Assessing the role of a medication-indication resource in the treatment relation extraction from clinical text

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.   **OXFORD** UNIVERSITY PRESS

Cosmin Adrian Bejan[1], Wei-Qi Wei[1], Joshua C Denny[1,2]

RESEARCH AND APPLICATIONS

ABSTRACT

Objective To evaluate the contribution of the MEDication Indication (MEDI) resource and SemRep for identifying treatment relations in clinical text.

Materials and methods We first processed clinical documents with SemRep to extract the Unified Medical Language System (UMLS) concepts and the treatment relations between them. Then, we incorporated MEDI into a simple algorithm that identifies treatment relations between two concepts if they match a medication-indication pair in this resource. For a better coverage, we expanded MEDI using ontology relationships from RxNorm and UMLS Metathesaurus. We also developed two ensemble methods, which combined the predictions of SemRep and the MEDI algorithm. We evaluated our selected methods on two datasets, a Vanderbilt corpus of 6864 discharge summaries and the 2010 Informatics for Integrating Biology and the Bedside (i2b2)/Veteran's Affairs (VA) challenge dataset.

Results The Vanderbilt dataset included 958 manually annotated treatment relations. A double annotation was performed on 25% of relations with high agreement (Cohen's $\kappa = 0.86$). The evaluation consisted of comparing the manual annotated relations with the relations identified by SemRep, the MEDI algorithm, and the two ensemble methods. On the first dataset, the best F1-measure results achieved by the MEDI algorithm and the union of the two resources (78.7 and 80, respectively) were significantly higher than the SemRep results (72.3). On the second dataset, the MEDI algorithm achieved better precision and significantly lower recall values than the best system in the i2b2 challenge. The two systems obtained comparable F1-measure values on the subset of i2b2 relations with both arguments in MEDI.

Conclusions Both SemRep and MEDI can be used to extract treatment relations from clinical text. Knowledge-based extraction with MEDI outperformed use of SemRep alone, but superior performance was achieved by integrating both systems. The integration of knowledge-based resources such as MEDI into information extraction systems such as SemRep and the i2b2 relation extractors may improve treatment relation extraction from clinical text.

INTRODUCTION

Recognition of treatment relations between medical concepts described in clinical documentation is a challenging task for current natural language processing (NLP) applications. Accurately extracting such relations could improve both clinical and research tasks, such as enabling a more comprehensive understanding on a patient's treatment course,[1,2] improving adverse reaction detection,[3] and assessing healthcare quality.[4,5] The purpose of this study was to evaluate several methods of extracting treatment relations from clinical notes.

Treatment relations (denoted as TREATS) between two medical concepts describe situations when one of the concepts is a treatment for the other. Such relations often link medications to

their indications, but other types of medical concepts can be arguments of treatment relations. For instance, in example (1) below, the relation TREATS (*carvedilol*, *hypertension*) specifies a medication(*carvedilol*) as a treatment for a disease(*hypertension*), whereas in (2), TREATS (*left total hip arthroplasty*, *arthritis*) corresponds to a procedure (*left total hip arthroplasty*) for a specific disease (*arthritis*). Both of these categories of treatment relations are of interest for this study. Notably, the context in which the medical concepts are described plays a key role in identifying treatment relations in text. As observed in (3), although a medication(*nitroglycerin*) was prescribed for a patient's medical problem (*chest pain*), the treatment did not cure or improve the medical condition of the patient. Therefore,

For numbered affiliations see end of article.

since the context indicates that the treatment of the patient's medical problem did not result in a positive outcome, the two concepts from example (3) do not form a treatment relation.

(1)	*His* [*hypertension*] *was controlled with* [*carvedilol*] *while in the hospital.*
(2)	*He underwent* [*left total hip arthroplasty*] *for* [*arthritis*] *on the date of admission.*
(3)	[*Nitroglycerin*] *drip was started, but* [*chest pain*] *did not improve.*

In this work, we evaluated two systems for identifying treatment relations in clinical documents. First, we evaluated the treatment relations extracted by SemRep,[6] a rule-based system used in biomedical information extraction. Then, we investigated the impact of a previously developed medication-indication resource (MEDI) on treatment relation extraction. Our hypothesis relied on the fact that providers often record the reasons (ie, the indications) for therapeutic interventions in their clinical notes; hence, many treatment relations encoded in text correspond to medication-indication pairs. To test this hypothesis, we implemented a simple algorithm based on MEDI,[7] a large database of medication-indication linkages generated by combining four sources of medication-indication information. Our ultimate goal was to create an automatic extraction system that not only will improve treatment relation extraction from clinical text, but also will expand MEDI with new medication-indication pairs.

BACKGROUND

The task of treatment relation extraction was recently studied as a part of the 2010 Informatics for Integrating Biology and the Bedside (i2b2)/Veteran's Affairs (VA) challenge.[8] However, the treatment relations for the i2b2 challenge were annotated for a more specific scope. Specifically, the classification criteria for such a relation requires the corresponding treatment for a patient's medical problem to meet one of the following conditions: (a) the treatment cures or improves the problem (TrIP); (b) the treatment worsens the problem (TrWP); (c) the treatment causes the problem (TrCP); (d) the treatment is administered for the problem (TrAP); or (e) the treatment is not administered because of the problem (TrNAP). All the other concept pairs occurring in the same sentence and not meeting one of the conditions mentioned above were not assigned a relationship. The best performing systems solving the 2010 i2b2 task on relation extraction used machine learning technologies.[9–12]

SemRep

SemRep was developed at the US National Library of Medicine (http://semrep.nlm.nih.gov) with the purpose of extracting semantic relations (eg, DIAGNOSES, CAUSES, LOCATION_OF, ISA, TREATS, PREVENTS) from biomedical literature. In a first step, this tool employs MetaMap[13] to extract the Unified Medical Language System (UMLS) concepts from a specified input text. Next, SemRep extracts the semantic relations that exist between the

UMLS concepts using linguistic and semantic rules specific to each relation. For example, rules of the form "X was controlled with Y" and "X for Y" could be employed to identify the treatment relations in (1) and (2), respectively. Additionally, SemRep uses inference rules to increase its coverage in identifying semantic relations. For treatment relation extraction, the current version of SemRep, V.1.5, includes the TREATS(INFER) and TREATS(SPEC) rules described in (4) and (5), respectively.[14,15]

| (4) | TREATS(X, Y) $\land$ OCCURS_IN(Z, X) $\rightarrow$ TREATS(INFER)(X, Z) |
| (5) | TREATS(Y, Z) $\land$ ISA(X, Y) $\rightarrow$ TREATS(SPEC)(X, Z) |

While SemRep was originally developed for processing text from the biomedical research literature,[16–18] a few studies have used it for processing clinical text. One such study reported improvements in SemRep performance of extracting treatment relations from Medline citations when using drug disorder co-occurrences computed from a large collection of clinical notes.[15] Another study focused on labeling concept associations from clinical text based on the semantic relations extracted by SemRep from Medline abstracts.[19] To the best of our knowledge, prior studies have not evaluated SemRep for treatment relation extraction from clinical text.

MEDI

MEDI[7] is also a publicly available resource (http://knowledge map.mc.vanderbilt.edu/research/) designed to capture both on-label and off-label (ie, absent in the Food and Drug Administration's approved drug labels) uses of medications. The information stored in MEDI was aggregated from four resources: RxNorm, Side Effect Resource (SIDER) 2,[20] MedlinePlus, and Wikipedia. Since MedlinePlus and Wikipedia encode the information linking medications to their corresponding indications in narrative text format, additional processing of these resources was performed. Development of MEDI included the use of KnowledgeMap Concept Indexer[21,22] and custom-developed section rules (eg, analyzing sections such as 'Why is this medication prescribed' and ignoring sections such as 'Precautions') to extract the mentions describing indications and to map them into the UMLS database.[7]

In MEDI, medications are associated with RxNorm concept unique identifiers (RxCUIs) and indications are represented by *International Classification of Diseases*, 9th edition (ICD9) codes, concept unique identifiers (CUI) in the UMLS Metathesaurus, and descriptions in narrative text. The current version of MEDI contains 3112 medications and 63 343 medication-indication pairs (termed "MEDI-ALL" in this study). Additionally, a more accurate subset of medication-indication pairs was selected from MEDI. In this subset, called MEDI high-precision subset (MEDI-HPS), an indication has to be either in RxNorm or to be contained in at least two of the three other resources. MEDI-HPS contains 2136 medications involved in 13 304 medication-indication pairs and has an estimated precision of 92. The paper that introduced MEDI[7] describes in detail the evaluation of the medication-indication pairs in this resource.

