別紙４

令和 6 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

「公的に標準化された医療情報を活用した感染症流行状況と一般診療状況を把握するための分析手法の開発および評価方法に関する研究」

分担研究報告書

DPCデータを用いた感染症流行状況・診療状況の把握手法の開発

研究分担者　高嶋 隆太　　東京理科大学 創域理工学部経営システム工学科 教授
研究分担者　新城 大輔　　東京医科歯科大学大学院 医療政策情報学分野 准教授
研究分担者　伏見 清秀　　東京医科歯科大学大学院 医療政策情報学分野 教授
研究代表者　佐藤 大介　　藤田医科大学大学院 病院経営学・管理学 教授

研究要旨:

〇研究目的

本研究は、SIRD モデルをベースとしたパンデミックリスク管理モデルを開発することを目的する。特に、モデルのパラメータ推定に DPC データを用いることで、症状別の回復率や死亡率を把握する。また、入院日数等の予測に対する機械学習の適用可能性について検討する。

〇研究方法

2020 年～2022 年の DPC データを用いて、人口の状態を S(感受性)、I(感染性)、R(回復)、D(死亡)に分けてその推移を微分方程式で表した SIRD モデルのパラメータを推定し、入院患者数や新規陽性者数等を示した。また、DPC データを用いた患者の入院日数の予測のシステムを構築し、乳がんに関する基礎データを用いて、本システムの検証を行なった。

〇研究結果

SIRD モデルのパラメータ推定から、軽症、中等症、重症それぞれに対する回復率、死亡率は整合的であることが示された。また、入院患者数と新規陽性者数それぞれの推移の比較から、重症化率をモデルに組み込むことの重要性を明らかにした。さらに、機械学習をベースとした予測システムにより、乳がん患者の予後予測、その予測根拠をルールによって説明することが可能となり、DPC データを用いた入院日数予測への応用可能性を示した。

〇結論

DPC データにより SIRD モデルのパラメータを推定し、感染症流行状況を把握するための方法論を開発した。また、学習分類子システムを用いることにより、入院日数の予測や特徴量の評価が可能であることが示唆された。

研究協力者
後藤 允　東京理科大学 創域理工学部 経営システム工学科 准教授
原田 拓　東京理科大学 創域理工学部 経営システム工学科 准教授
伊藤 和哉　東京理科大学 創域理工学部 経営システム工学科 助教

A. 研究目的

2019 年 12 月 1 日に中国の武漢で最初の感染者が発見された新型コロナウイルス感染症（COVID-19）は、その後またたく間に世界中に拡大し、2020 年 3 月 11 日に WHO がパンデミック（世界的大流行）を宣言するに至った。2023 年 5 月の WHO 緊急事態宣言終了時点におけるわが国の累計感染者数は 3,380 万人、累計死亡者数は 7 万 4 千人にのぼる。社会生活や経済に甚大な影響が及んだことは明白で、2023 年の経済産業省発表によると、日本経済の損失は対 GDP 比でマイナス 6.1%であり、約 30 兆円以上であると推定されている。また、2024 年 7 月には第 11 波を記録したことは記憶に新しく、新型コロナウイルス感染症は今後も注視すべき脅威のままである。

今回の事態に対して日本政府がとった政策は、医療体制の確保やワクチン接種などの感染症に直接的なもの以外は、給付金や助成金、キャンペーンなどであり、経済政策といえるものは実質無利子・無担保融資、いわゆる「ゼロゼロ融資」くらいである。このような未曾有の状況下では、政府が事前に政策対応することは難しく、事後的な補償が中心にならざるを得ない。病院の立場としては、政府に頼るだけでなくパンデミックに対するリスク管理を自ら実行することが肝要であり、定性的な評価だけでなく定量的な評価も欠かすことができないが、これらの実行は非常に困難である。

新型コロナウイルス感染症に対するリスク管理が難しい原因のひとつとして、パンデミックの予測が困難であったことが挙げられる。新型コロナウイルス感染症の分析で広く知られるようになった感染症疫学モデルとして、人口の状態を S（感受性）、I（感染性）、R（回復）に分けてその推移を微分方程式で表した SIR モデルが挙げられる。SIR モデルは、Kermack and McKendrick [2] によって開発され、現在に至るまで多種多様な応用が研究されている。

