Supplementary Material: Transfer Learning from Medical Literature for Section Prediction in Electronic Health Records

Anonymous ACL submission

We include as supplementary material a more detailed breakdown of the data in section A and results in Section B. In Section C we show the results of using Wikipedia, the dataset we will publicly release, as the only medical literature source.

A Data

Our datasets consisted of sections from the 11 common headers found in Medical Literature (MedLit) and Electronic Health Records (EHRs):

The 11 classes are:

Allergies (Alrg)
Assessment and Plan (A&P)
Chief Complaint (CC)
Examination (Exam)
Family History (FHX)
Diagnostic Findings (Find)
Medications (Meds)
Past Medical History (PMH)
Personal and Social History (PSH)
Procedures (Proc)
Review of Systems (ROS)

As described in the main paper, we created a short list of common phrases for each of the 11 section classes by looking at labels used in prior work as well as section headers labeled in our annotated EHR datasets. The list of headers that we searched for are shown in Table 1. We ultimately used exact matches with stemming to extract passages with higher precision for the common categories. In the less frequent categories (e.g. *Allergies, Family History, Personal and Social History*), we allowed additional words to be present in the title. Case was ignored for all phrases except for acronyms (e.g. CC, SUBJ) which had to be in all caps). Although we explored each separately, we ultimately combined laboratory tests with findings to create the Diagnostic Findings class and History of Present Illness was merged with Chief Complaints. Due to the very small size of Health Maintenance we decided to exclude it. The EHRs and MedLit that did not match to one of these classes were excluded from our experiments.

We provide a more detailed breakdown by class of the sizes of our Medical Literature (MedLit), i2b2, and Cleveland Clinic (ClvC) data sets in Table 3 and 4. The statistics are broken down for our 11 classes listed as well as the total (as found in the main paper). Please refer to the main paper for more details about the sections and a summary of the tables. Additional examples from the WikipediaMedical corpus are shown in Table 2 with their associated labels.

B Results

We include full sentence results in Tables 5 (ClvC RNN), 6 (ClvC BERT), 7 (i2b2 RNN), and 8. Full section results are in in Tables 11 (ClvC RNN), 12 (ClvC BERT), 9 (i2b2 RNN), and 10. Finally, we also include a breakdown of the results on the MedLit dev set in the top two rows of Table 13. We provide a full breakdown by showing the F1-score for each of the 11 classes (See A). We also show the Micro and Macro average F1 score.

Please refer to the main paper for a summary of the tables, description, and analysis of the results.

C Wikipedia Results

Several of the data sources we use in the the main paper are licensed and cannot be released publicly. Therefore, in order to provide a fair comparison for future work we show the results on training with *just* Wikipedia Medical as our only medical literature data. Wikipedia Medical was extracted from the full Wikipedia corpus by considering all sub categories of Clinical Medicine (https://en.wikipedia.org/

Class	Header Phrases
Allergies	Allergies, Allergies and adverse reactions, Allergic, Allergen
Assessment and Plan	Impression, Assessment and Plan, Analysis assessment and plan, A/P, AP, Assessment, DISCHARGE INSTRUCTIONS, Follow up
Chief Complaint	Chief complaint, Reason for Admission, CC, signs and symptoms, SUBJ
Examination	Examination, Physical examination, PE, Physical Exam on Discharge, Physical, Physical Examinations, Physical Exam, Vital signs, Vital sign, VS, General examination, Dermatologic examination, Lymph nodes/heme examination, HEENT examination, Cardiovascular examination, Gastrointestinal examination, Chest examination, Genitourinary examination, Neurological examination, Psychological examination, Musculoskeletal examination, Extremity examination, Exam, EXAM DETAILS, Comparison, Contrast, OBJ, inspection, palpation, auscultation, percussion
Family History	Family History, Family medical history, FHX, Family Histories, Parent medical history, Sibling medical history, Child medical history
Findings	Findings, Procedural Findings, Indications, DIAGFIND, Diagnostic Findings,Laboratory imaging and pathology results, Laboratory results, imaging results, pathology results, Radiology Reports
Health Maintenance	Health Maintenance, health care maintenance,
History of Present Illness	Present Illness, History of Present Illness, HPI, History of Current Illness, History of Illness
Laboratory Tests	Laboratory Tests, LABS, Laboratory Examinations, Intraoperative Pathology Exam, Microscopic Exam, Pathologic Specimens
Medications	Medications, medication, MEDS, MED, Discharge Medications, Prescription History
Past Medical History	PROB, Medical History, Past medical history, PMH, PSH, past medical history and past surgical history, P/P, CONDITION BEFORE ADMISSION, Clinical History, Gynecological History
Personal and Social history	Personal and Social History, Personal Histories, Social History, SHX, SOC, Substance use history
Procedures	past surgical history, Procedures, PROC, Procedure Descriptions, Procedures Performed, HOSPITAL COURSE, Course of Care, Anesthesia, Diagnostic Studies, studies, study, interventions and practices considered
Review of Systems	Review of systems, systems enquiry, systems review, ROS, Constitutional symptoms, Eyes, Ears, nose, mouth, and throat, ENT,Cardiovascular, Respiratory, Gastrointestinal, Genitourinary, Musculoskeletal, Integumentary , Breast, Neurological, Psychiatric, Endocrine, Hematologic, lymphatic, immunologic

