Leveraging Information Redundancy of Real-World Data Through Distant Supervision

Anonymous submission

Abstract

We explore the task of event extraction and classification by harnessing the power of distant supervision. We present a novel text labeling method leveraging the redundancy of temporal information in a data lake. This method enables the creation of large programmatically annotated corpus, allowing the training of transformer models in a distant supervision manner. This aim to reduce expert annotation time, a scarce and expensive resource. Our approach utilizes temporal redundancy between structured sources and text, enabling the design of a replicable framework applicable to diverse real-world databases and use cases. We employ this method to create multiple silver datasets to reconstruct key events in cancer patients' pathways, using clinical notes from a cohort of 380,000 oncological patients. By employing various noise label management techniques, we validate our end-to-end approach and compare it with a baseline classifier built on expert-annotated data. The implications of our work extend to accelerating downstream applications, such as patient recruitment for clinical trials, treatment effectiveness studies, survival analysis, and epidemiology research. While our study showcases the potential of the method, there remain avenues for further exploration, including advanced noise management techniques, semi-supervised approaches, and a deeper understanding of biases in the generated datasets and models.

Keywords: Distant Supervision, Electronic Health Records, Programmatic Annotation

1. Introduction

Natural language processing (NLP) techniques applied to health care notes have already shown satisfactory results in the literature, in particular with supervised learning approaches based on machine learning (ML) (Esteva et al., 2019; Hahn and Oleynik, 2020; Wu et al., 2020). However, this good performance depends strongly on the existence of many annotated records and, moreover, these annotations must be performed by domain experts. This annotation task is in practice a bottleneck for the development of research because the experts' available time is a scarce and expensive resource (Ratner et al., 2020, 2016). Furthermore, the majority of annotated datasets issued from clinical notes could not be shared and reused due to patient privacy regulations (Carlini et al., 2021; Lehman et al., 2021).

On the other hand, the recent digitization of health records and their collection, in a near real-time basis, in Clinical Data Warehouses (CDWs) offer new perspectives for research, steering activities and policy making. Electronic Health Records (EHRs) contain extensive clinical data that have been collected as part of routine care and may be analyzed leveraging big data technologies (Kim et al., 2019; Wang et al., 2018). The collected data generated at the hospital come from multiple sources (biology, demographic, drug prescription, claim data, clinical notes, images, etc.) and the information becomes partially redundant in multiple cases (Suri et al., 2020; Cabitza et al., 2019, 2005). Additionally, it could be partial or incomplete if we consider each source separately. Even if promising, CDWs gather the information of millions of EHRs and taking advantage of them is still a current challenge, on the one hand because of the limitations and biases of Real-World Data (RWD), and on the other hand because of the volumetry (Big Data) and the impossibility of manually reviewing each record (Hanauer et al., 2015; Newgard et al., 2012). This makes a fertile ground for the development of algorithms and ML.

The distant supervision approach allows the use of multiple data sources to build annotated datasets automatically, consequently, much faster than what can be produced by a manual annotation (Ratner et al., 2016). However, this programmatic annotation is imperfect, producing a silver standard dataset with unreliable labels, also called noisy labels. Deep neural networks (DNNs) are susceptible to overfit on noisy labels due to the large number of model parameters (Zhang et al., 2017) and therefore several efforts and methods have been developed to be able to learn from noisy labels with DNNs, even though most of these methods have been developed for image applications (Song et al., 2022).

We address in this work the task of event extraction and classification. We leverage information redundancy present in a portion of EHRs to build a large programmatically annotated corpus which allows us to fit a transformer model in a distant supervision fashion. This model is therefore used for inference where information is not redundant or lacking. Overall, we propose a new text labeling method that leverages temporal information redundancy from external sources without using any other text information apart from dates itself, providing a framework that can be replicated for other use cases or in other real-world databases. Then,

we test this method to build multiple silver datasets for classification in order to reconstruct key events of cancer patients' pathway using clinical notes of a cohort of 380,000 patients. We train a classifier using different methods for noise label management, we validate the end-to-end approach and we compare it with a baseline classifier using an expert annotated corpus.

2. Related Work

NLP applications in the clinical text domain encompass various tasks and use cases. Named Entity Recognition (NER) and normalization are among the most common, as recently demonstrated in a comprehensive review of concept extraction by Fu et al. (2020). For instance, Lerner et al. (2020) employed a neural network for clinical entity recognition, while Zhang et al. (2019) utilized a deep learning NER model for a breast cancer application. Text classification and entity characterization tasks have also been widely explored. For example, Wu et al. (2014) delved into the classification of negation and hypothesis forms within sentences, which holds significant relevance when analyzing clinical text. Lastly, relation extraction is a well-established area in clinical NLP. It encompasses diverse applications such as temporal relationship extraction, disease-gene associations, and drug-dose relationships, as evidenced by Lv et al. (2016); Bose et al. (2021).

