



(51) International Patent Classification:

G06F 19/00 (2018.01) G06Q 50/00 (2012.01)
G06F 19/24 (2011.01) G16H 10/60 (2018.01)

(21) International Application Number:

PCT/US2018/056339

(22) International Filing Date:

17 October 2018 (17.10.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/574,116 18 October 2017 (18.10.2017) US
62/626,625 05 February 2018 (05.02.2018) US

(71) Applicant: **MEMORIAL SLOAN KETTERING CANCER CENTER** [US/US]; 1275 York Avenue, New York, New York 10065 (US).

(72) Inventor: **HUERGA, Iker**; c/o Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, New York 10065 (US).

(74) Agent: **KHAN, Shabbi S.** et al.; Foley & Lardner LLP, 3000 K Street NW, Suite 600, Washington, District of Columbia 20007 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME,

(54) Title: PROBABILISTIC MODELING TO MATCH PATIENTS TO CLINICAL TRIALS

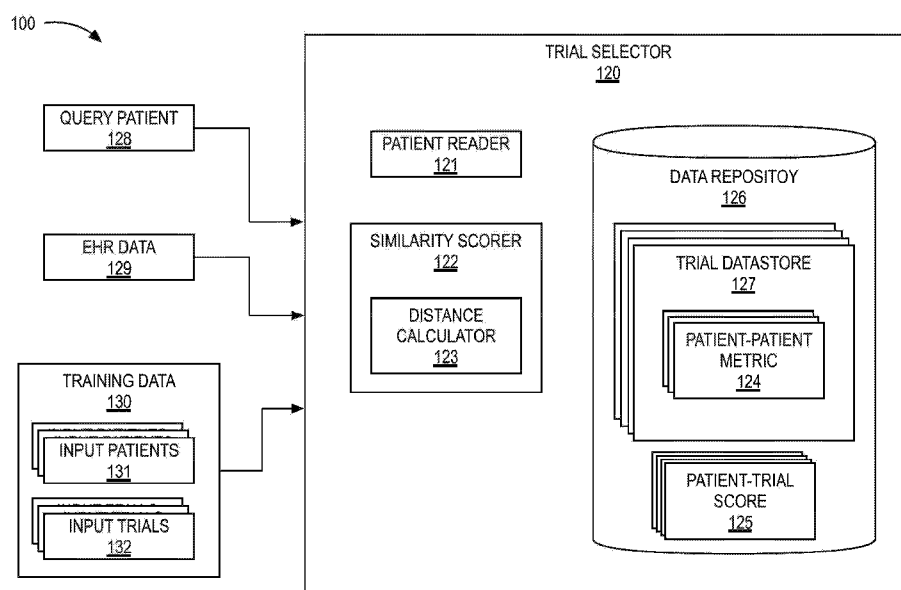


FIG. 1

(57) Abstract: A system to identify clinical trials using probabilistic modeling can include a trial selector. The trial selector can include one or more processors and memory to execute a patient reader and a similarity scorer. The trial selector receives input trials and generates, for each input patient of the respective input trials, a patient vector. The trial selector calculates, for each of the input trials, a patient-patient metric that can indicate a similarity between the respective plurality of input patients. The trial selector can generate, for a query patient, a query patient vector. The trial selector can calculate, for each of the plurality of input trials, a patient-trial similarity score indicating a similarity between the query patient and the respective plurality of input patients. The trial selector can select, from the plurality of input trials, a candidate trial based on the patient-trial similarity score for each of the plurality of input trials.

MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ,
OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA,
SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

PROBABILISTIC MODELING TO MATCH PATIENTS TO CLINICAL TRIALS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims the benefit, under 35 USC § 119(e), of the filing of U.S. Provisional Patent Application 62/574,116, filed October 18, 2017, and U.S. Provisional Patent Application 62/626,625, filed February 5, 2018, both of which are incorporated herein by reference for all purposes.

BACKGROUND

[0002] Clinical trials can advance medical science and provide patients with promising therapy before it is widely available; however, patient accrual is a major challenge. Existing approaches to help clinicians screen patients for clinical trials rely on text analysis and hand-made matching rules or on generic similarity metrics.

SUMMARY

[0003] The present system can formalize clinical trial recommendation as one-shot learning and learn task-specific patient similarity metrics, using them to recommend clinical trials to patients and vice-versa. The system can use a parametrization of input data followed by shallow or deep model architectures that can outperform strong supervised shallow classification models as well as deep metric learning models in this challenging healthcare setting.

[0004] Randomized controlled clinical trials can play an important role in oncology, not only advancing medical science but also providing patients with novel therapies that may offer increased chance of cure or prolonged survival. However, clinical trial participation is very low, with only 3-5% of the adult cancer population enrolling onto a trial. A major challenge clinicians face when finding appropriate clinical trials for their patients is that establishing a match is a largely manual, labor-intensive process. The system can address this challenge via a machine learning-powered clinical trial recommendation system to help clinicians efficiently screen patients for trials at the point of care.

[0005] Whether a patient can be enrolled onto a clinical trial is dictated by the clinical trial eligibility checklist, typically captured in text. There can be more than a hundred different eligibility criteria for a given trial, with some criteria being as granular as dictating specific

genetic mutations a patient must have. In some implementations, systems can apply NLP techniques to create structured rules from the text criteria and then apply these rules to the electronic health record (EHR) to automatically identify eligible patients. However, due to the complex syntactic structure of the eligibility criteria and the unstructured and noisy nature of data in EHRs, their success has been limited.

[0006] Some systems can generate recommendations based on the degree of similarity between the patient of interest, the query patient, and patients already enrolled on clinical trials. However, these approaches have been limited in their ability to use modern machine learning methods due to small datasets of under 20 clinical trials covering different medical fields. The present system can use a significantly larger and more focused dataset. For example, the dataset can include about 1,518 oncology clinical trials and about 39,915 cancer patients.

[0007] Patient-clinical trial recommendation presents a unique combination of methodological and data challenges that make applying widely used machine learning techniques difficult. First, there is numerous clinical trials the system needs to be able to recommend. These trials vary in size significantly and new trials are started every day. To address these challenges, the system can build on the successes of recently proposed methods in supervised metric learning and training objectives specifically designed for one-shot learning. Second, healthcare data is very different from the types of data (e.g., images) frequently considered in recent literature on one-shot learning: it can include many heterogeneous data types, has a significant amount of missing data, individual features can be very informative, and the total amount of training data is often much smaller. The system can include novel architectures that outperform previous state-of-the-art approaches in this setting. The present disclosure describes an end-to-end system that analyzes raw EHR data and produces patient-trial and trial-patient recommendations and generalizes to unseen clinical trials as well as new patients.

[0008] According to at least one aspect of the disclosure, a system to identify clinical trials using probabilistic modeling can include a trial selector that can include one or more processors and memory to execute a patient reader and a similarity scorer. The trial selector can receive a plurality of input trials. Each of the plurality of input trials can have a respective plurality of input patients. The trial selector can generate, for each input patient of the

respective plurality of input patients, a patient vector comprising a plurality of features. The trial selector can calculate, for each of the plurality of input trials, a patient-patient metric indicating a similarity between the respective plurality of input patients. The patient-patient metric for each of the plurality of input trials can be based on the patient vector for each input patient of the respective plurality of input patients. The trial selector can generate, for a query patient, a query patient vector that can include the plurality of features. The trial selector can calculate, for each of the plurality of input trials, a patient-trial similarity score that can indicate a similarity between the query patient and the respective plurality of input patients. The patient-trial similarity score can be based on the query patient vector and the patient-patient metric for each of the plurality of input trials. The trial selector can select, from the plurality of input trials, a candidate trial based on the patient-trial similarity score for each of the plurality of input trials.

[0009] In some implementations, the trial selector can calculate the patient-patient vector using at least one of a Siamese Networks, Matching Siamese Networks, AND-OR Support Vector Machine (SVM) models, Matching AND-OR SVM models, and AND-OR Attention models. The plurality of features can include at least two of a patient diagnosis code, a procedure code, demographic information, laboratory values, medications, genetic mutations, and textual assessment notes.

[0010] The trial selector can receive a second plurality of input trials. The second plurality of input trials do not include enrolled patients. The trial selector can generate a trial feature vector for each of the second plurality of input trials. The trial feature vector can be based on metadata of each of the second plurality of input trials. The trial selector can calculate a second patient-trial similarity indicating a similarity between the query patient and each of the second plurality of input trials. The trial feature vector can be one-hot encoded.

[0011] The trial selector can extract metadata from a patient record associated with the query patient. The trial selector can generate, for the query patient, the query patient vector based on the metadata. The trial selector can identify, for each of the plurality of input trials, one or more eligibility requirements associated with each of the respective one of the plurality of input trials. The trial selector can determine, for the query patient, whether the query patient has the one or more eligibility requirements associated with each of the respective one of the plurality of input trials.

[0012] According to at least one aspect of the disclosure, a method to identify clinical trials using probabilistic modeling can include receiving, by a trial selector, a plurality of input trials, each of the plurality of input trials having a respective plurality of input patients. The method can include generating, by the trial selector, for each input patient of the respective plurality of input patients, a patient vector that can include a plurality of features. The method can include calculating, by the trial selector, for each of the plurality of input trials, a patient-patient metric indicating a similarity between the respective plurality of input patients. The patient-patient metric for each of the plurality of input trials is based on the patient vector for each input patient of the respective plurality of input patients. The method can include generating, by the trial selector, for a query patient, a query patient vector comprising the plurality of features. The method can include calculating, by the trial selector, for each of the plurality of input trials, a patient-trial similarity score indicating a similarity between the query patient and the respective plurality of input patients. The patient-trial similarity score can be based on the query patient vector and the patient-patient metric for each of the plurality of input trials. The method can include selecting, by the trial selector, from the plurality of input trials, a candidate trial based on the patient-trial similarity score for each of the plurality of input trials.

