診療テキストの利活用に向けたデータベースの自動構築とカルテスクリーニングへの適用

柴田大作 [1] 辻川剛範 [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,masahirokubo}@nec.com

概要

医療現場で日々作成される診療テキストの利活用に向け，機械学習による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可能性について検討を行う．その結果，構築した診療データベースをカルテスクリーニングに用いることで，カルテスクリーニングに要する時間を約 80% 短縮 (174 時間から 34 時間) できる可能性があることが明らかとなり，その有効性が示唆された．

1 はじめに

医師によって作成される経過記録や病理レポートなどの診療テキストには患者の訴え，所見や診断に関する情報などが記載されており，臨床研究や診断支援などへの利活用が期待されている．しかし，診療テキストは医師によってフリーテキストで作成されるため，情報が構造化されておらず利活用における大きな障壁となっている．現状は医師が必要な情報を手動で診療テキストから抽出し，データベース化 (以後，診療データベース) することで診療テキストを構造化データへ変換し使用している．診療データベースはカルテスクリーニング (臨床試験の選択基準を全て満たす患者を検索する作業) やがん登録などの情報登録作業に活用が可能であるが，診療データベースの作成には大量の診療テキストを確認する必要があり，加えて運用には定期的なアップデート作業が必要不可欠であるため，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，診療データベースを自動で作成することができれば，医師の負担軽減や臨床試験などの更なる促進に繋がると考えられる．そこで本研究では，機械学習による診療データベースの構築と評価，また構築した診療データベースのカルテスクリーニングへの適用可能性について検討を行う．

2 関連研究

診療テキストからの情報抽出を行う研究はいくつか報告されており，Fei ら [1] は日本語で作成された 1,000 件の読影レポートと 156 件の病歴レポートに出現する固有表現と関係に対してアノテーションを行い，それらを用いてテキストに出現する医療表現，モダリティと関係の分析を高精度で実行可能なパイプライン型のシステムを構築し，公開した．土橋ら [2] は大規模言語モデル (Large Language Models: LLMs) により，日本語の読影レポートからモダリティ，主病変やリンパ節転移の有無などの情報について抽出精度の評価を行い，LLMs により一定の精度で読影レポートからの情報抽出が可能であることを報告した．Yan ら [3] は診療テキストからの固有表現抽出と関係抽出の精度を LLaMA と BiomedBERT で比較した結果，未知のデータの固有表現抽出においては LLaMA が F1 スコアで 7%，関係抽出においても同様に LLaMA が 4%高いが，LLaMA はより多くの計算資源を必要とし，推論時間は BiomedBERT の最大 28 倍遅いことを報告した．

自然言語処理や機械学習によりカルテスクリーニングを自動で行う研究もいくつか報告されている．Beattie ら [4] は Generative Pre-trained Transformers (GPT)-3.5 Trubo と GPT-4 により，202 名の患者記録に 13 個の選択基準の情報がアノテーションされたデータセット (うち 20 件はプロンプトエンジニアリングに使用) を用いて選択基準の自動判定を行った結果，GPT-4 では F1 が 0.86，GPT-3.5 Turbo では 0.79 であり，一定の精度で判定が可能であることを確認した．また Ni ら [5] は自然言語処理と機械学習に基づき，構造データと非構造データの自動解析を行うことで，患者の臨床試験への登録適格性を自動で判定するシステムを開発し，システムを用いることで手動によるスクリーニングと比較して 34%の時

間短縮ができることを明らかにした．これら以外にも同様の研究が多数報告 [6, 7, 8] されている．

このようにいくつかの関連研究が報告されているが，診療テキストからの情報抽出において，実際の臨床現場で作成されているデータベースの構築まで踏み込んだ研究は我々の知る限り報告されていない．そのため本研究では，診療テキストからの情報抽出だけではなく，実際に臨床現場で作成されている診療データベースの構築まで踏み込んだ検討を行い，また診療データベースの適用先の一つである臨床試験に着目し，診療データベースを用いたカルテスクリーニングの有効性について評価を行う．[1)]．

3　本研究で取り組むタスク

固有表現抽出 (Named Entity Recognition: NER)，関係抽出 (Relation Extraction: RE) とルールベースに基づく診療データベースの構築，構築した診療データベースの評価，そして診療データベースを用いたルールベースによるカルテスクリーニングの評価を行う．本研究の概要を図 1 に示す．