In contrast to traditional NLP approaches reliant on supervised learning, the weak supervised approach is a burgeoning field that aims to reduce or eliminate the need for human annotation. Health applications have witnessed innovations in this domain, including the use of external ontologies for deep learning NER on EHR, as demonstrated by Fries et al. (2021). Weak supervision and deep representation have also been applied to clinical text classification (Wang et al., 2019), for medical image classification using clinical notes to weakly label images (Dunnmon et al., 2020), and for the relation extraction task (Wang et al., 2022; Amin et al., 2020; Mintz et al., 2009; Zhao et al., 2020). In addition, this approach has been used for phenotype extraction from EHR while minimizing the annotation effort (Agarwal et al., 2016; Halpern et al., 2016) .

Simultaneously, the ML community is actively developing novel methods for training models with noisy labels. Conventional regularization techniques, such as data augmentation, weight decay, dropout, and batch normalization, have seen extensive use to mitigate overfitting to noise. However, they may not entirely resolve the issue. As a result, achieving strong generalization capabilities in the presence of noisy labels remains a significant challenge, as highlighted by Hedderich et al. (2021) and Song et al. (2022). Active research directions in enhancing noise label robustness include sample selection, robust regularization, robust loss functions, robust architecture, and others.

3. Methodology

We consider a task of event extraction and classification, from an organizational data lake containing text and other sources of information about events. All these sources can be noisy or incomplete. Multiple facets of the same real-world event could be represented in different data sources of an organization (company, hospital, public institution, etc.), producing a potential redundancy of information. In the example illustrated by Figure 1, the information about the yellow type event (A) is stored both in a structured form and in an unstructured textual form. For some reason, the second yellow event (B) is not collected in the structured database, but it is systematically recorded in the text. Under the assumption that both yellow type events are mentioned using a similar language distribution, we will leverage the text around the mention of the first event to train a model which allows us to identify from text the second yellow event. We use the time dimension as a pivot to align complementary information present in different data sources.

In the rest of this section, we describe our temporal text alignment algorithm for date labeling (programmatic annotation). Then, we briefly introduce the noise management methods used and the adopted modelling.

3.1. Programmatic Annotation

3.1.1. Positive examples

We develop a general method for date alignment between an external source of structured data including event dates and a paired text corpus (e.g., linked by a person/client/user identifier). The algorithm is based on the hypothesis that if the event date (known from a structured source) is mentioned on a text of the same person/client/user, there is a probability P_0 that the text context around the identified date refers to the event. Inherently, the text context could not refer to the targeted event with probability $1 - P_0$. We assume that for these latter cases, the topic distribution of text context is not systematic (and ideally random), producing therefore a label noise that could be statistically distinguished. Our method should lead to a P_0 greater than annotation performed through random selection.

The algorithm consists of a first step of NLP rule-based dates recognition (named entity recognition task - NER) in the text corpus using the open source

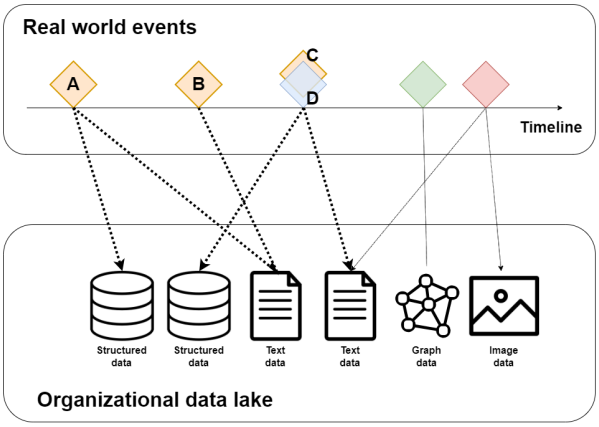

Figure 1: Illustration of information redundancy in an organizational data lake. The same facet of a real-world event could be represented in different data sources and formats. In the example, some information of the first yellow type event (A) is stored in a structured form and some other in an unstructured text form. For some reason, the second yellow event (B) is not collected in the structured database, but it is systematically recorded in text. Furthermore, multiple events could take place at the same moment (C and D) and both recorded in multiple databases.

edsnlp implementation (Dura et al., 2022). This algorithm recognizes absolute and relative dates in text presented in multiple forms (e.g. “01/01/2010”, “the first of January”, “01/01”, “1.1.2010”, “last Tuesday”, etc) and normalizes them. Secondly, for the alignment step, we select only dates mentioned once in the document (e.g., we discard date mentions corresponding to event C and D of Figure 1 if multiple mentions are found). The rationale behind this selection is to prevent erroneous matches caused by other events mentioned in the text sharing the same date, we aim to be precise and not exhaustive (we want to maximize P_0). Then, for each person/client/user, we join the dates from the structured data to the ones extracted from the text. This results in “aligned dates”, i.e., a set of labeled dates with respect to their event type. As we work with normalized dates, our algorithm allows a temporal approximate matching with a given threshold.

