Methods for Clinical Feature Abstraction and Driver Mutation Annotations Improve Utility of Real-World Cancer Data

by

Thinh Ngoc Tran

A Dissertation

Presented to the Faculty of the Louis V. Gerstner, Jr.

Graduate School of Biomedical Sciences,

Memorial Sloan Kettering Cancer Center

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

New York, NY

May, 2024

4/11/2024

Nikolaus Schultz, PhD
Dissertation Mentor

Date

For my parents

Dành tặng bố mẹ

ABSTRACT

Clinical and genomic real-world data (RWD) in oncology can offer valuable insights about patient outcomes and cancer biology, but challenges such as unstructured data fields and incomplete genomic annotations hinder their potential use. In this study, we developed and evaluated methods to address two main challenges in oncology RWD: unstructured clinical information and genomic variant interpretation.

Clinical data from electronic health records include rich information about disease and treatment; however, information about therapy outside of the host institution is only available in free-text initial consult notes, requiring manual curation to be usable in research. To overcome this bottleneck, we trained and validated natural language processing models using initial consult notes to identify whether patients had received treatment at external institutions and studied the impact of these putative treatments on tumor genomics. When deployed on 48,447 patients at Memorial Sloan Kettering Cancer Center, the model predicted 25% of patients to have received prior external treatment. Patients predicted to have received external treatment, similar to known pre-treated patients, had higher frequencies of resistance mutations and worse overall survival compared to treatment-naïve patients.

Genomics data is another frequently collected data modality in oncology. Identifying cancer driver mutations from sequencing data, however, is challenging, with 80% of somatic missense mutations labeled as variants of unknown significance (VUSs) in the multi-institutional pan-cancer AACR GENIE

cohort consisting of ~180,000 tumors. We explored the potential of 11 computational variant effect predictors (VEPs) in improving the identification of potential oncogenic variants. AlphaMissense, a deep-learning method based on the protein structure prediction tool AlphaFold2 and trained without human-curated labels, emerged as most accurate in annotating known oncogenic variants. Validation in two cohorts of non-small cell lung cancer patients showed associations between reclassified pathogenic VUSs in *KEAP1* and *SMARCA4* and worse overall survival. Reclassified pathogenic VUSs also occurred preferentially in binding sites and were mutually exclusive with known driver mutations in RTK/RAS and NRF2 pathways, further suggesting biological validity.

Taken together, this work demonstrates that machine learning methods can mitigate the challenges associated with RWD in oncology by enabling automatic abstraction of clinical variables stored in unstructured format and broadening the subset of annotated pathogenic variants in cancer, thus allowing RWD to be used more efficiently in studying tumor biology and patient outcomes.

BIOGRAPHICAL SKETCH

Thinh Ngoc Tran was born and raised in Ho Chi Minh City, Vietnam, but also calls West Chester, Pennsylvania, where she attended boarding school since she was fifteen, her second home. She attended New York University Abu Dhabi (NYUAD) in the United Arab Emirates as part of its fourth inaugural class, where she majored in Biology and dabbled in Sociology, Computer Science, Creative Writing and Arts History. She spent semesters abroad in London, Florence, Buenos Aires and New York City, where she tried to juggle her Biology major with her other academic interests and love for traveling. Since her sophomore year at NYUAD, she had worked in the lab of Dr. Claude Desplan to study the underlying genetic mechanisms of how neuronal stem cells give rise to an impressive number of cell types in the *Drosophila* optic lobes. After graduating *cum laude* with her Bachelor of Science, Thinh attended the University of Toronto for her Master's degree in Molecular Genetics, where she learned about single-cell RNA sequencing, pathway analysis and computational biology from Dr. Gary Bader. During her time in Toronto, Thinh developed Tempora, a trajectory inference method for time-course single-cell RNA sequencing data. Thinh joined Gerstner Sloan Kettering Graduate School in 2019 and Dr. Nikolaus Schultz's lab in 2020, where she worked on a range of projects related to cancer genomics, treatment responses and real-world cancer data.

