一般口演
# 一般口演6
# 医療データ分析２（ DWH,DPC,診断）
2018年11月23日(金) 14:20 ～ 15:50 F会場 (5F 502+503)

[2-F-2-3] 入院レセプトの主傷病名推定に有効な説明変数の検討
○山下 英俊[1], 倉沢 央[2], 河添 悦昌[3], 大江 和彦[1,3] （1.東京大学大学院 医学系研究科 医療情報学分野, 2.日本電信電話株式会社, 3.東京大学医学部附属病院 企画情報運営部）

【背景】国民の医療状況を正確に把握する上で国のレセプト等 DB（ NDB）の外来レセプトデータは悉皆性に優れるものの、半数以上の記録は過去の主傷病名が複数記載されており、患者の現在の主たる病名を把握できない。一方、 DPC病院における入院レセプトは、主傷病名（主病名）や医療資源を最も消費した傷病名（医療資源病名）を含め、最大13区分の病名（病名リスト）を記載する形式を用いて DPC病院が厳密に運用している。【目的】入院レセプトを用いて、病名リスト、診療行為等から主病名を推定する学習モデルを開発し精度を評価する。また推定に影響を与える変数を検討し、外来レセプトを用いた主病名推定に応用可能か考察する。【方法】入院レセプトから病名リスト、検査や手術等の診療行為、医薬品の情報を抽出し、これらを意味分類したコード体系に変換した上で、その全分類を含むホットベクトルで有無情報を表現した1593次元の説明変数を作成する。目的変数は主病名の ICD10コードに対応するワンホットベクトルとし、正則化項付きロジスティック回帰による分析を行った。【結果】主病名の推定精度は約8割であり、この結果は他の機械学習手法でも同程度であったことから、学習モデルによる差は小さいと考えられた。正則化により選択された説明変数は、病名リストに含まれる癌や心疾患等の病名が上位を占めており、説明変数として病名リストのみを用いても正解率の低下は限定的だった。また説明変数を診療行為と医薬品に限定した場合でも、病名リストの中から主病名を分類するタスクとすることで約7割の正解率が得られた。【考察】入院レセプトを用いた主病名の推定において、病名リストが精度に大きく貢献したことから、外来レセプトにおいても同様に、病名リストを用いて主病名を推定できる可能性がある。今後は年齢・性別等の患者情報を説明変数に加え、外来レセプトで評価を行う予定である。

入院レセプトの主傷病名推定に有効な説明変数の検討

山下英俊[*1]、倉沢央[*2]、河添悦昌[*3]、大江和彦[*1,3]
*1 東京大学大学院 医学系研究科 医療情報学分野、*2 日本電信電話株式会社、
*3 東京大学医学部附属病院 企画情報運営部

Investigation of the Explanatory Variables from Health Insurance Claim Databases of inpatients to Extrapolate Main Diseases in Japan

Hidetoshi Yamashita [*1] , Hisahi Kurasawa[*2] , Yoshimasa Kawazoe[*3] , Kazuhiko Ohe[*1,3]
*1 Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo,
*2 Nippon Telegraph and Telephone Corporation,
*3 Department of Healthcare Information Management, The University of Tokyo Hospital

Health insurance claim databases of outpatients, such as the National Database (NDB), provide useful data for building reliable evidence based on nationwide medical activities owing to their universal coverage; however, more than half of these records contain non-updated main disease names, making the identification of actual main disease names of outpatients difficult. We built machine learning models for extrapolating the main disease names from disease name lists and medical activities including prescriptions from inpatient health insurance claim data to evaluate the accuracy, effective explanatory variables, and dimension reduction effectiveness of objective variables by these lists for better extrapolation. We included 20,139 receipt records of 12,128 inpatients containing data from April 2016 to December 2016 from the University of Tokyo Hospital. We examined a regularized logistic regression model to investigate the explanatory variables, and this model achieved 89% accuracy on a multiclass classification task. We estimated the significance of regression coefficients using the bootstrap method, and we identified 125 important explanatory variables for these coefficients such as clinical tests and drugs for internal organs. We found that dimension reduction of objective variables by disease name lists was ineffective in improving classification accuracy of main diseases when disease name lists were used as explanatory variables. Therefore, disease name lists, medical activities, and drugs of inpatients are useful explanatory variables in extrapolating the main diseases of inpatients, and these outpatient variables are also expected to be useful in extrapolating the main diseases of outpatients. However, dimension reduction of objective variables by disease name lists may not be effective in improving classification accuracy of main diseases among outpatients.

