Title	臨床・ゲノムデータを含む多層データの臨床応用基盤構築と有用性の検証
Sub Title	Construction of platform for clinical application of cancer omics data
Author	北川, 雄光(Kitagawa, Yuko) 岡林, 剛史(Okabayashi, Koji) 川久保, 博文(Kawakubo, Hirofumi) 松田, 諭(Matsuda, Satoru) 北郷, 実(Kitago, Minoru) 阿部, 雄太(Abe, Yuta) 林田, 哲(Hayashida, Tetsu)
Publisher	
Publication year	2023
Jtitle	科学研究費補助金研究成果報告書 (2022.)
JaLC DOI	
Abstract	がんゲノム医療推進のため、クリニカルシークエンスが保険適応となったが、DNA情報に加え、RNA情報・エピジェネティック情報等を同一検体から収集して多層データを構築し、腫瘍の全体像を統合的に捉える次世代型クリニカルシークエンスが求められている。これら情報の理解は容易ではないが、医療者の持つ知識に異存せず、馴染みのある思考法に沿うような環境を作り出すため、臨床情報・ゲノム情報・遺伝子発現情報を視覚的に解析可能な統合解析プラットフォームの開発を施行した。自然言語処理を用いた診療録の構造化システムの開発を行い、がんゲノム医療に必須の臨床情報を効率的かつ永続的に付加する基盤の構築を行った。 Clinical sequencing is now covered by insurance to promote cancer precision medicine. In addition to DNA, RNA, and clinical information are collected from the same specimen to construct omics data, and next-generation clinical sequencing is required to obtain an integrated portrait of the entire tumor. Understanding this information is not easy, but in order to create an environment that is not alien to the knowledge possessed by medical professionals and that follows a familiar way of thinking, we have developed an integrated analysis platform that enables visual analysis of clinical information, genomic data, and gene expression data. We developed a system for structuring medical records using natural language processing, and constructed a platform for efficiently and permanently adding clinical information essential for cancer precision medicine.
Notes	研究種目 : 基盤研究 (B) (一般) 研究期間 : 2020～2022 課題番号 : 20H03747 研究分野 : 消化器癌・乳癌
Genre	Research Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KAKEN_20H03747seika

様　式　C－１９、F－１９－１、Z－１９　（共通）

科研費 KAKENHI

科学研究費助成事業　　研究成果報告書

令和　５年　６月　５日現在

機関番号：３２６１２
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０３７４７

研究課題名（和文）臨床・ゲノムデータを含む多層データの臨床応用基盤構築と有用性の検証

研究課題名（英文）Construction of platform for clinical application of cancer omics data

研究代表者
北川　雄光（KITAGAWA, Yuko）

慶應義塾大学・医学部（信濃町）・教授

研究者番号：２０２０４８７８

交付決定額（研究期間全体）：（直接経費）　14,400,000円

研究成果の概要（和文）：がんゲノム医療推進のため、クリニカルシークエンスが保険適応となったが、DNA情報に加え、RNA情報・エピジェネティック情報等を同一検体から収集して多層データを構築し、腫瘍の全体像を統合的に捉える次世代型クリニカルシークエンスが求められている。これら情報の理解は容易ではないが、医療者の持つ知識に異存せず、馴染みのある思考法に沿うような環境を作り出すため、臨床情報・ゲノム情報・遺伝子発現情報を視覚的に解析可能な統合解析プラットフォームの開発を施行した。自然言語処理を用いた診療録の構造化システムの開発を行い、がんゲノム医療に必須の臨床情報を効率的かつ永続的に付加する基盤の構築を行った。

研究成果の学術的意義や社会的意義
ゲノム情報の解析には通常プログラミング言語を使用し、大型のサーバーを用いたバイオインフォマティクス解析が行われる。そのため、臨床現場の医療従事者においては、解析技術を有するものは非常に少なく、その習得には膨大な時間を有するため、ゲノム情報を適切に処理し、理解することは不可能である。本研究で取り組んだように、臨床情報・ゲノム情報を含むオミックスデータを臨床医が視覚的・直感的に理解できる形式に変換することができれば、データが指し示す結果を参考に治療決定をしていくことが可能となる。これにより、本邦のがんゲノム医療を推進し、公に資することが可能になると考えられた。

研究成果の概要（英文）：Clinical sequencing is now covered by insurance to promote cancer precision medicine. In addition to DNA, RNA, and clinical information are collected from the same specimen to construct omics data, and next-generation clinical sequencing is required to obtain an integrated portrait of the entire tumor. Understanding this information is not easy, but in order to create an environment that is not alien to the knowledge possessed by medical professionals and that follows a familiar way of thinking, we have developed an integrated analysis platform that enables visual analysis of clinical information, genomic data, and gene expression data. We developed a system for structuring medical records using natural language processing, and constructed a platform for efficiently and permanently adding clinical information essential for cancer precision medicine.

