(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 May 2003 (15.05.2003)

PCT

(10) International Publication Number
WO 03/040879 A2

(51) International Patent Classification[7]: G06F

(72) Inventors: **RAO, R., Bharat**; 2060 St. Andrews Drive, Berwyn, PA 19312 (US). **SANDILYA, Sathyakama**; 28-12 Phesant Hollow Drive, Plainsboro, NJ 08536 (US).

(74) Agents: **PASCHBURG, Donald, B.** et al.; Siemens Corporation, Intellectual Property Dept., 186 Wood Ave. South, Iselin, NJ 08830 (US).

(21) International Application Number: PCT/US02/35303

(22) International Filing Date:
4 November 2002 (04.11.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/335,542 2 November 2001 (02.11.2001) US

(71) Applicant: **SIEMENS MEDICAL SOLUTIONS USA, INC.** [US/US]; 51 Valley Stream Parkway, Malvern, PA 19355 (US).

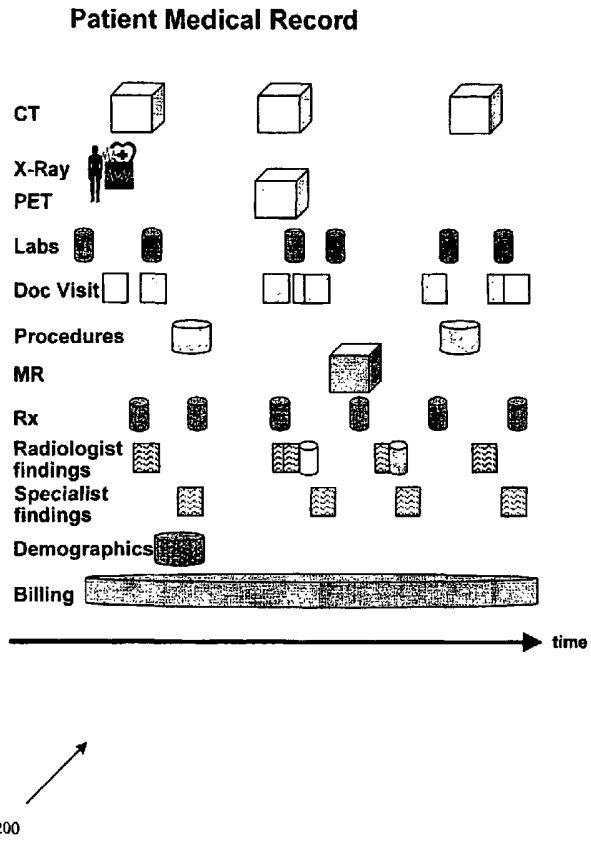
(81) Designated States *(national)*: CA, CN, JP.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *without international search report and to be republished upon receipt of that report*

[Continued on next page]

(54) Title: PATIENT DATA MINING WITH POPULATION-BASED ANALYSIS



(57) Abstract: A system and method for analyzing population-based patient information is provided. The method includes the steps of data mining a plurality of patient records using a domain knowledge base relating to a disease of interest; compiling the mined data into a plurality of structured patient records; inputting at least one patient criteria relating to the disease of interest; and extracting at least one structured patient record matching the at least one patient criteria. The system includes a data miner for mining information from the plurality of patient records using a domain knowledge base relating to a disease of interest and for compiling the mined data into a plurality of structured patient records; an interface for inputting at least one patient criteria relating to the disease of interest; and a processor adapted for extracting at least one of the structured patient records matching the at least one patient criteria.

WO 03/040879 A2

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

PATIENT DATA MINING WITH POPULATION-BASED ANALYSIS

Cross Reference to Related Applications

This application claims the benefit of U.S. Provisional Application Serial No. 60/335,542, filed on November 2, 2001, which is incorporated by reference herein in its entirety.

Field of the Invention

The present invention relates to medical information processing systems, and, more particularly to a computerized system and method for performing outcome analysis on a patient based on population-based information derived from various sources and for conducting retrospective studies on the population-based information.

Background of the Invention

The proper care of medical patients is essential for optimal treatment of their medical conditions. Typically, a patient having a particular condition/ailment is prescribed a medicine or treatment based upon established treatment guidelines. The treatment guidelines outline, *inter alia*, the specific dosages of medicines, the frequency in which dosages should be administered, instructions on how dosages should be administered and the time-lines for therapeutic treatments. Oftentimes, treatment for new patients is administered directly from the treatment guidelines with little variation. These guidelines are typically derived through prospective medical studies. Prospective medical studies, namely, randomized clinical trials, are studies wherein researches empirically test hypothesis in near ideal conditions by

screening the patient population, ensuring that patient care diligently follows the guidelines and recording all relevant data. Such practices fail to take full advantages of historical medical data, rather, relying only on success rates for the patients that rigidly adhered to the treatment guidelines. Additionally, clinical trials are very expensive to conduct.