4　実験データ

東北大学病院において作成された診療テキストを実験データとした使用した．NER と RE の学習データについて節 4.1，診療データベースの構築と評価に使用するデータを節 4.2，カルテスクリーニングに用いるデータについて節 4.3 でそれぞれ説明する．

4.1　固有表現抽出と関係抽出

東北大学病院で作成された 9 診療科の経過記録に出現するの固有表現と関係に対してアノテーションを行った．アノテーションは篠原らによって考案されたアノテーション方法 [9] を参考に，経過記録用に内容をいくつか修正し実施した．

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．5 名のアノテーターでアノテーションを行い，126 種類の固有表現と 39 種類の関係がアノテーションされた経過記録 9,698 文を作成した．学習データの統計情報を Appendix の表 3 に，アノテーション例を Appendix の図 3 に示す．

4.2　診療データベースの構築と評価

診療データベースの項目は実際に東北大学病院の婦人科で作成されている診療データベースの項目に基づいて決定した．項目の詳細を以下に示す．なお本研究では子宮体癌を対象としたため，病理情報や手術実施情報は子宮体癌に関連する情報である．

- 病理情報: 組織型，組織異型度，TNM，進行期，脈管侵襲，筋層浸潤，腹水細胞診
- 再発リスク分類: 再発リスク
- 手術実施の有無: 子宮全摘，BSO，PLN，PAN，大網切除
- 化学療法: 化学療法の有無，レジメン，回数，開始日，終了日
- その他: 経妊，経産

東北大学病院の婦人科において 2023 年に作成された 46 名分の子宮体癌の診療データベースを用い，診療データベースの構築と評価を行う．NER と RE の結果からルールベースにより診療データベースの各項目を決定する (例えば，TNM は ent: 病名と ent: TNM が rel: value_of で結ばれるなどのルールをアノテーションガイドラインから作成する)．そのため，46 名のうち 28 名をルール作成用，18 名を評価用に分割して使用した．1 患者ごとに経過記録と病理レポートをそれぞれ最大で 2 年分取得し使用した．なお，1 患者あたり経過記録は平均 77 件あり文字数は平均 7,246 文字，病理レポートは 5.8 件で 2,090 文字であった．

4.3　カルテスクリーニングの評価

2018 年から 2020 年の間に東北大学病院の婦人科を受診した一部の患者 (348 名) を対象とし，手術オーダーから各患者の手術日を特定し，初診日から手術前日までの経過記録と病理レポートを使用した．これは本研究で対象とした臨床試験が術前の情報を使用するためである．348 名のうち，6 名は婦人科領域で実施されたある臨床試験に参加しており，診療データベースからルールベースにより選択基準の各項目を満たすか満たさないかを判定する．なお，1 患者あたり経過記録は平均 15 件あり文字数は平均 3,749 文字，病理レポートは 9.1 件で 1,063 文字であった．

1)　倫理的配慮: 本稿で説明する内容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．また本研究で使用した診療テキスト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，データの利用を行った．