研究分野：消化器癌・乳癌

キーワード：がんゲノム医療　バイオインフォマティクス　オミックスデータ　自然言語処理　医学用語の構造化

１．研究開始当初の背景

次世代シークエンサー（NGS）を中心とする技術を用いた網羅的なゲノム解析が行われることで、悪性腫瘍における特徴的なドライバー遺伝子と言われる多数の遺伝子変異が短期間に同定可能である。これにより個々の患者毎に有効と思われる治療を選択することが可能となり、遺伝子治療に基づく個別化医療（プレシジョンメディシン）を強力に推進している。本邦ではがんゲノム医療推進のため、慶應義塾大学病院を含む 11 施設を「がんゲノム医療中核拠点病院」と定め、癌領域におけるクリニカルシークエンスが保険適応となった。加えて今後は全エクソン領域を対象にした、より網羅的な検査への移行が予想される。
これまでに、我々の施設では手術時に採取された腫瘍組織検体において、DNA の構造変異のみならず、RNA の発現情報、CHIP-seq などの epigenetic な異常を同一検体から多層的データ（オミックスデータ）として収集・統合し、腫瘍の全体像を把握する次世代型クリニカルシークエンスの開発に取り組んでいる。その理由として、NGS を用いた検査法ではゲノム変異情報が一度にわかるという利点の反面、構造情報のみであることから治療選択上の問題点として、タンパク質の機能異常を引き起こすか否かの判定が困難な変異（VUS 変異）が存在すること。さらに変異情報に基づいた薬剤により治療を行う臨床試験の結果、同一薬剤に対して、患者ごとあるいは臓器ごとに異なる反応を引き起こすという結果が多く確認される。そのため、このような状況下において実施されているクリニカルシークエンスは、治療選択という観点で考えると、現時点では未完成な診療形態であると考えている。
そのため、我々は DNA 情報に加え、RNA 情報・エピジェネティック情報等を同一検体から収集して多層データ（オミックスデータ）を構築し、腫瘍の全体像を統合的に捉える次世代型クリニカルシークエンスの構築を開始している。この際にゲノム情報と合わせて、患者の転帰を含めた臨床情報を収集することは、より精度の高いゲノム医療の構築のために最も重要な役割を果たす。しかし通常は臨床情報とゲノムオミックス情報を統合解析するためには複雑なバイオインフォマティクスの知識とプログラミング技術が求められ、医療従事者の適切な理解の上で大きな障害となっている。
この多層的データにおいて最も重要な情報は、患者の臨床病理学的情報・予後・転機・経過情報であり、これら臨床データを持たない DNA・RNA 情報は意味のない文字情報であるとすらいえる。しかしながら、ゲノム情報と臨床情報を合わせて収集し、統合・解析される上では大きな問題がある。一つは、臨床情報の収集には定期的なアップデートが重要であり、観察期間に応じて予後情報が変化するために、そのデータの収集と要約には多くの人的資材と労力を要し、欠損値を多く含むデータになる傾向がある。もう一つは、ゲノム情報の解析には通常プログラミング言語を使用し、大型のサーバーを用いたバイオインフォマティクス解析が行われる。現在の臨床現場の医療従事者においては、解析技術を有するものは非常に少なく、その習得には膨大な時間を有するため、ゲノム情報を適切に処理し、理解することは不可能である。
これまでのがんの薬物療法に対する医学的エビデンスは、大規模臨床試験を行い、ある患者集団における統計学的な優劣を薬剤ごとに蓄積していくことが主流であった。しかし、がんゲノム医療においては、全ての遺伝子変異ごとに臓器別の大規模臨床試験を行うことは不可能であるため、「一例報告」の集合である臨床情報とゲノム情報が記載されたデータベースに治療法の根拠を委ねざるを得ない。ゲノム医療時代にはこのように新たなエビデン構築のシステムが必要とされている。人工知能を用いたアルゴリズム開発が多く提案されることは間違いないが、データ取得項目に漏れがある場合、優れたアルゴリズムを用いても正解を得られることはない。加えて古典的な統計学とバイオインフォマティクスは最新のアルゴリズム取得の前に医療者全員が簡便に行うべきであると考えられた。

