

特集 1：未分類疾患の情報集約に関する研究

<ノート>

未分類疾患の発見プロセスに関する確率論的考察

緒方裕光¹⁾, 奥村貴史¹⁾

¹⁾ 国立保健医療科学院研究情報センター

A Probabilistic Approach for Detecting Undiagnosed Diseases

Hiromitsu OGATA¹⁾, Takashi OKUMURA¹⁾

¹⁾Center for Information Research and Library, National Institute of Public Health

抄録

既存の診断基準等によって特定できない疾患、すなわち未分類疾患は世界中に多数存在している。このような未分類疾患については、症状や原因などに関する情報が集積されておらず、診断技術・治療方法などに関する研究を進めることがきわめて難しい。

本報告では、既存の疾患および症候に関する情報に基づいて未分類疾患を見出すための方法について、確率論的な観点から検討を行った。症候（結果）から疾患（原因）を推定する際に、両者の関係に関してベイズの定理を適用したベイズ診断の方法がすでに提示されている。この考え方を未分類疾患発見プロセスに応用して、仮想データに基づいて、情報の蓄積と未分類疾患の確率（特定の症候が既分類疾患に分類されない確率）との関係について考察を加えた。情報の蓄積にともない、特定の症候が未知の原因に起因する確率は一般に減少するが、この未知の原因には、情報不足や非疾患の場合も含まれていることを考慮する必要がある。

キーワード：未分類疾患、希少疾患、確率、診断

Abstract

There are many diseases that cannot be specified by existing diagnostic criteria, and for many of them, the disease concept has not been established. For these diseases, it is difficult to advance the research on diagnosis and treatment because information about symptoms is insufficient. To presume a disease from symptoms, Bayesian diagnosis based on Bayes' theorem has already been presented. But the relationship between the process of discovering an undiagnosed disease and the accumulation of information is not clear. In this study, the method of finding an undiagnosed disease based on symptoms was examined from the viewpoint of probability theory using virtual data. The paper also shows that the probability of unidentified cause decreases along with the accumulation of information.

Keywords: undiagnosed disease, rare disease, probability, diagnosis

I. はじめに

既存の診断基準等によって特定できない疾患、すなわち未分類疾患は世界中に多数存在している可能性があり、そ

れらの多くはその疾患概念が確立されていない。このような疾患については、症状や原因などに関する情報が集積されておらず、診断技術・治療方法などに関する研究を進めることがきわめて難しい。本報告では、未分類疾患の分類

連絡先：緒方裕光

〒351-0197 埼玉県和光市南 2-3-6

TEL: 048-458-6203 FAX: 048-469-0326

E-mail: ogata@niph.go.jp

2-3-6 Minami, Wako-shi, Saitama 351-0197, Japan.

方法を確立することを最終的な目標とし、その一端として既存の症候に関する情報に基づいて未分類疾患を見出すための方法について、確率論的な観点から検討を行った。

II. 方法

症状から疾患名を診断することは、ある結果（症候）が特定の原因（疾患）によるものであるかどうかを判定することに相当する。この原因と結果との因果関係に関してベイズの定理を応用することによって、この診断のプロセスを定式化することができる（ベイズ診断）。このベイズ診断を実際に行うためには、特定の疾患によって特定の症候が現れる確率があらかじめ分かっている必要はない。通常、疾患と症候は複数あるので、それぞれの数を m , n とすれば $m \times n$ の組み合わせに対応した確率の表（疾患－症候行列）が必要となる。

本報告では、この疾患－症候行列が得られているという前提で、特定の症状のパターンが見られた場合に、それが未分類疾患である可能性を表現することを試みる。方法の概略は以下のとおりである。

1) 未知の原因を考慮しない場合（ベイズ診断）

$D_1, D_2, \dots, D_m$ を疾患（ある症候の原因）とするとき、症候 S が見られたときその原因が D_i である確率は、ベイズの定理により

$$P(D_i | S) = \frac{P(D_i)P(S | D_i)}{\sum_{k=1}^m P(D_k)P(S | D_k)}$$

と表わされる。ただし、 $D_1, D_2, \dots, D_m$ はすべての疾患を網羅しており、同時に2つの疾患が原因になることはない（すなわち $D_1, D_2, \dots, D_m$ はそれぞれ互いに排反である）と仮定する。ここで、 $P(D_i)$ は各疾患の事前確率、 $P(S | D_i)$ は疾患 D_i によって特定の症候 S が現れる確率である。通常、症候は複数あるので、疾患－症候行列は $P(S_j | D_i)$ ($i=1, 2, \dots, m, j=1, 2, \dots, n$) と表わされる。一方、それぞれの疾患 S_j が現れたときその原因が D_i である確率は $P(D_i | S_j)$ と表わされる。

