) is a temporal relation extraction dataset that contains 310 discharge summaries from Partners Healthcare and the Beth Israel Deaconess Medical Center. It contains inline annotations for each discharge summary in four categories: clinical concepts, clinical departments, evidentials and occurrences. In our experiments, it is used as an NER dataset with the following entity labels:

1. Medical Problem (PR)
2. Medical Treatment (TR)
3. Medical Test (TE)
4. Clinical Department (CD)
5. Evidental (EV)
6. Occurrence (OC)
7. None (NO)

Some samples from the training dataset are provided in Figure 1.

i2b2-2014 (Stubbs, Kotfila, and Uzuner 2015) consists of two sub-tasks: De-identification and heart disease risk factors identification. In our experiments, we focus on the de-identification task in which the goal is to remove protected health information (PHI) from the clinical notes. The data in this task contain over 1300 patient records and has inline annotations for PHIs in each note. Similar to i2b2-2012, this task is also framed as NER with 22 labels. Figure 2 shows some examples taken from the training subset of the dataset.

https://doi.org/10.1017/S1351324923000542 Published online by Cambridge University Press
4.2.3. ISARIC clinical notes (ICN)

The ISARIC COVID-19 Clinical Database^e^ consists of data from patients hospitalised with COVID-19 who are enrolled in an ISARIC Partner clinical study (Garcia-Gallo et al. 2022). The data (which are standardised to CDISC STDM format^f^) include hospital admission and discharge records, signs and symptoms, comorbidities, vital signs, treatments and outcomes. Non-prespecified terms related to one of five categories; medical history, complications, signs and symptoms, new diagnosis at follow-up or category not recorded. The non-prespecified terms may consist of one or multiple diagnoses or clinical events within the same category.

In December 2021, when the initial stratified sample of non-prespecified (free text) medical terms from the ISARIC COVID-19 Clinical Database was extracted, the database comprised of data from over 708,231 participants. The sample was formed of 125,381 non-prespecified terms and all five of the aforementioned categories were represented. We have released a relevant subset of this dataset along with this work and a copy of ICN, with necessary modifications for patient confidentiality, is available at https://github.com/nlpie-research/Lightweight-Clinical-Transformers/tree/main/ICN.

For the experiments in this work, each instance of the data consists of non-prespecified (free text) terms describing clinical and adverse events along with a classification label from a number of possible choices. An annotator with clinical training used the following three labels to annotate the free-text notes:

1. Malignancy
2. No Malignancy
3. Possible Malignancy (Ambiguous)

This annotation scheme is associated with the task of AD van Aken et al. (2021) as explained in Section 4.1.4. We refer to this annotated portion as ISARIC Clinical Notes (ICN) cancer classification dataset. Table 3 contains a few examples from ICN.

5. Experimental setup

5.1. Pre-training details and hyperparameters

We pre-train all of our models on the MIMIC-III dataset for a total of three epochs using either the MLM objective or Knowledge Distillation. We follow the same pre-processing used in Alsentzer et al. (2019) for MIMIC and use the BERT tokeniser from the Huggingface with a max length of

---

^e^The ISARIC COVID-19 Data Platform is a global partnership of more than 1700 institutions across more than 60 countries (Group 2021). Accreditation of the individuals and funders that contributed to this effort can be found in the supplementary material. These partners have combined data and expertise to accelerate the pandemic response and improve patient outcomes. For more information on ISARIC, see https://isaric.org. A subset of data is available for access via application to the Data Access Committee at www.iddo.org/covid-19.

^f^https://www.cdisc.org/standards/foundational/sdtm

https://doi.org/10.1017/S1351324923000542 Published online by Cambridge University Press
Table 3. Some sample clinical notes along with their annotation from the ICN dataset

<table>
<thead>
<tr>
<th>Clinical Note</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung cancer and hypertension</td>
<td>Malignancy</td>
</tr>
<tr>
<td>DVT, now on heparin</td>
<td>No malignancy</td>
</tr>
<tr>
<td>Arthritis, cataracts, under investigation for colon cancer</td>
<td>Possible malignancy (ambiguous)</td>
</tr>
</tbody>
</table>

256 tokens. The details of the hyperparameters used for pre-training and fine-tuning our models are available in Tables 7 and 8.

5.2. Results
We evaluated the proposed models and the baselines on five datasets: MedNLI, i2b2-2010, i2b2-2012, i2b2-2014 and ICN. All reported results are median of three runs with different seeds. As shown in Table 4, compact clinical models significantly outperform their general and biomedical baselines and achieve competitive results against BioBERT-v1.1 and BioClinicalBERT. ClinicalDistilBERT and ClinicalMobileBERT, which are trained using continual learning, obtain the best average results among all compact models (Table 4). ClinicalMiniALBERT outperforms both DistilClinicalBERT and TinyClinicalBERT in terms of average results among our distilled models.

