様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １７ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ０９１６１

研究課題名（和文）電子的診療情報からの高次元特徴量による患者状態の表現と機械学習の適用に関する研究

研究課題名（英文）Development of EHR Based Phenotyping with High Dimensional Patient Information

研究代表者
河添　悦昌（Kawazoe, Yoshimasa）

東京大学・医学部附属病院・特任准教授

研究者番号：１０６２１４７７

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：電子的診療情報により患者の状態を高次元の特徴ベクトルによって表現し機械学習を適用することで、院内がん登録業務への応用可能性を検討した。特徴ベクトルを、登録病名、投薬オーダの医薬品、検体検査項目の3種のカテゴリによって構成した。100,313件の症例を含むデータセットを構築し、がん症例と非がん症例を2値分類するタスクと、がん種別を多値分類するタスクの精度を評価した。前者の精度は、院内がん登録で行われる1次スクリーニングの精度に比べ若干良いと思われたが、後者の精度は十分ではなく、手術式等を含む医科診療行為コードや病理診断病名を特徴量として追加することが、精度を向上させる一つの方法と考えられた。

研究成果の学術的意義や社会的意義
病院における症例の登録業務は人手によるインテンシブな作業が必要であるため、機械学習等の技術を活用して人手による労力を軽減することが期待される。本研究は、日々の診療で発生する電子的診療情報を利用して、院内がん登録業務で行われるがん症例のスクリーニングとがん種別の分類を機械学習によって行った場合の精度を評価し、がん登録業務への応用可能性を検討したことが社会的意義としてあげられる。

研究成果の概要（英文）：This study aimed to develop EHR phenotyping algorithms which describes the patient characteristics by high-dimensional features utilizing all items contained in physician's order entry without manually selecting features. We constructed a dataset containing 100,313 patients, and evaluated the performance of machine learnings on the task of binarizing cancer and non-cancer cases, and the task of multi classification of cancer types. The former task showed better precision than the primary screening performed in hospital cancer registration, but the latter task does not seem to have sufficient accuracy. To improve the accuracy, it was considered to add surgical procedure codes and pathological diagnoses as features.

研究分野：医療情報学

キーワード：EHR Phenotyping　院内がん登録　電子的診療情報

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

1．研究開始当初の背景

（１）　電子的カルテに含まれる情報（EHR）を積極的に活用し、疾患の解明や臨床研究に役立てることが求められている。一例として、ドラッグリパーパシングは、既存薬の薬効や副作用を利用し別の疾患への適用を探るものであるが、電子カルテのデータを使ってこれを行うことで、より安価に迅速に研究仮設を提示できる可能性がある。米国の Mayo Clinic と Vanderbilt 大学では、各病院の電子カルテからがんを有する症例を抽出し、2 型糖尿病の有無と処方薬の組で層別化した症例群の予後を比較することで、糖尿病治療薬であるメトホルミンの抗種瘍効果を示唆する研究を行った。別の例として、米国 NIH が主導する eMERGE プロジェクトでは、ゲノムワイド関連解析のために多施設の EHR を活用し、効率的な症例コホートの抽出を行っている。このような研究を行うためには、特定の疾患を有する症例を抽出する必要があり、その際にキーとなる有力な情報は病名情報であるが、電子カルテに登録される病名は保険請求目的の病名であるため、実際に患者さんにその病気が存在したかどうかということと関連が低い点が問題となる。例えば、2 型糖尿病症例の抽出を目的とした場合、2 型糖尿病の病名が登録された症例であっても、検査を行うために病名が登録されただけで、実際には 2 型糖尿病を有さない場合や、逆に 2 型糖尿病の病名が登録されずとも 2 型糖尿病を有する場合がある。また、単に糖尿病とだけ登録され、1 型か 2 型かの区別がつかない場合も多い。このようなことは、社会制度の異なる諸外国においても同様の傾向にあり、心不全症例を同定するにあたり、診断コード（ICD-9 もしくは ICD-10 コード）のみを利用するだけでは不十分であったとの結果も報告されている。このような事情から、レセプト病名や処方歴、検査値歴、診療録など複数種類の診療情報を組み合わせ、これらに含まれる幾つかの項目を変数として目的の疾患を有する症例を同定する e-Phenotyping は重要な技術となる。

