Cross-Lingual Knowledge Transfer for Clinical Phenotyping

Jens-Michalis Papaioannou[*], **Paul Grundmann**[*], **Betty van Aken**[*], **Athanasios Samaras**[†]
Ilias Kyparissidis[‡], **George Giannakoulas**[†], **Felix Gers**[*], **Alexander Löser**[*]
[*]DATEXIS, Berliner Hochschule für Technik (BHT), Germany,
[†]First Department of Cardiology, AHEPA University Hospital, Aristotle University of Thessaloniki, Greece
[‡]Laboratory of Medical Physics and Digital Innovation, Aristotle University of Thessaloniki, Greece,
{michalis.papaioannou, paul.grundmann, bvanaken, gers, aloeser}@bht-berlin.de
{asamaraa, ggiannakoulas}@auth.gr
{iliaskypkok}@for.auth.gr

Abstract

Clinical phenotyping enables the automatic extraction of clinical conditions from patient records, which can be beneficial to doctors and clinics worldwide. However, current state-of-the-art models are mostly applicable to clinical notes written in English. We therefore investigate cross-lingual knowledge transfer strategies to execute this task for clinics that do not use the English language and have a small amount of in-domain data available. We evaluate these strategies for a Greek and a Spanish clinic leveraging clinical notes from different clinical domains such as cardiology, oncology and the ICU. Our results reveal two strategies that outperform the state-of-the-art: Translation-based methods in combination with domain-specific encoders and cross-lingual encoders plus adapters. We find that these strategies perform especially well for classifying rare phenotypes and we advise on which method to prefer in which situation. Our results show that using multilingual data overall improves clinical phenotyping models and can compensate for data sparseness.

Keywords: Document Classification, Text categorisation, Multilinguality, Neural language representation models

1. Introduction

Clinical phenotyping from text. Clinical information extraction has the potential to support clinicians in their daily work. Phenotyping—the extraction of patient conditions from text—in particular, can help clinicians to summarize patient states and to find similar patients more easily. Recent years have seen many advancements in automatic phenotyping from clinical text (Mulyar et al., 2019; Zhang et al., 2021), most of them based on neural networks. Especially domain-specific pre-training of such networks has shown to be of great benefit in clinical information extraction (Li et al., 2020; Gu et al., 2020). However, current state-of-the-art models are mostly exclusively applicable to English text, due to the large amount of both labeled and unlabeled clinical text resources in English.

The need for cross-lingual transfer. We want to investigate how to leverage the clinical knowledge encoded in English and other languages to the benefit of (low-resource) target languages. This scenario is commonly described as cross-lingual knowledge transfer (Ponti et al., 2020). Medicine is increasingly transformed by globalisation (Labonté et al., 2011) and knowledge collected in different languages has significant impact on local decisions (as seen during the Covid-19 pandemic). However, cross-lingual knowledge transfer in the clinical domain remains largely unexplored. Our work takes the perspective of clinicians working with non-English clinical notes. In particular we evaluate on medical texts based on Spanish and Greek in which there are far less such resources openly available.

Which cross-lingual approach works best in the clinical domain? Differences between languages (and clinics) regarding patient distributions, documentation styles or typologies make clinical knowledge transfer challenging. We evaluate three current approaches for cross-lingual transfer to find out which method handles these challenges best: 1) Translating Greek and Spanish notes into English before using a medical-specific English encoder, comparable to work by Isbister et al. (2021). 2) Using multi-lingual encoders (Conneau et al., 2020), which allow knowledge transfer between languages, but are not pre-trained on medical data. 3) Expanding multi-lingual encoders with adapters (Pfeiffer et al., 2020b), which have shown promising results on other cross-lingual tasks (Le et al., 2021) while being more resource efficient.

Cross-lingual transfer for rare phenotypes. Some phenotypes are more common in some parts of the world than in others. Distributing such knowledge to other regions is an important goal of medical knowledge transfer. Cross-lingual transfer is especially promising for these scenarios in which few examples exist in the target language. Therefore, we analyse in particular how the long-tail of phenotypes benefits from all evaluated cross-lingual approaches.

Contributions.

1. We compare state-of-the-art methods for cross-lingual knowledge transfer regarding their performance in the clinical domain. To this end, we analyse whether the incorporation of clinical notes written in different languages improves the results of clinical phenotyping.

2. We evaluate methods on three datasets in Spanish, Greek and English (for transfer only) and infer which methods are best suited for different scenarios when working with non-English clinical notes. Our analysis further reveals which combinations of languages are most beneficial.

3. We show in particular that the performance on the long tail of phenotypes benefits from the usage of additional data in different languages.

We publish the code for all experiments to ensure reproducibility[1].

2. Related Work

Phenotyping from clinical notes. Clinical phenotyping is the task of extracting patient conditions from Electronic Health Record (EHR) data. In this work, we focus on clinical notes–discharge summaries in particular–as information source, since they summarize the current state of a patient. Solt et al. (2009) proposed rule-based algorithms to solve the task, while Yao et al. (2019; Wang et al. (2019) applied Convolutional Neural Networks. Recently, large pre-trained language models have shown to outperform these earlier approaches (Mulyar et al., 2019). However, such language models are mostly used to process English clinical text data, for which large corpora such as PubMed Central exist. In contrast, we examine ways of solving the task in languages without a variety of openly available medical text data and evaluate how to transfer (clinical) knowledge from one language to another.

