Identification of Clinical Characteristics of Large Patient Cohorts through Analysis of Free Text Physician Notes

by

Alexander Turchin

B.A., New York University, 1995
M.D., The Johns Hopkins University School of Medicine, 1999

Submitted to the
Harvard-MIT Division of Health Science and Technology
in Partial Fulfillment of the Requirements of the Degree of

Master of Science in Medical Informatics
at the
Massachusetts Institute of Technology

May 2005 [June 2005]

Signature of author: _____
Alexander Turchin, MD
Children's Hospital Informatics Program

Certified by: _____
Isaac S. Kohane, MD, PhD
Lawrence J. Henderson Associate Professor of Health Sciences and Technology
Director, Children's Hospital Informatics Program
Thesis Supervisor

Accepted by: _____
Martha L. Gray, PhD
J W Kieckhefer Professor of Electrical Engineering
Co-director, Harvard-MIT Division of Health Sciences and Technology

Identification of Clinical Characteristics of Large Patient Cohorts through Analysis of Free Text Physician Notes

by

Alexander Turchin

B.A., New York University, 1995
M.D., The Johns Hopkins University School of Medicine, 1999

Submitted to the
Harvard-MIT Division of Health Science and Technology
in Partial Fulfillment of the Requirements of the Degree of

Master of Science in Medical Informatics

Abstract

Background

A number of important applications in medicine and biomedical research, including quality of care surveillance and identification of prospective study subjects, require identification of large cohorts of patients with specific clinical characteristics. Currently used conventional techniques are either labor-intensive or imprecise, while natural language processing-based applications are relatively slow and expensive.

Specific Aims

In this thesis we describe the design and formal evaluation of PACT – a suite of rapid, accurate, and easily portable software tools for identification of patients with specific clinical characteristics through analysis of the text of physician notes in the electronic medical record.

Methods

PACT algorithm is based on sentence-level semantic analysis. The major steps involve identification of *word tags* (e.g. name of the disease or medications exclusively used to treat the disease) specific for the clinical characteristics in the sentences of the physician notes. Sentences with word tags and *negative qualifiers* (e.g. "*rule out* diabetes") are excluded from consideration. PACT can also identify quantitative (e.g. blood pressure, height, weight) and semi-quantitative (e.g. compliance with medical treatment) clinical characteristics.

PACT performance was evaluated against blinded manual chart review (the "gold standard") and currently used computational methods (analysis of billing data).

Results

Evaluation of PACT demonstrated it to be rapid and highly accurate. PACT processed 6.5 to 8.8×10^5 notes/hour (1.0 to 1.4 GB of text / hour). When compared to the gold standard of manual chart review, PACT sensitivity ranged (depending on the patient characteristic being extracted from the notes) from 74 to 100%, and specificity from 86 to 100%. κ statistic for agreement between PACT and manual chart review ranged from 0.67 to 1.0 and in most cases exceeded 0.75, indicating excellent agreement. PACT accuracy substantially exceeded the performance of currently used techniques (billing

data analysis). Finally, index of patient non-compliance with physician recommendations computed by PACT was shown to correlate with the frequency of annual Emergency Department visits: patients in the highest quartile for the index of non-compliance had 50% as many annual visits as the patients in the lowest quartile.

Conclusion

PACT is a rapid, precise and easily portable suite of software tools for extracting focused clinical information out of free text clinical documents. It compares favorably with computation techniques currently available for the purpose (where ones exist). It represents an important advance in the field, and we plan to continue to develop this concept further to improve its performance and functionality.

Thesis Supervisor: Isaac S. Kohane, MD, PhD
Title: Lawrence J. Henderson Associate Professor of Health Sciences and Technology
Director, Children's Hospital Informatics Program

Acknowledgements

I would like to express my gratitude to my senior colleagues, Dr. Isaac S. Kohane and Dr. Merri L. Pendergrass for their council and encouragement during this project.

I am very grateful to Dr. Lee-jen Wei for his advice on statistical evaluation and to Dr. Nikheel Kolatkar for his help with manual chart review for validation of the index of non-compliance.

I would also like to thank Partners Research Computing and Hewlett-Packard staff, including Dennis Gurgul, Lance Davidow, and Werner Hahn for their unwavering IT support of my endeavors and tolerance of my occasionally successful attempts to bring the system down.

This research was supported in part by the NLM Training Grant 5-T15-LM-07092-11 and NLM grant 1U54LM008748-01

Contents

List of Figures and Tables

Figures

Tables

Introduction

The goal of this project was to develop a suite of software tools that could rapidly and accurately extract focused clinical information from free text physician notes in the electronic medical record.

Medical record contains many classes of information: laboratory results, radiology reports, billing data, encounter records, procedure reports, etc. Of all of these, physician notes are the most valuable because they document the conclusions of the integrated evaluation of the patient's health by that takes into account all other sources of information, both in- and outside of the medical record.

At the same time physician notes are frequently the least structured component of the medical record, and automatic extraction of specific information from them is technically challenging. A number of efforts have been made, most utilizing full or partial parsing of the syntactic structure of the text. Many of these suffer from low accuracy, and all are slow, costly (when commercialized) or unavailable to researchers outside of the designer institution (when academic).

Consequently, we aimed to develop a suite of software tools for extraction of clinical information from free text physician notes that would be characterized by the following:

1. High accuracy
2. High speed enabling enterprise-wide application
3. Low cost
4. Adaptability to different queries
5. Portability across different operating platforms

Specific Aims

The goal of this project was to develop and evaluate a suite of software tools which can identify patient cohorts with specific quantitative, semi-quantitative and qualitative clinical characteristics through analysis of free-text physician notes. This goal was accomplished through the following specific aims:

A. To develop and formally evaluate a tool that can identify the following quantitative patient characteristics through analysis of free-text physician notes:
   a) blood pressure
   b) height
   c) weight
   d) body mass index

B. To develop and formally evaluate a tool that can identify and quantify patient compliance – a complex but clinically important patient characteristic that has been hypothesized to play a crucial role in response to treatment in many diseases.

C. To develop and formally evaluate a tool that can identify patients with the following qualitative clinical characteristics through analysis of free-text physician notes:
   a) diabetes mellitus
   b) hypertension
   c) overweight

Background

A number of important applications in medicine and biomedical research require identification of large patient cohorts with specific clinical characteristics. These include, among others, quality of care surveillance[1], identification of prospective subjects for research studies[2] and clinical decision support[3].

Several approaches have been used to identify patients with specific conditions, including death certificates[4], billing data[5-7], and surveys[8]. Each of these methods has its own shortcomings and sensitivity remains relatively low. Consequently, manual chart review remains the gold standard for identification of individuals diagnosed with a particular disease. This is a labor-intensive and expensive process[9] that is not scalable to the level needed in a medium to large-size healthcare facility.

As most elements of the medical record are increasingly computerized, more data becomes available for computer-assisted analysis. In particular, physician notes are a very rich source of clinical information[10], and are now commonly available in digital format. However, the information in physician notes is unstructured and its analysis presents a technical challenge[11].

