有害事象情報の自動抽出における学習データ量がモデル性能に与える影響の分析

野呂和正[1] 二見光[1] 野村基雄[2] 藤本晃司[3] 田中幸介[4] 松本繁巳[4]
[1]キヤノンメディカルシステムズ株式会社 [2]京都大学大学院医学研究科 腫瘍内科学講座
[3]京都大学大学院医学研究科 高度医用画像学講座 [4]京都大学大学院医学研究科 リアルワールドデータ研究開発講座
[1]{kazumasa1.noro, hikaru.futami}@medical.canon
[2,3,4]{mnomura, kfb, kosuketanaka, motocame}@kuhp.kyoto-u.ac.jp

概要

薬物療法施行中のがん患者の治療経過を把握するために，医師は電子カルテに蓄積される膨大な診療情報を参照する必要があり，多大な労力を要する．本研究では，薬物療法の治療経過を把握する際に医師が参照する情報のうち，有害事象に着目し，診療録及び看護記録から有害事象を自動抽出するモデルの構築及び精度検証を実施した．有害事象抽出モデルの学習データ量と精度の関係性を検証し，学習データ作成に大きな作業コストを必要としない有害事象抽出モデルの構築を実現できることが示唆された．

1 はじめに

がん患者に対する薬物療法は，レジメンと呼ばれる薬剤の種類，投与量，投与スケジュールを集約した治療計画で管理される．レジメンの実施段階において，医師は悪心や下痢といった治療に伴う有害事象の有無を判断し，レジメンにおける薬剤の投与量の減量，投与間隔の変更，中止等の必要性について随時検討する．レジメンの内容変更を検討するための有害事象有無の判断は問診やカルテ記事の参照を通じて行われる．薬物療法の期間は数年に渡る場合があり，この過程で併存疾患の治療も行われると，カルテ記事には様々な診療科での検査結果や患者の主訴，医師の所見や処方オーダの転記等が混在することになる．その結果，医師は過去の治療における有害事象関連の情報へのアクセスに多大な時間と労力を要する．また，臨床研究において，医師は研究に必要なデータの収集のため電子カルテを参照して有害事象の有無を確認するが，症例数は数百の単位に及ぶ場合があり，情報検索の負荷が非常に高い．

近年，OpenAI 社の Chat-GPT や Google 社の Gemini といった，大規模言語モデル（LLM）を基盤とした対話型の AI 技術が普及したことにより，情報検索や文書の要約といった作業が効率的に行えるようになってきた．医療現場においても，これら技術の活用が望まれているが，医療現場への適用においては，個人情報の取扱いや大規模な計算資源の確保といった課題への対応が必要である．

病院内のクローズドな環境において限られた計算資源で動作する小規模な言語モデルは解決策の一つになると考えられる．しかし，小規模言語モデルでは，各診療科で用いられる専門的な用語に対応するための学習データの作成負荷が懸念される．そこで本研究では，有害事象抽出モデルの学習における学習データ量とモデル性能の関係性を分析し，学習データの準備コスト並びに有害事象抽出モデルの実臨床での有用性の観点で検証した．

2 関連研究

人名や地名等の固有表現を文章中から抽出する固有表現抽出は自然言語処理における基本的な技術であり，医療テキストを対象とした固有表現抽出に関する研究も盛んに行われている．柴田ら [1]は診療記録中の疼痛表現を，医療テキストを用いて事前学習した BERT 等を利用して抽出し，文脈を考慮する BERT の有効性と，事前学習に用いるテキストのドメインの重要性を示している．また，加藤ら [2]はカルテ記載中の抗がん剤によって起こる皮膚障害に関する有害事象の抽出とグレードの判定を行う機械学習とルールベースを組み合わせたシステムを構築し，その有効性について報告している．

