一般口演
# 一般口演7
# 医療データ分析３（レセプトデータ・治験）
2018年11月23日(金) 14:20 〜 15:50 H会場 (福岡サンパレスHパレスホール)

[2-H-2-5] レセプト情報・特定健診等情報データベース（ NDB）に対する死亡決定ロジックの手法開発ー R言語による決定木分析を用いてー

○久保 慎一郎[1,2], 野田 龍也[1], 西岡 祐一[1,3], 明神 大也[1,4], 降簱 志おり[5], 東野 恒之[6], 瀬楽 丈夫[5], 今村 知明[1] （1.奈良県立医科大学　公衆衛生学講座, 2.奈良県立医科大学附属病院　看護部, 3.奈良県立医科大学　糖尿病学講座, 4.奈良県立医科大学　病理診断学講座, 5.（株）三菱総合研究所　ヘルスケア・ウェルネス事業本部, 6.（株）三菱総合研究所　経営イノベーション本部）

【目的】レセプト情報・特定健診等情報データベース(以下， NDB)とは、日本の保険診療の悉皆データである。NDBには、死亡した患者のレセプトの「転帰区分」に死亡フラグが付与されるが、医療機関の付与忘れや付与間違い等によってすべての患者が正確とはいえなかった。診療行為や薬剤等から死亡を推定することで死亡転帰の有効性を高めることを目的とした。【方法】4年分の奈良県 KDBレセプトと3年分の NDBレセプトを用いた。はじめに、 KDBの保険者マスターに記載されている死亡転帰を教師データとし、 KDBの死亡転帰の正解率や必要となる決定木の診療行為を洗い出した。分析には R言語による決定木分析を用いた。その仕組みを用いて NDBでも検証し、その有効性を検証した。死亡数を人口統計と比較した。【結果】 KDBでの死亡フラグの陽性的中率は96.2%であった。決定木として用いられる診療行為は看取りに関連するコード、呼吸心拍監視に関連するコード、酸素が多かった。決定木より枝分かれした患者を基に的中率が83%以上確保できている場合を死亡決定とし、死亡フラグを付与した。感度・特異度を集計すると感度が92.9%、特異度が99.7%となった。これらを NDBのデータを用いて同様のロジックを抽出したが、ほぼ同様の傾向を示した。【結論】 KDBを教師データとして NDBの死亡フラグを策定した。現状ではそのすべての死亡を完全に追うことはできないが、死亡したアウトカムを正確に付与できれば NDBで日本のコホート研究が大きく前進する。

レセプト情報・特定健診等情報データベース（NDB）に対する死亡決定ロジックの手法開発

－R 言語による決定木分析を用いて－

久保慎一郎[*1*2*3]、野田龍也[*1]、西岡祐一[*1*4]、明神大也[*1*5]、降簱志おり[*6]、東野恒之[*7]、瀬楽丈夫[*6]、今村知明[*1]
*1 奈良県立医科大学 公衆衛生学講座、*2 奈良県立医科大学附属病院 看護部、
*3 奈良県立医科大学附属病院 医療情報部、*4 奈良県立医科大学 糖尿病学講座、
*5 奈良県立医科大学 病理診断学講座、*6（株）三菱総合研究所 ヘルスケア・ウェルネス事業本部、
*7（株）三菱総合研究所 経営イノベーション本部

Development of a Logic-based Method for Determining Mortality for the National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB)

- Using Decision-Tree Analysis in R -

Kubo Shinichiro [*1*2*3], Noda Tatsuya [*1], Nishioka Yuichi [*1*4], Myojin Tomoya [*1*5], Furihata Shiori [*6], Higashino Tsuneyuki [*7], Seraku Takeo[*6], Imamura Tomoaki [*1]
*1 Department of Public Heath, Health Management and Policy, Nara Medical University,
*2 Department of Nursing, Nara Medical University,
*3 Department of Medical Informatics, Nara Medical University,
*4 Department of Diabetology, Nara Medical University,
*5 Department of Diagnostic Pathology, Nara Medical University,
*6 Haelthcare and Energy Division, Mitsubishi Research Institute, Inc. ,
*7 ICT Innovation Division, Mitsubishi Research Institute, Inc.

The National Database of Health Insurance Claims and Specific Health Checkups of Japan (hereinafter, the NDB) is a complete inventory of data relating to healthcare services provided by health insurance in Japan. While patient deaths are described in the NDB, these lack accuracy. To address unspecified deaths, a mortality estimation logic was created using machine learning (decision-tree analysis) from data points such as courses of medical treatment and prescription medication, with reference to entries for Nara Prefecture in the National Health Insurance Database of Japan (the “Kokuho Database”; a database of healthcare services provided by health insurance companies in Japan similar to the NDB), which have few omissions. The positive prediction rate of this mortality logic was 96.2%. Among the medical treatments/services most likely to predict mortality were “additional caregiving,” “electrocardiogram monitor and pulse,” and “oxygen inhalation.” Sensitivity was 92.9% and specificity was 99.7%. Application of this logic to three years of data in the NDB confirmed the same trends. Although mortality in the NDB cannot currently be specified completely, accurate specification of death will contribute greatly to cohort studies making use of the NDB.

