

生成AIによる情報統合型希少疾患診断支援システム「ZebraSeek」の開発と予備評価

○吉桑実弥¹ 岡崎鉄平² 申在紋³
Orion Buske⁴ 土肥栄祐⁵ 三嶋博之⁶

山口敦子² Tzung-Chien Hsieh⁷
千葉啓和³ 藤原豊史³ 五斗進^{1,3}

連絡先: yoshikuwa-naoya107@g.ecc.u-tokyo.ac.jp

1 東京大学 新領域創成科学研究科 メディカル情報生命専攻
2 東京都市大学 大学院情報データ科学研究科 情報データ科学専攻
3 国立遺伝学研究所 バイオデータ研究拠点 ライフサイエンス統合データベース部門
4 PhenoTips, カナダ・オンタリオ州トロント
5 国立研究開発法人国立精神・神経医療研究センター 神経研究所
6 長崎大学 原爆後障害医療研究所
7 ボン大学病院 ゲノム統計学・バイオインフォマティクス研究所

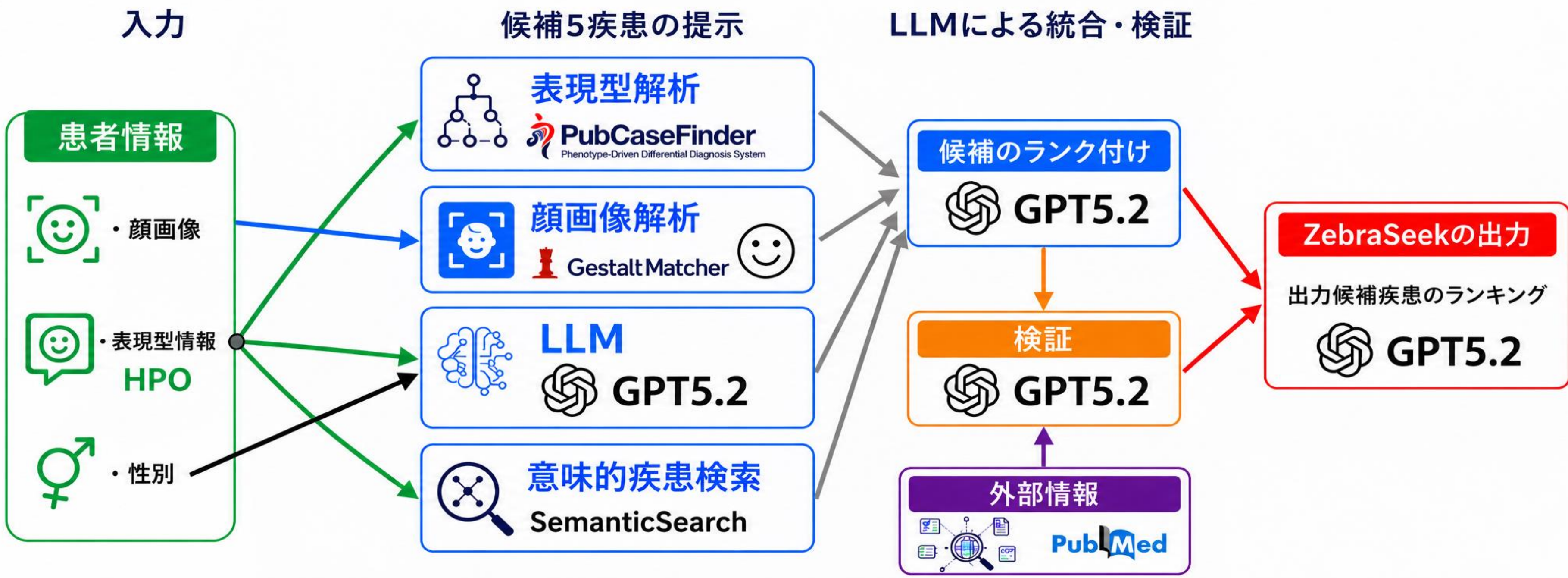
背景・目的

希少疾患では、個々の疾患の症例数が少ないことに加え、診断に有用な情報が患者の表現型、顔貌、症例報告、疾患データベースなどに分散している。そのため、患者の症状から適切な鑑別疾患を絞り込み、確定診断に至るまでに長い時間を要する Diagnostic Odyssey が課題となっている。

現在は、HPO（Human Phenotype Ontology／ヒト表現型オントロジー）を用いた表現型検索、顔画像を用いた疾患検索、症例報告や疾患記述に対する検索など、異なる情報に基づく診断支援手法が利用できる。一方、それぞれが利用する患者情報や得意とする症例は異なり、単独の手法だけでは候補疾患を十分に拾えない場合がある。