Table 1: The complete list of phrases that we used to match headers found in the Medical Literature

wiki/Category:Clinical_medicine)) up to depth 10 with a manual selection of the subcategories to exclude categories that did not contain true medical content. Such categories can appear due to cross references among Wikipedia categories. We searched this corpus of approximately 30,000 documents for all of the common EHR phrases found in Table 1 which left us with 2,658 sections. Each data instance consists of the section's passage including the header, and a class label which was assigned based on its header. Wikipedia Medical (WikipediaMed) makes up approximately 30% of our corpus.

A breakdown of the results for the MedLit dev set for Wikipedia is shown in the bottom two rows of Table 13. Note, that the dev set here consists of only the sections from Wikipedia. The Micro results are better than the entire MedLit set indicating that a single source is probably more consistent in style. However, the Macro average score is lower because some classes don't exist at all in the Wikipedia dev set.

The results on the WikipediaMed corpus are shown in Tables 14 and 15 with the results using the full Medical Literature corpus (All MedLit) from the main paper for comparison. It is clear that when training *only* on WikipediaMed the results are significantly worse than All MedLit. However, when fine-tuning is applied, the gap tends to shrink. When fine-tuning on i2b2 training data the results are very similar for both RNN and BERT and in some cases slightly higher. For example, the section and sentence level RNN results for testing on ClvC with WikipediaMed + TR i2b2 vs All Medlit + TR i2b2 have a few point difference. When fine-tuning on ClvC training data on the section level, the RNN model performance using all the data is much greater than using just Wikipedia. For example, the results of All MedLit + TR ClvC vs Wikipedia + TR ClvC is 20 points. The difference when using BERT is much smaller (at most 5 points, and sometimes no difference). We expect the gap is larger for ClvC due to its smaller size. Finally, although there are difference between the WikipediaMed and All Medlit results, our findings in the main paper still

Label	Document Title	Header	Passage (Abbreviated)
CC	hypoxemia	signs and symptoms	in an acute context hypoxemia can cause symptoms such as those in respiratory distress . these include breathlessness an increased rate of breathing use of the chest and abdominal muscles to breathe and lip pursing . chronic hypoxemia may be compensated or uncompensated . the compensation may cause symptoms to be overlooked initially however further disease or a stress such as any increase in oxygen demand may finally unmask the existing hypoxemia . in a compensated state blood vessels supplying less ventilated areas of the lung may selectively contract to redirect the blood to areas of the lungs which are better ventilated . however in a chronic context and if the lungs are not well ventilated generally this mechanism can result in pulmonary hypertension ...
Find	bariatric surgery	indications	biliopancreatic diversion . a medical guideline by the american college of physicians concluded . surgery should be considered as a treatment option for patients with a bmi of 40kg m or greater who instituted but failed an adequate exercise and diet program with or without adjunctive drug therapy and who present with obesity related comorbid conditions such as hypertension impaired glucose tolerance diabetes mellitus hyperlipidemia and obstructive sleep apnea . a doctor patient discussion of surgical options should include the long term side effects such as possible need for reoperation gallbladder disease and malabsorption ...
ROS	angiotensin	cardiovascular	they are potent direct vasoconstrictors constricting arteries and veins and increasing blood pressure . this effect is achieved through activation of the gpcr at1 which signals through a gq protein to activate phospholipase c and subsequently increase intracellular calcium . angiotensin ii has prothrombotic potential through adhesion and aggregation of platelets and stimulation of pai 1 and pai 2 . when cardiac cell growth is stimulated a local autocrine paracrine renin angiotensin ...
Meds	gastroesophageal reflux disease	medications	the primary medications used for gerd are proton pump inhibitors h2 receptor blockers and antacids with or without alginic acid . proton pump inhibitors ppis such as omeprazole are the most effective followed by h2 receptor blockers such as ranitidine . if a once daily ppi is only partially effective they may be used twice a day . they should be taken one half to one hour before a meal . there is no significant difference between agents in this class . when these medications are used long term the lowest effective dose should be taken . they may also be taken only when symptoms occur in those with frequent problems . h2 receptor blockers lead to roughly a 40 improvement .

