経理様式１別添

令和6年度　成果報告書

基本情報（公開）

事業名 プログラム名		戦略的イノベーション創造プログラム（SIP） 課題「統合型ヘルスケアシステムの構築」
研究開発課題名		D-2 統合型の医学概念・知識連結データベースの構築及び医療文書の自動分析基盤の整備
研究開発担当者＊1	機関名	奈良先端科学技術大学院大学
	所属	先端科学技術研究科
	役職	教授
	氏名	荒牧 英治

実施期間＊2	令和5年9月22日～令和8年３月31日

＊1 委託研究開発契約書に定義

＊2 年度の契約に基づき、本委託研究開発を行った期間又は中止までの期間

1. 研究開発テーマ概要

1.1 研究開発内容

SIP「統合型ヘルスケアシステムの構築」のミッションである「知識発見」と「医療提供」の循環を実現するためには，情報を収集・統合・分析する必要がある．これには，電子カルテ（EHR）や患者記録（PHR）など様々なテキストから構造化された医療知識を抽出する必要がある．そこで本テーマ（D-2）では，既存の医療・医学関連の辞書やリソースをまとめた大規模な医学概念・知識連結データベース（以下，「医学概念 DB」という）を40万語規模で構築する．さらに，電子カルテなど，病院内で作成される自然言語で書かれる医療情報を医学概念に変換する機構も併せて開発し，病院内のデータを医療デジタルツインに投入するための医療言語処理基盤とする．最終的には，FHIR 準拠の医療情報システムに医学概念 DB や医療言語処理基盤を搭載することで，主要な国内ベンダー4 社以上の電子カルテで活用可能として医療者の業務効率化を支援するとともに，病院データや文書を患者が理解しやすいように言い換えるといった患者向けのアプリケーションを開発し，社会還元する． 具体的には，大きく3つの目標の達成に向けて逐次的に研究開発を遂行する．

目標1：医学概念 DB 開発

まず，医学概念 DB の構築を概ね完了する．これまで辞書開発に実績の多い奈良先端科学技術大学院大学（以下，「NAIST」という）・荒牧らのグループ（以下，総称して「NAIST・荒牧グループ」という）が中心となって，これまでのノウハウや開発体制（アノテーター・作業環境）を最大限活かし，かつ，社会実装を想定しながら行う． すでに本邦には，万病辞書など大規模な辞書や，ICD-10 に紐付け可能な標準病名マスターという海外の SNOMED-CT などの巨大オントロジーと比べることができるリソースが存在している．問題は，どのように概念 DB を使うかの部分であり，目標3に掲げた直接の知識の紐付け先としてだけでなく，知識グラフとして多様な言語処理タスクの精度を向上させ，患者相談分類や FAQ 検索を実用レベルにするなど，多くの社会実装を想定した開発を行う．