Keywords: health insurance claims in Japan, patient identification, personal identifiers, rezept, KDB

1. はじめに

レセプト情報・特定健診等情報データベース(以下, NDB)とは、病院や診療所から国に送信される電子レセプトデータと特定健診等のデータを個人が特定されないように一部の情報を匿名化・削除した上で、格納・構築されているデータベースである。国民皆保険制度をとる日本における保険診療の悉皆データであり、2009 年 4 月～2016 年 12 月診療分で約 128 億 8,400 万件(2017 年 3 月末時点)[1]のレセプトデータが蓄積されるなど、世界最大級の健康関連データベースといえる。これを有効に活用することで日本全体の臨床研究(コホート研究等)や政策研究が強力に推進できると期待されている。

ただし、NDB にはさまざまな問題点や障壁が存在する。まず、政策用途を目的としたデータであるため、利用するためには申請が必要である。申請し有識者の審査を受けてデータが切り出されるには早くても半年程度の時間がかかる。そのうえ、著者らが行った先行研究では、NDB の巨大なサイズによる処理の限界と、診療報酬請求のために設定されているレセプトの構造がそのままでは研究目的での利用に適さない形式となっていることから研究者が個人で分析をするのが困難である問題点が挙げられた [2]。加えて、最も致命的なものとして NDB に付与されている 2 種類の ID が経過とともに変化しやすいという問題が挙げられる。なお、NDB には 2018 年 9 月現在で一生涯不変の個人 ID は含まれていない。代わりに、個人紐付け用の匿名変数として、「ID1」と「ID2」が用意されている。ID1 は保険者番号、被保険者証等記号・番号、生年月日、性別から個人情報保護のためにハッシュ関数と呼ばれる関数を用いて変換された英数字列であり、ID2 は氏名、生年月日、性別から同様に変換された英数字列である。ところが、同一患者でも、就職・転職等で保険者は変化し、医療機関での表記ゆれ（例：渡辺と渡邉）や結婚・離婚等のライフイベントで氏名表記は変化するため、ID1、ID2 ともに容易に変わり得ることが分かった [3]。筆者らの先行研究により、ID1 と ID2 を組

み合わせて一つの個人 ID を作成することに成功した[4) 5)]。これらは NDB をコホート化する上での基盤技術となっており、さまざまな研究班でこの名寄せロジックが用いられ始めている。

患者数を集計することは可能となったが、NDB にはまだ課題が存在する。それは、アウトカム指標が乏しいことである。死亡や罹患のようにアウトカムにできる指標が乏しいため、治療の成果などが明らかにできない。

「死亡」に関してはレセプトの転帰区分の「死亡」を用いることで死亡を抽出することができる。ただし、これには①医療機関が正しく入力していること、②医療機関外の死亡が補足できることの二つが重要となる。ただし、どちらも精度が高いとは必ずしも言えない状況である。特に、医療機関の付与忘れや付与間違い等は一定発生していると推察される。

本研究では、NDB の転帰死亡の精度を向上させ、診療行為や薬剤等から死亡を推定することで死亡転帰の有効性を高めることを目的とした。また、実際の死亡データと比較してその有効性を検証した。

2. 方法

本研究は医療計画策定に係る評価指標作成の一環として（5 年生存率など死亡に関連する指標のため）行った。2013 年 4 月～2017 年 3 月までの 4 年分の奈良県 KDB レセプトと、2013 年 4 月～2016 年 3 月までの 3 年分の NDB レセプトを用いた。はじめに、KDB には被保険者マスターに記載されている死亡転帰(保険者の持っている死亡情報、レセプトが発生していない患者を含む)とレセプトの死亡転帰(レセプトが発生している患者の死亡。以下医療内死亡)の2種類の死亡転帰が存在する。前者の被保険者マスターに基づいた死亡は死亡届が出された情報に基づいて作成されているため正確である。この被保険者マスター死亡をアウトカムとし、診療行為や薬剤などのレセプトの内容について機械学習を行い、レセプト死亡転帰の精度が高めることを目指した。機械学習は R 言語による決定木分析(R-part)を用いた。

決定木分析を行うに当たっては、条件設定や変数の選択が重要となる。まずは様々な情報を組み合わせて、どのような決定木を採用することが死亡転帰の精度が高まるか繰り返し決定木を作成した。作成した決定木を基に NDB で死亡決定が行えるよう死亡推定ロジックを作成した。その仕組みを用いて NDB に適用させた後、レセプトの死亡転帰に比べてどの程度付与できたか検証した。また、人口統計の死亡数と比較しその有効性を検証した。

3. 結果

3.1 決定木の条件と説明変数の絞り込み

決定木の作成においては説明変数に何の変数をどのような分類で入れるかによって決定木の出方が異なる。そのため、数々の変数を条件に含めたり外したりしながら、適切な条件を決定する必要がある。

KDB のレセプトの内、「全入院レセプトから 1/8 サイズになるように無作為抽出して決定木分析を行った場合」と、「病床機能別（DPC／一般／精神／療養／障碍者施設）でそれぞれ分割し、決定木分析を行った場合」の結果は表 1 通りとなった。なお、説明変数には、退院日または死亡日に算定された診療行為および医薬品を使用した。診療行為は小分類で集約した。ただし基本診療料、医学管理料、在宅、精神を除外した。医薬品は薬価基準コードの先頭 4 桁で集約した。