3.1.2. Negative Examples

We propose two methods for the selection of alternative examples (i.e., a negative class in the case of binary classification):

- Random Selection (RS): A naive random selection among all non-matched dates in the previous step;
- Proximity Selection (PS): A random selection weighted by a score of distance in text with respect to the matched dates.

The second approach is more likely to choose dates close to the ones already identified in the text, which are presumably more challenging examples. However, this method also has the potential for a higher rate of false negatives.

3.1.3. Dataset Creation

For each labeled date, we extract a textual context surrounding the date, using a window of k_1 words before and k_2 words after it. Consequently, each training example in the dataset includes a text snippet, an offset denoting the start and end of the date entity, and a label that results from the alignment process. Additionally, the date to be classified is substituted with a mask token. This entire approach generates a dataset for classification with labels that may have some noise (referred to as silver labels).

3.1.4. Illustrative Example

We illustrate our approach with a task consisting in extracting the dates of three event types from clinical reports: biopsy[1], surgery[2] and an "other" class. Table 1 shows several text examples, together with their automatic silver label and the ground truth. Example 1 represents event A of Figure 1, i.e., the event took place in the hospital and it was recorded both in a structured database and in clinical reports. On the other hand, example 2 represents event B of Figure 1, i.e., the biopsy event was done outside the hospital (ambulatory care) and therefore the only reference to this event is available in text (see Figure 2 and Section 4 for more details about our application use case).

One notable distinction between our approach and relation extraction methods, which focus on predicting relationships between entity pairs identified in text (such as the connection between a date and a concept) (Bach and Badaskar, 2007), is that our annotation method is not reliant on any NER task for identifying the concepts. No previous knowledge vocabulary dictionary or model is used. This allows the model by itself to learn the label of date entities from context. This approach is particularly interesting when concepts are not explicitly mentioned, or vocabulary could be very

[1]A biopsy is a medical procedure that involves the extraction of sample cells or tissues for examination to determine the presence or extent of a disease, the tissue is generally examined under a microscope by a pathologist.

[2]Medical procedure that could occasionally lead to an extraction of tissue for analysis done by a pathologist.

ID	Text	Silver label	Gold label
1	The patient was biopsied on <masked date> at the hospital	Biopsy	Biopsy
2	The patient showed the results of the biopsy done outside the hospital on <masked date> [..]	Other	Biopsy
3	digestive endoscopy on <masked date>. The pathological analysis indicates [..] On 17.08.2019 the patient underwent abdominal surgery	Biopsy	Biopsy
4	digestive endoscopy on February 15th, 2019. The pathological analysis indicates [..] On <masked date> the patient underwent abdominal surgery	Surgery	Surgery
5	Mr. Smith came to consultation on <masked date>	Other	Other
6	Mrs. Dupont came to consultation on <masked date>	Biopsy	Other

Table 1: Examples of produced data points. The silver label is the product of the programmatic annotation and the gold label is the one given by an expert. For example 2 a biopsy is mentioned in the text, but it is not in the structured records, producing a false negative label. For the example 6, the date corresponds to a biopsy date but also to another event of the same day, leading to a false positive label.

heterogeneous. In the example 3 of Table 1, no explicit mention of biopsy is done, there is a mention of an endoscopy[3] followed by a mention of anatomical pathology analysis[4], the combination of these two mentions implies an implicit biopsy, however none of the concepts separately implies a biopsy. Example 4 illustrates the lexical similarity between two types of different events (Biopsy and Surgery).

Example 5 corresponds to a correctly labeled alternative example extracted by a random selection among all non-matched dates. Example 6 represents a case when multiple events happen at the same day. A date is labeled as a biopsy with the alignment procedure, but the text does not mention any biopsy, leading to a mislabeled example.

Note that for each snippet, only the date to be classified is masked, for which we know the position in text.

3.2. Modelling

As our experiments are carried out on texts in French, we use CamemBERT (Martin et al., 2020) for the text encoder. To classify the date spans, we add a RoBERTa token classification head (Liu et al., 2019) that takes as input the vector representation at the masked position (illustration in supplementary materials). Nevertheless, it's possible to utilize a different contextual encoder.

The rationale behind masking the date is that it is not the date in isolation that signifies the event we aim to detect, but rather the context surrounding it. Consequently, the vector representation of the masked date incorporates the surrounding text context.

[3]Medical procedure to look inside the body.

[4]The laboratory examination of organs and tissues for the diagnosis of disease.

3.3. Noise Management

In order to learn with a dataset with noisy labels, we test three methods selected from literature. We choose them considering the generalization capability to be extended to other applications (e.g., we avoided methods based on specific robust architectures), the implementation simplicity, the need for gold label annotated data, the existence of theoretical demonstration and the proved performances in other contexts. With these criteria and following the taxonomy presented by Song et al. (2022), we select a method from the family of Robust Loss Functions, one from the Sample Selection family and a last from the Label Refurbishment family.