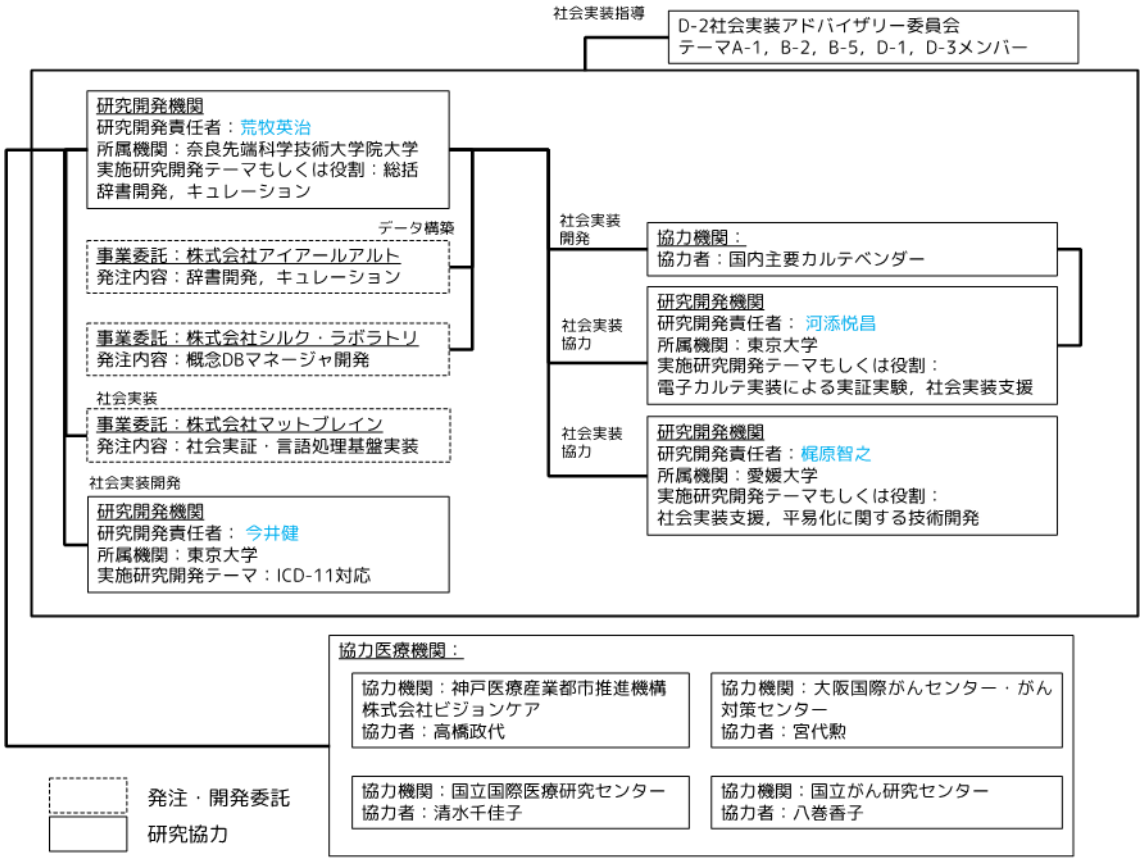
目標2：医療言語処理基盤開発

次に，自然言語で記述された医療データから構造化情報を取得するための医療言語処理基盤を開発する．テーマ D-1 と連携し，提供されたレポートデータ（放射線レポート，病理レポート，内視鏡，エコー）をもとに実証実験を行う．テーマ E-2 で開発される大規模言語モデルや B-3 から提供される症例の構造解析技術を導入し，精度向上を行う．

