厚生労働科学研究費補助金
政策科学総合研究事業（臨床研究等ICT基盤構築・人口知能実装研究事業）

総括研究報告書

薬局ヒヤリハット事例に対する安全管理対策評価に関するAI開発

研究代表者　岡本　里香
京都大学大学院医学研究科　人間健康科学系専攻　ビッグデータ医科学系分野
特定准教授

研究要旨

独立行政法人医薬品医療機器総合機構（以下，PMDA）では，公益財団法人日本医療機能評価機構（以下，評価機構）が薬局ヒヤリ・ハット事例収集・分析事業（以下，本事業）に基づき収集・分析・公表した「ヒヤリ・ハット事例」に対し，医薬品の名称・包装等の観点から安全対策を講じる必要がないか検討を行っている．本事業は，全国の薬局からの事例を収集・分析し，薬局における医療安全対策に有用な情報を共有するなど，医療安全対策の一層の推進を図ることを目的として行われている．収集される事例は，薬局で発生した調剤や疑義照会等に関するヒヤリ・ハット事例であるが，例えば，調剤に関する事例のうち，薬剤の名称の視覚的，音韻的な類似に起因したことで薬剤取違えた等の事例の場合，PMDA では製造販売会社に対し，薬剤の取違えを防ぐための注意喚起の必要性等について指導する，といった医薬品の物的要因に対する安全管理対策の評価・検討している．しかしながら，評価機構の薬局ヒヤリ・ハット事例報告数が急激に増加しており，令和 2 年前期の評価機構の報告数（令和元年 5 月から 12 月までの 8 カ月分）は 9 万 7 千超であり，この中から対策の必要性を検討しなければならない事例を抽出するだけでも，かなりの労力と言える．

PMDA では「薬局ヒヤリ・ハット事例」に対して，現在，人による目で，評価を 5 段階に分類し，安全対策の必要な事例を抽出している．本研究では，この分類を人工知能（以下，AI）が行えるようにすることを目的としている．過去の PMDA における評価では，評価機構が公表したの報告に対して，対策を検討する事例は，評価機構が公表した報告の約 0.5%程度であり，ほとんどがヒューマンエラーや情報不足の事例であることから，これらを 1 次スクリーニングとして AI で分類するだけでも，PMDA における労力は軽減され，対策を検討しなければならない事例に注力して安全管理対策を講じることが可能となる．

（研究分担者）
中津井　雅彦
山口大学大学院医学系研究科・医学部附属病院AIシステム医学・医療研究教育センター・特命教授
小島　諒介
京都大学大学院医学研究科・人間健康科学系専攻・ビッグデータ医科学分野・特定講師

A. 研究目的

評価機構が公表するヒヤリ・ハット事例は年々増加傾向にあり，PMDA においてこれらに対する安全対策要否の検討・評価は負担が大きくなっている．（図 1）

（図 1）

薬局ヒヤリ･ハット報告数と評価対象事例数の推移

	H25	H26	H27	H28	H29	H30 前期	H30 後期	R1 前期	R1 後期	R2 前期	R2 後期
評価機構の報告数	6,497	5,387	4,923	4,865	5,151	3,457	20,684 (*1)	34,373 (*2)	92,741 (*3)	97,707 (*4)	
PMDA評価対象	3,670	3,019	2,959	2,893	3,356	2,424	評価対象事例の抽出方法の検討のため評価事例なし	1,255	2,369	2,416	
対策検討事例数	0	6	12	12	20	11	-	57	122	161	
評価対象事例に占める割合	0.2%	0.2%	0.4%	0.4%	0.6%	0.5%	-	4.5%	5.1%	6.7%	