[0013] In some implementations, the method can include calculating, for each of the plurality of input trials, the patient-patient metric further comprising determining the patient-patient metric with at least one of a Siamese Networks, Matching Siamese Networks, AND-OR Support Vector Machine (SVM) models, Matching AND-OR SVM models, and AND-OR Attention models. The plurality of features can include at least two of a patient diagnosis code, a procedure code, demographic information, laboratory values, medications, genetic mutations, and textual assessment notes.

[0014] In some implementations, the method can include receiving a second plurality of input trials. The second plurality of input trials do not include enrolled patients. The method can include generating a trial feature vector for each of the second plurality of input trials. The trial feature vector based on metadata of each of the second plurality of input trials.

[0015] The method can include calculating a second patient-trial similarity that can indicate a similarity between the query patient and each of the second plurality of input trials. The trial feature vector can be one-hot encoded.

[0016] In some implementations, the method can include extracting metadata from a patient record associated with the query patient. The method can include generating the query patient vector based on the metadata.

[0017] In some implementations, the method can include identifying, for each of the plurality of input trials, one or more eligibility requirements associated with each of the respective one of the plurality of input trials. The method can include determining, for the query patient, whether the query patient has the one or more eligibility requirements associated with each of the respective one of the plurality of input trials.

[0018] The foregoing general description and following description of the drawings and detailed description are exemplary and explanatory and are intended to provide further explanation of the invention as claimed. Other objects, advantages, and novel features will be readily apparent to those skilled in the art from the following brief description of the drawings and detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] The accompanying drawings are not intended to be drawn to scale. Like reference numbers and designations in the various drawings indicate like elements. For purposes of clarity, not every component may be labeled in every drawing. In the drawings:

[0020] FIG. 1 illustrates a block diagram of an example trial selector.

[0021] FIGS. 2A–2C illustrate different model architectures for use in the trial selector illustrated in FIG. 1.

[0022] FIG. 3 illustrates an example method to classify a patient into a clinical trial using the system illustrated in FIG. 1.

[0023] FIG. 4A illustrates a plot of the trial area under the curve (AUC) as a function of the number of positive examples.

[0024] FIG. 4B illustrates a plot of Patient Quantile Index (PQI) as a function of the number of positive examples.

DETAILED DESCRIPTION

[0025] The various concepts introduced above and discussed in greater detail below may be implemented in any of numerous ways, as the described concepts are not limited to any particular manner of implementation. Examples of specific implementations and applications are provided primarily for illustrative purposes.

[0026] As will be appreciated by one skilled in the art, aspects of the present disclosure may be embodied as a system, method or computer program product. Accordingly, aspects of the present disclosure may take the form of an entirely hardware embodiment, an entirely software embodiment (including firmware, resident software, micro-code, etc.) or an embodiment combining software and hardware aspects that may all generally be referred to herein as a "circuit," "engine," "module" or "system." Furthermore, aspects of the present disclosure may take the form of a computer program product embodied in one or more computer readable medium(s) having computer readable program code embodied thereon. Aspects of the present disclosure may be implemented using one or more analog and/or digital electrical or electronic components, and may include a microprocessor, a microcontroller, an application-specific integrated circuit (ASIC), a field programmable gate array (FPGA), programmable logic and/or other analog and/or digital circuit elements configured to perform various input/output, control, analysis and other functions described herein, such as by executing instructions of a computer program product.

[0027] Any combination of one or more computer readable medium(s) may be utilized. The computer readable medium may be a computer readable signal medium or a computer readable storage medium. A computer readable storage medium may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples (a non-exhaustive list) of the computer readable storage medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing. In the context of this document, a computer readable storage medium may be any tangible medium that can contain or store a

program for use by or in connection with an instruction execution system, apparatus, or device.

[0028] A computer readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband or as part of a carrier wave. Such a propagated signal may take any of a variety of forms, including, but not limited to, electro-magnetic, optical, or any suitable combination thereof. A computer readable signal medium may be any computer readable medium that is not a computer readable storage medium and that can communicate, propagate, or transport a program for use by or in connection with an instruction execution system, apparatus, or device.

[0029] Program code embodied on a computer readable medium may be transmitted using any appropriate medium, including but not limited to wireless, wireline, optical fiber cable, RF, etc., or any suitable combination of the foregoing. Computer program code for carrying out operations for aspects of the present disclosure may be written in any combination of one or more programming languages, including an object oriented programming language such as Java, Smalltalk, C++ or the like and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The program code may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider).

[0030] Aspects of the present disclosure are described below with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems) and computer program products according to embodiments of the disclosure. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer program instructions. These computer program instructions may be provided to a processor of a general-purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the

computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0031] These computer program instructions may also be stored in a computer readable medium that can direct a computer, other programmable data processing apparatus, or other devices to function in a particular manner, such that the instructions stored in the computer readable medium produce an article of manufacture including instructions which implement the function/act specified in the flowchart and/or block diagram block or blocks. The computer program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other devices to cause a series of operational steps to be performed on the computer, other programmable apparatus or other devices to produce a computer implemented process such that the instructions which execute on the computer or other programmable apparatus provide processes for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0032] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of code, which comprises one or more executable instructions for implementing the specified logical function(s). It should also be noted that, in some alternative implementations, the functions noted in the blocks may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts, or combinations of special purpose hardware and computer instructions.

[0033] Reference in the specification to “one embodiment” or “an embodiment” of the present principles, as well as other variations thereof, means that a particular feature, structure, characteristic, and so forth described in connection with the embodiment is included in at least one embodiment of the present principles. Thus, the appearances of the phrase “in one embodiment” or “in an embodiment”, as well as any other variations, appearing

in various places throughout the specification are not necessarily all referring to the same embodiment.

[0034] The present solution, explores how metric learning can be applied to build a patient-trial recommendation system. The system can use a matching training procedure to improve performance over a simple presentation of just positive and negative pairs. The present solution can provide high-performance results even when there is large variability in the underlying data as well as the number of target classes.

[0035] According to an aspect of the disclosure, a system to identify clinical trials using probabilistic modeling can include a trial selector. The trial selector can include one or more processors and memory to execute a patient reader and a similarity scorer. The trial selector can receive a plurality of input trials. Each of the plurality of input trials can include a respective plurality of input patients that are enrolled in the respective input trial. The trial selector can generate, for each input patient of the respective plurality of input patients, a patient vector that can include a plurality of features. The trial selector can calculate, for each of the plurality of input trials, a patient-patient metric that can indicate a similarity between the respective plurality of input patients. The patient-patient metric for each of the plurality of input trials can be based on the patient vector for each input patient of the respective plurality of input patients. The trial selector can generate, for a query patient, a query patient vector comprising the plurality of features. The trial selector can calculate, for each of the plurality of input trials, a patient-trial similarity score indicating a similarity between the query patient and the respective plurality of input patients. The patient-trial similarity score can be based on the query patient vector and the patient-patient metric for each of the plurality of input trials. The trial selector can select, from the plurality of input trials, a candidate trial based on the patient-trial similarity score for each of the plurality of input trials.

[0036] The present solution can improve the performance of a related field and technology. The present solution can improve the identification and selection of clinical trials in which a patient can be enrolled. The selection of clinical trials for a patient was previously only performed by a trained clinician. Finding matching clinical trials for a patient was a manual, labor-intensive process. The present solution enables the automated selection of clinical trials for a patient. The system can parameterize input data about the patient and about also candidate clinical trials. The system can use machine learning to select clinical trials in which

the patient can be enrolled based on the features of the patient and the features of the clinical trials.

[0037] FIG. 1 illustrates a block diagram of a system 100 to match patients to one or more trials. The system 100 can include an example trial selector 120. The trial selector 120 can select or otherwise identify one or more trials (e.g., clinical trials) for which a query patient 128 is eligible for enrollment. The trial selector 120 can include a patient reader 121 and a similarity scorer 122. The similarity scorer 122 can include a distance calculator 123. The trial selector 120 can include a data repository 126 that can store one or more trial data stores 127. For each of the trial datastores 127, the data repository 126 can include a plurality of patient-patient metrics 124. The data repository 126 can include a plurality of patient-trial scores 125 that correspond to the query patient 128. In some implementations, the patient-patient metrics 124 and/or the patient-trial similarity score 125 can be generated on demand and not stored in the data repository 126. The trial selector 120 can select a clinical trial or generate a ranked list of clinical trials for a given query patient 128. The trial selector 120 can make the selection based on input EHR data 129 and training data 130. The training data 130 can include input patients 131 and input trials 132.

[0038] The trial selector 120 can include multiple, logically-grouped servers and facilitate distributed computing techniques. The logical group of servers may be part of a data center, server farm, or a machine farm. The servers can be geographically dispersed. A data center or machine farm may be administered as a single entity, or the machine farm can include a plurality of machine farms. The servers within each machine farm can be heterogeneous – one or more of the servers or machines can operate according to one or more type of operating system platform. The trial selector 120 can include servers in a data center that are stored in one or more high-density rack systems, along with associated storage systems, located for example in an enterprise data center. The trial selector 120 with consolidated servers in this way can improve system manageability, data security, the physical security of the system, and system performance by locating servers and high-performance storage systems on localized high-performance networks. Centralization of all or some of the trial selector 120 components, including servers and storage systems, and coupling them with advanced system management tools allows more efficient use of server resources, which saves power and processing requirements and reduces bandwidth usage.