ACKNOWLEDGEMENTS

First and foremost, my deepest gratitude goes to my mentors, Niki Schultz and Justin Jee, without whom this work would not have been possible. Niki first introduced me to the captivating world of cancer genomics and has continued to guide me through it many years later, with impressive attention to both detail and the big picture. Justin, with his unwavering support and fearless approach to science, has not only encouraged me to work hard and persevere, but also helped me realize the rewarding aspects of an otherwise challenging process. Their dedication to science and mentorship has been invaluable in allowing me to explore interesting questions, fail safely and learn from those mistakes.

Beyond my two mentors, I would like to thank many individuals and groups at Memorial Sloan Kettering who have played crucial roles in my academic and professional journey. I am thankful to my committee members, Michael Berger and David Solit, for their guidance, expertise and invaluable input throughout my research endeavors. I have been extremely fortunate to work with Karl Pichotta, Chris Fong, Siyang Liu, Michele Waters and many other brilliant people in the Clinical Data Mining team – I have learned so much from them not only on the science front, but also on how teamwork can fuel innovation. I am thankful to the GENIE BPC working group, in particular Greg Riely and Ronglai Shen, for introducing me to the amazing resource that is GENIE BPC data and teaching me how to work with it. As a PhD student in the Schultz lab, which is the intersection of many great teams in the Cancer Data Science Initiative, I have also been lucky to collaborate with and learn from amazing people, including

Debyani Chakravarty and Moriah Nissan from OncoKB, Ino de Bruijn and the cBioPortal team, dedicated members of the Translational Research Network, as well as brilliant labmates. Finally, I want to acknowledge the unsung heroes behind the scenes who contribute to the smooth running of my PhD – Dean Michael Overholtzer, David McDonagh, Thalyana Stathis, Linda Burnley, Tom Magaldi, Desiree Lara, Keshia Pierre, Nicole Scagliola, as well as other members of the administration at GSK Graduate School, Computational Oncology and Epidemiology/Biostatistics. Without them keeping track of my progress, reminding me of deadlines and funding opportunities, and finding time for me to meet with extremely busy people at MSK, this process would have been so much more difficult.

Moreover, I want to acknowledge those individuals who have inspired and supported my decision to pursue science. From NYUAD, I extend my thanks to Professor Claude Desplan, Nikos Konstantinides and members of the Desplan Lab both in New York and Abu Dhabi for their generosity and patience during my earliest days in the lab, as well as to Professor Joseph Gelfand, Professor Duncan Smith, Joe Koussa and Ibrahim Chehade for their dedication to teaching and mentoring. From the University of Toronto, I am grateful to Professor Gary Bader, who instilled in me the love for data and algorithms, as well as Professor Gordon Keller, Maria Abou Chakra and Brendan Innes for their inspiration, encouragement, and contributions to my growth as a scientist.

Last but not least, I am blessed to have a steadfast support system behind me during this journey. Thank you Monique and Carmel for keeping me sane, quite literally. Thank you, family and friends, for sticking with me through 20+ years of education and pretending to understand when I explained my work. Cảm ơn bố mẹ rất nhiều vì đã tin tưởng và ủng hộ con. Olive, thank you for always keeping my feet on the ground, cheering me up and dispelling my anxieties one at a time. Finally – all my love and appreciation go to my tiny loving household: Gabe, thank you for always keeping me company when I work late and waiting for me by the door whenever I come back. Nghiêm, thank you for believing that I am capable of more than I give myself credit for, and for convincing me to believe the same too. I'm so happy we get to do this together.

It really does take a village. Thank you all so much, I am forever grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	artificial intelligence
AACR	American Association of Cancer Research
AUROC	area under the receiver operating curve
BPC	Biopharma Collaborative
BRCA	breast cancer
CNN	convolutional neural network
CRC	colorectal cancer
CV	cross validation
DNA	deoxyribonucleic acid
EHR	electronic health records
FDA	Food and Drug Administration
FN	false negative
FP	false positive
GENIE	Genomics Evidence Neoplasia Information Exchange
HGMD	Human Gene Mutation Database
IC	initial consult
IPTW	inverse probability of treatment weighting
KM	Kaplan-Meier
LR	logistic regression
MAVE	multiplexed assays of variant effect
MSA	multiple sequence alignment
MSI	microsatellite instability status
MSK	Memorial Sloan Kettering Cancer Center
NLP	natural language processing
NSCLC	non-small cell lung cancer
OG	oncogenes
OS	overall survival
PANC	pancreatic cancer
PH	proportional hazard
PPI	protein-protein interaction
PRAD	prostate cancer
RCT	randomized controlled trial
ROC	receiver operating curve
RWD	real-world data
RWE	real-world evidence
SEER	Surveillance, Epidemiology, and End Results
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
TF-IDF	Term Frequency - Inverse Document Frequency
TMB	tumor mutational burden
TPR	true positive rates
TSG	tumor suppressor gene
VEP	variant effect predictor
VUS	variants of unknown significance