METHODS

Our study compared three methods to extract treatment relations from clinical text. The first was to use SemRep. The second was a simple rule based algorithm using MEDI. The third consisted of combinations of our MEDI rules and SemRep.

The methodology we devised for extracting treatment relations using MEDI was based on the assumption that any two concepts corresponding to a medication-indication pair in MEDI are likely to be in a treatment relation. Although this assumption does not take into account the context in which the two concepts are mentioned, our review of documents before this study revealed that it holds true for the majority of the cases. For instance, the concepts in (3) show an example when our assumption is invalid. Despite the fact that these concepts correspond to a medication-indication pair in MEDI, they do not form a treatment relation in this particular context.

To increase the coverage of MEDI, we expanded the initial set of medication-indication pairs by using ontology relationships from the UMLS and RxNorm databases. For instance, because the medications in MEDI are mapped to generic ingredients,[7] the resource contains pairs involving RxCUI#723 (*Amoxicillin*), but it does not include pairs involving medications containing amoxicillin ingredients such as RxCUI# 315367 (*Amoxicillin 125 MG*) or brand names of amoxicillin such as RxCUI#203169 (*Amoxil*).

As shown in figure 1, we proposed three levels of expansion for medications in MEDI. Starting from the initial set of core medication-indication pairs (Level 0), we employed the HAS_INGREDIENT relationship to include pairs having medications that contain in their composition the ingredient concepts from MEDI (Level 1). Our intuition on validating the expanded pairs from Level 1 for identifying treatment relations is derived from the heuristic rule in (6) below. Therefore, in addition to TREATS(*Amoxicillin*, *Pneumonia*) from the core set of pairs,

Level 1 contains relations such as TREATS(*Amoxicillin 125 MG*, *Pneumonia*). Similarly, based on the rule in (7), we used TRADENAME_OF to include pairs with medication brand names in Level 2. Finally, Level 3 of medication expansion contains pairs derived from concepts that have as ingredients brand name medications (or pairs with brand name medications that have as ingredients medication concepts derived from Level 1).

The contribution of each medication expansion level for both MEDI-ALL and MEDI-HPS is listed in table 1. Additionally, a graphical representation of the medication expansion distribution for MEDI-ALL is depicted in figure 2. Of note, the number of the initial pairs in MEDI (63 343 and 13 304 in MEDI-ALL and MEDI-HPS, respectively) is different from the number of core MEDI pairs in table 1 (53 095 and 12 905 in MEDI-ALL and MEDI-HPS, respectively) because we collapsed the pairs having the same identifier (ie, RXCUI→CUI) in the database. For instance, the first three rows in MEDI correspond to the same medication-indication identifier, RXCUI#77 (*Mesna*)→ CUI#C0018965, but have different ICD9 codes and indication descriptions. The ICD9 codes and indication descriptions of these three rows are (599.7, *Hematuria*), (599.70, *Hematuria; unspecified*), and (791.9, *Other nonspecific findings on examination of urine*), respectively.

(6)	TREATS(X, Y) $\land$ HAS_INGREDIENT(Z, X) $\rightarrow$ TREATS(Z, Y)
(7)	TREATS(X, Y) $\land$ TRADENAME_OF (Z, X) $\rightarrow$ TREATS(Z, Y)
(8)	TREATS(X, Y) $\land$ ISA*(Z, Y) $\rightarrow$ TREATS(X, Z)

To expand the list of MEDI indications, we used the UMLS Metathesaurus relationships from MRREL. Specifically, based on the rule in (8), we computed the transitive closure of the ISA relation, ISA*, to include the entire set of hyponym concepts from SNOMED-CT for each indication from MEDI. As figure 1 shows, the hyponym-based expansion will enable the

Figure 1: The MEDication Indication (MEDI) expansion process.
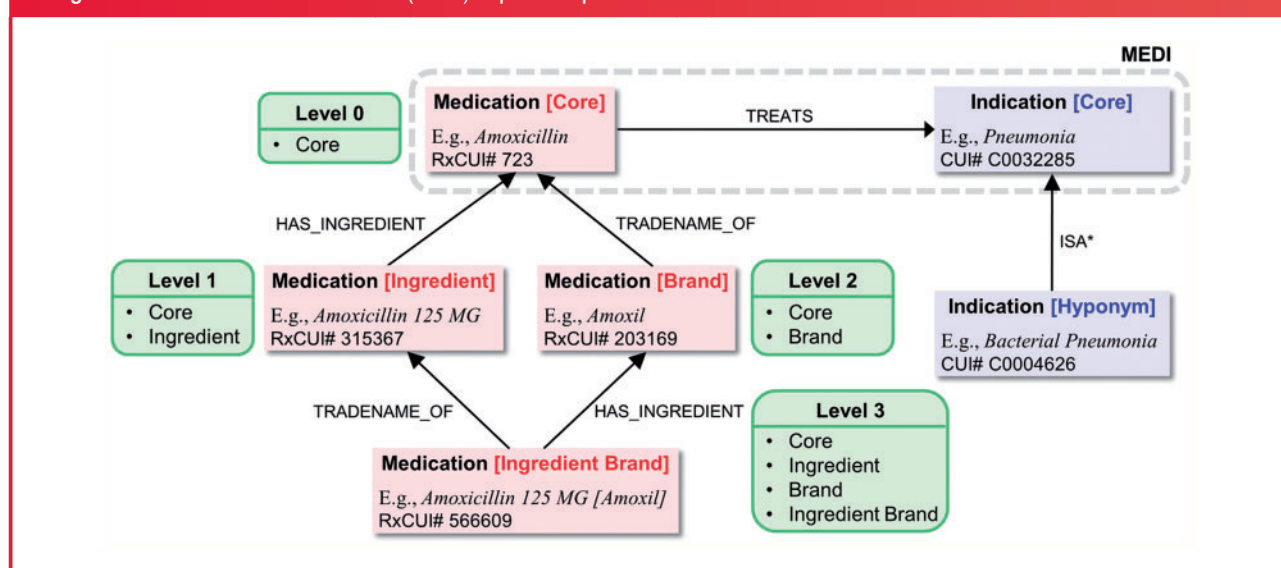
identification of a treatment relation between RxCUI#723 (*Amoxicillin*) and CUI#C0004626 (*Bacterial Pneumonia*) since the bacterial pneumonia concept is a hyponym of the pneumonia concept. The last row in table 1 lists the amount of the new medication-indication pairs introduced by this expansion. We made publicly available all the extended medication-indication pairs at http://knowledgemap.mc.vanderbilt.edu/research/.

Evaluation

To evaluate the three methods, we first constructed a dataset with manually annotated treatment relations. Additionally, we constructed a second dataset of treatment relations from the dataset of manually annotated relations used in the 2010 i2b2 challenge.[8] Then, we measured how accurately our proposed methods identified the manually annotated relations.

Our dataset included 6864 randomly selected discharge summaries from the Vanderbilt Synthetic Derivative, a de-identified version of the electronic medical record. We choose discharge summaries because they contain diverse sources of clinical exposures and treatment histories; nevertheless, the methodologies we performed to process these documents can be applied to any narrative report.

In the first processing step, we employed the OpenNLP sentence detector (http://opennlp.apache.org/) to split the content of each report into sentences. After sentence splitting, we filtered out the duplicate sentences from the reports corresponding to the same patient. The output generated by this process consisted of 290 911 sentences. Then, we parsed these sentences with SemRep, V.1.5, the version currently available at the time of this study.