Models with cross-lingual abilities. In recent years, a variety of multi- and cross-lingual language models have been introduced (Hu et al., 2020). For this work, cross-lingual models are of special interest due to their inherent ability to transfer knowledge between languages (Mann and Yarowsky, 2001). To this end, Chung et al. (2020) propose to cluster and merge the vocabularies of similar languages for joint learning. The LASER model (Artetxe and Schwenk, 2019) is trained on parallel data of 93 languages with a shared BPE vocabulary. XLM (Conneau and Lample, 2019) additionally pre-trains BERT with parallel data. Building on these approaches, the XLM-R model (Conneau et al., 2020), pre-trained on 100 languages, performs especially well on low-resource languages, which leads us to examine whether it is also of use in the clinical domain. Pfeiffer et al. (2020b) introduced language adapters to cross-lingual models. Using adapters has shown to improve performance on multiple tasks while being more parameter-efficient. Thus, we investigate whether they are also beneficial to cross-lingual clinical phenotyping.

Language transfer via translations. In contrast to model-based approaches, Isbister et al. (2021) propose the use of translations for knowledge transfer between languages. While Artetxe et al. (2020) highlight that translations can have a regulating effect, they also show how translation errors can accumulate throughout the pipeline. These contrasts lead us to evaluate both the use of cross-lingual and translation based approaches to improve clinical phenotyping of low-ressource datasets.

Cross-lingual knowledge transfer in the medical domain. While there is a number of monolingual pre-trained language models for the medical domain in languages other than English (Carrino et al., 2021), the effect of cross-linguality on medical text has been rarely studied. Schäfer and Friedrich (2020) enrich a translated version of the Spanish CodiEsp dataset with English data to improve the performance in ICD code prediction. Schneider et al. (2020) use a pre-trained multi-lingual BERT model to further train on Brasilian clinical notes. We are the first to conduct an extensive analysis over multiple cross-lingual approaches regarding their usefulness for the clinical domain.

3. Task and Datasets

This section describes the clinical phenotyping task and introduces the three datasets used for evaluating cross-lingual transfer methods.

3.1. Phenotyping Task

The goal of clinical phenotype prediction is to support medical doctors to automatically categorise patients by the clinical conditions reported in a clinical note. This task is crucial e.g., for understanding the full picture of occurring diseases at a clinic, enabling further studies using specific patients, and predicting disease development and outcome. Here, we use the pre-defined CCSR[2] categorization where each category represents a set of ICD codes. Each category may represent a disease or a set of e.g., different arrhythmias or ill defined diseases.

3.2. Datasets

We use three datasets namely MIMIC-III (Johnson et al., 2016), CodiEsp (Miranda-Escalada et al., 2020) and AHEPA-cardio. These datasets consist of clinical notes that originate from different clinics in different countries i.e. USA, Spain and Greece. The languages in which the notes are written belong to the same language family (indo-european) but are classified into different language branches (Germanic, Italic and Greek). Greek is also typologically different from the other two. Furthermore, we choose medical data from diverse medical contexts. MIMIC-III originates from the ICU, CodiEsp from different clinics and AHEPAcardio from the cardiology department. The narrative and style of

---

[1] https://github.com/neuron1682/cross-lingual-phenotype-prediction

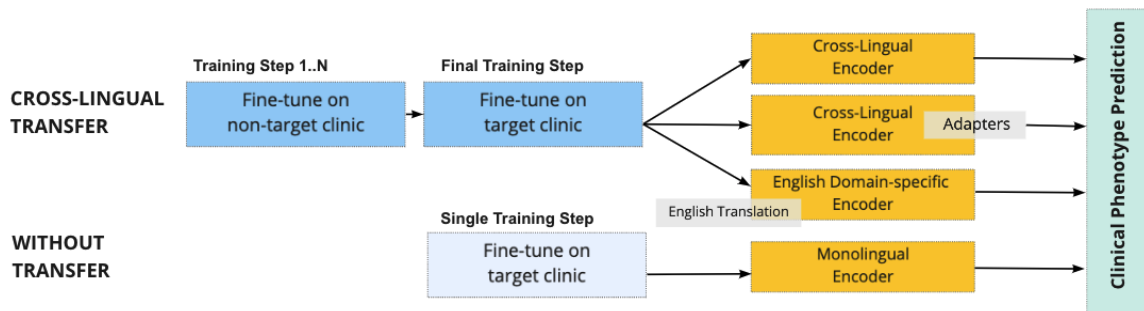[2] https://hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp

Figure 1: Schematic demonstration of our approaches. We compare cross-lingual with monolingual approaches. For the knowledge transfer we use sequential transfer learning starting from Mimic (high-resource dataset). We distinguish between cross-lingual encoders, cross-lingual encoders plus adapters and English domain-specific encoders with prior translation.

the texts is therefore different. This is also reflected in the length of the texts (see Table 1). Our rationale is that the medical knowledge in these datasets is diverse and complementary.