There have been a number of attempts to identify clinical characteristics of patients from the text of physician notes. Most of the early reports were characterized by low sensitivity and specificity[12-15]. More recently, both academic[16] and commercial[17] tools were reported to have attained high accuracy in identifying clinical concepts from free text. However, these tools require extensive manual training on the data set and are slow (1 to 30 seconds to process a note)[17, 18]. Additionally, commercially available software is expensive and most academic systems are not freely available to the public.

Consequently there is an acute need for an inexpensive, accurate and rapid method for identification of specific patient characteristics from unstructured medical data. In this proposal we aim to address this need by developing and formally evaluating PACT (PAtient Cohort identification through Text analysis) – a suite of software tools that can identify patient cohorts with specific clinical characteristics through analysis of free text physician notes.

Quality of care improvement and surveillance would be a particularly important application of at tool like PACT. As more advanced electronic medical record systems become available, health care facilities increasingly rely on structured patient health data to monitor patient outcomes and implement clinical decision support tools[11]. Additionally, as healthcare payors are starting to move towards pay-for-performance schemes where a fraction of reimbursement is dependent on the patients' clinical outcomes[19], availability of structured patient data can affect healthcare facilities financially as well.

We therefore selected for prototype development of PACT patient characteristics that are likely to have the greatest influence on clinical outcomes. Obesity and hypertension have reached epidemic proportions in the United States[20, 21]. Both of these are independent risk factors for cardiovascular disease – a number one cause of mortality in this country[22]. Obesity is also a major risk factor for diabetes mellitus – itself a leading cause of death[22], but also frequently leading to development of coronary artery disease, peripheral vascular disease, renal failure, and blindness. Together, obesity, hypertension and insulin resistance (the pathophysiologic mechanism behind the more common form of diabetes mellitus – type 2) are components of what has been termed "metabolic

syndrome" or "syndrome X" – an increasingly common phenomenon that has been shown to significantly increase both cardiovascular and overall mortality[23].

The choice of qualitative patient characteristics lead to the selection of the quantitative patient characteristics to be extracted. Knowledge of the patient's blood pressure measurements is necessary to diagnose and manage hypertension, and height and weight values are needed to calculate body mass index (seldom documented by itself) which is a major assessment variable for obesity.

Diabetes mellitus, hypertension and obesity are all chronic medical conditions that require daily lifelong treatment that includes both medications and lifestyle changes. Unsurprisingly, patient compliance with rigorous treatment regimens for these conditions is commonly less than perfect[24-27], and non-compliance has been linked to a deterioration in clinical outcomes[28], while successful interventions to improve treatment adherence have been shown to improve outcomes[29]. However, no effective method for assessing patient compliance currently exists. Most commonly used techniques include patient surveys and electronic medication dispensation devices, neither of which scales up well, or pharmacy records which are often fragmented across multiple vendors[30]. These approaches are difficult to use in research settings and impossible to implement in routine healthcare. We therefore included a tool that could rapidly obtain a measure of individual patient compliance with physician recommendations across an entire healthcare facility in the PACT suite.

Methods

Data

Patient data (physician notes and billing information) was obtained from Research Patient Data Registry (RPDR) – a database maintained by Partners Healthcare System which contains data on more than 1.5 million patients treated at Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH) in Boston, MA from 1993 until present.

Hardware and Software

PACT was implemented in Perl 5.6.1 to insure cross-platform compatibility. Perl regular expressions engine was used for the string and pattern matching detailed in the Design section. It was tested under Linux OS on a AMD Opteron 32/64 2.1 GHz system with 4 GB of RAM.

Design

The overall design of PACT is schematically represented in Figure 1. PACT takes as input one or more files that contain the text of all physician notes for the patient population being studied. It subsequently performs the following steps:

1. *Identification of individual notes*
   Note headers were used to identify individual notes.

2. *Identification and removal of duplicate documents*
   For each extracted document a unique document ID was created by appending the number of characters in the document to a sequence of 10 letters that were located in

positions L/10, 2*L/10, …, L where L is the number of characters in the document. Documents that were less than 10 characters in length were to be rejected; however, there were no such documents in our data set because each document had a header which alone was more than 10 characters in length.

3. *Identification of individual sentences*
   A boundary between two sentences was identified if the following sequence of characters was detected:
   a. a whitespace character
   b. a sequence of alphanumeric (A..Z, a..z, 0..9, _) characters which had to contain either a vowel or a digit
   c. a period, exclamation mark, or a question mark
   d. one or more whitespace characters
   e. any character but a lowercase letter

   While this method of sentence boundary detection was less than perfectly accurate due to high prevalence of abbreviations and frequent absence of punctuation in physician notes, it was empirically found to be sufficient.

4. *Location of word tags specific for the patient characteristic in the sentences*
   For each of the patient characteristics we have compiled a list of *word tags* - morphemes, words, phrases, or acronyms that are specific for the characteristic. These included the full name of the characteristic, related adjectives and acronyms, medications and procedures used exclusively to treat the disease, etc. The list of word tags for each of the patient characteristics we sought to identify is given in Table 1.

5. *Identification of the specific word + numeric pattern for quantitative patient characteristics*
   The following numeric patterns were required to identify quantitative patient characteristics:
   a. Blood pressure
      The following pattern was required to identify blood pressure value in the text of a note:
      - two or three digits followed by forward slash followed by two digits
      - this pattern has to follow the blood pressure word tag
      - the number before the forward slash (systolic blood pressure) has to be larger than the number after the forward slash (diastolic blood pressure)
      - this pattern cannot be either preceded or followed by a forward slash (as in a date)
      - systolic blood pressure has to be between 50 and 300
      - diastolic blood pressure has to be between 20 and 200
   b. Height
      The following pattern was required to identify height in the text of the note:
      - height word tag (*height / hgt /ht*) followed by
      - *today* (optional) followed by
      - *is / was / of* (optional) followed by

- *X'XX''* or *X feet XX inches* or *XX inches* or *XXX cm* where 'X' represents a digit

c. Weight
The following pattern was required to identify weight in the text of the note:
- weight word tag (*weight / wght / wt*) followed by
- *today* (optional) followed by
- *is / was / of* (optional) followed by
- two or three digits

d. Body Mass Index (BMI)
The following pattern was required to identify BMI:
- BMI word tag (*BMI / body mass index*) followed by
- *today* (optional) followed by
- *is / was / of* (optional) followed by
- two optionally followed by a period with 1-2 digits

6. *Identification of units of measurement for quantitative patient characteristics which can be expressed in more than one unit*
Both height and weight can be expressed in several different units: meters, centimeters, feet, or inches for height, and kilograms vs. pounds for weight. We empirically found that units are documented for the majority of height values. We therefore required that the units be present in the sentence in order to be able to quantify a patient's height. On the other hand, units were frequently not documented for weight values. Therefore PACT recorded a weight value if one of the following criteria was met:
   a. Units were documented for that specific weight value
   b. Units were not documented for that specific weight value but weight units were documented elsewhere in the note apparently referring to the specific weight value (e.g. 'Patients weight today is 120 and is *down by 20 lbs*').
   c. Units were not documented for that specific weight value but the same note had a height value with documented units and a BMI value, from which the units for weight could be calculated
   d. Units were not documented for that specific weight value but were documented for the same patient in other notes. In this case the mean value of the patient's recorded weights was calculated using weight values with documented units from other notes (which were converted to kilograms when recorded) and compared to the current value. If the difference between the mean of recorded weight values (which were all in kilograms) and the current value was less than 25%, kilograms were assumed as units. Otherwise, if the difference between the mean of recorded weight values and the current value multiplied by 0.4536 was less than 25%, pounds were assumed as units.

