診療データベースを用いたカルテスクリーニング

柴田 大作 [1] 辻川 剛範 [1] 渋谷 恵 [1] 森田 智子 [1] 久保 雅洋 [1] 中川 敦寛 [2] 重田 昌吾 [2] 島田 宗昭 [2]
[1] 日本電気株式会社 [2] 東北大学病院
{daisaku-shibata,tujikawa,kei-shibuya,tomoko-morita,masahirokubo}@nec.com
{atsuhiro.nakagawa.b2,shogo.shigeta.a4,muneaki.shimada.b7}@tohoku.ac.jp

概要

医師によって作成される初診時記録や経過記録などの診療記録の利活用に向け，機械学習とルールベースに基づく診療データベースの構築，そして診療データベースを用いたカルテスクリーニング (臨床試験の選択基準に適合する患者の選定作業) について検討を行う．開発した機械学習モデルは固有表現抽出が Micro-F1 で 0.884，関係抽出が 0.768 と一定の精度で診療記録からの情報抽出が可能であった．また診療データベースを用いたカルテスクリーニングの精度は，診療記録を直接用いた際の精度よりも最大で 4.2 倍改善することが明らかとなった．本研究で構築した診療データベースはカルテスクリーニングに有用である可能性が示唆された．

1 はじめに

医師によって作成される経過記録などの診療記録には患者の訴えや所見などの情報が多く含まれているため，臨床研究や診断支援などにおける利活用が期待されているが，診療記録はフリーテキストで記載される非構造なデータであるため直接的に利用することは難しい．現在は医師が研究に必要な情報をそれぞれ診療記録から手動で抽出し，データベース化 (以後，診療データベース) しているが，日々の診療業務で多忙な医師にとっては負担の大きい作業である．そのため，診療記録からの情報抽出を行う研究がいくつか報告されている．Patel ら [1] は 5,160 件の診療記録に出現する 13 種類の固有表現と 11 種類の関係に対してアノテーションを行い，Conditional Random Field (CRF) による固有表現抽出を行い，高い精度で抽出できることを明らかにした．Hu ら [2] は ChatGPT を用いた退院サマリからの固有表現抽出の性能について調査し，Fine-tuning を行った事前学習済みモデルには劣るものの、訓練データなしでも一定の精度で固有表現抽出が可能であることを報告した．日本語の診療記録を対象とした研究もあり，矢田ら [3] は 3,769 件の読影レポートと診療録に出現する 40 種類の固有表現と 11 種類の関係についてアノテーションを行い，事前学習済みモデルを用いた情報抽出の精度評価を行った．

診療記録へのアノテーションや情報抽出の精度評価は重要な知見であるが，診療記録から抽出した情報の利活用まで踏み込んだ研究は少なく検討の余地がある．そのため本研究では，診療記録からの情報抽出に基づく診療データベースの構築，加えて診療記録の利活用が期待される先の一つである臨床試験において構築した診療データベースがどの程度活用可能かについて検討を行う．

2 本研究で取り組むタスク

2.1 診療データベースの構築

診療データベースには統一されたフォーマットがないため，用途に応じて任意にカラムを追加，変更もしくは削除できることが望ましいと考えられる．そのため固有表現抽出 (Named Entity Recognition: NER) と関係抽出 (Relation Extraction: RE) により診療記録の解析を行い，それらを事前に定めたルールに基づいてデータベースへ変換する機械学習とルールベースを併用した手法による診療データベースの構築を行う．概要を図 1 の青枠部分に示す．

2.2 カルテスクリーニング

臨床試験では，試験の目的を正確に評価するため，試験に参加する患者を選定する選択基準が設けられている．選択基準には，適格基準 (全て満たさないといけない条件) と除外基準 (一つも抵触してはいけない条件) の 2 種類があり，これらの基準に適合する患者の選定作業を一般にカルテスクリーニングと言う．カルテスクリーニングは初診時記録，経過記録や病理レポートなどの診療記録，MRI や