（２）　e-Phenotyping のアルゴリズムを作成する方法としてルールベースを用いる方法と機械学習を用いる方法が考えられる。ルールベースの例として、前述の eMERGE プロジェクトでは各種疾患に対する Phenotyping アルゴリズムを開発し、その成果として 2 型糖尿病や慢性心不全など約 30 種類のアルゴリズムを公開している。ルールベースによる方法は、医療の専門家の知識が必要となるため、これを行えるものが少ないことがボトルネックとなる。一方、機械学習による方法はルールベースに比べて解釈性が劣るものの、正例・負例を付与したデータセットの作成と適用する機械学習アルゴリズムの開発を分けて行うことが可能となるため、効率的なアルゴリズムの開発が期待される。ここで、機械学習による方法の精度を上げるためには有効な特徴量をあらかじめ特定しておく必要があり、疾患領域ごとに異なる専門家にこれを依頼することは、迅速な機械学習モデル開発の支障となる。そこで、本研究申請者は疾患ごとに特徴量を取捨選択するのではなく、項目コードによって表現される診療情報の全項目を特徴量として用いた、高次元の特徴ベクトルによって患者の状態を表現する方法が、より多くの疾患に対する Phenotyping アルゴリズムを開発する上で有効であるとの着想に至った。

2．研究の目的

（１）　電子的診療情報を用いて患者の疾患や状態を高次元の特徴ベクトルによって表現するための方法を検討し、院内がん登録業務におけるがん症例のスクリーニングならびにがん種別の分類を機械学習によって行うことの実用可能性を検討する。

3．研究の方法

（１）　研究用データベースの構築

申請者が所属する東大病院の SS-MIX2 標準化ストレージから患者基本情報、診断病名、投薬情報、検体検査結果を抽出し、院内がん登録情報からがんの種別を抽出し、これらデータの患者 ID を仮名化し、氏名、住所、保険番号等情報を削除した上でデータベースを構築する。

（２）　患者の疾患や状態を表現する特徴量の設計

機械学習の入力となる患者の特徴量を、登録病名、投薬オーダの医薬品、検体検査項目によって構成する。これらのコードをそのまま用いると粒度が細かくなりすぎ、特徴量の数が不当に多くなる。そのため、既存の診療用マスタと外部知識を用いて、医学的に意味を保つ程度の適当な粒度に抽象化して扱うことを検討する。

（３）　がん症例を同定・分類する機械学習モデルの開発ならびに精度評価

院内がん登録業務では 2 段階の方法によりがん症例の登録を行っている。1 段階目では保険病名として登録される ICD-10 コードと病理診断名によって候補症例をスクリーニングし、2 段階目ではがん登録の専門スタッフが候補症例の診療録を目視することでがんの種別を分類している。ここで、上記の特徴量を入力とする機械学習により、1 段階目のスクリーニングと 2 段階目のがん種別の分類をどの程度の精度で行うことができるかを検討する。

4. 研究成果

（1） 研究用データセットの要約

SS-MIX2 ストレージに含まれる診療情報から、登録病名(PPR-01)、処方・ 注射オーダ(OMP-01, OMP-02)、検体検査結果 (OML-11)を抽出した。また、院内がん登録情報と突合し、がん登録の対象となった患者についてはがんの種別を含めデータセットを作成した。このデータセットには、304,000 件の症例が含まれたが、極端に情報が少ない症例を除外するために、2 回以上の受診歴、1 回以上の登録病名、処方・注射オーダ、検体検査結果をそれぞれ有することを基準としてデータセットを構築した。表 1 に構築したデータセットをがん症例と非がん症例を層別化した上での要約を示す。