7. *Identification of negative qualifiers in the sentences*
For each of the qualitative and semi-quantitative patient characteristics we compiled a list of *negative qualifiers* – words or phrases which, when encountered in the same sentence as a disease word tag, made it unlikely that the sentence asserted that the

patient had the disease. We have identified four main categories of negative qualifiers:

a) references to another disease (e.g. diabetes *insipidus*)
b) family history
c) non-confirmations (e.g. *rule out* hypertension)
d) negations

The first category of negative qualifiers is characteristic-specific (Table 1) while the other three are common for all qualitative and semi-quantitative characteristics (Table 2).

8. *Output*
   a) Quantitative characteristics
      PACT determines the value of each of the quantitative characteristics for each note where they were documented as well as the average value of each of the quantitative characteristics over a given period of time. If several different values were reported for the same quantitative characteristic in one note, all were recorded. The exception from this rule was blood pressure which can vary depending on the patient's anxiety, position, etc.; therefore the lowest blood pressure value for each note was recorded.
   b) Semi-quantitative characteristics
      PACT calculates an "index of non-compliance" which is the ratio of the number of times word tags specific for non-compliance were found in the patient's physician notes to the number of notes.
   c) Qualitative characteristics
      PACT derives the conclusion about presence or absence of a qualitative patient characteristic (e.g. diagnosis of a disease) based on the number of times word tags specific for that characteristic were found in the patient's physician notes as well as on the average value of a related quantitative characteristic (if applicable). The thresholds required to make a conclusion with respect to a particular characteristic were determined empirically and are detailed in Table 3.

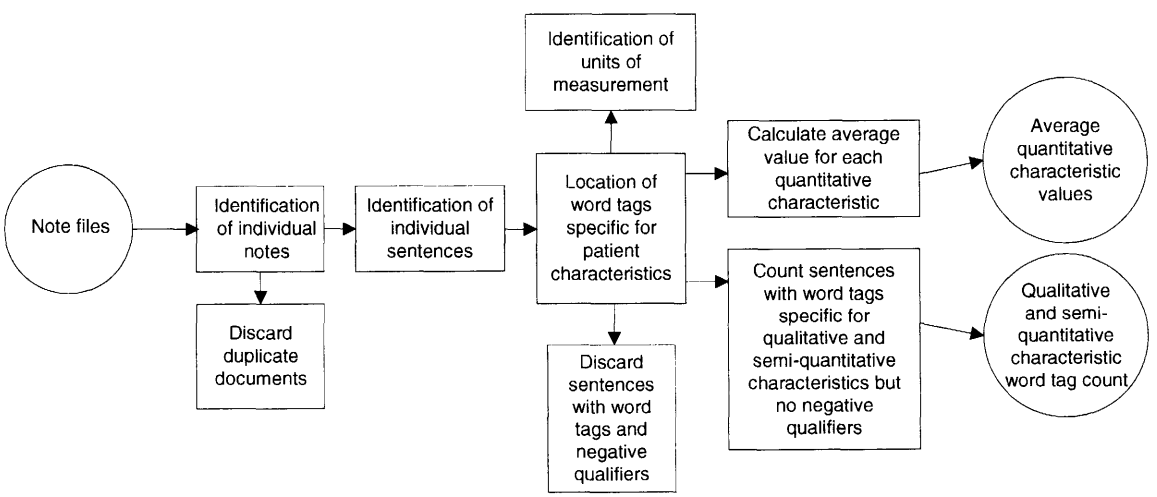

Figure 1. Schematic representation of the PACT design

Table 1.
Patient Characteristic-Specific Word Tags and Negative Qualifiers.

Patient Characteristic	Word Tags	Characteristic-specific negative qualifiers
Blood pressure	blood pressure BP B/P Vital signs[1] V.S.[1] Physical exam(ination)[1] exam[1] xm[1]	N/A
Height	height hgt ht	N/A
Weight	weight wght wt	N/A
Body Mass Index	body mass index BMI	N/A
Hypertension	hypertens* HTN high [BP] elevated [BP]	pregnancy pregnant [pulm] [HTN] portal [HTN] Endocrine-Hypertension[2]

[1] It was empirically found that many physician notes documented a blood pressure in a phrase like "Vital signs: 120/70, hr 68, wt 135". We hypothesized – and empirically confirmed – that a finding of two or three digits followed by a forward slash followed by two digits is sufficiently specific not to require an explicit declaration of a "blood pressure" tag in presence of a more general "vital signs" or "physical examination" tag.

[2] "Endocrine-Hypertension" is the name of a Division of the Department of Medicine at the Brigham and Women's Hospital. Physicians at this Division see many patients with diagnoses other than hypertension (i.e. patients with endocrine problems) and therefore it is not specific for diagnosis of hypertension.

Patient Characteristic	Word Tags	Characteristic-specific negative qualifiers
Overweight	obes* overweight high BMI elevated BMI increased BMI gastric bypass gastric banding sibutramine	N/A
Treatment Compliance	non-compliant no show DNKA[3] did not show did not come did not keep appointment missed appointment did / does not wish did / does not want declined refused not taking did not take not interested is reluctant uncooperative	N/A

[3] Acronym for "Did Not Keep Appointment"

Patient Characteristic	Word Tags	Characteristic-specific negative qualifiers
Diabetes mellitus	diabet* *IDDM glyburide micronase diabeta prestab glynase glucovance glipizide glucotrol metaglip glimepiride amaryl chlorpropamide diabinese tolbutamide orinase toltab talazamide tolinase repaglinide prandin nateglinide starlix acarbose precose miglitol glyset rosiglitazone actos lantus glargine glucometer	insipidus gestational pregnancy pregnant

All word tags are case insensitive.
* wildcard expansion allowed in this direction
[HTN] can be either "HTN" or "hypertension"
[BP] can be either "BP" or "blood pressure"
[pulm] can be either "pulm" or "pulmonary"

Table 2.
Negative Qualifiers Common for All Qualitative Patient Characteristics

Family history	Non-confirmation	Negation
FH*	work up for	no
family	check for	not
parents	at risk for	doesn't
daughter	screen	didn't
son	rule out	denies
children		unknown
child		N/A
sibling		negative
brother		unlikely
mother		
father		
sister		
nephew		
niece		

Table 3.
Criteria Used to Establish Diagnoses Using PACT Data

Disease	Criteria
Diabetes mellitus	WT ≥ 2
Hypertension	(WT ≥ 1 OR BP ≥ 1) AND (WT+BP) ≥ 2
Overweight	WT ≥ 1 OR BMI ≥ 1 OR WE ≥ 1

WT: Number of sentences with disease word tags but without negative qualifiers required to establish diagnosis of the disease

BP: Number of blood pressure measurements where either systolic blood pressure ≥140 or diastolic blood pressure ≥90

BMI: Number of sentences with body mass index reported > 24.9

WE: Number of sentences with weight reported greater than that which would result in a BMI > 24.9 in a person who was less than two standard deviations above the mean height for the gender.