近年では LLM を用いた固有表現抽出も行われており，風間ら [3]は日本語症例報告コーパスを対象に GPT-4 と Gemini を用いて病変・症状の固有表現を抽出し，適切なプロンプトの設計により医療テキストの固有表現を効果的に抽出可能であることを示

している．西山ら [4]は医療テキスト中の病名，症状等の固有表現抽出精度を LLM と BERT とで比較し，少量データの場合には LLM の精度が高い一方，十分な学習データ量でファインチューニングした場合には BERT の精度が高くなることを示している．LLM を用いた医療テキスト中の固有表現抽出の報告はあるものの，実臨床で使える高精度な有害事象抽出モデルが構築可能か，BERT 等小規模な言語モデルがどの程度の学習データを準備する必要があるかを検討している研究は少ない．

3 方法

本研究は，京都大学大学院医学研究科・医学部及び医学部附属病院の倫理委員会の承認を得て，関連する倫理指針及び規則に従って実施した(審査番号: R2466-4）．

有害事象抽出モデルの学習及び検証手順を図 1 に示す．本手順は，公開データセットで学習した有害事象抽出モデルを医療機関に導入し，医療機関での運用中に追加学習することを模擬している．はじめに，有害事象アノテーションを付与した公開データセットを用いて有害事象抽出モデルをファインチューニングし，検証用の公開データセットで精度を検証した．次に，公開データセットと同様の基準で有害事象アノテーションを付与した実臨床データセットを用いて，公開データで調整されたモデル（調整済モデル）に対し，追加のファインチューニングを行った．実臨床データで再調整されたモデル（再調整済モデル）の精度検証では，学習データ量を段階的に増加させていき，学習データ量の増加に伴う精度の推移を検証用の実臨床データを用いて計測した．最後に腫瘍内科医による有害事象の評価を付与した正解付実臨床データセットを用い，再調整済モデルの精度を評価した．

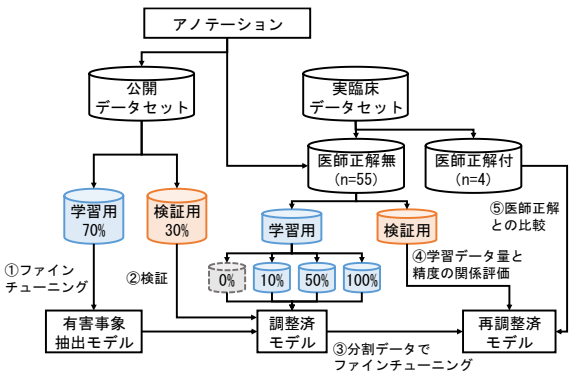

図 1 学習・検証手順

3.1 有害事象抽出モデル

有害事象抽出モデルには，固有表現抽出に適した手法として用いられている BERT [5]と CRF [6]を組み合わせた BERT-CRF [7]を選択した．BERT は，形態素解析により分割されたトークンを入力とし，各トークンを特徴量に変換するエンコーダーモジュールである．本研究では，東北大学自然言語処理研究グループが公開する CC-100 データセット [8], [9]の日本語部分（74.3GB）及び Wikipedia（4.9GB）で事前学習された BERT [10]（パラメータ数：0.1B）を用いた．CRF は，BERT の出力である各トークンの特徴量を入力とし，各トークンに対して有害事象のラベルを分類するモジュールである．ファインチューニング時のバッチサイズは 8，最長出力トークンは 256，最適化アルゴリズムは AdamW，徐々に学習率を上げていく Warmup は 0.1，学習率は 0.0001，エポック数は 30 とした．

3.2 データセット

有害事象抽出モデルのファインチューニングのため，2 種類のデータセットを使用した．

3.2.1 公開データセット

148 件の症例報告に病名や部位名などのアノテーションを付した公開データセット MedTxt-CR: 症例報告 (Case Reports) コーパス [11]を用いた．公開データセットに含まれるアノテーションは使用せず，本研究にて作成した有害事象アノテーション基準（3.2.3）に基づき，開発者が有害事象のアノテーションを新たに付与した．

