第 1 部

論文集

第１部 論文集
原著

レセプトデータセットを用いたがん登録のケースファインディングに有効な統計的手法の判別能の評価

小原 仁[1] 水谷 駿介[2] 小森園 康二[3]

久留米大学データセンター [1]
南風病院医事課 [2]
南風病院肝臓内科 [3]

要 旨

【目的】調査実施施設が有するレセプトデータからデータセットを新たに生成し、我々が開発したがん登録のケースファインディングに有効な統計的手法を適用させた判別能を評価する。
【方法】レセプトデータセットの対象は 2015 年 1 月から 2015 年 12 月までの間に、国立がん研究センターから提供されているがん登録の対象となるICD10コード、またはがん登録の対象となる可能性のある ICD10 コードに該当した 15,357 件とした。レセプトデータセットをもとに予測モデルと追加抽出ロジックを構築した。構築した統計的手法の判別能は、未知となる評価用データセットに本手法を適用して求めた判別結果の感度と特異度をもとに評価した。
【結果】レセプトデータセットから 11 項目の説明変数を有する予測モデルが構築された。説明変数には ICD10 の C コードの該当を示す変数のほか、病理診断や悪性腫瘍特異物質治療管理料などの算定実績を示す変数が選択された。予測モデルの ROC 曲線から求めた AUC は 0.978 であった。評価用データセットから求めた予測モデルによる判別の感度は 93.6%、特異度は 94.5%であった。また予測モデルの結果に追加抽出ロジックを加えた最終的な判別能の評価となる感度と特異度はそれぞれ、96.3%（95%CI：94.0%-97.9%）、80.2%（95%CI：79.3%-81.1%）となった。
【考察】レセプトデータセットから構築した予測モデルの判別能は、感度 93.6%、特異度 94.5%の判別成績を得た。また本予測モデルに追加抽出ロジックを加えた結果、最終的な感度は 95%を上回る成績を得た。これは施設類型や規模に関係なく、レセプトデータを有する多くの施設で効率的かつ精度の高いがん登録のケースファインディングを実現できる可能性を示唆した。
【結語】がん登録のケースファンディングに有効な統計的手法は、保険診療を行う施設で生成可能なレセプトデータセットにも有効である。

１．はじめに

2016 年 1 月から全ての病院と一部の診療所を対象に、日本でがんと診断されたすべての人のデータを、国で１つにまとめて集計・分析・管理する全国がん登録事業が開始された [1)]。全国がん登録事業では医療機関等からの症例の届出を義務化することで、高い悉皆性を期待できる制度的な裏付けを実現した [2)-3)]。こうした全国がん登録事業による情報の蓄積

が進むことによって、今後のがん対策の立案や評価への活用が期待されている[4)-7)]。

一方、がんの診断や治療を担う医療機関においては、がん病名の付与されていない症例やがん病名が付与されていても登録対象とならない症例をひとつひとつ選別するケースファインディングという作業を行う必要がある。多くの場合、日々の外来や入院患者のなかから、がん登録症例を効率的に勘良く見つけ出すことは容易でない[8)]。また、がん登録症例を行う実務者の経験や能力にも依存してしまうため、精度の高いケースファインディングをいかに効率よく実施できるかが実務において重要となる。

筆者らは、候補病名が付与された病名開始日前後の診療行為実績からがん登録症例を判別する予測モデルとこの予測モデルを補完する抽出ロジックを組み合わせたケースファンディング法を開発した[9)]。評価用データを用いた本手法の判別能は感度 99.3%、特異度 93.6%という成績を得た。

しかしながら、このケースファインディング法を用いた判別能は、調査実施施設の院内システムから取得した悪性腫瘍の病理診断結果や腫瘍マーカーといった検査結果の診療情報をもとに構築したデータセットから得られた成績である。これは当該施設独自のデータセットを用いた判別能の評価であるため、他施設においても同等の再現性を有しているかは明らかでない。

そこで本論文では、他施設における本手法の検証と適用可能性を高めることを目的に、多くの医療施設で標準的に生成・蓄積されているレセプトデータからデータセットを構築し、本データセットにがん登録のケースファインディングに有効な統計的手法を適用させた判別能の評価を行った。