[0039] The trial selector 120 can, for a query patient 128, recommend one or more clinical trials for the enrollment of the query patient 128. The clinical trials to which the trial selector 120 can match the query patient 128 can include the input trials 132. The trial selector 120 can match query patients 128 to trials based on a similarity between the query patient 128 and the input patients 131 associated with each of the input trials 132. For example, the trial selector 120 can determine which trial's enrolled patients are the most similar to the query patient 128. The trial selector 120 can match the query patient 128 to a clinical trial based on the trial's enrolled patients rather than matching information about the query patient 128 (e.g., text containing keywords, diagnosis, test results, etc.) to text found in the trial's eligibility criteria. The trial selector 120 can classify a query patient 128 as eligible for a specific trial using the input patients 131 as positive examples and patients who were manually screened and selected for the trial. The input patients 131 that were deemed ineligible for a specific trial can serve as negative examples. The trial selector 120 can use one-shot learning to determine whether a query patient 128 should be eligible for a specific trial.

[0040] The patient reader 121 can be any script, file, program, application, set of instructions, or computer-executable code, that is configured to enable a computing device on which the patient reader 121 is executed to generate patient vectors. The patient reader 121 can generate respective patient vectors for each of the query patient 128 and the input patients 131.

[0041] The patient reader 121 can generate a patient vector by identifying a patient identifier in the query patient 128 or input patient 131. The patient identifier can be a unique patient identifier (e.g., a string or integer value) used by a hospital or other medical facility to uniquely identify a patient. Using the patient identifier, the patient reader 121 can select one or more features from the electronic health record (EHR) data 129. The features from the EHR data 129 can include patient diagnosis and procedure codes, demographic information, laboratory values, medications, genetic mutations, and textual assessment notes. The patient reader 121 can convert the features into the patient vectors. For example, the patient reader 121 can generate a dictionary for each of the respective features, where the results for a respective feature are uniquely encoded based on order of appearance. The patient reader 121 can convert the integer-based character vectors into binary-based character vectors by one-hot encoding the character vectors.

[0042] The patient reader 121 can generate a patient vector for each of the input patients 131 to which the similarity scorer 122 compares the query patient 128. The input patients 131 can be included in the training data 130. The training data 130 can include input trials 132 that can indicate a plurality of clinical trials. The input trials 132 can also be the clinical trials to which the query patient 128 is matched (or ranked). The training data 130 can include input patients 131. Each of the input patients 131 can correspond to one or more of the input trials 132. An input patient 131 can correspond to an input trial 132 when the patient identified by the input patient 131 is enrolled in the input trial 132. In some implementations, the training data 130 can be a database of clinical trials where each row, for example, indicates an input trial 132 and each column indicates an input patient 131 enrolled in the respective input trial 132.

[0043] The patient reader 121 can store each of the patient vectors for a given input trial 132 into a respective trial data store 127. Each trial data store 127 can store a one-hot encoded patient vector for each of the input patient 131 enrolled in the respective clinical trial. The patient reader 121 can update the trial data stores 127 at predetermined intervals. The patient reader 121 can update the trial data stores 127 between about 1 day and about 4 days, between about 1 day and about 7 days, between about 1 day and about 10 days, or between about 1 days and about 14 days. For example, the patient reader 121 can update the trial data stores 127 to add patients recently added to the clinical trial or to remove patients that were removed from the clinical trial.

[0044] The similarity scorer 122 can be any script, file, program, application, set of instructions, or computer-executable code, that is configured to enable a computing device on which the similarity scorer 122 is executed to match or recommend a query patient 128 to a clinical trial (e.g., one of the input trials 132). The similarity scorer 122 can break the classification task into two components. The similarity scorer 122 can learn or otherwise calculate a problem-specific patient-patient metric 124, which can indicate the similarity between the input patients 131 enrolled in a given input trial 132. The similarity scorer 122 can use the patient-patient metric 124 to calculate a patient-trial similarity score 125 indicating a score of the query patient's match to a given input trial 132. The similarity scorer 122 can calculate the patient-trial similarity score 125 for a given query patient 128 and input trial 132.

[0045] Once the patient reader 121 processes the EHR data 129 to generate patient vectors for each of the input patients 131 in the input trials 132, the similarity scorer 122 can match the query patient 128 to an input trial 132. The similarity scorer 122 can calculate similarities between the query patient 128 and the patients already enrolled in the respective clinical trials (e.g., each of the input patients 131 in the respective input trials 132). The input patients 131 enrolled in the same input trial 132 can share certain characteristics. For example, the input patients 131 enrolled in the same input trial 132 can have features in the EHR data 129 that are similar or related. The input patients 131 that are enrolled on different trials that were available at the time of their enrollment may not share these characteristics (e.g., EHR data 129 features).

[0046] The similarity scorer 122 can include the distance calculator 123 that calculates the distances between patients. For each input trial 132, the distance calculator 123 can calculate the distances between each of the pairs of input patients 131 in the respective input trial 132.

[0047] The distance calculator 123 can include different types of model architectures to calculate the patient-patient metric 124 and the patient-trial similarity score 125. The model architectures can include Siamese Networks, Matching Siamese Networks, AND-OR Support Vector Machine (SVM) models, Matching AND-OR SVM models, or AND-OR Attention models.

[0048] The distance calculator 123 can use the model architectures to combine the raw input of two patients (e.g., the patient vectors generated from the EHR data 129) and produce a similarity metric $sim(p_i, p_j)$ (which can also be referred to as a patient-patient similarity score) that can indicate the distance between the two patients p_i and p_j . Each of the model architectures can learn a weighted distance function of their inputs via cross-entropy or hinge loss, that is then used as the patient-to-patient similarity metric 124 (e.g., $sim(p_i, p_j)$). The similarity scorer 122 can use two training procedures, either presenting the models with positive and negative pairs or using support sets. FIGS. 2A–2C illustrate block diagrams of the example architectures that can be used by the distance calculator 123.

MODEL ARCHITECTURES

[0049] The distance calculator 123 can use a Siamese Network. FIG. 2A illustrates an example Siamese Network model. The similarity scorer 122 can train the Siamese network

such that a pair of inputs are mapped into a hidden feature space by two neural networks with tied weights. In some implementations, rather than using contrastive loss, the distance calculator 123 can follow and combine the output of the two Siamese networks via a weighted L1 distance: $d(s_1, s_2) = \sum w_i |s_{1i} - s_{2i}| + b$.

[0050] The distance calculator 123 can use sigmoid activation to restrict the values between 0 and 1. The distance calculator 123 can use cross-entropy loss to discriminate between similar and dissimilar pairs. The weighted distance is used as the patient-patient similarity metric.

[0051] For training the Siamese Network, the similarity scorer 122 can construct positive pairs from patients that are enrolled on the same trial. For example, each patient of a positive pair can be enrolled in the same input trial 132. The similarity scorer 122 can construct negative pairs from patients that are enrolled in different trials. For example, the first patient for an example pair can be from the input trial 132(1) and the second patient for the example pair can be from the input trial 132(2). The different trials from which the similarity scorer 122 can select the negative pairs (e.g., input trial 132(1) and input trial 132(2) in the above example) can be available at the same time. For example, each of the input trials 132 can be enrolling or accepting patients at the same time. This process can simulate the decision-making process whereby a clinician considered a patient for several trials but chose a first trial (e.g., input trial 132(1)) as the most appropriate. In some implementations, the input trials 132 can vary in the number of enrolled patients. To address this, the similarity scorer 122 can oversample small input trials 132 (e.g., the input trials 132 including a relatively small number of input patients 131) and under sample large input trials 132 (e.g., the input trials 132 including a relatively large number of input patients 131). In some implementations, the similarity scorer 122 can keep the positive and negative pair ratio substantially equal.

[0052] The distance calculator 123 can use a matching Siamese Network. In some implementations, the matching Siamese Network can employ a different training procedure. Instead of using pairs of similar and dissimilar items, the matching Siamese Network can use a support set of k examples of item-label pairs $S = \{(x_i, y_i)\}_{i=1}^k$ that is mapped to a classifier that learns $P(y|x, S)$, a probability distribution over outputs y given a test example x . The matching Siamese Network model is specified as $\hat{y} = \sum_{i=1}^k a(\hat{x}, x_i) y_i$ where x_i are

examples, and y_i their associated one-hot encoded labels, from the support set. In the model, a can be an attention kernel that can take on many forms. For example, the attention kernel can be in the general form that uses a softmax over a distance function d :

$$a(\hat{x}, x_i) = \frac{e^{d(f(\hat{x}), g(x_i))}}{\sum_{j=1}^k e^{d(f(\hat{x}), g(x_j))}}$$

[0053] The similarity scorer 122 can use the Siamese network architecture from above to embed $\hat{x}$ and x_i (e.g., setting $f = g$) and use the learned distance as d . Then, mimicking the evaluation procedure, the similarity scorer 122 can construct training examples by taking each query patient 128 in the training set and randomly sampling up to n additional input patients 131 from their respective input trial 132 clinical trial as positive examples. The similarity scorer 122 can select up to n input patients 131 from each of the trials that the query patient 128 could have potentially been enrolled onto. For example, the potential trials can include the input trials 132 that were available at the time of the query patient's potential enrollment. Within each trial in the support set, the similarity scorer 122 can average the distances $d(\hat{x}, x_i)$, and calculate trial probabilities using a softmax on the averaged distances. The predicted class can be the argmax over that distribution.

