

厚生労働科学研究費補助金
（政策科学総合研究事業（臨床研究等ICT基盤構築・人工知能実装研究事業））
総括研究報告書

課題名 : 新薬創出を加速する症例データベースの構築・拡充/創薬ターゲット推定アルゴリズムの開発
 研究代表者名 : 夏目やよい（国立研究開発法人医薬基盤・健康・栄養研究所 AI 健康・医薬研究センター）

研究要旨

本事業では、「創薬ターゲットの枯渇問題」を克服すべく、動物からではなくヒトの情報から創薬ターゲット分子を探索する AI の開発実装を目的とする。

令和 2 年度は、肺がん手術検体及びバイオプシー検体のオミックス解析を行うとともに、神奈川循環器呼吸器疾患センターで IPF を含む間質性肺炎患者の臨床情報（オミックスデータ及びそれと紐づけられた診療情報）を収集した。本事業で開発された患者層別化 AI を用い、大阪大学で収集された臨床情報の解析を行うことで IPF の特徴に該当する診療情報の項目に紐づけられる血中タンパク質をデータ駆動的に複数個見出した。これらのタンパク質のバイオインフォマティクス解析により、IPF の創薬標的候補を提示した。さらに、本事業の成果であるデータベースや AI を広く共有するための基盤であるオープンプラットフォームのプロトタイプ構築や格納するデータの匿名加工条件設定、運用体制を策定するための先行事例調査やニーズ調査を実施した。

A. 研究目的

医薬品開発において、近年国内外を問わず創薬ターゲットの枯渇が問題となっている。現在残されているのは高難易度の創薬ターゲットのみであるがために、新薬の研究開発には多額の費用が必要となっており、これが高薬価、ひいては医療費の高騰の要因となっている。

更に、臨床試験段階で期待していた薬効が得られず開発が中断する例が増えていることも問題点として挙げられる。特に医薬品開発の 70～80%が Phase2 で中止となっており、この約 60%が、薬効が得られなかったことが原因との報告がある。つまり、「動物では効くが、ヒトでは効かなかった」という事案が多発している。これは現在の創薬研究開発スキームの限界であると考えられる。

このような現状を打開する解決策として、人工知能 (AI; Artificial intelligence) が注目されている。AI のパフォーマンスと可能性に創薬・医療・ヘルスケア分野が大きな期待を寄せており、今後国際競争が激化することが必至である。

これらの現状を背景に、本事業では、「創薬ターゲットの枯渇問題」を克服すべく、動物からではなくヒトの情報から創薬ターゲット分子を探索する AI の開発実装を目的とする。つまり、臨床情報 (=電子カルテを始めとする診療情報+オミックスデータ) を収集・利用して創薬ターゲットを探索する AI 手法の開発をおこなう。本事業では、対象疾患として難病指定の IPF (特発性肺線維症) を含む間質性肺炎及び部位別がん死亡者数 1 位である肺がんを選択し、これらの臨床情報収集とそれを支援する基盤構築、異種かつ大量のデータを統合して創薬ターゲット候補となる生体分子群を自動的に抽出する AI 手法の開発を行う。また、本事業で作成される IPF/肺がんの疾患統合データベース、機能分子を特定するための AI 及び知識ベース等を多くの研究者等に利用してもらうための環境 (オープンプラットフォーム) の構築を目指す。

B. 研究方法

1. 既知 (背景) 情報の収集と知識ベース化: 公共・市販のデータベースを TargetMine システムに追加統合した。追加したデータベースは下記の通りである。

- ・ MGeND: 疾患名・年齢・性別などの臨床情報と遺伝子変異データとを統合的に扱うデータベース
- ・ PoSSuM: タンパク質上の低分子結合サイトを推定するためのデータベース

論文からの自動知識抽出を可能にするモデル構築のため、臨床系文献のアノテーション基準となるガイドラインの作成・修正を行った。対象となる文献 (167 件) は当所専門家により選抜され、アノテーション作成は産総研人工知能研究センターのアノテーター 3 名が独立して業務に当たることにより頑健なアノテーションを実施し、随時アノテーション中のデータを解析して一つのエンティティに複数の異なる ID がアサインされていないか確認・修正をおこなった。このプロセスは産総研人工知能研究センターのプロトコルに則って

実施された。アノテーションガイドライン作成は当所専門家及び産総研人工知能研究センターの担当研究者・アノテーターと議論の上、本事業の対象疾患である呼吸器疾患の専門用語を適切にアノテーションできるように実践に即して策定した。モデル構築は産総研人工知能研究センターへの委託により実施した。

2. 臨床情報の収集と機械可読表現の開発：肺がん手術検体及びバイオプシー検体のオミックス解析を行うとともに、大阪大学医学部呼吸器・免疫内科バイオバンク、大阪大学医学部附属病院医療情報部大阪大学病院バイオバンク及び神奈川循環器呼吸器疾患センターで IPF を含む間質性肺炎患者の臨床情報を収集した。