In the second processing step, we identified treatment relations in the 6864 discharge summaries using the MEDI knowledge-based approach. To focus the evaluation on the differences between SemRep and MEDI relations (instead of concept identification techniques), we restricted the MEDI matching procedure to concepts identified by MetaMap, applied when extracting SemRep relations. Furthermore, since SemRep is designed to identify semantic relations at sentence level, we applied the MEDI algorithm for any pair of concepts mentioned in the same sentence.

To create a gold standard, two reviewers annotated all the sentences in which SemRep and the MEDI algorithm identified at least one treatment relation. The MEDI algorithm configuration included MEDI-ALL, the Level 3 medication expansion, and the core indication expansion. We decided on this set of sentences to cover as many SemRep and MEDI relations as possible and, at the same time, to minimize the annotation effort. One limitation of this approach, however, is that the evaluation of both methods on the resulting dataset will overestimate recall. To facilitate the annotation process, we loaded these sentences into the web-based BRAT (brat rapid annotation tool)[23] and preserved only the concept annotations; therefore,

Table 1: Statistics describing the medication and indication expansions for the MEDI-ALL and MEDI-HPS datasets

	MEDI-ALL		MEDI-HPS	
	Pairs	%	Pairs	%
Medication expansion				
Level 0 (Core)	53 095	0.65	12 905	0.64
Level 1	2 178 824	26.79	505 298	25.05
Level 2	849 046	10.44	210 645	10.45
Level 3	5 053 148	62.12	1 287 958	63.86
Indication expansion				
Core	53 095	5.31	12 905	2.2
Hyponym	946 553	94.69	573 475	97.8

HPS, high-precision subset; MEDI, medication indication.

Figure 2: The contribution of the medication-indication pairs from MEDI-ALL to each level of medication expansion.

RESEARCH AND APPLICATIONS
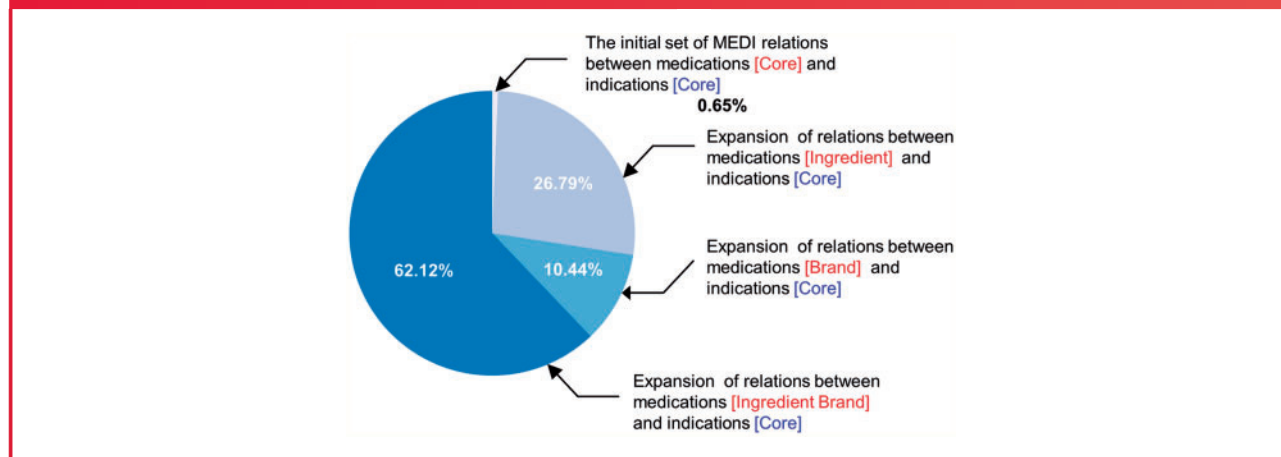
the manual annotation of treatment relations was blinded to the relations extracted by SemRep and the MEDI algorithm.

The annotation process consisted of the user manually linking pairs of concepts that represent treatment relations inside every sentence by dragging and dropping a concept onto another (BRAT allows easy inversions of direction errors if the process was inadvertently inverted). Once a sentence was annotated, the remaining combinations of concepts not representing treatment relations in the sentence were automatically marked as non-treatment relations. During this process, reviewers double annotated 25% of the data obtaining a percentage agreement of 97.9 and a Cohen's κ of 0.86. Disagreements were adjudicated by an experienced, board-certified clinician. Overall, reviewers annotated 958 treatment relations. The remaining concept combinations consisted of 9628 non-treatment relations.

To create the second dataset of treatment relations, we mapped the relation categories defined for the 2010 i2b2 challenge into treatment and non-treatment relations. After analyzing the dataset used for this competition, we identified two relation categories that could be mapped to treatment relations: (1) relations where the treatment of a patient's medical problem cures or improves the problem (TrIP); and (2) relations where the treatment is administered for a medical problem (TrAP). We mapped the remaining concept pairs occurring within the same sentence (and not being labeled as TrIP or TrAP) into non-treatment relations.

During the evaluation of the relations derived from the 2010 i2b2 dataset, we selected only the ones whose corresponding arguments match the concepts identified by MetaMap while running SemRep. This is because the annotated concepts in the 2010 i2b2 dataset are not mapped to concepts from the UMLS database. Therefore, neither SemRep nor the MEDI algorithm is able to identify relations with concepts that could not be mapped to MetaMap concepts. During this process, we also filtered out the i2b2 sentences that were further split by SemRep.

We measured the performance of our systems in terms of precision, recall, and F1-measure. Additionally, to determine whether the differences in performance between the MEDI related systems and SemRep are statistically significant, we employed a randomization test based on stratified shuffling[24] with Bonferroni correction for multiple testing.

RESULTS

After processing the set of 6864 discharge summaries, SemRep identified 943 306 concepts and 3386 treatment relations in 2841 sentences. These treatment relations include 49 TREATS(SPEC) and 5 TREATS(INFER) relations. The Venn diagrams (A–D) in figure 3 show the relationship between the SemRep

