API-based curation of biomarker–therapy relationships for WGS-guided precision oncology

Mohammad Naushad Alam[1,2,3], Muhammad Junaid[1,2,3], Karolina Hanna Prazanowska[1,2,3], and Su Bin Lim[1,2,3]*

[1]*Department of Biochemistry & Molecular Biology, Ajou University School of Medicine*

[2]*Department of Biomedical Sciences, Graduate School of Ajou University*

[3]*BK21 R&E Initiative for Advanced Precision Medicine, Ajou University School of Medicine*

**Corresponding author: sblim@ajou.ac.kr*

The rapid growth of biomedical literature and genomic data complicates timely identification of clinically relevant biomarkers. In this work, we present a hybrid literature-mining pipeline that couples deterministic rules with selective OpenAI API–based structured extraction to accelerate curation of biomarker–therapy relationships for precision oncology. We retrieved records from PubMed and PubMed Central using NCBI E-utilities and converted them into JSON format for efficient parsing, and applied structured prompt-based extraction to standardize entities such as gene symbols, variants, drugs, cancer types, assay methods, and clinical outcomes. Performance was strengthened through rule-based thresholds, lightweight domain-specific filters, and reproducibility logs employing harmonized vocabularies. At ~100,000 articles, our pipeline produces ranked evidence tables (gene–variant–drug–cancer, resistance/response, assay/context) that map directly to whole-genome sequencing (WGS) VCFs for patient-level annotation and reporting within hours. This rules-first, LLM-augmented design maintains transparency and reproducibility while improving coverage of non-canonical expressions. The framework operationalizes literature evidence for WGS-guided decision support in precision oncology.