実際には、ある特定の疾患 D_i に関して、その発生に関する統計的頻度が分かっているならばその値を確率に置き換えて $P(D_i)$ とおく。また、その疾患に関してどのような症候がどの程度の割合で現れるかが把握されている場合には、その割合を確率に置き換えて $P(S_1 | D_k), P(S_2 | D_k), P(S_3 | D_k), \dots$ などと置く。

2) 未知の原因を考慮する場合

上記の1)において、 $D_1, D_2, \dots, D_m$ 以外のすべての原因を U とおく。この U （未知の原因）には、「未分類疾患」、「現時点で情報が入手されていない既分類疾患」および「疾患ではない原因」が含まれる。症候 S が現れたときその原因が U である確率は $P(U | S)$ で表わされ、複数の症候 S_j に関しては、 $P(U | S_j)$ と表現される。これらの確率を求めるためには、 U が原因で S_j が現れる確率

$P(S_j | U)$ が必要であり、1)のプロセスを応用して $P(S_j | U)$ および $P(U | S_j)$ を求めることができる（注1参照）。 $P(D_i)$ および $P(S_j | D_i)$ が疾患ごとにそれぞれ互いに独立した情報として得られるのに対して、 U の中には $D_1, D_2, \dots, D_m$ 以外のすべての原因が含まれるため、 $P(U)$ および $P(S_j | U)$ の値は $P(D_i)$ および $P(S_j | D_i)$ に依存する。したがって、最終的に求めようとしている $P(U | S_j)$ （ある疾患が未知の原因に起因する確率）は、既存の既分類疾患の情報に依存することになる。

3) 症状がないことを情報として含める場合

ある症候について、症状があることだけでなく症状がないことも情報として利用するには、症候 S について、症状があるとき $a=1$ 、ないとき $a=0$ とおき、 a が現れたときにその原因が U である確率 $P(U | a)$ を求める必要がある。症候が複数ある場合には、症候 S_j に関して a_j とおき、同様に症状があるとき $a_j=1$ 、ないとき $a_j=0$ とおき、症候 $(a_1, a_2, \dots, a_n)$ が現れたときその原因が U である確率 $P(U | a_1, a_2, \dots, a_n)$ を求める（注2参照）。なお、以下では、これらの $P(U | a_j)$ または $P(U | a_1, a_2, \dots, a_n)$ を中心に議論しているが、その計算方法は前述の1)および2)のプロセスを基礎としている。

4) 仮想データによる検討

疾患－症候行列について仮想データを設定し、このデータを用いて $P(U | a_1, a_2, \dots, a_n)$ を算出する。その際、疾患あるいは症候に関する情報が不足している状態から情報が追加されていく過程において、ある症候が未知の原因に起因する確率の変化について調べた。本報告では、仮想データとして疾患数が $m=10$ ($D_1, D_2, \dots, D_{10}$)、症候数が $n=5$ ($S_1, S_2, \dots, S_5$) の場合の疾患－症候行列を想定した。表1にその疾患－症候行列を示す（文献1より一部 $S_1 \sim S_5$ のデータを引用）。さらに、疾患の発現確率の仮想データとして、 $P(S_1)=0.1, P(S_2)=0.2, P(S_3)=0.15, P(S_4)=0.05, P(S_5)=0.03$ とした。なお、表1の仮想データおよび $P(S_1) \sim P(S_5)$ は任意に設定されたものであるが、これらのデータを利用して得られる結論に関して、少なくとも本報告の議論の中で一般性を失うことはない。

表1 疾患－症候行列 $P(S_j | D_i)$ (仮想データ)

疾患	$P(D_i)$	症候				
		S_1	S_2	S_3	S_4	S_5
D_1	0.100	0.01	0.49	0.50	0.01	0.00
D_2	0.081	0.10	0.50	0.50	0.02	0.01
D_3	0.005	0.30	0.60	0.10	0.20	0.10
D_4	0.001	0.10	0.20	0.70	0.30	0.10
D_5	0.027	0.20	0.50	0.30	0.15	0.05
D_6	0.005	0.10	0.40	0.50	0.01	0.01
D_7	0.001	0.20	0.70	0.10	0.65	0.10
D_8	0.018	0.50	0.48	0.02	0.30	0.65
D_9	0.001	0.10	0.45	0.45	0.22	0.44
D_{10}	0.054	0.40	0.55	0.05	0.25	0.25