Keywords: main disease name, health insurance claims, machine learning, logistic model, bootstrap

1. 緒論

国民の医療状況を正確に把握する上で国のレセプト等 DB (NDB)の外来レセプトデータは悉皆性に優れるものの、半数以上の記録は過去の主傷病名が複数記載されており、患者の現在の主たる病名を把握できない [1]。したがって特定の病名の患者数 [2]や医療費、その増減を正確に把握することができない。一方、DPC 病院における入院レセプトは、主傷病名や医療資源を最も消費した傷病名を含め、最大 13 区分の病名を記載する形式を用いて厳密に運用されている。このことから主傷病名が明確な入院レセプトを用いて、これを推定する適切な機械学習モデルが構築できれば、外来と入院とでレセプトの情報量と質の違いはあるものの、このモデルを外来レセプトの主傷病名の推定に応用できる可能性がある。主傷病名を推定するにあたり、その種類の多さが推定を困難にすると予想されることから、目的変数の選択範囲に病名区分による制約を設けることで、主傷病名の推定精度が向上するかどうかを検討した。

2. 目的

入院レセプトを用いて、病名区分、診療行為等から主傷病名を推定する機械学習モデルを開発し、病名区分による制約を設けることが、主傷病名の推定に有効かどうかを評価する。そして推定に影響を与える変数を検討する。また本方法が外来レセプトを用いた主傷病名の推定に応用可能かを考察する。

3. 方法

3.1 方法の概要

入院および外来レセプトから病名、検査や手術等の診療行為、医薬品の情報を抽出し、これらの次元を削減し、推定に有効な変数を検討する利便性から、意味分類したコード体系に変換することで類似のコードをグルーピングし、それらの全コードに対する有無(0 または 1)を表すホットベクトルを説明変数とした。目的変数は主傷病名の国際疾病分類第 10 版(ICD10)コードに対応するワンホットベクトルとした。入院レセプトを用いて学習したモデルの外来レセプトへの外挿性を評価するため、入院および外来レセプトにおける説明変数および目的変数を構成する各コードの出現頻度をカウントし、両者の間で順位相関係数を算出し比較検討した。主傷病名を推定する機械学習モデルは L1 正則化ロジスティック回帰を採用し、主傷病名の推定に有効な説明変数を検討した。ここでロジスティック回帰は線形モデルであるため、十分な適合が得られない可能性を考慮し、非線形モデルである SVM を用いて同様の推定を行い、精度を比較した。また主傷病名の推定に有効な説明変数であるかどうかを回帰係数の総和の大小により評価したが、この際にブートストラップ標本を用いてモデルの作成を複数回行い、各モデルにおける回帰係数の 95%信頼区間を算出することで行った。主傷病名の推定における病名区分による制約の効果は、主傷病名の推定時に目的変数の選択範囲として病名区分による制約を設けた場

合とそうでない場合の 2 種類の方法による推定を行い、Accuracy を用いた比較により評価した。

3.2 データソース

東京大学医学部附属病院における 9 ヶ月間(2016 年 4 月から 12 月)に来院された患者のレセプトを用いて解析を行った。入院および外来レセプトの 1 レコードは一ヶ月毎に患者単位で構成されており、各々20,139 レコードと 321,714 レコードを解析に用いた。なお同じ月に 2 回以上入院した患者のレセプト(いわゆる総括レセプト)は本解析の対象外とし、それらは全入院レセプトのうち 8%であった。レセプトから病名として ICD10 コード、検査や手術等の診療行為として診療行為コード、医薬品として医薬品コード(支払基金コード)、を抽出した(表 1)。なお一般的にレセプトは、年齢や性別等の患者情報を含むが、本研究では用いていない。

表 1　レセプトの記載コードとマッピング情報

記載コード	対応表・一覧表	マッピング先
ICD10 コード	標準病名マスター	
診療行為コード	医科診療行為マスター	診療報酬点数
医薬品コード	医薬品マスター	薬効分類コード

本研究で用いた入院レセプトに欠損値は無かったが、主傷病名を機械学習法で推定する際に交差検定法を用いて評価することから、主傷病名の出現回数が 5 未満の 2,156 レコード(1,057 種類の主傷病名を含み、全入院レセプトの 11%に相当するレコード)については、機械学習の解析対象外とした。本研究に用いたレセプトに関する個人情報保護については、所属機関のウェブサイトへの掲示により周知を行い、本研究への情報提供を希望しない患者のレセプトは用いていない。解析には抽出時にシステムにより、連結不可能匿名化されたデータを使用した。本研究は、東京大学医学部倫理委員会の承認を得て実施した(審査番号:11939)。