表 1 構築したデータセットの要約統計

	がん症例	非がん症例
症例数 (%)	21475 (21.4)	78838 (78.6)
年齢　中央値 (範囲)	66 (0-99)	54 (0-101)
性別 (%)		
男性	12,582 (58.6)	34,596 (43.9)
女性	8.893 (41.4)	44,272 (56.1)
観察期間　中央値	953 (2-2554)	1330 (2-2552)
登録病名数　中央値	12 (2-197)	7 (2-138)
処方・注射薬品数　中央値	43 (2-1264)	19 (2-2689)
検体検査結果　中央値	30 (2-646)	9 (2-876)

（2） 特徴量の設計

機械学習の入力となる患者の特徴量を、登録病名、投薬オーダの医薬品、検体検査項目の 3 種類のカテゴリによって構成した。登録病名は ICD-10 の分類コードを特徴として用い、その出現頻度を特徴の値として用いた。登録病名には 4 桁もしくは 5 桁の ICD-10 コードが付与されるが、がん症例かどうかの判別にがんの詳細分類までの情報は不要と考え、これを 3 桁のコードに集約した。例えば、C16.0 は胃の悪性新生物・噴門部、C16.1 は胃の悪性新生物・胃底部であるが、これらを C16 胃の悪性新生物として扱った。医薬品は処方オーダと注射オーダの区別なく、医薬品の成分を特徴として扱うために、HOT コードに対応付けた KEGG コードを用い、その出現頻度を特徴の値として用いた。例えば、HOT コード 1090932 の 5-FU 錠 50mg と 1091083 の 5-FU 注 250mg はそれぞれ、KEGG コード D00584 のフルオロウラシルに対応付けることができるため、これらを同一の特徴として扱った。検体検査項目は、標準コードとして JLAC10 の分類コードが用いられるが、抽出したデータセットにおいては標準コードの付与割合が高くなかった。そのため、病院のローカルコードを用い、その出現頻度を特徴の値として用いた。結果、対象データセットには、保険登録病名が 1,411 種類、医薬品が 910 種類、検体検査項目が 1,740 種類出現し、機械学習モデルへの入力となる特徴量の種類は合計 4061 となった。

（3） がん症例と非がん症例を 2 値分類するタスクの精度評価

表2 アルゴリズムの評価結果

	登録病名		医薬品		検体検査		全特徴	
	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
SVM	.954	.865	.914	.800	.937	.843	.965	.900
RF	.960	.877	**.931**	**.836**	.942	.852	.963	.895
MLP	.961	.885	.922	.815	.941	.845	.969	.912
SDA	.963	.887	.923	.814	.944	.850	.970	.915
CNN1	**.966***	**.896***	**.931**	.832	**.950***	**.868***	.971	.918
CNN2	–	–	–	–	–	–	**.974***	**.924***

太字は各データセットでの最高値。添字＊は最高値とその他をDunnettの方法で比較し有意差を示したもの

前述のデータセットと設計した特徴量を元に、がん症例か非がん症例かを 2 値分類するタスクとして、Support Vector Machine (SVM), Random Forest (RF)、Stacked Auto-Encoder (SAE), Convolution Neural Network (CNN)の 4 種類の機械学習モデルを用いて分類性能を比較した。ここで、CNN は 4061 の特徴に 0 パディングを施し 64×64 の行列としたデータを 1ch で入力するもの(CNN1)と、登録病名、医薬品、検査項目の 3 つのカテゴリにそれぞれ同様に 0 パディングを施し、42×42 の行列としたデータを 3ch で入力するもの(CNN2)を用いた。結果、各モデルの分類精度は、AUC-ROC で 0.965-0.974、AUC-PR で 0.90-0.924 といずれも高い精度を示した（表 2)。特に、入力データを 3 チャネル 2 次元とする Convolution Neural Network (CNN2) を用いたモデルの精度が最も高かったことから、入力データを工夫することでより精度の高いモデルの開発が可能と考えられた。利用する特徴量の種類について、(1) 全特徴を用いた場合に最も高い AUC-ROC と AUC-PR を示し、続いて(2)登録病名のみ、(3)検体検査項目のみ、(4)医薬品のみを使用した場合の順に高い傾向を示した。このことから、登録病名のみならず、検体検査項目と投薬された医薬品を特徴として加える事で、性能が向上すると言えた。