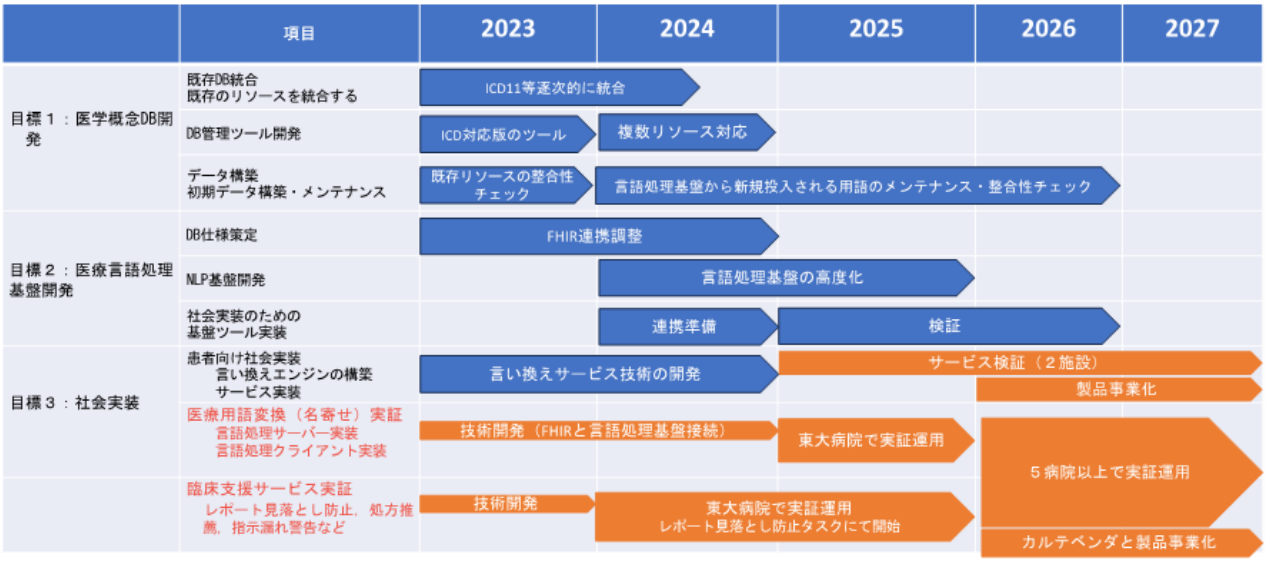
目標3：社会実装

上記，医学概念 DB や医療言語処理基盤を電子カルテ主要ベンダー4社以上にて検証を行う．具体的には，テーマ D-1 が整備する FHIR Terminology サーバなど既存のフレームワークの辞書コンテンツに医学概念 DB を搭載し，前処理ソフトウェアとして医療言語処理基盤を搭載して機能拡張することで，電子カルテ主要ベンダーが容易に接続可能とする．このように電子カルテに搭載するだけでなく，患者向け用語平易化サービスを構築し，医療者のみならず患者も社会実装を体験できるようにする．

1.2 研究開発実施体制



1.3 研究推進のスケジュール



2. 本年度の成果・進捗の概要

目標1：医学概念 DB 開発

概ね順調である. 臨床テキストに出現する表現を 100 万用語以上抽出した. 頻度によるフィルタリングにより, 病名辞書 53.5 万用語について, 頻度の高いものを中心に, 標準病名マスターの標準病名, ICD-10, UMLS, ICD-11, Lilak, 医学会医学用語辞典への自動コーディングをほぼ完了した. このように病名部分はベースとなる万病辞書があり順調である. 同様に, 医薬品(11.3 万用語), 部位(4 万用語), 検査表現(4.6 万用語)についても作業を完了しているが, 医薬品における薬効分類, 部位における正規形など, 不十分な点もある. これらを 2024 年度に補完する.

その他, 辞書管理ソフトを構築し, 社会実装アドバイザ委員の先生方への共有なども行い, 開発項目は多いものの優先順位を立てながら, 着実に進めているものと考えている.

さらに, 2024 年度には, 正規形など十分でなかった部分を LLM による半自動精査によって修正を開始した.これにより, 大幅に精査効率が向上し, 従来頻度上位１万語を対象にしていたものを全件対象とすることができた.

目標2：(辞書開発のための)医療言語処理基盤(解析基盤)

概ね順調である. 解析基盤は, まずは, 辞書の用語抽出に用い, 以降は自動メンテナンスや社会実装に用いる. 今回, 症例報告文書から 100万語以上の多くの語を抽出できたことから, 未知の医療用語をサルベージする目的に対しては十分な精度があると考えている.

2024 年度より, WebAPI を開発して公開している.ただし, 仕様が複雑なため, 利用には一定の開発力が必要な状態であった.そこで従来の MedNERN の辞書に追加する形で公開した.

目標3：(辞書を用いた)サービスの社会実装

社会実装は, カルテに組み込む病院向け社会実装と患者が利用する患者向けサービスの社会実装の2つを予定している.

前者のカルテ組み込みの社会実装では, 本格的なカルテ組み込みは, テーマ D-1 の FHIR 連携と連動して行う. 当初は, 2024 年度の予定であったが, D-1 に合わせて 3 年目とする. ただし, FHIR 連携以前でも, 読影レポート中の悪性腫瘍警告などを題材に, 社会実装を試みた. この結果, 解析基盤の精度が十分でないなどの問題も明らかになった. この問題を解決するために, アノテーション作成人材を雇用し, 東大のチューンした解析基盤の構築を来年から予定している.