Evaluation

Performance of PACT with respect to quantitative and semi-quantitative patient characteristics was compared to manual chart review as no other established computational method for identification of these characteristics currently exists. Correlation of semi-quantitative characteristics with patient outcomes was also determined.

Performance of PACT with respect to qualitative patient characteristics was compared to two existing standards: manual chart review (the "gold standard") and

computational ascertainment of diagnosis from billing codes (the currently used method for identification of large patient cohorts with a particular diagnosis).

1. *Quantitative Characteristics*

   150 physician notes (the number used in previous reports[18]) were randomly selected to assess the sensitivity and specificity of PACT identification of blood pressure. 150 nutritionist notes were randomly selected to assess the sensitivity and specificity of PACT identification of height, weight and body mass index. The nutritionist notes were used because the low frequency of documentation of height, weight and body mass index in the notes of physicians of other specialties would have decreased the power of the study. Manual review by a board certified internist (Alexander Turchin) blinded to the results of PACT was used as the gold standard for calculation of sensitivity and specificity.

2. *Semi-Quantitative Characteristics*

   The evaluation of PACT with respect to extraction of a semi-quantitative patient characteristic (non-compliance with physician recommendations) from physician notes was performed on three levels: individual sentences, patient notes, and relationship with clinical outcomes.

   For sentence-level validation 150 sentences were randomly selected from the set of sentences where PACT detected a word tag specific for patient non-compliance with physician recommendations. These sentences were manually reviewed by a board-certified internist (Alexander Turchin) who either confirmed or rejected the finding of documentation of patient non-compliance.

   For note-level validation 30 patient records were randomly selected to assess the accuracy of PACT quantification of patient compliance. 10 notes were randomly selected from each record to be scanned by PACT (all notes were selected if the record had a total of 10 or fewer notes). The same notes were also manually reviewed by a board certified internist who did not participate in the design of the program, and who was blinded to the results of PACT. The reviewer recorded the number of events of patient non-compliance with physician recommendations documented in each note. Correlation between the average number of word tags specific for patient non-compliance per note as determined by PACT with the number of events of patient non-compliance per note abstracted by the manual review was estimated using linear regression.

   For clinical outcome-level validation index of non-compliance was calculated for 4,415 patients who had a primary care physician at one of four Brigham and Women's Hospital outpatient clinics who had either one billing code of diabetes mellitus (ICD-9-CM code of 250.xx), one serum glucose value > 199 mg/dL or one hemoglobin A1C measurement of any value; patients with more than one billing code of a neoplasm (ICD-9-CM codes between 140.xx and 239.xx) were excluded[4]. These patients were then stratified into quartiles according to the value of their index of non-compliance and the average number of annual Emergency Department visits was calculated for each quartile using the encounter data.

[4] Our hypothesis was that patients who do not follow their physicians recommendations are more likely to visit the Emergency Department. This relationship was unlikely to hold for patients with malignant neoplasms as their clinical course is frequently suboptimal despite maximum medical interventions.

3. *Qualitative Characteristics*

In order to increase the power of evaluation of sensitivity and specificity of PACT results we used patient populations at high risk for the characteristics that PACT identifies. Specifically, for evaluation of PACT identification of patients with a documented diagnosis of diabetes, hypertension or overweight we used a population of all adult patients of four primary care practices at BWH with at least one of the following:

a. A billing ICD-9-CM code of 250.xx (diabetes mellitus)
b. A serum glucose measurement ≥ 200 mg/dL
c. A measurement (of any value) of hemoglobin A1C level

While each of these criteria identifies patients at high risk for diabetes mellitus they are not highly specific, and the overall prevalence of diabetes mellitus in this population was c. 30%.

The thresholds for identification of patients with each of the diagnoses using PACT were determined empirically and are detailed in Table 3. At least two ICD-9-CM billing codes over a two-year period (2002-2003) were required to establish the diagnoses using billing code data – a commonly accepted standard[31]. The following billing ICD-9-CM codes were used to establish the diagnoses:

a. Diabetes mellitus: 250.xx
b. Hypertension: 401.xx
c. Overweight: 278.0x

For each of the diseases, 150 patients records were randomly selected for comparison between PACT, billing code data and manual review. The patient records were reviewed by a board-certified internist (Alexander Turchin) who was blinded to the results of both PACT and billing code analysis. To ensure comparability between PACT and billing code data, only notes over the same two-year period were used for PACT analysis. The entire chart (including physician-maintained problem list not scanned by PACT as well as notes outside of the date range studied) was used in manual review to determine whether the patient had the disease during the period of the study to establish the "gold standard". For example, if the diagnosis of diabetes or hypertension was documented in the notes in 2001 (i.e. one year prior to the beginning of the study period) but not in the notes in 2002-2003, the patient was still recorded as having the diagnosis because neither of this conditions typically resolves over one year except under extraordinary circumstances.

Statistics

Kappa statistic was used to evaluate agreement between the PACT analysis and manual chart review[32]. McNemar's test was used to determine statistical significance of the difference in performance between PACT and billing code analysis with respect to qualitative patient characteristics[33]. Linear regression was used to estimate correlation between PACT index of non-compliance and the results of manual chart review. Student's t-test was used to determine the statistical significance of the difference between average number of annual Emergency Department visits for patients in different quartiles with respect to index of non-compliance.

IRB

The study protocol was reviewed and approved by Partners Human Research Committee.

Results

Quantitative Characteristics

Speed

To determine the speed of the PACT vitals tool, the tool was used to scan notes of primary care providers of patients with diabetes. PACT processed 293,037 notes totaling 10,713,499 lines and 476.1 MB[5] in 1,613 seconds, with an average speed of 6.5×10^5 notes / hour or 1.04 GB of text / hour.

Blood Pressure Accuracy

The breakdown of validation of PACT extraction of blood pressure by manual chart review is given in Figure 2. Both PACT and manual chart review detected the same lowest blood pressure value in 52.3% of the notes; PACT reported a different lowest blood pressure value from manual chart review in 2.7% of the notes; PACT did not detect blood pressure reported by the manual chart review in 1.7% of the notes; both PACT and manual chart review did not detect a blood pressure value in 43.3% of the notes. There were no cases of a false positive report of a blood pressure value by PACT. Overall PACT and manual chart reviewed gave the same results for 95.5% of the notes. PACT sensitivity for detection of blood pressure was 95.2% (95% confidence interval 92.7% to 97.6%) and specificity 100% (95% confidence interval 99.4% to 100%). κ statistic for agreement between PACT and manual chart review was 0.91 ($p < 0.0001$ that $\kappa > 0.8$) which represents excellent agreement.