肺がん手術検体・バイオプシー検体のマルチオミックスデータ収集

国立がん研究センターにおいて収集された肺がん手術検体及びバイオプシー検体（検体採取・処理・保存は国立がん研究センターにおけるプロトコルに則って行われた）を用いて、下記のオミックスデータを収集した。オミックスデータ取得はタカラバイオ株式会社への外部委託により実施した。

- ・ 全ゲノム解析
- ・ 全エクソーム解析
- ・ DNA メチル化解析
- ・ ヒストン修飾（ChIP-seq）解析
- ・ トランスクリプトーム（RNA-seq）

大阪大学コホートにおける臨床情報収集

昨年度までに大阪大学医学部呼吸器・免疫内科のバイオバンクより、同意取得済みの IPF を含む間質性肺炎患者血清と、血清にひもつく診療情報（診察記録、CT 画像およびその読影所見、血液検査値、患者基本情報および初診時間診票）を大阪大学病院医療情報部より収集した。また、検査の結果、器質的な呼吸器疾患を有しないと診断された方々を健常者とし、同様に血清および診療情報を収集した。なお、血清はバイオバンク番号、診療情報は、いずれも研究用の匿名化 ID が付与された形で入手した。今年度は、上記診療情報の中で、病院側の COVID-19 対応のため、入手が遅れていた CT 画像の残り欠損データを同様に匿名化 ID が付与された形で入手した。また免疫染色を行うために、肺がんを合併して手術に至った IPF 患者肺における線維化部位および正常部位のホルマリン固定標本も入手した。呼吸機能検査値およびフローボリュームカーブについては、匿名化作業を行うため、医療情報部から呼吸内科へ転送することとした。

i. 診療情報の収集と構造化

診療情報のうち、診察記録は、必要な情報 102 項目をあらかじめリストして作成したテンプレートを用いて医師が入力もしくは自由記載事項テキストデータよりマニュアルでキュレーションを行って構造化データとした。読影所見は、マニュアルあるいは自然言語処理手法を用いて、重要表現にタグ付けを行い、部位と病変のペアとそれらのペアが認められる（positive）、認められない（negative）および疑われる（suspect）の 3 つの場合分けからなる構造化データを得た。血液検査値は、173 項目の重要項目を選択してキュレーションを行い構造化した。初診時間診票、基本情報は、診察記録のテンプレート項目に情報をキュレーションして追加した。また、構造化にあたり、欠損値補填の手法について、欠損の意味を確認するとともに、主として健常者基準値を用いて補填を行った。なお、呼吸機能データについては、匿名化処理の手法構築を行った。

ii. プロテオーム解析

血清 200 μ l から MagCapture isolation kit (Fujifilm Wako) を用いて細胞外小胞を精製した。細胞外小胞中のタンパク質をトリス(2-カルボキシエチル)ホスフィンによる還元・ヨードアセトアミドによるアルキル化を行い、トリプシン消化、脱塩を行った。前処理したサンプルを Data independent acquisition (DIA) 法を用いた LC-MS/MS 解析を実施した。データ解析は、DIA 解析ソフト Spectranout を用いて実施し、欠損値には run-wise imputation を行った。Quality control として、15 検体毎に市販血清 1 検体を加えて、サンプル調製からデータ解析までの品質保証を行った。また質量分析の Quality control として HeLa 細胞消化物の DIA 解析を実施した。

神奈川県立循環器呼吸器病センターコホートにおける臨床情報収集

神奈川循環器呼吸器疾患センターにおいても、下記の臨床情報（診療情報＋オミックスデータまたは肺組織や血液）を収集した。診療情報及び血液は同意を得られた対象疾患患者全員より取得し、採血のタイミングから最も近い診療記録と紐づけた。肺組織は、診断の上でクライオバイオプシーが必要と判断された場合及び外科手術検体が得られた場合においてのみ収集した。クライオバイオプシーを収集する場合、1 患者より

数カ所からクライオバイオプシーの採取を行い、縦に半割して一方を病理診断、もう一方をオミックスデータ取得に用いた。採取した肺組織及び血液は、下記のプロトコルで処理・保存し、1ヶ月に1回当所に輸送（三井倉庫ホールディングス株式会社に委託）し、オミックスデータ取得はタカラバイオ株式会社に委託した。