患者向けの社会実装では辞書と解析基盤をベースにしたチャットボットを開発し, 事業化を目指す.

2024年度は, 以下のように3つの社会実装先での準備を進めた. 特に, がん相談 FAQ チャットボットは 2024年度, 患者パネルを用いてのパイロットスタディを行った.

●遺伝カウンセリングチャットボットプロタイプを構築し, ビジョンケアとの共同研究を開始し, チャットボットの安全性に関する特許を申請した.

●がん相談 FAQ チャットボットを構築し, 難解な医療用語を含む文書をわかりやすく説明する技術(医療テキスト平易化)の研究に取り組んだ.テキスト平易化モデルを開発するための日本語の言語資源は乏しいため, 2024年度は主にコーパスの整備に取り組んだ.この結果, テキスト平易化

モデルを訓練するためのパラレルコーパス 1.6 万文対を構築や医療ドメインに特化したパラレルコーパス 1,425 文対を構築した.これらの研究成果は, 査読付き論文誌および査読付き国際会議に採択され, 言語処理学会年次大会において若手奨励賞を受賞した.

また, 医学概念 DB 開発の一環として, 医療用語の患者向け難易度推定の研究にも取り組んだ.まず, 医療用語の難易度アンケートを大規模に実施し, 20 代から 50 代までの男女から回答を得た.そして, 収集したデータに基づき 37 万語の医療用語(病名および症状表現)に対する患者向け難易度を約 8 割の精度で推定した.これらの研究成果は, 査読付き国際会議に採択された.

2024 年度からは, これを用いてカルテ情報を構造化する共同研究の実施に向けた調整を開始し, 具体的な実施内容やスケジュール等について関係者と協議を進めた.また, 国立循環器病センターでの脳卒中データベース自動構築や都立駒込病院でのインシデントレポート解析などの研究を行なった.

また, 辞書の縮小版を JMED-DICT mini としてオープンソースで公開し, 現在まで約 100 のダウンロードがある.すでに, 商用製品に搭載されたものもあり, 短期間ながら普及効果があるものと考えている.

3. 成果物の公表

3.1 論文など(原著論文、学位論文、プロシーディングス、総説、解説、速報など)

論文数(総数)	(内国際誌)	(内国内誌)
6	6	0

1. Yoshimasa Kawazoe, Masami Tsuchiya, Kiminori Shimamoto, Tomohisa Seki, Emiko Shinohara, Shuntaro Yada, Shoko Wakamiya, Shungo Imai, Eiji Aramaki, and Satoko Hori: Natural language processing of electronic medical records identifies cardioprotective agents for anthracycline induced cardiotoxicity, Scientific Reports, 15:6678 (2025/2/24)
2. Xinbai Li, Shaowen Peng, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki: GenKP: generative knowledge prompts for enhancing large language models, Applied Intelligence, Vol. 55, No. 464 (2025/2/19)
3. Takuya Fukushima, Masae Manabe, Shuntaro Yada, Shoko Wakamiya, Akiko Yoshida, Yusaku Urakawa, Akiko Maeda, Shigeyuki Kan, Masayo Takahashi, and Eiji Aramaki: Evaluating and Enhancing Japanese Large Language Models for Genetic Counseling Support: Comparative Study of Domain Adaptation and the Development of an Expert-Evaluated Dataset, JMIR Medical Informatics 2025;13:e65047 (2025/1/16)
4. Gabriel Herman Bernardim Andrad, Shuntaro Yada, and Eiji Aramaki: Is Boundary Annotation Necessary? Evaluating Boundary-Free Approaches to Improve Clinical

Named Entity Annotation Efficiency: Case Study, JMIR Medical Informatics, 12:e59680, 2024 (2024/7/2)