Height Accuracy

The breakdown of validation of PACT extraction of height values by manual chart review is given in Figure 3. Both PACT and manual chart review detected the same height value in 32.3% of the notes; PACT did not detect height reported by the manual chart review in 4.4% of the notes; PACT reported a height value not actually documented in the note in 0.6% of the notes; both PACT and manual chart review did not detect a height value in 62.7% of the notes. There were no cases of a disagreement on the actual value between PACT and manual chart review when both reported a height value. Overall PACT and manual chart reviewed reported the same results for 94.5% of the notes. PACT sensitivity for detection of height was 87.9% (95% confidence interval 82.9% to 93.0%) and specificity 99.0% (95% confidence interval 97.5% to 100%). κ statistic for agreement between PACT and manual chart review was 0.89 ($p = 0.01$ that $\kappa > 0.8$) which represents excellent agreement.

[5] 2^{20} bytes was used as the definition of one MB (megabyte) and 2^{30} bytes was used as the definition of one GB (gigabyte)

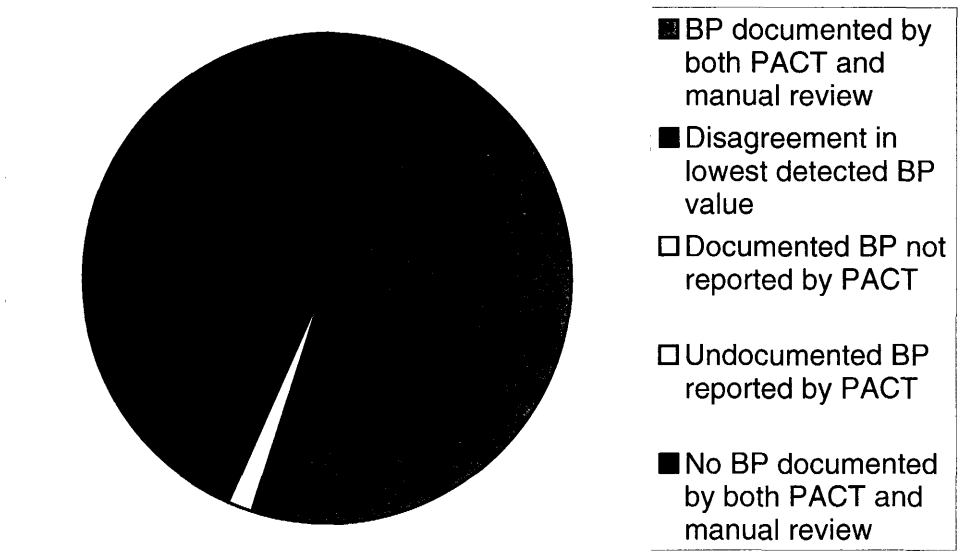

Figure 2.
Validation of blood pressure extraction by PACT by manual chart review

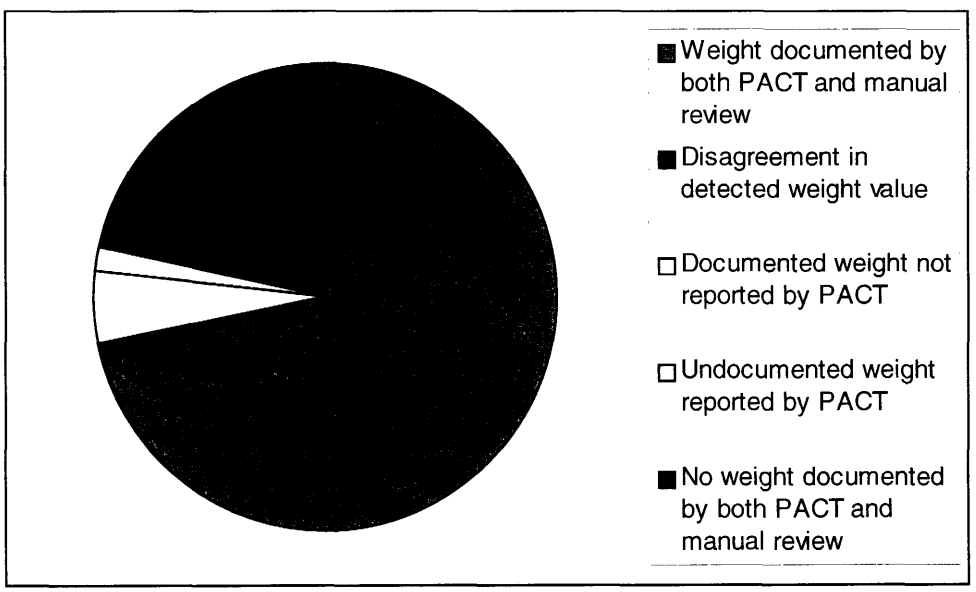

Figure 3.
Validation of height extraction by PACT by manual chart review

Weight Accuracy

The breakdown of validation of PACT extraction of weight values by manual chart review is given in Figure 4. Both PACT and manual chart review detected the same weight value in 71.3% of the notes; PACT reported a different weight value from the one abstracted by manual chart review in 1.3% of the notes; PACT did not detect weight reported by the manual chart review in 5.0% of the notes; PACT reported a weight value not actually documented in the note in 1.9% of the notes; both PACT and manual chart

review did not detect a weight value in 21.9% of the notes. Overall PACT and manual chart reviewed reported the same results for 91.9% of the notes. PACT sensitivity for detection of weight was 91.8% (95% confidence interval 87.6% to 96.1%) and specificity 92.1% (95% confidence interval 87.9% to 96.3%). κ statistic for agreement between PACT and manual chart review was 0.79 which represents excellent agreement.

BMI Accuracy

The breakdown of validation of PACT extraction of BMI values by manual chart review is given in Figure 5. Both PACT and manual chart review detected the same weight value in 20.9% of the notes; PACT reported a different weight value from the one abstracted by manual chart review in 1.3% of the notes; both PACT and manual chart review did not detect a BMI value in 79.1% of the notes. There were no cases of disagreement between PACT and manual chart review. PACT sensitivity for detection of BMI was 100% (95% confidence interval 98.8% to 100%) and specificity 100% (95% confidence interval 98.8 to 100%). κ statistic for agreement between PACT and manual chart review was 1.0 which represents excellent agreement.

