様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ９ 日現在

機関番号：１４４０１
研究種目：基盤研究(B)（特設分野研究）
研究期間：2015～2019
課題番号：１５ＫＴ００１７

研究課題名（和文）医薬品候補化合物の副作用発症確率を予測する数理モデルの創成

研究課題名（英文）Modeling for Prediction of Serious Adverse Events Probabilities of Drug Candidates

研究代表者
高木　達也（TAKAGI, Tatsuya）

大阪大学・薬学研究科・教授

研究者番号：８０１４４５１７

交付決定額（研究期間全体）：（直接経費）　　11,200,000円

研究成果の概要（和文）：我々は、化学構造から、希少かつ重篤な有害事象を起こしやすい医薬品を予測する目的で、様々な機械学習法を適用した。化学構造記述子、物理化学記述子に加え、ATCコード記述子を導入することにより、幾つかの機械学習法（ロジスティック回帰、Support Vector Machine、Random Forest、k-nn法、ナイーブベイズ、ニューラルネットワーク法に加え、スタッキング法を用いた）で、血小板減少症、悪性症候群を起こしやすい医薬品の予測が「可能である」と言える程度の予測性を得た。この結果により、前臨床の段階での開発の中断や、市販医薬品に対する注意を行うことができる。

研究成果の学術的意義や社会的意義
希少有害事象はこれまで、ともすれば、見過ごされがちであった。理由としては、極めて珍しいこと、殆どの医薬品に見られるので、これを咎めると、医薬品が市場に無くなってしまうことが考えられる。しかしながら中には命に関わる重篤なものもあり、かつ、一部の医薬品で高確率で見られ、回収に繋がったものもある事から、このまま見過ごすことはできないと考えた。今回の結果により、より高確率で希少重篤有害事象を引き起こす医薬品の化学的特徴が明らかとなり、将来的には、医薬品開発の様々な段階で援用され、服薬指導時に注意を促すための材料となり、有害事象から命を落としたり、後遺症に悩む人々の現出を予防できることが期待される。

研究成果の概要（英文）：We tried to make some models for predicting drugs which show relatively high probabilities of rare and severe adverse events using chemical structure information and machine learning methods. As results, several machine learning methods (Logistic Regression, Random Forest, Support Vector Machine, Artificial Neural Network, etc) with Stacking method showed enough prediction abilities when ATC code was introduced for malignant syndrome and thrombocytopenia. These results can be utilized as a powerful tool for Drug Development and drug administration guidance.

研究分野：計量薬学

キーワード：有害事象　機械学習　ATCコード　Stacking　化学記述子

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

従来の安全対策は個々の医薬品に着目し、医薬品毎に発生した有害事象を収集・評価し、臨床現場に添付文書の改訂等により注意喚起する「警報発信型」「事後対応型」が中心であった。しかし、有害事象は原疾患とは異なる臓器で発現する可能性があることや、多くの重篤な有害事象は発生頻度が低く、臨床現場において医療関係者が遭遇する機会が少ないことから、場合によっては有害事象の発見が遅れ、重大化することがある。そこで厚生労働省では従来の安全対策に加え、医薬品の使用により発生する有害事象疾患に着目した対策の整備を行うとともに、有害事象の発生機序を解明する研究を推進することにより、「予測・予防型」の安全対策への転換を図ろうとしている。

医薬品の重大な有害事象が発症すれば、患者の QOL (quality of life) を損ねるだけでなく、二次的医療費の発生、医薬品の回収作業など、国と製薬会社の経済的損失も大きい。発生頻度が低く、重度の後遺症が残るような重大な有害事象を発症する確率が相対的に高い医薬品を市販される以前に発見することができれば、社会的な影響がきわめて深刻である薬害の発症確率を低下させることができ、医療費の削減などを含む、経済的な波及効果も期待できる。