Mimic III - English Language. We use the freely-available MIMIC-III v1.4 database (referred to as *Mimic* in the following). It contains de-identified Electronic Health Records (EHR) data including clinical notes in English from the Intensive Care Unit (ICU) of Beth Israel Deaconess Medical Center in Massachusetts between 2001 and 2012. We focus our work on discharge summaries in particular and the diagnosis information associated with an admission. Similar to previous work (van Aken et al., 2021) we use discharge summaries and only keep the most informative sections regarding the phenotyping task including *Chief Complaint*, *History of Present Illness*, *Physical Exam*, *Social/Family History* and *Brief Hospital Course*. We further filter out re-admissions, duplicates and notes about newborns.

CodiEsp - Spanish Language. The CodiEsp dataset consists of 1,000 clinical case studies manually selected by doctors and cover a diverse set of medical specialties, including oncology, urology, cardiology, pneumonology and infectious diseases. The case studies are annotated regarding the diagnoses mentioned in the text. The notes are provided in both the original Spanish language and an English translation. For the translation of the notes the authors used a machine translation system which was fine-tuned on biomedical data (Soares and Krallinger, 2019).

AHEPAcardio - Greek Language. The AHEPAcardio dataset (referred to as *Ahepa* in the following) is a collection of around 2,400 discharge summaries and originates from the cardiology clinic of the AHEPA University Hospital in Greece.[3] The discharge summaries date from 2013 to 2020 and are annotated regarding clinical diagnoses found in the text by a team of medical doctors.

Mappings to 334 CCSR Target Labels. We map all three datasets into a common label space of CCSR codes using the official mapping[4]. We limit our CCSR target space to labels in the Spanish dataset CodiEsp, since it has the biggest variety of CCSR codes. This results in 334 clinical disease categories for CodiEsp, from which 60 appear in Ahepa and 326 in Mimic. We use iterative stratification (Sechidis et al., 2011) to split each of the datasets in train/dev/test depending on the medical conditions of the patients. This leads to a representative group of patients in each of the splits for every clinic separately. In Table 1 we note the number of clinical notes i.e., the number of patients in each of the sets. Note that not all labels appear in each of the splits. The occurence of labels ranges from one to thousand. For example, several types of cancers appear only once and medical conditions that appear most are *Other general signs and symptoms*, *Essential Hypertension* and *Respiratory signs and symptoms* in CodiEsp, Mimic and Ahepa respectively.

Clinical Note Statistics				
	Train	**Dev**	**Test**	**Ø Length**
CodiEsp	656	165	175	351
Ahepa	1,592	402	393	257
Mimic	24,758	6,187	6,182	649

Table 1: Split details of each dataset: CodiEsp, Ahepa and MIMIC. Average length of each note is measured in words.

4. Methods

The task of clinical phenotyping has been rarely investigated in a multilingual scenario. We compare and discuss approaches for cross-lingual transfer that have already been proven successful in other domains. We restrict our approaches to sequential transfer learning, since it allows to share models across clinics without

---

[3]The publication of parts of the data is currently under review.

[4]`https://hcup-us.ahrq.gov/toolssoftware/ccsr/DXCCSR_v2021-2.zip`

having to share patient data explicitly. An alternative approach would be to combine different datasets. However, in this case clinics are forced to share their data which is often difficult due to data regulations. Fig 1 overviews the evaluated approaches.

Translation into English for monolingual models. The first approach we evaluate is to translate all notes into English and predict phenotypes with monolingual language models. While the translation may result in erroneous results, translation services are widely available. Therefore this approach is a decent baseline for common practice.

First, for the task of translating notes from Greek into English, we use pre-trained Opus-MT translation models[5] (Tiedemann and Thottingal, 2020) via EasyNMT[6]. The machine translation of CodiEsp is provided with the corpus and was performed with a system adapted to the biomedical domain. Next, we apply PubMed-BERT (Gu et al., 2020) a state-of-the-art model trained from scratch on a large medical corpus in English and optimized for tasks like ICD code prediction. As a result, we expect the model to produce higher quality representations for domain specific terms. We fine-tune PubMedBERT sequentially on Mimic and on translated data in different combinations.

Pre-trained cross-lingual models. The second approach is the use of pre-trained cross-lingual models or multi-lingual models with cross-lingual capabilities such as XLM-R (Conneau et al., 2020). Cross-lingual models do not require to translate the low- resource data into English. They can represent the same information from different languages close to each other in the embedding space. In contrast, multilingual models are typically pre-trained on only a language modeling task with data from multiple languages. This pre-training schema does not enforce the model to map semantically similar texts close to each other in the embedding space. Nevertheless, Conneau et al. (2020) have shown that even multilingual pre-trained models can have cross-lingual properties and even outperform cross-lingual models in cross-lingual downstream tasks. We use a pre-trained multilingual model with cross-lingual capabilities and fine-tune it sequentially on different clinical datasets. By keeping the label space identical, we force the model to reuse knowledge from one dataset in one language for the next dataset. We first fine-tune on the high-resource dataset and then continue training on the low-resource data. This reflects a scenario of closed data silos (Asiimwe et al., 2021), where only the pre-trained model is accessible, but not the training data.

