Leveraging Information Redundancy of Real-World Data Through Distant Supervision

Supplementary Material

1 Cohort Selection

We consider patients with at least one ICD-10[1] diagnostic code (principal or related) of cancer definied in eTable 1 and coded between between January 1st, 2017 and August 8th, 2023. Patients with multiple cancer types are excluded.

Cancer type	ICD-10 code
Anus	C21, D013
Biliary Duct	C221, C23, C24, D015, D376
Bowel	C17, D014, D372
Breast	C50, D05, D486
Colon	C18, C19, D010, D011, D374, D373, C20, D012, D375
Head & Neck	C0*, C10, C11, C12, C13, C14, C30, C31, C32, D000, D020, D370, D380
Kidney	C64, C65, D410, D411
Lung	C33, C34, D021, D022
Oesophagus	C15, D001
Other Gynecology	C51, C52, C57, C58, D071, D072, D073, D392, D397, D399
Other Pneumology	C37, C38, C39, D023, D024, D382D383, D384, D385, D386
Pancreas	C250, C251, C252, C253, C255, C256, C257, C258, C259
Prostate	C61, D075, D400
Thyroid	C73, D440

eTable 1: ICD-10 for cohort selection

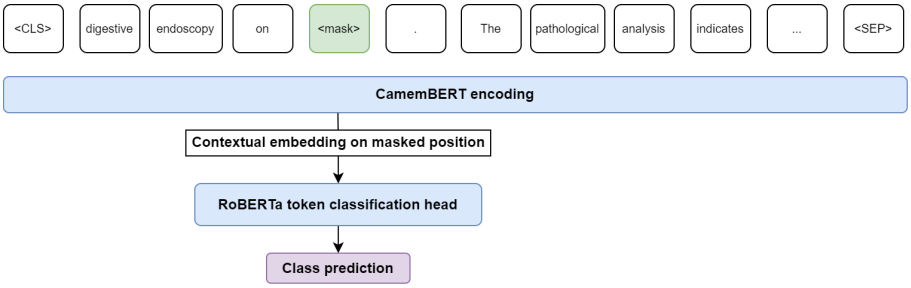
[1] International Classification of Diseases,Tenth Revision

2 Modelling & Training

To classify date spans, we add a RoBERTa token classification head that takes as input the contextual embedding at the masked position (eFigure 1). For all experiments, we use a pre-trained CamemBERT language model trained on French clinical reports. Nevertheless, it's possible to utilize a different contextual encoder.

For training, we use the AdamW optimizer with the following parameters: Transformer Encoder learning rate = 5e-5 (with a linear warmup of 10% of steps), RoBERTa token classification head learning rate = 5e-4. Both learning rates are decreased with a linear scheduler. All rest of hyper-parameters are set to default ones.

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 evaluate the different hyper-parameter combinations, results are shown in eTable 2.



eFigure 1: Model schema

Method	Hyper-parameter	Mean F1	N experiments
NCE RCE	$\alpha = 0.1$, $\beta = 1$	**0.85**	3
NCE RCE	$\alpha = 1$, $\beta = 1$	0.80	3
NCE RCE	$\alpha = 1$, $\beta = 0.1$	0.77	3
NCE RCE	$\alpha = 0, 1$, $\beta = 10$	0.79	3
O2U	fr=30% ; NCE-RCE best	**0.89**	3
O2U	fr=40% ; NCE-RCE best	0.86	3
O2U	fr=50% ; NCE-RCE best	0.82	3
O2U	fr=60% ; NCE-RCE best	0.83	3
O2U	fr=70% ; NCE-RCE best	0.83	3
O2U	fr=30% ; CE	**0.89**	3
O2U	fr=40% ; CE	0.83	3
O2U	fr=50% ; CE	0.83	3
O2U	fr=60% ; CE	0.86	3
O2U	fr=70% ; CE	0.86	3
LRT	$\delta = 1.3$; NCERCE best	**0.80**	5
LRT	$\delta = 1.4$; NCERCE best	0.79	5
LRT	$\delta = 1.5$; NCERCE best	0.79	5

eTable 2: Mean F1-score for different hyper-parameter combinations. The forget rate hyper-parameter of the O2U method is indicated as *fr*.