別紙 3

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

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： 水口賢司（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・センター長）

要旨の作成
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。
2019 年度は、創薬標的探索支援データウェアハウス TargetMine に 7 種類の市販/公共データベースおよび知識ベースを統合・拡充した。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（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。miRNA-seq のプロトコルの妥当性評価を行い、適切にデータ収集が可能であることを確認した。昨年度にプロトタイプを開発した患者層別化 AI の改良により、実際に本事業にて収集を進めている臨床情報からデータ駆動的に患者層別化が可能であることを確認した。

参画研究者： 水口賢司（医薬基盤・健康・栄養研究所）、熊ノ郷淳（大阪大学大学院医学系研究科）、小倉高志（神奈川県立循環器呼吸器病センター）、高村大也（産業技術総合研究所）、西村紳一郎（北海道大学大学院先端生命科学研究）、浜本隆二（国立がん研究センター研究所）、西岡安彦（徳島大学大学院医歯薬学研究）、奥野恭史（京都大学医学研究科）

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（特発性肺繊維症）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。

- WHO-clinical Trials（臨床データベース）
- HGMD（ヒト遺伝子変異データベース）
- IPF キュレーションベース（定型化された医療情報データベース）
- IPF 情報ベース（論文から抽出した知識ベース）
- PharmaProject（医薬品データベース、臨床試験情報データベース）
- BioExpress（遺伝子発現情報データベース）
- CTOD（治験情報データベース）

本業務は三菱スペースソフトウェア株式会社への外部委託により実施した。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、基礎生物科学系文献のガイドライン・アノテーション作成および自動キュレーション、および及び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を行った。対象となる文献（IPF・基礎分子系：121 件、肺がん・基礎分子系 100 件、肺がん・臨床系：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。
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を、また大阪大学医学部附属病院医療情報部より、上記血清にひもつく診療情報（採血日に最も近い日程における診察記録、CT 画像およびその読影所見、血液検査値、呼吸機能検査値およびフローボリュームカーブ、患者基本情報および初診時問診票）を収集した。また、検査の結果、器質的な呼吸器疾患を有さ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、研究用の匿名化 ID が付与された形で入手した。
診療情報のうち、診察記録は、必要な情報 120 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、部位と病変のペアからなる構造化データを得た。血液検査値は、276 項目の重要項目を選択して構造化した。初診時問診票、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。呼吸機能は、匿名化処理の手法構築を行った。

<u>プロテオーム解析</u>
血清 200μl 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。

診療情報

- 診療記録（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

1. 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)シークエンス解析

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 vl.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。

2. 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 Normalizaion を実施して解析を行なった。

3.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)シークエンス解析

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 vl.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。

<u>miRNA-seq のフィージビリティースタディー（FS）</u>

神奈川循環器呼吸器疾患センターにおいて収集された血漿 24 症例から 250µl ずつプールし、FS 用サンプルを調製した。これを容量（100 または 200µl）、凍結融解の回数（1,2,3 回）、スパイク（QIAseq miRNA Library QC Spike-Ins：RNA 抽出前に添加）の有無について検討した。RNA 抽出は QIAGEN miRNeasy micro kit を用いた。QIAseq miRNA

Library QC Spike-Ins（nuclease free water 500µl で懸濁したもの）と QIAzol を混合し mixture を作成したのちに、各サンプルに 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。

収集された臨床情報は、当所で構築を進めている疾患統合データベースに順次格納する。当該データベースは 2018 年度に MongoDB（オープンソースソフトウェアのドキュメント指向データベース）を用いて構築した（三井情報株式会社への委託)。2019 年度は、当該データベースに格納されたデータの統計量（データの特徴を要約した数値）を表示する機能を実装した（ライフマティックス株式会社への委託)。