２．研究の目的

これら背景から、本研究課題において解決すべき問題を、「臨床情報・ゲノム情報を含むオミックスデータを臨床医が視覚的・直感的に理解できる形式に提示可能か？」と設定した。これが可能となれば、カンファレンスやエキスパートミーティングの場において、それぞれの専門知識を有する臨床医が、蓄積された癌腫別・遺伝子変異別の薬剤感受性データと臨床データをその場でスムーズに解析することで、深い専門知識の交流を促し、データが指し示す結果を参考に治療決定をしていくことが可能となる。
臨床データは集める情報が膨大で、動的なデータであるために、「情報の収集と利用のシステム」そのものを根本的に変え、構築する必要がある。このように収集され常に最新に更新された臨床情報をゲノム情報と合わせて当たりまえのように参照しながら、全ての医療従事者やバイオインフォマティシャンがお互いの専門知識を超えてリテラシーを高めることで、本邦のがんゲノム医療を推進することにより、がんゲノム医療中核拠点病院として公に資することを目的とした。最も重要なのは情報を最新の状態で収集し、匿名化を経てデータベース化を行った後に、多くの医療従事者が閲覧可能な環境を整備することにあるが、サーバーを含めた資材、システム構

築のための技術者の確保はなされており、本研究へ円滑に統合可能である。
国立がん研究センターでは、がんゲノム情報管理センター (C-CAT)を立ち上げ、先進医療や今後保険収載される予定の遺伝子パネル検査情報を収載するプロジェクトを遂行中である。しかしながら日常臨床で常用するレベルまで運用されているデータベースは存在しない。さらにはそのデータベースを容易に即時に解析するようなエキスパートパネルをサポートするシステムは存在していない。C-CAT は本邦のがんゲノム医療を推進する上で、大変期待されるプロジェクトであるが、その性質上、全国の医療機関から集められたゲノム情報・臨床情報の収集がなされる予定である。蓄積された情報を全ての医療者がその場で解析し結果を参照できる環境は現在なく、本研究では慶應大学にて進行され、すでに開発されているシステムを統合することで、いちはやく新たなエビデンスレベルの構築を提案可能と考える。これらのシステムによって得られた成果物は基本的にゲノム医療の向上のためにデータベース化後、公開を予定しており、日本が今後がんゲノム医療のイニシアチブをとるためには、幅広い公共性をもって世界中で常用される、質の高いサービスを提供することとしている。

３．研究の方法
本研究では遺伝子変異データ・臨床病理学的データを適切に収集するシステムを用いて統合されたデータベースを即時にその場で容易に解析可能な環境を構築し、臨床医が有用な情報を理解可能な状態で提示することを目指す。対象とする癌腫は
・食道・胃・肝胆膵・大腸を含む消化器癌
・乳癌
としている。慶應義塾大学病院では、がんゲノム医療の均てん化のため、より簡便かつ迅速に実施できるクリニカルシーケンスのシステムを確立することを目的とし、悪性固形腫瘍の診断・治療を目的として採取された腫瘍組織に対して、同意が得られた全症例で遺伝子パネル検査（PleSSision Rapid）が実施されている。上記の癌腫は臨床情報がすでに蓄積されており、かつこれに対応する症例で､PleSSision Rapid 検査が研究立案時点ですでに 800 例以上に施行され、これらの癌種においては研究期間中 PleSSision Rapid を含むオミックスデータの蓄積が行われた。
研究期間中に臨床情報とゲノム情報を統合し解析可能なアプリケーションの開発を行う。このアプリケーションは臨床情報と DNA 変異情報に加え、遺伝子発現情報を含むオミックスデータを解析可能であり、さらには The Cancer Genome Atlas、SEER などの公共のデータベースも解析可能である。本ソフトはプログラミングやバイオインフォマティクスの技術がなくともマウスのクリック操作のみで直感的で容易に解析可能な仕様を目差し、研修医を含めた全ての医師が容易にクリニカルシークエンス情報と臨床情報をインプットし自ら解析することが可能な状況を確立する。
本研究の最初のステップとして、このアプリケーション開発を行い、慶應義塾大学病院において既に蓄積されているオミックスデータと臨床データを匿名化した形でセキュリティーが担保されているサーバー上に実装する。今後、新たに PleSSision 検査を施行された患者の治療決定を議論する擬似的なエキスパートパネル開催し、本システムを用いて、過去の患者のゲノム変異、臨床データを合わせて閲覧し、その場でオンタイムに解析を行い、治療選択の提案に関する議論を行う。さらに治療データが蓄積された場合にシステマティックなレビューを行い、クリニカルシークエンス以外のオミックスデータが有用な症例の算出を計画した。
また、研究途上において、臨床データをさらに充実させるため、自然言語処理を用いた診療録の構造化システムの開発と知識データベースの充実化の必要性が提案された。個別の疾患の全体像を把握するために、自由記載形式で形成された診療録の単語を検索する際は、一つの言葉における多様な表現が、検索の精度を低下させてしまう。共同研究を行っている IT 企業と、診療録における問題について検討を行い、これら自由記載における言葉の「ゆらぎ」、すなわち類義語を認識して一つの医学用語に集約・統合することで構造化するシステムの開発を行う。自由記載の診療録の記述をもとに、人工知能（AI）による「疾患の診断」・「検査項目追加の指示」・「重度の疾患や副作用の示唆」などに利用することが可能である。
これらの開発を行った後、
①システム導入前後での、全ての医療従事者のゲノム医療に対する理解度がどの程度向上するかの検証
②システム導入前後で、治療選択に変化を及ぼしうるかの検証
③システム導入により、患者の治療成績、生存期間に変化をもたらしうるかの検証
を目指していく。