Figure 3: The diagrams in panels A–D show the connection between the relations identified by the MEDI algorithm and SemRep for each medication expansion level. The diagrams in panels E–H depict the connection between the relation types generated by the two resources. Finally, the diagrams in panels I–L consider the set of relations with their type generated by both resources. MEDI algorithm configuration: MEDI-ALL and the core indication expansion.
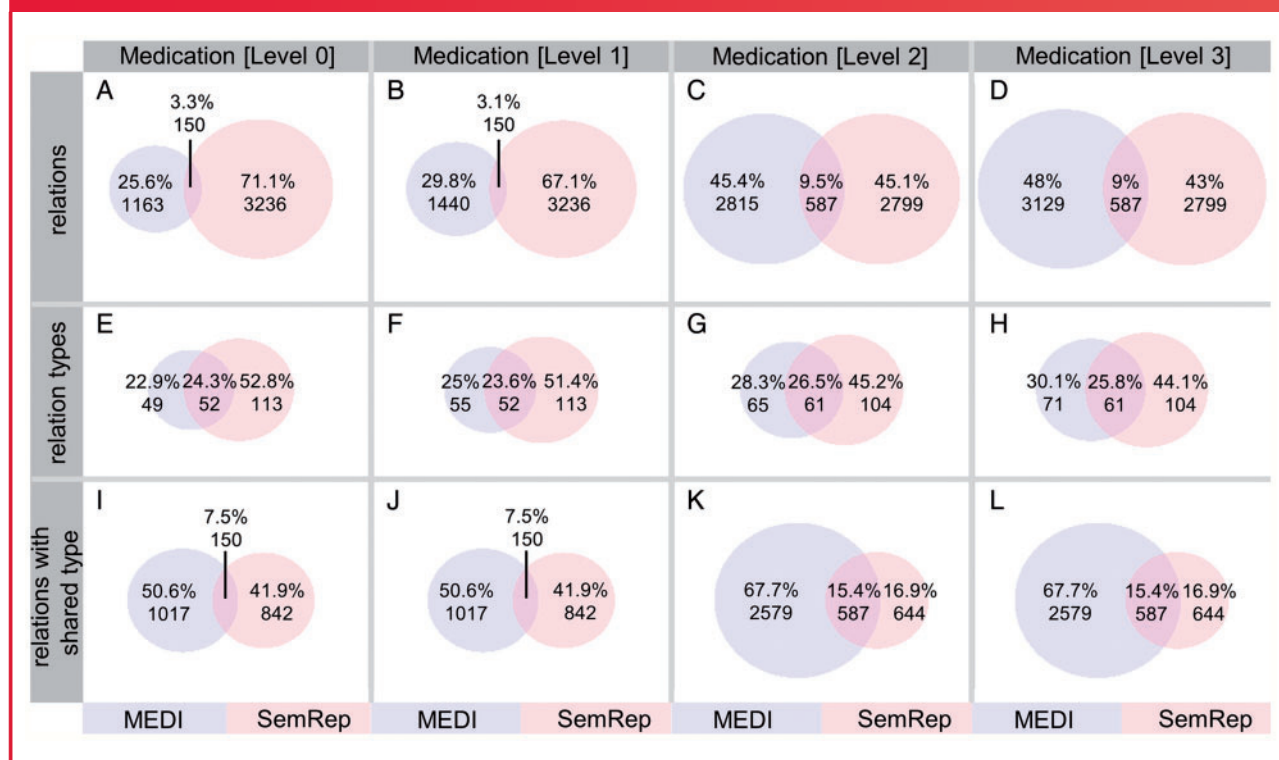
and MEDI relations extracted from discharge summaries. As observed, the number of relations derived from MEDI varies from 1313 (Level 0) to 3716 (Level 3). The diagrams (E–H) illustrate the connection between the sets of unique relation types generated by SemRep and MEDI. Furthermore, the diagrams (I–L) show the relationship between the SemRep and MEDI relations whose corresponding types are common for both resources. For instance, the diagram (K) includes only those relations with types belonging to the 61 common types shown in diagram (G). In this study, the type of a treatment relation is represented by the UMLS semantic types associated with the relation arguments. For instance, the relation type of the treatment relation in (1) is denoted as orch, phsu→dsyn, where orch and phsu are the semantic types of carvedilol, and dsyn is the semantic type of hypertension. In this notation, orch, phsu, and dsyn represent the abbreviations of the 'organic chemical', 'pharmacologic substance', and 'disease or syndrome' UMLS semantic types, respectively. As observed, arguments can have more than one semantic type. This is because a concept can be assigned to multiple semantic types in the UMLS Metathesaurus.

As a result of evaluating the MEDI algorithm on the set of manually annotated relations, we found that medication expansions significantly improved recall with only minimal decreases in precision (figure 4). For instance, in the first bar plot from the left (ie, the configuration using MEDI-ALL and the core indication expansion) the recall increased from 21.92 to 76.62

from Level 0 (core MEDI) to Level 3 (MEDI expanded with brand name medications and related ingredients), while the precision dropped from 82.03 to 80.48. The most significant gains in recall were yielded when including medication brand names (ie, Levels 2 and 3). The moderate recall increases when adding the ingredient medications (ie, from Level 0 to 1 and from Level 2 to 3) were mainly caused by the MetaMap preference for identifying shorter medication expressions (eg, preference for identifying "[*dolasetron*] *100 mg*" instead of "[*dolasetron 100 mg*]"). When comparing the results obtained by the two MEDI databases (ie, MEDI-ALL vs MEDI-HPS), we observed higher precision values for MEDI-HPS at the cost of small decreases in recall. Finally, a study on the indication expansion levels showed an increase in recall for the hyponym indication expansion and comparatively higher precision values for the core indication level. To understand better the results in these plots, it is worth recalling the relationships that exist between the set of pairs corresponding to Level 0 (core MEDI), Level 1 (ingredient medications), Level 2 (brand name medications), and Level 3 (brand name medications and related ingredients): Level 0 ⊆ Level 1, Level 0 ⊆ Level 2, and Level 1 ∪ Level 2 ⊆ Level 3. Similarly, for indications, the hyponym expansion includes the initial set of MEDI pairs.

The results listed in table 2 correspond to a more detailed set of experiments using the manually annotated relations. In addition to SemRep and the MEDI algorithm, we included two simple ensemble methods. These methods (denoted as 'MEDI

Figure 4: Precision and recall results achieved by the MEDI algorithm for each expansion configuration. The elements of the horizontal axis in each plot correspond to results for each medication expansion level. The results of the algorithm using MEDI-ALL and MEDI-HPS are shown in the first and last two plots from the left, respectively. The results for the core indication expansion are shown in the first and third plot from the left. Finally, the results for the hyponym indication expansion are shown in the second and fourth plot from the left.
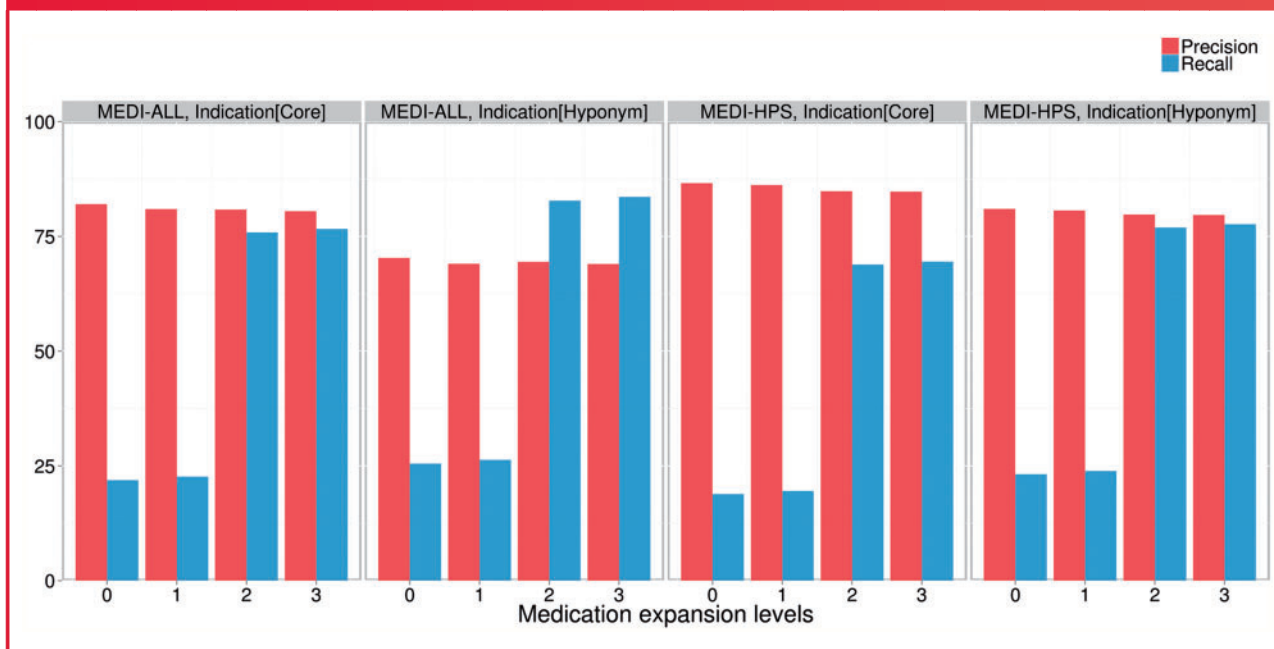
RESEARCH AND APPLICATIONS

Table 2: Performance results for evaluating the treatment relations predicted by SemRep, the MEDI algorithm, and two ensemble methods of these resources