5. Yukiko Ohno, Riri Kato, Haruki Ishikawa, Tomohiro Nishiyama, Minae Isawa, Mayumi Mochizuki, Eiji Aramaki, and Tohru Aomori: Using the natural language processing system MedNER-J to analyze pharmaceutical care records: Natural language processing analysis, JMIR Formative Research, 8:e55798, 2024 (2024/5/9)
6. Zhouqing Zhang, Kongmeng Liew, Roeline Kuijer, Wan Jou She, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki: Differing Content and Language Based on Poster-Patient Relationships on the Chinese Social Media Platform Weibo: Text Classification, Sentiment Analysis, and Topic Modeling of Posts on Breast Cancer, JMIR Cancer, 10:e51332, 2024 (2024/5/9)

3.2 学会発表など((国内・国際)学会口頭発表・ポスター発表、招待講演で成果を公表したもの)

1. Seiji Shimizu, Shuntaro Yada, Lisa Raithel, and Eiji Aramaki: Improving Self-training with Prototypical Learning for Source-Free Domain Adaptation on Clinical Text, In Proceedings of the 23rd Workshop on Biomedical Language Processing (BioNLP), pp. 1-13 (2024/8/16, Bangkok, Thailand)
2. Lisa Raithel, Philippe Thomas, Bhuvanesh Verma, Roland Roller, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Shoko Wakamiya, Eiji Aramaki, Sebastian Möller, and Pierre Zweigenbaum: Overview of #SMM4H 2024 – Task 2: CrossLingual Few-Shot Relation Extraction for Pharmacovigilance in French, German, and Japanese, In Proceedings of the 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks (#SMM4H 2024), pp. 170-182 (2024/8/15, Bangkok, Thailand)
3. Soichiro Sugihara, Tomoyuki Kajiwara, Takashi Ninomiya, Shoko Wakamiya,and Eiji Aramaki: Semi-automatic Construction of a Word Complexity Lexicon for Japanese Medical Terminology, In Proceedings of the 6th Clinical Natural Language Processing Workshop (Clinical NLP Workshop 2024) (2024/6/20-21, Mexico City, Mexico)
4. Lisa Raithel, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Philippe Thomas, Tomohiro Nishiyama, Sebastian Möller, Eiji Aramaki, Yuji Matsumoto, Roland Roller, and Pierre Zweigenbaum: A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages, In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 395-414 (2024/5/23, Torino, Italy)