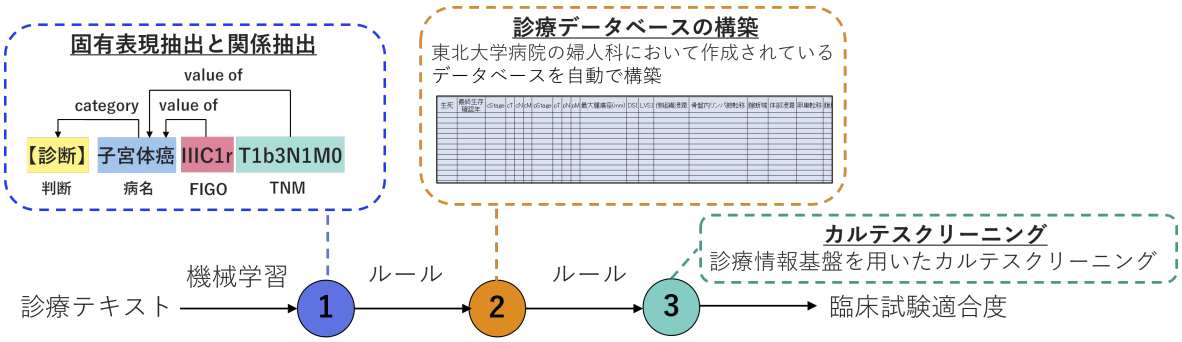

図 1　本研究の概要: 1 は教師あり学習，2 と 3 はルールベースにより実施する．

5　実験設定

5.1　固有表現抽出と関係抽出

NER には Bidirectional Encoder Representations from Transformers (BERT)[10] に条件付き確率場を接続したモデルを，RE には Head Selection 問題として定式化するモデル [11, 1] を使用した．入力データの最大トークン数は 400 に設定し，NER の学習では学習率を 1e-5，バッチサイズを 64，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 NER と同じ設定とし，それ以外の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 Wikipedia で事前学習されたモデル [12] を使用した．評価は 10 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．評価方法の詳細は先行研究 [13] と同様であるため省略する．

5.2　診療データベースの構築と評価

28 名分の経過記録と病理レポートに対して NER と RE を行い，それらの結果とアノテーションガイドラインを参照し，診療データベースの各項目を抽出するルールを作成した．その後，18 名分の経過記録と病理レポートから診療データベースを自動で作成し (BERT+ルール)，評価を行った．

評価においては，各項目が正解データと一致した場合に正解，それ以外を不正解とし正解率を算出した．なお，経過記録と病理レポートで記載が異なる場合があったため，病理レポートと経過記録からそれぞれ各項目に対応する情報を 1 つずつ抽出し，どちらか一方でも正解データと一致していれば正解として扱った．

5.3　カルテスクリーニングの評価

5 つの選択基準を対象とし，選択基準を満たす割合を示す適合率を患者ごとに算出し，適合率ごとに集約することで Recall と候補患者数を算出する (臨床試験に参加した患者の適合率は 1.0 になると仮定する)．カルテスクリーニングにおいては抽出漏れを減らすことが重要であるため，Recall が 1.0 となる時の候補患者数で評価を行う．評価のイメージを図 2 に示す．

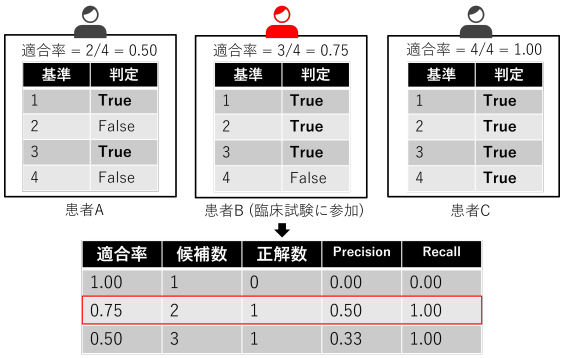

図 2　カルテスクリーニングの評価方法

本研究で使用した選択基準 (実験のため一部文言を変更) を以下に示す．PS は節 4.2 の項目にはないが，本実験において追加した (正解率は未評価である)．判定では，例えば基準 1 であれば診療データベースの組織型が類内膜腺癌，粘液性腺癌，漿液性腺癌，明細胞腺癌，未分化癌，混合癌のいずれかの文字列を含んでいれば True，そうでなければ False とし，文字列一致による判定を行った．

1. 組織型が類内膜腺癌，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
2. 進行期が IB 期，II 期，IIIA 期，IIIB 期，IIIC 期のいずれかである
3. 腹膜播種，遠隔臓器転移，鼠径リンパ節転移を

認めない (TNM が M1，進行期が IVB 期でない)
4. 膀胱浸潤，直腸浸潤のいずれも認めない (TNM が T4 でない，進行期が IVA 期でない)
5. Performance status (PS) が 0 または 1

6 結果

6.1 固有表現抽出と関係抽出

実験により，NER は F1 で 0.910，RE は 0.792 であり，先行研究の結果 [14, 15, 16] と比較しても遜色ない精度が得られていることが確認された．

6.2 診療データベースの構築

実験結果を表 1 に示す．なお参考値として，GPT-OSS-20B (GPT)[17] を用いて診療テキストから診療データベースを直接生成した場合の結果も併せて示す．表 1 から，BERT+ルールベース (BERT) の正解率の平均値は 0.899，GPT は 0.841 であり，提案手法は 1 患者あたり 35 秒，GPT は 113 秒の処理時間を要することが確認された．