収集した診療情報

- ・ 診察記録（97 項目）
- ・ 所見
 - 超音波（5 項目）
 - 気管支鏡（19 項目）
 - 外科生検（16 項目）
 - CT 画像検査（19 項目）
- ・ 検査
 - 血液＋尿（103 項目）
 - 生理機能（35 項目）
 - CT 画像
 - 肺組織
- ・ クライオバイオプシー：採取後チューブに回収し、速やかに-80℃で保存した。
- ・ 外科手術検体（VATS）：採取後チューブに回収し、速やかに-80℃で保存した。
- ・ 上記と紐づけられた病理所見
- ・ 上記と紐づけられた病理画像からの特徴量
- ・ 血液
 - 血清（プロテオーム解析用）：採血後、室温で静かに 5 回転倒混和し、室温で 30 分間静置した後にスイング式ローターを用いて、室温、1,300×g で 10 分遠心した。上清をチューブに 250μL ずつ分注して速やかに-80℃で保存した。
 - 血漿（miRNA トランスクリプトーム解析用）：採血後、室温で静かに 10 回転倒混和し、スイング式ローターを用いて、室温、1,300×g で 10 分遠心した。上清をチューブに 250μL ずつ分注して速やかに-80℃で保存した。
 - 末梢血単核細胞（PBMC、全ゲノム解析用）：採血後、室温で静かに 5 回転倒混和し、スイング式ローターを用いて室温、1,600×g で 20 分遠心した。まとめて採血する場合、最初の検体から最後の検体の時間が 1 時間以内になるようにした。遠心後、ゲルバリアの上にある細胞層を乱さないようにして血漿層を約半分吸引し、パスツールピペットを用いて、細胞層全量をチューブに回収して速やかに-80℃で保存した。

肺組織からの核酸（DNA・RNA）同時抽出

肺組織サンプル（56 症例、80 検体）について、DNA・RNA 同時抽出キット（NucleoSpin RNA、NucleoSpin RNA/DNA Buffer Set）を用いて核酸抽出を行った。抽出された核酸はバイオアナライザーを用いて RNA 分画における DNA 混入の有無を確認し、RNA 分画は NucleoSpin RNA/DNA Buffer Set のマニュアル通りに DNase 処理を行なった。

肺組織からのマルチオミックスデータ収集

i. 全ゲノム解析（depth 30）

(1) TruSeq DNA ライブラリの作成

アコースティックソルビライザーCovaris(コバリス社)を用いて物理的に数百 bp に断片化した DNA の両末端を平滑化とリン酸化処理をした後、3'-dA 突出末端処理を行い、Index 付きアダプターを連結した。その後、アガロースゲル電気泳動によりアダプターを連結した DNA をサイズ選別し、それを鋳型とし、PCR による増幅を行い、シーケンスライブラリとした。ライブラリの品質は Agilent2100 BioAnalyzer を用いて測定した。

(2) シーケンス解析

Illumina 社の NovaSeq6000 を用いて 150base 長ペアエンドシーケンスを行い、シーケンサー付属ソフト（NovaSeq Control Software v1.6.、Real Time Analysis (RTA) v3.4.、bcl2fastq2 v2.20）を用いて塩基配列（リード配列）を得た。タグ配列に基づき塩基配列（リード配列）を検体ごとに分類して fastq ファイ

ルを取得した。

ii. DNA メチル化解析

Illumina 社の Infinium MethylationEPIC BeadChip を用いて、Infinium HD Methylation Protocol Guide, Manual Protocol (15019519 v01)に従い、ゲノム DNA を用いてメチル化解析を実施した。

(1) Bisulfite 処理

ゲノム DNA (250ng) を EZ DNA Methylation Kit (ZYMO RESEARCH 社)を用いて Bisulfite 変換し、精製、回収を行った。

(2) 全ゲノム増幅、断片化、及び精製

Bisulfite 処理した DNA はアルカリ処理後に酵素による全ゲノム増幅を行なった。その後、酵素による断片化、イソプロパノール沈殿による精製の後、バッファーに再懸濁した。これを熱変性させ、Infinium MethylationEPIC BeadChip にアプライして、48℃で約 23 時間ハイブリダイゼーションを行なった。

(3) 一塩基伸長反応とスキャニング

ハイブリダイゼーション後、BeadChip をバッファーで洗浄し、一塩基伸長反応によってプローブ末端に一塩基の標識ヌクレオチドを導入、ハイブリダイズした DNA を変性・除去し、取り込ませた標識ヌクレオチドに対する蛍光色素標識抗体染色を行い、蛍光イメージを取得した。

(4) データ解析

取得した蛍光イメージデータを GenomeStudio/Methylation Module を用いて、Background Subtraction 及び Normalization を実施して解析を行なった。

iii. トランスクリプトーム (RNA-seq)