（４）　CNN2 における Precision Recall 曲線

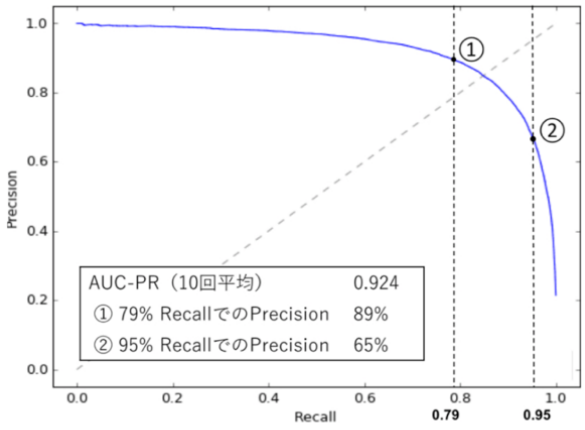

図1　CNN2におけるPrecision-Recall曲線

図 1 に最も高い性能を示した CNN2 における Precision-Recall 曲線を示す。Recall と Precision の和を最大化する曲線上のポイントは、Recall 79%、Precision 89%（図１①）であった。この分類器をがん登録の１次スクリーニングに応用することを考えた場合、症例の取りこぼしがないことを優先し高い Recall が要求されると考え、その Recall を 95%と設定した場合の Precision は約 65%（図１②）であった。

（５）　がんの種別を多値分類するタスクの精度評価

表３ がんの種別ごとの件数（上位１０位まで掲載）

順位	ICD	がん名称	件数
1	—	非がん症例	78838
2	C16	胃の悪性新生物	2621
3	C18	結腸の悪性新生物	1839
4	C61	前立腺の悪性新生物	1775
5	C22	肝及び肝内胆管の悪性新生物	1755
6	C34	気管支及び肺の悪性新生物	1409
7	C50	乳房の悪性新生物	1357
8	C20	直腸の悪性新生物	798
9	C15	食道の悪性新生物	796
10	C42	造血器の腫瘍	762

がん症例と非がん症例の２値分類に続き、がんの種別を多値分類する機械学習モデルを作成し精度評価を行った。上記（１）で構築したデータセットを用い、各症例に付与される ICD-0 3 桁のコードを分類するタスクとした。データセットには ICD-0 3 桁コードが 68 種類出現し、このうち症例数が 10 症例以下のがん種を除く 61 種類のがん種を対象とした。出現頻度の高い上位 10 位までのがん種ごとの症例の件数を表 3 に示す。機械学習モデルは L2 正則化項付きのロジスティック回帰モデルを用い、評価指標を Accuracy として 5 分割交差検定により評価した。非がん症例を 1 クラスとして扱い合計 62 クラスを分類するタスクの 5 回平均の精度は Accuracy=0.89 であった。このタスクにおける、5 分割交差検定のうち 1 回の検定における分類結果の混同行列を表 4 に示す。