４．研究成果
（1）　臨床情報・ゲノム情報・遺伝子発現情報を統合解析するアプリケーションの開発とその

運用

本研究では、医療者の持つバイオインフォマティクスの知識に異存せず、馴染みのある医療者の思考法に沿うような環境を作り出すため、臨床情報・ゲノム情報・遺伝子発現情報を視覚的に解析可能な統合解析プラットフォームの開発を施行した。これによりゲノム、臨床データを即座にマウス操作で解析が可能であり、患者一人一人の解析はもちろん、何千人という患者群の中から臨床病理学的因子に基づく分類、例えば「乳癌のエストロゲン受容体陽性」症例のみを抽出し、遺伝子変異の傾向や、選択した遺伝子発現状況のクラスター解析などをマウス操作のみで行うことが可能となった。また、これら臨床病理学的因子をもとに、通常の医学統計解析である生存曲線や単変量・多変量解析などをリアルタイムに行うことが可能である。このプラットフォームはセキュリティの完備された IBM クラウド上に常駐し、各種データをアップロードすることで解析を行うことができる。（図 1）テスト環境として、まずは The Cancer Genome Atlas および SEER データベースなどの公共データベース上の症例を使用して動作確認を行った。その後に約 100 例の乳がん患者のオミックスデータを使用して解析を行った。これらの症例は、前述の PleSSision Rapid によるゲノム情報を有する患者で、かつ遺伝子発現情報である Curegest95GC を施行した患者群であり、cell file と呼ばれれるマイクロアレイデータが取得されている。また、全症例は当院にて手術が施行された症例であるため、充実した臨床病理学的因子に関するデータベースが存在している。これらも前述の IBM クラウド上にあり、日常的な臨床業務で使用が可能となっている。