Ⅲ. 結果

1) 既知の疾患の確率および未知の原因の確率

原因として未知の原因 (U) を含め, 症候 ($S_1, S_2, \dots, S_5$) に関してそれぞれ症状の有無を情報として考慮した場合 (症候 S_j について症状があるとき $a_j=1$, ないとき $a_j=0$), a_j が見られたときその原因が U である確率 $P(U|a_j)$ を求めた. その結果を表 2 に示す. 「症状がなかった」という情報よりも, 「症状があった」という情報の方が, 未知の原因であると判断される確率を小さくしている.

表 2 未知の原因 (U) を考慮し, かつ症状がないことを情報として含めた場合に症候から推定される原因の確率 $P(D_i|a_j)$ および $P(U|a_j)$ (表 1 のデータに基づく計算結果)

疾患	症候*									
	S_1		S_2		S_3		S_4		S_5	
	$a_1=0$	$a_1=1$	$a_2=0$	$a_2=1$	$a_3=0$	$a_3=1$	$a_4=0$	$a_4=1$	$a_5=0$	$a_5=1$
D_1	0.10	0.01	0.06	0.27	0.06	0.37	0.10	0.02	0.10	0.00
D_2	0.08	0.10	0.05	0.22	0.05	0.30	0.08	0.04	0.08	0.03
D_3	0.00	0.02	0.00	0.02	0.01	0.00	0.00	0.02	0.01	0.02
D_4	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00
D_5	0.02	0.06	0.02	0.07	0.02	0.06	0.02	0.09	0.03	0.05
D_6	0.01	0.01	0.00	0.01	0.00	0.02	0.01	0.00	0.01	0.00
D_7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
D_8	0.01	0.11	0.01	0.05	0.02	0.00	0.01	0.12	0.01	0.40
D_9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02
D_{10}	0.04	0.26	0.03	0.16	0.06	0.02	0.04	0.31	0.04	0.46
U	0.73	0.44	0.82	0.20	0.78	0.23	0.72	0.36	0.73	0.04

* $a_j=0$ (症状なし), $a_j=1$ (症状あり)

複数の症候 ($S_1, S_2, \dots, S_5$) に関して同時に情報が得られた場合は, その情報は 5 つの数値 (それぞれ 0 または 1) の組み合わせ ($a_1, a_2, \dots, a_5$) となる.

ここで, ($a_1, a_2, \dots, a_n$) が現れた場合にその原因が D_i または U である確率を求めた. このとき, ($a_1, a_2, \dots, a_n$) のパターンは $2^5=32$ 通りあり, それぞれのパターンについて $P(D_i|a_1, a_2, \dots, a_5)$ または $P(U|a_1, a_2, \dots, a_5)$ を求めた. その結果の一部を表 3 に示す. これらの確率の大小に関するパターンはきわめて複雑であるが, 症候 (症状の有無を含めて) の組み合わせによってはひじょうに高い確率で原因を特定することができる (すなわち未知の原因である確率は低くなる).

表 3 複数の症候の情報が同時に得られた場合に症候から推定される原因の確率 $P(D_i|a_1, a_2, \dots, a_5)$ および $P(U|a_1, a_2, \dots, a_5)$ の抜粋 (表 1 のデータに基づく計算結果の一部)

疾患	症候 $S_1 \sim S_5$ のパターン (a_1, a_2, a_3, a_4, a_5)*				
	(0,0,0,0,0)	(0,0,0,0,1)	(0,0,0,1,0)	...	(1,1,1,1,1)
D_1	0.038	0.000	0.014	...	0.000
D_2	0.027	0.028	0.019	...	0.006
...	...	...	...	...	...
D_{10}	0.012	0.414	0.139	...	0.566
U	0.908	0.137	0.720	...	0.000

* $a_j=0$ (症状なし), $a_j=1$ (症状あり)

2) 未分類疾患発見のプロセス

現実には, あらゆる疾患と症候に関して網羅的に情報が得られるわけではなく, 通常, 疾患-症候行列は不完全なものである. たとえば, 原因である $D_1, D_2, \dots, D_{10}$ のうち D_{10} に関する情報が存在していなければ, 本来ならば D_{10} が原因である確率は, U が原因である確率 (すなわち未知の原因である確率) に含まれることになる. さらに, 症候についても, 実際にすべての検査を行うことは不可能である.