表 1　各病床機能別の決定木分析結果

	感度	特異度	偽陽性率	偽陰性率	AUC
全入院から無作為抽出	78.0%	95.5%	4.5%	22.0%	0.93
DPC	74.1%	96.5%	3.5%	25.9%	0.90
一般	83.2%	96.2%	3.8%	16.8%	0.93
精神	74.8%	98.2%	1.8%	25.2%	0.92
療養	59.3%	95.4%	4.6%	40.7%	0.85
障碍者施設	76.9%	92.3%	7.7%	23.1%	0.87

特異度は全入院データと比較して変わりはなかったが、感度は一般病床を除き全入院データの方が高かった。AUC に関しても全入院データの方が高く、レセプトの決定木を作成する上では病床を分類して検証しない方がいいことが分かった。別途全データで決定木分析を行った際に最初に分岐する項目が「食事加算」であり、入院と外来の2パターンに分かれることが分かった。

入院と外来で改めて決定木分析を行ったところ、表 2 のような結果となった。なお、説明変数には、退院日または死亡日に算定された診療行為および医薬品を使用した。診療行為は小分類で集約した。ただし医学管理料を除外した。医薬品は薬価基準コードの先頭 4 桁で集約した。傷病名も使用したが、処理の関係上、入院はICD10(4桁)、外来はICD10(3桁)で検証した。

表 2　入院・外来の決定木分析結果

	感度	特異度	偽陽性率	偽陰性率	AUC
入院	85.0%	97.7%	2.3%	15.0%	0.94
外来	43.9%	99.9%	0.1%	56.1%	0.79

入院は概ね精度が保てているものの、外来になると著しく精度が低下した。外来において決定木にかかる診療行為を確認すると、「往診料」が最上位でその後「在宅患者訪問指導料」と「心停止」に分岐しており、往診と往診以外(通院等)で分かれていることが分かった(図 1)。

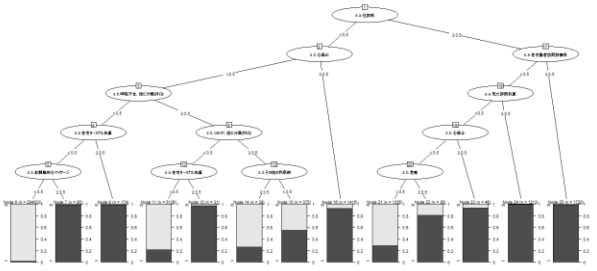

図 1　外来の決定木分析結果

決定木分析では「生存」としたにも関わらず実際に「死亡」と判別した事例を確認したところ、医療機関外死亡(検死)により死亡したものと考えられた。そのため、検死を除外する条件を設定した。具体的には、KDB の被保険者マスターの中に、資格喪失年月日が記載されている。マスターで死亡が付与されている患者の中で資格喪失年月日とレセプトの最終発生年月日の差をとり、2 日以上空いている場合は決定木の分析対象から除外する対応を行った。

また、年齢に関して、39 歳以下のレセプトが極端に少なく、39 歳以下で決定木分析を行うと、入院の生存者に出産が出現し、若年者特有の疾患に偏る(外来の生存者に風邪や花粉症の患者が多く出現する)ため正しく結果が出なかった。今回は 40 歳以上を分析の対象とした。

さらに、外来のロジックでは「往診料」の有無と、「非開胸的

心マッサージ」の有無が上位に分かれる結果となった。そのため、「往診」と「往診以外」でデータを分けることでその精度を向上させることとした。

加えて、KDB には先述の医療内死亡(医療機関が報告した死亡転帰)が存在しており、その情報が正しいか検証した。マスター死亡の患者の内、死亡転帰が入力されている患者を比較すると、入院が 9 割以上の一致率であった。一方、外来は 7 割以上であるが、先述の医療機関外死亡を除外すると 9 割以上となり、比較的精度高く死亡転帰が付与されていることが分かった。そのため、死亡転帰は死亡決定のための重要なフラグとなりうるが、1割程度の過誤を検出するために、死亡転帰が入っている患者に限定し、決定木分析を行った。

3.2 入院の決定木と感度特異度

上記の理由により、入院は二種類の決定木を作成し、転帰死亡の有無にデータを分けて分析した。対象データは医科入院のすべてで、決定木を作成する対象患者は40歳以上のデータを用いた。マスターの死亡転帰より、死亡者数と生存者は表 3 のとおりであった。これらをそれぞれ学習用データと検証用データに 2:1 で分類し、学習用データで決定木を作成したものを検証用データで抽出すると、マスター死亡がどのように振り分けられるかを検証した。説明変数には、退院日または死亡日に算定された診療行為および医薬品、傷病名を説明変数に使用した。診療行為は小分類で集約し、医学管理料および基本診療料を除外した。医薬品は薬価基準コードの先頭 4 桁で集約した。傷病名は ICD10 の 4 桁で集約した。転帰が生存および死亡でそれぞれ、表 4、表 5 とおりの推定結果となった。また、感度・特異度・AUC を表 6 に示した。