Language and task adapters. As a third approach, we use adapters (Pfeiffer et al., 2020b) to enhance a multilingual model with additional language- and task

specific parameters. Adapters consist of feed-forward layers that are added to each encoder layer in a transformer model. Typically, only the parameters of the adapters are trained while the rest of the parameters of the base model stay frozen. In consequence, they are an efficient method to fine-tune models since it involves significantly fewer parameters compared to fine-tuning the entire model. Furthermore, adapters can prevent catastrophic forgetting, because the model parameters are frozen and only the adapters are trained. Hence it preserves the ability of a cross-lingual model to map similar inputs from different languages close to each other in the same embedding space even though we train with an unequal number of samples in different languages. We use pre-trained language adapters from AdapterHub (Pfeiffer et al., 2020a) in combination with a cross-lingual pre-trained model to enhance the representation on the low-resource datasets. By reusing the fine-tuned weights of the task adapter for each dataset, we enforce that knowledge from previous fine-tunings is kept and reused for further trainings. In contrast to the task adapters, we keep the language adapters frozen during the training and choose a pre-trained language adapter that matches the language of the training dataset.

Monolingual baseline without knowledge transfer. Finally, we use mono-lingual, domain-specific pre-trained language models which provide a baseline for the training without any knowledge transfer. Domain-specific language models are usually monolingual due to the sparse data available in the subject domain. They are also more common in high-resource languages. In return, the models usually provide an improved representation for technical terms in the specific domain. The most common approach is to train the downstream task on the clinical notes in their native language. For each dataset, we separately fine-tune a different, state-of-the-art pre-trained language model and measure the performance on each dataset individually. However, in this scenario, the model only learns the underlying distribution of labels of the given dataset. We therefore expect this approach to perform worse, especially for long-tail or zero-shot predictions. Furthermore, we also perform experiments with adapters and translation on each target, low-resource dataset in combination with cross-lingual (XLM-R) as well as monolingual pre-trained models as a baseline.

4.1. Experimental Setup

For the monolingual and translation setting, we use PubMedBERT (Gu et al., 2020). For Spanish, we use Spanish-BERT (Cañete et al., 2020) and Spanish Biomedical Clinical RoBERTa (Carrino et al., 2021). Since no domain specific biomedical or clinical language model for Greek has been released to our knowledge, we use Greek-BERT (Koutsikakis et al., 2020), which has been pre-trained on non clinical or biomedical language. We use pre-trained models from Huggingface (Wolf et al., 2020). Due to the limited context

---

Model	Clinical Phenotyping	
	Macro-AUC [%]	Macro PR-AUC [%]
Single Dataset Training		
Monolingual Spanish BERT (C)	82.00	25.91
Spanish Biomedical Clinical RoBERTa (C)	84.58	29.89
XLM-R (C)	56.64	5.28
XLM-R + Adapters (C)	61.96	6.43
Translation + PubMedBERT (C_T)	83.45	29.54
Multi Dataset Training		
XLM-R (M $\rightarrow$ C)	83.52	25.96
XLM-R (M $\rightarrow$ A $\rightarrow$ C)	83.82	25.96
XLM-R + Adapters (M $\rightarrow$ C)	85.63	34.41
XLM-R + Adapters (M $\rightarrow$ A $\rightarrow$ C)	83.90	32.22
Translation + PubMedBERT (M $\rightarrow$ C_T)	**90.95**	**43.13**
Translation + PubMedBERT (M $\rightarrow$ A_T $\rightarrow$ C_T)	90.40	41.98

Table 2: Performance for **CodiEsp**. *M*: Mimic, *A*: Ahepa and *C*: CodiEsp. The order represents the fine-tune order. The subscript $_T$ means that the English translation of the texts is used and otherwise the original language. The approach which yields the strongest results is the sequential fine-tuning of the **Domain specific Encoder** first with Mimic and then with the **English translation** of CodiEsp.

Model	Clinical Phenotyping	
	Macro-AUC [%]	Macro PR-AUC [%]
Single Dataset Training		
Monolingual Greek BERT (A)	90.18	56.22
XLM-R (A)	60.45	12.31
XLM-R + Adapters (A)	56.60	10.30
Translation + PubMedBERT (A_T)	83.15	37.10
Multi Dataset Training		
XLM-R (M $\rightarrow$ A)	89.87	50.23
XLM-R (M $\rightarrow$ C $\rightarrow$ A)	90.03	51.15
XLM-R + Adapters (M $\rightarrow$ A)	90.15	54.45
XLM-R + Adapters (M $\rightarrow$ C $\rightarrow$ A)	**91.50**	**57.63**
Translation + PubMedBERT (M $\rightarrow$ A_T)	86.20	45.14
Translation + PubMedBERT (M $\rightarrow$ C_T $\rightarrow$ A_T)	88.75	49.90

Table 3: Performance for **Ahepa**. *M*: Mimic, *A*: Ahepa and *C*: CodiEsp. The order represents the fine-tune order. The subscript $_T$ means that the English translation of the texts is used and otherwise the original language. The approach which yields the strongest results is the sequential fine-tuning of the **Cross-lingual Encoder plus Adapter** on Mimic, CodiEsp and Ahepa in **original language**.

