US 20150324527A1

(19) **United States**

(12) **Patent Application Publication**

Siegel et al.

(10) **Pub. No.: US 2015/0324527 A1**

(43) **Pub. Date: Nov. 12, 2015**

(54) **LEARNING HEALTH SYSTEMS AND METHODS**

(71) Applicant: **Northrop Grumman Systems Corporation**, Falls Church, VA (US)

(72) Inventors: **Neil G. Siegel**, Rolling Hills Estates, CA (US); **Sam S. Shekar**, Potomac, MD (US); **Jeffrey C. Yu**, Castro Valley, CA (US); **Sreelatha Ghanta**, Ashburn, VA (US); **Sanjiv Desai**, Falls Church, VA (US); **H. Morgan Crafts, JR.**, Oakton, VA (US)

(21) Appl. No.: **14/803,963**

(22) Filed: **Jul. 20, 2015**

Related U.S. Application Data

(63) Continuation-in-part of application No. 13/837,370, filed on Mar. 15, 2013.

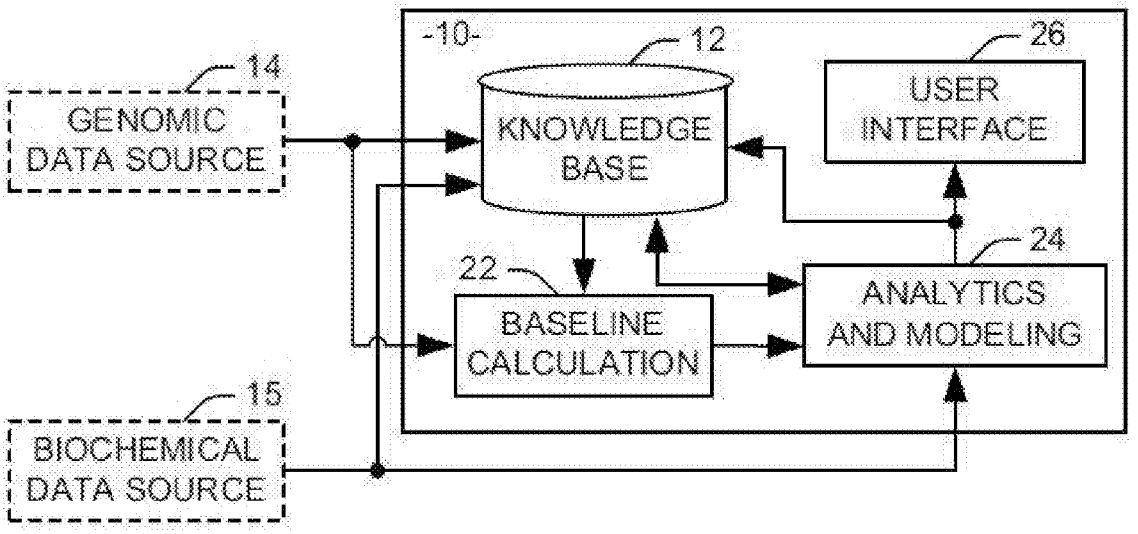
Publication Classification

(51) **Int. Cl.**
G06F 19/00 (2006.01)

(52) **U.S. Cl.**
CPC ***G06F 19/322*** (2013.01)

(57) **ABSTRACT**

A method for providing personalized healthcare for a patient is provided. The method may include receiving information indicative of a patient, retrieving a record associated with the patient, receiving a query identifying a healthcare related issue associated with the patient, performing analytics via a statistical discovery component and a natural language processing component configured to interface with respective portions of heterogeneous data sources to selectively identify correlations between genomic profile information of the patient and selected data of the data sources, applying a selected risk model based on the query; and providing a response to the query including information associated with clinical decision support tailored to an identity of the user, the user being a selected one of a patient, a researcher and a clinician.



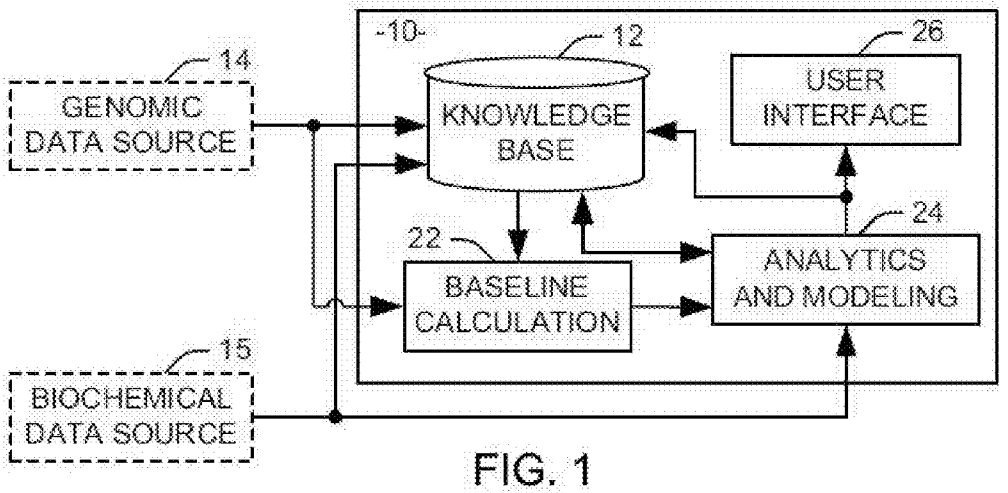

10
12
14
GENOMIC DATA SOURCE
KNOWLEDGE BASE
26
USER INTERFACE
24
22
BASELINE CALCULATION
ANALYTICS AND MODELING
15
BIOCHEMICAL DATA SOURCE

FIG. 1

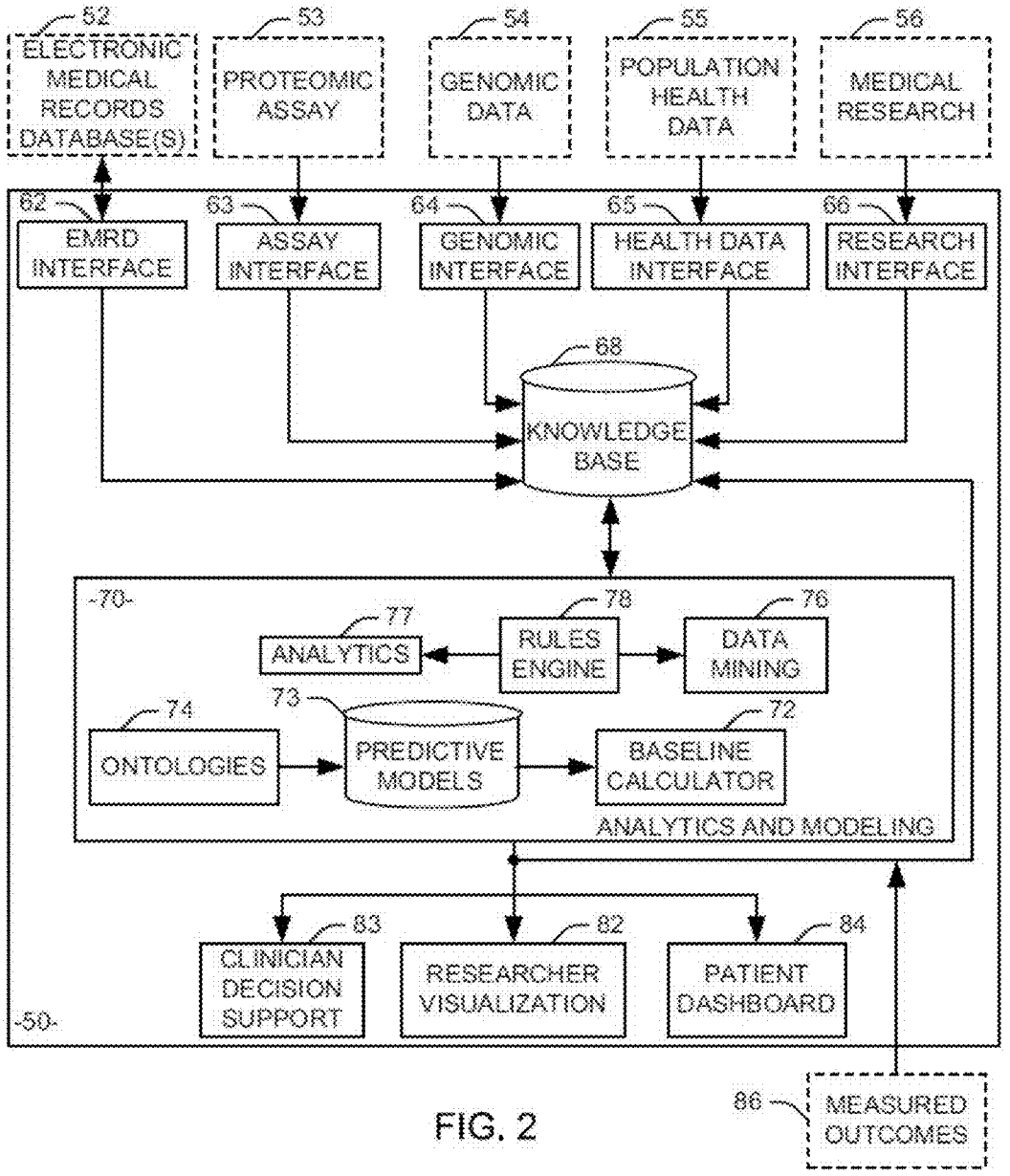

52
ELECTRONIC MEDICAL RECORDS DATABASE(S)
53
PROTEOMIC ASSAY
54
GENOMIC DATA
55
POPULATION HEALTH DATA
56
MEDICAL RESEARCH
62
EMRD INTERFACE
63
ASSAY INTERFACE
64
GENOMIC INTERFACE
65
HEALTH DATA INTERFACE
66
RESEARCH INTERFACE
68
KNOWLEDGE BASE
-70-
77
ANALYTICS
78
RULES ENGINE
76
DATA MINING
74
ONTOLOGIES
73
PREDICTIVE MODELS
72
BASELINE CALCULATOR
ANALYTICS AND MODELING
83
CLINICIAN DECISION SUPPORT
82
RESEARCHER VISUALIZATION
84
PATIENT DASHBOARD
-50-
86
MEASURED OUTCOMES