表 3　入院の対象患者数

	転帰区分生存	転帰区分死亡
死亡者	35,211	32,202
生存者	174,653	3,892

表 4　入院の推定結果

		正解		合計
		死亡	生存	
推定	死亡	14,958	1,883	16,841
	生存	2,647	80,612	83,259
合計		17,605	82,495	100,100

表 5　入院(転帰区分が死亡)の推定結果

		正解		合計
		死亡	生存	
推定	死亡	15,832	893	16,725
	生存	269	1,053	1,322
合計		16,101	1,946	18,047

表 6　入院全件の感度・特異度

	入院全件	転記区分死亡
感度	85.0%	98.3%
特異度	97.7%	54.1%
偽陽性率	2.3%	45.9%
偽陰性率	15.0%	1.7%
陽性的中率	88.8%	94.7%
陰性的中率	96.8%	79.7%
AUC	0.942	0.877

また、決定木に関しては入院全件と転帰が生存および死亡でそれぞれ、図 2、図 3 のとおりとなった。図 2 の転帰区分が生存であるものに関する変数としては、「食事の有無」や「大腸検査(内視鏡)」、「酸素吸入」、「人工呼吸」が頻出しており、例えば、「入院最終日に食事を摂取しなくて内視鏡検査をしておらず酸素吸入をしている人」は死亡に振られている、など概ね臨床の体感に合う結果となった。図 3 の転帰区分に死亡があるものに関しては、「食事の有無」や「療養病床かどうか」、「酸素吸入」、「人工呼吸」が上位にあり、説明変数は似ているものの、ツリーの構造としてはシンプルな結果となった。

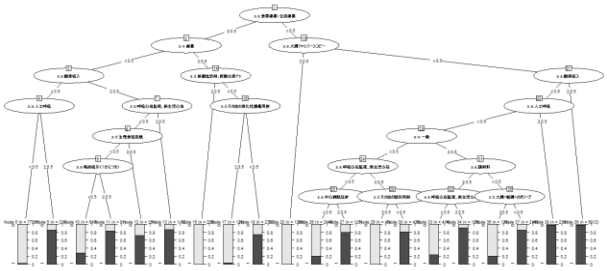

図 2　入院(転帰区分が生存)の決定木分析結果

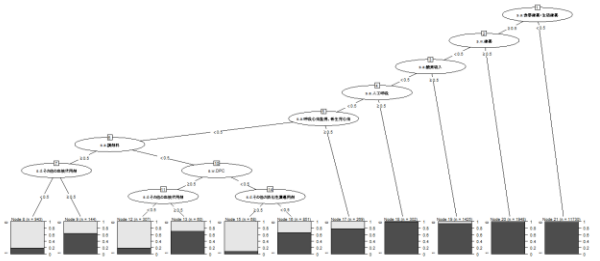

図 3　入院(転帰区分が死亡)の決定木分析結果

3.3 外来の決定木と感度特異度

外来においても外来全件の決定木と転帰区分が死亡の場合の二種類の決定木とした。さらに決定木を作成していく中で「往診」に関する診療行為と「往診以外」に関する診療行為で大きく分かれることがわかったため、これらも二分することにした。データの分割に当たっては、NDB との互換が保持できるようにレセプト最終日に「往診料」が算定されている患者を往診とした。これら、合計 4 種類の決定木を作成した。往診の場合は対象データは医科外来のすべてで、決定木を作成する対象患者は 40 歳以上のデータを用いた。往診以外の場合は生存者数が非常に多く死亡者とのデータの不均衡が生じるため、生存者のみ 1/30 無作為抽出とした。マスターの死亡転帰より、死亡者数と生存者は表 7 のとおりであった。これらをそれぞれ学習用データと検証用データに2:1で分類し、入院と同様に検証した。説明変数には入院同じ手法で、退院日または死亡日に算定された診療行為および医薬品、傷病名を説明変数に使用した。集計したところ、死亡と生存・往診と通院でそれぞれ表 8～表 11 の推定結果となった。また、感度・特異度・AUC は表 12 のとおりであった。

表 7　外来の対象患者

	往診		往診以外	
	転記区分生存	転記区分死亡	転記区分生存	転記区分死亡
死亡者	5,997	4,613	2,727	2,025
生存者	2,213	908	570,202	650

表 8　外来往診の推定結果

		正解		合計
		死亡	生存	
推定	死亡	1,909	26	1,935
	生存	94	713	807
合計		2,003	739	2,742

表 9　外来往診(転帰区分が死亡)の推定結果

		正解		合計
		死亡	生存	
推定	死亡	1,467	23	1,490
	生存	74	280	354
合計		1,541	303	1,844

表 10　外来往診以外の推定結果

		正解		合計
		死亡	生存	
推定	死亡	720	52	772
	生存	191	6,318	6,509
合計		911	6,370	7,281

表 11　外来往診以外(転帰区分が死亡)の推定結果

		正解		合計
		死亡	生存	
推定	死亡	640	55	695
	生存	36	162	198
合計		676	217	893

表 12　外来の感度・特異度

	往診		通院	
	転記区分 生存	転記区分 死亡	転記区分 生存	転記区分 死亡
感度	95.3%	95.2%	79.0%	94.7%
特異度	96.5%	92.4%	99.2%	74.7%
偽陽性率	3.5%	7.6%	0.8%	25.3%
偽陰性率	4.7%	4.8%	21.0%	5.3%
陽性的中率	98.7%	98.5%	93.3%	92.1%
陰性的中率	88.4%	79.1%	97.1%	81.8%
AUC	0.977	0.961	0.917	0.921