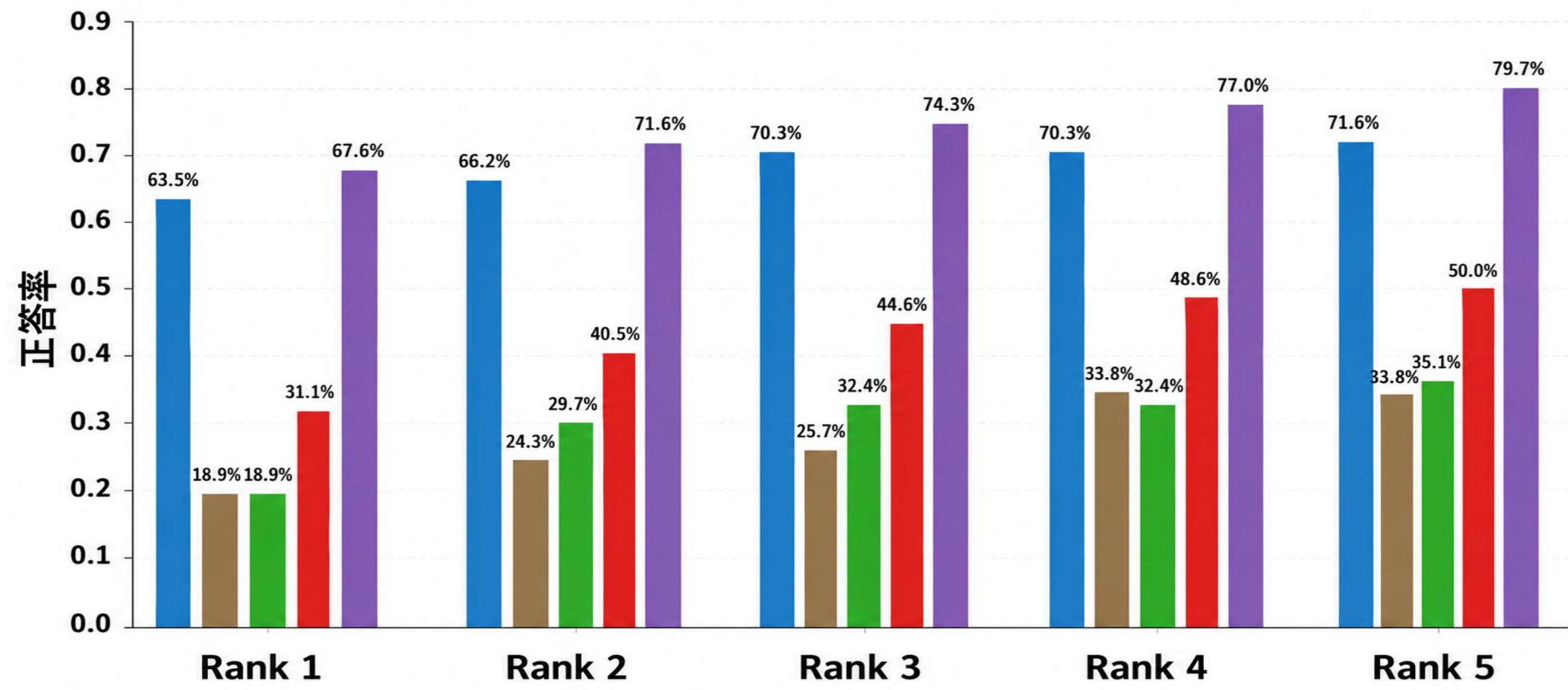
そこで本研究では、異なる情報を利用する複数の診断支援手法を組み合わせ、それぞれが取得した候補疾患を生成AIで統合・検証する希少疾患診断支援システム「ZebraSeek」を開発した。