FIG. 2

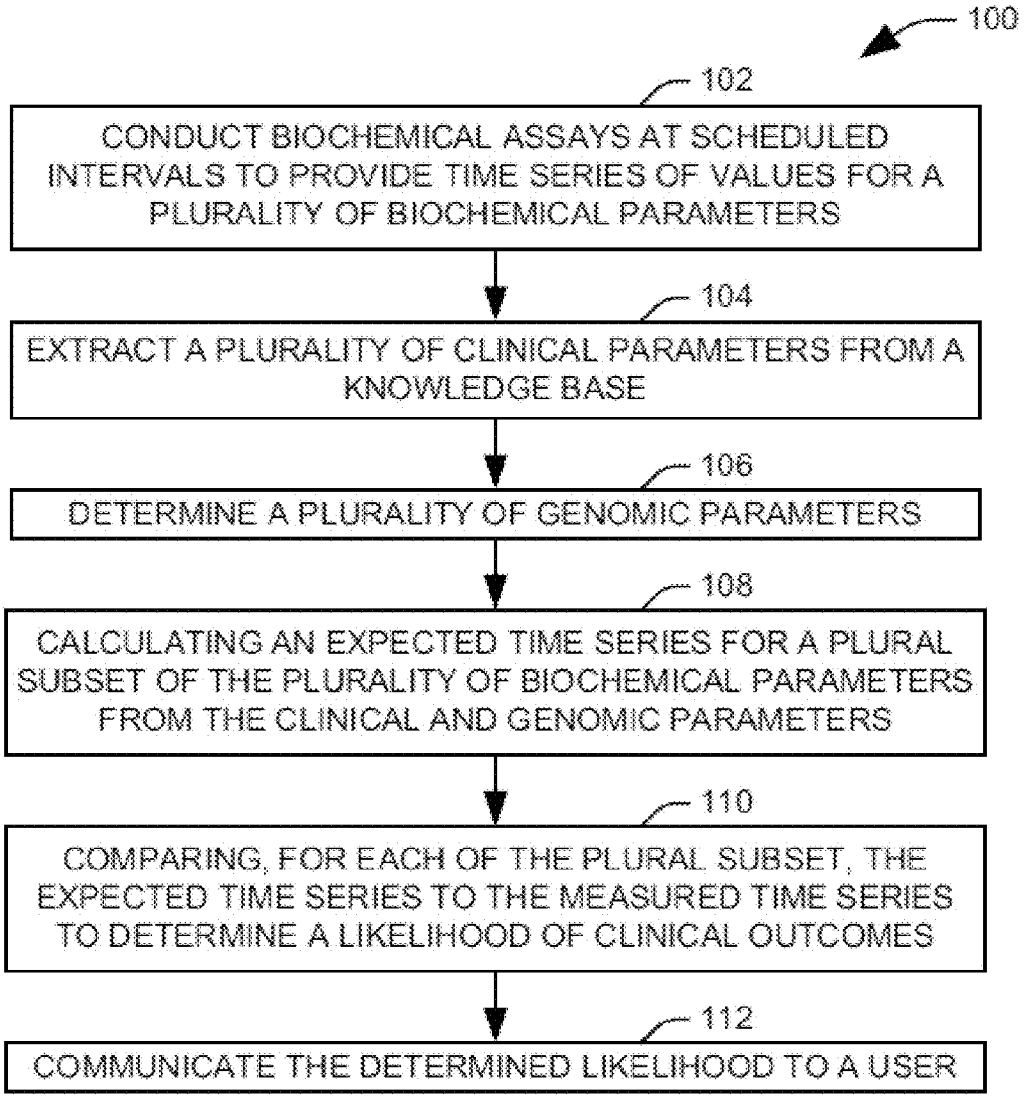

100
102
CONDUCT BIOCHEMICAL ASSAYS AT SCHEDULED INTERVALS TO PROVIDE TIME SERIES OF VALUES FOR A PLURALITY OF BIOCHEMICAL PARAMETERS
104
EXTRACT A PLURALITY OF CLINICAL PARAMETERS FROM A KNOWLEDGE BASE
106
DETERMINE A PLURALITY OF GENOMIC PARAMETERS
108
CALCULATING AN EXPECTED TIME SERIES FOR A PLURAL SUBSET OF THE PLURALITY OF BIOCHEMICAL PARAMETERS FROM THE CLINICAL AND GENOMIC PARAMETERS
110
COMPARING, FOR EACH OF THE PLURAL SUBSET, THE EXPECTED TIME SERIES TO THE MEASURED TIME SERIES TO DETERMINE A LIKELIHOOD OF CLINICAL OUTCOMES
112
COMMUNICATE THE DETERMINED LIKELIHOOD TO A USER

FIG. 3

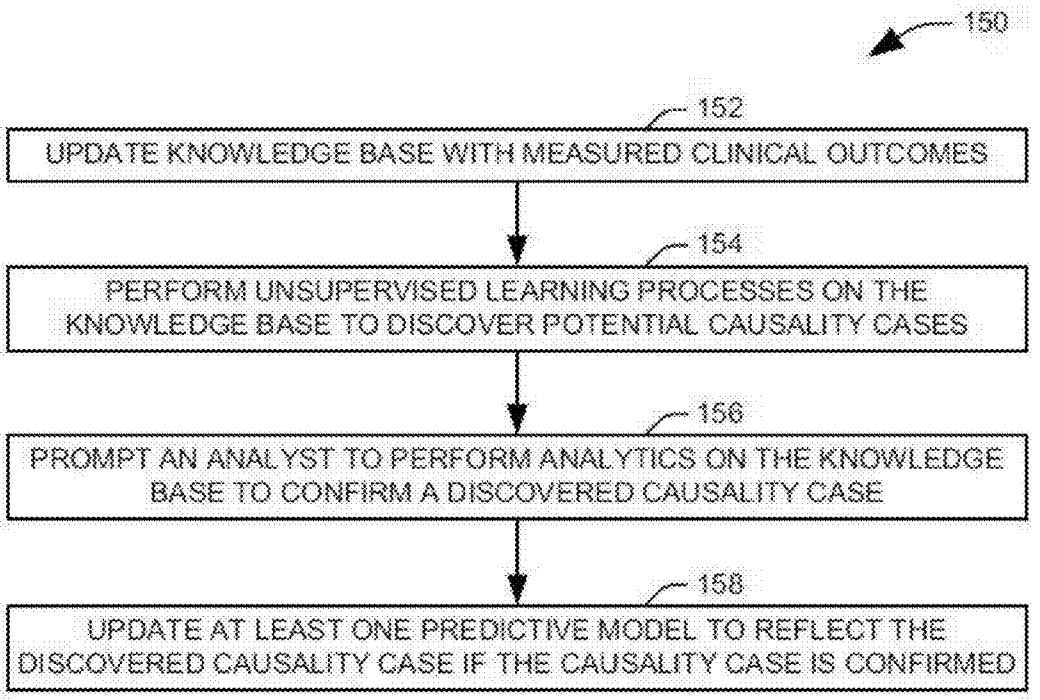

150
152
UPDATE KNOWLEDGE BASE WITH MEASURED CLINICAL OUTCOMES
154
PERFORM UNSUPERVISED LEARNING PROCESSES ON THE KNOWLEDGE BASE TO DISCOVER POTENTIAL CAUSALITY CASES
156
PROMPT AN ANALYST TO PERFORM ANALYTICS ON THE KNOWLEDGE BASE TO CONFIRM A DISCOVERED CAUSALITY CASE
158
UPDATE AT LEAST ONE PREDICTIVE MODEL TO REFLECT THE DISCOVERED CAUSALITY CASE IF THE CAUSALITY CASE IS CONFIRMED

FIG. 4

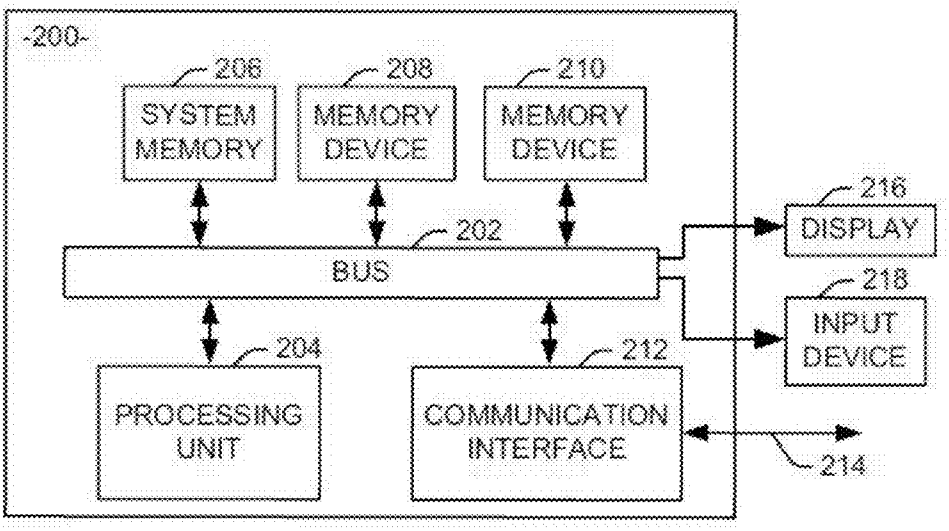

200
206
SYSTEM MEMORY
208
MEMORY DEVICE
210
MEMORY DEVICE
202
BUS
216
DISPLAY
218
INPUT DEVICE
204
PROCESSING UNIT
212
COMMUNICATION INTERFACE
214

FIG. 5

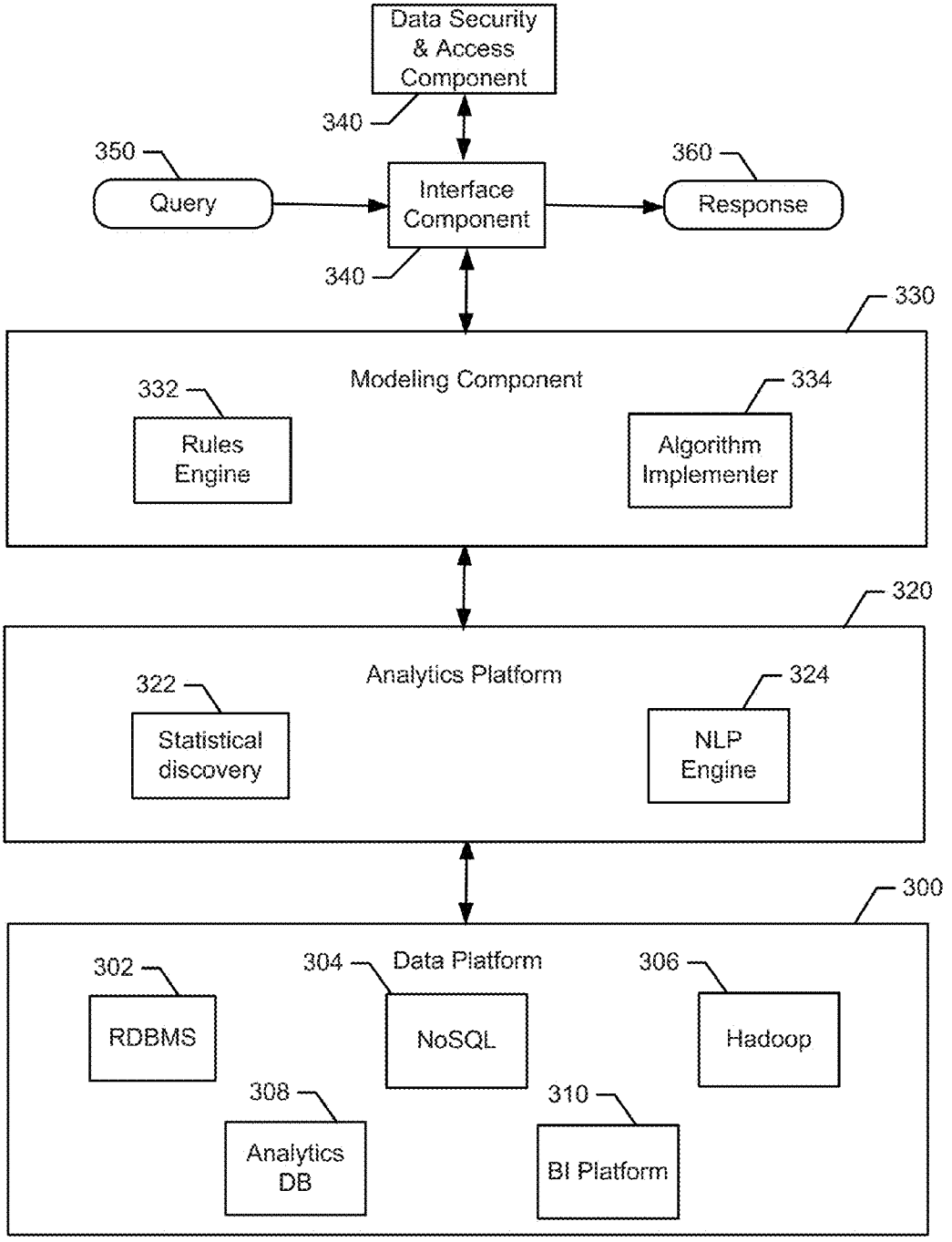

Data Security & Access Component
340
350
Query
Interface Component
360
Response
340
330
Modeling Component
332
Rules Engine
334
Algorithm Implementer
320
Analytics Platform
322
Statistical discovery
324
NLP Engine
300
302
RDBMS
304
Data Platform
NoSQL
306
Hadoop
308
Analytics DB
310
BI Platform