決定木に関しては図 4～図 7 のとおりとなった。図 4、図 5 に関して往診で転帰区分が生存/死亡の違いで出てきた変数としては、両者ともに「血液検査の有無」や「内服の有無」、「在宅訪問診療料」が上位であった。しかし、死亡転帰ありの場合の決定木には変数の出現が少なかった。図 6、図 7 の往診以外で転帰区分が生存/死亡の共通で出てきた変数としては、「非開胸的心臓マッサージ」が上位にあり、転帰が生存の場合は「在宅訪問診療料」や「在宅ターミナル加算」、転帰が死亡の場合は薬の処方に関連して「処方せん料」や「調剤料」がそのほかの説明変数としてあらわれた。

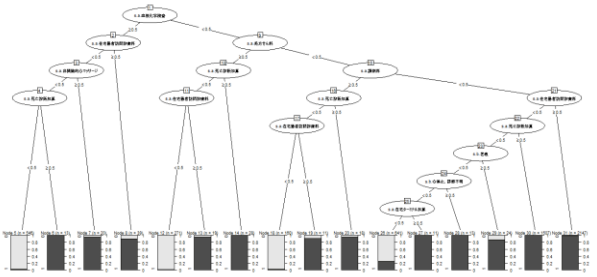

図 4　外来往診の決定木分析結果

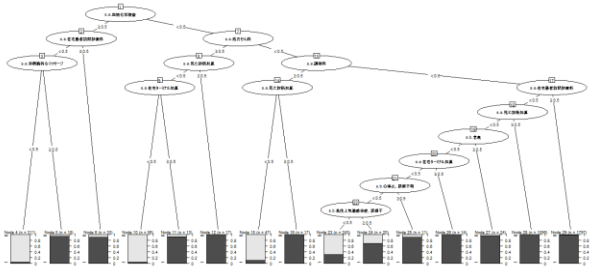

図 5　外来往診(転帰区分が死亡)の決定木分析結果

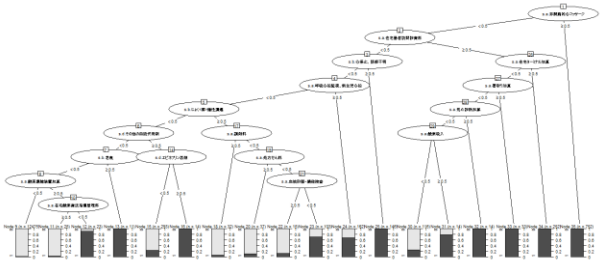

図 6　外来往診以外の決定木分析結果

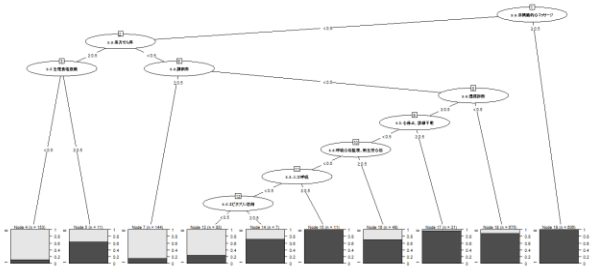

図 7　外来往診以外(転帰区分が死亡)の決定木分析結果

3.4 NDB への適用

これら 6 つの決定木を基として、NDB で死亡を付与するプログラムを作成した。具体的には、決定木は各枝の末端ごとに「死亡」「生存」が割り振られており、枝の通る道に分岐として説明変数が記されている。各枝の説明変数ごとに条件式を設けて振り分ける SQL を記載し、集計を行った。入院において、決定木で死亡を付与した場合と、転帰区分で死亡がつけられている場合で分類し集計した結果、表 13 のとおりとなった。決定木による死亡付与は 3 年分で 8,199,968 名となった。これにかかる決定木の末端の内訳は表 14 のとおりとなった。

外来に関しては往診と往診以外で振り分けているが、死亡の殆どが往診であるため、今回は往診のみ記載した。往診以外は死亡者数が少ないため、精度を高める必要があり、今回は表には掲載しない。死亡者数を表 15 に、これにかかる決定

木の末端の内訳は転帰死亡がなしの場合を表 16、死亡転帰ありを表 17 に示した。No.は各決定木の末端の番号を示しており、最終条件を示しているが、各条件に至るまでは別の経路をたどっており、項目が一緒であっても患者層は異なる。

表 13　入院の死亡者数

決定木による区分	転帰区分	件数	患者数(生存)	患者数(死亡)
生存	生存	15,624,699	15,637,078	-
生存	死亡	12,379		
死亡	生存	5,835,903	-	8,199,968
死亡	死亡	2,364,065		