しかしながら、新型コロナウイルスのような未知の感染症の予測は不確実性が高く、標準的な SIR モデルでは十分に対応できなかったといえる。SIR モデルの応用の中でも不確実性に着目し、拡散項を導入した確率的疫学モデルを分析したのが Ji and Jiang [3]、 Dieu et al. [4]などである。これらの確率的疫学モデルを利用すればパンデミックに対するリスク分析が可能となるが、そのような例はまだ見当たらない。

本研究の目的は、確率的 SIR モデルをベースとしたパンデミックリスク管理モデルを開発することである。特に、モデルのパラメータ推定に DPC データを用いることで、症状別の回復率や死亡率を把握することが可能となる。これによって、パンデミックリスクの定量的評価とパンデミックリスク管理が可能となり、予測不可能なリスクに対して頑健な経営システムの確保につながる。

また、DPC データを用いて患者の入院日数を予測することは、医療施設での病床の確保計画の基礎データとして有用であり、パンデミックリスク管理の一助となる。そこで、このための基礎として、まず、乳がんに関する基礎データを用いて乳がん患者の予後予測を行い、さらに、その予測根拠をルールによって説明する。

B. 研究方法

厚労省オープンデータを用いた確率的 SIS モデル

ベンチマークとして人口状態のうち R（回復）を除いた SIS モデルによる分析を行う。データは厚労省が提供する COVID-19 の新規陽性者数の日次データを用いる。確率的 SIS モデルは

$$dS_t = -(\beta S_t - \gamma) I_t dt - \sigma S_t I_t dW_t$$

$$dI_t = (\beta S_t - \gamma)I_t dt + \sigma S_t I_t dW_t$$

と表すことができる。ただし、βは感染率、γは回復率、σは拡散パラメータ、W_tは標準ブラウン運動である。SIS モデルの状態推移をまとめたのが図 1 である。

データ期間は 2020/02/01–2022/09/08 であり、COVID-19 の回復には平均して 14 日かかるとされているため、回復率には日次 1/14、月次 2.17 を用いる。データを加工した感染者数の推移は図 2 のとおりであり、各ピークを捉えた結果となっている。

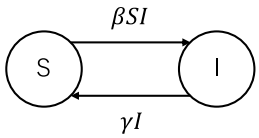

図 1　SIS モデルの状態推移

図 2　確率的 SIS モデルの感染者数推移

表 1　確率的 SIS モデルの月次パラメータ

感染率β	2.59
回復率γ	2.17
拡散パラメータσ	0.39

感染率と拡散パラメータの推定には、R 言語の YUIMA パッケージによる擬似最尤法を用いる。推定結果は表 1 のとおりである。このとき、基本再生産数は

$$R_0 = \frac{\beta - \sigma^2/2}{\gamma} = 1.01 > 1$$

となり、感染拡大のリスクがあることが分かる。

DPC データを用いた SIRD モデル

次に、人口状態に D（死亡）を加えた SIRD モデルによる分析を行う。DPC データを用いることによって、症状別の回復率や死亡率を把握することが可能となる。確定的 SIRD モデルは

$$dS_t = -\beta S_t I_t dt$$
$$dI_t = (\beta S_t - \gamma - \delta)I_t dt$$
$$dR_t = \gamma I_t dt$$
$$dD_t = \delta I_t dt$$

と表すことができる。ただし、δは死亡率である。

DPC データは入院患者のみを対象としているため、人口状態を I（入院）、R（退院）と読み替える。さらに、症状別に入院者を軽症I^l、中等症I^m、重症I^hと分類し、

$$dS_t = -(\beta_1 + \beta_2 + \beta_3)S_t dt$$
$$dI_t^l = (\beta_1 S_t + \alpha_{21} I_t^m + \alpha_{31} I_t^h - (\alpha_{12} + \alpha_{13} + \gamma_1 + \delta_1) I_t^l) dt$$