FIG. 6

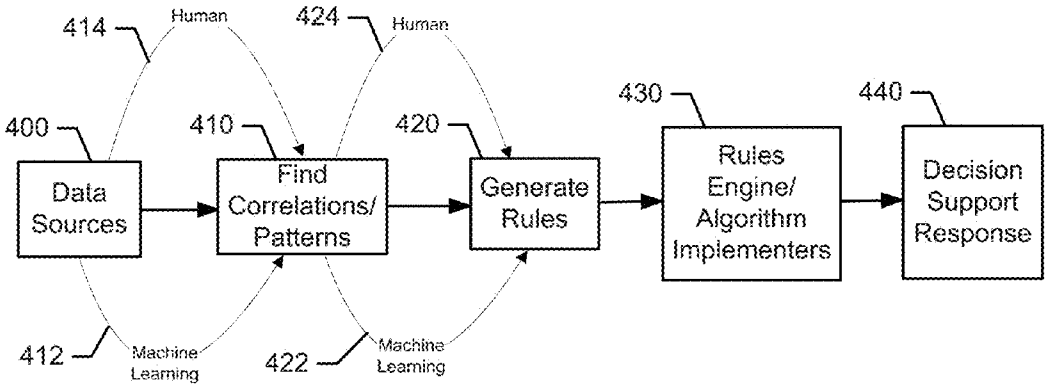

414
Human
424
Human
400
410
420
430
440
Data Sources
Find Correlations/ Patterns
Generate Rules
Rules Engine/ Algorithm Implementers
Decision Support Response
412
Machine Learning
422
Machine Learning

FIG. 7

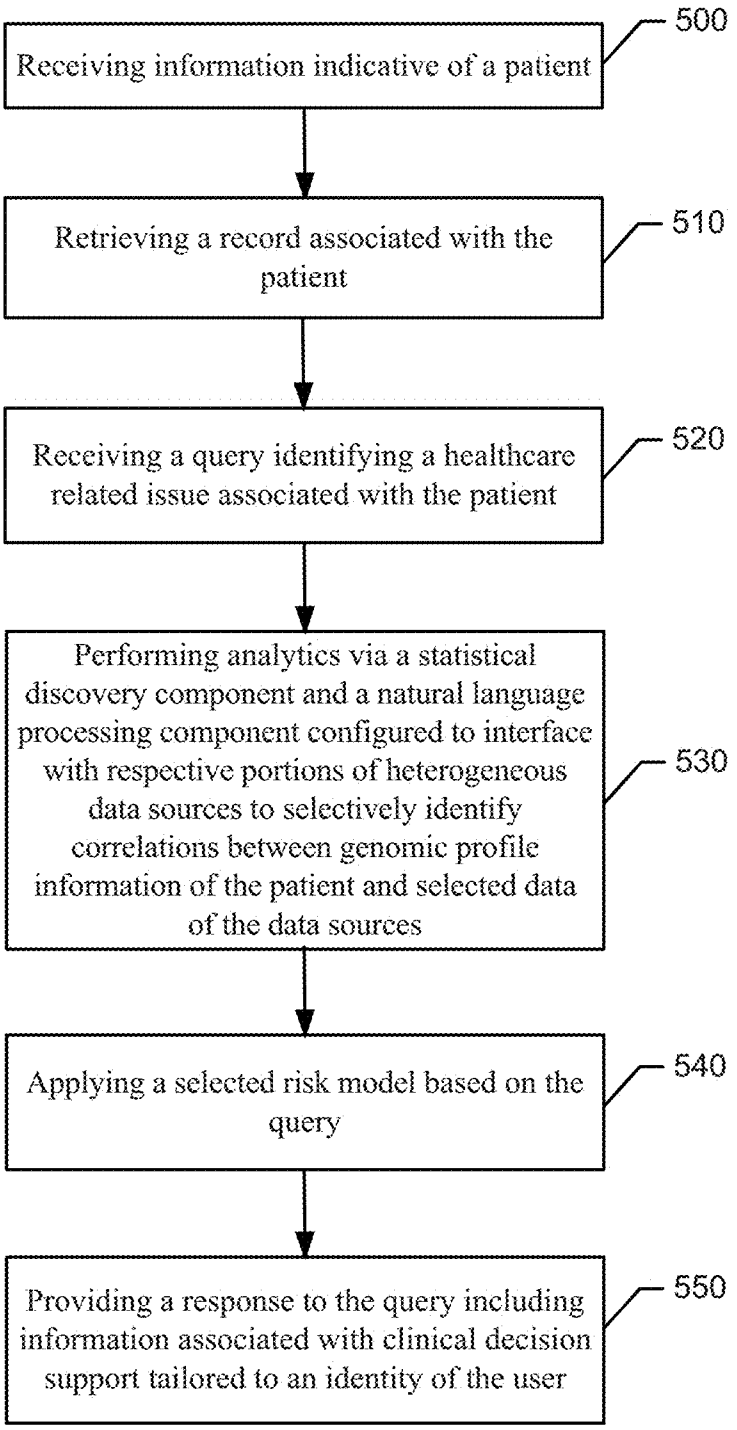

Receiving information indicative of a patient
500
Retrieving a record associated with the patient
510
Receiving a query identifying a healthcare related issue associated with the patient
520
Performing analytics via a statistical discovery component and a natural language processing component configured to interface with respective portions of heterogeneous data sources to selectively identify correlations between genomic profile information of the patient and selected data of the data sources
530
Applying a selected risk model based on the query
540
Providing a response to the query including information associated with clinical decision support tailored to an identity of the user
550

FIG. 8

LEARNING HEALTH SYSTEMS AND METHODS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is a continuation-in-part of U.S. patent application Ser. No. 13/837,370, entitled "LEARNING HEALTH SYSTEMS AND METHODS," filed Mar. 15, 2013, the entire contents of which are hereby incorporated herein by reference.

TECHNICAL FIELD

[0002] Example embodiments generally relate to healthcare information management, and more particularly, to learning health systems and methods.

BACKGROUND

[0003] The healthcare industry provides goods and services to treat patients with curative, preventive, rehabilitative, and palliative care. The modern healthcare sector is divided into many sub-sectors, and depends on interdisciplinary teams of trained professionals and paraprofessionals to meet health needs of individuals and populations. The healthcare industry is one of the world's largest and fastest-growing industries. Consuming over ten percent of gross domestic product (GDP) of most developed nations, healthcare can form an enormous part of a country's economy. Currently, the United States spends over seventeen percent of GDP on healthcare, and this amount is expected to grow at a nearly six percent annual rate. Many attempts have been made to slow down, and eventually reverse, this increase in healthcare spending, however, most attempts have failed or have not had an impact as yet. Contributing to the cost impacts, healthcare is often provided at later stages of illness—based on current technologies and applications—rather than on earlier stages of illness, where care would be less intensive and costs would be much lower.

SUMMARY

[0004] In accordance with an example embodiment, a personalized healthcare system is provided. The system may include a data platform, an analytics platform, a modeling component and a user interface component. The data platform may be scalable to include a plurality of data sources. The data sources may include at least a clinical research database, genomic data, and a patient health record database. The patient health record database may include a record for each of a population of patients, a plurality of genetic markers, and a plurality of clinical parameters associated with the patients. The analytics platform may include at least a statistical discovery component and a natural language processing component configured to interface with respective portions of the data sources to selectively identify correlations based on analysis of contents of the data sources responsive to a query. The modeling component may be configured to apply a selected risk model based on the query. The user interface component may be configured to enable a user to provide the query, and to generate a response to the query. The response may provide information associated with clinical decision support tailored to an identity of the user, the user being a selected one of a patient, a researcher and a clinician.

[0005] In accordance with another aspect of the present invention, a method is provided for enhancing personalized healthcare. The method may include receiving information indicative of a patient, retrieving a record associated with the patient, receiving a query identifying a healthcare related issue associated with the patient, performing analytics via a statistical discovery component and a natural language processing component configured to interface with respective portions of heterogeneous data sources to selectively identify correlations between genomic profile information of the patient and selected data of the data sources, applying a selected risk model based on the query; and providing a response to the query including information associated with clinical decision support tailored to an identity of the user, the user being a selected one of a patient, a researcher and a clinician.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] Having thus described some example embodiments in general terms, reference will now be made to the accompanying drawings, which are not necessarily drawn to scale, and wherein:

[0007] FIG. **1** illustrates block diagram of a learning healthcare system in accordance with an example embodiment;

[0008] FIG. **2** illustrates a block diagram of one implementation of a learning healthcare information processing system in accordance with an example embodiment;

[0009] FIG. **3** illustrates a block diagram of a method for providing personalized healthcare support in accordance with an example embodiment;

[0010] FIG. **4** illustrates a block diagram of a method for discovering and applying new causality cases in a learning healthcare system in accordance with an example embodiment; and

[0011] FIG. **5** is a schematic block diagram illustrating an exemplary system of hardware components capable of implementing examples of the systems and methods disclosed in FIGS. **1-4** in accordance with an example embodiment;

[0012] FIG. **6** is a block diagram of a cloud-based platform for implementing an example embodiment;

[0013] FIG. **7** illustrates a block diagram of the mechanisms and platforms associated with practicing example embodiments; and

[0014] FIG. **8** illustrates a method of providing personalized healthcare in accordance with an example embodiment.

DETAILED DESCRIPTION

[0015] Healthcare in the many nations is driven by medical protocols, which are guidelines for when and how to perform diagnostic and clinical activities on an individual. These protocols, however, are created with, at best, superficial reference to any significant knowledge of the individual. The inventors have found that genomics can be helpful in customizing care, and that genomics data can sometimes be supplemented with other data that provide more insight into ones health condition at the time of the measurement to further customize care. Genomics can be helpful in relation to identifying risk relative to development of a condition, while other measurements may be about present health status.

[0016] Accordingly, the inventors have determined that data about an individual—derived from proteomics and other sources—can allow for a new type of medical protocol. This protocol adapts to deep medical knowledge of an individual, both their current medical and proteomic state and their own trend and history over time, as a replacement for today's medical protocols that are rigid and rely on generalizations

based on populations, rather than the medical state of an individual. The practice of medicine in accordance with such new, individualized medical protocols is expected to provide significant cost savings while simultaneously improving average individual health.

[0017] In one implementation, a low-cost, minimally intrusive, proteomic-based test can be periodically given to each of a population of patients, for example, on respective schedules informed by individual genomics. This periodic, proteomic-based test can be used for a new approach to personalized healthcare. Results from this periodic test are used to create individualized longitudinal medical data, vastly improving the efficacy of any later diagnosis. The results can also be used provide individualized medical data that can be used to provide a personalized medical protocol using deep medical knowledge of an individual and his/her own trend and history over time, and provide early indications of onset of specific conditions that may require treatment or life-style modification. In essence, the proteomic-based test acts as the "gatekeeper" to personalized care. This approach shifts the current medical model from a reactive, symptom-based approach to a predictive/preventive approach based on personalized information.