6. Discussion and analysis
6.1. Effect of different initialisations
Following the work of Alsentzer et al. (2019), we explore the effect of different initialisations for our continually learned models, as shown in Table 5. We find that initialising ClinicalDistilBERT with the trained weights from a pre-existing biomedical model significantly improved the model’s average performance, particularly on the MedNLI dataset. However, we discovered that initialising the ClinicalMobileBERT with a similar biomedical checkpoint did not result in a significant performance boost.

6.2. Performance evaluation through latency and GMACs
Latency and Giga Multiply-Accumulate Operations per second (GMACs) are two important metrics used to evaluate the efficiency of machine learning models. Latency measures the amount of time it takes for a model to process a single input and produce an output and is typically measured in milliseconds. GMACs, on the other hand, stands for Giga Multiply-Accumulate Operations and is a measure of how many computational operations a model needs to perform its tasks. GMACs are expressed in terms of billions of operations and provides a way of quantifying the computational resources required by a machine learning model.

The results of the latency and GMACs of the models, as well as the model sizes, are presented in Table 6. These measurements were conducted using a V100 32G GPU. The results show that there is a trade-off between accuracy and efficiency of the models. The ClinicalBioBERT model, which is listed only as representative of a class of larger-sized models (110m parameters, as listed in Table 4), has the best performance on the test sets but has the highest latency and GMACs, making it less suitable for deployment in resource-constrained environments. On the other hand,

---

This is in line with our expectation to achieve comparable results, rather than beat significantly larger models.
GMACs and smaller sizes compared to the ClinicalBioBERT model. The BioClinicalBERT and

models offer a good balance between performance and efficiency with relatively lower latency,
not be as good as that of ClinicalBioBERT.

................................. ............................... ................................ ... ................................ ............................... ...............................

................................. ............................... ................................ ... ................................ ............................... ...............................