Historical medical data represents a valuable source in the analysis of the patient care process and medical outcomes. As indicated, treatment guidelines have been generated based solely upon the results of treatment on patients who rigidly adhered to the treatment guidelines. However, the number of variables from patient and professional medical care having an impact on the results of patient care is exceedingly high. Moreover, the relationship between these variables is virtually unknown. Accordingly, the ability to fully learn from past medical data could greatly improved patient health care.

Retrospective studies, for example, the analysis of historical medical patient records from a hospital, are complementary to prospective clinical trials. Health-care organizations are accumulating vast stores of patient data, which are a vital tool for knowledge management. Analyzing this already-collected information may lead to insights that can be subsequently verified in a prospective trial. Most importantly, retrospective studies can measure, in a least two ways, the impact of guidelines in real-life clinical settings. First, retrospective studies can determine the effectiveness of the treatment for a patient population that was excluded from clinical trials. For example, patients above 65, or those with other diseases may be excluded in a clinical trial – however, the guideline validated in that trial is now used to treat *all* hospital patients. Second, patient treatment in a hospital may differ from that in a

trial. For instance, the colon cancer guideline mandates commencing chemotherapy within 6 weeks of surgery, which is rigorously enforced in the clinical trial. However, in a hospital, some patients may begin chemotherapy up to 10 weeks after surgery (e.g., they may be too sick or miss appointments). The impact of this delay on a patient's outcome can only be determined via retrospective analysis since it is not ethical to conduct a clinical trial that would test the impact of this delay – in effect, withholding the accepted standard of care.

However, analyzing hospital data is hard for many reasons. First, medical data is very complex to analyze because of its rich structure. Many traditional statistical methods are ill-suited to data with structure, time-sequenced events (medical data has important temporal components) and/or no structure such as free text, images, etc. Second, because the hospital patient data was collected to treat the patient (as opposed to collected for analysis in a clinical trial), it is imperfect in many ways, for example, missing/incorrect/inconsistent data; key outcomes/variables not recorded; bias in data collection, e.g., sick patients get more tests than well ones, (this is perfectly natural from the medical point of view, but has inherent assumptions that may cause problems for many algorithms); and variables collected/treatments change over time, which particularly impacts some long-term diseases whose treatment can span decades. Lastly, there is wide variation in practice among medical professionals – determining if a patient is on a guideline and treated properly is difficult to tell.

In view of the above, there exists a need for techniques to collect population-based patient information from a variety of sources, to perform outcome analysis on the collected information, and to conduct retrospective analysis on a large quantity of medical information derived from various sources in a rapid manner.

Summary of the Invention

A system and method for analyzing population-based patient information is provided.

According to one aspect of the present invention, a method for analyzing patient records is provided including the steps of data mining a plurality of patient records using a domain knowledge base relating to a disease of interest; compiling the mined data into a plurality of structured patient records; inputting at least one patient criteria relating to the disease of interest; and extracting at least one structured patient record matching at least one patient criteria.

According to another aspect of the present invention, a system for analyzing a plurality of patient records includes a data miner for mining information from the plurality of patient records using a domain knowledge base relating to a disease of interest and for compiling the mined data into a plurality of structured patient records; an interface for inputting at least one patient criteria relating to the disease of interest; and a processor adapted for extracting at least one of the structured patient records matching at least one patient criteria.

In a further aspect of the present invention, a method for conducting a retrospective study on a plurality of patient records is provided. The method includes the steps of data mining the plurality of patient records using a domain knowledge base relating to a disease of interest; compiling the mined data into a plurality of structured patient records; inputting a plurality of patient criteria forming a hypothesis relating to the disease of interest; and extracting a plurality of structured patient records matching the plurality of patient criteria. The method further includes

the steps of determining patient outcomes from the plurality of structured patient records and validating the hypothesis by comparing the patient outcomes to a suggested outcome.

In another aspect of the present invention, a program storage device readable by a machine, tangibly embodying a program of instructions executable by the machine to perform method steps for analyzing patient records is provided including the method steps of data mining a plurality of patient records using a domain knowledge base relating to a disease of interest; compiling the mined data into a plurality of structured patient records; inputting at least one patient criteria relating to the disease of interest; and extracting at least one structured patient record matching the at least one patient criteria.

Brief Description of the Drawings

The above and other aspects, features and advantages of the present invention will become more apparent from the following detailed description when taken in conjunction with the accompanying drawings in which:

FIG. 1 is a block diagram of a computer processing system to which the present invention may be applied according to an embodiment of the present invention;

FIG. 2 illustrates an exemplary computerized patient record (CPR); and

FIG. 3 illustrates an exemplary data mining framework for mining high-quality structured medical information;

FIG. 4 illustrates a block diagram of an exemplary analysis system according to an embodiment of the present invention; and

FIG. 5 illustrates a flow diagram for analyzing large amounts of medical information according to an embodiment of the present invention.