3.2.1 病名区分

入院レセプトは表 2 のように 13 種類の病名区分[3]を含み、外来レセプトは表 3 のように 2 種類の病名区分を含む。同表に、本研究に用いた入院レセプトにおける 1 レセプトあたりの、各病名区分に対する平均記載病名数および最大の記載病名数を示す。入院レセプトの全てのレコードにおいて、1 レコードに 1 つの主傷病名が記載されていた。全病名区分に記載されていた病名の数は、入院レセプトでは 1 レコードあたり平均 4.41、外来レセプトでは 10.42 だった。

表 2　入院レセプトの病名区分と記載状況

#	病名区分	1 レコードあたりの記載数		
		平均	中央値	最大値
1	傷病名(医療資源病名)	1.00	1	1
2	副傷病名	0.02	0	1
3	主傷病名	1.00	1	1
4	入院の契機となった傷病名	1.00	1	1
5	医療資源を2番目に投入した傷病名	0.22	0	1
6	入院時併存病名(1)	0.76	1	1
7	入院時併存病名(2)	0.59	1	1
8	入院時併存病名(3)	0.46	0	1
9	入院時併存病名(4)	0.36	0	1
10	入院後発症傷病名 (1)	0.48	0	1
11	入院後発症傷病名 (2)	0.31	0	1
12	入院後発症傷病名 (3)	0.22	0	1
13	入院後発症傷病名 (4)	0.16	0	1
	全 13 病名区分	4.41	4	12

表 3　外来レセプトの病名区分と記載状況

#	病名区分	1 レコードあたりの記載数		
		平均	中央値	最大値
1	主傷病名	3.38	2	49
2	副傷病名	7.04	5	79
	全 2 病名区分	10.42	8	99

3.2.2 病名コード

病名コードとして ICD10 コードを用いた。ICD10 コードの全コードのリストを MEDIS の標準病名マスター(V4.04)[4] より取得し、用いた。ICD10 コードは 7,400 種類のコードで構成され、本研究で対象とした入院レセプトには 2,179 種類の ICD10 コードが出現し、そのうち 692 種類が主傷病名に出現した。ICD10 コードは階層的な分類体系となっており、その上 1 桁(大分類に相当)に対応する分類名を表 4 に示す。また本研究で用いた入院レセプトの主傷病名および 13 区分の病名における ICD10 コードの出現頻度も同表に示した。

表 4　ICD10 コード(上 1 桁)の分類名と 1 レコードあたりの出現頻度

ICD10	分類名	出現頻度	
		主傷病名	13 区分内の病名
A	感染症および寄生虫症	1%	5%
B		0%	9%
C	新生物	30%	38%
D	血液および造血器の疾患ならびに免疫機構の障害	6%	22%
E	内分泌, 栄養および代謝疾患	4%	32%
F	精神および行動の障害	1%	6%
G	神経系の疾患	3%	17%
H	眼および付属器の疾患、耳および乳様突起の疾患	5%	8%
I	循環器系の疾患	13%	37%
J	呼吸器系の疾患	4%	17%
K	消化器系の疾患	7%	43%
L	皮膚および皮下組織の疾患	1%	9%
M	筋骨格系, 結合組織の疾患	7%	21%
N	尿路性器系の疾患	4%	12%
O	妊娠, 分娩および産褥	3%	4%
P	周産期に発生した病態	2%	2%
Q	先天奇形, 変形, 染色体異常	3%	4%
R	症状, 徴候および異常臨床所見・異常検査所見で他に分類されないもの	1%	18%
S	損傷, 中毒およびその他の外因の影響	3%	4%
T		2%	7%
	総レコード数	20,139	

3.2.3 診療行為コード

抽出されたレセプト電算処理コードの診療行為コード(全 7,185 種類の 9 桁のコード)を平成 29 年の診療報酬点数の医科点数表の区分番号(全 4,238 種類の 7 桁の分類コード)へ

社会保険診療報酬支払基金の医科診療行為マスター[5] に従いマッピングした。医科診療報酬点数はコードの桁数に応じて階層的に診療行為が分類されている。その上 1 桁(大分類に相当)に対する分類名と、本研究で用いた入院レセプトにおける1レコードあたりの出現頻度を表5に示す。本研究では医科診療報酬点数の上 4 桁(全 1,392 種類)を用い、本研究で対象とした入院レセプトには、それらのうち 692 種類が出現した。例えば B 型肝炎の検査である HBs 抗原検査、HBs 抗体検査は、7 桁のコードでは同じ"D013030"であり、HBe 抗原検査、HBe 抗体検査はそれとは異なるコード"D013040"である。上 4 桁のコードに変換した場合、いずれも同じコード"D013"となる。

