

データ駆動・AI 駆動を中心としたデジタルトランスフォーメーションによる生命  
科学研究の革新

2022 年度採択研究代表者

2022 年度  
年次報告書

荒牧 英治

奈良先端科学技術大学院大学 先端科学技術研究科  
教授

リアルワールドテキスト処理の深化によるデータ駆動型探査

主たる共同研究者:

河添 悦昌 (東京大学 大学院医学系研究科 特任准教授)

堀 里子 (慶應義塾大学 薬学部 教授)

## 研究成果の概要

本研究の基盤となる1)データベース(DB)の構築, 2)ツールの整備, 3)創薬シーズの抽出を行った. 構築したデータベースは医療テキスト(いわゆる電子カルテ)を処理する研究としては世界最大規模である. 今後, この DB を用いた研究を網羅的に行う予定である.

### データベースの構築

過去 18 年間に東大病院に入院した, 入院時 16 歳以上 100 歳未満の患者を対象として, 1) DPC (診断群分類包括評価)データ, 2) 処方・注射オーダー, 3) 検体検査結果, 4) 退院サマリの症状・所見を DB 化した. 退院サマリテキストに対しては, 後述する NLP-DX プラットフォームを適用し, 抽出された患者の症状や所見文字列で構成した. この DB から DPC レコードをがん症例(138,656 症例)に絞り込んだサブセットを抽出し, 研究用 DB として NAIST・慶應大と共有した.

### NLP-DX プラットフォームの整備

上記 DB の構築に使用した電子カルテを処理する医療用語抽出システム MedTXT-CR-JA を公開した.

### 探薬シーズの抽出

抗がん剤誘発性の肝機能障害, 末梢神経障害, 手足症候群, 嘔気嘔吐, 口内炎を対象として, 既報の原著論文の網羅的探索を行い(検索語で抽出した論文 900 報), 基礎研究の知見とそのメカニズムに基づき, リアルワールドテキストでの検証に値するシナジー薬の有力候補の選定を行なった.

今後, DB から得られたシーズ薬候補の効果検証を進めていく.