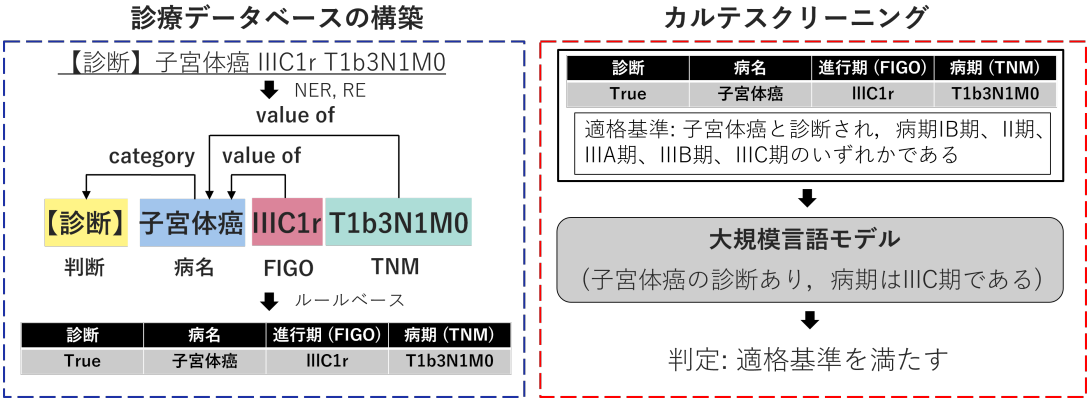

図 1　本研究の概要: 青枠が診療データベースの構築，赤枠がカルテスクリーニングを示す．

CT などの医療画像，現病歴や検査結果などの構造化データを参照して実施されるが，1 患者につき 30 分から 80 分程度の時間が必要であり，コンピューターによる支援が期待されている一方，基準は臨床試験ごとにそれぞれ数件から数十件程度設定されるため，一般的な機械学習やルールベースによる判定は難しい．そこで大規模言語モデル (Large Language Model: LLM) と診療データベースを用いた適格・除外基準の自動判定について検討を行う．概要を図 1 の赤枠部分に示す．

3　倫理的配慮

本研究は東北大学病院医学系研究科倫理委員会の承認 (承認番号: 2024-1-091) を得て実施された．本研究で使用した診療記録は事前に東北大学病院内で仮名加工化した上で，NEC 内部へ転送し，データの分析を行った (詳細は付録 A に示す)．

4　実験データ

東北大学病院の婦人科において作成された診療記録を実験データとして使用した．診療データベースの構築で用いる NER と RE の実験データの詳細を節 4.1，カルテスクリーニングで用いる実験データの詳細を節 4.2 にそれぞれ示す．

4.1　固有表現抽出と関係抽出

診療データベースの構築において，NER と RE は教師あり学習によって行う．婦人科において 2021 年に作成された経過記録に対して固有表現と関係アノテーションを行った．アノテーションにおいては篠原ら [4] によって考案された症例報告ベースのアノテーションガイドラインを経過記録向けにいくつか変更したものを使用した．

アノテーションでは経過記録を S/O/A/P 単位にそれぞれ分割し，分割後の各セグメントを 1 文として扱い，1 文ごとにアノテーションを行った．4 名のアノテーターでアノテーションを行い，117 種類の固有表現と 38 種類の関係がアノテーションされた経過記録 851 文書 (2,741 文) を作成した．

4.2　カルテスクリーニングの評価データ

婦人科領域で行われている臨床試験である JCOG1412[1)] の基準を用いて評価データの作成を行う．JCOG1412 では 11 個の適格基準 (表 5) と 15 個の除外基準が設定されているが，除外基準に抵触する患者は少ないため本研究では適格基準を対象とし，適格基準の中から初診時記録もしくは経過記録から判定が可能であるものを 5 つ選択し，一部文言を変更したものを評価に使用する基準とした．

婦人科を 2017 年もしくは 2018 年に受診した患者の中から 142 名を選択し (詳細は付録 B に示す)，各患者の 2017 年もしくは 2018 年に作成された初診時記録と経過記録を全て取得し，適格基準を満たす場合は True，満たさないもしくは記載なしの場合は False として各基準に対するアノテーションを行った．アノテーションは看護師 1 名により実施した．評価データで使用する適格基準とその統計情報を表 1 に示す．

5　実験設定

5.1　診療データベースの構築

診療データベースの構築は NER，RE，そしてルールベースによる表形式への変換の 3 つのステップで

1) https://jrct.niph.go.jp/latest-detail/jRCTs031180269

表 1 評価用データの統計情報: 適格数はその基準を満たす患者数を示す.