(1) TruSeq Stranded mRNA ライブラリの作製

検体より PolyA+RNA を単離し、断片化により得られる RNA を鋳型とした逆転写反応により一本鎖 cDNA を合成した。これを鋳型として、dUTP を取り込ませた二本鎖 cDNA を合成した。得られた二本鎖 cDNA の両末端を平滑化・リン酸化処理した後、3'-dA 突出処理を行い、Index 付きアダプターを連結した。アダプターを連結した二本鎖 cDNA を鋳型とし、dUTP を持つ鎖を選択的に増幅しないポリメラーゼにより PCR 増幅を行い、シーケン斯拉イブラリとした。ライブラリの品質は Agilent2100 BioAnalyzer を用いて測定した。

(2) シーケンス解析

Illumina 社の NovaSeq6000 を用いて 150base 長ペアエンドシーケンスを行い、シーケンサー付属ソフト (NovaSeq Control Software v1.6.、Real Time Analysis (RTA) v3.4.、bcl2fastq2 v2.20) を用いて塩基配列 (リード配列) を得た。タグ配列に基づき塩基配列 (リード配列) を検体ごとに分類して fastq ファイルを取得した。

血液からのマルチオミックスデータ収集

i. 全ゲノム解析 (depth 30)

肺組織を取得していない症例においては、PBMC (末梢血単核細胞) を用いた全ゲノム解析を実施した。プロトコルは、「肺組織からのマルチオミックスデータ収集 1. 全ゲノム解析」と同一である。

ii. トランスクリプトーム (miRNA-seq)

miRNA トランスクリプトーム解析には血漿を用いた。RNA 抽出は QIAGEN miRNeasy micro kit を用いた。各サンプルに QIAzol 700µl を添加して RNA 抽出後、14µl の nuclease free water で溶出した。ライブラリ調製キットは QIAseq miRNA Library Kit を使用し、input 量は 5µl とした。こうして調製したライブラリの quality check はバイオアナライザーと qPCR (2nM サンプルを使用) にておこなった。miRNA-seq は Nextseq 75bp シングルエンドで行い、1 サンプルあたり 2000 万リード以上取得することとした。RNA 抽出から miRNA-seq までの一連の作業は株式会社 DNA チップ研究所に委託した。

iii. プロテオーム (DIA)

プロテオーム解析には血清を用いた。エクソソーム抽出条件検討の結果、大阪大学コホートの血清中エクソソームプロテオーム解析と同じく MagCapture isolation kit (Fujifilm Wako) を用いることに決定した。

3. 統合データ解析の手法開発とそれを用いた新規創薬ターゲット候補の同定：本事業において理研 AIP との共同研究により開発された患者層別化 AI を用いて、これまでに収集した大阪大学コホートの臨床情報（602 件の血清中エクソソーム プロテオームデータ及びそれと紐づく診療情報）のデータ解析を実施した。診療情報の欠損値は、健常者の代表値を用いて補填してから解析に用いた。

患者層別化 AI を用いた臨床情報の解析

患者層別化 AI を用いた解析では、紐付けるデータ 2 種類を行列の形で入力する必要がある。そこで上記臨床情報の解析時には①胸部 CT 画像の読影所見（one-hot ベクトル化による構造化済みデータ）とプロテオームデータ、②基本患者情報などからなる診療情報とプロテオームデータ、③血液検査値とプロテオームデータ、の 3 通りの組み合わせで入力し、それぞれにおいて出力された結果の中から特に IPF の特徴（①「網状影」「牽引性気管支拡張」「蜂巣肺」、②「SpO2」「労作時呼吸器困難」など、③「KL-6」「肺サーファクタントプロテイン D」「アンギオ変換酵素」など）との連関が示唆されるタンパク質を探索した。

バイオインフォマティクス解析による出力結果解釈

上記解析によって見出されたタンパク質に関して、TargetMine や IPA (Ingenuity Pathway Analysis, QIAGEN 社) を用いた既存知識との照合による分析を行い、関与が疑われるパスウェイの推定や肺線維化に寄与する因子の絞り込みを行なった。

IPF 患者肺組織・病変部位におけるタンパク質の免疫染色

大阪大学呼吸器内科より入手したホルマリン固定済みの肺組織から FFPE (Formalin fixed paraffin embedded ; ホルマリン固定パラフィン包埋) ブロックを作製し、滑走式マイクロームを用いて、厚さ 4 μ m に薄切して未染色標本を作製した。未染色標本は HE (ヘマトキシリンエオジン) 染色、MT (マッソントリクローム) 染色を実施し、炎症および線維化を確認した。さらに、未染色標本を用いて、解析によって見出されたタンパク質の特異抗体を用いる IHC (Immunohistochemistry; 免疫組織化学) 法による染色を実施した。検出試薬は OptiView DAB ユニバーサルキットを使用し、また、抗体ごとに陰性コントロールとして抗体希釈液を用いた染色を同時に実施し、各抗体に対する IHC 染色の評価を行った。なお、染色標本は明視野スライドスキャナーを用いて、対物レンズ 20 倍の設定で撮影し、画像ファイルを SVS 形式で保存した。この標本作製、染色作業はジェネティックラボ株式会社（北海道）に委託して実施した。