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：昨年度にプロトタイプを開発した患者層別化 AI の改良を実施した。本事業で収集を進めている臨床情報は、ある項目が当てはまるか当てはまらないかを 0（当てはまらない）または 1（当てはまる）で表現した数字が記載されているものと実測値（例：SpO2）が記載されているものが混在する。2020 年度は、これを入力データとして受け入れることが可能となるように改良した。プログラムは python で実装した。大阪大学病院にて収集された IPF を含む間質性肺炎患者 363 症例の診療情報（日常診療の電子カルテ:110 項目、血液検査:276 項目、CT 画像読影所見:3342 項目。欠損値は下記の方法で補填した。①項目が当てはまるか否かを 1 または 0 で表現している場合、欠損値には 0 を補填、②実測値が記載される場合、欠損値には k-nearest neighbors で補填）及びそれを紐づけられた血清エクソソーム プロテオームデータ（1941 種類のタンパク質）を用い、改良した患者層別化 AI の動作確認をおこなった。

（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：計画通りに公共・市販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肺がん・基礎分子系要旨データ 100 件のエンティティ（disorder、cell、pharmacological substance）アノテーションはほぼ終了し、一貫性の確認・修正作業を実施している。IPF・基礎分子系要旨データに対して、生命現象イベント（gene expression、conversion、pathway、regulation）に関するアノテーション用ガイドラインを作成し、IPF・基礎分子系要旨データ 121 件に対してエンティティ（GGPs、

analytical entity、subject、cell、cell component、有機化合物、無機化合物）のアノテーションとノーマリゼーション ID における一貫性の確認を行った。イベントについても、gene expression など比較的シンプルなものに対してアノテーションを実施した。肺がん・臨床系論文 167 件のアノテーションについては、アノテーションガイドライン作成後に再開する計画へと変更し、現在中断している。言語リソース構築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、下記の臨床情報を取得した。

- 血清　602 症例
- 診療記録＋基本情報　594 症例
- 血液検査値　594 症例
- 読影所見　538 症例
- 初診時問診票　141 症例
- 画像　323 症例

血清は、エクソソームを分離したのち、エクソソーム内プロテオームデータを取得し、診療情報は、方法に記載した手法を用いてそれぞれ構造化した。初診時問診票、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た。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。

これらの臨床情報を用い、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）。大阪大学バイオバンクの臨床情報収集全般の詳細については、研究分担者（熊ノ郷淳）による報告書を参照されたい。

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（2020 年 2 月 20 日現在）。

- 研究同意取得　452 名
- 血液　370 名
- 肺組織（クライオバイオプシー・外科手術検体）　82 検体
- 臨床診断　310 名

肺組織 56 症例（80 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）

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。

<u>肺組織からの核酸（DNA・RNA）同時抽出</u>

神奈川循環器呼吸器病センターにおいて収集された肺組織 56 症例（80 検体）について DNA・RNA 抽出を行ない、シークエンスに必要量を確保することができるか否かを検討した。

<u>miRNA-seq のフィージビリティースタディー（FS）</u>

患者 24 人分の血漿をプールして調製したサンプル（容量 2 点（100μL, 200μL）、凍結融解 3 回、スパイクの有無について各 3 サンプル用意し、合計 36 サンプル）について