CHAPTER 1

INTRODUCTION

Patient data are routinely collected as part of clinical care, resulting in rich sources of real-world data (RWD). RWD comprises clinical data from electronic health records (EHR), insurance claims and registry data, and may include information about demographics, diagnosis, treatment, outcomes and more depending on the source[1]. In oncology, besides clinical data, multiple large-scale genomic datasets from real-world cohorts have also been collected as a result of routine sequencing to facilitate precision medicine[2–4]. Here, we provide an overview of RWD usage as well as its strength and limitations in oncology research, first focusing on clinical data, followed by genomic data.

1.1. Real-world clinical data in cancer

Real-world clinical data, with its ongoing collection on a population scale, is frequently more extensive and temporally complex compared to randomized controlled trials (RCTs). Therefore, they can supplement RCTs, contributing to the comprehensive study of drug efficacy and increasingly included as evidence in submissions for regulatory approvals of agents[5].

Long-term drug safety profiles, overall survival and other endpoints of interest are not fully characterized during RCTs due to the limited length and prohibitive cost of these trials, but can easily be studied using RWD[6,7]. The Food and Drug Administration (FDA) relies on real-world evidence (RWE) generated from RWD for postmarket safety monitoring and reporting through the Sentinel Initiative, which can result in label changes or withdrawal of agents if risks of

adverse events outweigh clinical benefits[8]. For example, gemtuzumab ozogamicin, a humanized anti-CD33 monoclonal antibody approved to treat relapsed acute myeloid leukemia, was assigned a black-box label after a year on the market because of evidences showing increased hepatic veno-occlusive disease in patients[5,9,10]. Finally, real-word evidence (RWE) generated from RWD can also support label expansion and off-label usage of approved FDA drugs[11]. Palbociclib, a CDK4/6 inhibitor approved to treat metastatic breast cancer in women, was approved as a treatment of men with breast cancer in 2019 based on the review of EHR and insurance claims data[12]. Finally, RWD allows for clinical efficacy studies of treatment in larger and more diverse populations, as participants enrolled in RCTs often need to meet certain demographics and clinical criteria such as age and performance status that may not be representative of the population[13,14].

Beyond its use in drug efficacy evaluation, RWD has empowered studies that shed light on cancer biology, epidemiology and social determinants of cancer outcomes. Together with genomic data, clinical RWD provide feature-rich training data and enable the development of multi-modal risk stratification/outcome prediction models that outperform traditional single-marker models and highlight novel prognostic factors[15,16]. National cancer registries such as the U.S. Surveillance, Epidemiology, and End Results (SEER) Program[17] empower systematic investigations into disparities in cancer outcomes, which helps establish the association between socioeconomic factors and poor prognosis for

patients with cancer, potentially due to later stage at diagnosis, environmental exposure, as well as poor access to healthcare, treatment and screening[18–20].

Challenges in analyses of RWD

Although RWD enables the cost-effective accumulation of a substantial amount of data that can be used to answer many research questions as discussed in the previous section, they are not specifically designed to support research[21,22]. Consequently, various considerations emerge when utilizing RWD to address research questions, especially those aiming to identify the causal effects of an intervention or a certain feature on an outcome of interest[23–25].