表 5　医科診療報酬点数(上 1 桁)の分類名と
入院レセプト 1 レコードあたりの出現頻度および
入院と外来のレセプトの出現頻度の順位相関係数

点数	分類名	出現頻度	順位相関係数
A	初・再診料、入院料等	93%	0.63
B	医学管理等	71%	0.92
C	在宅医療	4%	0.92
D	検査	96%	0.80
E	画像診断	77%	0.87
F	投薬	91%	−0.21
G	注射	59%	0.78
H	リハビリテーション	15%	0.68
I	精神科専門療法	1%	0.45
J	処置	42%	0.62
K	手術	45%	0.44
L	麻酔	29%	0.60
M	放射線治療	2%	0.86
N	病理診断	27%	0.76
X	食事代等	96%	1.00
	総レコード数	20,139	

3.2.4 医薬品コード

抽出されたレセプト電算処理コードの医薬品コード(全 20,654 種類の 9 桁の分類コード)を平成 29 年の薬効分類コード(全181種類の3桁の分類コード)へ、社会保険診療報酬支払基金の医薬品マスター[5] に従いマッピングした。薬効分類コードはその桁数に応じて階層的に医薬品が分類されている。その上 1 桁(大分類に相当)に対する分類名と、本研究で用いた入院レセプトにおける 1 レコードあたりの出現頻度を表 6 に示した。本研究では薬効分類コードの 3 桁(全 181 種類)を用い、本研究で対象とした入院レセプトには、それらのうち 126 種類が出現した。例えば、神経系及び感覚器官用医薬品(コード"1")には中枢神経系用薬(コード"11")等があり、さらにその中には抗てんかん剤や抗パーキンソン剤、全身麻酔剤等の区分があり、さらに全身麻酔剤(コード"111")内に様々な剤形や容量を含めて全 48 種類の医薬品を包含する。

表 6　薬効分類コード(上 1 桁)の分類名と
1 レコードあたりの出現頻度

コード	分類名	出現頻度
1	神経系及び感覚器官用医薬品	82%
2	器官系用医薬品	88%
3	代謝性医薬品	88%
4	組織細胞機能用医薬品	20%
5	生薬及び漢方処方に基づく医薬品	7%
6	病原生物に対する医薬品	64%
7	治療を主目的としない医薬品(診断薬等)	47%
8	麻薬	26%
9	薬効不明	0%
	総レコード数	20,139

3.3 ホットベクトル化

入院レセプトにおける各レコードの病名区分、診療行為、医薬品について、各々ICD10 コード(上 1 桁または全 5 桁)、診療報酬点数(上 4 桁)、薬効分類コード(全 3 桁)、に変換した上で、各分類コードの有無情報をホットベクトル(0 または 1 で構成されるベクトル)で表現した説明変数を作成した。各コードは順に 20(または 7400)、1392、181 分類あることから、説明変数は合計 1593 次元(または 8973 次元)となった。目的変数は、各 ICD10 コード(上 1 桁または全 5 桁)が主傷病名であるかどうかをワンホットベクトル(1 つの成分のみが 1 となるワンホットベクトル)で表現して解析に用いた。

3.4 機械学習

主傷病名の推定に用いたL1正則化ロジスティック回帰は、式(A)に示すロジスティック回帰モデルに、式(B)に示す L1 型の罰則項を加えた目的関数が最小となるように、変数選択と回帰係数の推定を行う[6]。ここで$\boldsymbol{x_i}$は説明変数を表すホットベクトル、y_iは目的変数を表す二値変数(ただし説明の都合上$y_i \in \{-1,1\}$とした)、$\mathrm{prob}(y_i = 1)$は対象となる分類への所属確率、$\boldsymbol{w}$は回帰係数を表すベクトル、λ は罰則項の効果を調整する正のハイパーパラメータ、m はレコード数、n は説明変数の数である。

$$\mathrm{logit}\big(\mathrm{prob}(y_i = 1)\big) = \mathrm{w}_0 + \boldsymbol{w}^T\boldsymbol{x_i} \qquad \text{(A)}$$

$$\min_w \sum_i^m \log(1 + e^{-y_i w^T x_i}) \; + \; \boldsymbol{\lambda}\, \Sigma_j^n |w_j| \qquad \text{(B)}$$