表4がんの種別を多値分類（非がん症例を1クラスとして含む）するロジスティック回帰モデルの混同行列

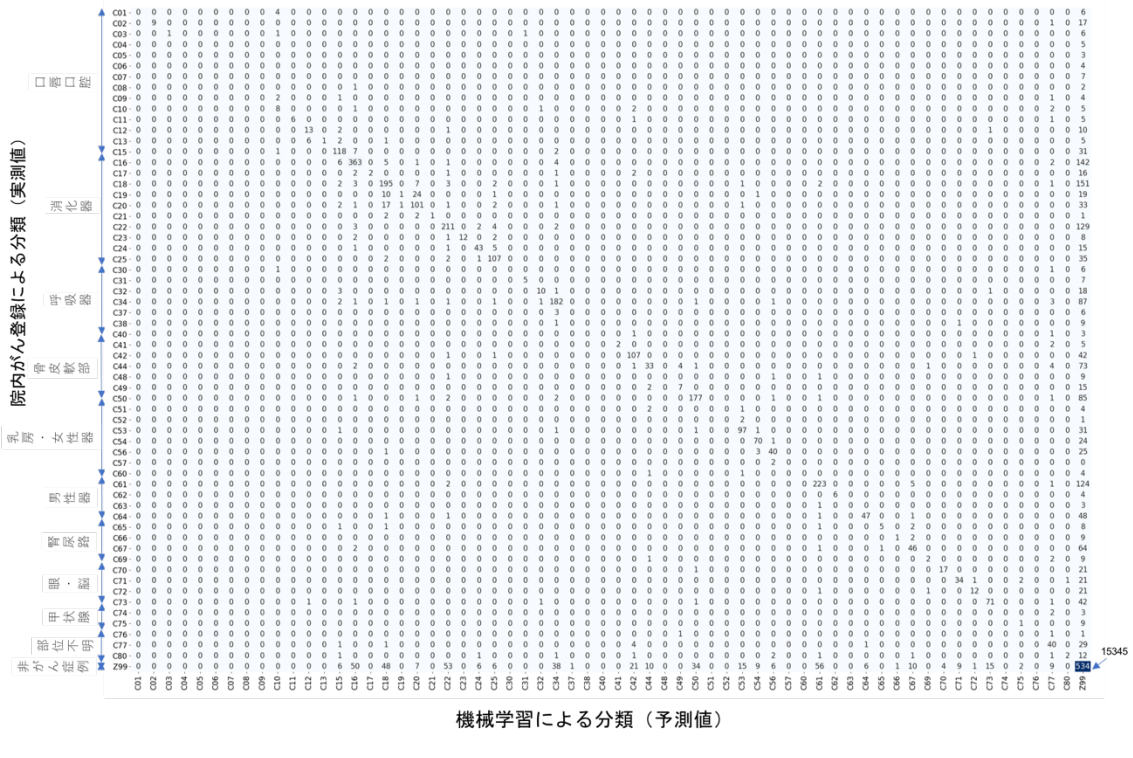

表 4 から機械学習モデルの誤分類先の頻度が高いクラスは非がん（Z99 でコード）であることがわかる。このデータセットは 76%の頻度で非がん症例を含むものであるため、機械学習による予測も非がんにふれる割合が高いためと考えられた。次に、非がん症例を含まない合計 61 クラスを分類するタスクの 5 回平均の精度は Accuracy=0.81 であった。このタスクにおける、5 分割交差検定のうち 1 回の検定における分類結果の混同行列を表 5 に示す。