重大な有害事象を誘発する恐れのある市販前の医薬品候補化合物を発見できるタイミングは、大別すると、1) 医薬品候補化合物の発見、合成の段階、2) 非臨床試験 (動物実験) 通過後、臨床試験開始前の段階、3) 臨床試験の段階である。本研究の目的は、医薬品候補化合物を段階 2) の時点で、希少かつ重大な有害事象を発症する確率を化学構造から予測するモデルを構築し、希少かつ重大な有害事象を発症する医薬品の特徴を発見することである。

２．研究の目的

医薬品の希少かつ重大な有害事象における詳細な発症メカニズムは未だほとんど解明されておらず、それら全てついて予測し、それらの原因となる薬の化学構造の特徴を抽出することは非常に困難である。その予測、または化学的特徴を抽出することができたならば、薬害を発症する確率を減少させることができ、社会的に大きな意義をもたらすことができる。また、電子カルテの導入が進むことで、医療用データ内の副作用報告データが増大し、現状よりもさらに大規模なデータベースを構築することが出来る。このデータベースから機械学習などの解析を行うことによって、有益な情報を抽出して知識化することの重要性が増すことが予想できる。そこで、本章では機械学習で予測しやすい有害事象や、化学構造の特徴を抽出しやすい有害事象を探索し、その有害事象に対して検討を行うことを目的とした。

３．研究の方法

医薬品の希少かつ重大な有害事象と化学構造との関連性を予測するにあたり、医薬品医療機器総合機構 (Pharmaceutical and Medical Devices Agency: PMDA) が無償で公開している副作用報告のデータベース JADER (Japanese Adverse Drug Event Report)を使用した。JADER に報告された有害事象の種類は 6895 種であったが、対象とする有害事象を表１の 7 種類に絞った。

表 １．対象有害事象

略語	有害事象名
Y1	スティーブンス・ジョンソン症候群(SJS)または皮膚粘膜眼症候群†
Y2	横紋筋融解症
Y3	白質脳症
Y4	悪性症候群
Y5	中毒性表皮壊死融解症または表皮壊死†
Y6	QT 延長症候群または心電図ＱＴ延長†
Y7	可逆性後白質脳症症候群
Y8	血小板減少症

†表記が異なるが、実質同じであると考えられる有害事象をまとめて処理した。

被疑薬の化学構造と対象有害事情の関係を解析するにあたり、該当医薬品の化学構造を入手し、その構造から物理、化学的な特徴（e.g. 分子量、極性表面積など）に数値化する必要がある。医薬品は構造が複雑で、3 次元コンフォメーションを求めるのは困難であるため、本研究では 2 次元構造から計算できる特徴（以下、記述子）（e.g. 二次元の部分構造など）を算出した。その際、2 次元構造でも入手不可能な医薬品及び合剤などを解析対象から除外した。対象被疑薬として報告されている医薬品の種類は合計で 1301 種類存在した。この医薬品の中で化学構造を入手できた医薬品数は 863 種類であり、合剤を除去すると 790 種類となった。さらに被疑薬が１つしかない症例に絞ると、医薬品の数は 577 種類となった。

解析手法として 7 種類の機械学習法を使用した。モデルの作成にはプログラミング言語 Python 向けの機械学習ライブラリである scikit-learn-0.18 を用いた。

LR	ロジスティック回帰（Logistic Regression）法
SVM	サポートベクターマシン（Support Vector Machine）法
KNN	K 近傍（k-Nearest Neighbor）法
DT	決定木（Decision Tree）法
RF	ランダムフォレスト（Random Forest）
GB	勾配ブースティング（Gradient Boosting）法
MV	多数決（Majority Voting）法
NB	ナイーブベイズ（Naïve Bayes）法
Stack	これらの幾つかを組み合わせた Stacking 法

機械学習の各手法のハイパーパラメータを調整するために、グリッドサーチを用いた。グリッドサーチでは与えられたハイパーパラメータの組み合わせの全てに対して網羅的な探索を行ない、最適なハイパーパラメータの組み合わせを探索する。モデルの性能評価には k 分割交差検証(k-fold cross-validation, k-CV) を用いた。最も AUC が高かったモデルをベストモデルとし、そのときのハイパーパラメータを最適なハイパーパラメータとした。k-CV では、まずトレーニングデータセットをランダムに k 個に分割する。そのうちの k-1 個のデータを用いてモデルの学習を行ない、残り 1 個のデータを予測する。これを分割された k 個のデータセットに対して繰り返すことで、全てのデータセットに対する予測を得られる。この予測値を用いてモデルの予測精度 AUC を評価する。本研究では k=10 とし、10-CV を使用した。