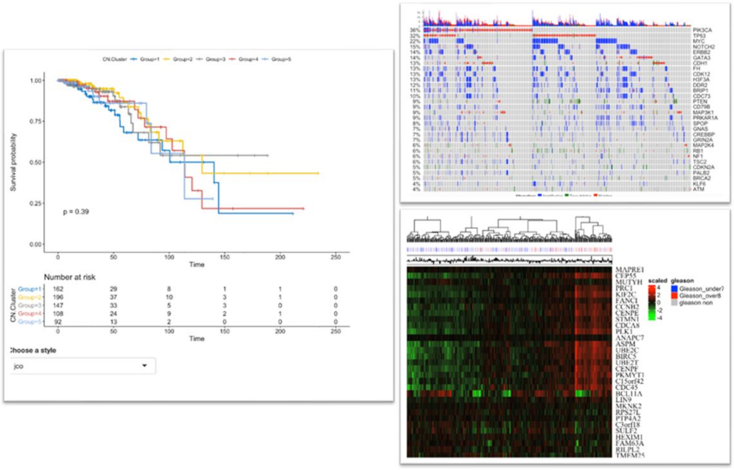

図 1：解析アプリケーション：医療統計解析・ゲノム解析・クラスター解析などを臨床情報とリンクさせた形式で直感的な操作で可能

（2） 自然言語処理を用いた診療録の構造化によるデータ取得システムの開発

研究の背景でも述べたとおり、臨床情報の収集においては、観察期間に応じて予後情報や薬物や手術などによる治療が変化するために、そのデータの収集と要約には多くの人的資材と労力を要し、ともすれば欠損値を多く含むデータになる傾向がある。これを解決するため、自然言語処理を用いた診療録の構造化システムの開発と知識データベースの充実化が本プロジェクトに重要であると考えた。すなわち、電子カルテなどの診療録に自由記載された文章から、コンピュータが医学・診療に関する言葉だけを抽出し、さまざまな臨床情報と結びつける「構造化」を行うことができる。これら自由記載における言葉の「ゆらぎ」、すなわち類義語を認識して一つの医学用語に集約・統合することで構造化するシステムを開発することで、文節ごとの医学用語を認識して構造化することが可能となる。本研究においてこのシステム構築を行ったが、開発当初はあくまで医学類義語辞書などを用いて構築を行ってきたため、実際に診療現場で用いられている用語・慣用語・略称・英文混じりの文章などに対応は不完全であると考えられる。そのため、本研究では自由記載による診療録の単語・文節を解析し、既存のシステムが未学習の未知語を抽出し、自由記載の「ゆらぎ」として学習させることで、自由記載構造化システムの精度上昇を図った。（図 2）これにより診療録のもつ新たな側面を IoT 技術によりハイライトし、さらなる付加価値を求めることが可能となり、構造化された臨床情報は、前述のゲノム情報と統合すること