表5 がんの種別を多値分類（非がん症例を除く）するロジスティック回帰モデルの混同行列

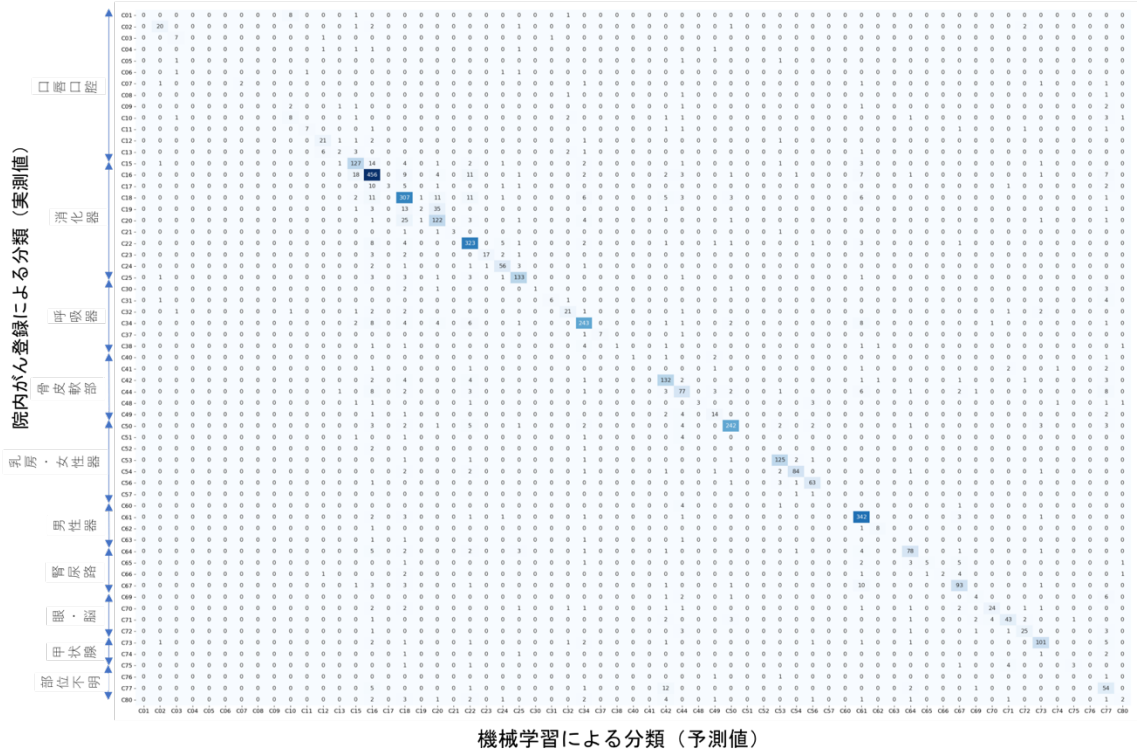

表 5 から消化器に分類される臓器のがん種間で誤分類の傾向がみられた。具体的には、C15（食道の悪性新生物）と C16（胃の悪性新生物）、ならびに C18（結腸の悪性新生物）と C19（直腸 S 状結腸移行部の悪性新生物）と C20（直腸の悪性新生物）をそれぞれ誤分類する傾向があることが読み取れた。消化器以外では、C67（膀胱の悪性新生物）を C61（前立腺の悪性新生物）と誤分類する傾向、ならびに C77（リンパ節の続発性及び部位不明の悪性新生物）を C42（造血器の悪性種瘍）と誤分類する傾向が見られた。この誤分類の原因として、近しい臓器の異なるがん種では似通った病名登録や薬品の投与が行われることが推察され、今回用いた特徴量のみでは分類精度を向上させることに本質的な限界があると思われ、手術の術式等を含む医科診療行為コードを特徴量として含めることが精度を改善させる一つの方法と考えられた。

（６）　主成分分析による次元削減の効果の検討

機械学習の入力となる患者の特徴量を、登録病名、医薬品、検体検査項目のカテゴリ毎に主成分分析を行い、寄与率の大きい特徴量を選択（累積寄与率が 99%を越えた時点で特徴量の組入を中止）し、各カテゴリの特徴量の次元を削減した上でこれらを連結して機械学習の入力とすることで、分類精度の向上がみられるかを検討した。L2 正則化項付きロジスティック回帰モデルを用いてがん症例と非がん症例を 2 値分類するタスクならびに、がんの種別を多値分類するタスク（非がん症例を含む、含まない場合）の精度を評価したところ、いずれの場合においても次元削減によって精度は向上せず、精度が変わらないもしくは若干の精度低下が見られた。この原因として、特徴量の次元数に比べてデータセットのサンプル数が十分多い（特徴量の数 4061 に対してがん症例数は 21475)ため、次元削減の効果が見られなかったものと考えられた。