２．方法

(1)　データソース

本研究のデータセットは、鹿児島県にある南風病院で記録された診療報酬請求明細情報（以下、レセプトデータ）とがん登録情報をもとに作成した。

南風病院は一般病床 338 床を有し、鹿児島医療圏を中心とする地域の患者に専門的かつ高度な医療を提供している施設である。また、鹿児島県のがん診療指定病院でもある。

(2) 対象

対象は 2015 年 1 月 1 日から 2015 年 12 月 31 日の間に、国立がん研究センターから提供されているがん登録の対象となる ICD10 コード、またはがん登録の対象となる可能性のある ICD10 コードに該当した症例とした 10)。すなわち、ICD10 コードの C コードに分類される新生物や D コードの一部に含まれる上皮内癌のほか、消化器系疾患の K コード、症状、徴候・所見に関する分類である R コードなどの一部の ICD 病名が与えられた症例も対象に含まれている。

なお、ICD10 コードは当該医療機関の電子カルテに登録されたテキスト病名をもとに診療報酬請求業務を行う担当者が付与し、主治医または担当医の確認が得られたものを用いている。

(3)　解析用レセプトデータセットの作成

レセプトデータをもとにした解析用のレセプトデータセットの項目は、当該症例の診断根拠や初回の治療情報に係る診療提供実績を反映する可能性が高いと考えられた項目を選定した。また粒度の細かいレセプト電算コード単位ではなく、解釈番号を用いた理由は診療報酬改定に伴う定義変更やマスタ保守の煩雑な作業を簡便に行うためである。以下にデータセット項目の明細を示す。①レセプト情報：患者 ID、解釈番号 A310 緩和ケア病棟入院料、同 B0013 悪性腫瘍特異物質治療管理料、同 B00112 がん性疼痛緩和指導管理料、同 D215 超音波検査、同 D325 肝臓カテーテル法、同 E000～E004 エックス線診断料、同 E100～E102 核医学診断料、同 E200～E203 コンピューター断層撮影診断料、同 K000～919 手

術、同 M000～M005 放射線治療管理・実施料、同 N000~N007 病理診断に係る各項目の該当情報と実施日、②がん登録情報：2014 年登録対象の患者 ID と診断日、2015 年登録対象の患者 ID と診断日、に加え、患者 ID、年齢、性別、ICD10 コード、病名開始日の患者属性情報を用いた。各レセプト項目の該当の有無については、対象病名の診断日の前後 120 日以内に 1 回以上の算定実績があったものを該当有り、算定実績が確認されなかったものについては該当無しの 2 値データとして、解析用データセットに反映させた。

レセプトデータセットは、各データファイルの患者 ID を突合させ、がん登録対象の有無情報と候補病名に関連する各診療行為の該当情報を有するレコード構造とした。

またレセプトデータセットは、予測モデルの構築に用いる訓練用データと作成した予測モデルの評価に用いる評価用データに無作為に 2 分割した。

(4)　がん登録症例を判別するための予測モデルの構築

予測モデルはロジスティック回帰分析をもとに構築した。目的変数を 2015 年がん登録の有無、説明変数は年齢、性別、ICD10（C コード）該当の有無、前年度がん登録対象の有無のほか、悪性腫瘍特異物質治療管理料、がん性疼痛緩和指導管理料、超音波検査、肝臓カテーテル法、エックス線診断料、核医学診断料、コンピューター断層撮影診断料、手術、緩和ケア病棟入院料、放射線治療管理・実施料、病理診断に関するレセプトの算定実績の有無を示す項目のなかから、ステップワイズ法による変数選択を適用させて選定した。なお、変数選択に用いた統計的有意水準は 20%未満とした。

構築したモデルから求めた各症例における予測確率のカットオフ値は、予測確率を 0.01 刻みごとに区分した分割表において、感度と特異度の和が最大となる値とした。

(5)　追加抽出ロジックの定義

予測モデルに用いたカットオフ値は感度だけでなく、特異度も同等に評価した水準となっている。そのため、特異度が考慮されたもとで、がん登録症例を拾い漏らしてしまう可能性がある。そこで、予測モデルから除外された症例のなかから、解釈番号 B0013 悪性腫瘍特異物質治療管理料、B00122 がん性疼痛緩和指導管理料、A310 緩和ケア病棟入院料、または ICD10 の C コードのいずれかに該当する症例を追加抽出するロジックを設定した。