3.3.1. Robust Loss Function

In the context of ML and optimization, a robust loss function refers to a loss function that is less sensitive or resistant to outliers or noisy data points, compared to regular clean points. The term "robust" here refers to the ability of the loss function to provide reliable and consistent optimization results in the presence of data that deviates from the underlying assumptions of the model. The Cross Entropy (CE) loss has been shown to be not robust against label noise in classification tasks (Ghosh et al., 2017). We used the Normalized Cross En-

tropy - Reverse Cross Entropy (NCE-RCE) loss, an Active-Passive loss introduced by Ma et al. (2020), with theoretical robustness properties.

3.3.2. Sample Selection

This category of methods aims to select true-labeled examples from a noisy training dataset. We tested the O2U-Net approach (Huang et al., 2019). The idea of the authors of the O2U-Net method is to introduce multiple rounds of status transfer in training, changing between underfitting and overfitting using a cyclical learning rate. A large learning rate is first applied, then linearly decreases, and is then reset to the original value to jump from overfitting to underfitting (O2U) cyclically. After the whole cyclical training, the average of the losses of every sample is computed. All the average losses are then ranked in descending order, and the top k% (forget rate) of samples are removed from the original dataset as noisy labels. Then a final step of training is done on cleaned data. We tested this approach in two forms: a O2U sample selection step followed by a regular training using the CE loss, and a second implementation using the NCE-RCE loss during the training step.

3.3.3. Label Refurbishment

This third method is based on a similar idea. When training a network on a noisy dataset, the value of probabilities granted by the network to the different possible labels for each sample is an indicator of the complexity - or noisiness - of the data. In their work, Zheng et al. (2020) give theoretical proof that when a classifier trained on noisy data has low confidence in the label of a sample, that label is likely to be false. First the method applies a Likelihood Ratio Test (LRT) on noisy classifier predictions to check label purity, the likelihood ratio is compared with a predetermined threshold δ; then, it corrects wrong labels for future training. To avoid overfitting of the DNN, they recommend the use of a robust loss function. For simplicity reasons, we adopt the NCE-RCE loss presented before.

4. Application

The CDW of a big hospital[5] contains the EHR of 11.4 million patients, including 380,000 patients with cancer and around 35,000 new cancer cases each year. We address the problem of cancer patient journey reconstruction, particularly we focus on the diagnosis date because of its importance in oncology research (e.g., implication in treatment effectiveness studies, survival analysis, epidemiology, etc.). Although cancer patients receive close monitoring at the hospital, their diagnosis, typically determined through a biopsy procedure, is often conducted in an ambulatory care setting (outside the hospital). As a result, this information is not readily available in a structured format. Nonetheless, clinicians do transcribe this crucial date in an unstructured and varied manner during systematic reporting in clinical documents. Ultimately, only around 30% of patients undergo their cancer diagnostic biopsy at the hospital. In such cases, the biopsy date information is duplicated, appearing in both the textual clinical report and a structured format. We will leverage the redundancy of information present for this subset of patients in order to build a programmatic labeled dataset of biopsy dates and train a model to classify a given date as corresponding to a biopsy or not (see Figure 2).

4.1. Experiment Setting

4.1.1. Initial Data Selection

After discussion with oncologists, we select two different external sources of information. For patients in the cohort (see supplementary materials), we extract biopsy dates present in an semi-structured form in pathology reports, and surgery dates extracted from CCAM[6] structured claim data. For both sources, we extract a table with the information of the date, the event type and a patient identifier. For these patients, we select their clinical reports[7] as the text corpus (in French) in order to apply our proposed methodology. Note that patients for which a structured event exists have only been used for the training step.

We conduct experiments using only the pathology reports source and using both sources (Section 4.1.3).

4.1.2. Test & Development Corpus

We reserve 101 and 60 documents for the test and development set, respectively. The selection of the documents for the development set is biased in order to find potentially alignable dates, in contrast to the test set sampled in a random way. This corresponds to 1,474 and 680 date entities annotated in a binary way for the test and development set respectively (4.5% and 10% correspond to biopsies). All dates present in both corpora are annotated by a senior oncologist. The development set (expert labeled - **EL**) is used for three purposes, to optimize hyper-parameters, to use it as a train set to the proposed approach against a baseline training with an

[5] Anonymized for review. This information will be provided in the final version.

[6] French medical procedures classification ("*Classification Commune des Actes médicaux*" - CCAM).

[7] Multidisciplinary meeting reports (MDM).