ZebraSeekの概要



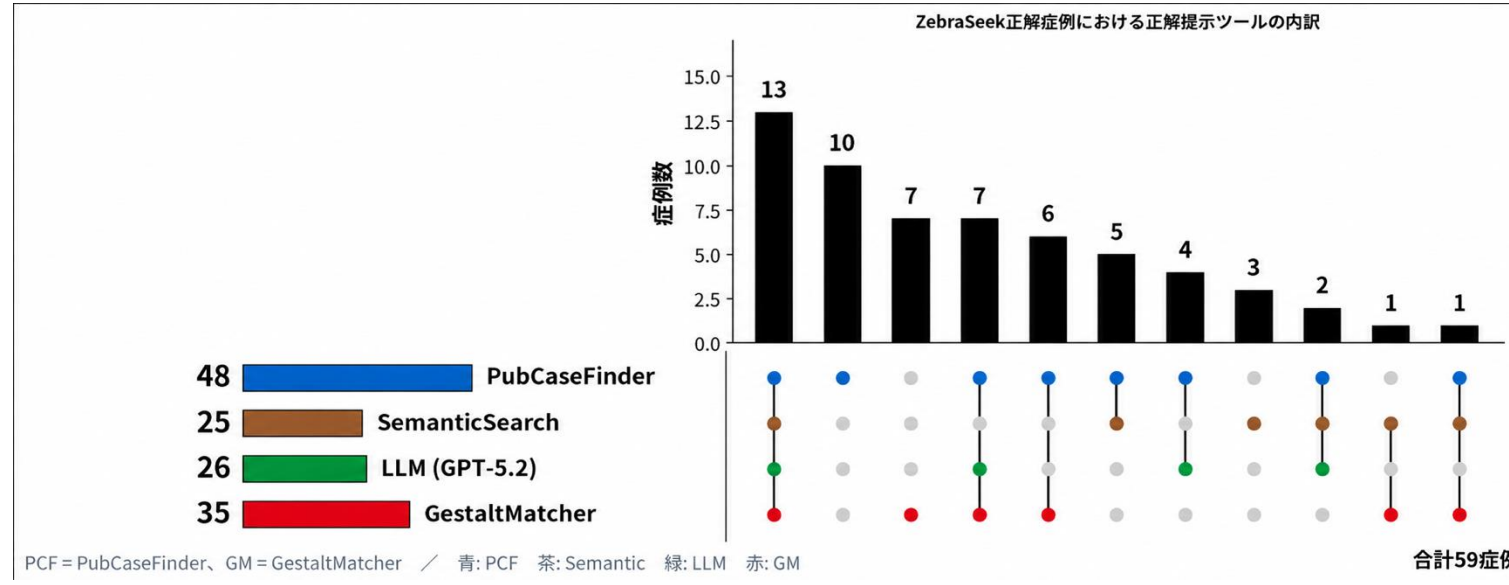
ZebraSeekでは、顔画像、HPO、性別を入力として、PubCaseFinder、GestaltMatcher、Semantic Search、LLM（Large Language Model／大規模言語モデル）から候補疾患を取得する。各ツールが生成した候補をそのまま並べるのではなく、LLMによって比較・再検証し、最終的な上位5疾患とその根拠を提示する。現行実装では、複数の候補生成手法から得られた最大20候補を統合・検証の対象としている。したがって、最終性能は「候補生成・候補選択の段階で正解疾患をLLMまで届けられるか」と「渡された候補の中から正解疾患を適切に上位へ順位付けできるか」の両方に依存する。

予備評価（74症例・19疾患）



Phenopacket StoreのHPO・性別・確定診断と GestaltMatcher Database の顔画像を対応付けた74症例・19疾患を用いて予備評価を行った。正解疾患が上位k候補以内に含まれる割合を Recall@k として評価したところ、ZebraSeek は Recall@1 67.6%（50/74症例）、Recall@5 79.7%（59/74症例）となり、Rank 1～5のすべてで比較した単独手法より高い値を示した。