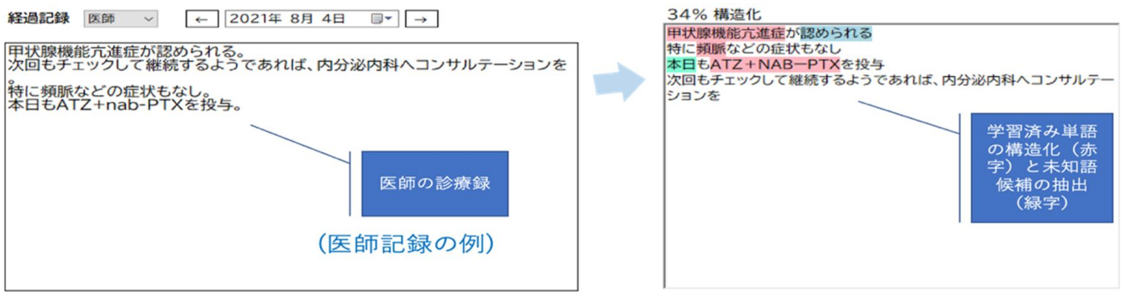

図 2：構造化の精度向上を目的とし、実際の診療録に用いられている未知語を学習させるシステム構築

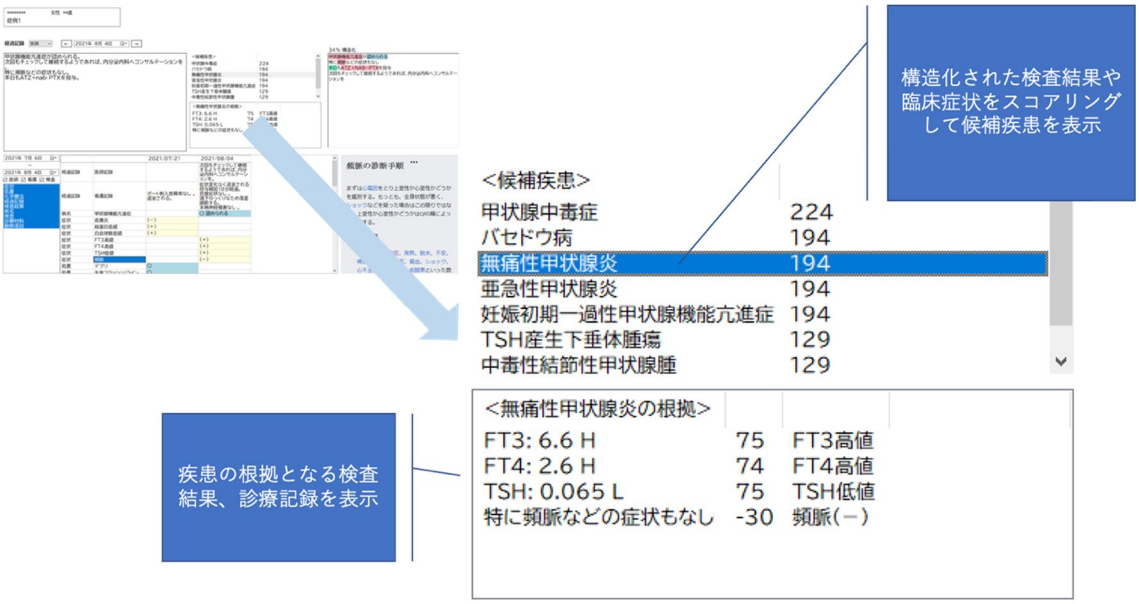

図 3：診療録の自然言語処理と疾患推定システムの開発：自由記載からなる診療録より医学用語のみの抽出を行い構造化、さらにその医学用語に紐付いた疾患の推定システムを知識データベースを作成し構築

が可能であった。この結果、層別化された群ごとの予後と遺伝子変異情報の比較などが実現可能となり、より一層の価値の向上が認められると考えられた。

さらに付加的な研究として、上記のように構造化された医学用語である臨床症状と疾患の紐付けを行う仕組みを開発し、癌の化学療法で生じる症状から副作用やその背景疾患などの推測を行うシステムを実現した。構造化された単語に対して医学的な意義を付加し、診断や医療安全に役立てることがすでに行われている。例えば、「腹痛」という症状を現す単語に、「胆石症」「胃炎」「腹膜炎」などの疾患を表す単語を紐付けることで、自由記載の診療録の記述をもとに、システムによる「疾患の診断」・「検査項目追加の指示」・「重度の疾患や副作用の示唆」などに利用することが可能である。すでに我々はがんに関連する疾患のガイドラインや教科書的な文献、および仮想症例を用いて、症状を示す単語に疾患を対応させることで本システムの基盤を整備してきた。(図 3)

欧米を中心に、カルテ記載などの自由記載に対して、自然言語処理による構造化を行って、診療情報の取得・整理を行う研究がすでに推進されており、その一部は実用化され、医学的エビデンスの構築に寄与している。本研究はこのようにすでに行われている検討について、日本語を対象として同様の研究を行うものであり、がんゲノム医療に必須の臨床情報を付加するだけではなくさらなる発展が見込まれると本研究から考察された。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 山口茂夫、林田哲、北川雄光	４．巻 28
２．論文標題 がん遺伝子変異情報を臨床現場で日常的に使うために　AIの役割	５．発行年 2021年
３．雑誌名 腫瘍内科	６．最初と最後の頁 436-440
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 林田哲、山口茂夫、北川雄光	４．巻 122
２．論文標題 【乳癌診療の現状と課題】がんゲノム医療の現状と課題	５．発行年 2021年
３．雑誌名 日本外科学会雑誌	６．最初と最後の頁 330-334
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計4件（うち招待講演　3件／うち国際学会　0件）

１．発表者名 林田　哲
２．発表標題 がんゲノム医療の仕組みと、乳がん診療における位置づけ
３．学会等名 日本乳癌学会学術総会（招待講演）
４．発表年 2021年

１．発表者名 林田　哲
２．発表標題 ゲノム医療の現状と課題
３．学会等名 第28回日本乳癌学会学術総会
４．発表年 2020年

1．発表者名 林田　哲
2．発表標題 乳癌診断・手術時にCGPを施行する有用性と今後の展開
3．学会等名 第30回日本乳癌学会学術集会（招待講演）
4．発表年 2022年

1．発表者名 林田　哲
2．発表標題 がん遺伝子変異情報を臨床の現場で日常的につかうために
3．学会等名 第18回日本臨床腫瘍学会学術集会（招待講演）
4．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	岡林　剛史 (OKABAYASHI Koji) (00338063)	慶應義塾大学・医学部（信濃町）・講師 (32612)	
研究分担者	川久保　博文 (KAWAKUBO Hirofumi) (20286496)	慶應義塾大学・医学部（信濃町）・准教授 (32612)	
研究分担者	松田　諭 (MATSUDA Satoru) (30594725)	慶應義塾大学・医学部（信濃町）・助教 (32612)	

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	北郷　実 (KITAGO Minoru) (70296599)	慶應義塾大学・医学部（信濃町）・准教授 (32612)	
研究分担者	阿部　雄太 (ABE Yuta) (70327526)	慶應義塾大学・医学部（信濃町）・講師 (32612)	
研究分担者	林田　哲 (HAYASHIDA Tetsu) (80327543)	慶應義塾大学・医学部（信濃町）・講師 (32612)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関