Appendix I: Definitions and Examples of Relation Types

1. Medical problem—treatment relations:
    a. Treatment improves medical problem (TrIP). Includes mentions where the treatment improves or cures the problem, e.g., *hypertension* was controlled on *hydrochlorothiazide*.
    b. Treatment worsens medical problem (TrWP). Includes mentions where the treatment is administered for the problem but does not cure the problem, does not improve the problem, or makes the problem worse, e.g., the *tumor* was growing despite the available *chemotherapeutic regimen*.
    c. Treatment causes medical problem (TrCP). The implied context is that the treatment was not administered for the medical problem that it ended up causing, e.g., *Bactrim* could be a cause of *these abnormalities*.
    d. Treatment is administered for medical problem (TrAP). Includes mentions where a treatment is given for a problem, but the outcome is not mentioned in the sentence, e.g., He was given *Lasix* periodically to prevent him from going into *congestive heart failure*.
    e. Treatment is not administered because of medical problem (TrNAP). Includes mentions where treatment was not given or discontinued because of a medical problem that the treatment did not cause. E.g., *Relafen* which is contraindicated because of *ulcers*.
    f. Treatments and problems that are in the same sentence, but do not fit into one of the above defined relationships are not assigned a relationship.
2. Medical problem—test relations:
    a. Test reveals medical problem (TeRP). Includes mentions where a test is conducted and the outcome is known. E.g., *an echocardiogram* revealed *a pericardial effusion*.
    b. Test conducted to investigate medical problem (TeCP). Includes mentions where a test is conducted and the outcome is not known. E.g., *an VQ scan* was performed to investigate *pulmonary embolus*.
    c. Tests and problems that are in the same sentence, but do not fit into one of the above-defined relationships, are not assigned a relationship.
3. Medical problem—medical problem relations:

a. Medical problem indicates medical problem (PIP). Includes medical problems that describe or reveal aspects of the same medical problem and those that cause other medical problems. E.g., *Azotemia* presumed secondary to *sepsis*.
b. Pairs of medical problems that are in the same sentence, but do not fit into one of the above defined relationships, are not assigned a relationship.

Appendix II: Related Work

In 2006, the first i2b2 shared-task challenge targeted automatic de-identification of discharge summaries and showed that supervised learning systems could identify private health information even in the presence of ambiguities and out-of-vocabulary terms [8]. The same year, a second task run in parallel with the de-identification challenge studied document classification for determining patient smoking status based on discharge summaries, and found that a small set of keywords went a long way towards addressing this task [9]. The second i2b2 challenge took document classification further and showed that systems could determine the presence of obesity and its co-morbidities in a patient by aggregating over all individual mentions of these diseases in individual discharge summaries [10]. The third i2b2 challenge expanded concept extraction from private health information to medications and medication-related information, and showed that medications, their dosages, frequencies, and routes were easier than reasons and durations to identify. The phrasal structure of durations and reasons made them difficult. Linking reasons to their correct medication was especially challenging [3].

Appendix III: Definitions for Evaluation Metrics

For concept extraction, we computed exact and inexact P, R, and F_1 metrics. For exact metrics, we defined TP, FP, and FN per class as:

TP: Annotations in class for which the system and the reference agree exactly in span, as marked by token offset.

FP: System annotations in class that do not exactly match in span with a reference annotation in class. E.g., system generates an annotation in class that the reference standard says should not exist in that class.

FN: Reference annotations in class that do not exactly match in span with a system annotation in class. E.g., system fails to generate an annotation in class that the reference standard includes.

For inexact metrics, we defined TP, FP, and FN per class as:

TP: System and reference annotations in class that overlap on span.

FP: System annotations in class that do not overlap in span with reference annotations in class.

FN: Reference annotations in class that do not overlap in span with system annotations in class.

If two system annotations overlap one reference annotation, one TP and one FP are counted.

For assertions and relations, we defined TP, FP, and FN per class as:

TP: System annotations in class that match reference annotations in class.

FP: System annotations in class that do not match any reference annotations in class

FN: Reference annotations in class that do not match any system annotations in class.

Appendix IV: Outline of Annotation Workflow

The training and test reports for the 2011 i2b2/VA challenge were annotated by a group of 12 annotators, including 8 clinicians. The annotation workflow was as follows: In primary double annotation, two annotators labeled the same document. Primary annotation was followed by a first review in which annotation differences were adjudicated and corrections suggested. A second review then involved the re-adjudication of the primary annotations with the corrections proposed in the first review. The output of this process was programmatically checked for consistency.

Appendix V: Definitions of external resources, medical experts, and methods

External resources ("open" or "closed"): The i2b2/VA challenge team created a list of publicly available resources. At the time of system submission, systems that had been limited to the resources on this list were marked as "closed"; those that had utilized any resources outside of the list were marked as "open".

Medical experts (“yes” or “no”): System outputs that were the product of a team that involved medical experts were marked as “Yes”.

Methods (“unsupervised”, “semi-supervised”, “supervised”, “rule-based”, or “hybrid”): System outputs were marked based on the approaches that the teams utilized.