Description of Preferred Embodiments

To facilitate a clear understanding of the present invention, illustrative examples are provided herein which describe certain aspects of the invention. However, it is to be appreciated that these illustrations are not meant to limit the scope of the invention, and are provided herein to illustrate certain concepts associated with the invention.

A system and method for analyzing population-based medical data is provided. According to an embodiment of the present invention, a computer-based system will compile population-based patient data from various sources, e.g., structured and unstructured, into a structured database for analysis. First, the system will assimilate information from both structured, e.g., financial, and unstructured, e.g., imaging, sources within a computerized patient record (CPR). These data can be automatically extracted, combined, and analyzed in a meaningful way.

The present invention allows for analysis of a large amount of information in a rapid manner, as opposed to the traditional method of medical personnel reviewing each record and transposing their findings. Since information is collected from a variety of sources containing different information relating to specific patients, various criteria or variables can be analyzed to determine their effect on a proposed treatment or guideline.

It is to be understood that the present invention may be implemented in various forms of hardware, software, firmware, special purpose processors, or a

combination thereof. Preferably, the present invention is implemented in software as a program tangibly embodied on a program storage device. The program may be uploaded to, and executed by, a machine comprising any suitable architecture. Preferably, the machine is implemented on a computer platform having hardware such as one or more central processing units (CPU), a random access memory (RAM), and input/output (I/O) interface(s). The computer platform also includes an operating system and microinstruction code. The various processes and functions described herein may either be part of the microinstruction code or part of the program (or combination thereof) which is executed via the operating system. In addition, various other peripheral devices may be connected to the computer platform such as an additional data storage device and a printing device.

It is to be understood that, because some of the constituent system components and method steps depicted in the accompanying figures are preferably implemented in software, the actual connections between the system components (or the process steps) may differ depending upon the manner in which the present invention is programmed.

FIG. 1 is a block diagram of a computer processing system 100 to which the present invention may be applied according to an embodiment of the present invention. The system 100 includes at least one processor (hereinafter processor) 102 operatively coupled to other components via a system bus 104. A read-only memory (ROM) 106, a random access memory (RAM) 108, an I/O interface 110, a network interface 112, and external storage 114 are operatively coupled to the system bus 104. Various peripheral devices such as, for example, a display device, a disk storage device(e.g., a magnetic or optical disk storage device), a keyboard, and a

mouse, may be operatively coupled to the system bus 104 by the I/O interface 110 or the network interface 112.

The computer system 100 may be a standalone system or be linked to a network via the network interface 112. The network interface 112 may be a hard-wired interface. However, in various exemplary embodiments, the network interface 112 can include any device suitable to transmit information to and from another device, such as a universal asynchronous receiver/transmitter (UART), a parallel digital interface, a software interface or any combination of known or later developed software and hardware. The network interface may be linked to various types of networks, including a local area network (LAN), a wide area network (WAN), an intranet, a virtual private network (VPN), and the Internet.

The external storage 114 may be implemented using a database management system (DBMS) managed by the processor 102 and residing on a memory such as a hard disk. However, it should be appreciated that the external storage 114 may be implemented on one or more additional computer systems. For example, the external storage 114 may include a data warehouse system residing on a separate computer system.

Those skilled in the art will appreciate that other alternative computing environments may be used without departing from the spirit and scope of the present invention.

Increasingly, health care providers are employing automated techniques for information storage and retrieval. The use of a computerized patient record (CPR) to maintain patient information is one such example. As shown in Fig. 2, an exemplary CPR (200) includes information that is collected over the course of a patient's

treatment. This information may include, for example, computed tomography (CT) images, X-ray images, laboratory test results, doctor progress notes, details about medical procedures, prescription drug information, radiological reports, other specialist reports, demographic information, and billing (financial) information.

A CPR typically draws from a plurality of data sources, each of which typically reflects a different aspect of a patient's care. Structured data sources, such as financial, laboratory, and pharmacy databases, generally maintain patient information in database tables. Information may also be stored in unstructured data sources, such as, for example, free text, images, and waveforms. Often, key clinical findings are only stored within physician reports, e.g., dictations.

Fig. 3 illustrates an exemplary data mining system for mining high-quality structured clinical information using data mining techniques described in "Patient Data Mining," by Rao et al., copending U.S. Patent Application Serial No. 10/___,___, (Attorney Docket No. 8706-600) filed herewith, which is incorporated by reference in its entirety. The data mining system includes a data miner (350) that mines information from a CPR (310) using domain-specific knowledge contained in a knowledge base (330). The data miner (350) includes components for extracting information from the CPR (352), combining all available evidence in a principled fashion over time (354), and drawing inferences from this combination process (356). The mined information may be stored in a structured CPR database (380). In this manner, all information contained in a CPR, whether from a structured or unstructured source, will stored in a structured fashion.