length of the pre-trained models we truncate all the notes to 512 tokens. We perform a Bayesian hyperparameter tuning for learning rate, accumulation gradients, warm-up steps, hidden dropout rate and attention dropout rate. For optimization we use AdamW (Loshchilov and Hutter, 2019) and for the learning rate we use a learning rate scheduler with warm-up. We run 50 trials for each model with 100 being the maximum number of epochs and use Early Stopping with patience of 5 epochs. We use AUROC as validation metric for the training. For each dataset, we perform a hyperparameter optimization and reuse the weights of the best model for the next dataset. This applies for all experiments where sequential fine-tuning is part of the training. In this manner we

transfer the knowledge from one dataset to the other. We start training from a high-resource dataset and finally fine-tune on the target dataset on which we evaluate the results. For example, in our results the term Mimic $\rightarrow$ CodieEsp (M $\rightarrow$ C) means that the model was first fine-tuned on Mimic and the next model was initialised with the best performing model on the same dataset and finally fine-tuned on CodiEsp.

5. Results

Tables 2 and 3 report the performance on the phenotype classification task on the low-resource datasets CodiEsp and Ahepa, respectively. Each table is separated in two segments, single and multi dataset train-

ing. In the first segment we report the scores of 1) the monolingual BERT of the respective target language, 2) the cross-lingual model (XLM-R), 3) the cross-lingual model with adapters (XLM-R with adapters) and 4) the English translation in combination with a domain-specific encoder (PubMedBERT). We report the performance of the mentioned models trained with only the target dataset for each task. Additionally, we report for the CodiEsp task the performance of a Spanish domain-specific encoder (Spanish Biomedical Clinical RoBERTa). In the second segment we report the scores of the models we use for cross-lingual knowledge transfer. We measure the approaches in macro-average AUROC and PR-AUC.

Knowledge transfer surpasses monolingual models. For both tasks, knowledge transfer incorporating other non-target datasets yields improved results compared to the monolingual counterparts trained on the target dataset only. Table 2 shows the results for CodiEsp. Interestingly, all knowledge transfer methods outperform all single dataset approaches. Translation in combination with a domain specific encoder works best on the CodiEsp dataset. It outperforms the monolingual Spanish BERT by a large margin (11%) in AUROC and the uplift for PR-AUC is even higher (62%). The approach also outperforms the domain-specific Spanish Biomedical Clinical RoBERTa (by 8% AUROC and 44% PR-AUC).

The cross-lingual encoder in combination with adapters is the best performing approach on Ahepa. Table 3 shows that the best model surpasses the performance of the monolingual Greek BERT. In contrast to the CodiEsp task the monolingual model shows high performance and the use of knowledge transfer results to a more moderate increase in both AUROC ($\approx$2%) and PR-AUC ($\approx$3%).

When trained on a single dataset, monolingual models outperform the cross-lingual encoders for both datasets, CodiEsp and Ahepa. This shows that the performance increase is indeed a result of training with clinical datasets in other languages.

Detection of rare conditions improves when using further languages. To test the performance for rare conditions, we aggregate labels according to their frequency in the training sets and report the macro-averaged AUROC in Figure 2 and 3.

For CodiEsp (Fig. 2) we observe that approaches using two or three languages work significantly better for low frequency labels (0-50 times) than single language approaches. The increase in AUROC is 13% in average for the lowest frequency interval (0-10 times). In the interval with medium label frequency (51-100 times) the performance gap decreases. And very frequent labels (>100) are best extracted by a single-language approach. For Ahepa (Fig. 3) multi-language approaches as well outperform the others regarding very low-frequency labels (0-10). Regarding more frequent labels (11-50, 51-100, >100) the monoligual Greek BERT performs
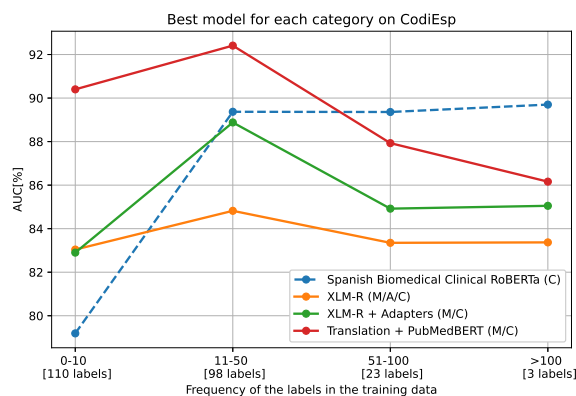