EMT 試験系の樹立

線維化の発症と進行には肺上皮細胞が上皮間葉転換 (Epithelial-mesenchymal transition ; EMT) により筋線維芽細胞に分化することが重要であることが広く示唆されている。そのため、解析により見出した分子の EMT への関与を評価する目的で、これらの分子の発現が報告されており、また TGF- β 添加により EMT が誘導されると報告されているヒト気道上皮細胞 BEAS-2B を用いて EMT 評価試験系の樹立に着手した。

EMT は、上皮性マーカーである E-cadherin の発現抑制と、間葉系マーカーである Fibronectin、 α -SMA 及び Snail の発現増強を指標として評価することとした。ATCC より購入した BEAS-2B 細胞を $1 \sim 7.5 \times 10^4$ cells/mL (培地 ; Bronchial epithelial cell basal medium、Lonza 株式会社) で 96 well プレートへ播種し、24 時間後に TGF- β (Proteintech) を終濃度 2.5、5、10、15 及び 20 ng/mL となるように添加した。TGF- β 添加 48 時間後に、SuperPrep® II Cell Lysis & RT Kit for qPCR (東洋紡株式会社) を用いて細胞溶解および RT を行い、上述の EMT マーカー遺伝子群の発現量を qPCR により測定した。qPCR には、THUNDERBIRD Probe qPCR Mix (東洋紡株式会社) を用い、各マーカープローブには TaqMan Probe (サーモフィッシュサイエンティフィック株式会社) を用いた。本試験系の樹立は株式会社ケーエーシー（京都）に委託して実施した。

血清中代謝物の網羅的測定 (メタボローム)

大阪大学より入手した血清のうち、IPF 確定診断 10 例および健常者 10 例について、キャピラリー電気泳動・フーリエ変換型質量分析計 (CE-FTMS) を用いて、糖リン酸、有機酸、アミノ酸、アミン等のイオン性代謝物質を網羅的に測定した。また液体クロマトグラフィー・飛行時間型質量分析計 (LC-TOFMS) を用いて遊離脂肪酸、アシルカルニチン、胆汁酸、ポリフェノール類等の中性代謝物を網羅的に測定した。なお、分析対象物質の物性に応じて、ポジティブモードとネガティブモードの 2 つのメソッドにて実施した。測定データは、検出ピークを標準データベース (約 1300 物質) と照合することにより帰属するとともに、糖代謝、アミノ酸代謝に関わる 110 物質に関しては濃度定量を行った。本測定は、ヒューマン・メタボローム・テクノロジーズ株式会社 (山形) に委託して実施した。

4. オープンプラットフォームの構築：

本事業で作成される IPF/肺がんの疾患統合データベース、機能分子を特定するための AI 及び知識ベース等を多くの研究者等に利用してもらうための環境（オープンプラットフォーム、以後 OPF）を構築するため、OPF システム基盤プロトタイプ of 構築・診療情報の個人特定リスク評価及び二次匿名加工条件検討・想定ユーザーへのヒアリングや先行事例調査を行なった。

OPF システム基盤プロトタイプ of 設計と構築

特定非営利活動法人 システム・バイオロジー研究機構（SBI）との連携により、OPF システム基盤のプロトタイプを構築した。SBI が開発したプラットフォームである Garuda (<http://www.garuda-alliance.org/>) で用いられている技術（SBI が特許保有）により、OS やプログラム言語に依存せずデータベースや AI を相互連結させることができるように設計・構築した。OPF へのアクセスはアカウント発行により制御し、OPF 上で閲覧・解析可能なデータはカテゴリーごと（「Medical Record」「miRNA」「mRNA」「Genome」「Epigenome」「Protein」）、プロジェクトごと（例えばコホートによる分類）に選択できるように設計し、AI は「ガジェット」として逐次追加可能な仕様とした。また、本システム基盤上にガジェット化した患者層別化 AI を搭載し、簡便な操作により実行可能となっていることを確認した。

診療情報の個人特定リスク評価及び二次匿名加工条件検討

OPF の構築及び運用にあたり、患者の診療情報など要配慮個人情報を取り扱う上でのリスク評価並びに匿名加工の必要性の評価、また匿名加工作業の実施が求められる。そのため、当該年度は IPF の診療情報の各項目について匿名加工作業の観点からの類型化を行い、匿名加工を必要とする項目の絞り込みを行なった。さらに、絞り込まれた項目に対して個人特定リスクの評価を行なった。この結果をもとに、匿名加工作業を実施する必要があると判断された各項目について、その個人特定リスクに応じた具体的な匿名加工手法を検討・設定した。