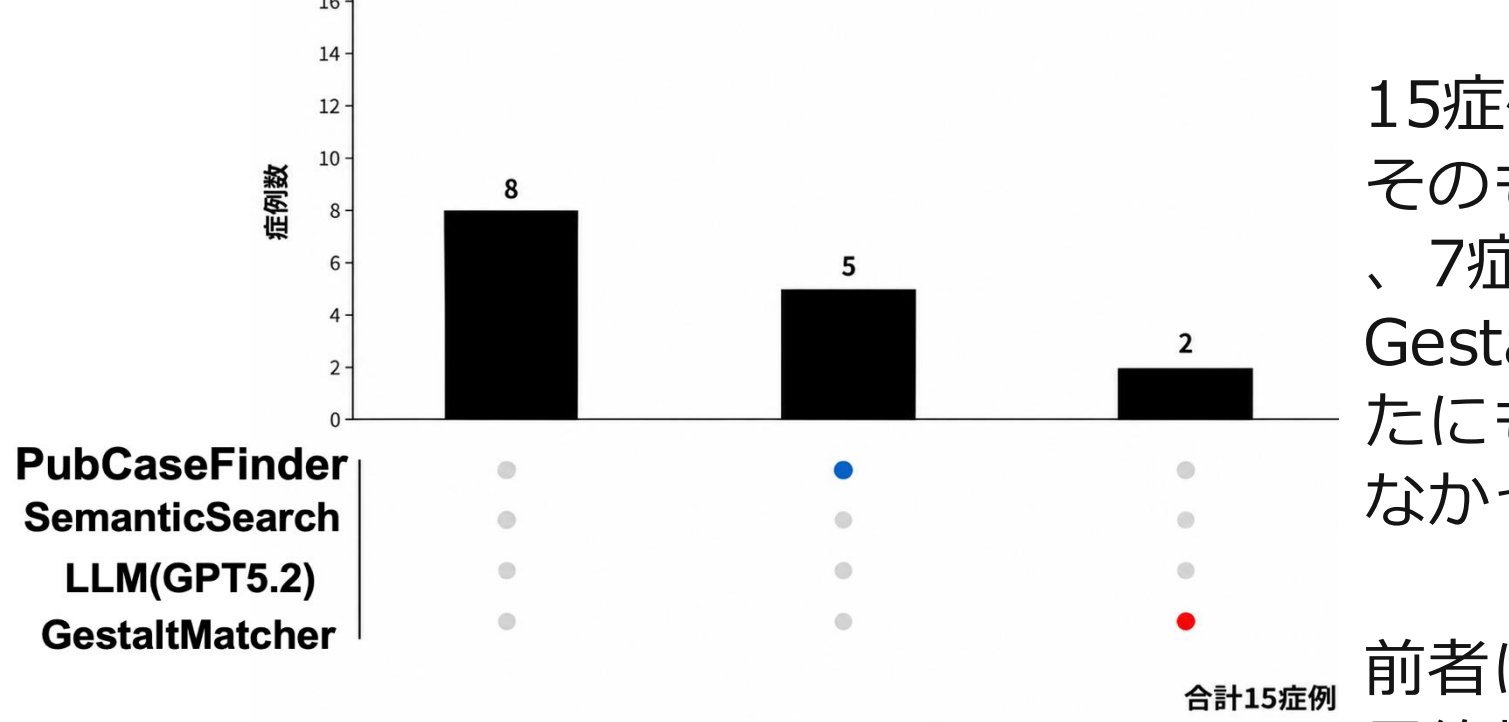
正解できた59症例について各ツールの結果を確認すると、PubCaseFinder のみが正解を含んでいた症例が10例、GestaltMatcher のみが7例、Semantic Search のみが2例存在した。すなわち、異なる症例を補い合っており、複数情報源を統合する設計自体に意味があることが分かった。



ツール間の補完関係

正解59症例の内訳を見ると、PCFのみ10例、GMのみ7例、Semantic Searchのみ2例が存在した。単一ツールでは拾えない症例を、他のツールが補っている。