ClinicalMobileBERT

ClinicalDistilBERT

80.66 93.39/84.92 81.28 96.91 95.80 88.90

TinyClinicalBERT 15M 71.30 91.55/76.91 77.46 92.15 89.71 83.17

ClinicalMiniALBERT 18M 78.90 93.43/82.93 82.34 95.11 93.73 87.74

Table 4. The results of the baselines (above the double line) and our pre-trained models on clinical down-
stream tasks. The metrics used for reporting scores are accuracy for the MedNLI, micro-averaged F1 for
i2b2-2010 (RE), macro-averaged F1 for ICN, and Exact F1 for the others. Bold numbers denote the best
performance and underlined numbers denote the second-best performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>#Params</th>
<th>MedNLI</th>
<th>i2b2-2010</th>
<th>i2b2-2012</th>
<th>i2b2-2014</th>
<th>ICN</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NLI</td>
<td>RE/NER</td>
<td>NER</td>
<td>NER</td>
<td>CLS</td>
<td></td>
</tr>
<tr>
<td>BERTbase</td>
<td>110M</td>
<td>78.27</td>
<td>92.75/77.18</td>
<td>80.19</td>
<td>96.77</td>
<td>90.88</td>
<td>87.55</td>
</tr>
<tr>
<td>BioBERT-v1.1</td>
<td>110M</td>
<td><strong>84.10</strong></td>
<td><strong>93.70/82.54</strong></td>
<td><strong>83.00</strong></td>
<td>96.69</td>
<td>93.06</td>
<td>88.84</td>
</tr>
<tr>
<td>BioClinicalBERT</td>
<td>110M</td>
<td>82.41</td>
<td>93.58/84.27</td>
<td>82.98</td>
<td><strong>96.72</strong></td>
<td><strong>93.30</strong></td>
<td><strong>88.87</strong></td>
</tr>
<tr>
<td>DistilBERT</td>
<td>65M</td>
<td>73.41</td>
<td>92.75/76.43</td>
<td><strong>79.15</strong></td>
<td>96.23</td>
<td><strong>93.75</strong></td>
<td><strong>85.28</strong></td>
</tr>
<tr>
<td>MobileBERT</td>
<td>25M</td>
<td><strong>76.16</strong></td>
<td><strong>93.16/77.65</strong></td>
<td>76.03</td>
<td><strong>96.87</strong></td>
<td>89.03</td>
<td><strong>84.81</strong></td>
</tr>
<tr>
<td>TinyBERT</td>
<td>15M</td>
<td>74.75</td>
<td>89.34/66.00</td>
<td>68.51</td>
<td>91.92</td>
<td>70.82</td>
<td>76.89</td>
</tr>
<tr>
<td>DistilBioBERT</td>
<td>65M</td>
<td>71.80</td>
<td>92.98/79.96</td>
<td>80.78</td>
<td>95.55</td>
<td><strong>93.44</strong></td>
<td>85.75</td>
</tr>
<tr>
<td>BioDistilBERT</td>
<td>65M</td>
<td>77.77</td>
<td>93.19/79.58</td>
<td><strong>81.11</strong></td>
<td>96.13</td>
<td>93.10</td>
<td><strong>86.81</strong></td>
</tr>
<tr>
<td>BioMobileBERT</td>
<td>25M</td>
<td><strong>78.97</strong></td>
<td><strong>93.56/81.21</strong></td>
<td>78.77</td>
<td><strong>96.90</strong></td>
<td>89.61</td>
<td><strong>86.50</strong></td>
</tr>
<tr>
<td>TinyBioBERT</td>
<td>15M</td>
<td>67.72</td>
<td>91.19/69.12</td>
<td>73.30</td>
<td>92.18</td>
<td>85.44</td>
<td>79.82</td>
</tr>
<tr>
<td>BioMiniALBERT</td>
<td>18M</td>
<td>75.52</td>
<td>92.89/78.32</td>
<td>80.04</td>
<td>95.28</td>
<td>92.31</td>
<td>85.72</td>
</tr>
<tr>
<td>DistilClinicalBERT</td>
<td>65M</td>
<td>78.05</td>
<td>93.38/83.12</td>
<td>82.06</td>
<td>95.45</td>
<td>90.70</td>
<td>87.12</td>
</tr>
<tr>
<td>ClinicalDistilBERT</td>
<td>65M</td>
<td><strong>81.29</strong></td>
<td>93.39/82.36</td>
<td>82.30</td>
<td>95.81</td>
<td><strong>97.75</strong></td>
<td><strong>88.81</strong></td>
</tr>
<tr>
<td>ClinicalMobileBERT</td>
<td>25M</td>
<td>80.66</td>
<td><strong>93.88/84.92</strong></td>
<td>81.28</td>
<td><strong>96.91</strong></td>
<td>95.80</td>
<td><strong>88.90</strong></td>
</tr>
<tr>
<td>TinyClinicalBERT</td>
<td>15M</td>
<td>71.30</td>
<td>91.55/76.91</td>
<td>77.46</td>
<td>92.15</td>
<td>89.71</td>
<td>83.17</td>
</tr>
<tr>
<td>ClinicalMiniALBERT</td>
<td>18M</td>
<td>78.90</td>
<td><strong>93.43/82.93</strong></td>
<td><strong>82.34</strong></td>
<td>95.11</td>
<td>93.73</td>
<td>87.74</td>
</tr>
</tbody>
</table>

Table 5. The effect of different initialisations on the continual learning of compact models.

<table>
<thead>
<tr>
<th>Model</th>
<th>isBio</th>
<th>MedNLI</th>
<th>i2b2-2010</th>
<th>i2b2-2012</th>
<th>i2b2-2014</th>
<th>ICN</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClinicalDistilBERT</td>
<td>✔</td>
<td>81.29</td>
<td>93.39/82.36</td>
<td>82.30</td>
<td>95.81</td>
<td><strong>97.75</strong></td>
<td><strong>88.81</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>76.72</td>
<td>93.32/80.30</td>
<td>81.37</td>
<td><strong>96.20</strong></td>
<td>96.13</td>
<td>87.33</td>
</tr>
<tr>
<td>ClinicalMobileBERT</td>
<td>✔</td>
<td><strong>80.66</strong></td>
<td><strong>93.88/84.92</strong></td>
<td><strong>81.28</strong></td>
<td>96.91</td>
<td>95.80</td>
<td><strong>88.90</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>80.23</td>
<td>93.71/84.55</td>
<td>80.39</td>
<td><strong>97.25</strong></td>
<td><strong>96.17</strong></td>
<td>88.71</td>
</tr>
</tbody>
</table>

the TinyClinicalBERT model has the lowest latency, GMACs and size, but its performance may not be as good as that of ClinicalBioBERT.