Medication expansion	MEDI dataset	Indication expansion	System	All annotated relations			Relations with common type			Relations with arguments in MEDI		
				P	R	F	P	R	F	P	R	F
1	2	3	4	5	6	7	8	9	10	11	12	13
Level 0	MEDI-ALL		SemRep	81.17	65.24	72.34	86.51	71.19	78.11	95.04	73.02	82.59
		Core	MEDI	82.03	21.92	34.60	83.61	24.09	37.40	82.03	28.61	42.42
			MEDI and SemRep	95.77	14.20	24.73	95.77	16.06	27.50	95.77	18.53	31.05
			MEDI or SemRep	79.07	72.96	75.90**	83.98	79.22	81.53*	89.97	83.11	86.40**
		Hyponym	MEDI	70.32	25.47	37.39	72.90	27.63	40.07	82.03	28.61	42.42
			MEDI and SemRep	92.12	15.87	27.07	92.07	17.83	29.87	95.77	18.53	31.05
			MEDI or SemRep	75.32	74.84	75.08*	80.33	80.99	80.66	89.97	83.11	86.40**
	MEDI-HPS	Core	MEDI	86.60	18.89	31.02	87.13	20.78	33.56	86.60	24.66	38.39
			MEDI and SemRep	98.31	12.11	21.56	98.31	13.70	24.04	98.31	15.80	27.23
			MEDI or SemRep	80.14	72.03	75.87**	84.89	78.28	81.45*	91.76	81.88	86.54**
		Hyponym	MEDI	81.02	23.17	36.04	82.44	25.50	38.95	86.73	26.70	40.83
			MEDI and SemRep	95.89	14.61	25.36	95.86	16.41	28.02	97.67	17.17	29.20
			MEDI or SemRep	78.73	73.80	76.19**	83.54	80.28	81.88*	91.68	82.56	86.88**
Level 1	MEDI-ALL	Core	MEDI	80.97	22.65	35.40	83.61	24.09	37.40	80.97	29.56	43.31
			MEDI and SemRep	95.77	14.20	24.73	95.77	16.06	27.50	95.77	18.53	31.05
			MEDI or SemRep	78.79	73.70	76.16**	83.98	79.22	81.53*	89.42	84.06	86.66**
		Hyponym	MEDI	69.04	26.30	38.10	72.90	27.63	40.07	80.97	29.56	43.31
			MEDI and SemRep	92.12	15.87	27.07	92.07	17.83	29.87	95.77	18.53	31.05
			MEDI or SemRep	74.74	75.68	75.21*	80.33	80.99	80.66	89.42	84.06	86.66**
	MEDI-HPS	Core	MEDI	86.18	19.52	31.83	87.13	20.78	33.56	86.18	25.48	39.33

(continued)

Table 2: Continued

Medication expansion	MEDI dataset	Indication expansion	System	All annotated relations			Relations with common type			Relations with arguments in MEDI		
1	2	3	4	P 5	R 6	F 7	P 8	R 9	F 10	P 11	R 12	F 13
			MEDI and SemRep	98.31	12.11	21.56	98.31	13.70	24.04	98.31	15.80	27.23
			MEDI or SemRep	80.09	72.65	76.19**	84.89	78.28	81.45*	91.55	82.70	86.90**
		Hyponym	MEDI	80.63	23.90	36.88	82.44	25.50	38.95	86.02	27.66	41.86
			MEDI and SemRep	95.89	14.61	25.36	95.86	16.41	28.02	97.67	17.17	29.20
			MEDI or SemRep	78.63	74.53	76.53**	83.54	80.28	81.88*	91.36	83.51	87.26**
Level 2	MEDI-ALL	Core	MEDI	80.87	75.89	78.30**	81.85	84.65	83.23**	80.87	99.05	89.04***
			MEDI and SemRep	95.04	55.95	70.43	95.04	63.28	75.97	95.04	73.02	82.59
			MEDI or SemRep	73.85	85.18	79.11***	77.70	92.56	84.48***	80.87	99.05	89.04***
		Hyponym	MEDI	69.44	82.78	75.52*	72.13	91.97	80.85	80.87	99.05	89.04***
			MEDI and SemRep	92.85	59.60	72.60	92.83	67.30	78.03	95.04	73.02	82.59
			MEDI or SemRep	65.30	88.41	75.12*	69.82	95.87	80.80	80.87	99.05	89.04***
	MEDI-HPS	Core	MEDI	84.83	68.89	76.04**	85.45	76.98	80.99	84.83	89.92	87.30**
			MEDI and SemRep	96.07	51.04	66.67	96.07	57.73	72.12	96.07	66.62	78.68
			MEDI or SemRep	76.61	83.09	79.72***	80.55	90.44	85.21***	84.87	96.32	90.24***
		Hyponym	MEDI	79.76	76.93	78.32**	80.98	85.95	83.39**	84.33	94.55	89.15***
			MEDI and SemRep	94.87	55.95	70.39	94.86	63.16	75.83	95.87	69.62	80.66
			MEDI or SemRep	73.16	86.22	79.16***	77.13	93.98	84.73***	84.19	97.96	90.55***

(continued)

RESEARCH AND APPLICATIONS

RESEARCH AND APPLICATIONS

Table 2: Continued

Medication expansion	MEDI dataset	Indication expansion	System	All annotated relations			Relations with common type			Relations with arguments in MEDI		
1	2	3	4	P 5	R 6	F 7	P 8	R 9	F 10	P 11	R 12	F 13
Level 3	MEDI-ALL	Core	MEDI	80.48	76.62	78.50***	81.85	84.65	83.23**	80.48	100.0	89.19***
			MEDI and SemRep	95.04	55.95	70.43	95.04	63.28	75.97	95.04	73.02	82.59
			MEDI or SemRep	73.61	85.91	79.29***	77.70	92.56	84.48***	80.48	100.0	89.19***
		Hyponym	MEDI	68.99	83.61	75.60*	72.13	91.97	80.85	80.48	100.0	89.19***
			MEDI and SemRep	92.85	59.60	72.60	92.83	67.30	78.03	95.04	73.02	82.59
			MEDI or SemRep	64.97	89.25	75.20*	69.82	95.87	80.80	80.48	100.0	89.19***
	MEDI-HPS	Core	MEDI	84.73	69.52	76.38**	85.45	76.98	80.99	84.73	90.74	87.63**
			MEDI and SemRep	96.07	51.04	66.67	96.07	57.73	72.12	96.07	66.62	78.68
			MEDI or SemRep	76.60	83.72	80.00***	80.55	90.44	85.21***	84.78	97.14	90.54***
		Hyponym	MEDI	79.66	77.66	78.65***	80.98	85.95	83.39**	84.15	95.50	89.47***
			MEDI and SemRep	94.87	55.95	70.39	94.86	63.16	75.83	95.87	69.62	80.66
			MEDI or SemRep	73.13	86.95	79.45***	77.13	93.98	84.73***	84.03	98.91	90.86***

The experiments were performed using the entire set of annotated relations (columns 5–7), only relations with types generated by both SemRep and the MEDI algorithm (columns 8–10), and relations with both arguments in MEDI (columns 11–13).

*p<0.05, **p<0.01,***p<0.001, all after Bonferroni correction; statistically significant differences in performance between the MEDI related systems and SemRep. Each emphasized result represents the best performance value in its corresponding column.

F, F1-measure; HPS, high-precision subset; MEDI, medication indication; P, precision; R, recall.

Table 3: Frequency counts of the relation sets used in table 2

Relation set	Treatment	Non-treatment
All annotated relations	958	9628
Relations with common type	847	606
Relations with arguments in MEDI	734	178

MEDI, medication indication.

Table 4: Evaluation of treatment relations derived from the i2b2 test set

Configuration	System	P	R	F
	UTD	76.08	83.83	79.77
	SemRep	84.50	19.16	31.23
MEDI-ALL				
Core	MEDI	77.23	27.42	40.47
	MEDI and SemRep	94.92	9.84	17.83
	MEDI or SemRep	76.84	36.73	49.70
Hyponym	MEDI	42.83	34.62	38.29
	MEDI and SemRep	92.00	12.13	21.43
	MEDI or SemRep	46.11	41.65	43.77
MEDI-HPS				
Core	MEDI	83.85	23.73	36.99
	MEDI and SemRep	94.12	8.44	15.48
	MEDI or SemRep	82.01	34.45	48.51
Hyponym	MEDI	59.73	30.76	40.60
	MEDI and SemRep	92.42	10.72	19.21
	MEDI or SemRep	62.64	39.19	48.22

The configuration of MEDI-related systems included the Level 3 medication expansion.

