Improving variant prioritization in exome analysis by entropy-weighted ensemble of multiple tools

Yanjie Fan[1], *YingZhou*[2], *HuiliLiu*[1], *XiaomeiLuo*[1], *TingXu*[1], *YuSun*[1], *TingtingYang*[1], *LinlinChen*[1], *XuefanGu*[1], *YongguoYu*[1]

1.Shanghai Institute of Pediatric Research,Xinhua Hospital affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China; 2.GeneDock Corporate, Beijing, China

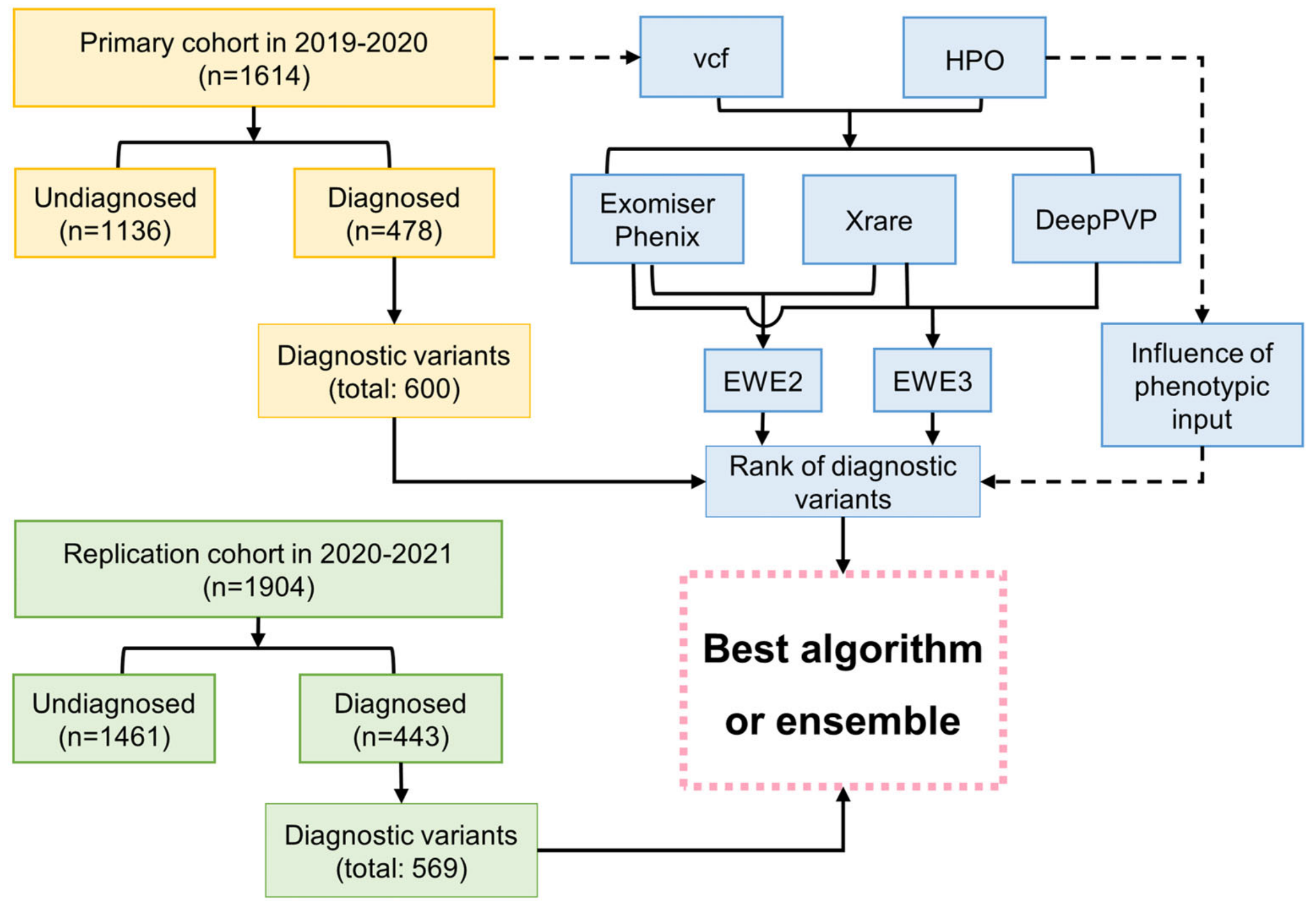
Yanjie Fan, Ying Zhou, Huili Liu, Xiaomei Luo, and Ting Xu contributed equally to this study.