The DistilClinicalBERT, ClinicalDistilBERT, ClinicalMobileBERT and ClinicalMiniALBERT models offer a good balance between performance and efficiency with relatively lower latency, GMACs and smaller sizes compared to the ClinicalBioBERT model. The BioClinicalBERT and
Table 6. Comparing the efficiency of the proposed models with ClinicalBioBERT. ↓ denotes that less is better for that particular metric

<table>
<thead>
<tr>
<th>Model</th>
<th>Latency (ms)</th>
<th>GMACs</th>
<th>Size (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClinicalBioBERT</td>
<td>94.61±6.11</td>
<td>0.970</td>
<td>410</td>
</tr>
<tr>
<td>ClinicalDistilBERT</td>
<td>58.27±3.98</td>
<td>0.588</td>
<td>248</td>
</tr>
<tr>
<td>DistilClinicalBERT</td>
<td>58.27±3.98</td>
<td>0.588</td>
<td>248</td>
</tr>
<tr>
<td>TinyClinicalBERT</td>
<td>16.92±2.74</td>
<td>0.123</td>
<td>52</td>
</tr>
<tr>
<td>ClinicalMobileBERT</td>
<td>85.18±6.06</td>
<td>0.184</td>
<td>94</td>
</tr>
<tr>
<td>ClinicalMiniALBERT</td>
<td>59.01±2.99</td>
<td>0.598</td>
<td>68</td>
</tr>
</tbody>
</table>

BioBERT-v1.1 models, with 110 million parameters, offer the highest performance but are also the most computationally intensive.

In a real-world setting, the choice of the appropriate model should depend on the specific requirements of the application, such as the required accuracy, computational resources and memory constraints. Based on the results presented in Table 4, the DistilClinicalBERT, ClinicalDistilBERT, ClinicalMobileBERT and ClinicalMiniALBERT models with 65 million, 25 million and 18 million parameters, respectively, provide a good balance between performance, latency, GMACs and model size.

6.3. ICN error analysis

6.3.1. Preparation of the error analysis subset

To perform error analysis, we chose three of the best-performing models, namely, BioBERT, BioClinicalBERT and our proposed ClinicalDistilBERT. In order to evaluate the models on a truly unseen dataset for this analysis, we selected the internal ICN dataset (Section 4.2.3). It consists of approximately 125,000 clinical notes, with 6000 of them having been labelled by clinicians from the ISARIC consortium. We fine-tuned the pre-trained models on the labelled section of the ICN and then used the resulting models to predict labels for all 125,000 samples. The samples in which at least two of the models disagree on the predicted label were identified as corner cases, resulting in approximately 1500 clinical notes.

To perform the error analysis, these corner cases were annotated by a clinician and the outputs of the three fine-tuned models were analysed and compared with the expert human annotations. More information about the specifics of this adjudication is provided in the Appendix C. Figure 4 provides the confusion matrices for performance of the models both on the test set and on the corner cases. Based on the information from the confusion matrices, ClinicalDistilBERT performed better than the rest of the models on these cases. BioBERT, on the other hand, fared comparatively poorly, which can be attributed to its lack of clinical domain knowledge. In the following section, we present the analysis and observations of the human expert annotator about the outputs of each model on the corner cases and investigate if there are any common mistakes or recurring patterns that have caused confusion for any of the models.

6.3.2. Patterns observed in error analysis

**Interpretation of ‘ca’ Abbreviation:** In the portion of ICN on which ClinicalDistilBERT was trained, the abbreviation ‘ca’ often refers to cancer. However, these two letters could also be used in other contexts; one such example was a free text term containing ‘low ca+’. It was assumed this free text was referring to low calcium levels, therefore a ‘No Malignancy’ label was
assigned by the human annotator. This was in direct contrast to the ‘malignant neoplasm’ label output of ClinicalDistilBERT. On this occasion, description of the level (i.e. low) preceding ‘ca’ and the addition sign succeeding this abbreviation indicated to human annotator that this term refers to a calcium measurement rather than cancer. This example shows how ClinicalDistilBERT has the potential to generate more reliable results if it is further developed to more accurately process contextual information in cases where free texts comprise abbreviations with multiple interpretations.

**Mislabelling of Benign Terms (Adenoma, Hemangioma, Angioma):** A frequently occurring inconsistency between the labels assigned by a human annotator and ClinicalDistilBERT related to certain free text terms containing the words ‘adenoma’, ‘hemangioma’ or ‘angioma’. Some examples of false positives in Clinical DistilBERT related to the term ‘adenoma’ are as follows: ‘bilateral adenomas’, ‘benign adrenal adenoma’, ‘benign ovarian cystadenoma’ and ‘fibroadenoma of right breast’. In all of these cases, the human annotator decided the note was ‘No Malignancy’ but the model labelled it as ‘Malignancy’. It is particularly interesting that ClinicalDistilBERT incorrectly labelled these, given the fact the some samples start with the word ‘benign’. The model may have been misled due to the similarities between the terms ‘adenocarcinoma’ (a type of malignancy) and ‘adenoma’ (a benign neoplasm).