No.	内容	適格数
2	術前病期 IB 期、II 期、IIIA 期、IIIB 期、IIIC 期のいずれかである	54
6	Performance status （PS）0 または 1	19
7	BMI（Body Mass Index）35 以下	76
8	腸管切除を伴う手術の既往がない	94
9	他のがん種を含めて化学療法や放射線治療の既往がない	61

行う.

5.1.1 固有表現抽出と関係抽出

NER には Bidirectional Encoder Representations from Transformers (BERT)[5] ベースの BERT-CRF、RE には Head Selection 問題として定式化するモデル [6, 7] を使用した．入力データの最大トークン数は 300 に設定し，NER の学習では学習率を 1e-5，バッチサイズを 32，最適化関数には AdamW を使用し，RE ではバッチサイズを 4 とし，それ以外は固有表現抽出と同じ設定とした．なお記載のないパラメーターは全てデフォルト値を使用した．事前学習済みモデルとしては Wikipedia で事前学習された luke-japanese-large-lite[8] を使用した．なお 1 文あたりの平均トークン数は 40.1 トークンで標準偏差は 42.7，1 文あたりに含まれる固有表現は 11.5 個で標準偏差は 11.7，1 文あたりに含まれる関係は 8.9 個で標準偏差は 10.7 であった.

評価は 5 分割交差検証により行い，訓練データの 20%を検証データとして使用し，検証データにおける Micro-F1 が最も高かった時のパラメーターを用いてテストデータに対する評価を行った．NER は IOB2 形式によるラベリングを行い，評価は CoNLL-2000[9] と同様の方法で実施した．RE は正解の関係ラベルとモデルの予測した関係ラベルが一致した場合を True Positive，本来は関係ラベル A が付与される箇所に誤って他の関係ラベルを付与した場合を False Negative，本来は関係ラベルが付与されない箇所に誤って何かしらの関係ラベルを付与した場合を False Positive として評価を行った [10].

5.1.2 表形式への変換

NER と RE によって得られた固有表現タグと関係ラベルを用い，ルールベースによる表形式への変換を行う．ルールはアノテーションガイドラインをベースに設定し，例えば図 1 の青枠部分では，ある疾患のステージに関する情報は「固有表現タグが病名である Entity と固有表現タグが FIGO もしくは TNM である Entity が関係ラベル value-of で関係付けられる」というルールを定義することで表形式へ変換している．最終的に病名，治療，検査と既往歴に関する情報を集約したテーブルをぞれぞれ作成するようにルールを定義した，なお本研究では作成したテーブル自体の評価は実施しておらず，これについては今後の課題とした.

5.2 カルテスクリーニング

診療データベースを用いて適格基準を満たすか否かの判定を行うプロンプトを作成し，LLM による判定を行う．基準ごとに判定を行い，各基準に関連するテーブルを用いてプロンプトを作成する (例えば，既往歴に関する内容の基準であれば既往歴と治療テーブルを使用する)．LLM には Llama-3.1-Swallow-70B-Instruct-v0.1 を使用し，評価用データに含まれないある患者の 1 か月分の診療記録を用いて事例を作成し One-Shot Learning による推論を行った．また診療データベースの代わりに，オリジナルの診療記録を入力した場合の判定も併せて実施する．プロンプトのイメージなどを含めた概要を付録 C の図 2 に示す.

患者の状態は時々刻々と変化するため，診療記録を 1 か月単位に分割し，1 か月ごとに適格基準の判定を行い，最終的に全期間の予測結果を結合して評価を行う．評価においては Precision と Recall を適格基準ごとに算出した．なお 1 患者あたり平均で診療記録が 55 件 (標準偏差は 33) あり，文字数は 10,648 文字 (標準偏差は 6,347) であり，月単位では 6.2 件 (標準偏差は 3.6) あり，文字数は 1,123 文字 (標準偏差は 514) であった.

6 実験結果

6.1 固有表現抽出と関係抽出

実験結果を表 2 に示す．NER は Micro-F1 で 0.884，RE は 0.768 であることが確認された．また参考値として先行研究で報告されている NER と RE の結果も併せて記載する.