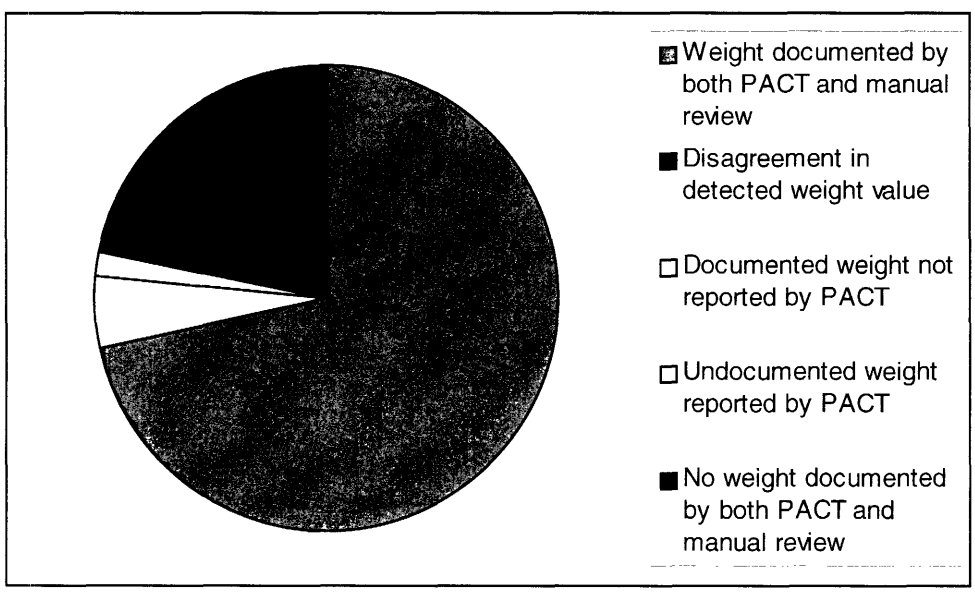

Figure 4.
Validation of weight extraction by PACT by manual chart review

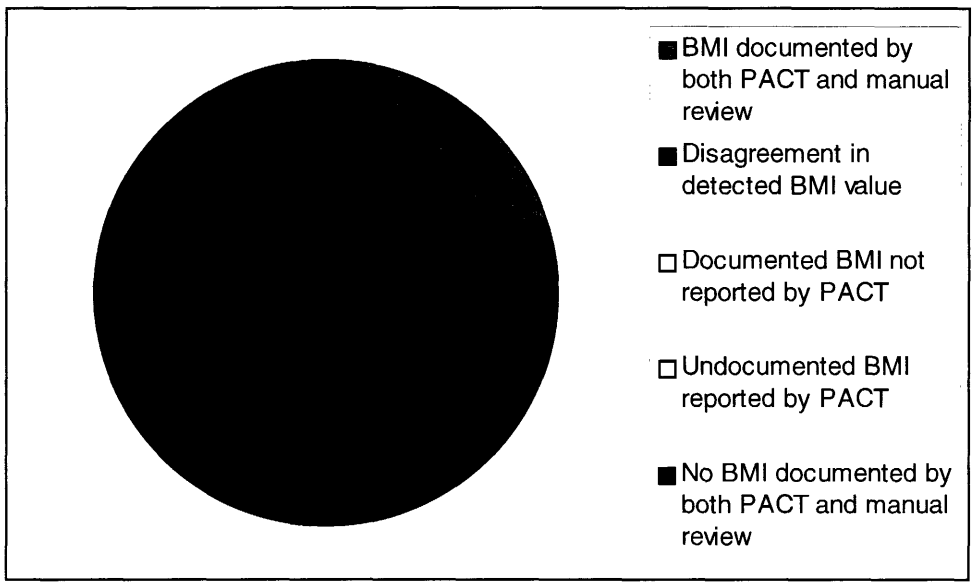

Figure 5.
Validation of BMI extraction by PACT by manual chart review

Semi-Quantitative Characteristics

Speed

To evaluate the speed of the PACT non-compliance tool, the tool was used to scan notes of primary care providers of patients with diabetes. PACT processed 293,037 notes totaling 10,713,499 lines and 476.1 MB in 1193 seconds, with an average speed of 8.8×10^5 notes / hour or 1.4 GB of text / hour.

Accuracy

On the sentence level, 150 randomly selected sentences reported by PACT to contain word tags specific for patient non-compliance with physician recommendations were reviewed. Documentation of non-compliance was confirmed on review for 140 sentences, resulting in a positive predictive value of 93.3% (95% confidence interval 89.3% to 97.3%).

On the patient level, a total of 286 notes of 30 randomly selected patients were manually reviewed in a blinded fashion. Pearson correlation coefficient between the PACT-generated index of non-compliance and manually abstracted number of documented instances of non-adherence to physician recommendations per note was 0.64 ($p = 0.0001$).

Correlation with Outcomes

4,415 patients in four primary care clinics of Brigham and Women's Hospital were stratified into quartiles according to the PACT index of non-compliance. Patients in the first quartile had less than 0.055 non-compliance word tags per note and patients in the fourth quartile had more than 0.203 non-compliance word tags per note. The annual frequency of Emergency Department visits was then calculated for patients in each

quartile (Figure 6). Patients in the fourth quartile were found to have almost twice as many annual ED visits (1.15 visits / year) than patients in the first quartile (0.65 visits / year); this difference was highly statistically significant ($p = 7 \times 10^{-18}$).

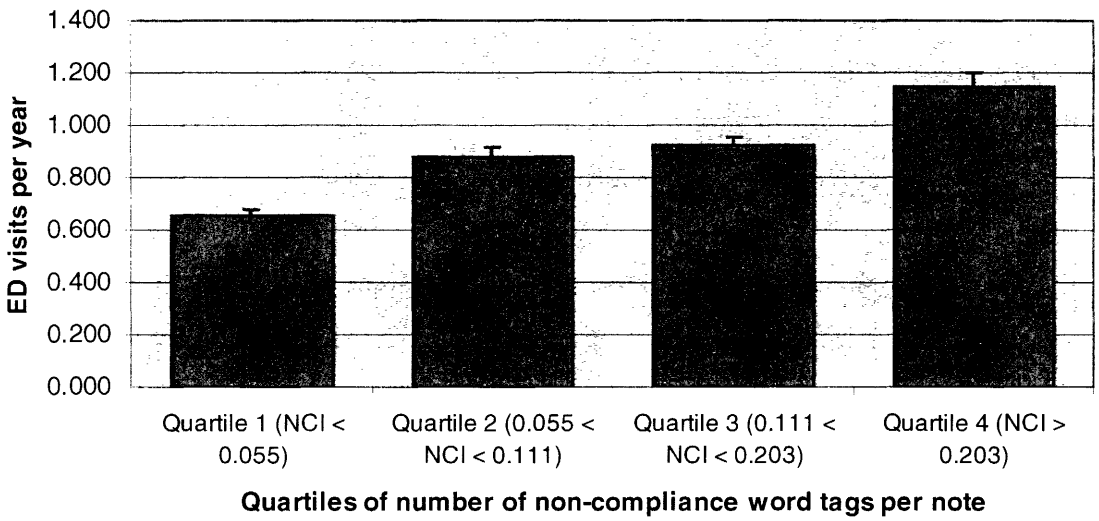

Figure 6.
Distribution of frequency of Emergency Department visits by quartiles of the number of non-compliance word tags per note.

Qualitative Characteristics

Speed

The speed of diagnosis detection by PACT was assessed on the example of the diabetes tool. PACT processed 2,295,461 notes totaling 102,832,425 lines and 4.3 GB in 11679 seconds, with an average speed of 7.1×10^5 notes / hour or 1.3 GB of text / hour.