(6)　予測確率モデルと追加抽出ロジックによる判別能の評価

判別能の評価は、評価用データに予測モデルを適用した判別結果とこの予測式に追加ロジックによる追加抽出を加えた感度と特異度を用いた。

なお、解析用のレセプトデータセットの生成とすべての統計解析は Stata/IC 14.0 を使用した。

３．結果

(1) 大腸がん集団検診事業の精度管理

がん登録のスクリーニング対象に該当した症例 15,357 件（100%）を抽出した。そのうち 2015 年がん登録対象は 819 件（5.3%）、非対象は 14,538 件（94.7%）であった。スクリーニング対象の全体に占めるがん登録症例割合は 5.3%であった。スクリーニング対象の内訳を表 1 に示す。

スクリーニング対象の内訳として、平均年齢はがん登録対象群 71.5 歳（標準偏差 11.9）、非対象群 65.6 歳（標準偏差 14.8）、ICD10 コードの C コードに該当した症例のうち、がん登録対象は 752 件（23.1%）であった。一方、C コード以外のコードに該当した症例に含まれていたがん登録対象は 67 件（0.6%）であった。

表 1　がん登録のスクリーニング対象の内訳

	全体	がん登録対象	がん登録非対象
患者数（割合）	15,357（100%）	819（5.3%）	14,538（94.7%）
ICD-10：Cコードあり（割合）	3,259（100%）	752（23.1%）	2,507（76.9%）
ICD-10：Cコードなし（割合）	12,098（100%）	67（0.6%）	12,031（99.4%）
平均年齢（標準偏差）	65.9歳（14.7）	71.5歳（11.9）	65.6歳（14.8）
性別：男性（割合）	7,473（48.7%）	491（60.0%）	6,982（48.0%）

スクリーニング対象に該当した 15,357 件のうち、訓練用データに無作為に割り付けられた 7,678 件をもとに、ステップワイズ法による変数選択を適用させたロジスティック回帰分析を実施した結果、11 項目の説明変数が選択された。

11 項目の説明変数とその β 係数は、年齢 0.027、男性-0.425、ICD10（C コード）4.143、前年度がん登録対象-4.076、病理診断 2.499、肝臓カテーテル法 0.313、手術 1.216、悪性腫瘍特異物質治療管理料 2.224、エックス線診断料 1.719、緩和ケア病棟入院料 1.101、コンピュータ断層撮影診断料-0.981 であった。ロジスティック回帰分析の結果を表 2 に示す。図 1 には構築したモデルによる ROC 曲線とその ROC 曲線から求めた AUC 値 0.978 を示している。

表 2　ロジスティック回帰分析の結果

説明変数		β	95%信頼区間			p 値
年齢	X_1	0.027	0.012	-	0.042	<0.001
性別：男性	X_2	-0.425	-0.800	-	-0.050	0.026
ICD10（Cコード）：あり	X_3	4.143	3.697	-	4.589	<0.001
前年度がん登録対象：あり	X_4	-4.076	-5.191	-	-2.961	<0.001
病理診断：あり	X_5	2.499	2.065	-	2.932	<0.001
肝臓カテーテル法：あり	X_6	0.313	-0.127	-	0.754	0.163
手術：あり	X_7	1.216	0.791	-	1.641	<0.001
悪性腫瘍特異物質治療管理料：あり	X_8	2.224	1.715	-	2.734	<0.001
エックス線診断料：あり	X_9	1.719	1.129	-	2.308	<0.001
緩和ケア病棟入院料：あり	X_{10}	1.101	0.170	-	2.031	0.020
コンピュータ断層撮影診断料：あり	X_{11}	-0.981	-1.402	-	-0.561	<0.001
切片	-	-9.466	-10.678	-	-8.255	<0.001