[0018] In one implementation, the system can receive data from a simultaneous assay of thousands of proteins from a single drop of blood. Combining these data with the data in electronic healthcare records and other sources can provide both current and longitudinal information about each individualized patient. Using a data base of current medical knowledge and best practices, in combination with a set of "causality cases", which relate the sensed medical signals to current and predicted conditions and diagnoses, can provide rapid, accurate, and personalized diagnoses and recommendations for a healthcare treatment course of action for thousands of conditions simultaneously, all from a single blood test. This approach improves outcomes by basing diagnosis and recommendations on far more data than are available today from any affordable diagnostic procedure, and decreases costs by substituting an inexpensive test for a series of expensive ones, enabling earlier detection and intervention, mechanizing the sharing of test results across specialists and institutions, reducing the variation in clinical decision practices, and significantly reducing the broad range of individuals who are currently and unnecessarily screened, tested, and treated. Further, by maintaining all of this data in a centralized knowledge base, research into new causality cases can be substantially facilitated.

[0019] Healthcare also is subject to huge variation in practice. It is not enough to have personalized data that indicate a patient's condition. The recommended course-of-action must reflect appropriate practices consistently and incorporate evidence-based standards. Data indicate, however, that unjustified variation in medication practice accounts for between thirty to fifty percent of the total U.S. healthcare spending, in addition to causing harm and even death. Improving quality starts with reducing variation, rather than simply improving the population mean, and this is addressed by a learning health system in accordance with an aspect of some embodiments. At the scale of a national healthcare system, reducing variation first entails creating evidence for best practices based upon new findings, and incorporating that evidence into recommendations for patients and healthcare providers.

[0020] Some embodiments of the proposed system provide a number of advantages over the traditional healthcare approach. For example, the system allows consumers to know, based on objective data, when they need to enter the healthcare system to seek detailed diagnosis and treatment. At present, people make this determination with only a minimum of information, much of which is subjective or unreliable (e.g., how they "feel," "what hurts," temperature, etc.). The system also permits the healthcare practitioner at the patient's point-of-entry to rapidly evaluate changes and off-nominal conditions in the patient across a wide range of conditions and factors, based on minimally invasive technologies and data sources with a high degree of certainty, and route the patient to appropriate tests, screening, specialist practitioners, and procedures, thereby saving time, money, and frustration. Variation in healthcare practice can be materially reduced by conjoining detailed diagnostic information with evidence of clinical effectiveness applied to specific patient strata, allowing healthcare systems to improve and target delivery of care. The efficiency of the healthcare system is also increased by more proactively identifying and monitoring sick people earlier in their disease course, so that they come into treatment more effectively and with reduced use of more intensive treatments. Finally, practitioners can become more efficient and effective in their practice through periodic incorporation of new "causality cases," that is, the latest information about measureable health indicators that indicate and predict health factors, diseases, and tendencies, into a computer database, which can then be linked automatically to personalized healthcare options.

[0021] FIG. **1** illustrates a learning health system **10** in accordance with an example embodiment. It will be appreciated that the system can be implemented as machine executable instructions stored on a set of at least one non-transitory computer readable medium and executed by an associated processor, dedicated hardware, or a combination of dedicated hardware and software components. The system **10** includes a knowledge base **12** storing a record for each of a population of patients. The knowledge base **12** can include data received from one or both of a genomic data source **14**, representing a genetic mapping of an individual to locate genetic markers, and a biochemical data source **15**, representing the levels of various biochemical parameters for the individual as derived from biochemical assays. In accordance with an aspect of an example embodiment, the biochemical assays can be scheduled at regular intervals, such that even healthy patients are encouraged to provide a usable time series of biochemical parameters.

[0022] Accordingly, each record can include a time series of values for each a plurality of biochemical parameters taken from biochemical assays performed at scheduled intervals, a plurality of genetic markers, and a plurality of clinical parameters associated with the patient. The plurality of clinical parameters can be extracted, for example from electronic health record databases and include previous diagnoses and procedures, clinical observations, longitudinal biometric parameters (e.g., age, weight, blood pressure, temperature, glucose levels, etc.), and a family medical history. It will be appreciated that the population of patients can include, for each of a plurality of conditions of interest, a set of patients having the condition and a set of patients not having the condition. In addition to patient records, the knowledge base **12** can also contain statistics representing incident rates and measured outcomes for various disorders as well as data on causal links between available parameters and conditions drawn from medical research. In one implementation, a

research interface (not shown) can be provided for extracting data from available medical research, including an information extraction component to reduce an unstructured source of research, such as a journal article, into a template compatible with the knowledge base.

[0023] A baseline calculation component **22** is configured to calculate, for a given patient, an expected time series for a patient's biochemical parameters from at least the clinical parameters and the genomic parameters associated with the patient. While the system **10**, evaluates patients for a large number of conditions in parallel based on the biochemical assays, it will be appreciated that not every biochemical parameter is relevant to every situation and patient. Accordingly, the baseline calculation component **22** may selectively calculate an expected time series for each of a plural subset of the available biochemical parameters to preserve processing resources.

[0024] An analytics and modeling component **24** is configured to determine a deviation of the time series of values from the calculated expected time series and apply the deviation as an input to one or more predictive models associated with respective conditions of the plurality of conditions. Each predictive model can be derived from data in the knowledge base **12** associated with each of the set of patients having the condition and the set of patients not having the condition. For example, the predictive models can include appropriate supervised learning algorithms, such as regression models, artificial neural networks, support vector machines, and statistical classifiers, trained on data from the knowledge base. Each predictive model predicts a likelihood of one of a plurality of disorders according to deviations between the measured biometric parameters and the baseline from the deviation. For example, the predictive model can operate on one or more of a distance metric (e.g., Euclidian, Mahalanobis, Manhattan), difference between the measured and expected time series can be used as a predictive feature. Alternatively, the difference in the time series across a number of most recent data points can be used as features. In general, it will be appreciated that a number of descriptive statistics representing differences between two time series can be calculated, and any of these measures may be useful as a predictive feature. It will be appreciated that a given model can include parameters beside the calculated deviation as well, and that these additional parameters can be drawn from the knowledge base. In one implementation, the results of the predictive modeling can be supplemented with an actual course of treatment and a measured clinical outcome and fed back to the knowledge base **12** for use in generating addition causality cases.

[0025] In one implementation, the analytics and modeling component **24** can include a data mining component (not shown) configured to perform a plurality of unsupervised learning algorithms on the knowledge base **12** to determine at least one causality case relating one of the clinical parameters and the genomic parameters to the condition. The determined causality case can, once confirmed by subject matter experts, be used to refine existing predictive models or generate new predictive models. To facilitate review of the newly generated causality cases, the analytics and modeling component **24** can also include an analytics component (not shown) available to the user through a user interface **26** and configured to retrieve data from the knowledge base **12** and an associated database (not shown). Under the guidance of a subject matter expert, the analytics component can run various queries on the knowledge base **12** and the associated database to provide evidence supporting or refuting a given causality case. In one implementation, the analytics and modeling component **24** also includes a rules engine (not shown) that evaluates causality cases determined by the data mining component, according to an associated set of rules, to determine which variables, associated with the causality cases, present a highest likelihood of providing actionable results if evaluated with the analytics component. By limiting the analysis to parameters believed to be relevant, this rules engine can be used to conserve processing resources and decrease the likelihood of false positives in determining interrelationships among the data stored in the knowledge base **12**.

[0026] The user interface **26** is configured to provide the determined likelihood that the patient has the condition to a user. The user interface **26** can include visualization tools to allow the user to see a graphical comparison of the expected time series of biochemical parameter values and an actual time series of biochemical parameter values. In one implementation, the user interface **26** includes a patient dashboard (not shown) configured to communicate each of the determined likelihood of the condition, a healthcare treatment course of action, and/or a scheduled next biochemical assay. Accordingly, the patient can be instructed to enter the healthcare system at an appropriate time based on the biochemical analysis. The patient dashboard may also include links to information about any diagnosed disorders and recommended treatment option.

[0027] The user interface **26** can also include a clinician decision support component (not shown) configured to communicate a recommended protocol of care to a clinician based on the determined likelihood that a patient has a condition. By making the data from the knowledge base **12** and predictive models available to all stakeholders in the healthcare system, the user interface **26** can ensure transparency of the recommended courses of actions to clinicians and patients and ensure that researchers have easy access to data stored in the knowledge base to allow for the generation of new causality cases and predictive models.

[0028] FIG. **2** illustrates one implementation of a learning healthcare information processing system **50** in accordance with an example embodiment. In the illustrated implementation, the system **50** receives data from a plurality of data sources **52-56** external to the system, indicated in a dashed outline, through respective data interfaces **62-65** and processes that data to provide recommendations to patients, clinicians, and researches based on accumulated data from these resources. A first data source **52** includes electronic medical record databases, with each electronic medical record database containing medical data for a plurality of patients comprising, for example, previous diagnoses and procedures, clinical observations, longitudinal biometric parameters, and a family medical history. Examples of electronic medical record databases that could be compatible with the information processing system can include the Armed Forces Health Longitudinal Technology Application (AHLTA), the Veterans Health Information Systems and Technology Architecture (VISTA), and similar such databases maintained by large healthcare organizations with a significant patient base. Records from these databases can be provided through an electronic medical record database (EMRD) interface **62** to convert the retrieved records to an appropriate format for a knowledge base **68** associated with the healthcare information processing system **50**. In one implementation, the full

record stored in the electronic medical record database is truncated by the interface to a set of clinically relevant observations.

[0029] The data sources can also include a biometric assay taken from a large population of patients. In the illustrated implementation, a proteomic assay **53** is utilized, but it will be appreciated that other biometric assays can also utilized, including pharmacogenomic assays, metabolomic assays, epigenomic assays, as well as interactomic, transcriptomic, and microbiomic data. In one implementation, the proteomic assay **53** can detect around ten thousand proteins and be administered at scheduled intervals to provide a time series of blood levels for each of the ten thousand proteins. An assay interface **63** may be configured to format the assay data for the knowledge base **68** and associate identifying information of the assays with corresponding patient records in the knowledge base. The assay interface **63** may also be configured to normalize the proteomic data to a scale utilized by the knowledge base **68**. In one implementation, the proteomic assay **53** can be reduced to a vector of clinically important features to be provided to the knowledge base **68**, with the full assay compressed and stored in a separate mass storage with time-stamped line from the patient file to the full assay.

[0030] The system **50** can also utilize genomic data **54** from a population of patients. For example, the genomic data **54** can be captured for each patient via an appropriate assay and provided to the system through a genomic interface **64**. The genomic interface **64** extracts known genetic markers from the genome, formats the extracted data for the knowledge base **68** and associates identifying information of the genetic information with corresponding patient records in the knowledge base **68**, for example, via a link from the patient record to the extracted markers.

[0031] Information and statistics from population health data sources **55** can be provided through a health data interface **65**. Population health data sources **55** include, for example, structured or semi-structured data representing incident rates and measured outcomes for various disorders. Examples of population health data sources **55** can include the Surveillance, Epidemiology, and End Results (SEER) program maintained by the National Cancer Institute, the Behavioral Risk Factor Surveillance System (BRFSS) maintained by the Centers for Disease Control and Prevention, the Healthcare Cost and Utilization Project (HCUP) maintained by the Agency for Healthcare Research and Quality, and the Food and Drug Administration Adverse Event Reporting System (FAERS). The health data interface **65** may be configured to convert the structured and semi-structured data maintained in these resources into an appropriate format for a knowledge base **68** associated with the system **50**.