We discovered a pattern of discrepancy between ClinicalDistilBERT labels and human annotator labels in samples containing words such as ‘hemangioma’ or ‘angioma’. For example, the model labelled ‘liver hemangioma’ and ‘right frontal cavernous angioma’ as ‘Possible Malignancy (Ambiguous)’ but the human annotator deemed these terms as ‘No Malignancy’. This may have occurred as the structure of these words is broadly similar to a number of conditions, such as ‘astrocytoma’ and ‘meningioma’, for which it is difficult to decipher the benign or malignant nature of the free text term; these terms were therefore previously seen by the model in its training samples associated with possible malignancy. More specifically, the aforementioned diagnoses with human and model annotation disparity end in ‘-oma’; free text terms ending in ‘-oma’ may often refer to a malignancy, but this is not universally correct. The model may have identified these samples incorrectly due to spelling similarities to malignancy terms encountered during training, but it has not yet developed the ability to differentiate between them consistently.

**Acronym Misinterpretation (CLL, ALL, NHL):** An analysis of label discrepancies among the three models revealed a pattern of incorrect labelling of certain acronyms by BioBERT and BioClinicalBERT. The models frequently mislabelled three-letter acronyms such as ‘CLL’, ‘ALL’ and ‘NHL’ as ‘No Malignancy’, while both ClinicalDistilBERT and human annotation identified them as malignant conditions. These acronyms can commonly be understood to refer to chronic lymphocytic leukaemia, acute lymphoblastic leukaemia and non-Hodgkin lymphoma, respectively. Despite the fact that these acronyms were included in the ‘Malignancy’ training samples, BioBERT and BioClinicalBERT often labelled free text terms containing these acronyms as ‘No Malignancy’. On the other hand, these models sometimes labelled certain acronyms such ‘gbs’ as ‘Malignancy’, while human annotation and ClinicalDistilBERT did not identify them as malignant.

**Impact of Text Length on Diagnosis (Term ‘Calculi’):** The word ‘calculi’ frequently led to inaccuracies in the diagnosis label of all three models. Out of the 1544 corner cases studied during the error analysis, 33 contained the term ‘calculi’ and were all annotated as ‘No Malignancy’ during human annotation. ClinicalDistilBERT, however, labelled 17 of these as malignant. These samples usually contained short free text terms, with 10 having 3 or fewer words and none having more than 5 words. Examples include ‘bladder calculi’, ‘left renal calculi’ and ‘staghorn calculi’. The 16 samples which were correctly labelled as ‘No Malignancy’ by ClinicalDistilBERT were frequently of a greater length, with 8 samples containing 4–20 words. Examples of these samples are:

\[\text{We assume this term refers to Guillain-Barré Syndrome.}\]
‘psoriasis/ depression/ renal calculi’, ‘gall bladder + common bile duct (cbd) calculi’ and ‘calculi in gallbladder and common bile duct’. BioBERT and BioClinicalBERT only correctly labelled one free text term containing ‘calculi’ as ‘No Malignancy’.

7. Conclusion and future works

In this work, we have presented five compact clinical transformers that have achieved significant improvement over their baselines on a variety of clinical text-mining tasks. To train our models, we utilised both continual learning and knowledge distillation techniques. For continual learning, we employed the BioDistilBERT and BioMobileBERT models, which are compact models derived using standard techniques and trained on large biomedical data. Our experiments showed that the average performance of these models can increase by up to 2% after training on the MIMIC-III clinical notes dataset, and they even outperformed larger baselines on the i2b2-2010 and ICN datasets (Table 4). In order to determine the best approach for knowledge distillation in clinical NLP, we explored a range of different methods including standard (Sanh et al. 2019), layer-to-layer (Jiao et al. 2020) and recursive distillation (Nouriborji et al. 2022).

Moreover, to confirm the efficacy of our methods on an unseen private dataset, we evaluated the performance of the top models on the ICN dataset by looking at the corner cases in the test set where at least two of the models disagreed and then asked an expert annotator with clinical training to adjudicate the corner cases. In this way, we managed to further assess the models on more complicated samples and provided a more in-depth analysis of where the models tend to fail and what recurring themes exist in the more challenging cases.