Accuracy

The results of the evaluation of diagnosis identification by PACT and billing data as compared to the manual chart review are found in Table 4. Sensitivity of PACT ranged between 74.2 and 96.2% and was invariably substantially higher than that of billing data analysis. Specificity of PACT ranged from 85.9% to 100% and was usually the same or comparable to the billing data analysis.

κ statistics for agreement between PACT-based diagnosis and manual chart review are listed in Table 5. The agreement for diabetes and hypertension had κ statistics > 0.75 (excellent agreement) while agreement for overweight had a κ of 0.67 (substantial agreement).

Table 4.
Diagnosis Identification by PACT vs. Billing Data Analysis

Disease	Test	PACT	Billing	p value
Diabetes	Sensitivity	96.2%	76.9%	0.024
	Specificity	98.0%	98.0%	
Hypertension	Sensitivity	90.7%	74.4%	0.078
	Specificity	85.9%	92.2%	
Overweight	Sensitivity	74.2%	14.4%	<0.001
	Specificity	100%	100%	

Table 5.
Agreement between PACT and Manual Chart Review

Disease	κ	p-value that κ > 0.75
Diabetes	0.94	< 0.001
Hypertension	0.77	0.053
Overweight	0.67	N/A

Discussion

The goal of this project was to respond to an acute need in Medical Informatics – to develop a tool that could rapidly, reliably and cheaply extract clinical information from free text physician notes. Clinical investigators and hospital administrators in need of structured patient data that could be used for research, quality of care monitoring or clinical decision support, often find themselves in the position of Coleridge's Ancient Mariner: there is "water, water everywhere, nor any drop to drink." The information they need is at their fingertips: hospitals and outpatient clinics throughout the country are sitting on terabytes of invaluable medical data locked in the free text of physician notes, unable to use it due to a void in technology.

PACT successfully fills this void. We have demonstrated in this project that it is possible to develop a suite of easily portable and adaptable software tools that can extract clinical data from free text documents with high speed and accuracy.

A number of other processors for free-text clinical documents have been developed. These have primarily followed two main technical approaches: Boolean keyword search[34, 35] and natural language parsers[16, 17, 36]. The Boolean keyword search approach can be used for focused queries. It is very fast, particularly with modern database engines. However, it lacks any information about the remainder of the clinical document and therefore typically unable to process qualifiers (e.g. negations). It has been applied successfully to relatively highly structured documents like radiology reports (which have an "Impression" section that summarizes the radiologist's conclusion)[34] but its performance on more complex clinical records has been less optimal[15, 35]. Natural language parsers are typically implemented using symbolic methods based on a combination of syntactic and semantic grammars. Their undisputed advantage is the ability to extract all information from the document not limited by a particular query. On the other hand, the sheer complexity of this task frequently leads to lower accuracy; they

are also usually slow, rendering them inadequate for operational use, and expensive (if commercial) or unavailable to users other than their designers (if academic).

In developing PACT we sought to draw on the advantages of both techniques. Similarly to Boolean keyword search systems, PACT is designed to answer specific queries. This allows us to avoid complete syntactic and semantic parsing and leads to a over 100-fold advantage in speed over the natural language parsers. Unlike the Boolean keyword search systems, PACT uses sentences rather than the entire clinical documents as the search units. This permits detection of negative qualifiers – a technique that has previously only been implemented in the natural language parsers, and leads to a substantially improved accuracy[16]. Additionally, programmatic control of Perl and greater compactness, readability and flexibility of Perl regular expressions compared to Boolean syntax allowed queries for very complex patterns representing different possible syntactic structures of the same message – a feature, again, previously only possible in natural language parsers.

We have tested the PACT paradigm on three classes of clinical data: quantitative (e.g. vital signs), semi-quantitative (patient compliance) and qualitative (presence of a particular diagnosis). PACT performed well in all three categories.

Sensitivity and specificity of detection of quantitative clinical data varied from 88% to 100%, and from 92% to 100% respectively. Mistakes were usually made in syntactically complex sentences, for example "*Ht and Today's Wt (vital signs): 74.75 in/198#*".

PACT analysis of semi-quantitative clinical data – index of patient non-compliance – was validated on three levels: sentence-wise, patient-wise and correlation with clinical outcomes. Positive predictive value of non-compliance concept identification by PACT at the sentence level was 93%; index of non-compliance calculated by PACT correlated highly with the number of documented instances of non-compliance abstracted from the patient record by an independent reviewer. Importantly, patients with a higher index of non-compliance calculated by PACT were shown to have 50% more Emergency Department visits per year than patients with lower index of non-compliance, confirming clinical relevance of the data extracted from the physician notes.

Performance of PACT in identification of patients with a particular diagnosis was to a significant extent related to underlying clinical characteristics of the disease. PACT performed best on the example of diabetes mellitus – a relatively well-defined, constant and treatable condition that can lead to a large number of complications and is therefore well documented in physician notes. Hypertension represents as substantially more labile condition whose main parameter – blood pressure – can be non-specifically elevated in cases of physical or emotional distress, leading in turn to decreased specificity of documentation of the diagnosis in the notes. Obesity, on the other hand, is characterized by the lack of effective treatment options. Consequently, despite the ease of diagnosis and the potential for numerous complications, it is seldom addressed during patients' visits[37] leading to the decreased sensitivity of PACT analysis. However, while imperfect, detection of the diagnosis of obesity by PACT compares very favorably to the traditional method of diagnosis detection – billing codes – since a disease that cannot be treated is billed for even less frequently than it is mentioned in the notes.

The ultimate goal of the development of PACT was to design a tool that could be also be used by other institutions and possibly for other queries. PACT was written in

Perl – an interpreted programming language that has been ported to all major operating systems (UNIX / Linux, Windows and Mac OS). It therefore requires minimal changes (mostly related to the format of the record separators in the text file containing physician notes and possibly accommodations for local differences in the medical vernacular[38]) for implementation in a different healthcare facility. Adaptation to a different query should also involve only small changes to the algorithm (primarily to the characteristic-specific negative qualifiers).

While a significant improvement over the existing techniques, PACT has its limitations. It is well known that a large fraction of patients with diabetes and hypertension have not had their diagnosis established medically, and therefore not documented in the notes; these patients would be missed by the tool. PACT is also not suitable for simultaneous identification of multiple diagnoses; this task is best handled by tools that implement comprehensive ontologies and complete lexical and syntactic parsing.

In summary, PACT is a highly accurate, rapid tool for identification of patient characteristics through analysis of free text physician notes in the electronic medical record. PACT can be used for investigations of quality of care in a specific healthcare facility, by researchers looking to identify potential subjects for a study, or to provide basis for clinical decision support. PACT can be easily adapted to a different healthcare facility or for detection of a different disease. It represents an important advance in the field, and we plan to continue to develop this concept further to improve its performance and functionality.