F, F1-measure; HPS, high-precision subset; MEDI, medication indication; P, precision; R, recall; UTD, University of Texas at Dallas.

or SemRep' and 'MEDI and SemRep') take the union and intersection of the results predicted by SemRep and the MEDI algorithm for each relation. Furthermore, to gain a deeper insight into the behavior of these systems when the manually annotated relations are better covered by MEDI, we performed the experiments on three relation sets: (1) the entire set of annotated relations, (2) relations with types generated by both the MEDI algorithm and SemRep, and (3) relations with both arguments in MEDI. Table 3 lists the dimensions of these three relation sets. Of note, the MEDI precision and recall values for the experiments including all annotated relations in table 2 (ie, the MEDI results from columns 5 and 6) correspond to the precision and recall values from figure 4.

For each set of experiments from table 2, unioning SemRep with MEDI and using the MEDI systems alone with the medication expansion levels 2 and 3 outperformed SemRep alone. The highest F1-measure for the experiment using all annotated relations (F1-measure = 80) was achieved by the union system in a configuration including MEDI-HPS with Level 3 medication expansion and core indications. Similarly, the best MEDI alone system (F1-measure = 78.65) used MEDI-HPS and the Level 3 medication expansion but also used the hyponym indication expansion instead of just core indications. Not surprisingly, the more restrictive system, the intersection of the MEDI algorithm and SemRep, achieved the highest precisions at the cost of decreased recall and lower F1 scores.

In tables 4 and 5, we evaluated our selected methods on the relations derived from the 2010 i2b2 test set and compared their results with the results of the top-ranked system[9] in the task of relation extraction at the 2010 i2b2 competition. This system was developed at the University of Texas at Dallas (UTD). First, we converted the relations from the i2b2 test set into 2685 treatment relations (ie, TrIP and TrAP relations) and 51 744 non-treatment relations. Next, we evaluated only the concept pairs that have a direct correspondence to the MetaMap concepts as described in the previous section. The final set of relations consisted of 569 and 12 924 treatment and non-treatment relations, respectively.

Table 4 lists the results on this final set of relations. When compared with the UTD system, both SemRep and the MEDI algorithm achieved better precision and significantly lower recall values. Notably, a comparison of the results from tables 2 and 4 shows an overestimation of the recall values obtained by SemRep and the MEDI algorithm on our annotated dataset;

this analysis confirms the limitation of our annotation procedure.

The results listed in table 5 correspond to the experiments in which the i2b2 relations were constrained to have both arguments in MEDI. The counts of the relation sets for each MEDI configuration are listed in the first column of table 5. As observed, for the MEDI-ALL[Core] and MEDI-HPS[Core] experiments, the UTD system and the MEDI algorithm achieved comparable F1-measure results.

Error analysis

Some of the most frequent false negative examples by SemRep occurred when treatment relations were characterized by long textual distances between arguments. Another common error occurred when multiple concepts with the same semantic type were described in sequential order. For instance, SemRep did not identify any treatment relations between the concepts emphasized in the following sentence: "*The patient was noted to have [**cellulitis**] of his right leg stump on admission and was treated with [**Dicloxacillin**] and [**Keflex**] as well as IV [**Ancef**]*" Although the valid relations characterized by the most frequent lexical patterns (eg, "*for*", "*was managed with*", "*treated with*", "*as needed for*") were correctly identified in our dataset, we found various treatment relations described by simple patterns (eg, "*well controlled with*", "*dramatically improved with*",

Table 5: Evaluation of treatment relations from i2b2 test set with arguments in MEDI

Treatment/ non-treatment relations	System	P	R	F
MEDI-ALL [Core]				
	UTD	91.10	85.26	88.08
	SemRep	94.92	35.90	52.09
156/46	MEDI	77.23	100.0	87.15
	MEDI and SemRep	94.92	35.90	52.09
	MEDI or SemRep	77.23	100.0	87.15
MEDI-ALL [Hyponym]				
	UTD	82.83	83.25	83.04
	SemRep	92.00	35.03	50.74
197/263	MEDI	42.83	100.0	59.97
	MEDI and SemRep	92.00	35.03	50.74
	MEDI or SemRep	42.83	100.0	59.97
MEDI-HPS [Core]				
	UTD	92.25	88.15	90.15
	SemRep	94.12	35.56	51.61
135/26	MEDI	83.85	100.0	91.22
	MEDI and SemRep	94.12	35.56	51.61
	MEDI or SemRep	83.85	100.0	91.22
MEDI-HPS [Hyponym]				
	UTD	88.17	85.14	86.63
	SemRep	92.42	34.86	50.62
175/118	MEDI	59.73	100.0	74.79
	MEDI and SemRep	92.42	34.86	50.62
	MEDI or SemRep	59.73	100.0	74.79

The configuration of MEDI-related systems included the Level 3 medication expansion.

F, F1-measure; HPS, high-precision subset; MEDI, medication indication; P, precision; R, recall; UTD, University of Texas at Dallas.

"*was retreated with*", "*therapy was instituted for*") that were not captured by SemRep.

In contrast, many false positive examples by MEDI occurred in complex sentences in which the context is critical in determining the relationship between concepts. One example incorrectly identified by MEDI is the relation between the emphasized concepts in the following sentence: "*He was begun on Unasyn for his* [*pneumonia*] *as well as* [*Ganciclovir*] *for his CMV culture*".

Our analysis on the i2b2 test set revealed that the majority of false negative examples by MEDI occurred when the relation arguments included procedures (eg, *chemotherapy*, *urostomy*), general medical expressions (eg, *therapy*, *treatment*, *management*, *disease*, *fluids*, *pain control*), and abbreviations

(eg, *A-*[*fib*] and [*DM2*] were incorrectly mapped by MetaMap to concepts having the gene or genome semantic type). Furthermore, more than 34% of the false positive examples by MEDI on the i2b2 test set were associated with TrCP (eg, "*She has* [*allergies*] *to* [*Codeine*]"), TrNAP (eg, "[*atrial fibrillation*] *for which he could not be on* [*Coumadin*] *because of history of GI bleed*"), and TrWP (eg, "[*oliguria*] *unresponsive to* [*Lasix*]") relation categories. Therefore, if the prevalence of these relation categories is higher on the i2b2 dataset, the precision values of the MEDI algorithm will be underestimated. This observation could explain the differences in precision between the results listed in table 2 and the results listed in tables 4 and 5.

A significant percentage of errors may be attributed to imperfect concept identification by MetaMap, which also represented one of the main challenges in the annotation of treatment relations. For instance, in "[*perirectal*] [*pain*] *was controlled with* [*morphine*]", the identification of "[*perirectal pain*]", which is not a concept in UMLS 2012AA used by SemRep in this study, would have been preferred. In this example, the concepts expressed by "[*morphine*]" and "[*pain*]" still represent a valid treatment relation because perirectal pain is a more specific concept of pain. Additional examples of concepts with imperfect annotations that could be selected as arguments in treatment relations are: "*oral* [*ciprofloxacin*]", "*left* [*stump pain*]", "*skull* [*osteomyelitis*]", "*occipital* [*hemorrhage*]", etc. However, there exist many expressions with imperfect concept annotations from which none of the incorrectly identified concepts can be used to annotate treatment relations. Examples of such expressions are: "[*shortness*] *of* [*breath*]", "[*increasing*] [*creatinine*]", "[*rapid*] [*heart rate*]", and "[*elevated*] [*WBC*]".

DISCUSSION

Our experiments indicated that MEDI is a broad and reliable resource on assessing the validity of the treatment relations extracted from clinical text. The experiments also showed that SemRep is able to extract treatment relations from patient reports with high precision, despite the fact that it was not developed primarily for the clinical domain.