Firstly, RWD is not randomized. The randomization in RCTs is carried out carefully to ensure both observable and unobservable confounders are balanced and allow analyses to identify the true effect of an intervention of interest[26–28]. Without randomization, observational RWD is subjected to effects of confounders and biases, i.e. factors that are correlated to both the treatment received and the outcome, thus obscuring the true causal effect[27,29]. For example, physical comorbidity in cancer both reduces response to treatment and increases mortality risk, making it an important confounder to account for in studies on treatment efficacy[30]. Methods such as propensity score matching and inverse probability of treatment weighting (IPTW) can alleviate confounding effects by matching or weighting treated and untreated subjects based on a number of observable characteristics to ensure that the “control” and “treatment” groups are comparable[29]. These statistical methods, though powerful, are heavily affected by the selection of variables to be adjusted for using propensity score weighting or

matching[29]. Incomplete or unstructured data fields in RWD, thus, limit which confounding variables can be accounted for. Another limitation is that propensity score adjustment cannot account for unobservable baseline characteristics. For more rigorous causal inference in highly skewed populations or in populations with few available confounders to adjust for, quasi-experimental methods such as difference-in-differences, which creates a pseudo-control group from an untreated population and compares the difference in outcome before and after treatment in the treated and untreated populations, are required[31,32]. Difference-in-differences have been widely used in evaluating healthcare policies as it can provide robust estimation of treatment effects using observational data, but it requires multiple strong assumptions to be met and can be biased in populations receiving treatment at different times[33,34].

Secondly, RWD is incomplete and contains unstructured data fields[35,36], requiring laborious manual curation to be usable in downstream analyses. Efforts such as AACR Project GENIE Biopharma Collaborative (GENIE BPC)[37] have been made to manually curate comprehensive clinical data elements, including treatment histories, pathology, biomarkers and outcomes, for a small number of patients, which results in feature-rich cohorts that lend themselves to diverse use cases in oncology research[38]. However, the number of curated cases remains small (less than 2,000) for all cancer types due to the high cost and unscalable manual labor. The small cohort size, thus, limits the use of these datasets in studying patient populations with rare clinical-genomic characteristics.

Finally, RWD can suffer from lack of interoperability and traditional data silos, which limit data integration across institutions and data modalities[39]. The increasing number of commercial RWD products also leads to inconsistencies in data structures and models, further hindering the integration of different datasets[40]. Studies have begun to challenge these traditional siloes, for example through the integration of tumor sequencing with antineoplastic data to uncover genomic modifiers of treatment response[13], or the integration of billing codes with tumor sequencing to uncover genomic alterations associated with specific organ sites of metastasis[14].

Addressing unstructured fields in clinical RWD: Automated abstraction of features from free-text reports

The large amount of information stored in free-text patient visit, radiology, histopathology and procedural notes is an important hurdle to the use of RWD in research[41]. Although abstraction of clinical variables from unstructured texts has traditionally required laborious manual abstraction, natural language processing (NLP) algorithms have allowed for automatic annotation of such features[42,43]. The early use of NLP in analyzing EHR involves tasks such as name entity recognition, e.g. medication names[44], classification e.g. identifying mammogram reports indicative of breast cancer[45,46], information retrieval e.g. adverse drug reactions[47], among other use cases[42,48]. These tasks are commonly achieved using three approaches: 1) rule-based methods, where a set of criteria such as matching specific terms is used to classify texts; 2) machine learning models, including logistic regression and random forest; and 3) deep learning models,

such as recurrent neural networks and convolutional neural networks[48,49]. Rule-based methods can be applied directly to unstructured texts, while machine learning and deep learning models require that words in free texts are first transformed into vectorial representations using different methods such as word2vec[50]. All of these approaches have demonstrated reasonable performance in their evaluated tasks, but have clear limitations. Rule-based approaches require significant domain and database-specific knowledge[51]. Recurrent and convolutional neural networks are computationally expensive to train and struggle with long-term dependencies, i.e. related words separated by a significant distance[52]. The fixed embeddings that serve as input to machine learning and other deep learning models also cannot handle words that do not previously exist in their vocabularies, as well as do not allow for the consideration of the same word in different language contexts, thus limiting their utility in more complex language tasks[53].