一般に情報は蓄積されていくものであり, この情報の蓄積過程に伴って不完全さの程度は減少していく. したがって, 情報の蓄積に伴い原因を特定できる可能性は高くなるはずである. しかし, ある症候に関して情報 (ここでは疾患-症候行列に関する情報) の蓄積によっても原因を特定できる可能性が高くならなければ, その症候パターン (複数の症候の組み合わせ) は, 未分類疾患を原因とする可能性が大きいことになる.

ある症候パターンが見られたときにそれが未知の原因によるものである確率について, 疾患に関する情報を追加していった場合 ($P(D_i)$ の大きい順に情報を追加していった場合) の変化の例を図 1 に示す. ここで, 疾患に関する情報が追加されるということは, 疾患-症候行列の各要素 (確率) に真実に近い値が埋められていくことを意味する. 症候パターンによっては, 情報が蓄積されても未知の原因である確率が 0 に近づかない場合がある (たとえば, (a_1, a_2, a_3, a_4, a_5) = (0, 0, 1, 1, 0)).

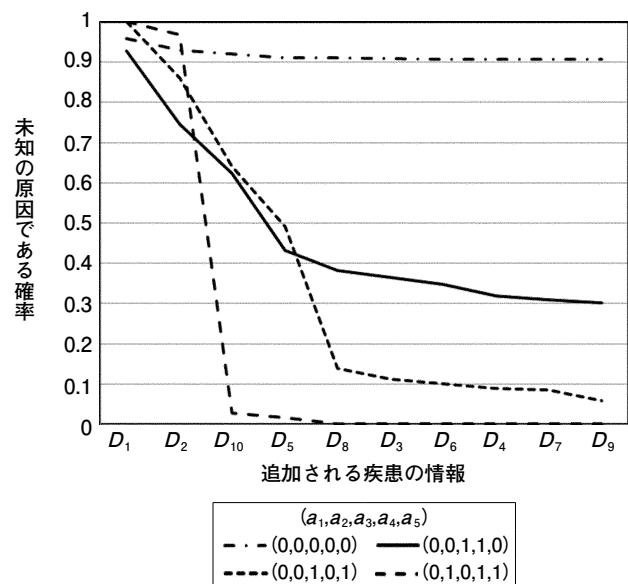


図 1 「疾患」に関する情報が追加されていくことによって, 特定の症候が未知の原因によるものである確率が減少していく例

また, 症候に関する情報が蓄積されていく場合 (ここでは, その例として症候 $a_1 \sim a_3$ の情報が既にあるときに, さらに $a_4=0, a_5=1$ が追加されていく場合を示した) について, 未知の原因である確率の変化を図 2 に示した. この場合も, 図 1 と同様に情報の蓄積にともなって未知の原因である確率が必ずしも 0 に近づかない場合があった.

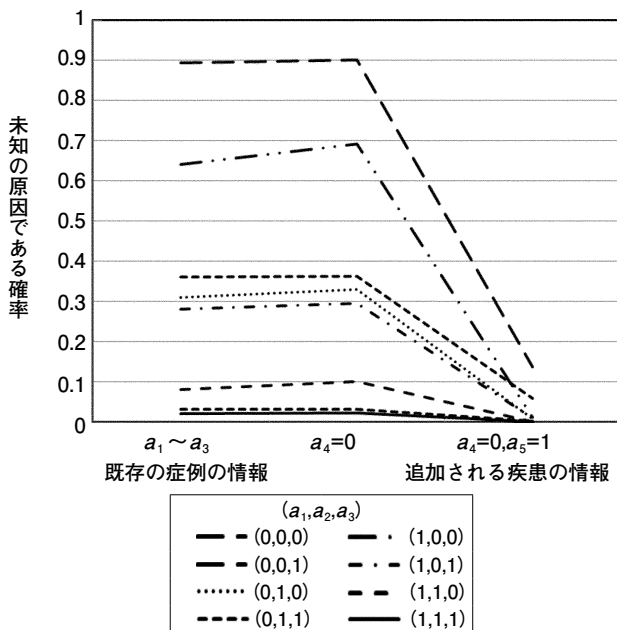


図2 「症候」に関する情報が追加されていくことによって、特定の症候が未知の原因によるものである確率が減少していく例

IV. 考察

本来、既分類疾患に関する情報がすべてそろっていないければ、ある症候が未分類疾患であるかどうかを判断することはできない。しかし、世界中に存在している既分類疾患に関する情報を完全に網羅的に集めることは技術的に限界がある。さらに、完全な情報を維持管理していくためには、時間の経過に伴う情報量の増加や内容の変更に対応しなければならない。したがって、われわれが持っている既分類疾患に関する情報はそもそも完全ではないという前提で、未分類疾患の分類や発見を行う方法を考える方が合理的である。