Figure 2: **CodiEsp**: Performance for different label frequencies in the training data. Baselines using only a single dataset are represented by dashed lines. The number of labels in brackets corresponds to the count of labels of each group in the target test dataset (CodiEsp).
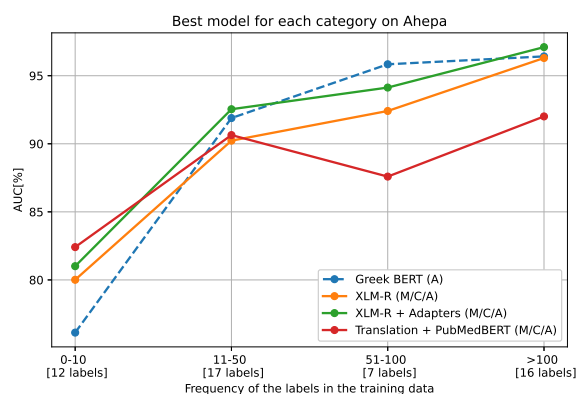

Figure 3: **Ahepa**: Performance for different label frequencies in the training data. Baselines using only a single dataset are represented by dashed lines. The number of labels in brackets corresponds to the count of labels of each group in the target test dataset (Ahepa).

on-par with the cross-lingual Encoder + Adapters approach trained on multiple languages. The differences in AUROC are more moderate than for CodiEsp, but still show that multi-lingual approaches are beneficial for long-tail phenotype detection.

Performance on high-frequency labels. In Figure 2 we observe a drop in performance of the cross-lingual models for the high-frequency labels. The amount of medical categories that fall into the high-frequency groups is very small and they consist of mostly general categories like e.g., *Other specified inflammatory condition of skin* and *Skin/Subcutaneous signs and symptoms*. In contrast the medical conditions in Ahepa are well distributed over the groups (see: number of labels in Fig. 3). In total, 68% of the codes in the CodiEsp test set benefit from knowledge transfer and 62% of the Ahepa test set codes.

Adapters are on par with XLM-R full fine-tuning. The sequential training of the task adapter on all datasets results are on par or have even better performance compared to the full fine-tuning of the XLM-R model for both datasets. Specifically for CodiEsp, both models perform approximately on par, and in the case of Ahepa it is ≈2% better.

6. Discussion

Our results show that cross-lingual transfer improves performance on the clinical phenotyping task–in particular for low-resource datasets. We further conduct a qualitative error analysis with medical doctors revealing specifics of cross-lingual transfer in the clinical domain. This section discusses selected results for applying our findings in a clinical setting.

6.1. Strengths and Limitations of Cross-lingual Transfer

Knowledge transfer is possible despite dataset differences. Clinical notes differ from clinic to clinic in a number of ways. They are written in different languages i.e. English, Spanish or Greek. Second, the form of documentation is different. The Mimic and CodiEsp clinical notes are written in a narrative style in contrast to the Ahepa notes which are much denser. The Ahepa notes are in average shorter, use a lot of abbreviations and some parts are not written in full sentences. Additionally, Greek is typologically different to English and Spanish. In spite of those differences, our experiments show that knowledge transfer between those datasets yields better results than the respective monolingual models. We hypothesise that some differences might, in fact, have a positive impact by adding variance to the data.

Limitations of cross-lingual knowledge transfer. Tables 2 and 3 show that cross-lingual transfer does not improve performance on all diagnoses groups. Our error analysis reveals that the effect of transferring knowledge for the same medical category can have opposite results for the two studied datasets. We observe this especially in broad categories such as *Other specified and unspecified skin disorders*. While these work well in Ahepa, the transfer fails for CodiEsp, which might be due to the diverse set of clinics included in CodiEsp resulting in a larger number of diagnoses per category. We also observe cases in which the monolingual Spanish and Greek encoder almost perfectly recognises relevant clinical cases and knowledge transfer has a negative impact. This is the case e.g. for *Pericarditis and pericardial disease*. Since every clinic has a different focus, the importance of clinical facts related to the medical condition is different. This has a direct impact on the amount of details a physician reports about a clinical fact. Consequently, to identify the same medical category different clinical textual cues are needed depending on the clinic it originates from.

The benefit for rare phenotype detection. Knowledge transfer for rare phenotypes is not only useful for languages with little labeled data but also for countries with smaller populations like Greece. For instance, *Takotsubo cardiomyopathy* is a disease with low occurrence (Y-Hassan and Tornvall, 2018). Although the frequency of the medical condition may not be different from clinic to clinic, low-resource datasets benefit from datasets with bigger volume of labeled data. In our results we observe an increase of 2pp for the category *Other and ill-defined heart disease* to which the condition belongs. Similarly, for the Ahepa task we observe an increase of 2pp for *Acute and unspecified renal failure*, where the cardiology clinic might profit from CodiEsp data from nephrology or pathology clinics. In both cases we observe that the data augmentation improves results for both low-resource datasets because of the higher text volume and added variance.

6.2. Design Choices in Clinical NLP Setting

Adapters and translation are suitable methods for cross-lingual knowledge transfer. Our results suggest that both techniques, adapters and translation to English, are suitable ways for clinical knowledge transfer. In the case of CodiEsp the translation to English and sequentially fine-tuning a large English domain-specific encoder yields the best results. For the Ahepa case, the pre-trained cross-lingual model with adapters trained with clinical notes in native language achieves the best results. In both cases, a transfer method outperforms all other corresponding methods for that data set.