事例のうち
✓ 規格・剤形間違い
✓ 薬剤取違え
✓ その他及び疑義照会

調査の結果，対策を講じたものの事例数は非常に少なく，ほとんどが情報不足，ヒューマンエラーの事例である

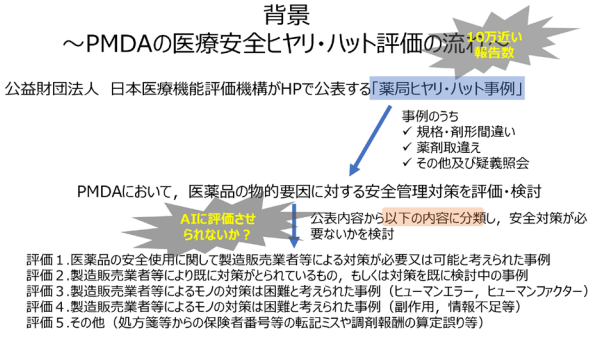
(*1) 増加背景として，H27年10月公表「患者のための薬局ビジョン」，H29年3月公表「「患者のための薬局ビジョン」実現のためのアクションプラン検討委員会報告書」の影響，さらに，平成30年度診療報酬改定における地域支援体制加算の新設が本事業への参加と報告の増加に影響していると考えられる。
(*2) これまで6月分の報告を対象としてきたが，H30年度後期のPMDA評価をスキップしたため，9カ月間分の報告対象（評価機構報告H30年1～9月分）
(*3) 7カ月間分の報告対象（評価機構報告H30年10～H31年4月分）
(*4) 8カ月間分の報告対象（評価機構報告R1年5～12月分）

PMDA における評価は，評価機構が公表する事例のうち，「規格・剤形間違い」「薬剤取違え」「その他及び疑義照会」として報告された事例を抽出し，各事例の内容を確認・評価し，次の評価 1～5 に分類している．

評価 1：医薬品の安全使用に関して製造販売業者等による対策が必要又は可能と考えられた事例
評価 2：製造販売業者等により既に対策がとられているもの，もしくは対策を既に検討中の事例
評価 3：製造販売業者等によるモノの対策は困難と考えられた事例（ヒューマンエラー，ヒューマンファクター）
評価 4：製造販売業者等によるモノの対策は困難と考えられた事例（副作用，情報不足等）
評価 5：その他（処方箋等からの保険者番号等の転記ミスや調剤報酬の算定誤り等）

本研究では，PMDA における評価 1～5 の分類を AI に実施させることにより，PMDA における本業務の負担を軽減し，ヒヤリ・ハット事例の増加に対しても一貫性のある評価を行うことを目的とする．（図 2）

（図 2）

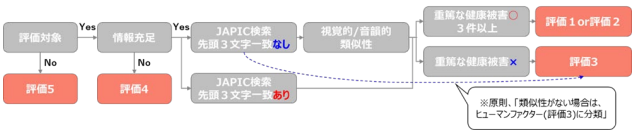


本研究の先行研究として実施した探索的研究において，本モデル開発での課題・問題点を検討した．

図 1 に示す，PMDA の「対策検討事例数」は，PMDA による評価 1 あるいは 2 に該当する事例である．評価機構の報告数に対して，令和元年前期以降は，評価機構の報告数の増加等で事例は増加しているものの，平成 30 年前期まででは 20 以下と，対策を検討した事例は非常に少ない．このため，「対策検討事例数」は学習データとしては数が不十分であり，評価 1 か評価 2 かを分類することはできないという問題がある．また，評価 1 及び評価 2 は安全管理対策が必要な事例であるため，モデルによる分類が誤ってこれらを評価 3，4 あるいは 5 の低リスクに分類してしまうと，安全対策が必要な事例を見逃すことになるという課題が挙げられた．

そこで，本研究におけるモデル開発では，分類は，「評価 1 及び 2」「評価 3」「評価 4」「評価 5」の 4 分類とすること，及び PMDA の評価 3～5 の事例がモデルで「評価 1 及び 2」に分類されることを許容することとした．PMDA の評価ルールを図 3 に示す評価スキームとし，各分類に際して必要な学習データ等を用いて，機械学習を実施することにより，評価分類モデルを作成し，アルゴリズムを検討する．

図 3：評価スキーム



B. 研究方法

PMDA において，評価機構の HP から「規格・剤形間違い」「薬剤取違え」「その他及び疑義照会」として報告されている事例を CSV 出力し，評価 1～5 に分類し，安全対策の要否を評価・検討した結果が PMDA の HP に報告されている．我々は，当該 PMDA が公表する「評価機構公表内容」＋「PMDA 評価結果」のデータ（以下，PMDA 公表データ）を入手し，これを対象として，評価分類モデルを作成およびモデルの精度向上を行う．（図 3）

＜方法＞

1st Step：評価分類モデル作成

評価機構公表内容の項目「事例の内容」「背景・要因」「改善策」「発生要因」のテキスト記述を特徴量化し，「関連する医薬品の情報」の項に対しては記載されている薬品名を抽出し，各薬品情報を参照できるようにし，評価分類モデルを作成する．