想定ユーザーへのヒアリングや先行事例調査

本事業の成果公開にあたり、既に OPF やそれに類似したシステムを運用している団体・企業の外部公開化の手法を把握・参考にする（先行事例調査）とともに、本事業 OPF の想定ユーザーとなる産業界などからのニーズを把握（想定ユーザーへのヒアリング）しておく必要がある。そのため、OPF の適切な運営体制の策定、ユーザーニーズの高いシステムの構築、持続運営可能なモデルの実現などに向けて情報収集を行なった。先行事例調査では、外部公開しているプラットフォームのシステム構築に関する全体像や構築方法、構築時の課題、法規制への対応状況、収益構造等を調査対象とした。想定ユーザーへのヒアリングでは、製薬・食品・ヘルスケア・診断分野・メディカル IT 分野の企業などを対象とし、各企業で求められているデータベースの内容、データベース利用に際して拠出可能な費用、データベースを利活用する上での社内・団体内での情報管理体制、データベースを利活用する上での課題点・問題点等を調査対象とした。さらに、医療用医薬品市場の市場規模および患者動向に関する情報に加えて、OPF 構築に必要な医療 IT 技術（インシリコ創薬・創薬支援システム、医療ビッグデータ分析システム、医療ビッグデータ向けクレンジングサービスなど）の市場規模についても併せて調査を行なった。

（倫理面への配慮）

本研究は、医薬基盤・健康・栄養研究所ならびに分担研究機関において倫理審査、承認を得た後、ヒトゲノム・遺伝子解析研究に関する倫理指針及び、人を対象とする医学系研究に関する倫理指針に従って遂行した。

C. 研究結果

1. 既知（背景）情報の収集と知識ベース化：

計画通りにデータベースの統合を完了した。

論文から分子・病態に関する知識抽出を行う AI 開発のために必要な言語リソースの構築において、産総研人工知能研究センターと共同でアノテーションガイドラインの作成を行い、肺がん・臨床系論文 167 件のアノテーションを完了した。当該年度までに構築した言語リソースは研究分担者（高村大也）が担当する論文からの知識自動抽出技術開発において学習データとして使用された。言語リソース構築・知識自動抽出技術に関する研究成果全般の詳細については、研究分担者（高村大也）による報告書を参照されたい。

2. 臨床情報の収集と機械可読表現の開発：

肺がん手術検体及びバイオプシー検体のオミックス解析の研究結果全般の詳細については、研究分担者（浜本隆二）による報告書を参照されたい。

大阪大学医学部呼吸器・免疫内科のバイオバンクおよび大阪大学医学部附属病院医療情報部より、IPFを含む間質性肺炎および器質的な呼吸器疾患がないと診断された健常者について、令和2年度はCT画像156名およびこれまでに入手した情報の中の欠損分を入手した。なお、呼吸機能データについては、匿名化処理の手法構築のみを行った。

これまでに入手済みの血清および採血日にひもついた診療情報を構造化して、患者層別化AIでの解析を試行した（本研究結果は次項「3. 統合データ解析の手法開発とそれを用いた新規創薬ターゲット候補の同定」を参照）。

神奈川循環器呼吸器病センターより、下記の臨床情報を取得した（2021年3月31日現在）。

- ・ 研究同意取得 令和2年度：376名（同意撤回4例）、累計：874名（同意撤回5例）
- ・ 血液 令和2年度：429例（累計：855例）
- ・ 肺組織（クライオバイオプシー・外科的肺生検・外科手術検体）令和2年度：111例
- ・ 臨床情報 累計：715例

肺組織126症例（176検体）を対象に下記のマルチオミックスデータを取得した。

- ・ 全ゲノム解析（depth 30）
- ・ DNAメチル化解析
- ・ トランスクリプトーム（RNA-seq）

PBMC197症例を対象に下記のマルチオミックスデータを取得した。

- ・ 全ゲノム解析（depth 30）

血漿301症例を対象に下記のマルチオミックスデータを取得した。

- ・ トランスクリプトーム（miRNA-seq）

また、COVID-19がSARS-CoV-2感染による間質性肺炎であると考えられる所見を得たことから、COVID-19患者のPBMC30症例を対象に下記のマルチオミックスデータを取得した。

- ・ 全ゲノム解析（depth 30）

神奈川循環器呼吸器病センターの臨床情報収集全般の詳細については、研究分担者（小倉高志）による報告書を参照されたい。

3. 統合データ解析の手法開発とそれを用いた新規創薬ターゲット候補の同定：

大阪大学コホートの臨床情報を患者層別化AIの入力データとして用いた解析の結果、患者基本情報を含む診療情報・胸部CT画像の読影所見・血液検査でIPF患者に認められる特徴と紐づけられるタンパク質を複数個見出すことに成功した。これらのタンパク質は、IPFや線維化に関する既知情報と照合して矛盾がなく、かつこれまでIPFとの関与が報告されていないパスウェイがIPF発症に寄与している可能性を示唆する結果が得られた。また、これらのタンパク質は血清中エクソソームだけでなくIPF患者の肺組織・病変部位においても量的変動が認められた。