'Rules-of-thumb' considerations. In order to improve the performance of clinical phenotyping some considerations are necessary to pick the best cross-lingual solution. Is there a well performing in-domain translation system[7] to English available for the datasets at hand? If this is the case, sequentially fine-tuning a large English domain-specific encoder is a promising solution. However, if no high quality machine translation is available, fine-tuning the adapters in native language is the better choice. In addition, further experiments revealed that the results vary depending on the order with which we fine-tune. Starting the training with the high-resource language mostly yields good results.

Adapters are efficient for cross-lingual transfer. The adapter approach proved to be an efficient way for cross-lingual knowledge transfer for the phenotyping task. This comes from the fact that, we only train the task adapter parameters. So, while the system adds additional parameters to the model to better represent low-resource languages, only a fraction of all the parameters is updated during training. Similar to Pfeiffer et al. (2020b), we observe on par or better results by training only the task adapter in comparison to the full

---

[7]The system should be deployable within the premises of a clinic due to regulations and sensitivity of patient data.

Original text	Failed machine translation	Doctor's translation
Ο ασθενής προσήλθε λόγω δύσπνοιας από ημερών	The patient came due to shortness of days	The patient was admitted due to current *dyspnoea* since days
ανοιχτός βοτάλλειος πόρος	open shell resource	patent ductus arteriosus (PDA)
ΑΥ	Authorize	Arterial Hypertension
la paciente presenta dolor y *defensa a la palpación profunda* de fosa lumbar derecha	the patient presented with pain and *defense of the deep fixation* of the right lumbar fossa	patient presented with pain and *defense to deep palpation* of the right lumbar fossa
realizó estudio *oftalmológico* e inició *con deflazacort*	performed the *diagnostic test*, and initiated treatment *with dementia*	performed an *ophthalmologic* study and started treatment *with deflazacort*
no fumadora *ni bebedora de alcohol*	non-smoker or *bicuspid bicuspid drinker*	non-smoker and *non-drinker of alcohol*

Table 4: Text excerpts where machine translation failed. We show examples of both the translation of the Greek and Spanish clinical notes. Note, that the Spanish machine translation system was fine-tuned on biomedical data and the Greek system was not.

cross-lingual model in both cases, Ahepa and CodiEsp, during sequential fine-tuning of those encoders. Thus, in case computational complexity is a limiting factor, a cross-lingual encoder like XLM-R plus adapters is the recommended approach.

6.3. Additional Observations

Mixed results on adding various languages. Tables 2 and 3 compare knowledge transfer methods fine-tuned on two or on three datasets. We found that adding more data does not necessarily improve results. The Ahepa task yields the strongest results when incorporating all three datasets. However, in the case of CodiEsp, we observe that fine-tuning on two languages performs on par or better than on three languages. The best result for the CodiEsp task was achieved using only English and Spanish text instead of additional Greek data. We leave further analyses on which language combinations work best to future work.

Language typology impacts transfer capabilities. When adding Greek as a third language in the CodiEsp task, knowledge transfer with XLM-R does not improve the performance. It even decreases when using the adapters approach. This may occur due to the different typologies of the languages and the limited shared vocabulary. Interestingly, when adding Spanish as a third language in the Ahepa task the performance for the adapters approach improves. This may be because the Greek texts contain some standard medical expressions, such as *STEMI*, which are also used in Spanish and English notes.

Impact of translation quality and consistency. The approach based on English translations yields lower performance increase for Ahepa than for CodiEsp. This might happen because we use two different machine translation systems: A domain-specific translator for

CodiEsp and a general one for Ahepa. Translating the same medical terms differently produces different representations and therefore may have a negative or at best no effect on the performance of the downstream task. In Table 4 we report examples of failed translations. Most of the problems we observe with the Greek machine translator are the medical terms and abbreviations doctors frequently use. In the Ahepa (Greek) notes we also notice that sometimes punctuation and accents on words are missing, which can be decisive when translating.

7. Conclusion

Our results demonstrate that cross-lingual knowledge transfer improves performance in the clinical phenotype prediction task. In particular we show that knowledge transfer yields the strongest uplift in performance on rare phenotypes. Thus, medical tasks whose targets can be mapped to a single scheme can benefit from adding data in different languages. This is an important result in the medical domain where worldwide standards for describing patient conditions exist. We also give recommendations on how to effectively and efficiently solve such a task that involves notes written in rather distant languages. Our future work will focus on 1) methods specifically targeting clinical cross-lingual transfer and 2) studying the effects of integrating more languages and different machine translation systems. 3) Finally we want to investigate the benefit of cross-lingual transfer for extreme low-resource languages.

Acknowledgments

8. Bibliographical References

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Artetxe, M., Labaka, G., and Agirre, E. (2020). Translation artifacts in cross-lingual transfer learning. In Bonnie Webber, et al., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7674–7684. Association for Computational Linguistics.

Asiimwe, R., Lam, S., Leung, S., Wang, S., Wan, R., Tinker, A., McAlpine, J. N., Woo, M. M. M., Huntsman, D. G., and Talhouk, A. (2021). From biobank and data silos into a data commons: convergence to support translational medicine. *Journal of Translational Medicine*, 19(1):493, Dec.