表 14　入院の決定木別死亡・生存者数の分類

No.	最終条件	生死フラグ※	KDB	NDB	NDB(内訳)		
			死亡率	死亡率	患者数	生存	死亡
#c8	人工呼吸していない	0	2%	1%	8,237,639	8,155,592	82,047
#c7	人工呼吸した	1	85%	39%	42,141	25,603	16,538
#c14	喀痰吸引（1日につき）していない	0	28%	8%	82,066	75,710	6,356
#c13	喀痰吸引（1日につき）した	1	82%	44%	9,841	5,466	4,375
#c11	生理食塩液類を処方した	1	72%	27%	34,660	25,211	9,449
#c9	呼吸心拍監視、新生児心拍した	1	86%	75%	108,152	26,981	81,171
#c15	無機塩製剤；炭酸水素ナトリウムを処方した	0	0%	0%	42,414	42,354	60
#c17	その他の消化性潰瘍用剤を処方した	0	3%	0%	38,346	38,229	117
#c18	その他の消化性潰瘍用剤を処方していない	1	74%	47%	702,710	372,060	330,650
#c19	大腸ファイバースコピーした	0	0%	0%	100,166	99,766	400
#c38	中心静脈注射していない	0	21%	2%	1,060,206	1,035,888	24,318
#c37	中心静脈注射した	1	79%	25%	6,199	4,661	1,538
#c35	その他の眼科用剤した	0	0%	0%	26,992	26,966	26
#c36	その他の眼科用剤していない	1	81%	22%	67,333	52,624	14,709
#c30	呼吸心拍監視、新生児心拍していない	0	24%	1%	5,968,442	5,923,293	45,149
#c29	呼吸心拍監視、新生児心拍した	1	91%	10%	354,185	317,774	36,411
#c31	大腸<結腸>のポリープの検査した	0	20%	4%	236,193	226,902	9,291
#c32	大腸<結腸>のポリープの検査していない	1	85%	6%	5,047,940	4,734,258	313,682
#c23	人工呼吸した	1	98%	91%	405,937	36,600	369,337
#c21	酸素吸入した	1	97%	81%	1,265,368	243,795	1,021,573

※生＝0、死＝1としている
※KDBの死亡率はマスター死亡より、NDBの死亡率は転記死亡より比較している

表 15　外来(往診)の死亡者数

決定木による区分	転帰区分	件数	患者数(生存)	患者数(死亡)
生存	生存	238,059	247,938	-
生存	死亡	9,879		
死亡	生存	99,404	-	359,123
死亡	死亡	259,719		

表 16　外来(往診)の転帰に死亡なし患者における決定木別死亡・生存者数の分類

No.	最終条件	生死フラグ※	KDB				NDB	
			総数	生存	死亡	死亡率	生存	死亡
c8	死亡診断加算を算定していない	0	546	528	18	3%	28,080	
c7	死亡診断加算を算定した	1	13	0	13	100%		428
c5	非開胸的心マッサージした	1	20	1	19	95%		1,303
c3	在宅患者訪問診療料を算定した	1	39	4	35	90%		1,603
c14	在宅患者訪問診療料を算定していない	0	271	268	3	1%	36,657	
c13	在宅患者訪問診療料を算定した	1	19	1	18	95%		976
c11	死亡診断加算を算定した	1	28	0	28	100%		710
c20	在宅患者訪問診療料を算定していない	0	160	155	5	3%	15,384	
c19	在宅患者訪問診療料を算定した	1	11	1	10	91%		311
c17	死亡診断加算を算定した	1	18	1	17	94%		409
c30	在宅ターミナル加算を算定していない	0	641	475	166	26%	157,938	
c29	在宅ターミナル加算を算定した	1	11	0	11	100%		132
c27	心停止，詳細不明の病名が付与された	1	13	0	13	100%		3,977
c25	老衰の病名が付与された	1	24	3	21	88%		5,359
c23	死亡診断加算を算定した	1	1507	6	1501	100%		37,370
c21	在宅患者訪問診療料を算定した	1	2147	31	2116	99%		46,826

※生＝0、死＝1としている
※KDBの死亡率はマスター死亡より、NDBの死亡率は転記死亡より比較している

表 17　外来(往診)の転帰に死亡あり患者における決定木別死亡・生存者数の分類

No.	最終条件	生死フラグ※	KDB				NDB	
			総数	生存	死亡	死亡率	生存	死亡
c6	非開胸的心マッサージしていない	0	311	296	15	5%	1,358	
c5	非開胸的心マッサージした	1	18	1	17	94%		1,172
c3	在宅患者訪問診療料を算定した	1	33	3	30	91%		1,996
c12	在宅ターミナル加算を算定していない	0	56	53	3	5%	612	
c11	在宅ターミナル加算を算定した	1	13	1	12	92%		1,235
c9	死亡診断加算を算定した	1	17	0	17	100%		1,469
c16	死亡診断加算を算定していない	0	47	42	5	11%	579	
c15	死亡診断加算を算定した	1	17	0	17	100%		773
c28	急性上気道感染症，詳細不明の病名を付与	0	243	168	75	31%	7,330	
c27	急性上気道感染症，詳細不明の病名を付与	1	20	6	14	70%		1,620
c25	心停止，詳細不明の病名を付与した	1	11	1	10	91%		941
c23	在宅ターミナル加算を算定した	1	14	0	14	100%		296
c21	老衰の病名を付与した	1	24	1	23	96%		1,296
c19	死亡診断加算を算定した	1	1056	5	1051	100%		90,371
c17	在宅患者訪問診療料を算定した	1	1797	28	1769	98%		158,550