[0054] The distance calculator 123 can use an AND-OR SVM model. FIG. 2B illustrates an example AND-OR SVM model. Siamese networks can indirectly try to learn the best representation of the input to minimize their loss. The distance calculator 123 can determine that, in the setting of noisy high-dimensional data, choosing an appropriate combination function of a pair of patient representations with tunable weights might improve performance. Using the AND-OR SVM model, the distance calculator 123 can take the original one-hot patient vectors in a given positive/negative pair and construct a new feature vector by applying an element-wise AND operation to them. The element-wise AND operation can result in only common entries being non-zero. The AND-OR SVM model can also include applying an OR operation, which can result in dissimilar elements being nonzero. The AND-OR model can concatenate the AND-OR vectors to double the original feature dimension.

[0055] The AND operation can enable the similarity scorer 122 to find regions of overlap between two input patients 131, such as common diagnosis codes for colon cancer. The AND operation may not be informative about features that are missing or different between the two

input patients 131, such as comorbidities like diabetes and heart failure which can be exclusion criteria for clinical trials. The non-overlapping but non-zero features can be found with the OR combination. For example, lab test for blood glucose can have as many as 5 different codes in EHR data 129. In this example, even if two input patients 131 have different blood glucose tests that are elevated, the downstream SVM can learn to classify the patients as being similar.

[0056] The similarity scorer 122 can train the Support Vector Machine with a linear kernel and L2 regularization using the AND-OR feature vector as an input. Since the similarity scorer 122 can use a linear kernel, the similarity scorer 122 can have that $d(x) = x \cdot W + b$, where x is the AND-OR feature vector. The system output of the decision function of the SVM as a similarity score.

[0057] In some implementations, the distance calculator 123 can use a matching AND-OR SVM model. Since the loss and training procedure can be general, the trial selector 120 can use the weighted AND-OR feature combination as the distance function d in the Matching Network training procedure. For example, the trial selector 120 can use support sets instead of pairs and cross-entropy loss instead of hinge-loss.

[0058] In some implementations, the similarity scorer 122 can use an AND-OR Attention model. FIG. 2C illustrates an example AND-OR Attention model. The AND-OR Attention model can detect potential nonlinear relationships between different variables and contextual effects. The trial selector 120 can take inspiration for the gating mechanism employed in Highway networks, which regulate how much of the input to transform and how much to pass through from layer to layer. A highway network exchanges the traditional layer of a neural network that has the form $x' = \sigma(x, W)$ (σ a nonlinear transformation) with $x' = \sigma_H(x, W_H) \cdot \sigma_T(x, W_T) + x \cdot (1 - (\sigma_T(x, W_T)))$ where $\sigma_T(x, W_T)$ learns how much of the original input to transform and $(1 - \sigma_T(x, W_T))$ - how much to let through unchanged.

[0059] In some implementations, x is the input AND-OR feature vector, and σ_H and σ_T are the sigmoid transformation. The system can use a single attention layer followed by a learned weighted combination of x' as the distance metric, i.e. $d(x') = x' \cdot W_D + b_D$ and minimize cross-entropy loss. The system can apply L2 regularization to W_H , W_T and W_D as in the AND-OR SVM model.

[0060] FIG. 3 illustrates an example method 300 to classify a patient into a clinical trial. The method 300 can include receiving training data (BLOCK 301). The method 300 can include training the trial selector (BLOCK 302). The method 300 can include receiving a query patient (BLOCK 303). The method 300 can include calculating patient-trial similarity metrics (BLOCK 304). The method 300 can include ranking the trials based on the patient-trial similarity metrics (BLOCK 305).

[0061] As set forth above, the method 300 can include receiving training data (BLOCK 301). Also, referring to FIG. 1, among others, the training data 130 can include a plurality of input trials 132. Each of the input trials 132 can include one or more input patients 131 that are enrolled in the respective input trial 132. The training data 130 can be associated with EHR data 129. The EHR data 129 can health record information for each of the respective input patients 131 from which the trial selector 120 can generate feature vectors. The EHR data 129 can include patient diagnosis and procedure codes, demographic information, laboratory values, medications, genetic mutations, and textual assessment notes. The trial selector 120 can generate a patient vector for each of the input patients 131. The patient vectors can be one-hot encoded to convert the features from the EHR data 129 into a binary encoded vector.

[0062] In some implementations, the training data or input data can include a plurality of input trials that do not yet include enrolled patients. For example, the input trials can be new trials that are seeking new patients. The trial selector can also receive data or metadata associate with the plurality of input trials. The trial selector can retrieve the data or metadata from a trial website, such as clinicaltrials.gov. The trial selector can process the information about the trial requirements with a natural language processor to extract the metadata for the trial. The metadata can include requirements for enrolling in the trial or features of the trial. The trial selector can generate a trial feature vector that includes the features, requirements, or metadata associated with the trial. The trial feature vector can be one-hot encoded.

[0063] The method 300 can include training the trial selector (BLOCK 302). For each input trial 132, the trial selector 120 can calculate a distance between pairs of input patients 131 within each of the given input trials 132. The distance between the pairs of input patients 131 can provide a similarity metric for the input patients 131 within each of the respective input trials 132. For example, the models can learn a weighted distance function of their inputs via

cross-entropy or hinge loss, that is then used as the patient-to-patient similarity function $\text{sim}(p_i, p_j)$ (e.g., the patient-patient similarity metric). A similarity function $\text{Sim}(T, p)$ can be used to score patients and trials. The trial-patient score $\text{Sim}(T_j, p)$ (e.g., the trial-patient similarity score) for patient and each of the input trials 132 T_j can be calculated by averaging the top- k highest similarities $\text{sim}(p, p_{ij})$ for patients p_{ij} on trial T_j . The value of k is chosen via cross-validation.

[0064] The method 300 can include receiving a query patient (BLOCK 303). Responsive to receiving the query patient 128, the trial selector 120 can generate a patient vector for the query patient 128. The query patient's patient vector can be generated in a method similar to the method for generating the patient vectors for the input patients 131. For example, the trial selector 120 can receive the query patient 128 as a patient identifier (e.g., a unique string or integer representation of the query patient 128). Using the patient identifier, the trial selector 120 can reference the EHR data 129 to generate a patient vector for the query patient 128. The trial selector 120 can one-hot encode the query patient's patient vector.

[0065] The method 300 can include calculating patient-trial similarity metrics (BLOCK 304). The trial selector 120 can use the learned metric to rank patient-trial combinations in a k -nearest neighbors-like framework. More formally, the similarity function $\text{Sim}(T, p)$ is used to score patients and trials and provide the patient-trial similarity score. The trial-patient score $\text{Sim}(T_j, p_q)$ for a query patient 128 p_q and each of the input trials 132 T_j can be calculated by averaging the top- k highest similarities $\text{sim}(p_q, p_{ij})$ for patients p_{ij} on trial T_j . The value of k is chosen via cross-validation.

[0066] In some implementations, the trial selector 120 can generate a patient-trial similarity metric between the query patient (e.g., the query patient's patient vector) and the trial feature vector. The patient-trial similarity metric can indicate a similarity to the query patient with a clinical trial or the metadata, features, or requirements thereof for a clinical trial that has few or no patients enrolled.

[0067] The method 300 can include ranking the trials based on the patient-trial similarity metrics (BLOCK 305). In some implementations, the method 300 can return a recommended trial from among the input trials 132 into which the query patient 128 can be enrolled. The recommended trial can be the input trial 132 with the highest patient-trial similarity score. In

some implementations, the trial selector 120 can return a ranked list of the input trial 132 into which the query patient 128 can be enrolled. In some implementations, the trial selector 120 can select one or more of the returned input trials for enrollment by the query patient. In some implementations, the trial selector 120 can select the highest ranked of the input trials.

[0068] In some implementations, the trial selector 120 can determine whether the query patient meets one or more eligibility requirements of the ranked trials. The trial selector 120 can determine whether the query patient meets the one or more eligibility requirements once the input trials are ranked or when the input trials are received. The eligibility requirements can be requirements for the patient to be enrolled in the trial. For example, the eligibility requirements can be the presence of a predetermined condition, presentation of the condition, age range, sex, or race. For example, the trial selector 120 can determine that the query patient is similar to patients enrolled in a given trial and may rank the given trial highly. However, the trial selector 120 can determine that the metadata for the given trial indicates that the age requirement for the given trial is between 18 and 45 years of age. The trial selector 120 can determine that the query patient is not within the age range, and the trial selector 120 can discard the given trial from the list of ranked trials.

[0069] Once one or more trials are selected for the query patient, the results of the selected trials can be provided back to the trial selector 120 and trial selector 120 can update its machine learning models. For example, an indication that a user of the system selected the first, third, and fourth trials for enrollment by the query patient can be provided back to the trial selector 120 and the trial selector 120 can update its machine learning models based on the feedback information.

EXPERIMENTS

DATA

[0070] The system was evaluated using data from a leading cancer center in the United States. This cancer center conducts over 500 therapeutic clinical trials at a time; most are focused on a single type of cancer (e.g. breast, lung). The test used electronic health record data from 39,915 patients enrolled on 1,518 therapeutic trials between 2004 and 2016. Missing data is common in the real-world clinical setting, and only demographic information is consistently available for all patients. The tests consider all patients that have at least one additional piece of information to demographics. The electronic health record captures

information about patients and their interaction with health providers. The system derived features from patient diagnosis and procedure codes, demographic information, laboratory values, medications, genetic mutations, and textual assessment notes. Data is longitudinal in nature; missingness and redundancies are commonly present, and dataset sizes are modest, which complicate the application of typical machine learning algorithms.