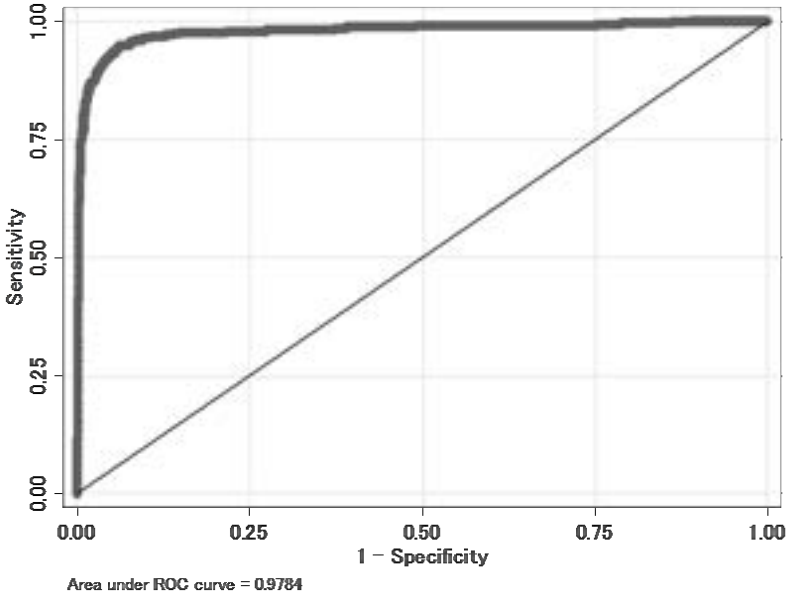

図 1　作成した予測式の ROC 曲線

構築モデルよりがん登録症例の予測確率は、次の式で推定される。

Log{ P(がん登録) / 1-P（がん登録) } = -9.466+0.027 × X1-0.425 × X2+4.143 × X3-4.076 × X4+2.499 × X5+0.313 × X6+1.216 × X7+2.224 × X8+1.719 × X9+1.101×X10-0.981×X11.

ただし、X1 = { 年齢 }, X2 = {1: 男性, 0: 女性 }, X3 = {1: ICD10（C コード）該当, 0: 非該当 }, X4 = {1: 前年度がん登録 該当, 0: 非該当}, X5 = {1: 病理診断 あり, 0: なし}, X6 = {1: 肝臓カテーテル法 あり, 0: なし, X7 = {1: 手術 あり, 0: なし}, X8 = {1: 悪性腫瘍特異物質治療管理料 あり, 0: なし}, X9 = {1: エックス線診断料 あり, 0: なし}, X10 = {1: 緩和ケア病棟入院料 あり, 0: なし}, X11 = {1: コンピューター断層撮影診断料 あり, 0: なし}.

カットオフポイントは 0.05 であった。判別式はこの予測モデルを用いて、以下で与えられる。

P（がん登録対象）＞　0.05
⇒　がん登録候補と判定
P（がん登録対象）≦　0.05
⇒　がん登録非候補と判定

最後に、作成した予測モデルによる判別結果を表 3 に、このモデルの判別結果に追加抽出ロジックによる抽出を加えた判別結果を表 4 に与える。表 3 より、評価用データを用いた予測モデルの感度は 93.6%（95%信頼区間：90.8%-95.8%）、特異度 94.5%（95%信頼区間：94.0%-95.0%）であった。また追加ロジックによる抽出を加えた最終的な感度と特異度はそれぞれ 96.3%（95%信頼区間：94.0%-97.9%）、80.2%（95%信頼区間：79.3%-81.1%）となった。

表３　予測式の判別能

判別結果	がん登録 非対象	がん登録 対象	計
非登録候補	6,873	26	6,899
n, (%)	(94.5%)	(6.4%)	(89.8%)
登録候補	399	381	780
n, (%)	(5.5%)	(93.6%)	(10.2%)
計	7,272	407	7,679
n, (%)	(100%)	(100%)	(100%)

感度　93.6%（95%信頼区間：90.8%-95.8%）
特異度　94.5%（95%信頼区間：94.0%-95.0%）
陽性的中率　48.8%（95%信頼区間：45.3%-52.4%）
陰性的中率　99.6%（95%信頼区間：99.4%-99.8%）

表４　予測式に追加抽出ロジックを加えた判別能

判別結果	がん登録 非対象	がん登録 対象	計
非登録候補	5,834	15	5,849
n, (%)	(80.2%)	(3.7%)	(76.2%)
登録候補	1,438	392	1,830
n, (%)	(19.8%)	(96.3%)	(23.8%)
計	7,272	407	7,679
n, (%)	(100%)	(100%)	(100%)