４．研究成果

１）血小板減少症

血小板減少症では、NB、KNN、RF、LR の 4 手法を組み合わせ、Stacking を行うことにより、良好な結果(ACC, AUC 共に 75%以上が目標)を得た（表 2)。この際、予測性に重要な化学記述子として得られたものを表 3 に示す。水素結合や分子の極性に関する記述子、a_acc、a_don、a_donacc、PEOE_VSA+2 などと共に、化学反応性が重要な記述子として登場している事が興味深い。

表 2．各手法を組み合わせして stacking を適用した際の結果

	Stacking
ACC test	0.77
ACC train	0.87
AUC test	0.77
AUC train	0.87

今回、医薬品のリスク評価の出発点を踏まえて、血小板減少症に対する医薬品有害事象発生の予測モデルの開発を行った。医薬品の化学物質の構造的な特性から記述子を算出した後、モデルを学習させる前に、カイ二乗検定によって説明変数の次元を削減し、モデルの学習を容易にした。その後、機械学習の手法を組み合わせることによって、良好な予測精度のモデルを構築した。本研究で使用した各手法は、予測対象（血小板減少症を起こす医薬品）の化合物に対して、予測精度に差が見られ、お互いに補い合うような組み合わせで、各手法を単独で使用した際のモデルよりも、予測精度が高くなった。今後も、より多くの有害事象についての患者医薬品使用情報を収集し、それらを基に随時モデルの再構築を行うことが、有害事象が未知であるデータに対してより良好な予測や、国際医薬品安全情報の収集、管理の迅速かつ高効率化に伴う予測結果の信頼性を向上させる上で重要になると考えられる。

表 3　血小板減少症の予測に重要な化学記述子

ast_fraglike	Astex のルールに関する記述子
a_acc	水素結合アクセプター原子の数（酸性原子を数えず、水素結合供与体と受容体（例えば-OH）の両方を数える)。
a_don	水素結合供与原子の数（基本原子を数えず、水素結合供与体と受容体（例えば-OH）の両方を数える)。
a_donacc	水素結合の供与原子とアクセプター原子の数の和
vsa_other	“その他”とタイプされた原子の VDW 表面積の合計の近似値。
Lip_violation	Lipinski's Rule of Five の違反回数
SMR_VSA1	R_i が（0.11,0.26]にあるような vi の和。
PEOE_VSA+2	PEOE 部分電荷計算公式により、範囲[0.10,0.15]にある qi と vi の和。qi を上記で定義した原子 i の部分電荷とする。 vi を原子 i の vdW 表面積とする（接続テーブル近似によって計算される）
Reactive	反応性基を持つかどうか

なお、今回行った計算は、医薬品有害事象の予測分野においては、巨大なシステムのほんの一部であり、アルゴリズムの予測精度および医薬品特性情報の統合にはまだまだ改善の余地がある。

医薬品有害事象の発生は、複雑な生理学的メカニズムとして、関連の深い因子について、医薬品の相互作用や原疾患からの影響など、解釈が困難なものが存在するため、今後ともさらなる検討する価値があると考えられる。目的に応じて適切な記述子及びモデルを構築・採用することが望ましい。

２）悪性症候群

既知情報として、悪性症候群を引き起こす医薬品には神経系薬が多いことから、全医薬品を対象とするモデルに加え神経系薬のみを対象とするモデル、神経系薬以外の医薬品を対象とするモデルも構築した。この際、医薬品の種類を分類する方法として解剖治療化学分類法（Anatomical Therapeutic Chemical Classification System: ATC コード）を用いた。ATC コードは世界保健機関（World Health Organization, WHO）の医薬品統計法共同研究センター（Collaborating Centre for Drug Statistics Methodology）によって管理されており、1976 年に発行が開始された。これらの記述子を用い、k-fold CV で変数選択した場合の結果を表 4 に示す。