3.2.2 実臨床データセット

2010 年 4 月 1 日から 2024 年 2 月 29 日までに京都大学医学部付属病院でがんと診断された患者 59 例の SOAP 形式 [12]で記載された診療録及び看護記録を病院情報システムより抽出した．1 患者あたりの平均抽出件数は，S（主観的情報），O（客観的情報），A（評価），P（計画）の単位で集計すると診療録で 2,731 件，看護記録で 606 件であった．1 患者あたりの診療録の抽出対象期間の平均は 1,728 日であった．59 例中 4 例は，腫瘍内科医と治験コーディネーターにより確認済の有害事象が記録されており，診療時の有害事象評価を確認できることから，モデルが抽出した有害事象の腫瘍内科医による評価用に

用いた．

3.2.3 有害事象アノテーション

診療録及び看護記録中の有害事象を抽出する有害事象抽出モデルの構築のため，臨床医学アノテーションガイドライン [13]，有害事象共通用語標準 JCOG CTCAE v5.0 [14]，医薬品副作用データベース（JADER） [15]及び腫瘍内科医の知見を参考に，有害事象アノテーション基準（A.2）を作成した．開発者が診療録と看護記載中から，CTCAE の定義または JADER のがん症例に紐づく有害事象にあたる文字列を探索し，該当する文字列に対して陽性（有害事象の発現への言及），陰性（過去に発現した有害事象への言及、発現した有害事象の消失），疑い（発現可能性のある有害事象の説明），一般（発現や消失の可能性の言及）の 4 種の属性を付与した．

3.3 検証方法

3.3.1 モデルの学習及び精度検証方法

公開データセットに含まれる 148 件の症例報告に対し，有害事象アノテーション基準に従い 130 件に計 1,223 個の有害事象のアノテーションを付与した．残りの 18 件には有害事象アノテーション基準に合致する有害事象が存在せず，使用しなかった．アノテーションには，本研究において作成したアノテーションツール（A.1）を用いた．856 個のアノテーションを含む 91 件（70%）の症例報告を学習用，367 個のアノテーションを含む 39 件（30%）を精度検証用とした．症例報告は句点で分割し，文単位で有害事象抽出モデルの学習と検証に用いた．

実臨床データセットは，各例の診療録と看護記録を日付昇順にソートし，日付の古い順に記載を参照し，公開データセットに対するアノテーションと同じ方法でアノテーションを付与した．本検証では，各例において 20 個を目安にアノテーションを付与し，目安に到達した段階でアノテーションを含む診療録または看護記録を学習用データとして保存した．続けて，10 個程度を目安にアノテーションを付与し，目安に到達した段階で検証用データとして保存した．本検証では腫瘍内科医による評価用の 4 例を除いた 55 例の診療録，看護記録に対し 2,541 個のアノテーションを付与し，学習用に 1,480 個，検証用に 1,061 個のアノテーションに分割した．

有害事象抽出モデルの精度は，①学習用の公開データセットのみ使用，②学習用の実臨床データセットを 10%（6 例）追加，③50%（28 例）追加，④全て（55 例）追加の 4 つの場合で検証した．まず①の場合において，検証用の公開データと実臨床データ（55 例）でそれぞれ精度を測定し，両者を比較した．次に検証用の実臨床データセットを用いて，②～④の場合で精度を測定し，精度の推移を確認した．予測された有害事象のうち，付与したアノテーションと完全一致したものを真陽性（TP），完全一致していないものを偽陽性（FP），付与したアノテーションのうち，未検出のものを偽陰性（FN）とし，再現率（$TP/(TP+FN)$）及び適合率（$TP/(TP+FP)$）の指標により精度を測定した．