[0032] Finally, data concerning causality factors for various disorders can be captured from medical research data **56** (or literature) and provided to the knowledge base **68** through a research interface **66**. Exemplary sources of medical research data (or literature) can include the Medline collection from the National Library of Medicine, the PubMed collection, the GenBank sequence database, and the Gene Expression Omnibus repository maintained by the National Center for Biotechnology, the ArrayExpress and InterPro databases maintained by the European Bioinformatics Institute, the ImmPort immunology database and the Database for Annotation, Visualization, and Integrated Discovery maintained by the National Institute of Allergy and Infectious Diseases, and the UniProt knowledge bases, as well as Internet publications, such as Wikipedia, WebMD, health organization websites, and similar information sources. Since the medical research data **56** can include unstructured data, the research interface **66** can include an information extraction component to reduce an unstructured source of research, such as a journal article, into a format compatible with the knowledge base **68**. The information extraction component may be configured to break down the unstructured source into individual words or phrases, interpret the context and meaning of the various words or phrases, and use the extracted information to generate a template representing the unstructured source. In one implementation, the generated template can be reviewed by a human expert in a field relevant to the unstructured source to ensure that the information provided to the knowledge base **68** is accurate.

[0033] The knowledge base **68** can be implemented as a massively parallel system to provide a low response time and significant scalability for increasing amounts of data. In one implementation, the knowledge base **68** can include a plurality of geographically remote regional caches, such that data associated with a given patient population is easily and quickly accessible to local clinicians. Each cache is operatively connected to a master knowledge base to allow for analysis of the data in aggregate for researchers, and can be fed data by the master knowledge base according to scheduled appointments. Requests from emergency rooms and other unscheduled sources of care can be prioritized to allow real-time or near real-time access to patient information. Information in the caches can be replaced such that data that has been least recently used is replaced. The knowledge base **68** may store any or all of clinical observations, proteomics, and genomics from various patients, including data for both a healthy population and a population of individuals that have disease syndromes, allergic reactions, or some other undesirable clinical outcome. The knowledge base **68** may include a mixture of active data in the knowledge base, for example, triggers supported by a notification subsystem, and a rule base using a scalable rules engine.

[0034] In accordance with an aspect of an example embodiment, an analytics and modeling component **70** can interact with the knowledge base **68** to determine relationships among the data. The function of the analytics and modeling component **70** can be roughly divided into what is referred to herein as "forward analytics," in which the likelihood of any of a variety of conditions for a given patient can be predicted by comparing data associated with the patient to data from the larger population, and "backwards analytics," in which data from a large population of patients is mined to determine relationships between clinical parameters and identified conditions.

[0035] In one example of a forward analytics process, a baseline calculator **72** can be configured to calculate, for a given patient, an expected longitudinal progression of a biometric parameter, such as the levels of clinically relevant proteins from the proteomic assays **53**. In general, the baseline is determined according to an amalgamation of biometric parameters recorded for cohorts of similarly situated patients, that is, patients who either live or work in the same location as the patient, have similar genetic markers, have similar medical histories, or otherwise have clinically relevant parameters in common with the patient. The baseline can be calculated, for example, via one or more statistical models that utilize this data to determine what an appropriate level or range of levels for each of a plurality of clinical relevant biometric param-

eters would be for the patient given his or her medical history, including not only diagnoses and conditions, but also longitudinally recorded parameters such as weight, blood pressure, and glucose levels, the patient's genetics, and the patient's biographical parameters, such as age and location of residence.

[0036] It will be appreciated that the knowledge base **68** is expected to include a large number of patient records. Accordingly, in one implementation, for each protein, the knowledge base **68** can simply be queried to return all or a predetermined number of records having all or a threshold number of biometric parameters relevant to establishing a baseline for that protein within a defined range around the patient's values for the biometric parameters. The time series for the protein can be averaged across all retrieved records to provide the baseline.

[0037] Once the baseline for biometric parameters has been calculated, each of the calculated baselines and a measured plurality of series of biometric parameters can be provided to a series of predictive models **73**. The predictive models **73** can include any of appropriate supervised learning algorithms, such as regression models, artificial neural networks, support vector machines, and statistical classifiers, which may be configured to predict a likelihood of one of a plurality of disorders according to deviations between the measured biometric parameters and the baseline. In one implementation, the predictive models **73** can include an analogical reasoning algorithm that compares the patient's measured biometric parameters, genetic markers, and clinical observation by a physician to sets of biometric parameters, genetic data, and observations from other patients for whom the presence or absence of a condition is known to determine a likelihood that the patient may experience the condition. The conditions evaluated by the predictive models **73** can be drawn from one or more disorder ontologies **74**. A disorder ontology can be compiled from existing resources such as the International classification of Diseases (ICD), the Diagnostic and Statistical Manual of Mental Disorders (DSM), the Medical Dictionary for Regulator Activities (MedDRA), BioOntology, and the Open Biological and Biomedical Ontologies.

[0038] It will be appreciated that the system is not limited to a rigid disorder ontology. Many pathological states are defined by symptoms, leading to imprecise classifications. For example, it is likely chronic fatigue syndrome is an umbrella class for a host of different, possibly unrelated pathologies. Other disorders, such as autism and schizophrenia, exist along a spectrum of symptom intensities, which may also group states with different underlying causes. To this end, the system can provide a complementary way to define pathologies by the underlying biological data, rather than these imprecise symptom presentations. Specifically, unique combinations of biological data (e.g., genomic, proteomic, metabolomic) will be statistically processed and associated with outcomes and symptoms to provide more precise pathological classifications. By linking the biological state directly with the pathological classification, treatments can be assigned that directly address the underlying biological cause of symptoms.

[0039] The backwards analytics performed by the system can include one or more data mining algorithms **76** that analyze data stored in the knowledge base **68** for connections between previously unconnected predictors. The connections determined from the data mining algorithms **76** can be utilized to define new causality cases for use in the forward analytics performed by the system. This process can be fully automated, with new causality cases integrated into the predictive models **73** automatically, or in a semi-supervised fashion, in which each newly discovered causality case is reviewed by a subject matter expert before being incorporated into the predictive models. The data mining algorithms **76** can include, for example, anomaly detection algorithms, association rule learning, clustering algorithms, and sequential pattern mining.

[0040] In one implementation, new causality cases are generated as treatments, protein expression changes, and outcomes and then iteratively input into the knowledge base as adjustments of any of correlations, scoring, recommendations, and weighing of causalities. This information allows researchers to evaluate hypotheses and suggests subsequent research, such as identifying new biomarkers. As the system ingests and process new data, interesting relationships will emerge as analytics and data mining algorithms are automatically run. Researchers will be able to log in and bring up an updated list of trends and statistically significant relationships that have emerged. These lists serve as an opportunity for researcher to explore the meaning behind relationships and develop hypotheses for future research projects, thereby accelerating research productivity.

[0041] The system **50** also includes an analytics component **77** configured to retrieve data from the knowledge base **68** to confirm causality cases identified by the data mining component **76** and researchers. To this end, the analytics component **77** can include integration with the Basic Local Alignment Search Tool to find commonalities between a given genetic sequence and library sequences as well as various custom analytics algorithms that automatically discover correlations between baseline protein assays and diagnosed diseases later in life, automatically discover correlations between baseline protein assays and genetic sequences, and discover new genetic markers by correlating genome with diseases or allergic reactions. Further, the analytics component **77** can include an algorithm for tracking protein level changes associated with clinical treatment outcomes to explore the biological relationship to the proteins and disease, relate to genetic mutations, and develop more effective drugs using knowledge of the causal biological interactions. Additionally, the analytics component **77** can include statistical analysis and analytic tools to assist researchers in confirming hypotheses generated by the data mining component **76** and the other analytic tools. In one implementation, the analytic tools can include advanced signal processing algorithms to extract correlations from noisy data and neural spike metrics.

[0042] Medications are often prescribed despite known side effects. The inventors have determined that the knowledge of who would be most likely to present with side effects is both within the capability of a learning healthcare information processing system **50** in accordance with an example embodiment and of considerable value, especially when alternative medications exist. Similarly, it would be possible to predict who may respond well and/or without side effects. To this end, the knowledge base **68** will be designed to collect outcome data fed back from the system **50**. Positive and non-adverse outcomes may be unique for specific genetic mutations or baseline protein levels, and can therefore serve as additional information for supporting practitioner treatment recommendations and suggest areas of research and discovery. Outcomes will therefore be linked to specific

genetic mutations and protein levels for individual patients to allow for prediction of patient response from proteomics and genomics.

[0043] It will be appreciated that the system may iteratively test hundreds to thousands of variables for significant correlations. While inclusion of more variables increases the probability of discovering insightful, actionable relationships, it also increases the probability of false positives. The standard approach to correct for this problem of "multiple comparisons" is to multiply significance test values by some corrective factor. For instance, in Bonferroni correction, the p value is multiplied by the number of independent tests performed. Unfortunately, this results in increasing the probability of false negatives. Therefore, the more independent significance tests run, the more interesting relationships will be buried into the background noise of non-significance.

[0044] In accordance with an aspect of an example embodiment, a rules engine **78** includes a mix of expert and machine-generated rules and weights are continuously deployed and tuned that learn which types of variables present the best probability for insightful or actionable results prior to analysis. The automated rules engine **78** is expected to supplement the efforts of expert researchers in determining what tests to run prior to a single research experiment. Reducing the overall number of tests will also optimize processing performance. Ultimately, the rules engine **78** mediates between statistical design and machine intelligence in developing healthcare-based statistical rules.

[0045] The results of the various analytics and modeling processes **70** can be provided to the knowledge base **68** to be added to the patient's record as well as any relevant medical databases **52**. These records will generally be supplemented with a treatment record and a patient outcome once these factors are known. The results are also provided to respective visualization components **82-84**. In one implementation, a researcher visualization component **82** presents the knowledge discovered by analytics component **77** (or analytics search engine) applied to the genetic and proteomic data collected in this system in a visual fashion that is readily comprehendible. The researcher visualization component **82** can provide a user interface for analytic search algorithms to discover correlations between protein assays, genetic sequences, and diagnoses. The researcher visualization component **82** can also include various display and graphical manipulation tools to view protein level changes associated with clinical treatment outcomes so that the researcher can explore the biological relationship to the proteins and disease, relate the outcomes and proteins to genetic mutations, and develop more effective drugs using knowledge of the causal biological interactions. The researcher visualization component **82** can also provide a periodic report of emergent statistical associations between variables across databases as outcome data is fed back into the system, as well as simply access to relevant data and findings from valuable scientific databases.

[0046] A clinician decision support component **83** allows a clinician to access results of forward analytic processes for a given patient and relevant support information. For example, the clinician decision support component **83** can display to the clinician a list of diseases consistent with the patient's clinical observations, a latest protein assay, geographic location, and relevant environmental factors in likelihood order. The clinician can also instruct the clinician decision support component **83** to display a comparison of the current protein assay with the measured or imputed baseline assay, and/or a comparison the patient's history of protein assays with the normal time series of expected protein assays. The clinician decision support component **83** can also display values significant in the calculated baseline assay, such as markers from the patient's genome and exogeneous variables such as gender, weight, and age. The decision support component **83** can also notify a clinician when a patient has not been in contact with the office for a predetermined period of time or has failed to provide a scheduled biochemical assay. In one implementation, this notification can be complied over a period of time and provided in list form to avoid overwhelming the clinician.

[0047] A patient dashboard **84** can present the results of forward analytic processes and supporting data to a patient. To this end, the patient can be presented with any findings of elevated risk, the genomic, biochemical, and clinical parameters supporting the findings, and links to information related to the disorder or outcome associated with the elevated risk, potential treatments, and the parameters supporting the findings. For example, a patient could be provided with a link to information about the side effects associated with a prescribed medication. Any recommendations on health screening results and potential courses of action provided to the patient can include certainty-weighting and risk-based weighting to facilitate informed decisions by the patient. The patient dashboard **84** can also provide an interface for the patient to ask questions, via an encrypted e-mail service, such as S/MIME, to a clinician to clarify information received during an earlier visit. The patient dashboard **84** can also provide reminders to the patient for scheduled biochemical assays, appointments with clinicians, or to take or refill medicines. In one implementation, the patient can record observations of symptoms through the patient dashboard **84** as well review, correct, and supplement data in the patient's electronic medical record.