表 4　各手法の悪性症候群に対する予測性能

		LR	RF	SVM	kNN
全医薬品の結果	**ACC**	0.55±0.11	0.87±0.02	0.97±0.00	0.62±0.13
	AUC	0.66±0.07	0.90±0.04	0.92±0.04	0.72±0.05
	TPR	0.63±0.16	0.76±0.09	0.42±0.07	0.65±0.15
神経系医薬品の結果	**ACC**	0.63±0.11	0.83±0.05	0.84±0.03	0.62±0.07
	AUC	0.70±0.07	0.90±0.03	0.89±0.03	0.66±0.06
	TPR	0.47±0.16	0.79±0.09	0.70±0.07	0.53±0.12
神経系の医薬品を除いた結果	**ACC**	0.68±0.20	0.63±0.06	0.96±0.01	0.79±0.08
	AUC	0.56±0.07	0.72±0.06	0.74±0.13	0.65±0.10
	TPR	0.36±0.16	0.72±0.15	0.04±0.06	0.35±0.18

表 4 より、神経系医薬品を除かなくとも十分な精度の予測が可能であることが示された。全ての評価指標が良好であった RF の利用が最も妥当だと考える。また、LR、kNN はすべてのデータセットにおいて標準偏差が大きく、不安定なモデルであった。神経系の医薬品を除いたモデルは他のデータセットを用いたモデルに比べ AUC、TPR が低かった。これは、目的変数の値が 1（被疑薬である）医薬品の半数以上が神経系薬であるため被疑薬数が少ないことに加え、被疑薬が抗炎症薬や抗菌薬など様々であり、被疑薬に共通する特徴の抽出が難しいためであると考えられる。次に、悪性症候群発症の判別に重要な特徴量を抽出した。最も利用が妥当であると考えられる RF モデルにおける重要度を用いた。まず、データセット 1 を用いた RF モデルで重要度の大きさが上位 10 位までに 10 回中 5 回以上選ばれた記述子を抽出し、表 5 にまとめた。

表５. RF における重要な記述子と上位 10 位以内に選ばれた回数

記述子	回数
AMID_O	10
159	8
140	7
SpMAD_A	6
ATSC3i	6
SMR_VSA6	6
nBase	5
AATSC1s	5

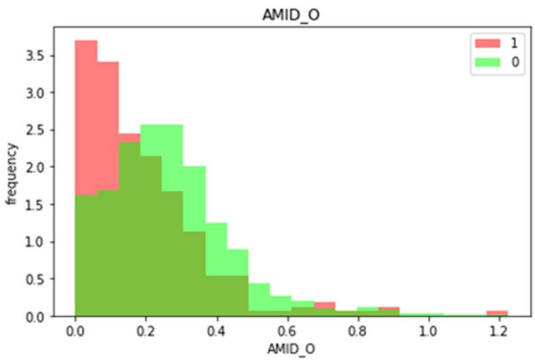

図１. 目的変数別の AMID_O ヒストグラム

AMID_O は酸素原子に結合している原子の ID 番号の平均値で、酸素原子に重い原子が多く結合しているほど（一般的には酸素原子、特に炭素鎖中の酸素原子が多いほど）大きな値となる記述子で、10 回すべてで選択された。また、重要度は 10 回中 8 回が 1 位であった。ID は結合の重みを用いた分子骨格のラベルであり、物理化学的解釈が容易ではないが、酸素原子数が０であれば AMID_O も 0 である。