表層（販売名）の類似性は次の複数の基準で

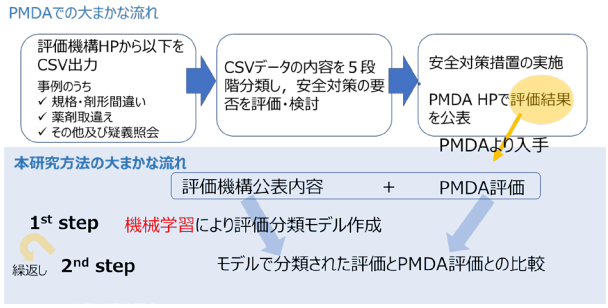
名称の類似度を計算した.

- 先頭 3 文字の一致の有無
- 先頭 5 文字の一致の有無
- 完全一致の有無
- 文字種の一致数
- 薬剤名称の平均長
- 最長 1 致文字数
- 最長 1 致文字数/名称の平均長
- 編集距離
- 編集距離/平均長
- ゲシュタルトパターンマッチング

2nd Step：評価分類モデルの精度確認
作成した評価分類モデルで，1st Step で使用していない PMDA 公表データを分類した結果と PMDA 評価結果とを比較する.

以上の 1stStep～2nd Step を繰り返すことにより，モデルの分類の精度を向上させる.

（図 4）



これまでに 1stStep～2nd Step の繰り返しは 4 回実施し，各回での評価機構データ，学習データセット，自然言語処理手法を図 5 に示す．また，4 回目では，サンプリング方法として Random Under Sampling と Random Over Sampling との比較も実施した.

（図 5）

	評価機構データ	学習データセット	自然言語処理手法	内容
1	H30年度前期まで	・各薬剤の薬効リスト ・類似薬剤名リスト	word2vec	試行
2	H30年度前期まで	上記１に以下を追加 ・薬剤知識ベース(KEGG)を用いて別名称追加 ・KEGG階層追加	word2vec	試行
3	H30年度後期及びR１前期	これまでの学習データセット	word2vec	試行及び事例分析
4	H30年度前期までの全データ	上記２に以下を追加 ・KEGGから「劇薬」追加	Universal Sentence Encoder（USE）→word2vecと比較	精度向上策の試行結果の確認

KEGG（Kyoto Encyclopedia of Genes and Genomes）は、遺伝子やタンパク質、代謝、シグナル伝達などの分子間ネットワークに関する情報を統合したデータベース

（倫理面への配慮）
本研究で使用するデータは個人情報を含まず，公表済のものを使用しているため，倫理面での配慮は特にない.

C. 研究結果

作成したモデル分類について，1stStep～2nd Step を繰り返し，モデルによる分類の精度は，precision，recall，F1-score を評価指標とした.

まず，Random Over Sampling の場合と Random Under Sampling の場合とで検出したところ，図 6 で示すように，全体として，Random Over Sampling の方が Random Under Sampling よりも精度がよい結果が得られた．また，Random Over Sampling の場合に，評価 1 及び 2 の分類について，precision は学習データセットを追加することにより 0.16→0.27→0.66→0.71 と高まる結果が得られた．しかし，同じく評価 1 及び 2 の分類における recall については，「テキスト情報」を学習させると，「表層類似度+一般名」→「表層類似度+一般名+テキスト情報」は 0.86→0.17 となり，低くなる結果となった．一方，評価 4 については，「テキスト情報」を追加することにより，recall は「表層類似度+一般名」→「表層類似度+一般名+テキスト情報」は 0.20→0.91 と高まった.