表 1 診療データベース構築の実験結果

項目	BERT	GPT	項目	BERT	GPT
組織型	0.833	0.833	PAN	1.000	1.000
異形度	1.000	0.944	大網切除	1.000	0.889
pT	1.000	1.000	経妊	0.944	0.722
pN	0.722	0.778	経産	0.889	0.722
pM	0.944	0.889	脈管侵襲	0.833	0.833
進行期	0.944	1.000	腹水細胞診	0.778	0.944
再発リスク	0.778	0.722	化学療法	1.000	0.500
筋層浸潤	0.667	0.889	回数	0.833	0.722
子宮摘出	1.000	1.000	レジメン	0.944	0.722
BSO	0.833	0.944	開始日	1.000	0.778
PLN	1.000	1.000	終了日	1.000	0.778

6.3 カルテスクリーニング

実験結果を表 2 に示す．Recall が 1.000 の時，候補患者数は 68 名であり，348 名から 68 名まで候補患者を絞り込めることが確認された．

表 2 カルテスクリーニングの実験結果

適合率	候補数	正解数	Precision	Recall
1.00	8	5	0.635	0.833
0.80	68	6	0.088	1.000
0.60	184	6	0.033	1.000
0.40	243	6	0.025	1.000
0.20	261	6	0.023	1.000
0.00	348	6	0.017	1.000

7 考察

7.1 固有表現抽出と関係抽出

婦人科の経過記録 2,741 文のみで学習した場合の精度 [13] と比較して，NER の F1 は+0.026 (0.884 から 0.910)，RE は+0.024 (0.768 から 0.792) の改善が確認された．一方，データの増加量 (2,741 文から 9,701 文) に比べ，精度の上昇量は微量であり，精度の上昇が鈍化している可能性がある．そのためアノテーション済みデータの確認だけではなく，基盤モデル自体の事前学習，もしくは追加事前学習なども含めた精度改善の手法を検討していく必要がある．

7.2 診療データベースの構築

正解率の平均値，1 患者あたりの処理に要する処理時間ともに BERT を用いた手法の方が高く，その有用性が示唆された．一方，GPT ではプロンプトや処理フローには改善の余地があり，GPT の方が正解率が高い項目もあることから継続した調査が必要である．

7.3 カルテスクリーニング

構造化データ (病名コードと手術情報) では 348 名まで候補患者を絞り込むことができたが，ここから診療データベースを用いることで Recall が 1.0 の状態で 68 名まで候補患者を絞り込むことができることが明らかとなった．カルテスクリーニングには 1 患者あたり 30 分から 80 分の時間が必要であるため [13]，最低でも 174 時間必要であった作業を 34 時間まで短縮できる可能性がある．実際に臨床試験に参加した患者のうち 1 名は適合率が 0.8 となったが，これは TNM に関する項目が抽出できなかったことが原因であり，NER，RE を行うモデルの改良や入力する情報源の追加 (例: 読影レポート) などによる対応を行う予定である．

8 おわりに

本研究では，NER，RE とルールベースによる診療データベースの構築と評価，診療データベースを用いたカルテスクリーニングの評価を行い，機械学習により実際の診療現場で作成されたデータベースを一定の精度で再現できることを確認した．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適用できる可能性が示唆された．

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の先生方，オープン・ベッドラボのスタッフの皆様，臨床研究推進センターバイオデザイン部門のスタッフの皆様に厚く御礼申し上げます．

参考文献

[1] Fei Cheng, Shuntaro Yada, Ribeka Tanaka, Eiji Aramaki, and Sadao Kurohashi. JaMIE: A pipeline Japanese medical information extraction system with novel relation annotation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 3724–3731, Marseille, France, June 2022. European Language Resources Association.

[2] 土橋大樹, 平田健司, 渡邊史郎, 竹中淳規, 若林直人, 木村理奈, 坂本圭太, 工藤與亮. 大規模言語モデルを用いた読影レポートからの情報抽出: Chatgpt3. 5, chatgpt4 および google bard の比較. 北海道放射線医学雑誌= Hokkaido Journal of Radiology, Vol. 4, pp. 7–12, 2024.

[3] Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K Keloth, Vincent J Zhang, Ruey-Ling Weng, Qingyu Chen, Xiaoqian Jiang, et al. Information extraction from clinical notes: are we ready to switch to large language models? **arXiv preprint arXiv:2411.10020**, 2024.