5. Seiji Shimizu, Lis Pereira, Shuntaro Yada, and Eiji Aramaki: QA-based Event Start-Points Ordering for Clinical Temporal Relation Annotation, In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 13371-13381 (2024/5/24, Torino, Italy)
6. 大槻 優佳, 矢田 竣太郎, 西山 智弘, 工藤 紀子, 川端 京子, 藤牧 貴子, 永井 宥之, 若宮 翔子, 荒牧 英治:大規模言語モデルを活用した大規模医療用語辞書メンテナンス の効率化, 言語処理学会第 31 回年次大会 (NLP2025), B9-3(2025/3/13, 出島メッセ長崎)
7. 永井 宥之, 西山 智弘, 大槻 優佳, 藤牧 貴子, 川端 京子, 工藤 紀子, 山崎 由佳, 白石 暖哉, 梶原 智之, 進藤 裕之, 河添 悦昌, 今井 健, 矢田 竣太郎, 若宮 翔子, 荒牧 英治:JMED-DICT: 大規模医療用語辞書の構築, 言語処理学会第 31 回年次大会 (NLP2025), B9-2(2025/3/13, 出島メッセ長崎)委員特別賞
8. 林 純子, 伊藤 和浩, 永井 宥之, 矢田 竣太郎, 若宮 翔子, 荒牧 英治:多様な客観的 解釈を反映した主体性コーパス構築と予備的分析, 言語処理学会第 31 回年次大会 (NLP2025), B8-2 (2025/3/13, 出島メッセ長崎)
9. 栗生 紗希帆, Dayeon Kim, Hyuk-Yoon Kwon, 若宮 翔子, 荒牧 英治:ソーシャル メディアテキストを用いた摂食障害の文化差比較, 言語処理学会第 31 回年次大会 (NLP2025), Q6-22(ポスター発表)(2025/3/12, 出島メッセ長崎)
10. 中岡 明義, 若宮 翔子, 荒牧 英治:行動分類のためのコーパス構築と行動分析への応 用, 言語処理学会第 31 回年次大会 (NLP2025), Q6-1(ポスター発表)(2025/3/12, 出島メッセ長崎)
11. 橋本清斗, 清水聖司 , 工藤紀子, 矢田竣太郎, 若宮翔子, 江本駿, 西村由希子, 荒牧英 治:質的研究の自動化:患者自由記述テキストからの潜在的トピックの発見, 言語処理 学会第 31 回年次大会 (NLP2025), P6-1(ポスター発表)(2025/3/12, 出島メッセ 長崎)
12. 清水 美緒奈, 林 純子, 久田 祥平, 若宮 翔子, 荒牧 英治, 大内 啓樹:場所表現の地 理的曖昧性を解消するための質問内容生成, 言語処理学会第 31 回年次大会 (NLP2025), E3-2(2025/3/11, 出島メッセ長崎)
13. Zhiwei Gao, 清水 伸幸, 藤田 澄男, Shaowen Peng, 若宮 翔子, 荒牧 英治:Cultural Adaptability of Multilingual Large Language Models: A Comparative Study in Japanese Workplace Contexts, 言語処理学会第 31 回年次大会 (NLP2025), Q2-10 (ポスター発表)(2025/3/11, 出島メッセ長崎)
14. 西岡 竜生, 若宮 翔子, 清水 伸幸, 藤田 澄男, 荒牧 英治:JHACE: Human-AI Collaboration の評価法の提案, 及び, 対人スキルの影響の調査, 言語処理学会第 31 回 年次大会 (NLP2025), Q2-1(ポスター発表)(2025/3/11, 出島メッセ長崎)
15. 辻 航平, 平岡 達也, Cheng Yuchang, 荒牧 英治, 岩倉 友哉:誤字に対する Transformer ベース LLM のニューロンおよびヘッドの役割調査, 言語処理学会第 31 回年次大会 (NLP2025), P2-11(ポスター発表)(2025/3/11, 出島メッセ長崎)
16. 伊藤 和浩, 矢田 竣太郎, 若宮 翔子, 荒牧 英治:イノベーティブな言語使用は集団的アイデンティティの指標になりうるか?, 言語処理学会第 31 回年次大会 (NLP2025), E2-1(2025/3/11, 出島