[0071] The system used a one-hot encoding of the raw data from the EHR. The system used binary features for genetic and laboratory tests by considering the cross-product of test and result (e.g. EGFR_positive, ERBB2_negative, Creatinine_abnormal, Hemoglobin_normal). Using separate features for positive and negative results captures the difference between a positive, negative and a missing result. Diagnosis and procedure billing codes are grouped from granular billing-based categories into clinically meaningful ones (reducing 84,904 diagnosis codes to high-level 283 groups and 54,019 procedure codes to 244 groups). Additional genetic mutations, functional status, and stage of cancer are extracted from the clinical notes using a simple named-entity-extraction algorithm. Drugs are represented via medication names (discarding dosage), and demographics information is left unchanged.

[0072] The decision whether a patient is eligible for a clinical trial is made using the latest available information as well as past medical history. To mirror this decision-making process, the system can group each data element by type and take the latest entry. For example, the patient's current cancer stage and lab tests are used rather than those documented 3 months ago. However, if they had a single diagnosis code for heart attack from 3 years ago, the system can retain that as well. The system can binarize the resulting "snapshot" representation and one-hot encode it.

[0073] Finally, the system one-hot encoded department names of the patient's treating clinician and concatenate it to the snapshot representation, producing a sparse binary vector with 7,768 features per patient.

DETAILS OF LEARNING ALGORITHMS

[0074] The system created a dataset as follows: clinical trials (and the patients enrolled on them) are stratified by cancer type and number of enrolled patients, randomly shuffled and split into train, validation and test set at 70/15/15 proportion. The system used TensorFlow to train all models. The system can also use scikit-learn.

[0075] When using a distance calculator 123 with a Siamese Network architecture, the system can use multilayer perceptrons (MLP) as embedding functions, batch normalization, dropout, and early stopping (based on the validation set) to improve performance and regularize the networks. The final model was trained for 25 epochs using the Adam optimizer with a learning rate of 0.001, batch size of 128, MLPs with a single hidden layer of size 400, and dropout of 0.1. During training, the system can set the number of generated pairs to be 22,656 per epoch, equal to the number of unique patients in the training set.

[0076] When using a distance calculator 123 with a Matching Siamese Network architecture, the system can keep the network configuration and training procedure identical to the Siamese network setup. The final model uses support size of 5 and is trained for 14 epochs.

[0077] When using a distance calculator 123 with a AND-OR SVM architecture, the system can use an L2 regularized linear SVM with 0.0001 as a regularization constant and train it for 11 epochs using the Adam optimizer with a learning rate of 0.001 and a batch size of 128.

[0078] When using a distance calculator 123 with a Matching AND-OR architecture, the system can use a support size of 5 and was trained for 3 epochs.

[0079] When using a distance calculator 123 with a AND-OR Attention architecture, the system is trained using the Adam optimizer with learning a rate of 0.001, batch size of 128, L2 penalty of 0.0005 on W_H and W_T and 0.00001 on W_D for 15 epochs.

BASELINES

[0080] For experiments matching patients to breast cancer trials, the baseline was manually constructed with a rules-based system for matching patients to breast cancer trials developed prior to pursuing the machine learning approach described herein. The system extracted demographic, cancer type, stage, genetic mutations, and lab requirements from text in trial eligibility criteria and attempted to match them to corresponding structured and NLP-extracted patient attributes using a manually tuned combination function. The system can use text extraction modules. The text extraction modules can include supervised learning on

manually annotated data. This approach is labor-intensive as individual NLP models need to be developed for important patient characteristics that vary by cancer type.

[0081] For the experiments, a regularized linear SVM was trained for each input trial. For each input trial, the system can consider enrolled patients as positive examples. The patients enrolled in other trials as negative examples (where the other trials are within the same time frame and cancer type). The system can then train a regularized linear SVM classifier for each trial, adjusting for imbalanced classes. The system can apply Platt scaling to the output of each SVM classifier to obtain probability estimates. The trial ensemble classifier (TEC) approach has to be trained using the same clinical trials as will be present at test time. The system can fit individual trial classifiers and use leave-one-out cross-validation to evaluate their performance.

EVALUATION METRICS

[0082] In order to make fair comparisons across different systems, the system can include a general framework for generating recommendations that is then used to evaluate every model. The system can use two metrics to evaluate models — one patient-specific and one trial-focused. Patient Quantile Index (PQI): all relevant clinical trials are ranked for each patient using the learned metric. The PQI is the percentile rank at which the patient's trial appears in the ranked list. The percentile ranks are averaged across all patients for a final score. Trial AUC (TAUC): for each trial, of the recommended patients, the system can use enrolled patients as positives and others as negatives. These patients are jointly ranked, and a trial AUC is calculated and averaged across all trials for a final score.

[0083] To reflect the real-world setting, patients are only compared to trials of the same cancer type and to trials that were available at the time of enrollment to their own trial. Additionally, to keep evaluations meaningful clinical trials with less than 2 enrolled patients and patients who cannot be evaluated for more than 4 trials are not considered. After applying these restrictions, the test set contains 144 unique clinical trials and 2,305 unique patients. After filtering, the PQI for each patient is calculated over 7 trials on average. The TAUC metric on average is over 10 patients on the trial and 70 off trial.

[0084] Since the breast cancer matching system is rule based and cannot generalize to other cancer types, the system can conduct a separate evaluation between it and the AND-OR SVM model on a subset of 38 breast-cancer focused clinical trials and 626 patients in the test set.

RESULTS

[0085] The AND-OR SVM outperforms the rules-based algorithm on the breast cancer subset (PQI: 0.86 vs. 0.70; TAUC: 0.83 vs. 0.70). This is an expected result since the rules-based algorithm is completely reliant on the quality of the text extraction models powering it, and has very limited ability to handle missing data.

[0086] The remaining models are compared on the full test set. Results are summarized in Table 1. Overall, models based on the AND-OR parameterization of the inputs perform much better than those using raw inputs. This supports the hypothesis that a problem-appropriate combiner function facilitates learning. The Matching training objective directly optimizes model performance for patient-trial recommendation resulting in the Matching AND-OR Model having higher PQI score than its simpler counterpart; trial-patient performance suffers slightly but still remains competitive with other models. Similar behavior in the Siamese network setting. However, the Matching variant does not improve on the PQI and suffers a significant drop on TAUC. While this is somewhat surprising, it is possible that a different Siamese architecture combined with the matching setup may produce better results.

Table 1: Performance comparisons on the test set

Model	Patient Quantile Index	Trial AUC
Trial Ensemble Classifier	0.840	0.820
Siamese Network	0.842	0.820
Matching Siamese Network	0.839	0.771
AND-OR SVM	0.872	0.833
Matching AND-OR Model	0.879	0.813
AND-OR Attention Model	0.876	0.841

[0087] Adding an attention mechanism to the base AND-OR model offers a modest increase in performance, both in PQI and TAUC. A possible reason for this is that clinical

department name does not provide enough context to drastically change how patient representations are weighted by the distance function.

[0088] Exemplar SVM is a strong baseline method in image applications. The downside of this method, however, is that its performance suffers considerably compared to that of the similarity models if the number of positive examples is small. To test the extent of this effect, the system artificially decreases the number of positive examples in the test set by randomly sub sampling a maximum number of enrolled patients per trial (1, 5, 10, 50, >50), re-train the TEC and re-apply the AND-OR SVM. The difference is most pronounced for 1 positive example, subsides for 5, and almost disappears as the maximum number of positive examples grows for both TAUC and PQI measures, as expected.

Table 2: Trial recommendations for a sample breast cancer patient.

Trials	Match Score
A Phase I/II Study of a Novel Indibulin Dosing Schedule for the Treatment of Metastatic Breast Cancer (NCT01113970)	0.860853
Xenogeneic HER2/NEU DNA Immunization for Patients with Metastatic or High Risk Breast Cancer: a Phase I Study to Assess Safety and Immunogenicity (NCT00393783)	0.279413
A Phase I Investigation of the Combination of MK-2206, Trastuzumab and Lapatinib in HER2+ Solid Tumors (NCT00963547)	0.275120
A Phase II Study of Cisplatin or Carboplatin for Triple-Negative Metastatic Breast Cancer and Evaluation of p63/p73 as a Biomarker of Response (NCT00483223)	0.175838
Adjuvant Paclitaxel and Trastuzumab for Node-Negative HER2-Positive Breast Cancer (NCT00542451)	-0.35151

[0089] In addition to quantitative evaluation, the system explored whether feature weights learned by the AND-OR SVM are clinically meaningful (Table 3). The top features can come from the AND operation – e.g., shared between patients. A lot of the feature weights make intuitive sense as prior treatment with Vorinostat is a strict inclusion criterion for a number of

lymphoma-focused clinical trials. HIV-focused trials look for patients with the condition; similarly, specific cancer types appear highly ranked as does positive status for particular mutations (MDM2 and IDH1). The system also analyzed recommendations the models return for specific patients. Table 2 shows trial recommendations, using AND-OR-SVM for a metastatic breast cancer patient with a HER2 positive mutation, sampled from the test set. Only 5 trials were available for this patient. The top most recommended trial is the trial the patient actually enrolled on. All trials except the last are for metastatic cancers. The patient is technically ineligible for the last trial as it is an early stage trial and gets a negative match score. All trials require HER2 positive status except the second to last which requires HER2 negative.