Carrino, C. P., Armengol-Estapé, J., Gutiérrez-Fandiño, A., Llop-Palao, J., Pàmies, M., Gonzalez-Agirre, A., and Villegas, M. (2021). Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *CoRR*, abs/2109.03570.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chung, H. W., Garrette, D., Tan, K. C., and Riesa, J. (2020). Improving multilingual models with language-clustered vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, Online, November. Association for Computational Linguistics.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In Hanna M. Wallach, et al., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, et al., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Isbister, T., Carlsson, F., and Sahlgren, M. (2021). Should we stop training more monolingual models, and simply use machine translation instead? In Simon Dobnik et al., editors, *Proceedings of the 23rd Nordic Conference on Computational Linguistics, NoDaLiDa 2021, Reykjavik, Iceland (Online), May 31 - June 2, 2021*, pages 385–390. Linköping University Electronic Press, Sweden.

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., and Androutsopoulos, I. (2020). GREEK-BERT: the greeks visiting sesame street. *CoRR*, abs/2008.12014.

Labonté, R., Mohindra, K. S., and Schrecker, T. (2011). The growing impact of globalization for health and public health practice. *Annual review of public health*, 32:263–83.

Le, H., Pino, J. M., Wang, C., Gu, J., Schwab, D., and Besacier, L. (2021). Lightweight adapter tuning for multilingual speech translation. In *ACL/IJCNLP*.

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., and Salimi-Khorshidi, G. (2020). Behrt: Transformer for electronic health records. *Scientific Reports*, 10.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mann, G. S. and Yarowsky, D. (2001). Multipath translation lexicon induction via bridge languages. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Mulyar, A., Schumacher, E., Rouhizadeh, M., and Dredze, M. (2019). Phenotyping of clinical notes with improved document classification models using contextualized neural language models. *CoRR*, abs/1910.13664.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., and Gurevych, I. (2020a). Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Pfeiffer, J., Vulic, I., Gurevych, I., and Ruder, S. (2020b). MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In Bonnie Webber, et al., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7654–7673. Association for Computational Linguistics.

Ponti, E., Glavavs, G., Majewska, O., Liu, Q., Vuli'c, I., and Korhonen, A. (2020). Xcopa: A multilingual dataset for causal commonsense reasoning. In *EMNLP*.

Schäfer, H. and Friedrich, C. M. (2020). Multilingual ICD-10 code assignment with transformer ar-

chitectures using MIMIC-III discharge summaries. In Linda Cappellato, et al., editors, *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., e Oliveira, L. E. S., Copara, J., Gumiel, Y. B., de Oliveira, L. F. A., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. (2020). Biobertpt - A portuguese neural language model for clinical named entity recognition. In Anna Rumshisky, et al., editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop, ClinicalNLP@EMNLP 2020, Online, November 19, 2020*, pages 65–72. Association for Computational Linguistics.

Sechidis, K., Tsoumakas, G., and Vlahavas, I. P. (2011). On the stratification of multi-label data. In Dimitrios Gunopulos, et al., editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III*, volume 6913 of *Lecture Notes in Computer Science*, pages 145–158. Springer.

Soares, F. and Krallinger, M. (2019). BSC participation in the WMT translation of biomedical abstracts. In Ondrej Bojar, et al., editors, *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, pages 175–178. Association for Computational Linguistics.

Solt, I., Tikk, D., Gál, V., and Kardkovács, Z. T. (2009). Research paper: Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. *J. Am. Medical Informatics Assoc.*, 16(4):580–584.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT - building open translation services for the world. In Mikel L. Forcada, et al., editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, pages 479–480. European Association for Machine Translation.

van Aken, B., Papaioannou, J., Mayrdorfer, M., Budde, K., Gers, F. A., and Löser, A. (2021). Clinical outcome prediction from admission notes using self-supervised knowledge integration. In Paola Merlo, et al., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 881–893. Association for Computational Linguistics.

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., and Liu, H. (2019). A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics Decis. Mak.*, 19(1):1:1–1:13.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue,

C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Y-Hassan, S. and Tornvall, P. (2018). Epidemiology, pathogenesis, and management of takotsubo syndrome. *Clinical autonomic research : official journal of the Clinical Autonomic Research Society*, 28(1):53–65, Feb. PMC5805795[pmcid].

Yao, L., Mao, C., and Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics Decis. Mak.*, 19-S(3):31–39.

Zhang, J., Bolanos Trujillo, L., Li, T., Tanwar, A., Freire, G., Yang, X., Ive, J., Gupta, V., and Guo, Y. (2021). Self-supervised detection of contextual synonyms in a multi-class setting: Phenotype annotation use case. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8754–8769, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

9. Language Resource References

Johnson, Alistair E.W. and Pollard, Tom J. and Shen, Lu and Lehman, Li-wei H. and Feng, Mengling and Ghassemi, Mohammad and Moody, Benjamin and Szolovits, Peter and Anthony Celi, Leo and Mark, Roger G. (2016). *MIMIC-III, a freely accessible critical care database*.

Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., and Krallinger, M. (2020). Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF ehealth 2020. In Linda Cappellato, et al., editors, *Working Notes of CLEF 2020 - Conference and Labs off the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.