4. オープンプラットフォームの構築：

オープンプラットフォーム（OPF）システムの設計と構築

OPFシステム基盤プロトタイプを医薬基盤・健康・栄養研究所サーバー上に構築し、搭載された患者層別化AIガジェットの動作確認を完了した。すなわち、テストデータを患者層別化AIガジェットで読み込み、基盤上でプログラムを実行して出力を画面上に表示させることができることを確認した。また、格納する臨床情報へのアクセス画面を設計・確認した。

所内で別途行っているデータ連携基盤とシステム等を共通とすべきかどうかを検討したが、コンピュータシステムの統合は行わず、運用体制だけを共通の方式で行うこととした。上記のごとくプロトタイプでシステムの動作が確認できたので、プロトタイプと同様に医薬基盤・健康・栄養研究所サーバー上にon premiseで本番環境を構築した。

また、事業終了後も OPF に AI やデータベースを追加することができるように、これらのガジェット化を行うための環境も用意した。

診療情報の個人特定リスク評価及び二次匿名加工条件検討

IPF の診療情報（大阪大学コホート、神奈川県立循環器呼吸器病センターコホート）の全項目について類型化、匿名加工の必要性の有無に関する評価、個人特定リスク評価、匿名加工が必要であると判断された項目に対する匿名加工条件設定を完了した。ただし、オミックスデータ、画像、医療テキストについては当該年度は対象外であり、次年度の実施を予定している。

想定ユーザーへのヒアリングや先行事例調査

先行事例調査については国内外の組織または機関 6 件を対象に調査を完了した。また、想定ユーザーへのヒアリングについては 25 社を対象に調査を完了した。本調査結果をもとに、情報セキュリティと利便性のバランスやエコシステム構築の重要性を熟慮しながら OPF 国内運用の基本構想策定を開始した。

D. 考察

当該年度の成果は、分担研究者との共同研究に関連するものを除くと大きく分けて 3 つである。すなわち、①大阪大学コホートの臨床情報解析による IPF 創薬標的候補の提示、②神奈川県立循環器呼吸器病センターコホートの臨床情報収集継続による疾患 DB 拡充、③OPF プロトタイプ構築と運用準備開始である。

①では、これまでに収集した臨床情報、開発した患者層別化 AI、結果解釈を支援するツールの拡充の全てを結集し、IPF の特徴（診療情報の項目）と連関のあるタンパク質及びその結果から導き出される IPF 創薬標的候補を見出すことに成功した。この結果は、IPF のように分子レベルで疾患発症・重症化メカニズムの理解が進んでいない難病であっても臨床情報を収集し適切な AI による解析を行うことで創薬標的探索が可能であることを示唆する重要な成果である。本解析によって見出されたタンパク質・IPF 創薬標的候補について、IPF との関連や創薬標的としての有望性をより詳細に検討するほか、その結果を AI 開発にフィードバックすることで更なる精度向上を目指す。

また、②では世界でも稀な IPF マルチオミックスデータベースの構築に向けて着実に症例数を追加しており、本データから創薬標的探索や発症メカニズム推定を行う AI 開発を並行して遂行していく。

③ではシステム基盤のプロトタイプ構築が完了したことから、今後は本基盤に搭載するデータや AI の加工（データの匿名加工や AI のガジェット化）を本格化し、令和 3 年度末国内運用開始を目指す。

E. 結論

コロナ禍により IPF 肺組織収集症例数を下方修正したものの、当該年度研究計画の達成状況は概ね良好である。AI による臨床情報の解析により、創薬標的候補を提示するに至った。本事業の成果であるデータベース、AI を広く共有するためのプラットフォームのプロトタイプが完成し、国内運用開始に向けた情報収集も予定通り完了した。

G. 研究発表

1. 論文発表

- 1) Esaki T., Horinouchi T., Natsume-Kitatani Y., Nojima Y., Sakane I., Matsui H., Estimation of relationships between gene expression and chemical substructures: Adapting canonical correlation analysis for small sample data by gathered features using consensus clustering, *Chem-Bio Informatics Journal* 20:58-61, 2020
- 2) Allendes Osorio R., Nyström-Persson J., Nojima Y., Kosugi Y., Mizuguchi K., Natsume-Kitatani Y., Panomicon: A web-based environment for interactive, visual analysis of multi-omics data, *Heliyon* 6(8): e04618, 2020
- 3) Nojima Y., Takeda Y., Maeda Y., Bamba T., Fukusaki E., Itoh M., Mizuguchi K., Kumanogoh A., Metabolomic analysis of fibrotic mice combined with public RNA-Seq human lung data reveal potential diagnostic biomarker candidates for lung fibrosis, *FEBS Open Bio* 11(10):2427-2436, 2020
- 4) Teranishi Y., Kuwahara H., Ueda M., Takemura T., Kusumoto M., Nakamura K., Sakai J., Kimura T., Furutani Y., Kawashima M., Imokawa G., Nogami-Itoh M., Sphingomyelin Deacylase, the Enzyme Involved in the Pathogenesis of Atopic Dermatitis, Is Identical to the β -Subunit of Acid Ceramidase, *International Journal of Molecular Sciences* 22(21):8789, 2020