References

1. Saydah SH, Geiss LS, Tierney E, Benjamin SM, Engelgau M, Brancati F. Review of the performance of methods to identify diabetes cases among vital statistics, administrative, and survey data. *Ann Epidemiol.* Aug 2004;14(7):507-516.
2. Schubart JR, Einbinder JS. Evaluation of a data warehouse in an academic health sciences center. *Int J Med Inform.* Dec 2000;60(3):319-333.
3. Hoelzer S, Boettcher H, Schweiger RK, Konetschny J, Dudeck J. Presentation of problem-specific, text-based medical knowledge: XML and related technologies. *Proc AMIA Symp.* 2001:259-263.
4. Sasaki A, Horiuchi N, Hasegawa K, Uehara M. The proportion of death certificates of diabetic patients that mentioned diabetes in Osaka District, Japan. *Diabetes Res Clin Pract.* Jun 1993;20(3):241-246.
5. Hebert PL, Geiss LS, Tierney EF, Engelgau MM, Yawn BP, McBean AM. Identifying persons with diabetes using Medicare claims data. *Am J Med Qual.* Nov-Dec 1999;14(6):270-277.
6. Goldstein LB. Accuracy of ICD-9-CM coding for the identification of patients with acute ischemic stroke: effect of modifier codes. *Stroke.* Aug 1998;29(8):1602-1604.
7. Raiford DS, Perez Gutthann S, Garcia Rodriguez LA. Positive predictive value of ICD-9 codes in the identification of cases of complicated peptic ulcer disease in the Saskatchewan hospital automated database. *Epidemiology.* Jan 1996;7(1):101-104.
8. Haapanen N, Miilunpalo S, Pasanen M, Oja P, Vuori I. Agreement between questionnaire data and medical records of chronic diseases in middle-aged and elderly Finnish men and women. *Am J Epidemiol.* Apr 15 1997;145(8):762-769.
9. McDonald CJ, Overhage JM, Dexter P, Takesue BY, Dwyer DM. A framework for capturing clinical data sets from computerized sources. *Ann Intern Med.* Oct 15 1997;127(8 Pt 2):675-682.
10. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. *J Am Med Inform Assoc.* Mar-Apr 2003;10(2):115-128.
11. McDonald CJ. The barriers to electronic medical record systems and how to overcome them. *J Am Med Inform Assoc.* May-Jun 1997;4(3):213-221.
12. Cooper GF, Miller RA. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. *J Am Med Inform Assoc.* Jan-Feb 1998;5(1):62-75.
13. Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. *J Am Med Inform Assoc.* Jan-Feb 2001;8(1):80-91.
14. Lowe HJ, Antipov I, Hersh W, Smith CA, Mailhot M. Automated semantic indexing of imaging reports to support retrieval of medical images in the multimedia electronic medical record. *Methods Inf Med.* Dec 1999;38(4-5):303-307.
15. Berrios DC. Automated indexing for full text information retrieval. *Proc AMIA Symp.* 2000:71-75.

16. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc.* Sep-Oct 2004;11(5):392-402.

17. Mamlin BW, Heinze DT, McDonald CJ. Automated extraction and normalization of findings from cancer-related free-text radiology reports. *AMIA Annu Symp Proc.* 2003:420-424.

18. Hripcsak G, Austin JH, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology.* Jul 2002;224(1):157-163.

19. Rosenthal MB, Fernandopulle R, Song HR, Landon B. Paying for quality: providers' incentives for quality improvement. *Health Aff (Millwood).* Mar-Apr 2004;23(2):127-141.

20. Mokdad AH, Serdula MK, Dietz WH, Bowman BA, Marks JS, Koplan JP. The spread of the obesity epidemic in the United States, 1991-1998. *Jama.* Oct 27 1999;282(16):1519-1522.

21. Stamler J. Setting the TONE for ending the hypertension epidemic. *Jama.* Mar 18 1998;279(11):878-879.

22. Leading causes of death and numbers of deaths, according to sex, race, and Hispanic origin: United States, 1980 and 2001. *National Center for Health Statistics.* [WWW]. Available at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/Health_US/hus03/Table031.xls. Accessed 10/22/04, 2004.

23. Lakka HM, Laaksonen DE, Lakka TA, et al. The metabolic syndrome and total and cardiovascular disease mortality in middle-aged men. *Jama.* Dec 4 2002;288(21):2709-2716.

24. Cramer JA. A systematic review of adherence with medications for diabetes. *Diabetes Care.* May 2004;27(5):1218-1224.

25. Rosen MI, Rigsby MO, Salahi JT, Ryan CE, Cramer JA. Electronic monitoring and counseling to improve medication adherence. *Behav Res Ther.* Apr 2004;42(4):409-422.

26. Feldman R, Bacher M, Campbell N, Drover A, Chockalingam A. Adherence to pharmacologic management of hypertension. *Can J Public Health.* Sep-Oct 1998;89(5):I16-18.

27. Fodor G, Cutler H, Irvine J, Ramsden V, Tremblay G, Chockalingam A. Adherence to non-pharmacologic therapy for hypertension: problems and solutions. *Can J Public Health.* Sep-Oct 1998;89(5):I12-15.

28. Morris AD, Boyle DI, McMahon AD, Greene SA, MacDonald TM, Newton RW. Adherence to insulin treatment, glycaemic control, and ketoacidosis in insulin-dependent diabetes mellitus. The DARTS/MEMO Collaboration. Diabetes Audit and Research in Tayside Scotland. Medicines Monitoring Unit. *Lancet.* Nov 22 1997;350(9090):1505-1510.

29. Friedman RH, Kazis LE, Jette A, et al. A telecommunications system for monitoring and counseling patients with hypertension. Impact on medication adherence and blood pressure control. *Am J Hypertens.* Apr 1996;9(4 Pt 1):285-292.

30. Vermeire E, Hearnshaw H, Van Royen P, Denekens J. Patient adherence to treatment: three decades of research. A comprehensive review. *J Clin Pharm Ther.* Oct 2001;26(5):331-342.

31. *HEDIS 2005. Volume 2: Technical Specifications.* Vol 2: National Committee on Quality Assurance.; 2005.

32. Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement.* 1960;20:37-46.

33. Hawass NE. Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients. *Br J Radiol.* Apr 1997;70(832):360-366.

34. Thomas BJ, Ouellette H, Halpern EF, Rosenthal DI. Automated computer-assisted categorization of radiology reports. *AJR Am J Roentgenol.* Feb 2005;184(2):687-690.

35. Aronsky D, Kasworm E, Jacobson JA, Haug PJ, Dean NC. Electronic screening of dictated reports to identify patients with do-not-resuscitate status. *J Am Med Inform Assoc.* Sep-Oct 2004;11(5):403-409.

36. Christensen L. HP, Fiszman M. MPLUS: a probabilistic medical language understanding system. Paper presented at: Workshop on Natural Language Processing in the Biomedical Domain, 2002.

37. Galuska DA, Will JC, Serdula MK, Ford ES. Are health care professionals advising obese patients to lose weight? *Jama.* Oct 27 1999;282(16):1576-1578.

38. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods Inf Med.* Jan 1998;37(1):1-7.