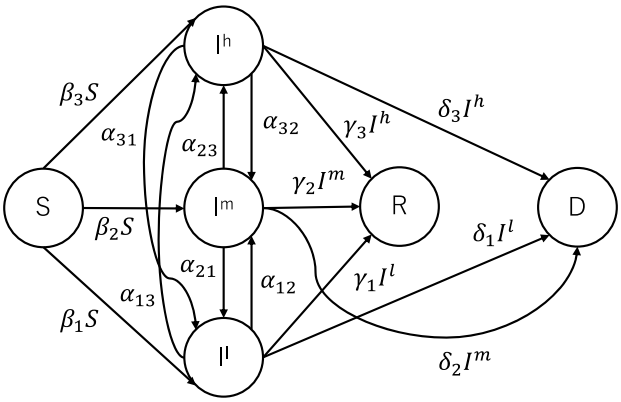

図 3　症状別に分類した SIRD モデルの状態推移（S: 感受性、I^l: 軽症、I^m: 中等症、I^h: 重症、R: 退院、D: 死亡、本来α_{ij}には推移元の I がかかる）

$$dI_t^m = (\beta_2 S_t + \alpha_{12} I_t^l + \alpha_{32} I_t^h - (\alpha_{21} + \alpha_{23} + \gamma_2 + \delta_2) I_t^m) dt$$
$$dI_t^h = (\beta_3 S_t + \alpha_{13} I_t^l + \alpha_{23} I_t^m - (\alpha_{31} + \alpha_{32} + \gamma_3 + \delta_3) I_t^h) dt$$
$$dR_t = (\gamma_1 I_t^l + \gamma_2 I_t^m + \gamma_3 I_t^h) dt$$
$$dD_t = (\delta_1 I_t^l + \delta_2 I_t^m + \delta_3 I_t^h) dt$$

と書き換える。ただし、パラメータの添え字は 1 が軽症、2 が中等症、3 が重症を表す。また、α_{ij}は症状の変化を表し、$i < j$ならば悪化を、$i > j$ならば良化を表す。症状別に分類した SIRD モデル

の状態推移をまとめたのが図 3 である。

学習分類子システム

これまでは、機械学習モデルを適用した結果に注目が集まっていたが、最近、その結果が得られた根拠を説明することに対するニーズが急速に高まっている。この説明は、機械学習モデルがブラックボックスであったため、得られた結果の信

表 2　確定的 SIRD モデルのパラメータ

	S	I^l	I^m	I^h	R	D
S	0.999996	2.64E-06	6.70E-07	7.54E-08	0	0
I^l	0	0.8888	0.0238	0.0019	0.0828	0.0027
I^m	0	0.1035	0.8421	0.0049	0.0452	0.0043
I^h	0	0.0674	0.0402	0.8627	0.0223	0.0074

S: 感受性、I^l: 軽症、I^m: 中等症、I^h: 重症、R: 退院、D: 死亡

例：I^lからDへの推移確率 0.0027 は、軽症患者の死亡率が 0.27%であることを示す。

頼性を説明し難いという問題点を解決するものである。このような、機械学習結果の説明を行う枠組みのことを説明可能 AI と呼ぶ。

本研究では、乳がんデータに対する機械学習結果が得られた根拠を説明するために、学習分類子システムを適用する。学習分類子システムは、if-then ルールを分類子と呼ばれる形式で表現する。分類子とは、入力データを条件として、その入力データから得られる結果を推定するルールである。また、分類子を生成する際には、分類子毎に適合度という分類子の評価値が計算される。適合度を参照することで、ルール集合内での各分類子の評価を確認することができるため、有用なルールを見つけやすいという特徴がある。

複数の特徴量、および、あるクラスの値をもつ患者データに対して、学習分類子システムを適用して If-Then ルールを生成する（図 4）。予測とは、ある特徴量の値を持った患者がどのクラスに属するのかを識別することを意味する。従って、基本的には、得られた分類子（If-Then ルール）は、多くの各患者のデータに当てはまるものとなっている。

C. 研究結果と考察

SIRD モデルによる分析

データ期間は 2020/02/11–2022/02/08 であり、図 5～9 はそれぞれ、全体のI_t、症状別のI_tと、S_t、R_t、D_tを表している。DPC データによる入院患者数の推移を表した図 4 と全新規陽性者数の推移を表した図 5 を比較すると、おおまかな傾向は同じであるが、第 5 波の入院患者比率が顕著に低いこ

図 4　学習分類子システムによる予測（例）

とが分かる。これは、第 5 波では重症化率が低下したこととも整合的であり、分析結果に齟齬はないといえる。

パラメータ推定にはマルコフ推移確率行列

$$P=\begin{pmatrix} * & \beta_1 & \beta_2 & \beta_3 & 0 & 0 \\ 0 & * & \alpha_{12} & \alpha_{13} & \gamma_1 & \delta_1 \\ 0 & \alpha_{21} & * & \alpha_{23} & \gamma_2 & \delta_2 \\ 0 & \alpha_{31} & \alpha_{32} & * & \gamma_3 & \delta_3 \end{pmatrix}$$