ZebraSeek不正解症例における正解提示ツールの内訳

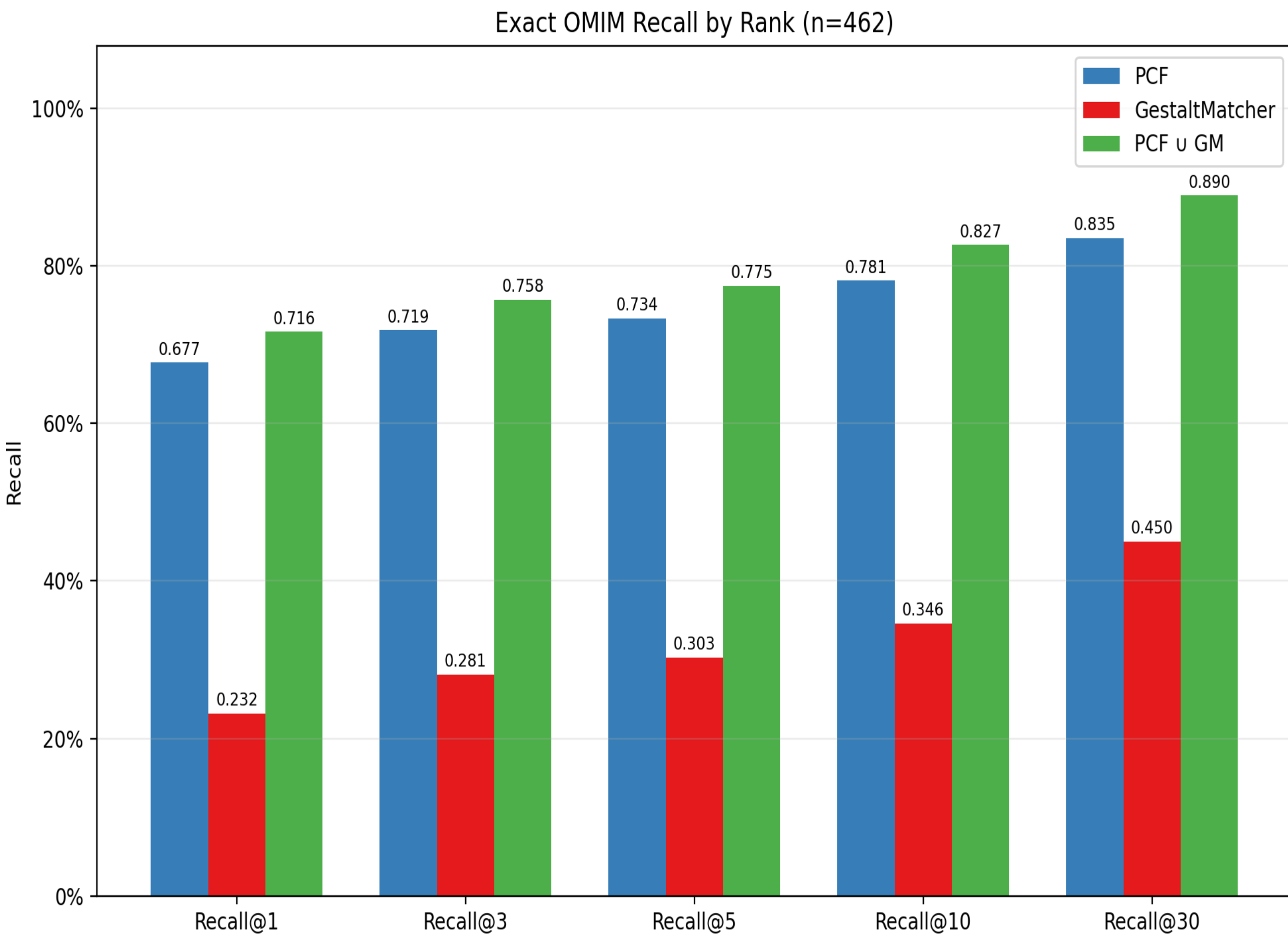


不正解15症例の解析

15症例を調べると、8症例では統合前の候補集合そのものに正解疾患が含まれていなかった。一方、7症例では PubCaseFinder または GestaltMatcher の候補に正解疾患が含まれていたにもかかわらず、最終的な上位5候補には残らなかった。

前者は現在のZebraSeekでは原理的に救えない。最終性能をさらに改善するには、まず統合前の候補集合に正解疾患を含める必要がある。そこで次に、各ツールが候補順位を広げたときに正解疾患をどこまで回収できるかを追加で調べた。

追加評価：統合前の候補集合にはどこまで正解疾患が含まれるか



Version 0.1.27

368 patients / 462 images / 54 disorders

目的

追加評価では ZebraSeek 全体の最終精度ではなく、LLM による統合前の候補集合に注目した。PubCaseFinder と GestaltMatcher について候補順位を1、3、5、10、30位まで広げた場合に、正解疾患がどこまで候補集合内に含まれるかを評価し、現在のツールでは原理的に拾えない症例と、より深い順位を見れば統合対象に含まれる症例を区別することを目的とした。

結果

PubCaseFinder は Recall@1 67.7% から Recall@30 83.5% まで上昇した。GestaltMatcher 単独の Recall@30 は45.0%であったが、両者の候補集合を組み合わせると Recall@30 は 89.0% となった。すなわち、PCF と GM の候補を十分広く取得すれば、約9割のデータで少なくともどちらかのツールが正解疾患を候補として保持している。

解釈と今後の方針

候補を広く見れば正解疾患はかなり拾えるが、PCF と GM からそれぞれ上位30件を取得すると、重複を除く前では最大60候補を LLM で比較・検証する必要がある。候補数が増えるほど入力トークン数、根拠検索の回数、推論時間、API 利用時の計算コストも増加する。したがって次の課題は、候補を増やすことなく、少ない候補数で正解疾患を候補集合に残すことである。

89.0%
PCF ∪ GM / Recall@30

今後は、さまざまな配分方法を比較し、Candidate Recall と最終 Recall@1 / Recall@5、ならびに推論時間・計算コストの関係を評価する。その後、HPO の情報量や顔貌情報、各ツールの出力スコアを用いて症例ごとに候補数を配分する方法を検討し、診断性能と計算効率の両立を目指す。