（図 6）

組み合わせ	評価	Over sampling precision	Over sampling recall	Over sampling F1-score	Under sampling precision	Under sampling recall	Under sampling F1-score
表層類似度	1及び2	0.16	0.83	0.27	0.04	0.85	0.07
	3	1.00	0.78	0.80	1.00	0.66	0.80
	4	0.76	0.20	0.29	0.62	0.19	0.29
	5	0.23	1.00	0.37	0.23	1.00	0.37
表層類似度＋**一般名**	1及び2	0.27	0.86	0.41	0.06	0.92	0.11
	3	1.00	0.80	0.89	1.00	0.70	0.82
	4	0.77	0.20	0.32	0.61	0.19	0.28
	5	0.23	1.00	0.37	0.23	1.00	0.37
表層類似度＋一般名＋**テキスト情報**	1及び2	0.66	0.17	0.27	0.03	0.91	0.06
	3	0.98	0.97	0.98	0.99	0.70	0.82
	4	0.78	0.91	0.84	0.58	0.85	0.69
	5	0.89	0.88	0.89	0.52	0.88	0.65
表層類似度＋一般名＋テキスト情報＋**KEGG情報**	1及び2	0.71	0.19	0.30	0.03	0.89	0.06
	3	0.99	0.98	0.98	0.99	0.72	0.83
	4	0.86	0.93	0.89	0.59	0.85	0.70
	5	0.91	0.88	0.89	0.53	0.88	0.65

＊ここでの一般名は，薬剤知識ベース（KEGG）を用いた別名称のことであり，製品名だけでなく，一般的名称に対しても対応し，当該対応により，後発医薬品の名称にも対応可能とする.
＊テキスト情報とは，背景情報である.

さらに，評価 1 及び 2 の分類精度を高めるために，規制区分情報として「劇薬」か否かの情報を学習データに追加した（図 7）．その結果，図 6 との結果と比較すると，Random Under Sampling では「劇薬情報」の追加による精度の変化は認められなかったが，Random Over Sampling での評価 1 及び 2 の precision は，「表層類似度」→「表層類似度+劇薬情報」が，0.16→0.22，「表層類似度+一般名」→「表層類似度+一般名+劇薬情報」0.27→0.32，「表層類似度+一般名+テキスト情報」→「表層類似度+一般名+テキスト情報+劇薬情報」0.66→0.69，「表層類似度+一般名+テキスト情報+KEGG 情報」→「表層類似度+一般名+テキスト情報+KEGG 情報+劇薬情報」0.71→0.74 と高まる結果が得られた.

（図 7）

組み合わせ	評価	Over sampling precision	Over sampling recall	Over sampling F1-score	Under sampling precision	Under sampling recall	Under sampling F1-score
表層類似度+劇薬情報	1及び2	0.22	0.83	0.34	0.04	0.87	0.08
	3	0.99	0.83	0.90	0.99	0.69	0.82
	4	0.76	0.20	0.32	0.65	0.19	0.29
	5	0.25	0.96	0.40	0.24	0.98	0.39
表層類似度+一般名+劇薬情報	1及び2	0.32	0.85	0.46	0.07	0.92	0.13
	3	0.99	0.84	0.91	1.00	0.74	0.85
	4	0.77	0.20	0.31	0.64	0.19	0.29
	5	0.25	0.96	0.40	0.24	0.98	0.39
表層類似度+一般名+テキスト情報+劇薬情報	1及び2	0.69	0.15	0.25	0.03	0.91	0.06
	3	0.98	0.97	0.98	0.99	0.70	0.82
	4	0.79	0.92	0.85	0.58	0.85	0.69
	5	0.89	0.89	0.89	0.52	0.88	0.65
表層類似度+一般名+テキスト情報+KEGG情報+劇薬情報	1及び2	0.74	0.19	0.30	0.03	0.91	0.06
	3	0.99	0.98	0.98	0.99	0.71	0.83
	4	0.86	0.93	0.89	0.59	0.85	0.70
	5	0.91	0.88	0.89	0.52	0.87	0.65

テキスト情報をベクトル化するための自然言語処理手法として，word2vec と USE とを比較検討した．当該比較は，図 3 の評価スキームに対して，次の NoStep アプローチと Step アプローチの 2 種類の各アプローチにおける word2vec と USE を用いた場合の「評価 1 及び 2」「評価 3」「評価 4」「評価 5」の 4 分類結果を F1-score および macro-F1-score で比較した（図 8）．

- NoStep アプローチ:

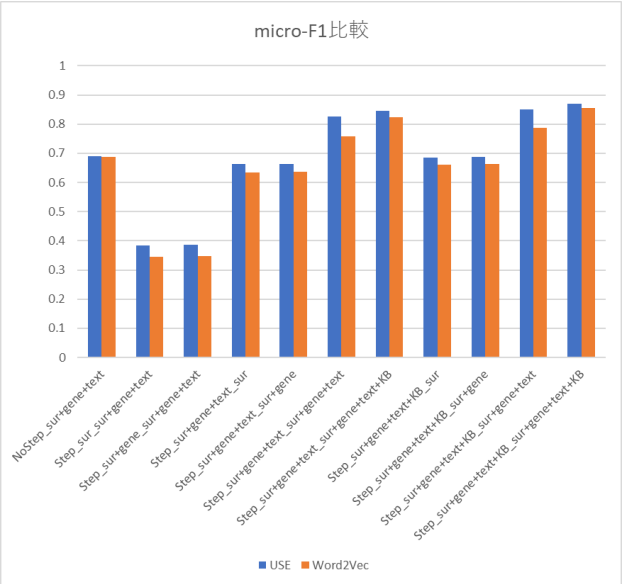
単一モデルで「評価 1 及び 2」「評価 3」「評価 4」「評価 5」に分類

- Step アプローチ：

モデルを 2 段階に分け，別のモデルを構築して，最終的に「評価 1 及び 2」「評価 3」「評価 4」「評価 5」に分類

- Step1：「評価 1 及び 2」「評価 3」は共に「薬剤取違い事例」であることから「評価 1，2 及び評価 3」を一つとして分類（→つまり，Step1 は 3 分類）
- Step2：Step1 の分類を実施した後に，「評価 1，2 及び評価 3」を「評価 1 及び 2」と「評価 3」に分類

図 8：特徴量の追加に対する word2vec と USE の比較（micro-F1）



sur:表層類似度，gene:一般名，text：テキスト情報，KB:KEGG 情報

結果として，NoStep アプローチでは差が見られず，Step アプローチでは，特に Step1 でテキスト情報を追加した場合に，USE の方が精度が良い傾向がみられた．

D. 考察

評価 1，2 及び 3 に分類される事例の記述の特徴の一つとして，テキスト部分に「XX として取り違えた」といった記述がされていることが多い．そのため，テキスト情報に対する特徴量を複数組み合わせることにより，precision が上がることに繋がったと考える．逆に，「XX として取り違えた」という記述が多いという特徴は，「記載が類似している」ということであり，評価 1，2 及び 3 では，テキスト情報に対する特徴量を複数組み合わせることは，誤分類のきっかけになり，recall 低下という結果になったと考える．word2vec と USE との比較においては，モデルを 2 段階に分け，Step1 で「表層類似度+一般名」に「テキスト情報」を追加した場合に，USE が word2vec よりも精度が高い傾向があることから，USE の「文を固定長のベクトルとして表現する，文脈により異なるベクトル表現を獲得可能」という特性が「テキスト情報」に対して有効であると考えられた．

E. 結論

今回検討した特徴量について，特徴量追加と precision，recall，F1-score の各評価指標は連動しておらず，評価機構データの特徴等に依存することが示された．今後，さらに評価機構データの特徴を精査し，新たに追加すべき特徴量を同定するために，図 8 の USE と word2vec との差分となった評価機構データの事例を分析する必要がある．しかし，モデルによる分類の精度を上げることはできても，分類される評価に対するデータ数が少ない，評価機構への報告記述のばらつき等により，モデルによる分類結果と PMDA 評価結果を 100%一致させることは現時点では困難である．今後の開発は，事例解析により，追加すべき特徴量の同定を行う一方で，PMDA の評価検討過程の 1 次スクリーニングとして使用する上で許容できるモデル分類の精度を検討し，許容できる精度を踏まえたアルゴリズムの決定や運用方法の検討をしていく必要があると考える．

また，課題として，評価機構への報告が 2020 年 3 月 17 日以降から新様式での報告となったことから，これまで旧様式をデータとして開発してきたモデルを新様式でのデータ対象に検証が必要となる．当該様式の変更による，これまでに開発したモデル分類への影響について，新様式では，これまでのテキスト記述から報告事例の区分や「発生要因に関する情報」が選択肢として選べるようになることから，分類の精度が上がることを期待し

ている.（新様式のデータは 2021 年度に PMDA より入手予定 ）

F. 健康危険情報　なし

G. 研究発表

1. 岡本里香，中津井雅彦，小島諒介．薬局ヒヤリ・ハット事例に対する安全管理対策評価に関する AI 開発．第 7 回日本医療安全学会学術総会（オンライン）2021 年 5 月 25 日

H. 知的財産権の出願・登録状況　なし