Objectives

- Assessing the efficacy of variant prioritization based on real clinical WES data
- Testing the influence of phenotypic quality among automate the variant prioritization tools
- Using entropy weighted ensemble of multiple tools to improve variant prioritization and accelerate molecular diagnosis in exome/genome sequencing.

Introduction

Variant prioritization is a crucial step in the analysis of exome and genome sequencing. Multiple phenotype-driven tools have been developed to automate the variant prioritization process, but the efficacy of these tools in clinical setting with fuzzy phenotypic information and whether ensemble of these tools could outperform single algorithm remains to be assessed. A large rare disease cohort with heterogeneous phenotypic information were recruited to assess the efficacy of variant prioritization and their ensemble.



Primary cohort in 2019-2020 (n=1614)

Undiagnosed (n=1136)

Diagnosed (n=478)

Diagnostic variants (total: 600)

vcf

HPO

Exomiser Phenix

Xrare

DeepPVP

EWE2

EWE3

Influence of phenotypic input

Rank of diagnostic variants

Replication cohort in 2020-2021 (n=1904)

Undiagnosed (n=1461)

Diagnosed (n=443)

Diagnostic variants (total: 569)

Best algorithm or ensemble

Figure 1:Schematic overview of the study design.

Materials

The following materials were required to complete the research:

- The primary cohort included 1614 consecutive patients referred to exome sequencing at Clinical Genetics Center of Xinhua Hospital in 2019–2020.
- The replication cohort included 1904 consecutive patients in 2020–2021.
- Clinical notes including narratives of the condition, test results, suspected diagnosis, and treatment by the referring clinician.

Methods

The study setup is shown in Figure 1. It involved two groups: a main group and a replication group to confirm findings. Variants identified in solved cases were used for evaluation. For variant prioritization, Clinical information was converted into HPO sets, and three tools were used to prioritize variants. Two ensemble results (EWE2 and EWE3), were tested for better decision-making. The effectiveness of the algorithms was assessed based on how well they ranked diagnostic variants.

Conclusion

This study showed that using phenotype to prioritize genetic variants potentially speeding up diagnosis after genetic sequencing, especially when combining multiple tools.

Performance evaluation of Exomiser, Xrare, and DeepPVP

Exomiser and Xrare performed similarly well, outperforming DeepPVP. However, DeepPVP missed the fewest diagnostic variants individually, suggesting it could be ensembled with the other two algorithms.