感度　96.3%（95%信頼区間：94.0%-97.9%）
特異度　80.2%（95%信頼区間：79.3%-81.1%）
陽性的中率　21.4%（95%信頼区間：19.6%-23.4%）
陰性的中率　99.7%（95%信頼区間：99.6%-99.9%）

４．考察

本論文では、多くの医療機関において標準的に生成されているレセプトから作成したデータセットに、がん登録のケースファインディングに有効な統計的方法を適用した判別能の評価を行った。

訓練用データから作成した予測モデルの説明変数は 11 変数であった。ICD10 の C コードの付与を示す変数のほか、病理診断や悪性腫瘍特異物質治療管理料の実施に該当する変数が選択された。

年齢と性別が選択されたことについては、当該医療機関におけるがん登録の施設特性が反映された可能性がある。当該医療機関は消化器系のがん登録が多く、婦人科系や小児、希少がんといったがん種は

ほとんどない。予測モデルの説明変数については、各施設のがん登録の状況によって異なることが想定できる。例えば婦人科系のがん登録が多い施設では男性を示す性別変数の回帰係数は負の値をもつ説明変数として、選択される可能性がある。

年齢と性別以外では、ICD10 病名の C コード、および病理診断や腫瘍マーカー検査といった病名と診療行為の提供実績は、がん登録症例の判別確率を高める正の値の回帰係数であった。一方で、前年度のがん登録症例の該当を示す変数の回帰係数は負の値であったことから、こうした症例を確率的にがん登録症例から除外する本予測モデルは妥当なモデルと解釈できる。また、このモデルから求めた ROC 曲線における AUC の値は 0.978 であった。これは先行研究で示された病理診断や生化学検査の結果から構築したデータセットを用いた予測モデルの AUC の値 0.983 と同等であった。[9]

追加抽出ロジックでは、予測モデルから除外される可能性があるがん登録候補症例の取りこぼしを軽減することを期待している。追加抽出ロジックの設定は、予測モデルから除外された症例で ICD10 の C コードが付与されている症例のほか、がん性疼痛緩和指導管理料、緩和ケア病棟入院料、悪性腫瘍特異的治療管理料のいずれかに該当する症例を再抽出している。これはつまり、病名付与後に積極的な治療を望まない症例や診療行為の乏しい症例の抽出を補完する役割となる。予測モデルと追加抽出モデルを組み合わせて実施することで、積極的な治療を行った症例だけでなく、そうでない症例の抽出を補完するケースファインディング法となっている。

予測モデルを適用した後に追加抽出ロジックを用いている理由として、がん登録のケースファインディングに係る作業の多くは、がん登録症例を見つける以上に、そうでない症例を除外するところに大きな負担を要している。そこで、まず予測モデルでがん登録の対象でない症例を大きく除外し、その後に追加抽出ロジックでがん登録症例の漏れがないか慎重に探索するという視点から本ケースファインディング法の開発を行った。つまり、一次スクリーニングの位置づけとなる予測モデルでは特異度を優先し、追加抽出ロジックによる二次スクリーニングでは、がん登録症例を拾い漏らさぬよう感度を重視するという、役割の異なる 2 段階の抽出法で効率化と精度向上の実現を期待している。

評価用データを用いた判別能では、予測モデルの感度は 93.6%、特異度は 94.5%であった。これは筆者らが病理診断結果や腫瘍マーカー検査値などから構築したデータセットの成績である感度 91.7%、特異度 97.0%と同等の判別精度であった。検査や診断結果の情報を持たないレセプトデータであっても、付与された病名とその前後の診療行為の提供実績からがん登録症例の抽出が可能であった。これはがん診療のガイドラインやクリニカルパスといった標準化された複数の診療提供実績を確率的に捉えられたことを示している。

また追加抽出ロジックによる抽出を加えた最終的な判別能は、感度 96.3%、特異度 80.2%であった。感度は 95%を上回る水準を示した一方で、特異度は約 80%の水準となった。これは病理診断や検査結果の情報を持たないレセプトデータでは、陰性症例を除外しにくいためである。