We also evaluated the models in terms of efficiency criteria such as latency and GMACs and compared the proposed models with BioClinicalBERT. We subsequently provided guidance on selecting the optimal model based on performance and efficiency trade-offs. We hope that by making our lightweight models public, we will make clinical text-mining methods more accessible to hospitals and academics who may not have access to GPU clusters or specialised hardware, particularly those in developing countries. The top-performing models produced in this study will be integrated into the data curation pipeline under development by ISARIC and Global.health, thereby facilitating the rapid aggregation and analysis of global data for outbreak response.

The work presented here has some limitations, however. Currently, our experiments are limited to datasets in English and it remains to be seen how the models would perform on datasets in other languages. We have also not tested our models on the task of connecting named entities to medical knowledge-bases such as SNOMED CT or ICD-10, which is an important task in clinical NLP. In future work, we aim to extend our research to more tasks and languages in order to address these limitations.

Another potential avenue for future work is the integration of information from other modalities such as images or electronic health records containing tabular data, including clinical laboratory test results and radiology images. This would allow us to train and evaluate multi-modal architectures, such as CLIP (Radford et al. 2021), and explore their utility for clinical NLP tasks.

Ethics statement. Ethics Committee approval for the collection and analysis of ISARIC Clinical Notes was given by the Health Organisation Ethics Review Committee (RPC571 and RPC572 on 25 April 2013). National and/or institutional ethics committee approval was additionally obtained by participating sites according to local requirements.

This work is a part of a global effort to accelerate and improve the collection and analysis of data in the context of infectious disease outbreaks. Rapid characterisation of novel infections is critical to an effective public health response. The model developed will be implemented in data aggregation and curation platforms for outbreak response - supporting the understanding of the variety of data collected by frontline responders. The challenges of implementing robust data collection

https://doi.org/10.1017/S1351324923000542 Published online by Cambridge University Press
efforts in a health emergency often result in non-standard data using a wide range of terms. This is especially the case in lower-resourced settings where data infrastructure is lacking. This work aims to improve data processing and will especially contribute to lower-resource settings to improve health equity.

**Competing interests.** The authors declare that they have no competing interests.

**Financial support.** This work was made possible with the support of UK Foreign, Commonwealth and Development Office and Wellcome [225288/Z/22/Z], Collection of data for the ISARIC Clinical Notes was made possible with the support of UK Foreign, Commonwealth and Development Office and Wellcome [215091/Z/18/Z, 222410/Z/21/Z, 225288/Z/22/Z, 220757/Z/20/Z and 222048/Z/20/Z] and the Bill & Melinda Gates Foundation [OPP1209135]; GIIHR Coronavirus Rapid Research Funding Opportunity OVI2170359 and was coordinated out of Sunnybrook Research Institute; was supported by endorsement of the Irish Critical Care- Clinical Trials Group, co-ordinated in Ireland by the Irish Critical Care- Clinical Trials Network at University College Dublin and funded by the Health Research Board of Ireland [CTN-2014-12]; grants from Rapid European COVID-19 Emergency Response research (RECOVER) [H2020 project 101003589] and European Clinical Research Alliance on Infectious Diseases (ECRAID) [965313]; Cambridge NIHR Biomedical Research Centre; Wellcome Trust fellowship [205228/Z/16/Z] and the National Institute for Health Research Health Protection Research Unit (HPRU) in Emerging and Zoonotic Infections (NIHR200907) at the University of Liverpool in partnership with Public Health England (PHE), in collaboration with Liverpool School of Tropical Medicine and the University of Oxford; the dedication and hard work of the Norwegian SARS-CoV-2 study team.

Research Council of Norway grant no 312780 and a philanthropic donation from Vivaldi Invest A/S owned by Jon Stephenson von Tetzchner; PJMO was supported by the UK’s National Institute for Health Research (NIHR) via Imperial’s Biomedical Research Centre (NIHR Imperial BRC), Imperial’s Health Protection Research Unit in Respiratory Infections (NIHR HPRU RI), the Comprehensive Local Research Networks (CLRNs) and is an NIHR Senior Investigator (NIHR201385); Innovative Medicines Initiative Joint Undertaking under Grant Agreement No. 115523 COMBACTE, resources of which are composed of financial contribution from the European Union’s Seventh Framework Programme (FP7/2007-2013) and EFPIA companies, in-kind contribution; Stiftungsfonds zur Förderung der Bekämpfung der Tuberkulose und anderer Lungenkrankheiten of the City of Vienna; Project Number: APCOV22BGM; Australian Department of Health grant (3273191); Gender Equity Strategic Fund at University of Queensland, Artificial Intelligence for Pandemics (A14PAN) at University of Queensland, The Australian Research Council Centre of Excellence for Engineered Quantum Systems (EQUIS, CE170100009), The Prince Charles Hospital Foundation, Australia; grants from Instituto de Salud Carlos III, Ministerio de Ciencia, Spain; Brazil, National Council for Scientific and Technological Development Scholarship number 303953/2018-7; the Firland Foundation, Shoreline, Washington, USA; The French COVID cohort (NCT04262921) is sponsored by INSERM and is funding by the REACTing (REsearch & ACtion emergING infectious diseases) consortium and by a grant of the French Ministry of Health (PHRC n°20-0424); the South Eastern Norway Health Authority and the Research Council of Norway; and a grant from the Oxford University COVID-19 Research Response fund (grant 0009109); Institute for Clinical Research (ICR), National Institutes of Health (NIH) supported by the Ministry of Health Malaysia; a grant from foundation Bevordering Onderzoek Franciscus.

