Abstract
The purpose of this study is to evaluate the feasibility of applying natural language processing in the automated extraction of medications information from unstructured electronic health records. Sixty-two documents containing medications were subjected to both manual and automated extraction. The latter consisted of algorithms which were modifications of a software application called NegEx and a UMLS-based drug lexicon. Both methods were able to identify over 90% medications from the 62 documents. The automated method identified more medications than manual review, 97% vs 92%. However, the automated extraction included a substantial percentage (17%) of non-medication items but none with manual review.

Introduction
In our institute, medication information is currently stored in a free-text format in electronic health records (EHRs). Such information is not coded and is limited value for research and analyses. Natural language processing (NLP) facilitates automatic identification and extraction of pre-defined information including medication from free-text documents using entries from a lexicon. The objective of the current study is to evaluate our ability to apply NLP in the automated extraction of medications information within the EHRs.

Materials and Methods
A sample of 62 transcribed documents of varying services and providers were selected. Each document was manually reviewed by a clinician to extract items that were considered to be “true” medications from all sections of the documents. We then applied algorithms which were modifications of a software application called NegEx 1 (developed by University of Pittsburgh for identifying negated findings in discharge summaries). Each document was tokenized into a list of phrases. The phrases were compared to the UMLS terms. The algorithm returned an XML string that included document identifier and medications associated with the document. This list was then compared with the list obtained manually.

Results
There were 823 items that were distinct medications among the 62 documents. Manual review identified 762 (93%) medications whereas the automated extraction identified 799 (97%) medications; 738 (90%) medications were identified by both methods. Out of 967 items identified by the automated extraction, 168 (17%) were not medications; on the contrary, all items identified manually were medications.

Conclusion
Both methods were able to identify over 90% medications from the documents. The automated extraction identified more medications than manual review, and therefore, was more “sensitive”. However, the automated extraction also included a substantial proportion of non-medication items, therefore, was less “specific” than manual review. Fine tuning of the lexicon may improve the performance of automated extraction.

Reference