Although our algorithm does not take into consideration sentence syntax or context, we empirically proved that a simple set of rules using the MEDI knowledge base yielded better results than SemRep for identifying treatment relations in clinical text. This confirms our hypothesis that a valid medication-indication pair expressed in a sentence corresponds to a treatment relation in the majority of cases. Additionally, our experiments provided evidence that the MEDI algorithm is able to better handle the treatment relations characterized by a longer distance dependency. In our experiments, the average word length between the arguments corresponding to the manually annotated treatment relations correctly identified by the MEDI algorithm and SemRep is 4.8 and 3.8, respectively.

An interesting observation drawn from our study is that MEDI and SemRep shared a relatively small number of relations (see figure 3). One possible explanation is that not all the

Table 6: The most frequent relation types identified by SemRep and the MEDI algorithm

All by SemRep			All by MEDI			Only by SemRep			Only by MEDI		
Relation types	Freq	%	Relation types	Freq	%	Relation types	Freq	%	Relation types	Freq	%
orch,phsu→sosy	425	12.6	orch,phsu→sosy	1582	42.6	topp→dsyn	324	15.0	clnd→sosy	270	49.1
topp→dsyn	324	9.6	orch,phsu→dsyn	616	16.6	topp→podg	264	12.3	orch,phsu→menp	48	8.7
topp→podg	264	7.8	clnd→sosy	270	7.3	diap,topp→podg	185	8.6	orch,phsu→ortf	25	4.6
diap,topp→podg	185	5.5	antb,orch→dsyn	194	5.2	topp→neop	182	8.5	clnd→dsyn	25	4.6
topp→neop	182	5.4	orch,phsu→patf	167	4.5	topp→fndg	117	5.4	antb,orch→fndg	22	4.0

MEDI algorithm configuration: MEDI-ALL, the Level 3 medication expansion, and the core indication expansion.

antb, antibiotic; clnd, clinical drug; diap, diagnostic procedure; dsyn, disease or syndrome; fndg, finding; menp, mental process; neop, neoplastic process; orch, organic chemical; ortf, organ or tissue function; patf, pathologic function; phsu, pharmacologic substance; podg, patient or disabled group; sosy, sign or symptom; topp, therapeutic or preventive procedure.

Freq, frequency; MEDI, medication indication.

SemRep relation types are included in MEDI. For instance, the relation between "*left total hip arthroplasty*" and "*arthritis*" in (2) is identified as a treatment relation by SemRep, but it does not have a corresponding matching pair in MEDI since MEDI does not include procedures. Conversely, not all the MEDI relation types were generated by SemRep. This can be explained by the fact that, in SemRep, the types of the relations extracted from text are constrained to match the types of their corresponding relations in the UMLS Semantic Network.[6,25,26]

Table 6 lists the top five most frequent relation types extracted by SemRep and the MEDI algorithm from our set of discharge summaries, and the top five relation types that were found by only one of these systems. Not surprisingly, the most frequent relation type for both SemRep and MEDI is orch, phsu→sosy, which represents the generic type medication→disease. The next four most frequent SemRep relation types correspond to the generic types procedure→disease, procedure→patient, and procedure→neoplastic process. These types are also the most frequent ones that are not found among the MEDI relation types (see 'Only in SemRep' column in table 6).

This finding is also reflected in the results of our selected methods on the entire set of treatment relations from the i2b2 test set. For instance, table 4 shows lower recall values of the MEDI algorithm on this set because the MEDI algorithm was not designed to process all the relation types from the i2b2 dataset (eg, relations involving procedures and general medical expressions). However, when the relation arguments are constrained to match the concept pairs in MEDI (see table 5), this algorithm achieved F1-measure results comparable with the ones of the top-ranked system, which performed extensive feature engineering and parameter optimization on the i2b2 training set.

Finally, to illustrate how treatment relation extraction could be used in clinical applications, we employed MEDI to extract and aggregate the reasons why medications were prescribed in our collection of 6864 discharge summaries. A select set of medication and indication distributions computed from these clinical reports is presented in figure 5. Similar distributions using the ICD9 codes of indications in MEDI were computed from a large cohort of patients.[27] Extracting such distributions could enable more accurate patient profiles and use of semantic search engines to answer complex clinical questions[28] for individual patients (eg, "what medication was used to treat his pneumonia?") or populations (eg, "what are the most common uses for levofloxacin?" or "how often are narcotics used for back pain?"). A richer linkage of drugs to their indications could also improve electronic medical record-based phenotyping[29–31] for clinical and genomic research by allowing searching not just for drugs relevant to a disease of interest but specifying which drugs were actually used to treat a disease. This explicit linkage would especially be helpful for broad-spectrum medications such as immunosuppressants (eg, steroids) and antibiotics.

In future research, we plan to develop a machine learning framework which will use MEDI as prior information to identify treatment relations. The feature set will also include the assertion values associated with the concepts of interest[32] and statistically significant features that capture the contextual information of these concepts.[33,34] Our goal was not only to improve treatment relation extraction, but also to discover new treatment relations that are not yet covered in MEDI. For instance, none of the MEDI algorithm configurations were able to identify the following treatment relations in our dataset: *ethambutol→infection*, *lidocaine→stump pain*, *famvir→oral ulcers*, *levaquin→pyuria*, *vesicare→bladder spasm*. Moreover, we plan to refine the distributions derived from extracting treatment relations by running the learning system over a larger set of clinical notes.

CONCLUSION

In this paper, we evaluated the impact of using a medication-indication resource for the task of identifying treatment relations in clinical text. Based on a set of heuristic rules and relationships from RxNorm and UMLS Metathesaurus, we expanded the initial set of medication-indication pairs in MEDI

RESEARCH AND APPLICATIONS

RESEARCH AND APPLICATIONS

Figure 5: Percentage distributions of medications and indications computed by running the MEDI algorithm over the set of discharge summaries. Each bar plot in red represents the distribution of all indications treated by a specific medication. Similarly, each bar plot in blue shows the distribution of all medications prescribed for a specific indication. In addition to the concept name, each plot title specifies the number of times the corresponding concept was identified in the dataset. For building these distributions, we converted all medication brand names into their corresponding generic names. MEDI algorithm configuration: MEDI-HPS, the Level 3 medication expansion, and the core indication expansion.
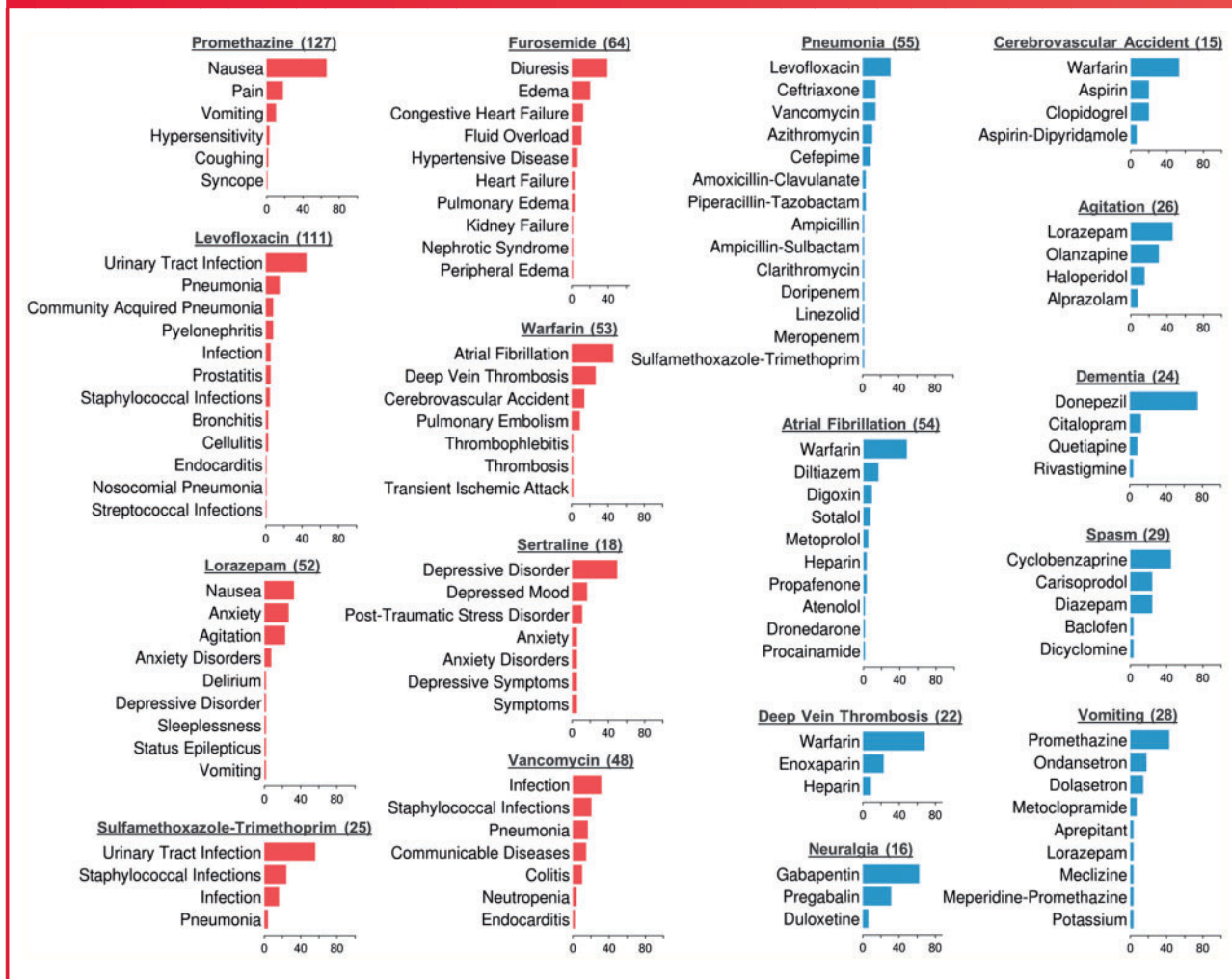
to provide better coverage of treatment relations. The most significant gains in recall for the MEDI algorithm were achieved when the set of medication-indication pairs was expanded to include medication brand names. Our investigation on SemRep demonstrates that this system was effective at extracting treatment relations from clinical text. However, both the MEDI algorithm and the union ensemble system significantly outperformed SemRep for this task. The experiments on the i2b2 test set showed better precision and significantly lower recall values of the MEDI algorithm when compared with the top-ranked results for this set. When the i2b2 relations were restricted to have both arguments in MEDI, the MEDI algorithm—which was not optimized on the i2b2 training set—achieved comparable F1-measure values with the ones of the best relation extraction