B.研究方法に記載した方法で miRNA-seq を実施した。その結果、シークエンスの Phred score（リード中の各ポジションにおける塩基のシークエンス QC スコア）が 75bp の全領域において 30（シークエンシングが良好に行われたことを示す基準値）を上回ることを確認した。2020 年 6 月現在、miRBase（miRNA のデータベース）に登録されているヒト miRNA の数は 2500 強であり、本研究では全てのサンプルにおいて 2000 以上の miRNA が検出された。総リードに含まれる miRNA の割合は、容量 100µl のサンプルでは 15-20%、容量 200µl のサンプルでは 20-30%であった。凍結融解の結果からは、１回目と２回目の間で多くの miRNA が分解されていることがわかった。更に、サンプルの凍結融解による miRNA 発現量への影響をより詳細に調査するため、凍結融解による分解を受けることがすでに報告されている miR-1（Glinge, C. et al (2017). Stability of circulating blood-based microRNAs - pre-analytic methodological considerations. PloS one, 12(2), e0167969.）の発現量を各サンプル間で比較した。その結果、容量 100µl・200µl 両方において miR-1 の発現量が減少する傾向が認められた。特に、容量 200µl の場合においては、凍結融解 1-2 回では各実験条件における miR-1 発現量の平均値が 150 前後であったのに対して凍結融解 3 回では平均値が 100 前後であり、１回目と３回目の間では p-value < 0.05（Welch's t test）で有意差があった。続いて、凍結融解の回数間で増減した miRNA の数を volcano plot により確認した。その結果、スパイクなしのサンプルでは、凍結融解数の違いにより有意に発現量が異なる miRNA が一定量検出されたのに対して、スパイクありでは発現量に有意差のある miRNA の数が減少し、凍結融解による影響が認められにくい傾向があった。

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：大阪大学医学部付属病院にて収集された IPF を含む間質性肺炎患者 363 症例の診療情報及びそれを紐づけられた血清エクソソームプロテオームデータを改良した患者層別化 AI に供したところ、「診療情報項目と紐づけられたタンパク質群」を出力し設計通りの動作をすることを確認した。

D. 考察

1. 既知（背景）情報の収集と知識ベース化：産総研人工知能研究センターより、「現時点のアノテーション・ツール（brat）では、一つのエンティティに複数の ID を付与できない。しかしながら、実際のアノテーションでは、一つのエンティティに複数の ID を付与しないといけないケースが多々あるので、どのように解決するか、アノテーター、システム開発の担当者とも議論する必要がある。」、「生命現象イベントは非常に複雑なので、文献データをどのようにテキスト・アノテーションしていくかが課題となる。複雑なアノテーションは、アノテーターによるアノテーションのコストも高くなり、時間を費やすだけで、ネットワークなどの関係抽出をする際にも精度が落ちる可能性があるので、アノテーションしやすいスキームを設計する必要がある。そのためにも、肺疾患の専門家、テキストマイニングの専門家の意見を取り入れることが望まれる。」と現状における課題が提示されており、当所肺疾患の専門家の意見を反映させた言語リソース構築を継続する必要がある。

2. 臨床情報の収集と機械可読表現の開発
大阪大学病院バイオバンクで収集した臨床情報は計画内容をほぼ完了し、2020 年度は患者層別化 AI による本解析を主軸としてデータ解析による知識抽出に注力する段階に到達していると言える。神奈川循環器呼吸器病センターにおける臨床情報収集では、特に間質性肺炎の中でもどのような症例を重点的に収集するべきかを随時収集状況に関する情報共有をしながら柔軟に対応する必要がある。

miRNA-seq のフィージビリティースタディー（FS）
miRNA はいずれのサンプルにおいても 2000 種類を超える miRNA を検出することができた。総リード数に対する miRNA の割合は容量100µlよりも200µlの方が高く平均26%であった。また、スパイクの有無では差がなかった。凍結融解を繰り返すと検出される miRNA の割合や特定の miRNA の発現量が減少することを確認した。また、スパイクの有無によりグローバルな miRNA 発現プロファイルに影響が出る可能性が示唆された。
スパイクの有無によって凍結融解の回数による影響の見え方が異なる点については、スパイクを入れることによって相対的に miRNA が占める割合が低下し、各サンプル間での差異が結果的に小さくなることによると考えられた。

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
大阪大学バイオバンクの全 602 症例データを用いた本解析により得られる出力については、本事業で拡充した TargetMine を用いたより多様な情報源を基にした情報検索・分析により結果解釈を実施する他、研究分担者（熊ノ郷淳）のグループに属する呼吸器内科の医師らによる結果解釈を取り入れることにより創薬標的探索に有益な情報を得られるものと期待される。