[4] J Beattie, S Neufeld, DX Yang, C Chukwuma, A Gul, NB Desai, M Dohopolski, and SB Jiang. Utilizing large language models for enhanced clinical trial matching. **International Journal of Radiation Oncology, Biology, Physics**, Vol. 120, No. 2, p. e611, 2024.

[5] Yizhao Ni, Monica Bermudez, Stephanie Kennebeck, Stacey Liddy-Hicks, and Judith Dexheimer. A real-time automated patient screening system for clinical trials eligibility in an emergency department: design and evaluation. **JMIR medical informatics**, Vol. 7, No. 3, p. e14185, 2019.

[6] Jacob Beattie, Dylan Owens, Ann Marie Navar, Luiza Giuliani Schmitt, Kimberly Taing, Sarah Neufeld, Daniel Yang, Christian Chukwuma, Ahmed Gul, Dong Soo Lee, et al. Large language model augmented clinical trial screening. **medRxiv**, pp. 2024–08, 2024.

[7] Ozan Unlu, Matthew Varugheese, Jiyeon Shin, Samantha M Subramaniam, David Walter Jacques Stein, John J St Laurent, Charlotte J Mailly, Marian J McPartlin, Fei Wang, Michael F Oates, et al. Manual vs ai-assisted prescreening for trial eligibility using large language models—a randomized clinical trial. **JAMA**, Vol. 333, No. 12, pp. 1084–1087, 2025.

[8] James Booker, Jack Penn, Kawsar Noor, Richard JB Dobson, Naomi Fersht, Jonathan P Funnell, Ciaran S Hill, Danyal Z Khan, Nicola Newall, Tom Searle, et al. Utilising natural language processing to identify brain tumor patients for clinical trials: Development and initial evaluation. **World Neurosurgery**, Vol. 197, p. 123907, 2025.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 柴田大作, 辻川剛範, 渋谷恵, 森田智子, 久保雅洋, 中川敦寛, 重田昌吾, 島田宗昭. 診療データベースを用いたカルテスクリーニング. 言語処理学会第 31 回年次大会. 言語処理学会, 2025.

[14] 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.

[15] Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). **Lang. Resour. Eval.**, Vol. 52, No. 2, p. 571–601, June 2018.

[16] 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.

[17] OpenAI. gpt-oss-120b gpt-oss-20b model card, 2025.

A NER と RE の学習データについて

NER と RE の学習データの統計情報を表 3 に，アノテーション例を図 3 に示す．

表 3 NER と RE の学習データの統計情報: 文字数，固有表現数，関係数は平均値を示す．

診療科	データ数	文字数	固有表現数	関係数
婦人科	2,654	60.3	11.5	8.9
消化器内科	2,016	90.6	17.5	14.0
呼吸器内科	1,672	77.2	10.8	8.6
呼吸器外科	1,143	75.3	13.2	11.1
循環器内科	726	52.6	10.8	8.3
泌尿器科	528	99.0	17.1)	14.1
耳鼻科	460	54.8	12.0	9.3
乳腺科	360	65.1	15.0	12.5
皮膚科	139	61.5	12.3	10.1

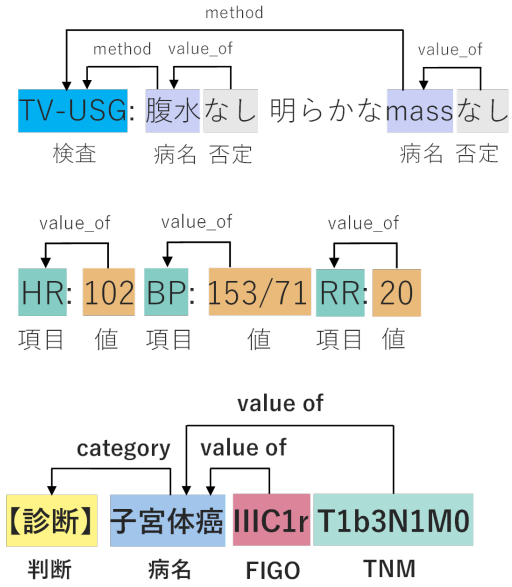

図 3 アノテーション例

B Limitation

他施設で作成された診療テキストに対する NER, RE とカルテスクリーニングの精度評価は未実施である．またカルテスクリーニングにおいては，選択基準を現時点で判定が可能と考えられる項目を用いて実施したため，今後，より複雑な選択基準 (医師の判断を要する重篤な合併症についてなど) について検討を行う必要がある．