[0048] It will be appreciated that, after a medical outcome is known for a given patient, the knowledge base **68** can be updated to reflect the new result. To this end, a set of measured clinical outcomes **86** can be provided to the knowledge base **68** to augment the existing patient data. The measured clinical outcome can reflect, for example, whether the patient has a condition of interest after a set period of time after the prediction. Along with new medical research and new patient records entering the system, these patient outcomes **86** can provide the knowledge base **68** with the basis for new causality cases to be discovered by the analytics and modeling component **70**.

[0049] In one example use case, a lab draws a patient's blood and provides the genomic **54** and proteomic **53** assays. In one implementation, the proteomic assay **53** can be performed using a low-cost, easily repeatable assay that can simultaneously determine levels for thousands of proteins from a small blood sample with a relatively low overhead for each testing site, allowing the test to be widely accessible. Since the test is designed to be low-cost and accessible, longitudinal data for a large population of individuals could be efficiently compiled. Once the data are normalized and processed, it can be determined if the patient's protein levels, taken in view of clinical observations of the patient, and genetic markers, indicate an enhanced likelihood of a given condition through the predictive models **73**. In this example use case, it is determined that the patient has a genetic marker associated with a high risk of a particular type of cancer and elevated proteins associated with that type of cancer. The

knowledge base **68** can include information indicating that a survival rate for this type of cancer is significantly higher when diagnosed within three months.

[0050] Once the enhanced risk of cancer is identified, a report is generated and the patient is notified. The patient can log into the patient dashboard **84** to view the report, which can include the diagnosis and links to information about the disorder, the proteomic and genetic data used to identify the elevated risk, and potential treatments. The report can also include a recommendation that the patient should schedule a visit with an oncologist. Similarly, a clinician associated with the patient, such as a family doctor and/or an oncologist treating the patient, can receive an alert through the clinician decision support component **83**. The alert can be linked to a summary report, including an overall risk score associated with the diagnosis, the specific genetic markers and proteins relied upon for the diagnosis, with links to pertinent research, and visualization tools for viewing this data. The clinician's treatment decisions and the clinical outcome can be fed back into the knowledge base **68**, along with information from follow-up visits, and comments from the patient and the clinician. These findings can then be made available to researchers, through the various tools available through the researcher visualization component **82**, for further analysis.

[0051] In a second example use case, a researcher might view a summary report showing recently emergent data trends and find a high prevalence of non-adverse Pramipexole response for patients with elevated proteins associated with food allergies. The researcher could then search text within available journal articles via a text miner in the researcher visualization component **82** as well as data within the knowledge base and affiliated data sources for known relationships between a genetic mutation shared by patients who respond well to Pramipexole and the elevated protein. Assuming no known relationship is found, the researcher could develop and conduct tests to search for unidentified proteins that may also be elevated, with the hypothesis that any identified proteins might be elevated in some patients with fibromyalgia and cause increased sensitivity to allergies in patients with the genetic mutation.

[0052] The researcher can provide the results of the research and the determined hypothesis to the knowledge base and request that the proteomics lab develop an aptamer for the identified protein. Once the aptamer is generated, results from multiple patients undergoing their scheduled proteomic assays can be aggregated to confirm or refute the researcher's hypothesis. It will be appreciated that other information from the knowledge base **68** can be mined or queried to provide evidence supporting or refuting the hypothesis. Assuming that it is confirmed, further research can be performed, for example, via queries of the knowledge base **68** through the researcher visualization component **82**, to find a drug that can be employed to reduce levels of this protein. This finding can then be fed back to the knowledge base **68** as a known relationship between the drug and fibromyalgia.

[0053] After all this has happened, a patient diagnosed with fibromyalgia might be determined by a clinician to be responding poorly to common medications. The clinician may wish to prescribe a dopamine agonist, but is concerned about efficacy and side effects. The clinician may instruct the patient to have blood drawn for a genomic or proteomic assay or utilize existing genomic and proteomic data from the scheduled assays for the patient. From this information, it might be determined that the patient shares the generic mutation associated with patients who respond to the dopamine agonist Pramipexole, but lacks a marker associated with patients who respond well to the dopamine agonist Ropinirole. The protein associated with increased sensitivity to allergies may also be found to be elevated in the patient. Information in the knowledge base can be automatically retrieved and provided to the clinician and the patient indicating that the protein expression level has been reduced in sixty percent of cases in which gluten has been removed from the diet.

[0054] All of this information can be provided to the clinician at the clinician decision support component **83** with a plurality of treatment options, each having an associated score representing the likelihood, generated from the predictive models **73**, that the treatment will lead to a favorable clinical outcome. Two high-score treatments might include placing the patient on a gluten-free diet and prescribing Pramipexole. Accordingly, the clinician might select either option or combine the options, with the dosage of Pramipexole reduced to account for any beneficial effects of the gluten-free diet. To the extent that Pramipexole is prescribed, levels of proteins associated with the side effects can be tracked, for example, with the frequency of the patient's proteomic assays increased until the effects of the drug are clear.

[0055] The patient can also be provided with a summary report with the diagnosis, the treatment decision made by the clinician, and an appointment schedule. This report can include links to information related to diagnosis and treatment, such as online resources that describe fibromyalgia, side effects and interactions associated with the drug, and advice for pursuing a gluten-free diet. Information can also be provided for genetic markers and protein levels used in the diagnosis. The patient can use the patient dashboard **84** to record symptom levels, such as pain and fatigue, over time. Additionally, the level for the relevant proteins can be tracked over time to maintain the patient's awareness of their progress and possibly encourage compliance. The patient's reported symptoms and the clinician's observations can be fed back into the knowledge base **68** for use in evaluating the efficacy of the selected treatment and the prevalence of any side effects.

[0056] The illustrated system **50** provides a number of advantages. For example, the system enables economy of scale by testing numerous causality cases from a single blood sample. The system is capable of quantifying, aggregating, and disclosing measurement and recommendation certainty, including biosensor variability and any other potential source of error to ensure that the confidence associated with recommendations is meaningful to the patient and clinician, and the system can improve recommendation accuracy over time. As a result, the system can have sufficiently high reliability, capacity, and availability to support mission-critical use and scale with expected data increases over time, both in the available causality cases and the inclusion of new target populations.

[0057] In view of the foregoing structural and functional features described above in FIGS. **1** and **2**, an example method will be better appreciated with reference to FIGS. **3** and **4**. While, for purposes of simplicity of explanation, the method of FIGS. **3** and **4** are shown and described as executing serially, it is to be understood and appreciated that example embodiments are not limited by the illustrated order,

as some actions could in other examples occur in different orders and/or concurrently from that shown and described herein.

[0058] FIG. **3** illustrates a method **100** for providing personalized healthcare support in accordance with an example embodiment. At **102**, biochemical assays are conducted, at scheduled intervals, on a blood sample taken from an individual to provide a time series of values for each of a plurality of biochemical parameters. In one implementation, the biochemical assay is a baseline protein assay measuring a large number of protein levels from a single drop of blood, such that the assay can be low-cost and easily performed outside of a clinical environment. Accordingly, patient access to the biochemical assay can be made convenient to encourage compliance in generating a complete time series of values.

[0059] At **104**, a plurality of clinical parameters, associated with the individual, from a knowledge base are extracted. The parameters can be categorical, such as diagnosed disorders or clinical observations of symptoms, as well as interval or ratio data, such as age, temperature, weight, blood pressure, cholesterol levels, and other such data. In one implementation, a plurality of cohort parameters can be extracted from respective series of biochemical assays in the knowledge base from record representing individuals who are associated with the individual. For example, the cohort parameters can include averaged time series of a given biochemical parameters across one or more of a set of people who are related to the patient, a set of people who live or work near the patient, and a set of people who share a condition or genetic marker in common with the patient.

[0060] At **106**, a plurality of genomic parameters are determined for the individual. In one implementation, this can be done from the same blood sample used to derive the biochemical parameters. It will be appreciated that each of the time series of values and the plurality of genomic parameters can be stored in the knowledge base such that the knowledge base contains biochemical assays, genomic parameters, and clinical parameters for a population of patients.

[0061] Chemical and biological analysis is typically used to determine characteristic features of a biological sample. The features could then be transformed into representative quantitative values and provided to an information processing system for calculation and statistical analysis including data mining, machine learning and other computational functions. Many methods are known to those skilled in the art of biochemistry for determining signature features derived from biomedical samples and for comparing the features against other samples or across reference data sets. For example, comparing multiple mass spectra from different biological samples and identifying common features across the samples can be used as a reference condition, whereas identifying distinguishing features could serve as potential biomarkers for detection of an anomalous condition. The features can be compared across individuals and/or temporally for a specific individual. As described herein, various types of biochemical parameters are known and are available for use in analytics. Some example embodiments produce a greatly improved biochemical signature feature by combining multiple biochemical assays of different types and including a temporal component to the signature.

[0062] At **108**, an expected time series is calculated for each of a plural subset of the plurality of biochemical parameters from at least the clinical parameters and the genomic parameters. For example, the expected time series can be determined as a weighted combination of time series values from patients having various characteristics associated with the clinical and genomic parameters of the patient, with the weight selected on a similarity, determined for example as a multivariate distance metric, between the patient and various other patients in the knowledge base. Alternatively, the knowledge base can be queried for patients having values for relevant biomedical parameters within a predefined range of the patient's values. The expected time series can be an unweighted average (e.g., mean or median) of the retrieved records.

[0063] In one embodiment, some example embodiments enable calculation of an expected time series by first representing the biochemical assays as feature vectors, each having a plurality of coefficients that correspond to a set of biochemical parameters. It then generates sets of clusters comprising pathological feature vectors derived from a large population of patients having a certain condition. The feature vector members of each specific cluster have signature similarities measured by a Euclidean distance calculation between the feature vector and the cluster centroid. Similarly, a well known unsupervised clustering method such as the K-means clustering algorithm can be used. Yet another alternative is to use a Mahalanobis distance for measuring similarity (correlation) with the advantage of being generally scale invariant. Furthermore, the combination of data sets and feature vectors that are associated with the biochemical assays can be represented in multiple dimensions as multivariate vectors or matrices and the clustering and distance calculations can be performed by fusing and correlating the multivariate vectors or matrices across the biochemical assay feature vector sets. There are many more distance measures and feature vector types that are known to those skilled in the art of statistical analysis. The embodiment described herein is shown only by way of example and it is understood that various alternatives can be used without a loss of generality.

[0064] The temporal aspect is now introduced where the sequences of cluster centroids are tracked over time and characterized by a cluster transition path. The time series value of an individual patient's biochemical assays can be compared to the expected time series by computing the distances of the associated feature vectors to the nearest-neighbor clusters, as each new blood sample is taken (e.g. on an annual basis). As an enhancement to the calculation, unnecessary features that are abundant in large bioinformatics data sets, and that do not materially contribute to system outcome/value, can be removed, thereby improving the results. Many other methods are available for performing supervised machine learning and data mining that are well known to those skilled in the art of data analysis.

[0065] At **110**, for each of the plural subset of biochemical parameters, the time series of values representing the individual is compared to the calculated expected time series to determine a likelihood of each of a plurality of conditions for the individual. For example, a significant deviation of the time series of values from the calculated expected time series can be determined and applied as an input to a predictive model associated with one of the plurality of conditions, with the predictive model being configured to determine the likelihood of the associated one of the plurality of conditions from at least one parameter derived from the significant deviation. In one implementation, predictive models can be generated and refined by unsupervised learning processes mediated by subject matter experts. For example, a data mining algorithm

can be applied to the knowledge base to identify at least one causality case relating one of the clinical parameters, the genomic parameters, and the cohort parameters to a condition. Once the causality case has been reviewed and verified by subject matter experts, for example, via the application of one or more analytic tools to retrieve evidence from the knowledge base, a predictive model can be refined or generated according to the identified causality case.