メッセ長崎）

17. 藤川 直也, 伊藤 和浩, 若宮 翔子, 荒牧 英治：書き手の孤独感を予測できるか？, 言語処理学会第 31 回年次大会 （NLP2025）, Q1-1（ポスター発表）（2025/3/11, 出島 メッセ長崎）委員特別賞
18. 橋本 清斗, 工藤 紀子, 矢田 竣太郎, 若宮 翔子, 荒牧 英治：LLM を用いた自由記述 アンケートの質的分析, 2024 年度 ナラティブ意識学ワークショップ「脳・言語・意識」（2024/9/24, 釧路市観光国際交流センター）
19. 祖父江 智子, 伊藤 和浩, 古賀 千絵, 吉村 有司, 若宮 翔子, 荒牧 英治：パブリック アートは人を幸せにするのか, 第 19 回 YANS シンポジウム, S4-P26（ポスター発表）（2024/9/6, 梅田スカイビル）
20. 峯 悠大, 西山 智弘, 谷 懿, 大竹 義人, 佐藤 嘉伸, 矢田 竣太郎, 若宮 翔子, 荒牧 英治：LLM を用いた CT 読影レポートの撮影目的分類, 第 19 回 YANS シンポジウム, S3-P25（ポスター発表）（2024/9/5, 梅田スカイビル）
21. 大槻 優佳, 矢田 竣太郎, 工藤 紀子, 川端 京子, 藤牧 貴子, 永井 宥之, 若宮 翔子, 荒牧 英治：自己改善する辞書: Sustainable Dictionary Grooming system (SDGs), 第 19 回 YANS シンポジウム, S3-P01（ポスター発表）（2024/9/5, 梅田スカイビル）
22. 橋本 清斗, 工藤 紀子, 矢田 竣太郎, 若宮 翔子, 荒牧 英治：LLM を用いた自由記述 アンケートの質的分析, 第 19 回 YANS シンポジウム, S2-P05（ポスター発表）（2024/9/5, 梅田スカイビル）企業賞（シェルパ・アンド・カンパニー賞）
23. 福島 拓也, 久田 祥平, 矢田 竣太郎, 若宮 翔子, 荒牧 英治：日本語医療 LLM 評価ベ ンチマークの構築と性能分析, 第 19 回 YANS シンポジウム, S1-P37（ポスター発表）（2024/9/5, 梅田スカイビル）
24. 福島 拓也, 久田 祥平, 矢田 竣太郎, 若宮 翔子, 荒牧 英治：日本語医療 LLM 評価ベ ンチマークの構築と性能分析, 情報処理学会 第 261 回自然言語処理研究発表会（2024/9/3, 梅田スカイビル）若手奨励賞
25. 清水 聖司, 矢田 竣太郎, 若宮 翔子, 荒牧 英治：RECORD TWIN: 病歴を保ちつつ表 現が異なる症例の生成, 情報処理学会 第 260 回 自然言語処理研究発表会（2024/6/28, 北陸先端科学技術大学院大学 （JAIST） 金沢駅前オフィス）優秀研究賞
26. 福島 拓也, 眞鍋 雅恵, 矢田 竣太郎, 若宮 翔子, 荒牧 英治, 吉田 晶子, 浦川 優作, 前田 亜希子, 寒重之, 高橋 政代：JGCLLM: 日本語遺伝カウンセリング大規模言語モデル, 2024 年度 人工知能学会全国大会（第 38 回）, 3S5-OS-7c-01（2024/5/30, アクトシティ浜松）
27. 堀口 航輝, 梶原 智之, 二宮 崇, 若宮 翔子, 荒牧 英治：日本語医療テキスト平易化の訓練用データセットの構築, 2024 年度 人工知能学会全国大会（第 38 回）, 3S1OS-7b-04（2024/5/30, アクトシティ浜松）
28. 眞鍋 雅恵, 矢田 竣太郎, 若宮 翔子, 荒牧 英治：エピソードバンク：当事者のナラティブをデータベース化する試み, 2024 年度 人工知能学会全国大会（第 38 回）, 3S1OS-7b-02（2024/5/30, アクトシティ浜松）
29. 大槻 優佳, 大塚 皇輝, 矢田 竣太郎, 若宮 翔子, 荒牧 英治, 尾原 信行, 吉江 智秀： 臨床研究 DX の試み：脳卒中リスク因子のテキスト構造化システム, 2024 年度 人工 知能学会全国大会（第 38 回）, 2S6-OS-7a-01（2024/5/29, アクトシティ浜松）

3.3 診療ガイドライン、省令、基準、日本薬局方、添付文書改訂、国の技術文書(通知)等への反映

該当なし

3.4 研修プログラム、カリキュラム、シラバス、教材、e-learning 等の公表

該当なし

3.5「国民との科学・技術対話」に対する取り組み

1. 荒牧 英治: Frontiers of Generative AI based Medical Applications, Diabetes Research Innovation Symposium, 2024 (2024/7/6, オンライン)(招待講演)
2. 荒牧 英治: LLM(大規模自然言語モデル)の医療応用:単なる情報抽出を超えて, どう研究・臨床に使うか?, 日本消化器病学会ビッグデータ・AI 研究会, 2024 (2024/6/23, ステーションコンファレンス東京)(招待講演)

3.6 その他

上記のほかに書籍出版や報道、展示会参加等のアウトリーチ活動実績がありましたら記載ください。

該当なし

以上