6.2 カルテスクリーニング

実験結果を表 3 に示す．診療データベースを用いてプロンプトを作成した方が，経過記録を用いてプ

表 2　NER と RE の実験結果

著者	言語	種類	データ数	固有表現タグ数	関係ラベル数	精度	
						NER	RE
本研究	日本語	経過記録	851	117	38	0.884	0.768
矢田ら [3]	日本語	読影レポート	3,380	40	11	0.867	0.918
		診療録	461	40	11	0.865	0.661
Patel et al.[1]	英語	診療録	5,160	11	13	0.916	-
Shibata et al.[10]	日本語	症例報告	183	113	36	0.912	0.759

ロンプトを作成するよりも Precision，Recall 共に高い値となることが明らかとなった．

表 3　カルテスクリーニングの実験結果

No.	LLM への入力			
	経過記録		診療データベース	
	P	R	P	R
2	0.046	0.217	0.202	0.569
4	0.029	0.084	0.508	0.750
6	0.087	0.259	0.331	0.390
7	0.117	0.343	0.276	0.536
8	0.054	0.130	0.101	0.141
平均値	0.067	0.207	0.284	0.477

7　考察

7.1　アノテーション一致率の観点から

本研究では 4 名のアノテーター (以後，作業者) により固有表現と関係アノテーションを行った．作業者ごとに習熟度に差があり，作業者 1 はアノテーション作業が 200 日を超えるベテランであり，作業者 2 も同様に 180 日を超えるベテラン，作業者 3，4 は約 80 日の若手である．ここで作業者 1 の実施したアノテーションデータを 50 件抽出 (作業者 1 が難易度の高いもの，普通のものをそれぞれ 25 件ずつ選択) し，作業者 2 から 4 がアノテーションを行い，作業者 1 のアノテーションを正解，作業者 2 から 4 のアノテーションを予測として評価した場合の結果をアノテーション一致率として表 4 に示す．表 4 から，アノテーション一致率の平均値は NER で 0.891，RE で 0.763 であり，実験結果に近い値となっていることが確認された．経過記録 2,741 文のうち，作業者 1 がアノテーションしたものが 1,140 文，作業者 2 が 1,177 文，作業者 3 が 237 文，作業者 4 が 187 文であり，作業者 1 と 2 がアノテーションしたデータが 80%以上であるものの，作業者 1 と作業者 2 のアノテーション一致率と実験結果の間には乖離があり，加えて作業者 1 と作業者 3，4 とのアノテーション一致率は低いことから全体的にアノテーションを見直し，品質を向上させることで性能の改善が可能であると考えられる．

表 4　作業者間のアノテーション一致率 (Micro-F1)

正解	予測	NER	RE
作業者 1	作業者 2	0.929	0.842
	作業者 3	0.879	0.765
	作業者 4	0.864	0.680
平均値		0.891	0.763

7.2　診療データベースと経過記録の比較

診療データベースを用いてプロンプトを作成した時の精度が，経過記録を用いて作成した時の精度よりも高い値となることが確認された．診療データベースを用いて作成したプロンプトの文字数の平均値は 1,674 文字である一方，経過記録を用いた場合は 3,065 文字であり，専門用語や略語などが多様される非文法的なテキストである診療記録を LLM で処理する場合，あらかじめ構造化データへ変換し，必要な情報のみを用いてプロンプトを作成することの有用性が示唆された．しかしながら，NER と RE で正しく情報抽出ができなかったためプロンプトに必要な情報を入力することができず，基準の予測を誤ってしまった事例も確認された．そのため診療データベースだけではなく，適宜オリジナルの診療記録も組み合わせて活用していくことが必要であると考えられる．

8　まとめ

本研究では NER，RE とルールベースに基づく診療記録からの診療データベースの構築と診療データベースのカルテスクリーニングへの適用について検討を行った．実験では一定の精度で NER と RE ができることが明らかとなり，また構築した診療データベースはカルテスクリーニングにおいて有用である可能性が示唆された．今後は構築した診療データベース自体の評価，本研究では使用しなかった適格・除外基準を用いた場合の評価や同様に本研究では扱わなかった病理レポートなども含めたカルテスクリーニングについて検討する予定である．

8.1 謝辞

本研究の実施にあたり，ご協力いただいた東北大学病院 臨床研究推進センターバイオデザイン部門[2)]の皆様に厚く御礼申し上げます．

参考文献

[1] Pinal Patel, Disha Davey, Vishal Panchal, and Parth Pathak. Annotation of a large clinical entity corpus. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2033–2042, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[2] Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. Zero-shot clinical entity recognition using chatgpt. **arXiv preprint arXiv:2303.16416**, 2023.