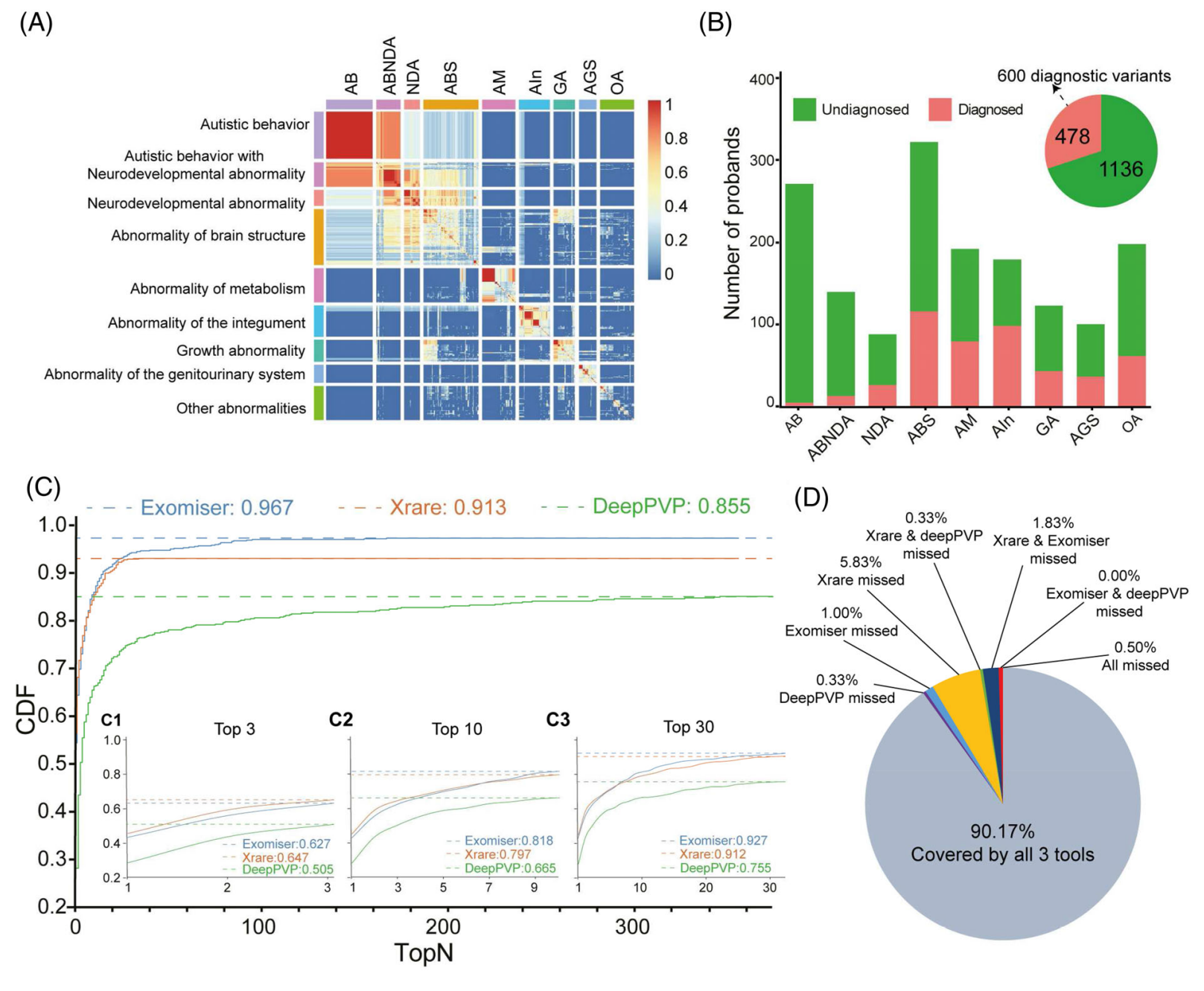


Figure 2:Phenotypic overview of the primary cohort and the performance evaluation of Exomiser, Xrare, and DeepPVP.

The influence of phenotypic input

Each algorithm performs best within specific premium ranges, depending on phenotypic attributes. When input is vague (IC<5), all three tools' performance declines. Exomiser shows greater stability across different IC ranges compared to Xrare and DeepPVP, indicating it can tolerate poor input better.