既に述べたように、未知の原因には「未分類疾患」、「現時点で情報が入手されていない既分類疾患」および「疾患ではない原因」が含まれる。このうち少なくとも「現時点で情報が入手されていない既分類疾患」に関する情報は、時間の経過とともに次第に増加していくはずである。本報告ではこのような状況を考慮したうえで、情報の蓄積と未分類疾患である可能性との関係を表現することを試みた。

医療の現場で通常行われている診断プロセスにおいて、症候からほぼ完全に疾患が特定できる場合もあれば、ある程度の不確実性を持って疾患が判定される場合がある。未分類疾患についても同様で、既分類疾患の情報が完全でない限りは、未分類疾患であるか否かという判断は、未分類疾患である可能性が大きいかどうかという判断をすることと同じである。その可能性の大きさを表現するために確率を用いることは1つの合理的な方法であろう。

本報告では、未分類疾患を見出すための1つのアプローチとして確率論的な観点から検討した。しかし、疾患と症

候の組み合わせの数は無数にあり現実の計算をどのように行うのか、症状の有無が0と1ではなく連続量で表される場合にどのようなモデルが適切なのか、臨床現場からの事例報告や研究者による研究論文など様々な形式の情報をどのように合理的に活用するのか、といった問題が残る。本報告で示した方法は、複雑なパターンを体系的に示すためのシミュレーションとしては十分ではないが、今後いくつかの問題点を整理することによって、多くの情報の中から未分類疾患を見出すための最初のステップとして有効な方法になりうると思われる。

V. 結論

特定の症候パターンが未分類疾患によるものかどうかという可能性と情報の蓄積との関係を確率論的に表現することを試みた。仮想データを用いた試算を通じて検討した結果、結論として以下のことが言える。1) ある症候パターンが未知の原因によるものである可能性は、情報を蓄積することによって一般に減少する。2) 情報の蓄積によっても未知の原因によるものである可能性が減少しなければ、その症候パターンは未分類疾患である可能性が高い。

参考文献

- 1) 松原望. 入門ベイズ統計. 東京: 東京図書; 2008.
- 2) 水野哲夫. 臨床・公衆衛生のための意思決定: 行動の統計学. 東京: 南江堂; 1973.
- 3) Gelman A, Meng XL, editors. Applied Bayesian modeling and causal inference from incomplete-data perspectives. Chichester: John Wiley; 2004.

注1)

症候 S が見られたときその原因が U である確率は

$$P(U|S) = \frac{P(U)P(S|U)}{P(S)}$$

ただし、

$$P(U) = 1 - \sum_{k=1}^m P(D_k)$$

$$P(S|U) = P(S) - P(S|\bar{U}) = P(S) - \sum_{k=1}^m P(D_k)P(S|D_k)$$

である。

注2)

$D_1, D_2, \dots, D_m$ がすべての疾患を網羅していると仮定して、症候 S について症状があるとき $a=1$ 、ないとき $a=0$ とおけば、 a が現れたときその原因が D_i である確率は、

$$P(D_i|a) = \frac{P(D_i)[aP(S|D_i) + (1-a)[1 - P(S|D_i)]]}{R}$$

である。ただし、

$$R = \sum_{k=1}^m P(D_k)[aP(S|D_k) + (1-a)[1 - P(S|D_k)]]$$

である. 同様に, 症候 S_j について症状があるとき $a_j=1$, ないとき $a_j=0$ とおけば, $(a_1, a_2, \dots, a_n)$ が現れたときその原因が D_i である確率は,

$$P(D_i | a_1, a_2, \dots, a_n) = \frac{P(D_i) \prod_{j=1}^n [a_j P(S_j | D_i) + (1-a_j)(1-P(S_j | D_i))]}{T}$$

ただし,

$$T = \sum_{i=1}^m P(D_i) \prod_{j=1}^n [a_j P(S_j | D_i) + (1-a_j)(1-P(S_j | D_i))]$$

以下, $P(U | a_1, a_2, \dots, a_n)$ は, 注 1 と同様のプロセスで算出できる.