The extraction component (352) deals with gleaning small pieces of information from each data source regarding a patient, which are represented as

probabilistic assertions about the patient at a particular time. These probabilistic assertions are called *elements*. The combination component (354) combines all the elements that refer to the same variable at the same time period to form one unified probabilistic assertion regarding that variable. These unified probabilistic assertions are called *factoids*. The inference component (356) deals with the combination of these factoids, at the same point in time and/or at different points in time, to produce a coherent and concise picture of the progression of the patient's state over time. This progression of the patient's state is called a *state sequence*.

The present invention can build an individual model of the state of a patient. The patient state is simply a collection of variables or criteria that one may care about relating to the patient. The information of interest may include a state sequence, i.e., the value of the patient state at different points in time during the patient's treatment.

Each of the above components uses detailed knowledge regarding the domain of interest, such as, for example, a disease of interest. This domain knowledge base (330) can come in two forms. It can be encoded as an input to the system, or as programs that produce information that can be understood by the system. The part of the domain knowledge base (330) that is input to the present form of the system may also be learned from data.

As mentioned, the extraction component (352) takes information from the CPR (310) to produce probabilistic assertions (elements) about the patient that are relevant to an instant in time or time period. This process is carried out with the guidance of the domain knowledge that is contained in the domain knowledge base

(330). The domain knowledge required for extraction is generally specific to each source.

Referring to FIG. 4, an exemplary analysis system 400 according to an embodiment of the present invention is illustrated. The analysis system 400 includes a processor 402 for extracting information from the structured database 380 and for performing different tasks on the extracted information. Additionally, the processor 402 is adapted to receive manually inputted patient criteria or variables 414 via an I/O interface which will be used to extract specific information from the database 380. Each task performed by the analysis system 200 is performed by an executable module residing either in the processor of the system 402 and/or in a memory device (e.g., RAM, ROM, external storage, etc.) of the system.

Referring to FIG. 5, a flow chart illustrating a method of analyzing population-based data is provided. For example, the problem of unsatisfactory outcomes (e.g., clinical, financial, and length of stay) in patients with diabetes who sustain a myocardial infarction can be examined for a particular hospital.

First, a plurality of computerized patient records is assembled during the course of treatment of a large number of patients over time, for example, in a particular hospital. This historical data is mined using a domain knowledge base relating to a disease of interest and compiled in a structured CPR database (step 502). For example, information is extract from a variety of sources to identify patients with a confirmed diagnosis of acute myocardial infarction (AMI). This will not be based on ICD-9 codes (which have about 90% accuracy), but on a combination of clinical, laboratory, and EKG findings that meet the MONICA criteria, the internationally accepted standard for identifying AMI patients.

One or more criteria or variables relating to the disease of interest is inputted into the system (step 504). The system extracts patient records from the structured database which conform to the criteria (step 506). For example, once the AMI patients are identified, the system will separate out a subset of patients with diabetes mellitus (e.g., the criteria), based on pharmacy data showing the need for administration of insulin or other anti-diabetic agents, and on lab data showing high blood sugars.

Then, the system determines patient outcomes for the extracted patient records (step 508) and outputs the results. At least one value of the patient criteria may be changed to determine how the change in value of the criteria effects the outcome (step 510). Finally, the system will compile and output the outcome results so the appropriate personnel can review (step 512). The system identifies differences in clinical outcomes, e.g. death, procedures (coronary bypass or angioplasty), infections, etc, and places these results in the context of the accompanying financial, case-mix, treatment, therapy and length of stay data. The output may be a chart, table, curve, etc. illustrating the effects of the changes in criteria against patient outcomes.

In another embodiment, the system and method of the present invention will perform outcome analysis on a particular patient, for example, a physician may want to determine the best prescription drug for lowering a patient's cholesterol level. The system will extract patient records for patient with a cholesterol level over a predetermined limit, e.g., 250. The physician will enter criteria or variables 414 related to a current state of the patient, e.g., age, blood pressure, LDL cholesterol, HDL cholesterol, etc. The processor 402 will then interact with the structured CPR

database 380 to extract patient records that match the criteria of the current patient and will output the patient outcomes versus drug treatments of the extracted records. The physician may change a value of one or more of the criteria or variables, e.g., use of a different drug, changes to the patients smoking habits, etc., to determine how the outcome is affected by the change, wherein the system will extract new patient records to reflect patient outcomes based on the new set of variables. Since the system can extract different patient records based on different criteria from a large volume of records, the system can perform outcome analysis much faster than in the traditional manner of trying to search by hand patient records with similar information.

Additionally, the system may be used to generate a hypothesis for a potential prospective clinical trial by correlating the inputted criteria to the determined outcomes.

In another embodiment, the system and method of the present invention may be employed to conduct a retrospective study. During a prospective clinical trial, a particular group of people, for example, males ages 25 to 40, may have been observed to determine the most appropriate guideline for treating a particular disease. The guideline developed from the clinical trial is later then applied to all age groups without further testing. The system and method of the present invention will allow a study to be conducted on people excluded from the trial by extracting patient records which match the guideline created during the actual trial but will be restricted by an inputted patient criteria, e.g., females ages 40-50. The system and method of the present invention allow a retrospective study be conducted on a large

population of people without the need for someone to manually review a large number of records.