ここでロジスティック回帰は線形のモデルであるため、十分な適合が得られない可能性を考慮し、非線形な機械学習モデルである SVM を用いて同様の推定を行い、精度を比較した。ハイパーパラメータである L1 正則化項の係数 λ はグリッドサーチによる探索を行い、目的変数および説明変数を層別サンプリングした 5×2 分割入れ子式交差検定法[7] による評価によって、主傷病名推定における Accuracy のマクロ平均[8] (以下 Accuracy)を最大化する変数選択と回帰係数の推定を行った。上記以外のハイパーパラメータは Python ライブラリの Scikit-learn version 0.19.1 のデフォルト値を用いており、具体的には多クラス分類の方法は One vs Rest、各クラスの重みは全て 1、最適解の算出は liblinear[6]、最適解の tolerance は 10^{-4} を用いた。

説明変数はすべて 0 または 1 で正規化されていることから、機械学習法として線形モデルを用いた場合、各説明変数に対する回帰係数を平等に評価することができる。そこで各説明変数に対する回帰係数の総和の大小により説明変数の寄与度を評価した。回帰係数の 95%信頼区間の算出には、ブートストラップ法を採用した[9]。入院レセプトの全レコードから重複を許したリサンプリングによる全対象者数と同数のブートストラップ標本を作成し、L1 正則化項目ロジスティック回帰による回帰係数の推定を 500 回繰り返し、95%ブートストラップパーセンタイル信頼区間を算出した。推定された回帰係数から各主傷病名に対するオッズ比を算出し、95%信頼区間の中

に 1 を含まない場合に有意な回帰係数と判定した[10]。

3.5 評価指標のベースライン

目的変数が全てのケースで最頻値の値をとる場合の Accuracy、または目的変数が取り得る選択肢からランダムに選択された場合の Accuracy の内、より大きな値を評価指標のベースラインとした。

3.6 病名区分を用いた病名の選択

入院レセプトから主傷病名を推定するタスクでは、全 ICD10 コードの中から 1 つの病名を主傷病名として推定するが、病名区分の情報を用いれば、主傷病名の選択肢は表 2 の 13 区分の病名に限定される。そこで ICD10 コードの中から 1 つの病名を主傷病名として推定するタスクで算出した各病名に対する所属確率の値を全 ICD10 コード間で比較するのではなく、13 区分の病名に限定して比較し、その中で最大の所属確率を持つ病名を主傷病名とした。このような病名区分による制約を設けて主傷病名を推定した場合と本制約を設けずに推定した場合の Accuracy を比較することで、外来レセプトにおいて複数ある病名の中から診療行為や医薬品に対応する実際の主傷病名の推定に応用する際の指標とする。病名区分の病名と説明変数、目的変数の関係を図 1 に示した。

4. 結果

4.1 変数の順位相関係数

入院レセプトの 13 区分の病名における ICD10 コード(5 桁)の出現頻度と、外来レセプトの 2 区分の病名(主傷病名および副傷病名)における ICD10 コードの出現頻度の Spearman の順位相関係数は 0.78 だった。一方、入院レセプトにおける主傷病名と 13 区分の病名の出現頻度についての順位相関係数は 0.75 だった。いずれも高い正の相関が認められた。

診療報酬点数(上 4 桁)について、入院レセプトと外来レセプトの出現頻度に関する順位相関係数は 0.62 であり、正の相関が認められた。この値が ICD10 コードの出現頻度における順位相関係数より若干低い原因は、“投薬(F)”が弱く逆相関していることや、診療報酬点数の 6 割以上の項目を占める“手術(K)”が弱い相関だったことに起因している(表 5)。

薬効分類コード(3 桁)について、入院レセプトと外来レセプトの出現頻度の順位相関係数は 0.95 であり、高い正の相関が認められた。

4.2 Accuracy

入院レセプトの病名区分の病名(ICD10 コード上 1 桁)、診療報酬点数、薬効分類コードを説明変数(1593 次元)とし、主傷病名(ICD10 コード上 1 桁)を目的変数として機械学習により主傷病名を推定した結果、表 7(左側)の Accuracy を得た。機械学習モデルとして L1 正則化ロジスティック解析と SVM で Accuracy にほぼ差は無かった。13 種類の病名区分による制約を設けずに推定した場合(表 7 左側)に対して、制約を設けて推定した場合の方(表 7 右側)が、Accuracy は同じか、わずかに改善した。

表 7 各機械学習法の Accuracy

機械学習法	制約無し		制約有り	
	Baseline	Accuracy	Baseline	Accuracy
ロジスティック回帰	30%	86%	47%	86%
SVM		85%		86%