※生＝0、死＝1としている
※KDBの死亡率はマスター死亡より、NDBの死亡率は転記死亡より比較している

4. 考察

決定木として適切なものの要件は、①決定木の説明変数の選択の妥当性、②AUC や感度特異度から判断した精度、③決定木に含まれる説明変数が臨床において説明できるか、④死亡数が死亡統計と比べて正確かの 4 点であるといえる。

4.1 決定木の説明変数の選択の妥当性

説明変数の選択においては、種々の診療行為、医薬品、病名等から判別をする必要がある。このうち、レセプトの最終日に診療行為などが発生している場合という条件にした理由は、最終日に行われている処置がその傾向をつかむうえで重要と判断したためである。検証段階で、集計対象に含める診療行為を期間別(最終日、3 日以内、5 日以内)で検証したが、期間が長くなれば長くなるほど余計なノイズ(日々の治療)が含まれてきており、死亡患者を説明しにくい状況となった。このため、今回は最終日に限定したところが重要な点である。

また、今回は対象を 40 歳以上の患者に限定した。これは先述のとおり、39 歳以下の患者では死亡の絶対数が少なく、周産期と小児が合わさって出現することから、正しいツリーを導けないと判断したためである。今回は分析の限界と言わざるを得ないが、今後小児や母性に特化した検証が必要であるといえる。

医学管理料および基本診療料を除外した理由については、すべての患者で継続的に発生しており説明変数としての役割が果たせなかったためである。このような変数はほかにも一定存在すると考えられるが、今回特定に至った部分はこのとおりであった。

上記のことより、決定木を作成する上でたたき台となる説明変数を用意することができたと考えるが、今後の精緻化は必要であるといえよう。

4.2 AUC や感度特異度から判断した精度

AUC は ROC 曲線を描いた際のモデル評価として使用されるものであるが、1 に近ければ精度の高いモデルであるといわれている。AUC による妥当性の評価は決定木の一面であり、参考程度に測定すると、入院の死亡転帰ありの決定木をのぞき概ね 0.9 を超えており、ある程度の精度は担保できていることが示唆された。なお、入院の死亡転帰ありの決定木に関しては、0.877 であったことが今後の課題である。ほかの決定木に比べ入院の死亡転帰ありの決定木は偽陽性率が 45.9%と高く、感度を高めることを優先した結果がこのようになったのではないかと推測された。また、この決定木は死亡転帰がある患者に対して、死亡されていないケースを除外するためのロジックとして今回採用されたものである。対象の集団にバイアスがかかるため、他の決定木に比べて精度が落ちてしまう可能性は否定できない。これらは今後に課題を残す結果となった。それ以外の部分についてはおおむね感度・特異度ともに概ね 80-90%以上となっており精度が担保されたといえる。

4.3 決定木に含まれる説明変数が臨床において説明できるか

AUC や感度特異度の精度が上がっても、決定木に含まれる説明変数が臨床に合うものでなければ懐疑的と言わざるを得ない。今回は、6つの決定木を決定しているが、これらが決定するまでに多くの決定木作成し、トライアンドエラーで実施している。たとえ AUC が高くても決定木が説明できないものは、対象から外す、という対応を行った。

その結果、入院であれば食事摂取の有無や、内視鏡検査、酸素吸入や退院処方の有無などが説明変数として抽出された。決定木の中に「入院で食事をしていない人が内視鏡

検査を受けたかどうか」という条件が存在する。これは絶食中でも内視鏡検査の場合は検査のみ受けて帰宅する(＝予後もいい)ということの表れであるといえる。外来は、往診は寝たきり患者の往診を示す在宅訪問診療料や、死亡診断加算、看取り加算などが説明変数として抽出された。通院に関しては、心臓マッサージやエピネフリンなどの救命的な加算・薬剤や、在宅訪問診療料や、死亡診断加算、看取り加算などが説明変数として抽出された。「調剤料」や「処方せん料」などが一定の項目で発生しているが、これは通院日に一定の内服ができる患者は生存している可能性が高いことを示している。また、病名において老衰や心停止が診断されている患者においても死亡に振られることが多いことが分かった。これらの診療行為や薬剤、病名は臨床としても違和感がないため、今回の決定木の説明変数としては可用性が高いものであるといえる。