Furthermore, a retrospective study may be conducted to validate the hypothesis generated by correlating the inputted criteria to the determined patient outcomes and, then, comparing the determined patient outcomes to a suggested patient outcome of the hypothesis.

The analysis system and method of the present invention provides for a collection of a large volume of data from various sources, i.e., structured and unstructured, to be analyzed in an efficient and rapid manner. The method and system will provide improve quality of care by allowing medical professionals to perform patient outcome analysis on population-based patient information, e.g., a large quantity of patients treated by a hospital, to determine the most appropriate treatment. Additionally, the system and method of the present invention will reduce costs to researchers and hospitals by allowing retrospective studies to be performed automatically by mining data from varied sources, as opposed to conventional individual review and analysis.

Although illustrative embodiments of the present invention have been described herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various other changes and modifications may be affected therein by one skilled in the art without departing from the scope or spirit of the invention.

What Is Claimed Is:

1. A method for analyzing patient records, the method comprising the steps of:

data mining a plurality of patient records using a domain knowledge base relating to a disease of interest;

compiling the mined data into a plurality of structured patient records;

inputting at least one patient criteria relating to the disease of interest; and

extracting at least one structured patient record matching the at least one patient criteria.

2. The method as in claim 1, further comprising the step of determining a patient outcome of the at least one structured patient record.

3. The method as in claim 2, further comprising the steps of changing a value of the at least one patient criteria and repeating the extracting and determining steps.

4. The method as in claim 1, wherein the plurality of patient records are stored in structured and unstructured sources.

5. The method as in claim 1, wherein the plurality of patient records are collected over a course of treatment of a plurality of patients.

6. The method as in claim 2, further comprising the step of correlating a plurality of criteria to a plurality of patient outcomes.

7. The method as in claim 6, further comprising the step of suggesting a hypothesis for a clinical trial based on the correlation.

8. The method as in claim 7, further comprising the step of validating the hypothesis by performing a retrospective study on the plurality of structured patient records.

9. A system for analyzing a plurality of patient records, the plurality of patient records being stored in structured and unstructured sources, the system comprising:

a data miner for mining information from the plurality of patient records using a domain knowledge base relating to a disease of interest and for compiling the mined data into a plurality of structured patient records;

an interface for inputting at least one patient criteria relating to the disease of interest; and

a processor adapted for extracting at least one of the structured patient records matching the at least one patient criteria.

10. The system as in claim 9, further comprising a database for storing the plurality of structured patient records.

11. The system as in claim 9, wherein the processor is further adapted to determine a patient outcome of the at least one structured patient record.

12. The system as in claim 11, wherein the processor is further adapted to correlate a plurality of criteria to a plurality of patient outcomes.

13. The system as in claim 12, wherein the processor is further adapted to suggest a hypothesis for a clinical trial based on the correlation.

14. A method for conducting a retrospective study on a plurality of patient records, the method comprising the steps of:

data mining the plurality of patient records using a domain knowledge base relating to a disease of interest;

compiling the mined data into a plurality of structured patient records;

inputting a plurality of patient criteria forming a hypothesis relating to the disease of interest; and

extracting a plurality of structured patient records matching the plurality of patient criteria.

15. The method as in claim 14, further comprising the step of determining patient outcomes from the plurality of structured patient records.

16. The method as in claim 15, further comprising the step of validating the hypothesis by comparing the patient outcomes to a suggested outcome.

17. A program storage device readable by a machine, tangibly embodying a program of instructions executable by the machine to perform method steps for analyzing patient records, the method steps comprising:

data mining a plurality of patient records using a domain knowledge base relating to a disease of interest;

compiling the mined data into a plurality of structured patient records;

inputting at least one patient criteria relating to the disease of interest; and

extracting at least one structured patient record matching the at least one patient criteria.

18. The program storage device as in claim 17, further comprising the method step of determining a patient outcome of the at least one structured patient record.

19. The program storage device as in claim 18, further comprising the method steps of changing a value of the at least one patient criteria and repeating the extracting and determining steps.

20. The program storage device as in claim 17, wherein the plurality of patient records are stored in structured and unstructured sources.