The investigators acknowledge the philanthropic support of the donors to the University of Oxford’s COVID-19 Research Response Fund; COVID clinical management team, AIIMS, Rishikesh, India; COVID-19 Clinical Management team, Manipal Hospital Whitefield, Bengaluru, India; Italian Ministry of Health “Fondi Ricerca corrente-L1P6” to IRCCS Ospedale Sacro Cuore-Don Calabria and Preparedness work conducted by the Short Period Incidence Study of Severe Acute Respiratory Infection; The dedication and hard work of the Groote Schuur Hospital Covid ICU Team, supported by the Groote Schuur nursing and University of Cape Town registrar bodies coordinated by the Division of Critical Care at the University of Cape Town.

This work uses data provided by patients and collected by the NHS as part of their care and support #DataSavesLives. The data used for this research were obtained from ISARIC4C. We are extremely grateful to the 2648 frontline NHS clinical and research staff and volunteer medical students who collected these data in challenging circumstances; and the generosity of the patients and their families for their individual contributions in these difficult times. The COVID-19 Clinical Information Network (CO-CIN) data were collated by ISARIC4C Investigators. Data and Material provision were supported by grants from: the National Institute for Health Research (NIHR; award CO-CIN-01), the Medical Research Council (MRC; grant MC_PC_19059) and by the NIHR Health Protection Research Unit (HPRU) in Emerging and Zoonotic Infections at University of Liverpool in partnership with Public Health England (PHE), (award 200907), NIHR HPRU in Respiratory Infections at Imperial College London with PHE (award 200927), Liverpool Experimental Cancer Medicine Centre (grant C18616/A25153), NIHR Biomedical Research Centre at Imperial College London (award ISBRC-1215-20013) and NIHR Clinical Research Network providing infrastructure support. We also acknowledge the support of Jeremy J Farrar and Nahoko Shindo.

This work was supported in part by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and in part by an InnoHK Project at the Hong Kong Centre for Cerebro-cardiovascular Health Engineering hardships.


Appendix A: I2B2-2014 labels

The following entity labels were used for framing the de-identification task as named entity recognition:

<table>
<thead>
<tr>
<th>City</th>
<th>Date</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>Fax</td>
<td>Medicalrecord</td>
</tr>
<tr>
<td>Phone</td>
<td>Bioid</td>
<td>Profession</td>
</tr>
<tr>
<td>Idnum</td>
<td>Zip</td>
<td>Healthplan</td>
</tr>
<tr>
<td>Device</td>
<td>Street</td>
<td>Hospital</td>
</tr>
<tr>
<td>Location</td>
<td>Patient</td>
<td>Organisation</td>
</tr>
<tr>
<td>Url</td>
<td>Country</td>
<td>Doctor</td>
</tr>
<tr>
<td>Age</td>
<td>Username</td>
<td></td>
</tr>
</tbody>
</table>

Appendix B: Details of the hyperparameters used in the experiment

In this section, we provide the details of the hyperparameters used in our experiments. Table 7 lists the hyperparameters used for pre-training models on the MIMIC-III dataset, and Table 8 shows the hyperparameters chosen for the purpose of fine-tuning on downstream tasks.
Table 7. Hyperparameters used for pre-training models on MIMIC-III

<table>
<thead>
<tr>
<th>Param</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate</td>
<td>{5e-4, 5e-5}</td>
</tr>
<tr>
<td>Scheduler</td>
<td>Linear</td>
</tr>
<tr>
<td>Optimiser</td>
<td>AdamW</td>
</tr>
<tr>
<td>Weight decay</td>
<td>1e−4</td>
</tr>
<tr>
<td>Total batch size</td>
<td>192</td>
</tr>
<tr>
<td>Warmup steps</td>
<td>5000</td>
</tr>
<tr>
<td>Epochs</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 8. Hyperparameters used for fine-tuning on downstream tasks