を用いる。ただし、1 行目は感受性、2 行目は軽症、3 行目は中等症、4 行目は重症を表し、1 列目は感受性、2 列目は軽症、3 列目は中等症、4 列目

は重症、5 列目は退院、6 列目は死亡を表す。各パラメータは行から列への推移確率を表す。また、*は行和が 1 になるような残余項を表す。

DPC データの加工については、以下のとおりである。

- 対象病院の人口カバー率が約 80%であることから、S の初期値を 1 億人とする
- 患者が入院するごとに症状に応じて S から I^l, I^m, I^h に追加する
- 症状の判定は別表のとおり
- 診療行為に応じて症状別に I^l, I^m, I^h を移動する
- 患者が退院するごとに I^l, I^m, I^h から R に追加する
- 患者が死亡するごとに I^l, I^m, I^h から D に追加する

パラメータ推定結果は表 2 のとおりである。

まず、$\beta_1 > \beta_2 > \beta_3$であることから、軽症、中等症、重症の順に入院率が高く、その差はそれぞれ 10 倍程度であることが分かる。また、$\gamma_1 > \gamma_2 > \gamma_3$であることから、軽症、中等症、重症の順に回復率が高いことが分かる。さらに、$\delta_3 > \delta_2 > \delta_1$であることから、重症、中等症、軽症の順に死亡率が高いことが分かる。以上のことから、推定結果に齟齬はないといえる。

次に、α_{ij}の比較から、悪化よりも良化の傾向が強く、特に重症から中等症よりも重症から軽症のほうが、傾向が強いことが分かる。このことから、適切な措置を施せば症状は良化していくといえる。