[0066] At **112**, the likelihood of at least one of the plurality of conditions is communicated to a user. In one implementation, the user is the individual and the communication can include any or all of a healthcare treatment course of action, based on the communicated likelihood of the at least one condition, an instruction to the individual when a next biochemical assay should be scheduled based on the communicated likelihood of the at least one condition, and a recommendation as to a type of healthcare practitioner from which the individual should seek treatment. In another implementation, the user is a clinician and the communication includes a recommended protocol of care to the clinician based on the communicated likelihood of the at least one condition.

[0067] In one implementation, the communication is provided through a user interface that is configured to display to the user, for a selected one of the plural subset of biochemical parameters, a graphical representation of each of the time series representing the individual for the selected biochemical parameter and the calculated expected time series for the selected biochemical parameter, such that the calculated expected time series can be easily compared to measured values from the scheduled biochemical assays. The user interface can allow a clinician to select a new value from a selected one of the parameters used to calculate the expected time series and alter the graphical representation of the expected time series to reflect the new value of the selected parameter. This can allow the clinician to determine the effects of possible treatments and lifestyle modifications on a patient's health. It will further be appreciated that these tools can be made available to researchers for assistance in searching for new causality cases.

[0068] FIG. **4** illustrates a method **150** for discovering and applying new causality cases in a learning healthcare system in accordance with an aspect of an example embodiment. At **152**, a knowledge base associated with the learning healthcare system can be updated with measured clinical outcomes for patients in the knowledge base. For example, the measured outcomes can be entered directly into the system via a user interface or retrieved from a medical records database. At **154**, unsupervised learning processes are performed on the knowledge base to discover potential causality cases. The unsupervised learning processes can include, for example, anomaly detection algorithms, association rule learning, clustering algorithms, and sequential pattern mining.

[0069] At **156**, an analyst is prompted to perform one or more analytics on the knowledge base to confirm a potential causality case. For example, a researcher might be provided with a summary report showing recently emergent data trends, with the appropriate supporting data available for review as text or a graphical representation. The researcher could then search text within available journal articles via a text miner or formulate one or more queries of related data in the knowledge base to develop a hypothesis for any emergent trends found to be of interest. The researcher could then develop and conduct tests to confirm the hypothesis, with the results of the research and the determined hypothesis provided to the knowledge base. If the hypothesis representing the causality case is confirmed, one or more predictive models are updated at **158** to reflect the new finding.

[0070] FIG. **5** is a schematic block diagram illustrating an exemplary system **200** of hardware components capable of implementing examples of the systems and methods disclosed in FIGS. **1-4**, such as the learning health system illustrated in FIGS. **1** and **2**. The system **200** can include various systems and subsystems. The system **200** can be a personal computer, a laptop computer, a mobile device, a tablet computer, a workstation, a computer system, an appliance, an application-specific integrated circuit (ASIC), a server, a server blade center, a server farm, etc.

[0071] The system **200** can include a system bus **202**, a processing unit **204**, a system memory **206**, memory devices **208** and **210**, a communication interface **212** (e.g., a network interface), a communication link **214**, a display **216** (e.g., a video screen), and an input device **218** (e.g., a keyboard and/or a mouse). The system bus **202** can be in communication with the processing unit **204** and the system memory **206**. The additional memory devices **208** and **210**, such as a hard disk drive, server, stand alone database, or other non-volatile memory, can also be in communication with the system bus **202**. The system bus **202** interconnects the processing unit **204**, the memory devices **206-210**, the communication interface **212**, the display **216**, and the input device **218**. In some examples, the system bus **202** also interconnects an additional port (not shown), such as a universal serial bus (USB) port.

[0072] The processing unit **204** can be a computing device and can include an application-specific integrated circuit (ASIC). The processing unit **204** executes a set of instructions to implement the operations of examples disclosed herein. The processing unit can include a processing core. Although one processing unit **204** is shown in FIG. **5**, it should be appreciated that the processing unit **204** may be distributed in some examples. Thus, for example, multiple instances of processing circuitry may be embodied at a plurality of different locations within an enterprise or within a network and the various instances of processing circuitry may communicate and combine their respective processing capabilities to embody the processing unit **204** of the system **200**. Similarly, other components of FIG. **5** should also be appreciated to have the potential for multiplicity and distribution in various different example implementations.

[0073] The additional memory devices **206**, **208** and **210** can store data, programs, instructions, database queries in text or compiled form, and any other information that can be needed to operate a computer. The memories **206**, **208** and **210** can be implemented as computer-readable media (integrated or removable) such as a memory card, disk drive, compact disk (CD), or server accessible over a network. In certain examples, the memories **206**, **208** and **210** can comprise text, images, video, and/or audio, portions of which can be available in formats comprehensible to human beings.

[0074] Additionally or alternatively, the system **200** can access an external data source or query source through the communication interface **212**, which can communicate with the system bus **202** and the communication link **214**.

[0075] In operation, the system **200** can be used to implement one or more parts of a learning health system in accordance with an example embodiment. Computer executable logic for implementing the composite applications testing system resides on one or more of the system memory **206**, and the memory devices **208**, **210** in accordance with certain

examples. The processing unit **204** executes one or more computer executable instructions originating from the system memory **206** and the memory devices **208** and **210**. The term "computer readable medium" as used herein refers to a medium that participates in providing instructions to the processing unit **204** for execution.

[0076] As discussed above, the system **200** may be configured to implement the methods and systems of FIGS. **1-4**, which generally may incorporate genomic data and proteomic data to improve healthcare outcomes. However, some example embodiments may also achieve improved healthcare outcomes using clinical data, genomic data and other relevant types of data. The clinical data may include patient health record information (e.g., electronic health record (EHR) information) and laboratory data. However, laboratory data may be considered to be distinct from clinical data in some cases since, for example, the laboratory data may have a different format (e.g., pdf) within an EHR. The genomic data may include sequenced genomes of patients. The other relevant types of data may include, for example, research data, and text or publication data from various biomedical literature sources (e.g., PubMed sources). Thus, for example, the system **200** may be configured to generate a personalized healthcare and informatics system that can use genetic data to drive healthcare outcome improvements. Other supplemental information may also be added. However, in accordance with this example embodiment, sizable gains in healthcare outcome improvement can be obtained by employing heterogeneous data sources that include genomic data, and then employing powerful analytic tools and visualization tools to improve patient healthcare and status.

[0077] In some embodiments, the system **200** may be employed to embody a closed feedback loop architecture for providing the data, analytics, modeling and interface capabilities to enable association of multiple data sources to provide insight into patient health risks and conditions, while also enabling the data sources to be dynamically updated and further used to support further research. The closed loop feedback architecture may be mainly constructed using open source components. However, proprietary solutions may be substituted for some components where desired. The analytic capabilities of the system **200** may be employed to associate clinical, genomic and proteomic biomarkers to patient health record data to provide the insight. In particular, the system **200** may provide massive amounts of genomics data that includes patient de-identified (e.g., by removing the identities of the individuals with which such data is associated) genomic data and identified genomic data along with identified and de-identified clinical data to be stored and analyzed to enable further research and healthcare decision support to be conducted using cloud based and scalable resources. The system **200** is dynamic, and thus is configured to discover and update clinical, genomic and/or proteomic interpretations and algorithms continuously. As discussed above, the data associated with the system **200** can be compared against clinical outcomes. Some embodiments may further provide a user interface to deliver visualizations of customized results or responses to queries to patients, physicians, and researchers to identify and diagnose emerging diseases and guide treatment interventions.

[0078] In an example embodiment, the architecture that the system **200** embodies or supports may have multiple layers including a storage/data layer, an analytics layer and an application layer supported on a cloud-based platform. A block diagram of such a platform is shown, for example, in FIG. **6**. As shown in FIG. **6**, a data platform **300** may be provided to support or embody the storage/data layer. Because the storage/data layer may include data of various different types and structures, the data platform **300** may include a broad variety of databases to support various different storage and retrieval mechanisms for corresponding different data types and structures. For example, the data platform **300** may include a relational database management system (RDBMS) **302** that is based on a relational database model. The data platform **300** may also include a NoSQL database **304** to provide a mechanism for storage and retrieval of data that is modeled in means other than tabular relations used in relational databases. In some cases, the data platform **300** may include a Hadoop component **306**, which may provide an open-source framework for distributed storage and processing of very large data sets on computer clusters. The Hadoop component **306** may be particularly useful in supporting "big data" analytics on genome data and/or proteomics data. In an example embodiment, the Hadoop component **306** may be supplemented with or replaced by Google Genomics for storage of massive amounts of genomic data. In some embodiments, the data platform **300** may include a Greenplum database or other analytics database **308**, and one or more composite data virtualization components and/or business intelligence and analytics platform (e.g., BI platform **310**) or other data integration platforms (e.g., Pentaho and/or SAS Access). Thus, it should be appreciated that the data platform **300** can support multiple types and structures of data and mechanisms for accessing such data. The data platform **300** can therefore be a scalable platform to provide data sources including structured and unstructured data that can be analyzed using an analytics platform **320** that may support or embody the analytics layer. The data may include patient health record information (e.g., EHR information), research data, genomic data (e.g., patients' sequenced genomes), and text or publication data from various biomedical literature sources (e.g., PubMed sources).

[0079] The analytics platform **320** may include analytics tools configured to interface with the various different data sources. Because the data sources are so diverse, the analytics tools must be equally diverse to be able to analyze the data and make correlations where appropriate. Moreover, the correlations to be made in the context of such massive amounts of data need to be made relative to user input (e.g., a query) in the form of a response that can be provided in real-time or near real-time. Thus, the analytics platform **320** provides analytical tools to respond to user queries by analyzing large and diverse data sets relative to a particular condition or medical issue to identify relevant correlations and/or patterns in the data based on the query received. Once the relevant correlations and/or patterns are identified, they can be processed according to human-defined and/or machine learned rules corresponding to risk models defined for various conditions or relative to certain issues. Thus, the analytics platform **320** is configured to perform fast analytics on massive amounts of data (e.g., multiple terabytes of data) to provide specific decision support responses that are germane to the queries provided.

[0080] The analytics platform **320** of an example embodiment may include at least a statistical discovery component **322** (e.g., SAS analytics and/or JMP, or a component designed using R (i.e., an open-source programming language for statistical computation)) and a natural language

processing component (e.g., NLP engine **324**). The statistical discovery component **322** may be configured to interface with portions of the data sources that include structured data (e.g., some EHR data, some research data, genomic data, etc.) to selectively identify correlations based on analysis of contents of the data sources and the query defined by the user. The NLP engine **324** may be configured to interface with portions of the data sources that include unstructured data (e.g., some EHR data, some research data, clinical data and publications, etc.) to selectively identify correlations based on analysis of contents of the data sources and the query defined by the user.

[0081] The analytics platform **320** may interface with a modeling component **330** configured to apply a selected risk model based on the query. The risk models may be any of a plurality of health models associated with different diseases, health issues or health conditions (e.g., cancer, heart disease, mental health, diabetes, pathogen detection, prescription drug therapy, arthritis, etc.). The modeling component **330** may include a rules engine **332** and/or one or more algorithm implementers (e.g., Bayes Net or components designed using R) **334** that provide risk models to which the analytics platform **320** output can be compared to place correlations and/or patterns identified in the data sources into a meaningful context relative to the query. The rules engine **332** may employ Drools to process rules.