Table 2: Examples of passages and their associated titles, headers, and label which was assigned based on the header. (See Table 1

remain: Combining a very small amount of in-domain EHR data with a large amount of automatically labeled, out-of-domain, out-of-genre Medical Literature data can perform as well as using a large amount of in-domain EHR data and combining out-of-domain, out-of-genre Medical Literature data with out-of-domain EHRs can provide significant improvement over using just out-of-domain EHRs depending on training data size.

	Source	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Total
Train	MedLit	148	1967	1759	886	189	791	388	163	145	335	271	7042
	i2b2	146	227	215	221	106	119	195	160	137	33	121	1680
	ClvC	17	39	42	39	19	28	26	23	18	18	25	294
Dev	MedLit	14	19	394	329	89	65	76	73	88	29	28	1204
	i2b2	140	216	210	209	99	105	176	153	132	35	116	1591
	ClvC	12	72	77	68	18	30	28	31	24	21	23	404
Test	i2b2	159	210	205	222	97	95	194	160	139	39	135	1655
	ClvC	14	38	38	31	7	15	17	12	10	12	10	204

Table 3: The number of sections per class in each of the three datasets for train, dev, and test sets.

	Source	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Total
Train	MedLit	1214	23724	18391	11235	1031	5010	4038	1125	878	3326	2219	72191
	i2b2	255	2824	2224	1874	167	527	605	614	401	102	500	10093
	ClvC	35	570	529	794	65	609	245	176	78	75	291	3467
Dev	MedLit	143	167	4408	4702	512	584	942	497	485	401	218	13059
	i2b2	221	2370	2171	1783	143	433	703	635	418	105	391	9373
	ClvC	28	666	680	804	74	327	178	122	124	65	214	3282
Test	i2b2	487	4776	4161	3973	473	868	1101	1413	792	233	833	19110
	ClvC	42	941	685	815	56	647	264	113	154	57	272	4046

Table 4: The number of sentences per class in each of the three datasets for train, dev, and test sets.

	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Micro	Macro
	0.67	0.62	0.45	0.66	0.81	0.66	0.78	0.43	0.60	0.64	0.56	0.61	0.63
MedLit	0.54	0.18	0.32	0.45	0.48	0.16	0.38	0.29	0.40	0.02	0.14	0.30	0.31
MedLit+TR ClvC	0.83	0.56	0.52	0.70	0.70	0.69	0.80	0.43	0.61	0.60	0.60	0.62	0.64
i2b2	0.74	0.58	0.49	0.67	0.67	0.37	0.71	0.36	0.58	0.67	0.58	0.56	0.58
MedLit+TR i2b2	0.70	0.55	0.53	0.68	0.71	0.11	0.66	0.30	0.60	0.56	0.53	0.53	0.54

Table 5: Sentence-Level results for training on ClvC with the RNN model. The last two columns are the Micro Avg F1 and Macro Avg F1.

	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Micro	Macro
CC	0.82	0.71	0.61	0.78	0.91	0.81	0.83	0.61	0.69	0.81	0.67	0.73	0.69
MedLit	0.80	0.11	0.37	0.56	0.68	0.05	0.51	0.26	0.59	0.03	0.13	0.36	0.34
MedLit + TR ClvC	0.77	0.73	0.61	0.82	0.92	0.79	0.84	0.57	0.71	0.74	0.72	0.74	0.68
i2b2	0.73	0.67	0.56	0.76	0.93	0.33	0.64	0.41	0.75	0.54	0.66	0.62	0.58
MedLit + TR i2b2	0.80	0.69	0.52	0.74	0.91	0.18	0.74	0.31	0.65	0.36	0.64	0.60	0.55

Table 6: Sentence-Level results for training on ClvC with the BERT model.