（７）　院内がん登録業務への応用の検討

院内がん登録業務では登録作業の 1 段階目として登録病名と病理診断病名を組み合わせたスクリーニングが行われ、2015 年症例におけるスクリーニングの Precision は約 60%と概算された。本研究の（３）で開発した機械学習モデルは、全特徴を用いた CNN2 において Recall を 95%と設定した場合の Precision は約 65%であったことから、実際に行われているスクリーニングに比べ若干の精度向上が見込まれると思われた。また、異なるカテゴリの診療情報を特徴量として追加することで性能が向上する傾向にあることから、病理診断病名を特徴量に加えることで、機械学習モデルがより実用的な精度を示す可能性が考えられた。ただし、データセットに含まれるがん症例と非がん症例の比率が、実際の受診者の比率と大きく異なっている場合には、精度を外挿することができず、この限りではないことが一つの限界として挙げられる。次に、登

録作業2段階目のがんの種別分類について､非がん症例を含まない合計61クラスを分類する機械学習モデルの精度はAccuracy=0.81であった。人手による分類精度が不明であるため比較は困難であるが、実際の利用を考えた場合にこの精度は十分であるとは言い難い。（６）でも述べたとおり、構築したデータセットは特徴量の次元数に比べ、サンプル数が十分に多いため、異なる特徴量の追加ががん種の分類精度の向上につながる可能性があり、手術の術式等を含む医科診療行為コードや病理診断病名を特徴量として追加することが一つの方法であると考えられた。

5．主な発表論文等

〔雑誌論文〕（計　6　件）

1. 山下 英俊, 倉沢 央, 河添 悦昌, 大江 和彦, 入院レセプトの主傷病名推定に有効な説明変数の検討. 医療情報学 38(Suppl.), pp.404-409, 2018.
2. Satoshi Iwai, Yoshimasa Kawazoe, Takeshi Imai, Kazuhiko Ohe. Effects of implementing tree model of diagnosis into a Bayesian diagnostic inference system. Stud Health Technol Inform. 245, pp.882-886, 2017.
3. Rina Kagawa, Yoshimasa Kawazoe, Emiko Shinohara, Takeshi Imai, Kazuhiko Ohe. The impact of “possible patients” on phenotyping algorithms: Electronic phenotype algorithms can only be reproduced by sharing detailed annotation criteria. Stud Health Technol Inform. 245, pp.432-436, 2017.
4. 河添 悦昌, 倉沢 央, 岩井 聡, 香川 璃奈, 大江 和彦. 状態空間モデルと深層ニューラルネットワークによる検体検査結果の欠損値推定精度の比較. 医療情報学 37(Suppl.), pp.820-824, 2017.
5. 香川 璃奈, 河添 悦昌, 篠原 恵美子, 今 井健, 大江和彦. 疾患横断的な e-phenotyping 手法開発を目的とした各疾患の特徴の検討. 医療情報学37(Suppl.), pp.754-759, 2017.
6. 河添 悦昌, 香川 璃奈, 山口 亮平, 桜井 亮太, 篠原 恵美子, 大江 和彦. 電子的診療情報からの高次元特徴データを用いた EHR Phenotyping アルゴリズムの開発. 医療情報学36(Suppl.), pp.672-675, 2016.

〔学会発表〕（計 2 件）

1. 河添 悦昌. 電子カルテデータの活用と e-Phenotyping. 第 22 回日本医療情報学会春季学術大会 シンポジウム 2018 in 新潟シンポジウム 大会企画セッション, 2018.
2. 河添 悦昌, 香川 璃奈, 今井 健, 大江 和彦. 診療情報による Phenotyping の現状・限界. 第37回医療情報学連合大会シンポジウム, 2017.

〔図書〕（計　3　件）

1. 河添 悦昌, 大江 和彦. AI と ICT が変える医療, 科学評論社 腎臓内科・泌尿器科 7 巻 2 号, 2019年.
2. 河添 悦昌, 大江 和彦. これまで行われてきた医療の分野での人工知能の利用と今後の課題, 東京医学社 小児内科51巻1号（1月号）, 2019年.
3. 河添 悦昌. 医療における人工知能技術の応用, 医歯薬出版株式会社 医学のあゆみ 264巻3号 p260, 2018年.

〔産業財産権〕
○出願状況（計0件）
○取得状況（計0件）

６．研究組織
(1)研究分担者 該当なし
(2)研究協力者 該当なし