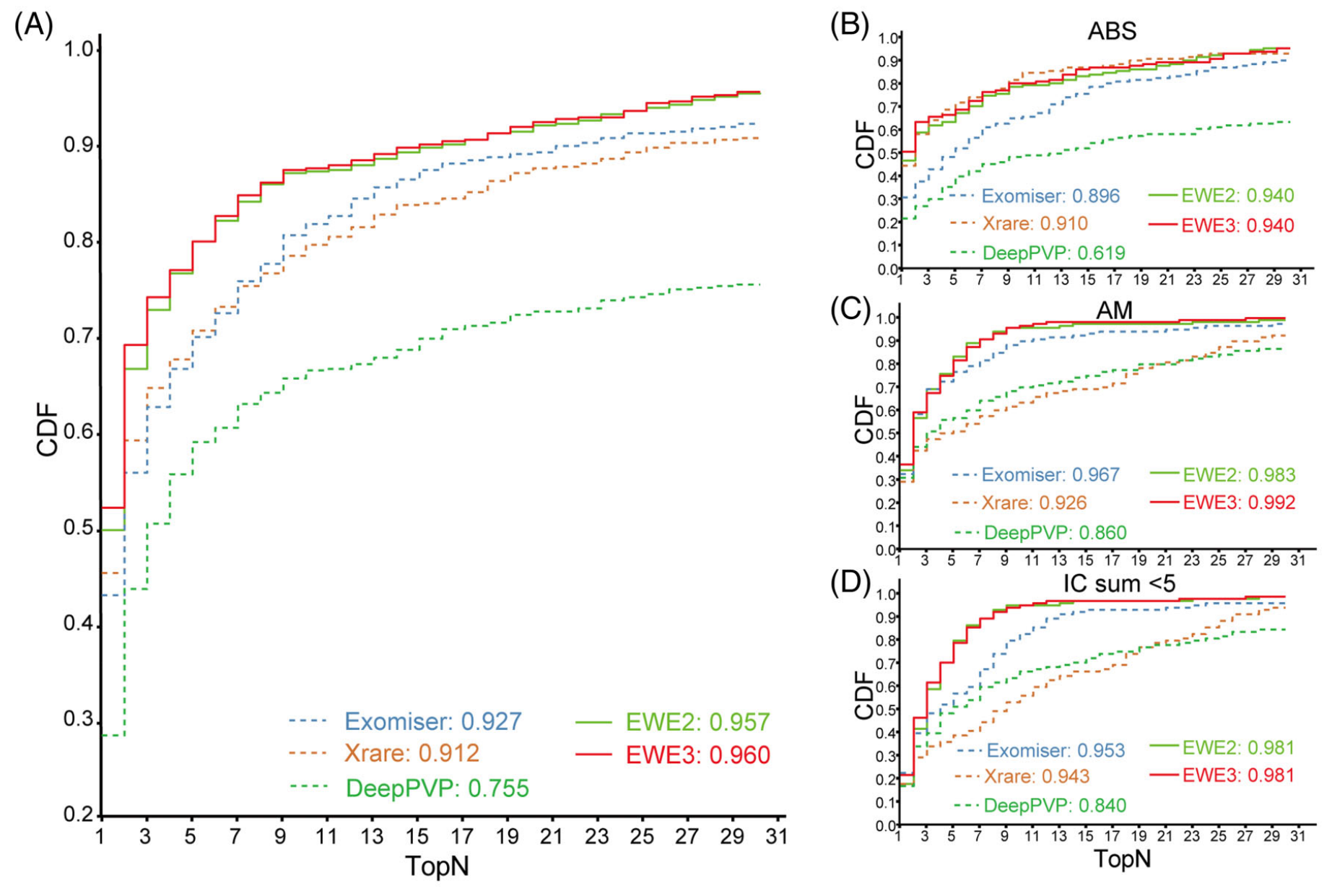


Figure 3:Ensemble of multiple algorithms and performance evaluation.

Ensemble of multiple algorithms improved the performance

As each algorithm has unique strength, we tested if the ensemble of multiple algorithms could outperform single algorithm. Ensemble of Exomiser and Xrare, were referred as “EWE2”; and the ensemble of all three algorithms—Exomiser, Xrare, and DeepPVP—were referred as “EWE3.” Our result showed that both EWE2 and EWE3 outperformed any single algorithm, and EWE3 achieved the highest efficacy (Figure3).

Evaluation of EWE3 in a replication cohort

EWE3 was further evaluated in a replication cohort. The results were similar to the findings in our primary cohort.