	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Micro	Macro
i2b2	0.68	0.65	0.63	0.78	0.54	0.42	0.70	0.40	0.62	0.25	0.51	0.64	0.56
medlit	0.45	0.25	0.34	0.48	0.25	0.13	0.28	0.05	0.10	0.08	0.08	0.30	0.23
MedLit+TR i2b2	0.69	0.64	0.64	0.80	0.56	0.48	0.71	0.38	0.63	0.36	0.55	0.65	0.58
CC	0.51	0.50	0.47	0.69	0.25	0.12	0.46	0.13	0.17	0.13	0.42	0.48	0.35
MedLit+TR ClvC	0.61	0.52	0.53	0.70	0.38	0.14	0.49	0.16	0.23	0.25	0.48	0.52	0.41

Table 7: Sentence-Level results for training on i2b2 with the RNN model. The last two columns are the Micro Avg F1 and Macro Avg F1.

	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Micro	Macro
i2b2	0.68	0.74	0.70	0.83	0.66	0.50	0.75	0.52	0.77	0.38	0.60	0.71	0.59
MedLit	0.70	0.28	0.45	0.63	0.54	0.13	0.46	0.20	0.56	0.09	0.09	0.41	0.34
MedLit+TR i2b2	0.63	0.73	0.69	0.83	0.70	0.52	0.74	0.52	0.79	0.41	0.60	0.71	0.60
CC	0.70	0.61	0.57	0.76	0.47	0.21	0.59	0.26	0.45	0.29	0.53	0.59	0.45
MedLit+TR ClvC	0.71	0.61	0.59	0.78	0.56	0.28	0.66	0.26	0.44	0.32	0.53	0.60	0.48

Table 8: Sentence-Level results for training on i2b2 with the BERT model. The last two columns are the Micro Avg F1 and Macro Avg F1.

	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Micro	Macro
CC	0.80	0.41	0.42	0.66	0.87	0.39	0.71	0.79	0.72	0.86	0.49	0.59	0.65
i2b2	0.62	0.59	0.61	0.83	0.93	0.42	0.81	0.75	0.74	0.83	0.81	0.71	0.72
MedLit	0.59	0.33	0.30	0.52	0.95	0.29	0.38	0.55	0.87	0.07	0.35	0.47	0.47
MedLit+TR ClvC	0.78	0.70	0.73	0.77	0.85	0.39	0.94	0.93	0.92	0.89	0.75	0.77	0.79
MedLit+TR i2b2	0.75	0.63	0.63	0.68	0.88	0.30	0.85	0.52	0.83	0.15	0.78	0.65	0.64

Table 9: Section-Level results for the i2b2 training set using the RNN model. The last two columns are the Micro Avg F1 and Macro Avg F1.

	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Micro	Macro
ClvC	0.96	0.90	0.82	0.90	0.98	0.73	0.94	0.93	0.94	0.91	0.91	0.89	0.83
i2b2	0.84	0.80	0.84	0.76	0.95	0.61	0.95	0.81	0.90	0.79	0.88	0.83	0.76
MedLit	0.93	0.44	0.20	0.39	0.95	0.27	0.82	0.59	0.89	0.16	0.83	0.55	0.54
Medlit ClvC	0.95	0.89	0.86	0.91	0.98	0.67	0.96	0.86	0.94	0.91	0.93	0.90	0.82
MedLit+TR i2b2	0.81	0.73	0.73	0.72	0.95	0.47	0.93	0.79	0.82	0.90	0.85	0.78	0.72

Table 10: Section-Level results for the i2b2 training set using the BERT model. The last two columns are the Micro Avg F1 and Macro Avg F1.

	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Micro	Macro
ClvC	0.90	0.30	0.40	0.79	0.61	0.25	0.71	0.61	0.14	0.34	0.50	0.53	0.51
i2b2	0.99	0.91	0.98	0.99	0.92	0.97	0.95	0.88	0.93	0.93	1.00	0.95	0.95
MedLit	0.60	0.56	0.44	0.73	0.83	0.81	0.87	0.63	0.87	0.58	0.57	0.68	0.68
Medlit+TR ClvC	0.90	0.76	0.74	0.76	0.83	0.76	0.91	0.68	0.86	0.97	0.93	0.81	0.83
MedLit+TR i2b2	0.98	0.88	0.95	0.99	0.93	0.98	0.95	0.78	0.94	0.62	0.99	0.93	0.91

Table 11: Section-Level results for training on ClVC with the RNN model. The last two columns are the Micro Avg F1 and Macro Avg F1.