3.3.2 腫瘍内科医による評価結果との比較

腫瘍内科医による有害事象の評価を正解として付与した 4 例を使用し，学習用の公開データセット及び実臨床データセット全て（55 例）でファインチューニングした有害事象抽出モデルの精度を腫瘍内科医の正解との比較により再現率，適合率の指標で評価した．腫瘍内科医による有害事象の評価結果には，好中球数減少といった JCOG CTCAE v5.0 の判定基準に基づき，臨床検査結果から機械的に判定される有害事象 [16]が含まれる．そのため，本検証では機械的に判断が可能な有害事象を評価対象から除外の上，腫瘍内科医の正解と比較した．

4 結果・考察

4.1 結果

4.1.1 学習データ量と精度の関係

公開データセット及び実臨床データ（55 例）を用いた有害事象抽出モデルの精度検証結果を図 2 に示す．①学習用の公開データセットのみを使用した場合，検証用の公開データセットで評価した結果は再現率 88.3%，適合率 80.5%であったが，検証用の実臨床データセットでの評価では，再現率 75.6%，適合率 36.1%と低下した．次に，段階的に学習データ量を増やしていく②～④の場合において精度を測定した結果，②学習用の実臨床データセットを 10%使用した場合には再現率 86.5%，適合率 61.5%，③50%使用した場合に再現率が 92.4%，適合率 78.5%，④全て使用した場合には再現率

92.4%，適合率 80.3%となり，学習データ量の増加により精度が向上することを確認した．

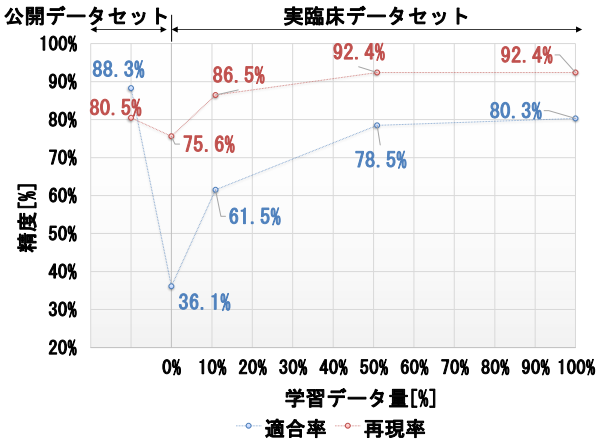

図 2 学習データ量と精度の関係

4.1.2 腫瘍内科医による評価結果との比較

腫瘍内科医による有害事象評価結果と有害事象抽出モデルによる有害事象出力との比較結果を表 1 に示す．4 例合計としての再現率と適合率は，それぞれ 96.6%及び 78.4%であった．

表 1 腫瘍内科医による評価結果との比較結果

症例番号	TP	FP	FN	再現率	適合率
1	16	2	0	100%	88.9%
2	36	13	2	94.7%	73.5%
3	5	1	3	62.5%	83.3%
4	143	39	2	98.6%	78.6%
合計	200	55	7	96.6%	78.4%

4.2 考察

公開データセットでファインチューニングした有害事象抽出モデルは，実臨床データセットでの追加学習がない場合には再現率，適合率ともに低かった．これは，公開データセットと実臨床データセット間でのドメインの差異によるものと考えられる．公開データセット（症例報告）には確認されなかったものの，実臨床データセット（診療録，看護記録）において頻出した記載のパターンとして，表 2 に示す「与薬指示」や「副作用の説明」が挙げられ，属性の誤検出（“一般”を“陽性”，もしくは“陰性”を“陽性”と誤検出）が確認された．こうした事例に対しては，実臨床データで追加のファインチューニングをすることで意図した属性で検出が可能となり，適合率を大幅に向上させる要因の一つとなった．また“陰性”として抽出されるべき「過去事象」がその時点で発現していることを示す“陽性”として抽出されていることも，有害事象抽出モデルの誤り事例の一つであった．「既往歴」や「7 年前に」のように明らかに過去の事象とわかる単語が近傍に存在する場合には判定が容易であるものの，「2014 年 10 月頃に」のようにある一期間または一時点を指す単語が近傍に存在する場合は，カルテ記載時点との相対的な時間関係によって過去の事象か現在も存続しているか判断が求められる場合があり，このパターンでは，実臨床によるファインチューニングを行っても誤検出の事例が存在した．