説明変数を病名区分の病名に限定した場合(表 9 の行 e)、または説明変数を診療行為と医薬品に限定した場合(表 9 の行 c)の Accuracy は各々73%, 74%に低下した。なお主傷病名および病名区分の病名として ICD10 コード(5 桁)を用いた場合(表 9 の行 b, d, f)では、1 桁のコードを用いた場合(表 9 の行 a, c, e)と同様に病名区分による制約を設定しても Accuracy は向上しなかった。説明変数に病名区分の病名を含まない場合(表 9 の行 d)は、Accuracy (43%)の低下が著しかったが、病名区分による制約を設けることで Accuracy は 84%に改善した。

4.3 混同行列

主傷病名の推定に関する混同行列を図 2 に示した(表 9 の行 a の制約無しの結果に対応)。混同行列の対角線要素以外の要素の値は全て 0.1 以下であり、各カテゴリーにおいて誤分類よりも正しく分離されたケースが最も多かった。一方、正解率が低かったカテゴリーは、順に「症状、徴候および以上臨床所見・異常検査所見で他に分類されないもの(R)」、「感染症および寄生虫症(ICD10 コード A,B)」、「皮膚および皮下組織の疾患(L)」であり、それらの誤分類先として「新生物(C)」、「循環器系の疾患(I)」、「筋骨格系および結合組織の疾患(M)」が選ばれた。なお前者のカテゴリー(R, A, B, L)は本研究で用いた入院レセプトにおいて出現頻度が比較的低く、後者(C, I, M)は出現頻度が比較的高かった。

4.4 説明変数に対する回帰係数

5×2 交差検定法で最適化された L1 正則化ロジスティック回帰の回帰係数につき、対応する説明変数ごとに総和を図 3 に示した。またブートストラップ法により、有意と判定された回帰係数(オッズ比の対数)に対応する説明変数は 125 個だった。これらの回帰係数の値を対応する説明変数のカテゴリー(上 1 桁)ごとに総和および対応する説明変数の個数につき、各カテゴリー上位 5 件を表 8 に示した。

表 8 有意な回帰係数に対応する説明変数(上位 5 件)

分類名	有意な回帰係数の絶対値	
	総和	個数
ICD10 コード		
消化器系の疾患	6.5	9
内分泌,栄養および代謝疾患	6.3	8
循環器系の疾患	5.1	7
尿路性器系の疾患	3.2	2
感染症および寄生虫症	3.0	1
診療報酬点数		
検査	16.1	22
初・再診料、入院料等	6.1	8
医学管理等	5.6	4
投薬	3.3	2
処置	2.1	4
薬効分類コード		
器官系用医薬品	12.2	16
代謝性医薬品	5.7	11
神経系及び感覚器官用医薬品	5.1	5
病原生物に対する医薬品	3.2	2
治療を主目的としない医薬品	2.5	4

4.5 病名区分を用いた病名の選択

説明変数に病名区分の病名が含まれている場合、主傷病名の推定において、病名区分の病名による制約を設定したとしても、制約を設定しない場合と比較して Accuracy は改善しなかった(表 7, 表 9)。

5. 考察

入院レセプトと外来レセプトにおける診療報酬点数の順位相関係数は病名区分の病名のそれと比較して若干低く、項目数の多い手術が外来レセプトより入院レセプトに多く出現したことが影響したと考えられるが、説明変数である手術の回帰係数の多くは有意でない小さい値だったことから、手術が外

来レセプトでの主傷病名推定に与える影響は限定的と考えられる。

機械学習モデルのハイパーパラメータをさらに調整する余地はあるが、L1 正則化ロジスティック回帰と SVM において同程度の Accuracy が得られたことから、本解析は線形モデルにより解析対象データに対して十分な適合が得られるタスクと考えられ、ロジスティック回帰を用いた説明変数の評価結果は一定の信頼度があると考えられる。そのようなロジスティック回帰を、ブートストラップ法による再標本化データに用いて得られた回帰係数の 95%信頼区間の推定により、診療報酬点数の“検査”、薬効分類コードの“器官系用医薬品”における複数の説明変数に対する回帰係数がより多く有意となり、主傷病名の推定に有効だった。なお類似の値を持つ説明変数は、正則化により推定における貢献度を実際より過小評価する可能性があるため、さらに回帰係数の精査が必要である。

入院レセプトにおいて病名区分の病名による制約を設けて主傷病名を推定しても、Accuracy が改善しなかったことから（表 9 の行 a, b, e, f）、外来レセプトにおいても病名区分の制約下における推定による Accuracy の改善は期待できないと思われる。しかし、説明変数に病名区分の病名を含めずに、診療行為と薬効分類コードのみを用いて主傷病名を推定したケースでは病名区分による制約によって、Accuracy が飛躍的に向上したため（表 9 の行 c, d）、同様の効果は外来レセプトにおいても期待できる。