4.4 死亡数が死亡統計と比べて正確か

NDB3 年間分の入院で 8,199,968 名、外来(往診)で 359,123 名の死亡者数となった。死亡統計における年次別の死亡者数を表 18 に示す。

表 18　死亡統計における死亡者数

年次	死亡者数
平成25年	1,268,436
平成26年	1,273,004
平成27年	1,290,428
３年累計	3,831,868

これは、入院、外来、医療機関外死亡(検死、自殺など)を含む結果である。単純に比較はできないが、医療機関等の施設死亡は 2015 年で 85.2%に達しており[6]、ほとんどが入院や通院による死亡と判断される。しかしながら、入院の死亡患者が約 820 万人と比較すると明らかに乖離している。原因は、入院の決定木の「大腸<結腸>のポリープの検査をしていない」「呼吸心拍監視、新生児心拍を算定している」「その他の消化性潰瘍用剤を調剤されていない」に当てはまった患者に関して死亡としているが、KDB 側ではおおむね 74%～91%の死亡率の精度であったのに対し、NDB の転帰区分のみをアウトカムとして死亡率を算出した場合、6～47%と大幅に精度が低下することが分かった。今回、KDB と NDB で別の傾向を示すものであると考えられたため、これらの決定木に関しては死亡としない、という調整を行った。調整により集計しなおした結果を表 19 に示す。

表 19　入院の死亡者数（調整後）

決定木による区分	転帰区分	件数	患者数(生存)	患者数(死亡)
生存	生存	21,048,791	21,460,602	-
生存	死亡	411,811		
死亡	生存	693,122	-	2,376,444
死亡	死亡	1,683,322		

転帰死亡で死亡としている人を生存とする数が増えたため、感度が低下していると考えられるが、死亡患者数の祖語を減少させた。3 年間の死亡患者数が 383 万人で、15%が医療機関外死亡であることを考えると、326 万人が施設で死亡していることとなる。入院患者の死亡患者数としては比較的近しい数値に近づいたといえるが、これらの除外したデータについては、今後決定木を組み替えるほかない。

また、外来の往診については施設外死亡が 19 万人存在している。死亡統計の医療機関外死亡患者は 57 万人存在しており、まだ一定数存在していると考えられる。往診患者と通院患者の定義をより精緻化して、さらなる検証が必要である。ただし、KDB の転帰区分からマスター死亡の比率(＝誤答率)を計算したところ、4.3%で、NDB においても転帰区分から付与した死亡フラグの誤答率を算出すると 3.7%で比較的近しいため、多角的な検証で精度を比較することが必要である。

5. 結論

本研究では、KDB のデータを用いて死亡患者の動向を機械学習し決定木分析を行うことで NDB の死亡フラグを策定した。現状ではそのすべての死亡を完全に追うことはできないうえ、決定木自体に課題を残す結果となった。今後は、この死亡ロジックの精度を高めることで正確性を担保した死亡ロジックを構築する必要がある。死亡したアウトカムを正確に付与できれば NDB で日本のコホート研究が大きく前進する。

謝辞

本研究は平成 30 年度厚生労働科学研究費補助金(地域医療基盤開発推進研究事業)「地域の実情に応じた医療提供体制の構築を推進するための政策研究」、平成 30 年度文部科学研究費助成事業(科学研究費補助金)基盤研究(A)(一般)「データ科学・疫学・臨床医学の融合による日本の保険診療情報(NDB)の全解析」、平成 28,29 年度国立研究開発法人日本医療研究開発機構(AMED)地域横断的 ICT 活用医療推進研究事業「レセプト等の大規模電子診療情報を活用した薬剤疫学研究を含む医療パフォーマンス評価に関する研究」、平成 28,29 年度国立研究開発法人日本医療研究開発機構(AMED)臨床研究等ICT基盤構築・人工知能実装研究事業「新たなエビデンス創出のための次世代 NDB データ研究基盤構築に関する研究」の一環として実施したものである。

参考文献

1) 中央社会保険医療協議会 総会(第 356 回) 議事次第. 横断的事項(その2)について. レセプト情報・特定健診等情報データベース(NDB)の概要. 厚生労働省保険局医療課, 2017. [http://www.mhlw.go.jp/file/05-Shingikai-12404000-Hokenkyoku-Iryouka/0000170931.pdf (cited 2018-Sep-03)].

2) 久保 慎一郎, 野田 龍也, 明神 大也, 加藤 源太, 今村 知明. NDB(ナショナルデータベース)の課題および留意点と今後の展望. 医療情報学連合大会論文集 2016 ; 36(1) : 272-275.

3) 久保 慎一郎, 野田 龍也, 明神 大也,東野 恒之, 松居 宏樹, 加藤 源太, 今村 知明.レセプト情報・特定健診等情報データベース (NDB)の臨床研究における名寄せの必要性と留意点. 日本健康開発雑誌 2017;38:11-18.

4) 野田龍也、久保慎一郎、明神大也、西岡祐一、東野恒之、松居宏樹、他. レセプト情報・特定健診等情報データベース(NDB)における患者突合(名寄せ)手法の改良と検証. 厚生の指標. 64(12), 7-13, 2017-10.

5) Kubo S, Noda T, Myojin T, et al. National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB): outline and patient-matching technique. bioRxiv. [https://www.biorxiv.org/content/early/2018/04/02/280008.full.pdf. (cited 2018-Sep-3)].

6) 平成 27 年人口動態調査 上巻 死亡 第5. 5表 死亡の場所別にみた年次別死亡数. 厚生労働省, 2015. [https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00450011&tstat=000001028897&cycle=7&year=20150&month=0&tclass1=000001053058&tclass2=000001053061&tclass3=000001053065&stat_infid=000031450149 (cited 2018-Sep-03)].