[0082] The modeling component **330** may interface with a user interface component **340**, which may be provided at the application layer, and which may be configured to enable a user to provide the query **350**, and to generate a response **360** to the query **350**. The response **360** may provide information associated with clinical decision support that is tailored to an identity (or role) of the user. Thus, for example, the same system can support access by multiple different types of users (e.g., patients, clinicians and researchers) to provide useful and potentially different levels of access and information extraction from the same massive repository of data to support various applications such as research, clinical trials, drug discovery and patient care. To this end, the infrastructure of the system **200** may further employ a data security and access component **370**. The data security and access component **370** may ensure that any information access restrictions that are appropriate for respective different data sources are enforced.

[0083] Some example embodiments may provide strong capabilities for a closed feedback loop for employment of specific analytics tools to discover correlations within data being analyzed based on the queries provided by the user. FIG. **7** illustrates a block diagram of the mechanisms and platforms associated with practicing example embodiments. As shown in FIG. **7**, various data sources **400** (e.g., genomic data, health record data, clinical research data, PubMed texts and publications, etc.) may be accessed to find correlations and/or patterns at operation **410**. The correlations may be found by machine learning **412** or by human intervention **414**. These correlations may be used to generate rules at operation **420**. Again, the generation may be guided or performed via machine learning **422** or by human intervention **424**. Thereafter, rules engines or algorithm implementors may operate at operation **430** to drive decision support responses based on queries received at operation **440**.

[0084] Example embodiments may be employed for analysis of genomic and/or clinical risk based on the genome data and/or clinical data as a portion of the data sources **400**. In this regard, there is over a terabyte of genomic data that is available for analysis and example embodiments may integrate the genomic data with patient health record data including genomic markers of specific patients to identify, by employing corresponding risk models for specific medical conditions or diseases, a risk score for the patient relative to a likelihood of having the corresponding medical condition or disease for which a query is received. Accordingly, the data sources **400** may be analyzed to identify a selected risk model based on a query and generate a response to indicate a degree of risk of the patient having a condition associated with the selected risk model. The selected risk model may be selected based at least in part on the genetic markers and clinical parameters of the patient and selected portions of the genetic data. The selected portions of the genetic data may be considered to be reference genetic data that is pertinent to the query (e.g., to the condition or disease of interest for a particular patient). As such, individual clinical data and genomic data (e.g., including genetic biomarkers) of the patient can be used along with identification of a specific disease, condition, drug or other query, to identify risks for the patient based at least in part on reference genetic data (and perhaps also reference clinical data) selected from among the massive amounts of patient de-identified data in the data sources **400**.

[0085] Thus, for example, the patient or a clinician may access a record associated with the patient. A query may be provided to request a risk score for a specific type of cancer. The risk score would then be the response to the query. The system **200** may access (among other things) information associated with the patient's genomic markers that are pertinent to risk for the specific type of cancer and the massive amounts of genomic data relating to other patients having and not having the corresponding specific type of cancer. Based on the pertinent information extracted from the analytics platform **320** and application of the modeling component **330**, a risk score may be calculated for the patient based on the correspondence between the genetic profile of the patient and genetic biomarkers associated with the genomic data of others having the cancer. In some cases, the risk score may be a composite risk score that further considers proteomic data, clinical data and/or the like. However, the data platform **300**, the analytics platform **320** and the modeling platform **330** may each be dynamically updateable. Thus, risk scores, models, profiles of various types and various other aspects of the system may be updateable to allow updated processing and decision support to be performed over time. Moreover, additional modules with different types of data sources and corresponding risk models can also be added to the scalable system provided by example embodiments.

[0086] In an example embodiment, the query may include identifying information indicative of a drug prescribed or in consideration for being prescribed for the patient. In such an example, the response to the query may include a risk score relative to the likelihood of one or more complications being experienced by the patient. Alternatively or additionally, the response may include an indication of drug variants and risks relative to a drug of interest (i.e., the drug prescribed) based on a pharmacogenomic profile generated for the patient based on gene variance analysis. Thus, the analytic platform **320** and the modeling component **330** may interact to identify, based on the genetic profile of the patient, a specific drug alternative that may be less likely to cause undesirable side effects for the patient. Alternatively or additionally, the information on drug variants may be directed to providing positive side effects instead of the avoidance of negative side effects. In this regard, the pharmocogenomic profile of the patient,

coupled with genetic data from many other patients with data indicating positive results or benefits of employing a particular drug or treatment regimen may be matched by the system to provide data that can be useful to a clinician in making healthcare decisions for the patient.

[0087] Ultimately, example embodiments enable heterogeneous data from a plurality of sources with different formats to be stored and analyzed from a single scalable system. Analytics, some of which is tailored specifically to the different types/structures of data in the data sources, may then be applied in real time by users that may have distinctly different uses for the information and desired outputs based on a query provided by a particular one of the users. Responsive to the query, the analytics may identify pertinent information and apply rules/models that are applicable to generate a response in the form of a useful visualization for the user. Thus, different types of users can get different types of responses out of the same data set and using the same system. However, the system can tailor the responses to the user by providing visualization tools and techniques that are tailored to the users. Essentially, the system packages information (e.g., genetic information and/or the like) into a form that can make it usable to support clinical decision making and information dissemination. The system can also be useful to process genetic information for different purposes such as finding drug variants or disease variants that are likely to impact a particular patient. Thus, the impact of a drug or disease on a patient may be studied on the basis of the genetic profile of the patient.

[0088] FIG. **8** illustrates a method for providing personalized healthcare for a patient is provided. The method may include receiving information indicative of a patient at operation **500**, retrieving a record associated with the patient at operation **510**, receiving a query identifying a healthcare related issue associated with the patient at operation **520**, performing analytics via a statistical discovery component and a natural language processing component configured to interface with respective portions of heterogeneous data sources to selectively identify correlations between genomic profile information of the patient and selected data of the data sources at operation **530**, applying a selected risk model based on the query at operation **540**, and providing a response to the query including information associated with clinical decision support tailored to an identity of the user, where the user is a selected one of a patient, a researcher and a clinician at operation **550**.

[0089] In an example embodiment, an apparatus for performing the method of FIG. **8** above may comprise a processor or processing circuitry configured to perform some or each of the operations (**500-550**) described above. The processor (e.g., processing unit **204**) may, for example, be configured to perform the operations (**500-550**) by performing hardware implemented logical functions, executing stored instructions, or executing algorithms for performing each of the operations.

[0090] Many modifications and other embodiments of the inventions set forth herein will come to mind to one skilled in the art to which these inventions pertain having the benefit of the teachings presented in the foregoing descriptions and the associated drawings. Therefore, it is to be understood that the inventions are not to be limited to the specific embodiments disclosed and that modifications and other embodiments are intended to be included within the scope of the appended claims. Moreover, although the foregoing descriptions and the associated drawings describe exemplary embodiments in the context of certain exemplary combinations of elements and/or functions, it should be appreciated that different combinations of elements and/or functions may be provided by alternative embodiments without departing from the scope of the appended claims. In this regard, for example, different combinations of elements and/or functions than those explicitly described above are also contemplated as may be set forth in some of the appended claims. In cases where advantages, benefits or solutions to problems are described herein, it should be appreciated that such advantages, benefits and/or solutions may be applicable to some example embodiments, but not necessarily all example embodiments. Thus, any advantages, benefits or solutions described herein should not be thought of as being critical, required or essential to all embodiments or to that which is claimed herein. Although specific terms are employed herein, they are used in a generic and descriptive sense only and not for purposes of limitation.

What is claimed is:

1. A personalized healthcare system, the system comprising:

a data platform scalable to include a plurality of data sources, the data sources including at least a clinical research database, genomic data, and a patient health record database, the patient health record database comprising a record for each of a population of patients, a plurality of genetic markers, and a plurality of clinical parameters associated with the patients;

an analytics platform comprising at least a statistical discovery component and a natural language processing component configured to interface with respective portions of the data sources to selectively identify correlations based on analysis of contents of the data sources responsive to a query;

a modeling component configured to apply a selected risk model based on the query; and

a user interface component configured to enable a user to provide the query, and to generate a response to the query, the response providing information associated with clinical decision support tailored to an identity of the user, the user being a selected one of a patient, a researcher and a clinician.

2. The system of claim **1**, wherein the statistical discovery component interfaces with portions of the data including structured data, and the natural language processing component interfaces with unstructured data.

3. The system of claim **2**, wherein the statistical discovery component interfaces with the genomic data and portions of the clinical research database and the patient health record database that include structured data, and the natural language processing component interfaces with portions of the clinical research database and the patient health record database that include unstructured data such as laboratory data.

4. The system of claim **1**, wherein the modeling component is configured to identify the selected risk model based on the query and generate the response to indicate a degree of risk of the patient having a condition associated with the selected risk model based at least in part on the genetic markers and clinical parameters of the patient and selected portions of the genetic data.

5. The system of claim **1**, wherein the modeling component further comprises a rules engine configured to evaluate the data sources relative to the patient and generate a risk score

based on clinical parameters and a genomic profile of the patient for a selected condition.

6. The system of claim **1**, wherein the modeling component further comprises a rules engine configured to evaluate the data sources relative to the patient and generate a risk score based on clinical parameters and a genomic profile of the patient for each of a plurality of conditions.

7. The system of claim **5**, wherein the risk score comprises a composite risk score for a condition based on the clinical parameters, genomic profile and a proteomic profile.

8. The system of claim **5**, wherein additional patient care results are integrated into the patient health record database to progressively generate an updated risk score.

9. The system of claim **1**, wherein the plurality of clinical parameters associated with the patient include at least one of an age, a weight, a blood pressure, and a temperature of the patient.

10. The system of claim **1**, wherein the user interface comprises a patient dashboard configured to communicate each of a likelihood of a condition, and healthcare treatment options associated with the condition.

11. The system of claim **1**, wherein the response comprises information on drug variants and risks based on a pharmacogenomic profile generated for the patient based on gene variance analysis.

12. The system of claim **10**, wherein the information on drug variants includes an identification of benefits correlated to a selected drug based on the pharmacogenomic profile of the patient.

13. The system of claim **10**, wherein the information on drug variants includes an identification of negative side effects correlated to a drug to be avoided based on the pharmacogenomic profile of the patient.

14. The system of claim **1**, wherein the selected risk model comprises a risk model that is selected based on a drug or disease identified in the query, and based on a genetic profile of the patient.

15. A method for providing personalized healthcare for a patient comprising:

receiving information indicative of a patient;

retrieving a record associated with the patient;

receiving a query identifying a healthcare related issue associated with the patient;

performing analytics via a statistical discovery component and a natural language processing component configured to interface with respective portions of heterogeneous data sources to selectively identify correlations between genomic profile information of the patient and selected data of the data sources;

applying a selected risk model based on the query; and

providing a response to the query including information associated with clinical decision support tailored to an identity of the user, the user being a selected one of a patient, a researcher and a clinician.

16. The method of claim **14**, wherein the statistical discovery component interfaces with genomic data, and the natural language processing component interfaces with unstructured data in a clinical research database and a patient health record database of the heterogeneous data sources.

17. The method of claim **14**, wherein providing the response comprises information on drug variants and risks based on a pharmacogenomic profile generated for the patient based on gene variance analysis.

18. The method of claim **17**, wherein the information on drug variants includes an identification of benefits correlated to a selected drug based on the pharmacogenomic profile of the patient.

19. The method of claim **17**, wherein the information on drug variants includes an identification of negative side effects correlated to a drug to be avoided based on the pharmacogenomic profile of the patient.

20. The method of claim **14**, wherein the selected risk model comprises a risk model that is selected based on a drug or disease identified in the query, and based on a genetic profile of the patient.

* * * * *