More recently, transformer architectures[54,55] have transformed the NLP landscape. Compared to recurrent neural network models, transformers can parallelize well, allowing it to be trained on large corpora of texts[55–57]. Transformers' positional encoding and self-attention mechanism allow these models to consider the position of words in a sentence, create contextualized representations of words and model longer-range relationships between words[54]. After pre-training, transformers can be fine-tuned using labeled data to perform specific language tasks. Transformers pre-trained on health record data[56,58] have shown promise at a number of medical tasks including predicting hospital

readmission[59], automated International Classification of Disease coding from clinical notes[60], answering health-related questions [61] and querying clinical trials[62].

In conclusion, the integration of clinical RWD in oncology research has proven invaluable, providing extensive insights into drug efficacy and safety, as well as enabling the development of predictive models and population-wide studies of health disparities. Despite challenges such as non-randomization and data incompleteness, innovative study designs and advanced statistical methods can effectively address these issues. Additionally, ongoing efforts in NLP offer promising solutions to automate the abstraction of key features from free-text reports, further enhancing the utility of clinical RWD in advancing oncology research.

1.2. Genomic data in cancer

Over their lifetime, somatic cells accumulate changes in their DNA through point mutations, copy number variations such as insertions or deletions (indels), and other structural variants such as fusions. Some of these changes can occur in genes that confer growth advantage by altering a multitude of processes, from sustained proliferative signaling to immune evasion, thus resulting in tumor development[63–65]. These genetic drivers of cancer have been studied using low-throughput approaches for decades[66], leading to two first success stories in identifying targetable mutations that respond to specific molecular inhibitors: *BCR-ABL* fusions in chronic myeloid leukemia, targetable by imatinib[67], and *EGFR* mutations in lung cancer, targetable by gefitinib[68]. These success stories,

along with the rapid development of high-throughput sequencing technologies, pioneered the new era of precision oncology and large-scale cancer genomics[66,69]. Since the first sequenced whole cancer genome[70], large multi-institutional efforts have been made to systematically characterize the cancer genomes pan-cancer and in specific tumor types, aiming to comprehensively identify cancer genes and alterations[66,71,72]. As more molecular targeted therapies are developed and approved[73], clinical sequencing to match patients with the right treatment according to their genomic aberrations have also become routine, giving rise to large genomic datasets of real-world patient cohorts, such as MSK-IMPACT[2] and AACR Project GENIE[74]. Along with this impressive amount of data comes great challenges in detecting variants, annotating their functions and identifying their mechanisms of oncogenicity, including the pathways they affect, among others[66,69]. Here, we discuss the challenges with variant interpretation in cancer.

Cancer driver vs. passenger mutations

One of the main goals and accomplishments of large-scale cancer genomics projects is the systematic identification of genes involved in cancer development, termed ‘cancer genes’. These cancer genes encompass three broad categories: oncogenes, tumor suppressor genes, as well as genes involved in signal transduction pathways, genome stability, chromatin maintenance, transcription, metabolism among other processes[64,66,75,76]. Oncogenes (OGs) are constitutively activated by gain-of-function alterations, such as amplification and other activating mutations, to promote cell growth and

survival, while tumor suppressor genes (TSGs) are inactivated by aberrations such as missense/truncating mutations and deletions, resulting in unchecked cell cycle progression and growth[65]. Over 300 cancer genes have been identified[76].

The high mutation rate and genomic instability characteristic of cancer cells give rise to a large number of mutations in these cancer genes as well as across the genome[77]. However, only some of these mutations have been causatively linked to neoplastic growth[76]. Mutations that confer fitness advantage and can drive cells towards neoplasm are known as driver mutations, while the rest of the observed mutations are termed passengers mutations.

Driver mutations are able to confer proliferative advantage and drive cells to neoplasms. Driver mutations are recurring[78,79] and positively selected for in cancer[80]. They affect the clinical phenotypes, prognosis, and in many cancer types, treatment responses. For example, among patients with non-small cell lung cancer (NSCLC), those with *KRAS* driver mutations have worse prognosis compared to those with *EGFR* drivers after adjustment for tumor stage[81,82]. In NSCLC, certain drivers in *EGFR*, *ALK*, *KRAS*, *RET*, *ROS1*, *BRAF* and *MET* are targetable, allowing patients carrying these drivers to benefit from treatment with targeted therapy that can significantly improve survival[83].