Table 3: Top Features by weight from the AND-OR SVM Model

Vorinostat	HIV infection (CCS 5)
Cancer of respiratory and intrathoracic systems (CCS 20)	Tacrolimus
Cancer of bladder (CCS 32)	Hodgkin's lymphoma (CCS 37)
Cancer of liver and intrahepatic bile duct (CCS 16)	Cancer of stomach (CCS 13)
Cancer of thyroid (CSS 36)	Cancer of pancreas (CCS 17)
Donor laboratory value: normal	Pamidronate
MDM2 gene positive	IDH1 gene positive

[0090] FIG. 4A illustrates a trial AUC and FIG. 4B illustrates a plot of PQI. The graphs illustrated in FIGS. 4A and 3B illustrate a comparison of the performance of TEC and AND-OR SVM as a function of the number of positive examples. The graphs illustrate that performance of the system increases as the number of positive examples increases.

[0091] The aspects of the present disclosure may be embodied as a system, method or computer program product. Accordingly, aspects of the present disclosure may take the form of an entirely hardware embodiment, an entirely software embodiment (including firmware, resident software, micro-code, etc.) or an embodiment combining software and hardware aspects that may all generally be referred to herein as a “circuit,” “engine,” “module” or “system.” Furthermore, aspects of the present disclosure may take the form of a computer program product embodied in one or more computer readable medium(s) having computer

readable program code embodied thereon. Aspects of the present disclosure may be implemented using one or more analog and/or digital electrical or electronic components, and may include a microprocessor, a microcontroller, an application-specific integrated circuit (ASIC), a field programmable gate array (FPGA), programmable logic and/or other analog and/or digital circuit elements configured to perform various input/output, control, analysis and other functions described herein, such as by executing instructions of a computer program product.

[0092] Any combination of one or more computer readable medium(s) may be utilized. The computer readable medium may be a computer readable signal medium or a computer readable storage medium. A computer readable storage medium may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples (a non-exhaustive list) of the computer readable storage medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing. In the context of this document, a computer readable storage medium may be any tangible medium that can contain, or store a program for use by or in connection with an instruction execution system, apparatus, or device.

[0093] A computer readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband or as part of a carrier wave. Such a propagated signal may take any of a variety of forms, including, but not limited to, electro-magnetic, optical, or any suitable combination thereof. A computer readable signal medium may be any computer readable medium that is not a computer readable storage medium and that can communicate, propagate, or transport a program for use by or in connection with an instruction execution system, apparatus, or device.

[0094] Program code embodied on a computer readable medium may be transmitted using any appropriate medium, including but not limited to wireless, wireline, optical fiber cable, RF, etc., or any suitable combination of the foregoing. Computer program code for carrying

out operations for aspects of the present disclosure may be written in any combination of one or more programming languages, including an object oriented programming language such as Java, Smalltalk, C++ or the like and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The program code may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider).

[0095] Aspects of the present disclosure are described below with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems) and computer program products according to embodiments of the disclosure. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer program instructions. These computer program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0096] These computer program instructions may also be stored in a computer readable medium that can direct a computer, other programmable data processing apparatus, or other devices to function in a particular manner, such that the instructions stored in the computer readable medium produce an article of manufacture including instructions which implement the function/act specified in the flowchart and/or block diagram block or blocks. The computer program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other devices to cause a series of operational steps to be performed on the computer, other programmable apparatus or other devices to produce a computer implemented process such that the instructions which execute on the computer or other programmable apparatus provide processes for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0097] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of code, which comprises one or more executable instructions for implementing the specified logical function(s). It should also be noted that, in some alternative implementations, the functions noted in the blocks may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts, or combinations of special purpose hardware and computer instructions.

[0098] Reference in the specification to “one embodiment” or “an embodiment” of the present principles, as well as other variations thereof, means that a particular feature, structure, characteristic, and so forth described in connection with the embodiment is included in at least one embodiment of the present principles. Thus, the appearances of the phrase “in one embodiment” or “in an embodiment”, as well any other variations, appearing in various places throughout the specification are not necessarily all referring to the same embodiment.

[0099] While operations are depicted in the drawings in a particular order, such operations are not required to be performed in the particular order shown or in sequential order, and all illustrated operations are not required to be performed. Actions described herein can be performed in a different order.

[0100] The separation of various system components does not require separation in all implementations, and the described program components can be included in a single hardware or software product.

[0101] Having now described some illustrative implementations, it is apparent that the foregoing is illustrative and not limiting, having been presented by way of example. In particular, although many of the examples presented herein involve specific combinations of

method acts or system elements, those acts and those elements may be combined in other ways to accomplish the same objectives. Acts, elements and features discussed in connection with one implementation are not intended to be excluded from a similar role in other implementations or implementations.

[0102] The phraseology and terminology used herein is for the purpose of description and should not be regarded as limiting. The use of “including” “comprising” “having” “containing” “involving” “characterized by” “characterized in that” and variations thereof herein, is meant to encompass the items listed thereafter, equivalents thereof, and additional items, as well as alternate implementations including the items listed thereafter exclusively. In one implementation, the systems and methods described herein can include of one, each combination of more than one, or all of the described elements, acts, or components.

[0103] As used herein, the term “about” and “substantially” will be understood by persons of ordinary skill in the art and will vary to some extent depending upon the context in which it is used. If there are uses of the term which are not clear to persons of ordinary skill in the art given the context in which it is used, “about” will mean up to plus or minus 10% of the particular term.

[0104] Any references to implementations or elements or acts of the systems and methods herein referred to in the singular may also embrace implementations including a plurality of these elements, and any references in plural to any implementation or element or act herein may also embrace implementations including only a single element. References in the singular or plural form are not intended to limit the presently disclosed systems or methods, their components, acts, or elements to single or plural configurations. References to any act or element being based on any information, act or element may include implementations where the act or element is based at least in part on any information, act, or element.

[0105] Any implementation disclosed herein may be combined with any other implementation or embodiment, and references to “an implementation,” “some implementations,” “one implementation” or the like are not necessarily mutually exclusive and are intended to indicate that a particular feature, structure, or characteristic described in connection with the implementation may be included in at least one implementation or embodiment. Such terms as used herein are not necessarily all referring to the same

implementation. Any implementation may be combined with any other implementation, inclusively or exclusively, in any manner consistent with the aspects and implementations disclosed herein.

[0106] The indefinite articles “a” and “an,” as used herein in the specification and in the claims, unless clearly indicated to the contrary, should be understood to mean “at least one.”

[0107] References to “or” may be construed as inclusive so that any terms described using “or” may indicate any of a single, more than one, and all the described terms. For example, a reference to “at least one of ‘A’ and ‘B’” can include only ‘A’, only ‘B’, as well as both ‘A’ and ‘B’. Such references used in conjunction with “comprising” or other open terminology can include additional items.

[0108] Where technical features in the drawings, detailed description or any claim are followed by reference signs, the reference signs have been included to increase the intelligibility of the drawings, detailed description, and claims. Accordingly, neither the reference signs nor their absence have any limiting effect on the scope of any claim elements.

[0109] The systems and methods described herein may be embodied in other specific forms without departing from the characteristics thereof. The foregoing implementations are illustrative rather than limiting of the described systems and methods. Scope of the systems and methods described herein is thus indicated by the appended claims, rather than the foregoing description, and changes that come within the meaning and range of equivalency of the claims are embraced therein.

CLAIMS

What is claimed:

1. A system to identify clinical trials using probabilistic modeling, comprising:
 - a trial selector comprising one or more processors and memory to execute a patient reader and a similarity scorer, the trial selector to:
 - receive a plurality of input trials, each of the plurality of input trials having a respective plurality of input patients;
 - generate, for each input patient of the respective plurality of input patients, a patient vector comprising a plurality of features;
 - calculate, for each of the plurality of input trials, a patient-patient metric indicating a similarity between the respective plurality of input patients, wherein the patient-patient metric for each of the plurality of input trials is based on the patient vector for each input patient of the respective plurality of input patients;
 - generate, for a query patient, a query patient vector comprising the plurality of features;
 - calculate, for each of the plurality of input trials, a patient-trial similarity score indicating a similarity between the query patient and the respective plurality of input patients, wherein the patient-trial similarity score is based on the query patient vector and at least one of a trial feature vector or the patient-patient metric for each of the plurality of input trials; and
 - select, from the plurality of input trials, a candidate trial based on the patient-trial similarity score for each of the plurality of input trials.
2. The system of claim 1, further comprising at least one of a Siamese Networks, Matching Siamese Networks, AND-OR Support Vector Machine (SVM) models, Matching AND-OR SVM models, or AND-OR Attention models to calculate the patient-patient vector.
3. The system of claim 1, wherein the plurality of features comprises at least two of a patient diagnosis code, a procedure code, demographic information, laboratory values, medications, genetic mutations, or textual assessment notes.