	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Micro	Macro
ClvC	0.96	0.85	0.72	0.95	0.83	0.75	0.96	0.80	0.65	0.62	0.93	0.84	0.75
i2b2	1.00	1.00	0.98	1.00	1.00	0.99	1.00	0.98	0.99	0.97	1.00	0.99	0.91
MedLit	0.97	0.63	0.57	0.74	0.83	0.90	0.94	0.72	0.84	0.45	0.65	0.76	0.69
Medlit+TR ClvC	0.97	0.96	0.86	0.96	0.82	0.94	0.97	0.84	0.83	0.82	0.96	0.92	0.83
MedLit+TR i2b2	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.99	0.91

Table 12: Section-Level results for training on ClVC with the BERT model. The last two columns are the Micro Avg F1 and Macro Avg F1.

	Alrg	A&P	CC	Exam	FHX	Find	Meds	PMH	PSH	Proc	ROS	Micro	Macro
RNN	0.65	0.14	0.68	0.75	0.65	0.40	0.59	0.55	0.47	0.53	0.29	0.65	0.52
BERT 1 ep	0.63	0.21	0.73	0.79	0.75	0.49	0.65	0.59	0.63	0.62	0.38	0.71	0.55
RNN	0.77	0.22	0.84	0.31	-	0.45	0.58	-	-	0.55	0.18	0.69	0.43
BERT 1 ep	0.72	0.60	0.88	0.42	-	0.53	0.64	-	0.31	0.63	0.33	0.74	0.42

Table 13: Sentence-Level results for the MedLit development set in F1-score for All Medical Literature (top rows) and Wikipedia Only (bottom rows). The last two columns are the Micro Avg F1 and Macro Avg F1.

Experiment	RNN F1	BERT F1
ClvC	0.59	0.89
All MedLit	0.47	0.55
WikipediaMed	0.25	0.53
All Medlit + TR ClvC	0.77	0.90
WikipediaMed + TR ClvC	0.57	**0.92**
i2b2	**0.71**	0.83
All Medlit + TR i2b2	0.65	0.78
WikipediaMed + TR i2b2	0.69	0.78

(a) Testing on ClvC

Experiment	RNN F1	BERT F1
i2b2	**0.95**	**0.99**
All MedLit	0.68	0.76
WikipediaMed	0.27	0.73
All Medlit + TR i2b2	0.93	**0.99**
WikipediaMed + TR i2b2	0.92	**0.99**
ClvC	0.53	0.84
All Medlit + TR ClvC	0.81	0.92
WikipediaMed + TR ClvC	0.49	0.87

(b) Testing on i2b2

Table 14: Section-Level results for testing on the (a) ClvC and (b) i2b2 EHRs with all MedLit and Wikipedia. F1-scores are micro averages across all classes. The best results in each column are highlighted in bold. The results are grouped by MedLit/Wikipedia, in-domain ((a) ClvC, (b) i2b2), and out-of-domain ((a) i2b2, (b) ClvC) training.

Experiment	RNN F1	BERT F1
CC	0.61	0.73
All MedLit	0.30	0.37
WikipediaMed	0.21	0.27
All MedLit + TR CC	**0.62**	**0.74**
WikipediaMed + TR ClvC	0.60	0.72
i2b2	0.56	0.62
All MedLit + TR i2b2	0.53	0.60
WikipediaMed + TR i2b2	0.55	0.60

(a) Testing on ClvC

Experiment	RNN F1	BERT F1
i2b2	0.64	**0.71**
All MedLit	0.30	0.41
WikipediaMed	0.13	0.26
All MedLit + TR i2b2	0.65	**0.71**
WikipediaMed + TR i2b2	**0.66**	**0.71**
CC	0.48	0.59
All MedLit + TR CC	0.52	0.60
WikipediaMed + TR ClvC	0.48	0.59

(b) Testing on i2b2

Table 15: Sentence-Level results for testing on the (a) ClvC and (b) i2b2 EHRs with all MedLit and Wikipedia. F1-scores are micro averages across all classes. The best results in each column are highlighted in bold. The results are grouped by MedLit/Wikipedia, in-domain ((a) ClvC, (b) i2b2), and out-of-domain ((a) i2b2, (b) ClvC) training.