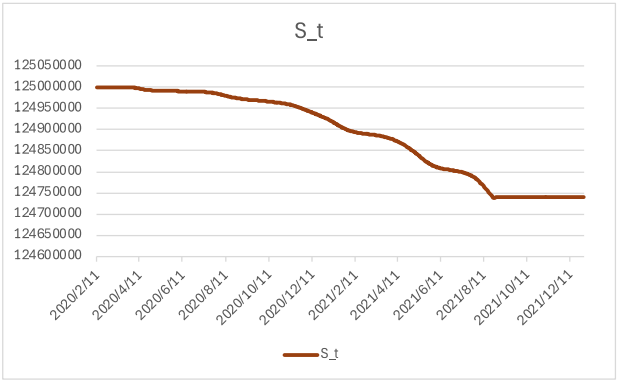

図 7　感受性者数の推移

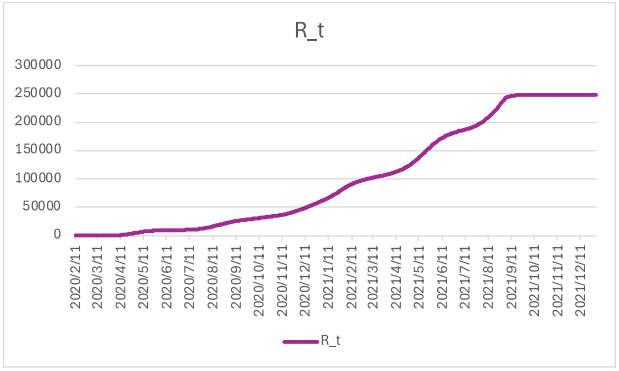

図 8　退院者数の推移

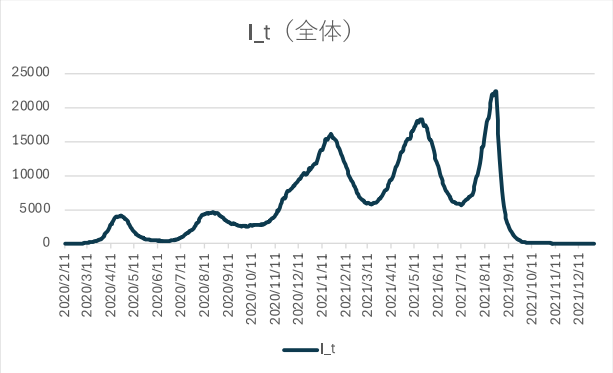

図 5　全体入院者数の推移

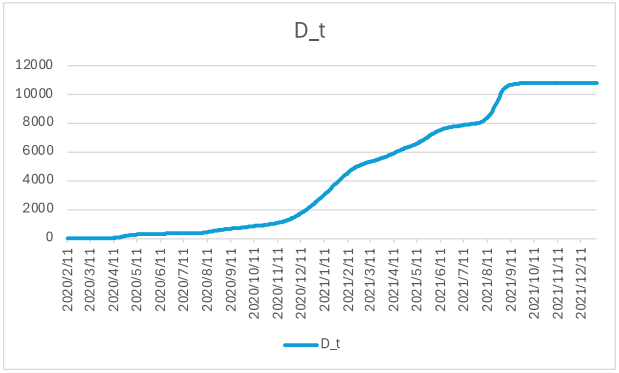

図 9　死亡者数の推移

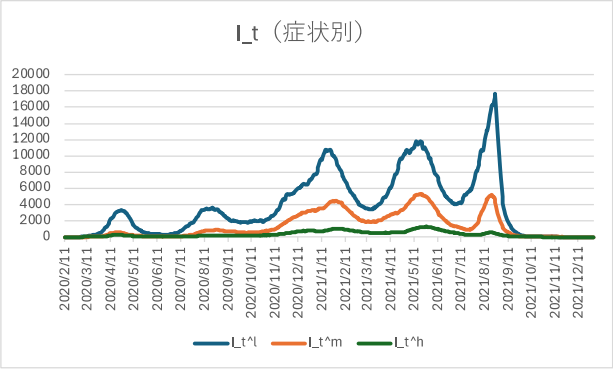

図 6　症状別入院者数の推移

学習分類子システム

乳がん患者の予後予測への学習分類子の適用

医療データとして METABRIC データセット[1]を使用する。METABRIC データセットは乳がん患者の臨床データと遺伝子データを含んでいる。データは原発性乳房腫瘍を持つ患者 2509 個で構成されており、数値データとカテゴリーデータが混在している。データセット内の特徴量の例としては、診断時の年齢、細胞の状態、乳がんに対しての治療の有無、腫瘍の状態などがある。本研究では予後予測を目的としているため、目的変数は全生存期間(Overall Survival (Months))を用いた。なお患者によって診断時期が異なるがデータの収集終了時期は同じため、診断時期が遅いために他の患者に比べて健康であっても全生存期間が短いことになっている患者も存在する。このような患者データを用いると適切に学習できなくなる可能性があるため、データ前処理で該当データを除外した。

クラス定義を表 3 に示す。今回は、予後を 10 年未満と 10 年以上の 2 クラスとして定義した。さらに、使用した特徴量を表 4 に示す。

表 3　クラス定義

クラス	予後
0	10 年未満
1	10 年以上

表 4　使用した特徴量

Age at Diagnosis
Type of Breast Surgery
Cellularity
Chemotherapy
Pam50 + Claudin-low subtype
ER Status
Neoplasm Histologic Grade
HER2 Status
Tumor Other Histologic Subtype
Hormone Therapy
Inferred Menopausal State
Integrative Cluster
Primary Tumor Laterality
Lymph nodes examined positive
Mutation Count
Nottingham prognostic index
Oncotree Code
PR Status
Radio Therapy
TMB(nonsynonymous)
Tumor Size
Tumor Stage

表 5　分類精度

LightGBM	XGBoost	Ours
0.6591	0.6248	0.6277

評価実験の結果、得られた分類精度を表 5 に示す。比較対象として決定木学習モデルである LightGBM および XGBoost による結果も示す。

さらに、学習分類子システムを適用した結果、得られたルールを表 6 に示す。Fitness は適合度であり、ルールが予測に貢献している程度を表している。Accuracy は正答率であり、ルールの条件に当てはまるデータの目的変数とルールの予測が一致している割合を表している。Match Count は、データの一致数であり、ルールに当てはまるデータの総数を表している。

表 6　得られたルール

Age at Diagnosis	32.8011999999999 94, 81.8788000000000 1
Neoplasm Histologic Grade	3
Lymph nodes examined positive	-3.375, 3.375
Mutation Count	-