system at the i2b2 competition. These studies highlight the importance of knowledge-based approaches, such as MEDI, in relation extraction. Future evaluations are needed to assess the usability of such systems in clinical applications.

ACKNOWLEDGEMENTS

CONTRIBUTORS

CAB and JCD conceived the study design and drafted the main manuscript. CAB implemented the algorithms described in the

paper and ran all the experiments. W-QW provided important guidance on using the MEDI resource. All authors approved the final manuscript.

COMPETING INTERESTS

None.

PROVENANCE AND PEER REVIEW

Not commissioned; externally peer reviewed.

REFERENCES

1. Cebul RD, Love TE, Jain AK, *et al*. Electronic health records and quality of diabetes care. *N Engl J Med*. 2011;365:825–33.
2. Ghitza UE, Sparenborg S, Tai B. Improving drug abuse treatment delivery through adoption of harmonized electronic health record systems. *Subst Abuse Rehabil*. 2011;2011:125–31.
3. Liu M, Wu Y, Chen Y, *et al*. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J Am Med Inform Assoc*. 2012;19:28–35.
4. Roth CP, Lim YW, Pevnick JM, *et al*. The challenge of measuring quality of care from the electronic health record. *Am J Med Qual*. 2009;24:385–94.
5. Roth MT, Weinberger M, Campbell WH. Measuring the quality of medication use in older adults. *J Am Geriatr Soc*. 2009;57:1096–102.
6. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform*. 2003;36:462–77.
7. Wei WQ, Cronin RM, Xu H, *et al*. Development and evaluation of an ensemble resource linking medications to their indications. *J Am Med Inform Assoc*. 2013;20:954–61.
8. Uzuner Ö, South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 2011;18:552–6.
9. Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. *J Am Med Inform Assoc*. 2011;18:594–600.
10. De Bruijn B, Cherry C, Kiritchenko S, *et al*. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc*. 2011;18:557–62.
11. Minard AL, Ligozat AL, Ben Abacha A, *et al*. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. *J Am Med Inform Assoc*. 2011;18:588–93.
12. Patrick JD, Nguyen DHM, Wang Y, *et al*. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *J Am Med Inform Assoc*. 2011;18:574–9.
13. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings AMIA Symposium*. 2001:17–21.
14. Fiszman M, Rindflesch TC, Kilicoglu H. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical texts. *Proceedings AMIA Symposium* 2003:239–43.
15. Rindflesch TC, Pakhomov SV, Fiszman M, *et al*. Medical facts to support inferencing in natural language processing. *Proceedings AMIA Symposium* 2005:634–8.
16. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. *HLT-NAACL Workshop on Computational Lexical Semantics* 2004:76–83.
17. Wilkowski B, Fiszman M, Miller CM, *et al*. Graph-based methods for discovery browsing with semantic predications. *Proceedings AMIA Symposium* 2011:1514–23.
18. Hristovski D, Friedman C, Rindflesch TC, *et al*. Exploiting semantic relations for literature-based discovery. *Proceedings AMIA Symposium* 2006:349–53.
19. Liu Y, Bill R, Fiszman M, *et al*. Using SemRep to label semantic relations extracted from clinical text. *Proceedings AMIA Symposium* 2012:587–95.
20. Kuhn M, Campillos M, Letunic I, *et al*. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*. 2010;6:343.
21. Denny JC, Smithers JD, Miller RA, *et al*. 'Understanding' medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc*. 2003;10:351–62.
22. Denny JC, Spickard AIII, Miller RA, *et al*. Identifying UMLS concepts from ECG impressions using KnowledgeMap. *Proceedings AMIA Symposium* 2005:196–200.
23. Stenetorp P, Pyysalo S, Topić G, *et al*. BRAT: a web-based tool for NLP-assisted text annotation. *Demonstrations Session at EACL* 2012:102–7.
24. Noreen EW. *Computer-intensive methods for testing hypotheses: an introduction*. 1st edn. New York: John Wiley & Sons, 1989.
25. Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In: Chen H, Fuller SS, Friedman C, *et al*., eds. *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. New York: Springer, 2005:399–422.
26. Ahlers CB, Fiszman M, Demner-Fushman D, *et al*. Extracting semantic predications from Medline citations for pharmacogenomics. *Pacific Symposium on Biocomputing* 2007:209–20.
27. Wei WQ, Mosley JD, Bastarache L, *et al*. Validation and enhancement of a Computable Medication Indication Resource (MEDI) using a large practice-based dataset. *Proceedings AMIA Symposium* 2013:1448–56.
28. Ely JW, Osheroff JA, Gorman PN, *et al*. A taxonomy of generic clinical questions: classification study. *BMJ*. 2000;321:429–32.

RESEARCH AND APPLICATIONS

29. Kho AN, Pacheco JA, Peissig PL, *et al*. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med*. 2011;3:79re1.

30. Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput Biol*. 2012;8:e1002823.

31. Newton KM, Peissig PL, Kho AN, *et al*. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc*. 2013;20:147–54.

32. Bejan CA, Vanderwende L, Xia F, *et al*. Assertion modeling and its role in clinical phenotype identification. *J Biomed Inform*. 2013;46:68–74.

33. Bejan CA, Xia F, Vanderwende L, *et al*. Pneumonia identification using statistical feature selection. *J Am Med Inform Assoc*. 2012;19:817–23.

34. Lin C, Karlson EW, Canhao H, *et al*. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. *PLoS One*. 2013;8:e69932.

AUTHOR AFFILIATIONS

[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA

[2]Department of Medicine, Vanderbilt University, Nashville, Tennessee, USA

RESEARCH AND APPLICATIONS