本研究では使用しなかったが、レセプトに含まれている年齢・性別等の患者情報、診療行為および医薬品の数量情報、を説明変数に加えることや、解析に用いるレコード数を増やすことで Accuracy はさらに向上すると考えられる。

6. 結論

入院レセプトにおける主傷病名（ICD10 コード 5 桁）を病名区分、診療行為、医薬品の情報を説明変数として L1 正則化ロジスティック回帰を用いて推定し、89%の Accuracy が得られた。病名区分の病名による制約を設けて主傷病名を推定しても、Accuracy は改善しなかった。これらの結果から、外来レセプトを用いた主傷病名の推定においても、病名区分の病名、診療行為、医薬品の情報を用いることで、主傷病名を精度良く推定できる可能性があるものの、病名区分による制約下における推定による Accuracy の改善は期待できないと思われた。

参考文献

1) 谷原真一，畝博．入院外レセプトにおける主傷病の記載状況について．厚生の指標, 2008: 15-20.
2) 武田理宏，三原直樹，村田泰三，真鍋史朗，松村泰志．レセプトデータを活用した医療機関ごとの疾患別受診患者数の推定．35th JCMI 2015: 388-391.
3) 社会保険診療報酬支払基金．レセプト電算処理システム 電子レセプトの作成手引き-DPC-（平成 28 年 7 月版）．[http://www.ssk.or.jp/seikyushiharai/rezept/iryokikan/iryokikan_02.html (cited 2017-8-16)]
4) MEDIS. ICD10 対応標準病名マスター. [https://www2.medis.or.jp/stdcd/byomei/index.html (cited 2018-Apr-1)]
5) 社会保険診療報酬支払基金．医科診療行為マスター.医薬品マスター[http://www.iryohoken.go.jp/shinryohoshu/downloadMenu/ (cited 2017-Sep-28)]
6) Rong F, Kai C, Cho H, Xiang W, Chih L. LIBLINEAR: A Library for Large Linear Classification. J of Machine Learning Research. 2008：9：1871-1874.
7) Sudhir V, Richard S. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 2006: 7：91-98.
8) Ming Z, Zhi Z. A Review on Multi-Label Learning Algorithms. IEEE Transactions on knowledge and data engineering. 2014: 26：8：1819-1837.
9) Efron B, Tibshirani R. An Introduction to the Bootstrap. Boca Raton：CRC press, 1994.
10) 林裕志，平松達雄，小出大介，田中勝弥，大江和彦．電子カルテデータベースからの LASSO ロジスティック回帰による医薬品副作用シグナルの検出:ケース・コントロール研究．薬剤疫学 2016：21：2：51-62.

図 1　病名区分と説明変数、目的変数の関係例

閲覧しやすいように説明変数が値を持つ欄の背景を灰色に、目的変数の正解または推測値の欄の背景を赤色に設定した。

表 9　入力変数と主傷病名の推定結果

#	入力変数					推定結果			
	目的変数		説明変数 (次元)			制約無し		制約有り	
	主傷病名(桁数)	主傷病名(次元)	病名区分内の病名	診療報酬点数	薬効分類コード	Baseline	Accuracy	Baseline	Accuracy
a	1	20	20	1,378	181	30%	86%	47%	86%
b	5	7,400	7,400			6%	89%	37%	89%
c	1	20		1,378	181	30%	74%	47%	85%
d	5	7,400				6%	43%	37%	84%
e	1	20	20			30%	73%	47%	73%
f	5	7,400	7,400			6%	85%	37%	85%