有害事象抽出モデルの精度は，50%の学習用実臨床データセットを使用して追加のファインチューニングを行った時点から再現率，適合率の推移幅が小さくなることが確認された．本検証では，開発者 1 名が学習用と検証用の実臨床データセットに対して計 2,541 個のアノテーションを付与するために約 15 時間を要した．再現率・適合率の推移幅が小さくなる 50%の学習用実臨床データセットには約 800 個のアノテーションを付与したため，アノテーションに要する時間は概ねその半分程度で十分であると考えられる．小規模な言語モデルを使用する場合でもそれほど多くの時間的コストをかけることなく実運用に足るシステムの構築ができると考えられる．

1 名でのアノテーションは再現性と信頼性の制約となるが，本研究では明確な有害事象のアノテーション基準を作成したことによって，再現性と信頼性を確保した．

表 2 有害事象抽出モデルの誤り事例

パターン	誤りの事例
与薬指示	**正解）**〇〇薬　**悪心・嘔吐**<一般>時. **予測）**〇〇薬　**悪心・嘔吐**<陽性>時.
副作用の説明	**正解）**副作用：**吐き気，食欲不振**<一般>について説明した. **予測）**副作用：**吐き気，食欲不振**<陽性>について説明した.
過去事象	**正解）**2014 年 10 月頃より**心窩部痛**<陰性>. **予測）**2014 年 10 月頃より**心窩部痛**<陽性>.

5 おわりに

本研究ではカルテ記載から有害事象を自動抽出する BERT ベースのモデル構築及び精度検証を行い，学習データ量と精度の関係性について分析した．実臨床データでファインチューニングすることで，小規模な言語モデルでも，アノテーションの大きな作業負担なく実運用に足る精度で有害事象の抽出が行えるシステムの構築が可能であることが示唆された．

謝辞

本研究を遂行するにあたり，研究計画や検証方法について有益な助言を頂いた京都大学大学院医学研究科 リアルワールド研究開発講座の坂本亮特定講師に感謝致します.