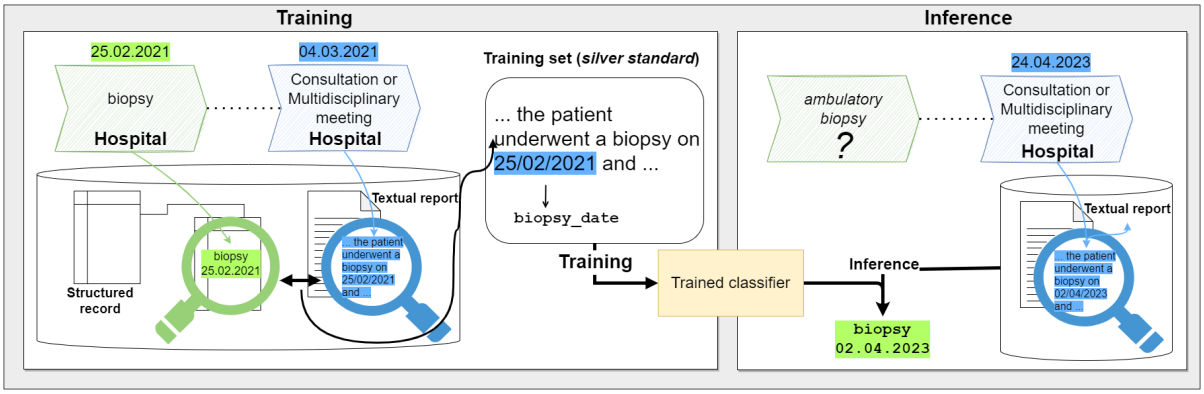

Figure 2: Representation of biopsy dates dataset generation and trained model application. The training set is constituted by means of redundant information from patients treated in the same hospital from the first examinations (here, a biopsy event is recorded, and mentioned later in the text reports). The trained model is applied to patient records whose history is only mentioned in the texts written at admission (here an ambulatory biopsy).

annotated corpus, and to evaluate the performance of the alignment strategy.

4.1.3. Programmatic Annotation

We apply our proposed programmatic annotation on four settings, product of the combination of different data sources and different methods of selection of alternative examples. These are:

1. **PR-RS**: Pathology Reports and Random Selection.
2. **PR-PS**: Pathology Reports and Proximity Selection.
3. **PR-SP-RS**: Pathology Reports, Surgery Procedures and Random Selection.
4. **PR-SP-PS**: Pathology Reports, Surgery Procedures and Proximity Selection.

PR-RS and PR-PS datasets are used to train a binary classifier (biopsy or not). Alternatively, the PR-SP-RS and PR-SP-PS datasets are used to train a three classes classifier (biopsy, surgery, other). If for a patient the same date has two different labels, priority is given to the surgery label. For each data point, we extract a window of 30 and 45 words before and after the date of interest, respectively. We perform a binomial test against the null hypothesis H_0 of a labeling by chance.

4.1.4. Training Configurations

For each produced dataset we test different training configurations:

1. Regular training with CE loss.
2. NCE-RCE loss.
3. O2U strategy with CE loss.
4. O2U strategy with NCE-RCE loss.
5. LRT strategy.

For all experiences, we use a pre-trained CamemBERT language model trained on French clinical reports (Dura et al., 2022). Also, we performed each experience with 5 different seeds in order to explore the robustness of each model. We use the AdamW optimizer (see supplementary material for details) (Loshchilov and Hutter, 2019). As no recommendation exists for the application of these strategies to NLP tasks, we use 15 documents (105 entities) randomly sampled from the development set to search for good hyper-parameters (see supplementary material). We set α to 0.1 and β to 1 for the NCE-RCE loss. We apply the O2U framework for 4 cycles of 6 epochs each and we set the forget rate to 30%. During the cycle step the CE loss was used. For the LRT training we start switching labels at the end of epoch 2, we set δ to 1.3 and we use the NCE-RCE loss from epoch 0.

We perform multiple training with different number of examples of the development set in order to measure performance as a function of the size of the expert annotated training data. We also compare the performance within patients with and without a biopsy done at hospital.

5. Results

Following our proposed distant annotation methodology, we produced four programmatically annotated datasets. The PR-RS and PR-PS are composed of 10,850 data points with two classes (50% biopsies). When combining with a third resource (PR-SP-RS and PR-SP-PS datasets), we obtain a dataset of 16,200 examples with three classes

(33% biopsies). As shown in Table 2, when comparing the four programmatically annotated datasets, the best results of F1-score are achieved when combining different sources of information. Besides, we obtain the best results when combining the label refurbishment method (LRT) with a robust loss function.

Table 3 shows the confusion matrix of the programmatic annotation for the PR-RS dataset measured over the development dataset. It allows us to estimate the degree of noise of the aligned dataset in a given setting. The precision of the method for the biopsy class obtained for this case is 0.91 and 0.96 for the alternative class, before application of any noise management strategy.

As shown in Figure 3, model performance increases in function of the number of examples used for training using an expert annotated dataset. However, performance increase is slower when using more than 20 documents. The distant supervised approaches perform comparably to the results obtained with 450 clean labeled entities (Table 2).

Lastly, in Table 4 , we present a comparison of model performance for two patient groups: those who underwent at least one hospital biopsy, and the others. Notably, we identify higher recall rates among the first group. Upon examining the randomly sampled test set, we discover that 29% of patients in the dataset underwent at least one hospital biopsy, and this subgroup contributes to 45% of the entities labeled as "biopsy" by the expert clinician.