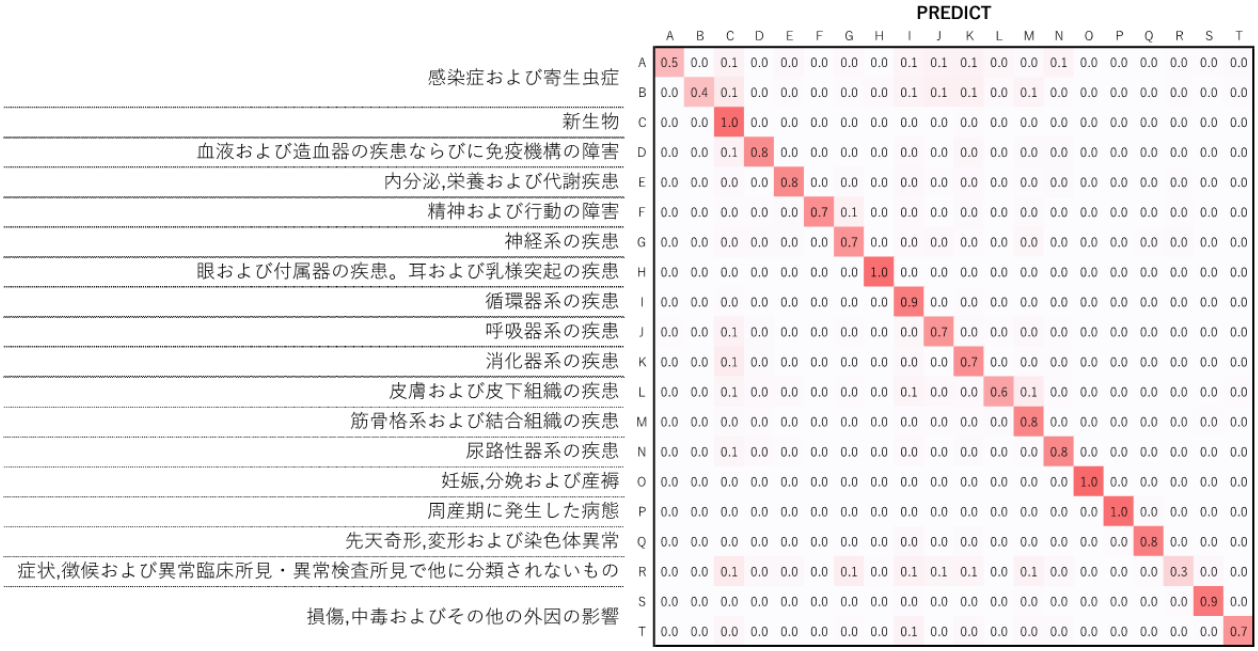

図 2　主傷病名(ICD10 コード 1 桁)の推定における混同行列

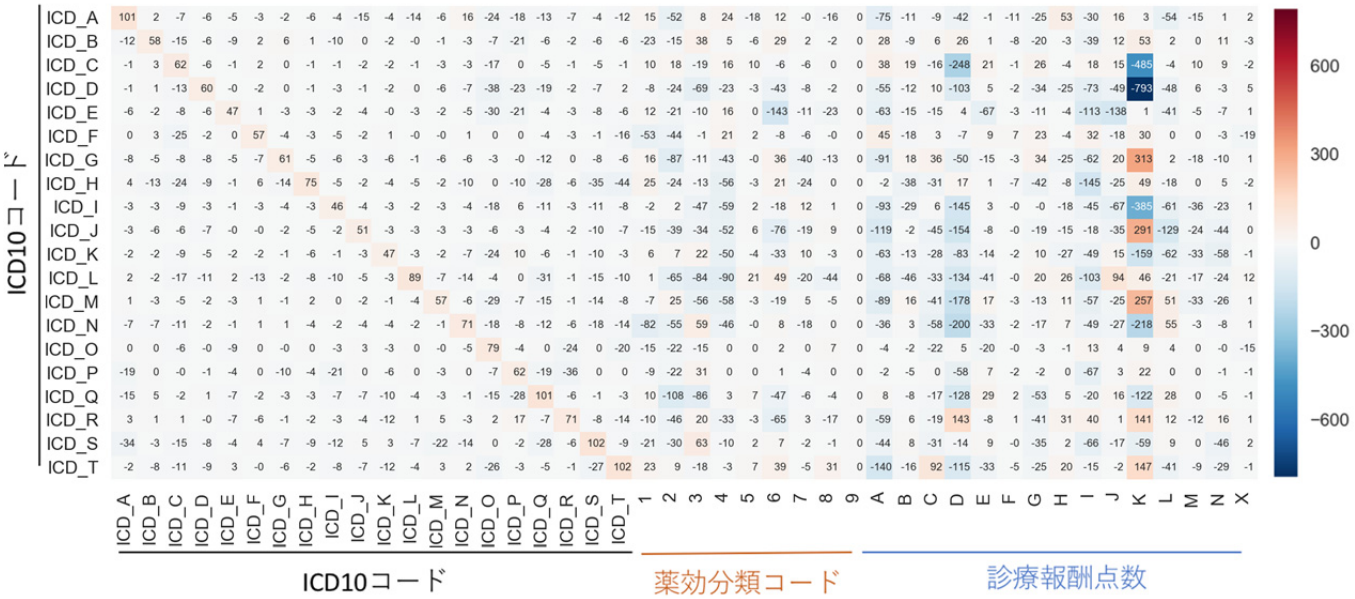

図 3　交差検定法の最適解における各カテゴリーの回帰係数の和

縦軸は目的変数、横軸は説明変数を表し、縦軸と横軸の組み合わせに対応する回帰係数の和を示した。