	6.425000000000001, 26.425
TMB (nonsynonymous)	-8.98918420975、 35.13953828975
Tumor Size	-33.435, 61.435
Tumor Stage	1
Type of Breast Surgery_MASTECTOMY	TRUE
Pam50 + Claudin-low subtype_Normal	FALSE
Tumor Other Histologic Subtype_Medullary	FALSE
Tumor Other Histologic Subtype_Mucinous	FALSE
Tumor Other Histologic Subtype_Other	FALSE
Inferred Menopausal State_Pre	FALSE
Integrative Cluster_2	FALSE
Integrative Cluster_3	FALSE
Integrative Cluster_4ER+	FALSE
Integrative Cluster_5	FALSE
Integrative Cluster_6	FALSE
Integrative Cluster_7	FALSE
Integrative Cluster_9	FALSE
Oncotree Code_IMMC	FALSE
Oncotree Code_MDLC	FALSE
Overall Survival Status (Months)	0
Fitness	0.556119545
Accuracy	1
Match Count	20

D. 結論及び今後の展開

SIRD モデルによる分析

DPC データにより SIRD モデルのパラメータを推定し、感染症流行状況を把握するための方法論を開発した。具体的には、DPC データの変数のうち、入院日、退院日（死亡日）、レセプト電算コードのみから症状別に分類した SIRD モデルのパラメータ推定が可能となった。ここから、さらに変数として年齢、性別、既往症などの属性別に分析を深化することで、幅広い知見を得ることが可能である。

今後の方向性として、推定した各パラメータを用いたシミュレーションによる入院患者数や死亡者の予測、パラメータを変更した際の感度分析などが挙げられる。例えば、より強い感染症が発生した場合の想定として、βがどのくらい大きくなれば、どのくらい入院患者が増えて、どのくらい死亡者数が増えるのかが予測可能となる。または、治療方法が発展した場合の想定として、γがどのくらい大きくなれば、どのくらい死亡者数が減るのかが予測可能となる。

これらを実行することで、DPC データを用いた粒度の高い分析結果が得られ、パンデミックリスクの定量的評価とパンデミックリスク管理が可能となる。したがって、本研究の分析により、DPC データを活用した入院患者数や死亡者数の予測とリスク管理は、十分に可能であると考えられる。

ただし、DPC データには死亡退院情報が含まれているが、レセプト情報には死亡情報が含まれていない可能性がある。その他のパラメータについてはレセプト情報で代替可能であると考えられるが、死亡情報についての取り扱いが課題となろう。

DPC データに対する学習分類子システムの適用

DPC データでの入院日数の予測に学習分類子システムを適用することを検討している。これは、例えば、入院日数を1週間単位でクラス定義し、入院時のレセプトデータを特徴量として、入院時点や入院中の時

点において、その後の入院日数のクラスを予測するルールを生成するものである。クラス定義の例を表 7 に示す。さらに、特徴量定義の例と表 8 に示す。

入院日数の予測のためのルールを生成することによって、予測した入院日数の根拠を説明することができる。さらに、重要な特徴量を評価し、これによって、入院中の治療計画の立案に寄与することが期待できる。

表 7 クラス定義(例)

クラス	入院期間
0	1 日間～7 日間
1	8 日間～14 日間
2	14 日間以上

表 8 特徴量定義(例)

特徴量1	ICUbed 系資源利用の有無
特徴量2	呼吸器関連資源利用の有無
特徴量3	薬剤資源利用の有無

参考文献

[1] cBioPortal FOR CANCER GENOMICS, “Breast Cancer (METABRIC、 Nature 2012 & Nat Commun 2016)”

[2] W.O. Kermack and A.G. McKendrick, “A contribution to the mathematical theory of epidemics,” *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, Vol. 115, No. 772, pp. 700–721 (1927).

[3] Ji, C. and Jiang, D. “Threshold behavior of a stochastic SIR model,” *Applied Mathematical Modelling*, 38 (2014) 5067-5079.

[4] N. T. Dieu, D. H. Nguyen, N. H. Du, and G. Yin, “Classification of Asymptotic Behavior in a Stochastic SIR Model,” *SIAM Journal on Applied Dynamical Systems*, Vol. 15, pp. 1062-1084 (2016).

E. 健康危険情報

特になし

F. 研究発表

該当なし

G. 知的財産権の出願・登録状況

1. 特許取得

特になし

2. 実用新案登録

特になし

3. その他

特になし