4. The system of claim 1, further comprising the trial selector to receive a second plurality of input trials, wherein the second plurality of input trials do not include enrolled patients.
5. The system of claim 4, further comprising the trial selector to generate the trial feature vector for each of the second plurality of input trials, the trial feature vector based on metadata of each of the second plurality of input trials.
6. The system of claim 5, further comprising the trial selector to calculate a second patient-trial similarity indicating a similarity between the query patient and each of the second plurality of input trials.
7. The system of claim 5, wherein the trial feature vector is one-hot encoded.
8. The system of claim 1, further comprising the trial selector to:
 - extract metadata from a patient record associated with the query patient; and
 - generate, for the query patient, the query patient vector based on the metadata.
9. The system of claim 1, further comprising the trial selector to identify, for each of the plurality of input trials, one or more eligibility requirements associated with each of the respective one of the plurality of input trials.
10. The system of claim 9, further comprising the trial selector to determine, for the query patient, whether the query patient has the one or more eligibility requirements associated with each of the respective one of the plurality of input trials.
11. A method to identify clinical trials using probabilistic modeling, comprising:
 - receiving, by at least one or more processors, a plurality of input trials, each of the plurality of input trials having a respective plurality of input patients;
 - generating, by the at least one or more processors, for each input patient of the respective plurality of input patients, a patient vector comprising a plurality of features;
 - calculating, by the at least one or more processors, for each of the plurality of input trials, a patient-patient metric indicating a similarity between the respective plurality of input patients, wherein the patient-patient metric for each of the plurality

of input trials is based on the patient vector for each input patient of the respective plurality of input patients;

generating, by the at least one or more processors, for a query patient, a query patient vector comprising the plurality of features;

calculating, by the at least one or more processors, for each of the plurality of input trials, a patient-trial similarity score indicating a similarity between the query patient and the respective plurality of input patients, wherein the patient-trial similarity score is based on the query patient vector and at least one of a trial feature vector or the patient-patient metric for each of the plurality of input trials; and

selecting, by the at least one or more processors, from the plurality of input trials, a candidate trial based on the patient-trial similarity score for each of the plurality of input trials.

12. The method of claim 11, wherein calculating, for each of the plurality of input trials, the patient-patient metric further comprising determining the patient-patient metric with at least one of a Siamese Networks, Matching Siamese Networks, AND-OR Support Vector Machine (SVM) models, Matching AND-OR SVM models, or AND-OR Attention models.
13. The method of claim 11, wherein the plurality of features comprises at least two of a patient diagnosis code, a procedure code, demographic information, laboratory values, medications, genetic mutations, or textual assessment notes.
14. The method of claim 11, further comprising receiving a second plurality of input trials, wherein the second plurality of input trials do not include enrolled patients.
15. The method of claim 14, further comprising generating the trial feature vector for each of the second plurality of input trials, the trial feature vector based on metadata of each of the second plurality of input trials.
16. The method of claim 15, further comprising calculating a second patient-trial similarity indicating a similarity between the query patient and each of the second plurality of input trials.
17. The method of claim 15, wherein the trial feature vector is one-hot encoded.

18. The method of claim 11, further comprising:
 - extracting metadata from a patient record associated with the query patient;
 - and
 - wherein generating, for the query patient, the query patient vector further comprising generating the query patient vector based on the metadata.
19. The method of claim 11, further comprising identifying, for each of the plurality of input trials, one or more eligibility requirements associated with each of the respective one of the plurality of input trials.
20. The method of claim 19, further comprising determining, for the query patient, whether the query patient has the one or more eligibility requirements associated with each of the respective one of the plurality of input trials.