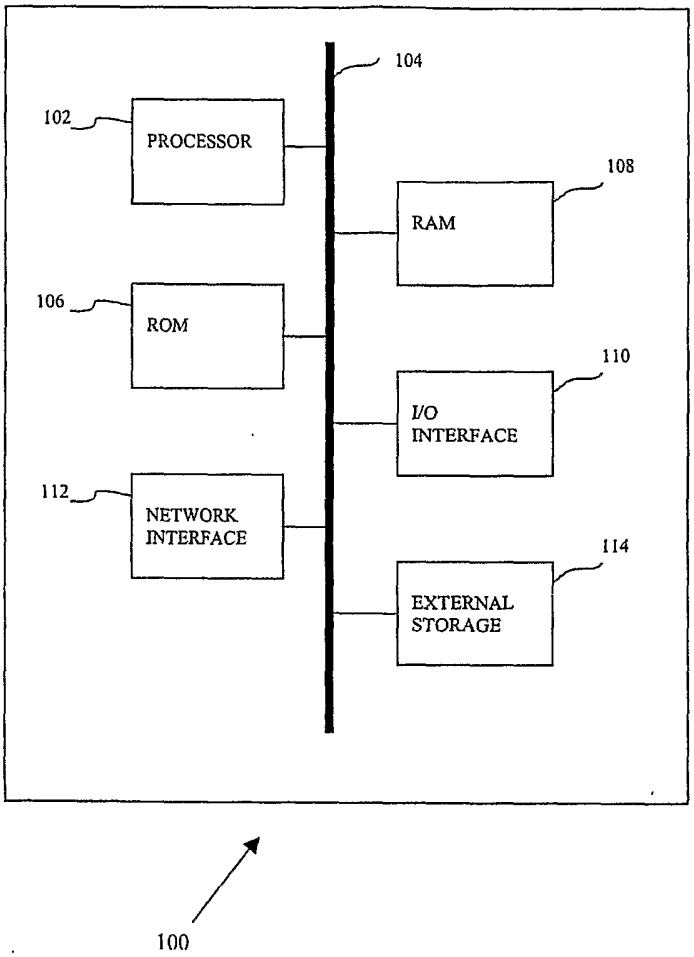

104
102
PROCESSOR
108
RAM
106
ROM
110
I/O
INTERFACE
112
NETWORK
INTERFACE
114
EXTERNAL
STORAGE
100

FIG. 1

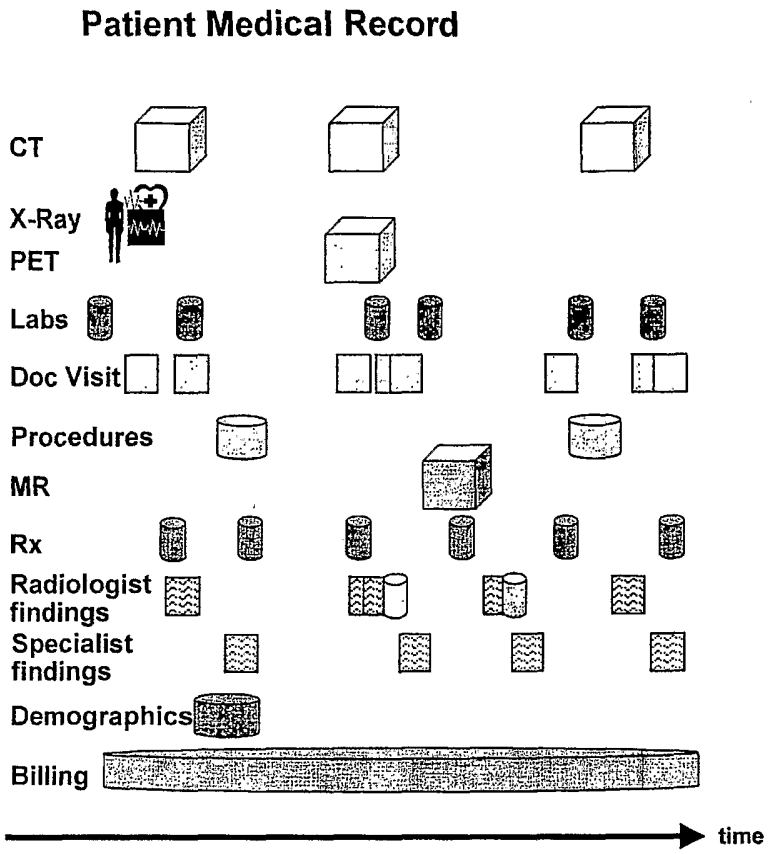

Patient Medical Record
CT
X-Ray
PET
Labs
Doc Visit
Procedures
MR
Rx
Radiologist findings
Specialist findings
Demographics
Billing
time