E. 結論
1. 既知（背景）情報の収集と知識ベース化：TargetMine の拡充が順調に進んだと判断し、2020 年度は公共データベース２点（PoSSuM、MGeND）の追加統合のみの実施とする。産総研人工知能研究センターとの委託研究について、2020 年度に計画内容を完了するべく、言語リソース構築についても完了・モデル構築において学習データとして利用可能な状態で産総研人工知能研究センターに提供することを目標とする。

2. 臨床情報の収集と機械可読表現の開発
国立がん研究センターにおける肺がんの手術検体及びバイオプシー検体を用いたマルチオミックス解析は、全エクソーム解析やトランスクリプトーム解析（RNA-seq）において世界最大規模のデータベース構築を達成したことから、2020 年度は欠損データの取得を実施することとする。大阪大学病院バイオバンクにおける IPF を含む間質性肺炎の臨床情報収集は、2020 年度に全 602 症例分の血清中エクソソームのプロテオームデータ及びそれと紐づけられた診療情報の収集をもって完了する。2019 年度に神奈川循環器呼吸器病センターにて収集された肺組織を用いたマルチオミックスデータは、2020 年度にプロセシング・解析を実施するほか、引き続き肺組織・血液・診療情報の収集を継続する。

miRNA-seq のフィージビリティースタディ

ー（FS）

血漿中 miRNA の網羅的測定にあたり、血液サンプルの収集プロトコルや RNA 抽出から miRNA-seq に至るプロトコルは適切に設定されていると判断した。

本研究結果を基に、

・ 容量：200μl
・ 凍結融解は 1 回に留める
・ スパイクは使用しない

という方針でプロトコルを設定することとする。miRNA-seq に至る全行程のプロコトルの妥当性を確認することができたと判断し、2020 年度において血漿中 miRNA の測定を実施する計画を具体化させることとした。

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：2019 年度に実施した患者層別化 AI の改良により、本事業にて収集を進めている臨床情報からデータ駆動的に患者層別化ルールを出力できることを確認した。これにより、臨床情報の収集・構造化が完了次第、当該 AI を用いた本解析を実施することとした。

F. 健康危機情報

なし

G. 研究発表

1. 論文発表

1. 伊藤眞里, 長尾知生子, 藤原大, 水口賢司, AI 創薬に向けた医療ビッグデータベースの統合と解析, **月刊ファームステージ**, Vol.19, No11 2 月刊 21－24, 2020
2. 夏目やよい, 水口賢司、新薬創出を加速する AI の開発、**Precision Medicine** Vol.3 No.5, 2020 (印刷中)

2. 学会発表

1. 水口賢司, Artificial intelligence-based drug discovery: challenges and applications to target identification and pharmacokinetic modelling, **第４回トランスオミクスシンポジウム**, 徳島, 2019.11.14
2. 水口賢司, 疾患・創薬研究におけるデータベースと AI 活用, **日本臨床試験学会　第１１回学術集会総会**, 東京, 2020.2.14
3. 水口賢司, 創薬研究への AI 活用**, ILSI Japan 先端技術シンポジウム**, 東京, 2020.2.21
4. 夏目やよい, Artificial intelligence for patient stratification based on clinical information, **The 1st International Symposium on Human InformatiX**, 京都, 2020.2.28
5. T. Fujiwara, T. Ogawa, H. Nakagawa, H. Hirata, I. Nagatomo, M. Itoh, Y. Takeda, A. Kumanogo, K. Mizuguchi, Multimodal feed-forward neural network coupled with convoluteional neural network for severity diagnosis of idiopathic interstitiall pneumonia using medical information and computed tomogramphy images, **The 1st International Symposium on Human InformatiX**, 京都, 2020.2.28

H. 知的財産権の出願・登録状況

1. 特許取得

該当事項なし

2. 実用新案登録

該当事項なし

3. その他

特記事項なし