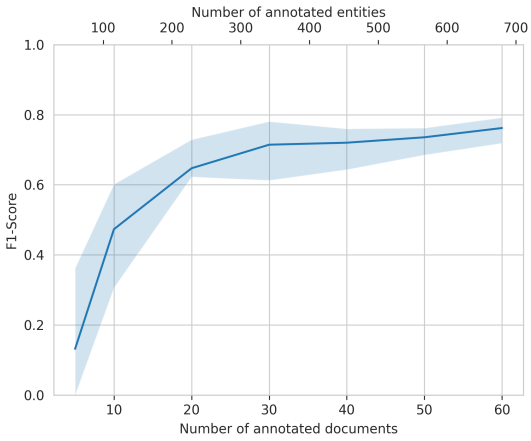

Figure 3: Median F1-score ([min-max] over 5 iterations) in function of training examples using an expert labeled dataset. The line represents the median F1-score of 5 iterations and the area represent the min and max F1-score of each configuration. For each iteration a random subset of documents of the corpus is selected and a different inital random seed is used for training. The lower X axis represent the number of annotated documents and the upper X axis represents the mean number of entities.

6. Discussion

We confirm that in a context of an organizational data lake such as a clinical data warehouse, it is possible to leverage information redundancy present in a portion of data to build a large programmatically annotated corpus which allows us to learn patterns from text in a distant supervision fashion. As shown in Table 2, the proposed approach achieves successful results (median F1-score of 0.70), comparable with models fitted on an expert annotated corpus (median F1-score of 0.76). As shown in Figure 3, a turning point between a supervised strategy and distant supervised one is found; This point will depend on each application use case. Moreover, combining the development set with the programmatically annotated one in order to train a model in a semi-supervised approach should be considered.

Our findings validate the presence of information redundancy in an organizational data lake that holds data related to real-world events. Notably, our algorithm relies solely on the knowledge of dates sourced from multiple data sources. In Table 3, we show the estimate of P_0 following our proposed methodology, proving that the annotation method based on date mentions performs much better than a random labeling (p-value $< 10^{-200}$). Moreover, because our method does not use other text information apart from dates itself, we assume that the produced label noise is independent from text.

Moreover, the benefit of using noise management methods is also demonstrated, we improve performance up to 59%, all approaches are distinctly better than a classic training using the CE loss, even in an non optimal setting of hyper-parameters. However, the choice between the different explored noise management methods and their hyper-parameters is still not clear. Different training experiences lead to large confidence intervals just varying the random seed. Further research or experiences should be done concerning stability on the convergence of training models in this setting.

In line with results shown by Suri et al. (2020), it has been experimentally shown that the use of multiple sources combined together reach better results (PR-* vs. PR-SP-* datasets). This could be explained by two possible reasons. On the one hand none of the external information sources are intrinsically reliable, so the combination of both in the labeling process seems to be more robust. On the other hand, as illustrated in section 3.1.4, the mention of biopsy and surgery events could share similar vocabulary; Therefore, giving the model the possibility to discriminate the vector space between

Method	PR-RS	PR-PS	PR-SP-RS	PR-SP-PS
CE	0.54 (0.53 - 0.54)	0.49 (0.44 - 0.54)	0.46 (0.45 - 0.48)	0.44 (0.42 - 0.45)
NCE-RCE	0.55 (0.53 - 0.60)	0.58 (0.55 - 0.60)	0.60 (0.58 - 0.70)	0.69 (0.40 - 0.74)
O2U - NCE-RCE	0.51 (0.46 - 0.58)	0.56 (0.50 - 0.61)	0.68 (0.67 - 0.71)	0.64 (0.63 - 0.71)
O2U - CE	0.58 (0.50 - 0.59)	0.56 (0.54 - 0.59)	0.67 (0.60 - 0.72)	0.68 (0.63 - 0.71)
LRT	0.56 (0.47 - 0.62)	0.58 (0.56 - 0.62)	0.65 (0.57 - 0.70)	**0.70** (0.65 - 0.70)

Table 2: Median F1-score ([min-max] over 5 iterations) comparison between methods using different training datasets. These are: i. PR-RS: Pathology Reports and Random Selection; ii. PR-PS: Pathology Reports and Proximity Selection; iii. PR-SP-RS: Pathology Reports, Surgery Procedures and Random Selection; iv. PR-SP-PS: Pathology Reports, Surgery Procedures and Proximity Selection.