FIG. 2

200

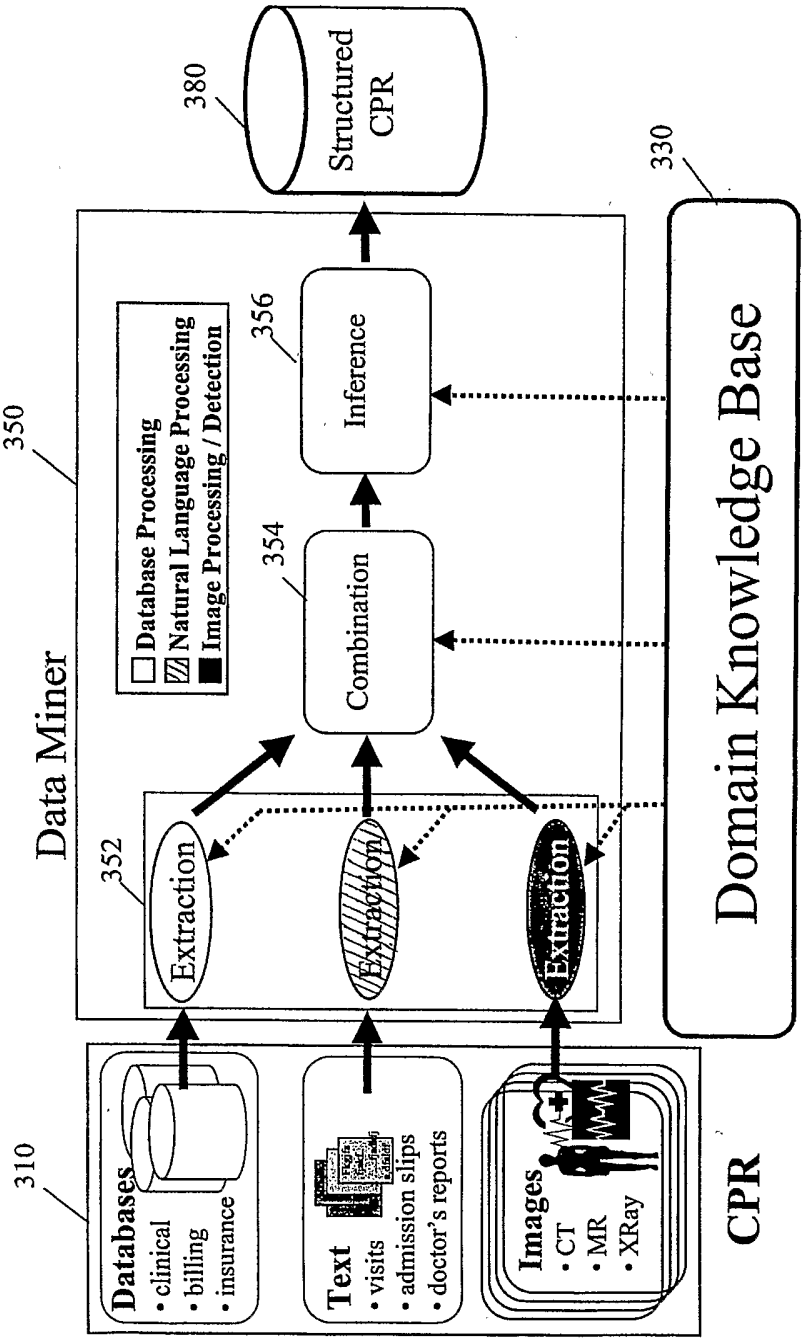

380
Structured CPR
330
Domain Knowledge Base
350
Data Miner
Database Processing
Natural Language Processing
Image Processing / Detection
356
Inference
354
Combination
352
Extraction
Extraction
Extraction
310
Databases
• clinical
• billing
• insurance
Text
• visits
• admission slips
• doctor's reports
Images
• CT
• MR
• XRay
CPR