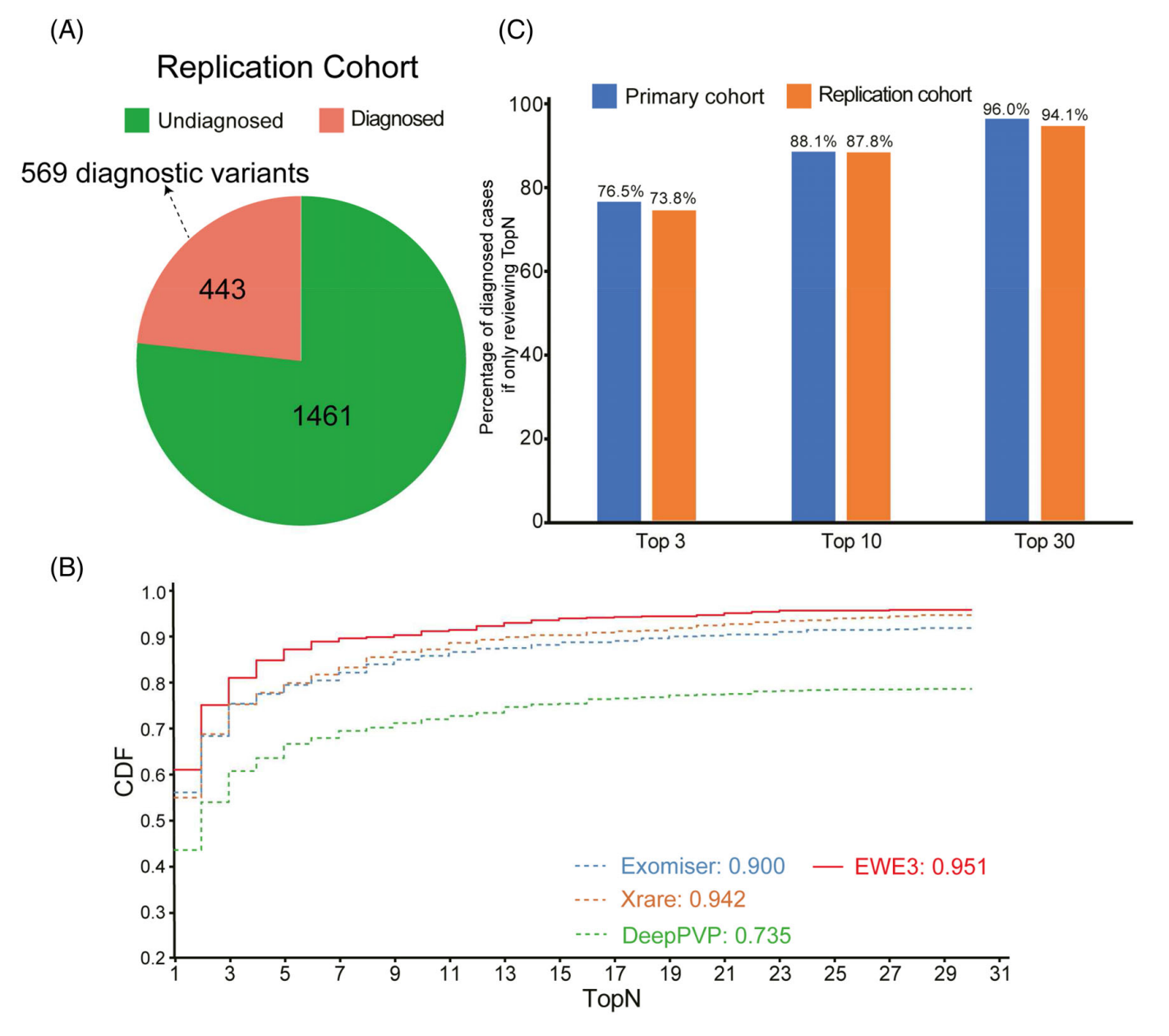


Figure 4:Evaluation of EWE3 in a replication cohort

My Contributions and Benefits from the project

I performed majority of the data analysis, plotted the figures and drafted the manuscript. Through attending this project, I learned basic model constructed methods and model evaluation. Moreover, I Understood clinical notes, mastered information extracted methods and mastered WES data analysis. I also practiced Writing skill.