Class	Precision	Recall	Support
Biopsy	0.91	0.62	34
Other	0.96	0.99	280

Table 3: Performance of the programmatic annotation for the PR-RS dataset, evaluated on development set.

	w/ biopsy		w/o biopsy	
Method	**Prec.**	**Rec.**	**Prec.**	**Rec.**
CE	0.38	0.90	0.27	0.53
NCE-RCE	0.61	0.90	0.59	0.69
O2U - NCE-RCE	0.54	0.93	0.57	0.69
O2U - CE	0.56	0.93	0.60	0.69
LRT	0.64	0.90	0.59	0.69
EL dataset - CE	0.69	0.93	0.62	0.75

Table 4: Median precision and recall (over 5 iterations) for patients with and without a biopsy done at the hospital, evaluated on test set. Model trained on PR-SP-PS and EL datasets.

these two events could explain the gain in performance (3 class model vs 2 class model).

Nevertheless, some drawbacks are found. First, all models, including the expert labeled one, underperform when evaluated on patients without biopsy procedures done at the hospital. Results presented in Table 4 evidence this, major differences were found on the recall of biopsy dates mentions and not on the precision, suggesting that clinicians refer differently in text according to the origin of the medical procedure. This result should be studied in a qualitative form. Second, our method helps to minimize the necessary annotation effort in a context when experts' available time is scarce, such as healthcare; however, we reduce expert annotation effort in exchange for an industry knowledge investment. Especially, the selection of the appropriate structured sources or terminologies could potentially not be straightforward. However, when a new project starts, annotating a small development set can be a good opportunity to better understand the task's complexity and its perimeter.

Finally, a novelty of our method is the absence of any NER procedure for the concepts to identify, this allows us to learn complex patterns from text. We find in our application case that biopsy acts are often non mentioned explicitly, but an expert could infer it from the subtext. Additionally, this makes our method language- and vocabulary-independent[8], and therefore directly applicable to other CDWs or to retrieve other important dates. The possibility to replicate the method in other healthcare institutions is a non-negligible strength, as the patient privacy regulations prevent researchers to share datasets and deep learning models. This distant supervision approach allows us to share a strategy to train a deep learning model without having to annotate a training dataset.

7. Conclusion

We successfully show that it is possible to leverage information redundancy of an organizational data lake to build a programmatically annotated corpus and train ML models, minimizing the required expert time for the annotation task. We develop a domain agnostic approach, particularly interesting in settings with scarce experts' available time, huge amounts of data collected and industry knowledge, such as healthcare. The development of efficient methods of information extraction from unstructured data for further use is, therefore, essential. A direct implication of the presented application case study is the precomputation of structured variables retrieving information from text in order to accelerate downstream applications as patient recruitment for clinical trials, treatment effectiveness studies, survival analysis or epidemiology studies. Also, the method allows the finetuning of deep learning models without sharing weights or datasets. Other techniques of noise management and semi-supervised approaches, as well as a better bias understanding, are still to be explored.

[8] except from the dates entity recognition task implemented in French.

8. Ethics Statement

The methods were performed in accordance with relevant guidelines and regulations and approved by the institutional review board of the institution.[9] Patients were informed and those who opposed the secondary use of their data for research were excluded from the study.

9. Acknowledgements

9

10. Funding

This study received funding from XXX[9]. The funder had no role in the analysis and interpretation of data.

11. Data availability

Access to the clinical data warehouse's de-identified raw data can be granted following the process described on its website:[9] A prior validation of the access by the local institutional review board is required.

[9] Anonymized for review. This information will be provided in the final version.

12. Bibliographical References

Vibhu Agarwal, Tanya Podchiyska, Juan M Banda, Veena Goel, Tiffany I Leung, Evan P Minty, Timothy E Sweeney, Elsie Gyang, and Nigam H Shah. 2016. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, 23(6):1166–1173.

Saadullah Amin, Katherine Ann Dunfield, Anna Vechkaeva, and Guenter Neumann. 2020. A Data-driven Approach for Noise Reduction in Distantly Supervised Biomedical Relation Extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 187–194, Online. Association for Computational Linguistics.

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature review for Language and Statistics II*, 2:1–15.

Priyankar Bose, Sriram Srinivasan, William C. Sleeman, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021. A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. *Applied Sciences*, 11(18):8319.

Federico Cabitza, Gunnar Ellingsen, Angela Locoro, and Carla Simone. 2019. Repetita Iuvant: Exploring and Supporting Redundancy in Hospital Practices. *Computer Supported Cooperative Work (CSCW)*, 28(1-2):61–94.

Federico Cabitza, Marcello Sarini, Carla Simone, and Michele Telaro. 2005. When once is not enough: the role of redundancy in a hospital ward setting. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work - GROUP '05*, page 158, Sanibel Island, Florida, USA. ACM Press.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. pages 2633–2650.

Jared A. Dunnmon, Alexander J. Ratner, Khaled Saab, Nishith Khandwala, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew P. Lungren, Daniel L. Rubin, and Christopher Ré. 2020. Cross-Modal Data Programming Enables Rapid Medical Machine Learning. *Patterns*, 1(2):100019.

Basile Dura, Charline Jean, Xavier Tannier, Alice Calliger, Romain Bey, Antoine Neuraz, and Rémi Flicoteaux. 2022. Learning structures of the French clinical language:development and validation of word embedding models using 21 million clinical reports from electronic health records. ArXiv:2207.12940 [cs, stat].