[3] 矢田竣太郎, 田中リベカ, Fei Cheng, 荒牧英治, 黒橋禎夫. 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. 自然言語処理, Vol. 29, No. 4, pp. 1165–1197, 2022.

[4] Emiko Shinohara, Daisaku Shibata, and Yoshimasa Kawazoe. Development of comprehensive annotation criteria for patients' states from clinical texts. **Journal of Biomedical Informatics**, Vol. 134, p. 104200, 2022.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Weipeng Huang, Xingyi Cheng, Taifeng Wang, and Wei Chu. Bert-based multi-head selection for joint entity-relation extraction. In **Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8**, pp. 713–723. Springer, 2019.

[7] Fei Cheng, Shuntaro Yada, Ribeka Tanaka, Eiji Aramaki, and Sadao Kurohashi. Jamie: A pipeline japanese medical information extraction system with novel relation annotation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 3724–3731, 2022.

[8] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Association for Computational Linguistics, 2020.

[9] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the CoNLL-2000 shared task chunking. In **Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop**, 2000.

[10] Daisaku Shibata, Emiko Shinohara, Kiminori Shimamoto, and Yoshimasa Kawazoe. Towards structuring clinical texts: Joint entity and relation extraction from japanese case report corpus. MEDINFO 2023—The Future Is Accessible, pp. 559–563. IOS Press, 2024.

2) https://www.edas.hosp.tohoku.ac.jp/about

A　情報公開

本研究の情報公開文書は https://www.med.tohoku.ac.jp/wp-content/uploads/2024/08/2024-1-091-1.pdf から参照可能である．

B　カルテスクリーニングで使用する評価データの抽出方法

本研究で対象とした臨床試験 (JCOG1412) において使用された適格基準を表 5 に示す．東北大学病院の婦人科で 2012 年から 2023 年において検体検査を受けた患者は 11,266 人おり，この中から以下の手順に従って評価データとして使用する患者 142 人を抽出し，アノテーションを行った．なお以下で使用した適格基準はカルテスクリーニングにおいて使用する基準からは除外した．

1. 年齢が 20 歳から 75 歳，性別が女性，検体の採取期間を 2017 年と 2018 年に限定 (適格基準 5): 11,266 人から 2,139 人に
2. 現病歴として子宮体癌もしくは子宮類内膜腺癌が登録されている患者に限定 (適格基準 1): 2,139 人から 357 人に
3. 特定の検査の検査値を満たす患者に限定 (適格基準 10): 357 人から 319 人に
4. 手術記録を参照し特定の治療を受けた患者に限定 (適格基準 1): 319 人から 142 人に

表 5　JCOG1412 の適格基準一覧

No.	内容	絞り込みに使用	除外
1	原発巣が子宮体癌（類内膜腺癌、粘液性腺癌、漿液性腺癌、明細胞腺癌、未分化癌、混合癌のいずれか）であることが組織学的に確認されている	✓	
2	術前病期 IB 期、II 期、IIIA 期、IIIB 期、IIIC 期のいずれかである。ただし、IIIC 期の場合には画像検査で以下のいずれかを満たす		
3	造影 CT で腹膜播種、遠隔臓器転移、鼠径リンパ節転移を認めない。胸部～骨盤造影 CT を同部位の PET/CT で代用してよい		✓
4	造影 MRI で膀胱浸潤、直腸浸潤のいずれも認めない		✓
5	登録日の年齢が 20 歳以上、75 歳以下である	✓	
6	Performance status （PS）0 または 1		
7	BMI（Body Mass Index）35 以下		
8	腸管切除を伴う手術の既往がない		
9	他のがん種を含めて化学療法や放射線治療の既往がない		
10	登録前 14 日以内の最新の検査値（登録日の 2 週間前の同一曜日は可）が、以下のすべてを満たす (以下，省略)	✓	✓
11	試験参加について患者本人から文書で同意が得られている		✓
12	二次登録の適格規準 (以下，省略)		✓

C　カルテスクリーニングの詳細な手順

カルテスクリーニングの実験手順を図 2 に示す．One-shot による推論を行い，与える事例は評価データに含まれない患者の 1 か月分の診療記録を用いて作成した．

図 2　カルテスクリーニングの概要図: 適格基準ごとに判定を行う．またプロンプトへ入力するテーブルは各適格基準と対応するものを使用する．