FIG 3

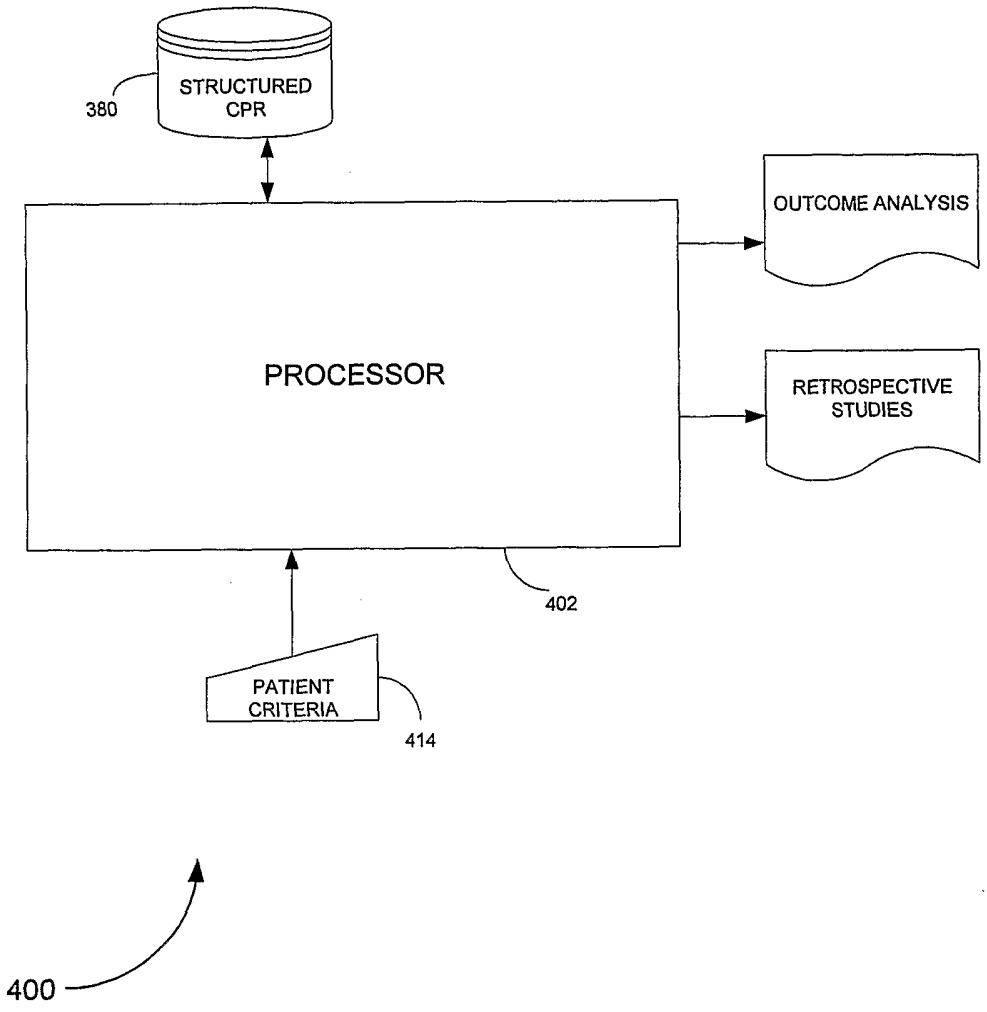

STRUCTURED CPR
380
OUTCOME ANALYSIS
PROCESSOR
RETROSPECTIVE STUDIES
402
PATIENT CRITERIA
414
400

FIG. 4

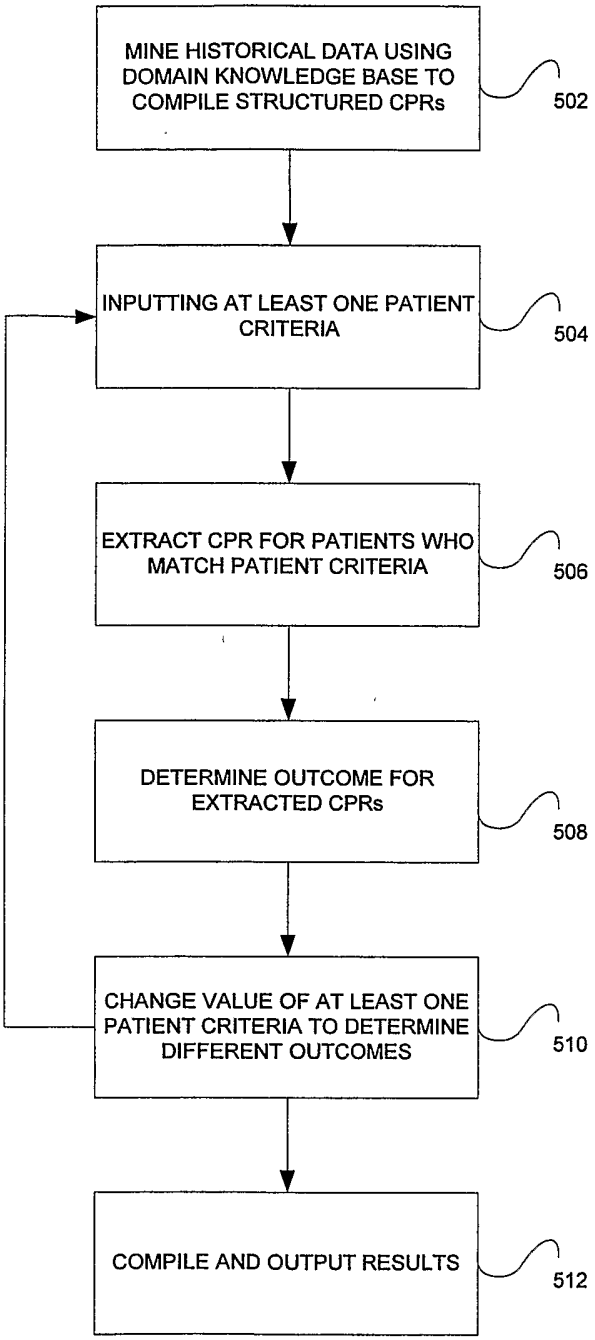

MINE HISTORICAL DATA USING DOMAIN KNOWLEDGE BASE TO COMPILE STRUCTURED CPRs
502
INPUTTING AT LEAST ONE PATIENT CRITERIA
504
EXTRACT CPR FOR PATIENTS WHO MATCH PATIENT CRITERIA
506
DETERMINE OUTCOME FOR EXTRACTED CPRs
508
CHANGE VALUE OF AT LEAST ONE PATIENT CRITERIA TO DETERMINE DIFFERENT OUTCOMES
510
COMPILE AND OUTPUT RESULTS
512

FIG. 5