参考文献

1. 診療記録で事前学習した BERT による疼痛表現の抽出. 柴田大作, 河添悦昌, 嶋本公徳, 篠原恵美子, 荒牧英治. 2, 2020 年, 医療情報学, 第 40 巻, ページ: 73-82.
2. 電子カルテ自由記述部分からの皮膚疾患における重症度抽出. 加藤由美, 平川聡史, 梶山晃平, 堀口裕正, 狩野芳伸. 2019 年, 言語処理学会 第 25 回年次大会.
3. プロンプト最適化を用いた大規模言語モデルによる医療文章からの固有表現抽出. 風間正弘, 太田満久, 西林孝. 2025 年, 2025 年度人工知能学会全国大会（第 39 回）.
4. 生成モデルは医療テキストの固有表現抽出に使えるか? 西山智弘, 柴田大作, 宇野裕, 辻川剛範, 北出祐, 久保雅洋, 矢田竣太郎, 若宮翔子, 荒牧英治. 2024 年, 言語処理学会 第 30 回年次大会.
5. Bert: Pre-training of deep bidirectional transformers for language understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019 年, Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, 第 1 巻.
6. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. John Lafferty, Andrew McCallum and Fernando C.N Pereira. 2001 年.
7. Portuguese Named Entity Recognition using BERT-CRF. Fábio Souza, Rodrigo Nogueira and Roberto Lotufo. 2019 年, arXiv preprint. arXiv:1909.10649.
8. Unsupervised Cross-lingual Representation Learning at Scale. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov. 2020 年, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), ページ: 8440-8451.
9. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin and Edouard Grave. 2020 年, Proceedings of the 12th Language Resources and Evaluation Conference (LREC), ページ: 4003-4012.
10. Tohoku NLP Group. Bert base japanese (unidiclite with whole word masking, cc-100 and jawiki20230102). (オンライン) 2023 年. (引用日: 2025 年 9 月 2 日.) https://huggingface.co/tohoku-nlp/ bert-base-japanese-v3.
11. Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task. Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya and Eiji Aramaki. 2022 年, In Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16), ページ: 285-296.
12. Weed Lawrence L. Medical Records, Medical Education, and Patient Care: The Problem-oriented Record as a Basic Tool.: Press of Case Western Reserve University, 1970.
13. 汎用的な臨床医学テキストアノテーション仕様およびガイドラインの策定：重篤肺疾患ドメインに着目して. 矢田竣太郎, 田中リベカ, Cheng Fei, 荒牧英治, 黒橋禎夫. 4, 2022 年, 自然言語処理, 第 29 巻, ページ: 1165-1197.
14. Japan clinical oncology group. Common terminology criteria for adverse events (ctcae) ver. 5.0. (オンライン) (引用日: 2025 年 9 月 2 日.) https://jcog.jp/assets/CTCAEv5J_20220901_v25_1.pdf.
15. 独立行政法人 医薬品医療機器総合機構. 医薬品副作用データベース　利用規約. (オンライン) (引用日: 2025 年 9 月 11 日.) https://www.pmda.go.jp/safety/info-services/drugs/adr-info/suspected-adr/0003.html.
16. 日本臨床腫瘍研究グループ. 共用基準範囲対応 CTCAE v5.0 Grade 定義表. (オンライン) (引用日: 2025 年 9 月 11 日.) https://jcog.jp/assets/JCOG_kyouyoukijunchi-CTCAE_50_20241128.pdf.

A 付録

A.1 アノテーションツール

本検証で用いたアノテーションツールの画面例を図 3 に示す．ブラウザ上で操作可能なツールを自作し，アノテーション作業に用いた．はじめに，病院情報システムから抽出した診療録及び看護記録を句点や改行により分割し，文章単位（学習させるレコード単位）で各行に表示させる．表示されたテキストに対して，マウスの選択操作により，有害事象に該当する文字列を選択する．有害事象に該当する文字列を選択する際に，有害事象の属性（A.2）を指定しておくと，指定した属性でアノテーションが付与される．

図 3 アノテーションツール

A.2 有害事象アノテーション基準

本研究で作成した有害事象アノテーション基準を表 3 に示す.

表 3 有害事象アノテーション基準

属性		属性判定基準
陽性 Positive		～を認めた、～がみられた、～と判断、～と診断、～の出現、～の症状あり、～の所見、～となった、～像を示す、～は増悪を認めない、主訴：～、～を自覚、～を指摘、～を呈した、現病歴：～、～を来した、～を発症、○○年前より～が続いており
陰性 Negative		～はなし、～はなかった、～はみられない、～は陰性であり、～は消失、～を認めなかった、～の再燃を認めなかった、～の再発は認めていない 既往歴：～、～の既往があり、○歳のときに～であり、○年前に～であり、
疑い Suspicious		～の疑い、～は改善し、～の消退、～の縮小、～は軽快した、～が推測された、～と考えた、～は寛解した、～の可能性が高い、～はおさまっており、～が示唆され、～と思われ、ほとんど消失
一般 General		～の（が）報告され、・・・の時（場合）には～である、～の原因は、
例外 Exception	血液・生化学検査の言及	血液検査では**血沈亢進**、ACE **高値**が持続し
	病理所見	**病理結果**では、線維性結合織の増生を伴う炎症細胞浸潤が認められた
	介入の判断が困難	筋緊張異常と**変形**・拘縮、呼吸障害