目的変数と AMID_O の関係性について調べるため、データセットを目的変数の値が 0 と 1 の医薬品に分け、それぞれのヒストグラムを作成し重ね合わせた。そのヒストグラムを図 10 に示す。横軸が AMID_O の値、縦軸が相対的な医薬品の数を示している。なお、ヒストグラム作成にあたり各ヒストグラムの面積を 1 に標準化した。図 10 を確認すると、AMID_O が低い値のときに目的変数の値が 1 の医薬品が多く存在し、逆に AMID_O が高い値のときに目的変数「0」の医薬品が多く存在していることが分かる。このことから AMID_O が悪性症候群を引き起こすかどうかに深い関係があることが推察され、AMID_O の値が低い場合、特に 0 原子を 1 つも含まない医薬品を扱

うときは注意が必要であることを示唆している。
また、フィンガープリント記述子の 159 が 10 回中 8 回、140 が 10 回中 7 回選択された。159 は O 原子が 2 つ以上存在するかどうか、140 は O 原子が 4 つ以上存在するかどうかを示す記述子であり、ともに O 原子数に関わる記述子である。
今回、FDA、PMDA が報告している副作用報告データベースを利用し、機械学習による悪性症候群発症の予測を行った。その結果、全医薬品を含むデータセット 1、神経系薬のみを含むデータセット 2 を用いて、ACC、TPR が 0.8 程度、AUC が 0.9 程度という非常に高い精度の予測モデルを構築することが出来た。この予測モデルを用いることで悪性症候群発症の抑制に貢献することが期待される。
そして、悪性症候群を惹き起こす医薬品の特徴の発見を目的に更なる解析を行った。その結果、特にフィンガープリント記述子 140、159 と AMID_O が悪性症候群に関連する可能性が示唆された。フィンガープリント記述子が示した悪性症候群に関連すると推察される化学構造を、医薬品設計の段階で考慮することで悪性症候群の被害抑制に貢献できると期待される。しかし、記述子等の特徴量を用いて構築した予測モデルは一般的に共通な問題点として、変数の説明が困難である。記述子は数式的に明確だとしても、理解しにくいものが多く、本研究も現段階では、予測性能に注力したが、今後、検出された重要な特徴量をベースに、フィンガープリント記述子以外の記述子についても悪性症候群を惹き起こす医薬品の構造や物理化学的な要因の有無を入念に検討することで、更なる医薬品設計における指針を示すことができると考えられる。
また、本研究で使用した副作用報告データベースは自発報告に基づくものであり、頻繁に用いられている医薬品は有害事象の発症報告数が多くなる傾向がある。医薬品の使用状況、使用頻度の情報は得ることができないため、このバイアスは自発報告データベースにおいて既知のものである。

３）その他の希少重篤有害事象
その他の希少重篤有害事象に関し、種々の機械学習法で予測を試みた結果を表 5 に示した。ここで Y3’は、データの偏りを防ぐために、Y3：白質脳症と Y7：可逆性後白質脳症症候群を同一機序によるものと仮定し、合同したものである。

表 5．有害事象と各手法におけるテストセットの AUC の中央値

AUC(テストセット)	LR1	LR2	SVM	KNN	DT	RF	GB	MV
Y1	0.65	0.65	0.65	0.63	0.56	0.66	0.64	0.67
Y2	0.52	0.53	0.52	0.55	0.52	0.53	0.58	0.56
Y3’	0.66	0.65	0.72	0.57	0.71	0.70	0.69	0.71
Y5	0.63	0.65	0.68	0.68	0.61	0.66	0.68	0.69
Y6	0.66	0.68	0.66	0.60	0.65	0.65	0.66	0.68

白質脳症系の 2 つ（Y3、Y7）は比較的良好な予測性を得ることができた。また、多数決法によれば、SJS、TEN、QT 延長に関しても、まずまずの予測性を得ることができている。ただ、多数決法の弱点は、重要な記述子が各手法で異なるため、どのような化学構造が有害事象に結び付くかがわからない点にある。今後、多数決法、Stacking 法を中心に、方法論的解決も重要になってこよう。白質脳症群の予測性が向上し、悪性症候群や血小板減少症に匹敵する結果を得た。多数決法でなく、SVM で良好な結果が得られたことの重要性は大きく、今後の解析に活かすことができよう。