- 5) Tripathi L., Itoh M., Takeda Y., Tsujino K., Kondo Y., Kumanogoh A., Mizuguchi K., Integrative Analysis Reveals Common and Unique Roles of Tetraspanins in Fibrosis and Emphysema, *Frontiers in Genetics* 11:585998, 2020

2. 学会発表

- 1) Koba T., Takeda Y., Narumi R., Nojima Y., Futami Y. N., Edahiro R., Itoh M. N., Shirai Y., Adachi J., Tomonaga T., Kumanogoh A., Targeted Proteomics of Serum Extracellular Vesicles to Identify Novel Biomarkers For COPD, ISEV2020, Web 開催, 2020.5.22
- 2) 伊藤眞里, 薬効薬理研究における動物実験の現状と課題, 第 67 回日本実験動物学会総会, 講演要旨集のみ発行, 2020.5.25
- 3) 夏目やよい, 全体概要・特発性肺線維症, 新薬創出を加速する人工知能の開発 令和元年度成果報告会, Web 開催, 2020.7.20
- 4) Koba T., Takeda Y., Narumi R., Nojima Y., Futami Y. N., Edahiro R., Itoh M. N., Shirai Y., Adachi J., Tomonaga T., Kumanogoh A., Targeted Proteomics of Serum Extracellular Vesicles To Identify Novel Biomarkers For COPD, ATS2020, Web 開催, 2020.8.5
- 5) 伊藤眞里, 機械学習を活用した創薬研究のテーマ創出, 技術情報協会セミナー, Web 開催, 2020.8.25
- 6) 夏目やよい, 臨床情報からのデータ駆動的な患者層別化に向けて, 第 9 回生命医薬情報学連合大会 IIBMP2020, Web 開催, 2020.9.1
- 7) 滝本宜之, 武田吉人, 木庭太郎, 玄山宗到, 広瀬雅樹, 新井徹, 伊藤眞里, 朝長毅, 熊ノ郷淳, 井上義一, エクソソームのプロテオミクスによるリンパ脈管筋腫症の新規血液バイオマーカーの同定, 第 60 回日本呼吸器学会学術講演会, Web 開催, 2020.9.20
- 8) 白井雄也, 武田吉人, 中山真菜, 木庭太郎, 二見悠, 足立淳, 朝長毅, 伊藤眞里, 水口賢司, 熊ノ郷淳, 血清エクソソームの最新定量プロテオミクスを用いた線維化マーカーの探索, 第 60 回日本呼吸器学会学術講演会, Web 開催, 2020.9.20
- 9) 安部祐子, 武田吉人, 吉村華子, 木庭太郎, 管泰彦, 福島清春, 白山敬之, 岩堀幸太, 平田陽彦, 足立淳, 伊藤眞里, 熊ノ郷淳, エクソソームの最新プロテオミクスによる悪性胸膜中皮腫の新規バイオマーカー探索, 第 7 回日本細胞外小胞学会, Web 開催, 2020.10.26
- 10) 木庭太郎, 武田吉人, 鳴海良平, 野島陽水, 伊藤眞里, 中山真菜, 吉村華子, 白水崇, 足立淳, 朝長毅, 熊ノ郷淳, 血清エクソソームのターゲットプロテオミクスによる新規 COPD バイオマーカー探索, 第 7 回日本細胞外小胞学会, Web 開催, 2020.10.26
- 11) 伊藤眞里, AI を用いる創薬研究テーマ探索:臨床情報収集と知識処理の自動化への挑戦, CBI 学会 2020 年大会, Web 開催, 2020.10.28
- 12) 夏目やよい, 官民研究開発投資拡大プログラム PRISM 創薬 AI 社会実装への展望, IQVIA ジャパンメディアセミナー, Web 開催, 2020.12.8

F. 知的財産権の出願・登録状況 (予定を含む。)

1. 特許取得

複数の項目を関係付けるための方法、システム、およびプログラム、夏目やよい、上田修功、2021.1.29、特願 2021-13264、国立研究開発法人医薬基盤・健康・栄養研究所、理化学研究所

2. 実用新案登録

なし

3. その他

なし