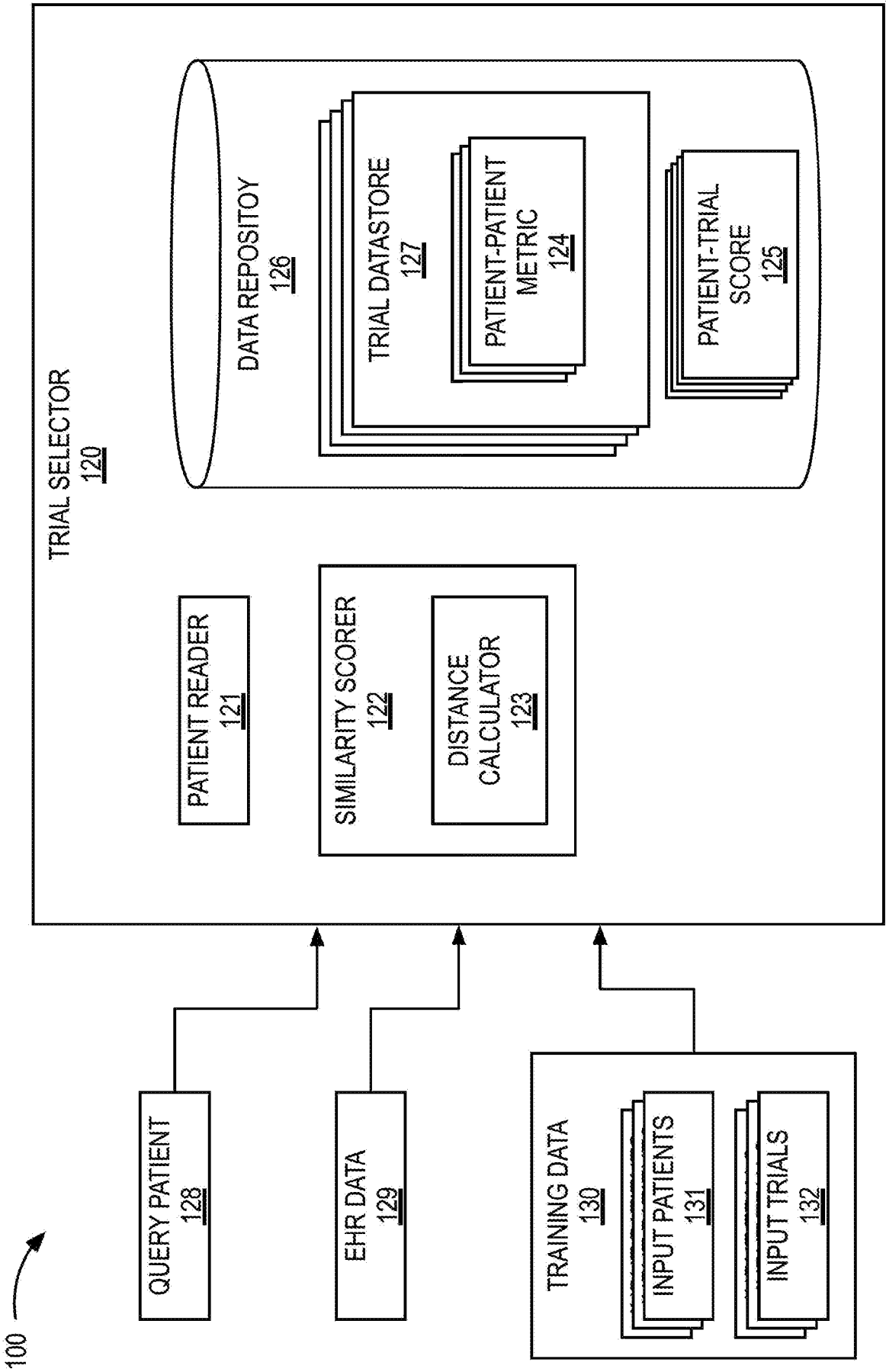


FIG. 1

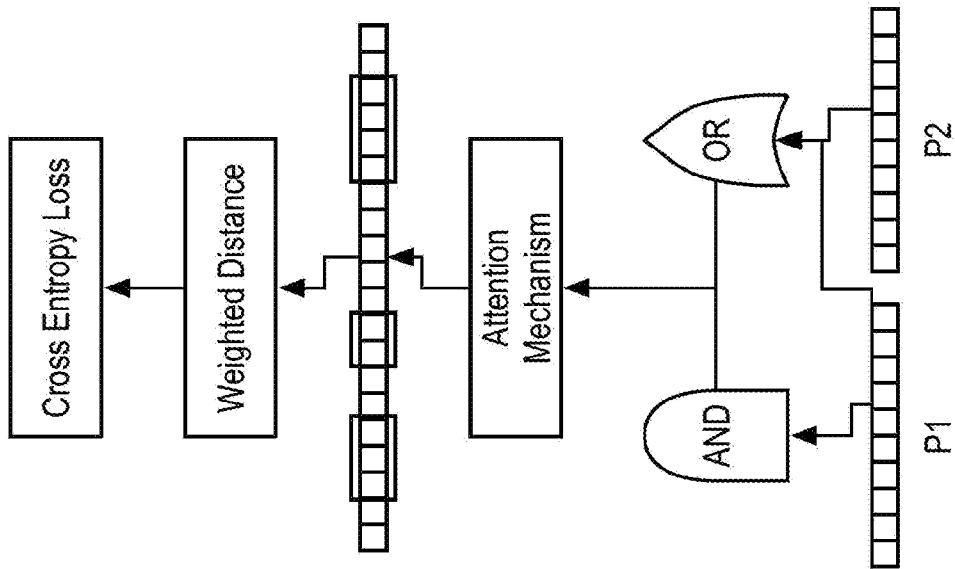


FIG. 2C

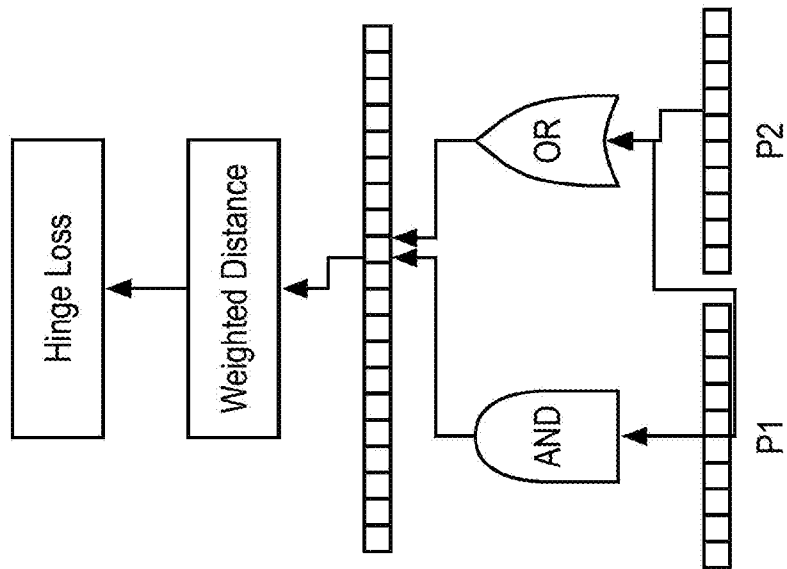


FIG. 2B

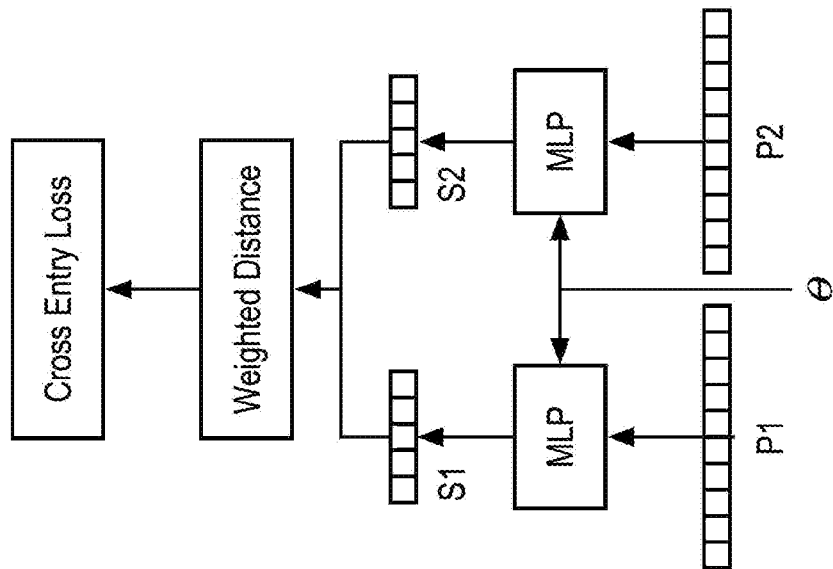
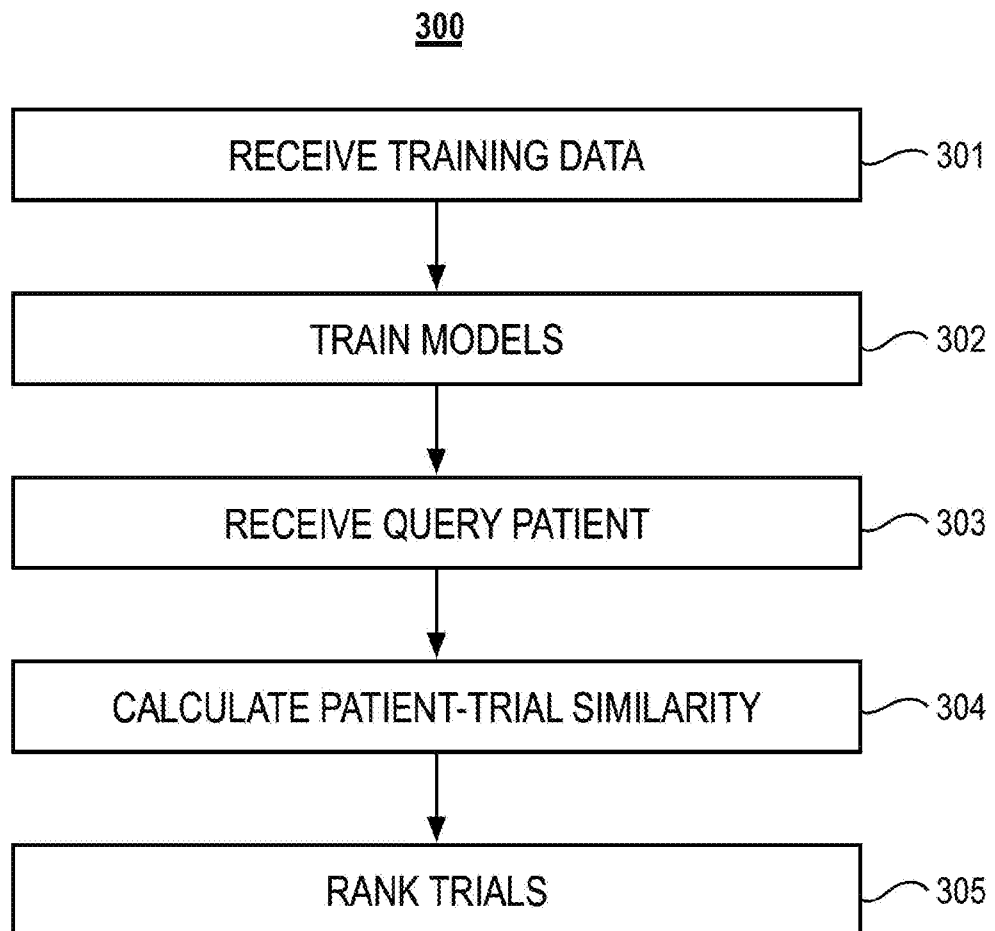


FIG. 2A

**FIG. 3**

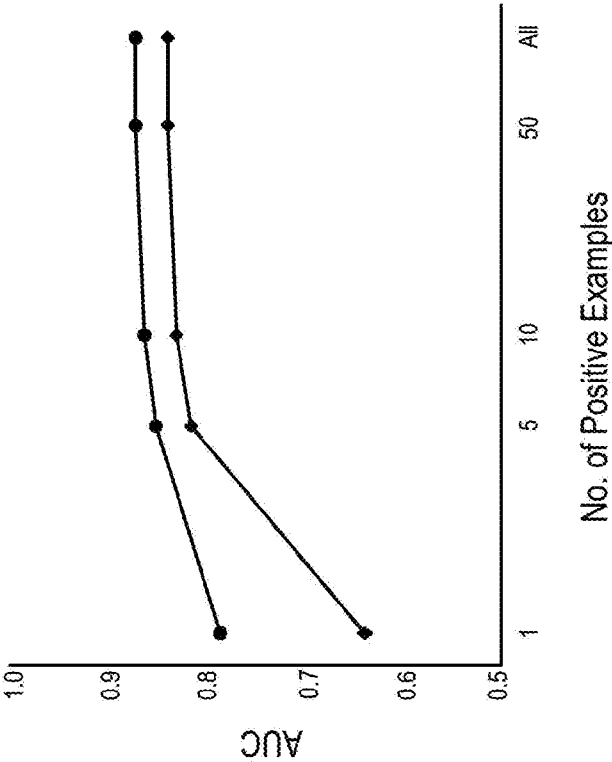


FIG. 4B

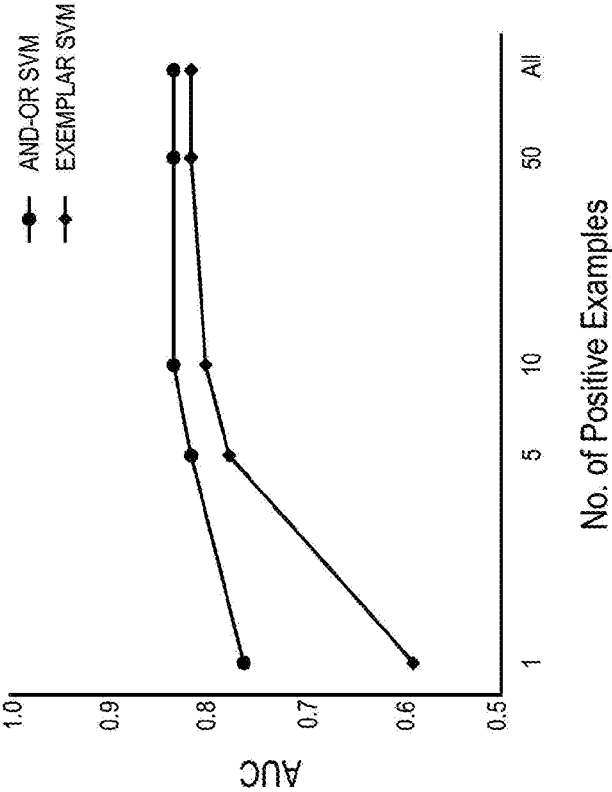


FIG. 4A

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2018/056339

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 19/00; G06F 19/24; G06Q 50/00; G16H 10/60 (2018.01)

CPC - G16H 10/20; G06F 19/24; G16H 10/60 (2018.08)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

USPC - 705/2; 705/3; 706/12 (keyword delimited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2010/0312798 A1 (DUTTA et al) 09 December 2010 (09.12.2010) entire document	1-20
Y	US 2017/0185730 A1 (CASE WESTERN RESERVE UNIVERSITY) 29 June 2017 (29.06.2017) entire document	1-20
Y	US 2013/0304484 A1 (MARTINEZ et al) 14 November 2013 (14.11.2013) entire document	4-10, 14-20
Y	US 2012/0121194 A1 (YAGNIK) 17 May 2012 (17.05.2012) entire document	7, 17
A	US 2008/0033658 A1 (DALTON et al) 07 February 2008 (07.02.2008) entire document	1-20
P,A	US 2017/0351816 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 07 December 2017 (07.12.2017) entire document	1-20
A	WO 2016/20345/ A1 (KONINKLIJKE PHILIPS N.V.) 22 December 2016 (22.12.2016) entire document	1-20

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

26 November 2018

Date of mailing of the international search report

04 DEC 2018

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, VA 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Blaine R. Copenheaver

PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774