Basile Dura, Perceval Wajsburt, Thomas Petit-Jean, Ariel Cohen, Charline Jean, and Romain Bey. 2022b. EDS-NLP: efficient information extraction from French clinical notes.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29.

Jason A. Fries, Ethan Steinberg, Saelig Khattar, Scott L. Fleming, Jose Posada, Alison Callahan, and Nigam H. Shah. 2021. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communications*, 12(1):2017.

Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J. Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen, Yiqing Zhao, Sunghwan Sohn, and Hongfang Liu. 2020. Clinical concept extraction: A methodology review. *Journal of Biomedical Informatics*, 109:103526.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust Loss Functions under Label Noise for Deep Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Udo Hahn and Michel Oleynik. 2020. Medical Information Extraction in the Age of Deep Learning. *Yearbook of Medical Informatics*, 29(01):208–220.

Yoni Halpern, Steven Horng, Youngduck Choi, and David Sontag. 2016. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4):731–740.

David A. Hanauer, Qiaozhu Mei, James Law, Ritu Khanna, and Kai Zheng. 2015. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *Journal of Biomedical Informatics*, 55:290–300.

Michael A. Hedderich, Dawei Zhu, and Dietrich Klakow. 2021. Analysing the Noise Model Error for Realistic Noisy Label Data. *Proceedings of*

the AAAI Conference on Artificial Intelligence, 35(9):7675–7684.

Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. 2019. O2U-Net: A Simple Noisy Label Detection Approach for Deep Neural Networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3325–3333, Seoul, Korea (South). IEEE.

Ellen Kim, Samuel M. Rubinstein, Kevin T. Nead, Andrzej P. Wojcieszynski, Peter E. Gabriel, and Jeremy L. Warner. 2019. The Evolving Use of Electronic Health Records (EHR) for Research. *Seminars in Radiation Oncology*, 29(4):354–361.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. 2021. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? *arXiv:2104.07762 [cs]*. ArXiv: 2104.07762.

Ivan Lerner, Nicolas Paris, and Xavier Tannier. 2020. Terminologies augmented recurrent neural network model for clinical named entity recognition. *Journal of Biomedical Informatics*, 102:103356.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Xinbo Lv, Yi Guan, Jinfeng Yang, and Jiawei Wu. 2016. Clinical Relation Extraction with Deep Learning. *International Journal of Hybrid Information Technology*, 9(7):237–248.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. 2020. Normalized Loss Functions for Deep Learning with Noisy Labels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6543–6553. PMLR.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219. ArXiv:1911.03894 [cs].

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, USA. Association for Computational Linguistics. Event-place: Suntec, Singapore.

Craig D. Newgard, Dana Zive, Jonathan Jui, Cody Weathers, and Mohamud Daya. 2012. Electronic Versus Manual Data Processing: Evaluating the Use of Electronic Health Records in Out-of-hospital Clinical Research: ELECTRONIC VS. MANUAL DATA PROCESSING. *Academic Emergency Medicine*, 19(2):217–227.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29(2-3):709–730.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data Programming: Creating Large Training Sets, Quickly. Publisher: arXiv Version Number: 3.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–19.

Sahaana Suri, Raghuveer Chanda, Neslihan Bulut, Pradyumna Narayana, Yemao Zeng, Peter Bailis, Sugato Basu, Girija Narlikar, Christopher Ré, and Abishek Sethi. 2020. Leveraging organizational resources to adapt models to new data modalities. *Proceedings of the VLDB Endowment*, 13(12):3396–3410.

Tianyin Wang, Jianwei Wang, and Ziqian Zeng. 2022. Learning with Silver Standard Data for Zero-shot Relation Extraction. ArXiv:2211.13883 [cs].

Yanshan Wang, Sunghwan Sohn, Sijia Liu, Feichen Shen, Liwei Wang, Elizabeth J. Atkinson, Shreyasee Amin, and Hongfang Liu. 2019. A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making*, 19(1):1.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and

Cheryl Clark. 2014. Negation's Not Solved: Generalizability Versus Optimizability in Clinical Natural Language Processing. *PLoS ONE*, 9(11):e112774.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, and Hua Xu. 2020. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. ArXiv:1611.03530 [cs].

Xiaohui Zhang, Yaoyun Zhang, Qin Zhang, Yuankai Ren, Tinglin Qiu, Jianhui Ma, and Qiang Sun. 2019. Extracting comprehensive clinical information for breast cancer using deep learning methods. *International Journal of Medical Informatics*, 132:103985.

Xinyu Zhao, Shih-ting Lin, and Greg Durrett. 2020. Effective Distant Supervision for Temporal Relation Extraction. Publisher: arXiv Version Number: 2.

Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen. 2020. Error-Bounded Correction of Noisy Labels. ArXiv:2011.10077 [cs].