<table>
<thead>
<tr>
<th>Param</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate</td>
<td>{5e-5, 2e-5, 1e-5}</td>
</tr>
<tr>
<td>Scheduler</td>
<td>Linear</td>
</tr>
<tr>
<td>Optimiser</td>
<td>AdamW</td>
</tr>
<tr>
<td>Weight decay</td>
<td>0.01</td>
</tr>
<tr>
<td>Batch size</td>
<td>{16, 32}</td>
</tr>
<tr>
<td>Epochs</td>
<td>3</td>
</tr>
</tbody>
</table>

Appendix C: Adjudication of classification outputs

The clinician who has assessed and adjudicated the classification differences, has opted to make the following decisions which are mentioned below for reference:

- If the terms used in the text were vague or unclear (e.g. cayn and loa), the note is judged as ‘no malignancy’.
- Indeterminate lymph nodes are assumed to be possible malignancy (ambiguous).
- Those notes that were written in a foreign language were excluded from the adjudication.
- ‘Glioma’ is typically indicative of possible malignancy (ambiguous), as it might be benign or malignant. However, if it has been described with a qualifier that signals benign (e.g. low-grade glioma or benign glioma), it has been deemed as ‘no malignancy’.
- ‘Angiorhyolipoma’, assumed to be angiomyolipoma was considered ‘no malignancy’.
- Text that refers to a possible sign/symptom of cancer, such as elevated ca125 or a mass causing bowel obstruction, has been labelled as ‘no malignancy’ if no indication of cancer or suspicion/investigation has been mentioned.
- As chemotherapy and radiotherapy are frequently used for cancer treatment, if either of these appear in free text without another cancer term listed, they are labelled ‘possible malignancy (ambiguous)’. 
Treatments that could be used for cancer but also have other applications, such as bone marrow, BMT, stem cell transplant, nephrectomy, prostatectomy and mastectomy, are labelled ‘no malignancy’ if no further reference to cancer is made in the text.

ADT (used for prostate cancer) is considered to represent possible malignancy (ambiguous).

‘Lump’ or ‘mass’ with no indication of cancer (e.g. breast lump) are labelled ‘no malignancy’.

Where resected year is redacted and ‘no recurrence’ is mentioned, the text is deemed ‘no malignancy’.

Grade one meningioma is ‘no malignancy’. Benign menigioma would also be ‘no malignancy’, but if only ‘meningioma’ is mentioned, this would be ‘possible malignancy (ambiguous)’.

Lynch syndrome is ‘no malignancy’. It increases the risk of malignancy but is not a cancer in itself.

‘Anti-cancer therapy’ is when mentioned without further context could refer to treatment or prophylactic so ‘possible malignancy (ambiguous)’ was used.

Kappa lambda light chain was considered as ‘no malignancy’.

Tumour lysis syndrome was classified as a ‘malignancy’ as the condition refers to release of the contents of cancer cells.

Figure 3. (A) Training loss of the ClinicalDistilBERT and the ClinicalMobileBERT when optimised for the MLM objective. (B) Training loss of the DistilClinicalBERT on the distillation objective as described in Section 3.1.1. (C) Training loss of TinyClinicalBERT on the distillation objective, as explained in Section 3.1.2. (D) Training loss of ClinicalMiniALBERT on the distillation objective, as introduced in Nouriborji et al. (2022).
Figure 4. (A) and (E) represent the confusion matrices for BioBERT on the test set and the corner cases (Section 6.3.1), respectively. (B) and (F) refer to the confusion matrices for ClinicalBioBERT on the test set and corner cases. (C) and (G) denote the confusion matrices for ClinicalDistilBERT on the test set and corner cases. ‘No M’ indicates ‘No Malignancy’, ‘P M’ represents ‘Possible Malignancy’ and ‘M’ signifies ‘Malignancy’.

Appendix D: ISARIC Clinical Characterisation Group


https://doi.org/10.1017/S1351324923000542 Published online by Cambridge University Press

https://doi.org/10.1017/S1351324923000542 Published online by Cambridge University Press

Cite this article: Rohanian O, Nouriborji M, Jauncey H, Kouchaki S, Nooralahzadeh F, ISARIC Clinical Characterisation Group, Clifton L, Merson L and Clifton D A. Lightweight transformers for clinical natural language processing. Natural Language Engineering https://doi.org/10.1017/S1351324923000542