今回の研究では医薬品の化学構造だけで有害事象を予測しようとしたが、副作用報告データベースには患者の年齢や性別、体重、投与経路、原疾患情報など様々な情報が報告されている。これらの情報を予測モデルに取り入れられたならば更なる予測精度の改善につながる可能性ある。また、今回の研究では併用薬については全く考慮に入れていないが、併用薬として報告の有無で説明変数を設定してモデルに組み込むことが可能である。特に SJS や TEN、横紋筋融解症など、今回の解析で十分な結果が得られなかった有害事象に関しては、併用薬の重要性も指摘されており、今後はそれも視野に入れて解析する必要があろう（言うは易し行うは難しで、データベースの構成から考え直す必要があるため、産官学の協力が必要になろうが、容易ではない）。さらに、悪性症候群の発症にはドーパミン神経系が関連していると考えられていることから、ドーパミン受容体との結合などの標的タンパクについての情報を説明変数としてモデルに組み込むことで予測精度の向上につながる可能性がある。本研究計画は、当初予定通りの結論を導き出せたと自負しているが、今後行わなければならない事も多い。しかしながら、有害事象に関する研究は、通常、薬学に従事する者ができれば避けたい分野でもあり、今まで十分な研究がなされてこなかったことは、自らも含めて反省材料である。本研究成果が、今後の有害事象に関する研究の発展に寄与することができれば、この上ない幸いである。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計5件（うち招待講演　1件／うち国際学会　2件）

１．発表者名 Ni Tao、高木達也、日比孝之、望月麻衣、森脇寛智、田雨時
２．発表標題 機械学習を用いた医薬品の有害事象の予測モデル構築の検討
３．学会等名 第46回構造活性相関シンポジウム
４．発表年 2018年

１．発表者名 Yushi-Tian, Hirotomo Moriwaki, Hiroaki Moriuchi, Satoshi Aoki, Nobuki Takayama, Norihito Kawashita, Takayuki Hibi, Tatsuya Takagi
２．発表標題 Prediction of Serious Adverse Events Using Machine Learning
３．学会等名 19th International Conference on Medicinal Chemistry and Multi Targeted Drug Delivery（招待講演）（国際学会）
４．発表年 2018年

１．発表者名 Tatsuya TAKAGI
２．発表標題 Prediction of Serious Adverse Events Using Machine Learning
３．学会等名 11th Asian Federation for Medicinal Chemistry's International Medicinal Chemistry Symposium（国際学会）
４．発表年 2017年

１．発表者名 Hiroaki MORIUCHI
２．発表標題 Constructing prediction models of adverse drug reactions using machine learning
３．学会等名 44th Symposium on Structure-Activity Relationships
４．発表年 2016年

１．発表者名
望月麻衣、福戸康平、田雨時、高木達也

２．発表標題
機械学習を用いた悪性症候群予測モデルの構築

３．学会等名
第47回構造活性相関シンポジウム

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
この他、"21st International Conference on Medicinal Chemistry and Multi Targeted Drug Delivery, Singapore, Dec 2019 (Key Note Lecture)", "PSWC2020, Montreal, May 2020"で発表予定でしたが、前者は会場の問題で、後者はCOVID-19パンデミックの影響で、何れも延期になっています。この学会でのディスカッションの結果を元に論文投稿を考えていたので、これも延期になっています。

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	日比　孝之 (Hibi Takayuki) (80181113)	大阪大学・情報科学研究科・教授 (14401)	
研究分担者	岡本　晃典 (Okamoto Kousuke) (70437309)	北陸大学・薬学部・講師 (33304)	
研究分担者	川下　理日人 (Kawashita Norihito) (00423111)	大阪大学・薬学研究科・助教 (14401)	

６．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	田　雨時 (Tian Yui-Shi) (60761252)	大阪大学・薬学研究科・助教 (14401)	
連携研究者	小田中　紳二 (Odanaka Shinji) (20324858)	大阪大学・サイバーメディアセンター・教授 (14401)	2017年2月12日逝去。所属等は当時のものです。