最終的な判別能の感度は 96.3%を示した一方で、がん登録症例全体の 3.7%にあたる 15 例が除外されている。これは、そもそもの病名登録に不備があった症例と年度をまたいで診断・治療をされている症例であった。病名登録の不備は、病名登録の精度管理に依存してしまうことから、日常での精度管理の維持が重要になる。年度をまたぐ症例については、本ケースファインディング法を継続的に適用することで次年度に抽出することが可能であったが、年末や年始に係る候補症例についてはがん登録を行う実務者が意識的に注意することで登録のタイムラグを回避することができる。

レセプトデータの特徴は、電子カルテやオーダリングなどのシステム環境が整備されていない施設であっても標準的に生成・蓄積されている点である。つまり、院内の情報化が進んでいる DPC 適用病院などの一部の病院だけでなく、あらゆる施設類型や規模でも生成可能なデータセットである。特に、全国がん登録への参加を期待されている在宅医療やがん患者の看取りを行っているクリニックなどにも適用可能である。こうした施設の多くは、がん登録専門のスタッフや経験豊富な知識を有しているスタッフが十分とはいえない。しかし、本手法によるケースファインディングはデータをもとに自動で判別するため、手作業で行う担当者の経験や知識不足に依存しないという利点がある。また、中小規模の施設やクリニックだけでなく、数千件、数万件といったケースファインディングを手作業で行っている大規模な施設においても、機械的に処理できることで作業効率の改善を期待できる。

ただし、今回得られた判別能の成果は、単一施設のレセプトデータセットを用いた評価用データによるものである。これから多施設による判別能の検証作業を行い、一般化可能性を検討していかなければならない。また、本予測モデルを使用するだけでは悉皆性が求められるがん登録の実現は困難である。予測モデルや追加ロジックによる質的な抽出が困難な症例については、熟練した院内がん登録実務者が補うことで、予測モデルによる量的な効率化と実務者による質的な精度確保をしていかなければならない。こうした補完関係のある本ケースファインディング法の運用が、実用性に最も適すると考えられる。

最後に、本研究で示したレセプトデータセットにがん登録のケースファインディングに有効な統計的手法を適用することで、がん登録業務に携わる関係者の量的な作業の効率化を期待できる。ケースファインディンに係る事務作業を効率化にするよって生じた時間を、質的な精度管理や追跡調査によるフォローアップといった新たな付加価値を生み出す業務に転換可能である。こうして全国から集められた質の高いがん登録情報が、今後さらに網羅的に偏りなく収集されていくことで、がん登録情報を活用した妥当性の高いがん対策がより推進されることを期待したい。

文献

1) 国立がん研究センターがん情報サービス．がん登録・統計．アクセス 2018 年 5 月 31 日．https://ganjoho.jp/reg_stat/can_reg/national/public/about.html.
2) 西野善一．わが国のがん登録の法制化．癌と化療 2015; 42(4);389-393.
3) 松田智大．全国がん登録の開始とがん登録情報利用の促進．日保医会誌 2015; 113(2): 71-83.
4) 柴田亜希子．がん登録等の推進に関する法律とがん登録．Surg Fronti 2014; 21(4): 367-371.
5) 柴田亜希子．がん登録等の推進に関する法律に基づく全国がん登録データベース．外科 2016; 78(5): 469-474.
6) 伊藤ゆり，中山富雄，宮代勲 他．大阪府がん対策推進計画の立案・評価における各種がん統計資料の活用．JACR Monograph 2013; 19: 19-28.
7) 松田智大．全国がん登録の開始に向けて．癌の臨床 2014; 60(5): 567-574.
8) 平林由香．院内がん登録における腫瘍見つけ出し.アクセス 2018 年 5 月 31 日．https://ganjoho.jp/data/hospital/training_seminar/cancer_registration/odjrh3000000hv0y-att/060829casefinder.pdf.
9) 小原仁，増田由佳，水谷駿介　他．がん登録のケースファインディングに有効な統計的手法の開発．日本診療情報管理学会誌 2018; 29(4): 31-35.
10) 国立がん研究センターがん情報センター．関連 ICD コード．アクセス 2018 年 5 月 31 日．http://ncc.ctr-info.com/new_toroku/?action=common_download_main&upload_id=1282.