Despite the importance of driver mutations in cancer biology, a saturation analysis suggests that the majority of mutations in cancer are passenger mutations or variants of uncertain significance (VUSs)[78]. Even though passenger mutations are assumed to be neutral, multiple evidences have suggested that their roles in cancer are not fully understood[84]. For example, ‘latent’ or ‘mini’

drivers are mutations that have neutral effects by themselves, but can drive cancer progression and resistance when occurring in combination with other mutations[85,86]. VUSs have been shown to accumulate in cancer genomes and decrease fitness of cancer cells[84,87]. Passenger mutations observed in cancer genomes have been predicted *in silico* to have more damaging effect on protein function compared to known neutral mutations[88]. Beyond the potentially accumulative damaging effect that all VUSs may have on cancer, a fraction of VUSs may be un-annotated drivers with clinical and therapeutic implications. Functional genomics characterization of 438 recurrent VUSs in 36 genes detected in a real-world cancer cohort reclassified 97 VUSs (22%) as potentially actionable[89]. Systematic efforts to characterize variants in *BRCA1* and *BRCA2* provided functional evidence that some VUSs disrupted protein function and thus should be classified as pathogenic[90–92]. It is therefore of great interest to look into the 'dark matter' of the cancer genomes in order to develop a more comprehensive understanding of how genomic alterations influence cancer biology.

Variant annotations in cancer

Even though systematic functional characterization of variants can provide valuable information about the biological effect of variants, it is labor-intensive and challenging, as appropriate functional assays and read-outs may not be available for all genes[90,91]. Most experimental investigations of variants, therefore, are limited to a few genes or variants and do not scale with the increasing number of variants detected[89,91,93]. Therefore, multiple

computational methods to predict variant effects have been proposed to bridge this gap and help prioritize variants for functional validation[94,95]. *In silico* annotation of variants is a challenging task due to the many ways that genetic alterations may affect protein functions[96]. In cancer, the complexity of this task is heightened by the variation in variants across different tumor contexts and the need for biological knowledge to sift through predictions. Identifying rare cancer drivers with biological and clinical validity amid potential spurious signals, which are unavoidable given the amount of data and their inherent heterogeneity, demands multifactorial analyses that integrate knowledge about mutational patterns and functional effects, which goes beyond assessing statistical significance[69,97].

Databases of variant effects

Multiple databases and knowledge bases have been developed to consolidate information about variants from experimental results, publications and clinical guidelines. These databases serve as valuable resources for researchers and clinicians, enabling them to access comprehensive and up-to-date information on variants. Additionally, these databases play a crucial role in fostering collaboration among the scientific community, promoting the sharing of data and insights that contribute to a more comprehensive understanding of genetic variation and its implications for human health.

ClinVar is a public repository of human variants as well as evidence for their functional impacts and clinical significance[98]. Variants are submitted by research groups, clinical research labs and expert panels worldwide. Multiple

submissions for each variant are then aggregated and reviewed to provide a consensus interpretation of clinical significance. As of January 2024, more than 2 millions unique variations with interpretation have been submitted, most of which are germline variants[98].

Due to the small number of somatic variants and lack of information about clinical actionability in ClinVar, it is often less informative for clinical use in oncology. Various oncology-specific knowledge bases, such as OncoKB[99], MyCancerGenome[100], JAX Clinical Knowledgebase[101] among others[102], further curate variants based on their functions in cancer and their actionability, i.e. whether they can be targeted by existing therapies. OncoKB is an FDA-approved precision oncology knowledge base that curates biological effects and clinical implications of variants in cancer[73,99]. Variants are classified as oncogenic, likely oncogenic, likely neutral or unknown according to manual review of public cancer variant databases and clinical guidelines, as well as variant-specific evidence such as recurrence status, *in silico* predictions and preclinical evidence from peer-reviewed literature[103]. Oncogenic mutations are further assigned a level of actionability depending on evidence for drug use in specific indications, ranging from standard card biomarkers to FDA-approved drugs (Levels 1 and 2) to investigational biomarkers with clinical and/or biological evidences supportive of response to drugs (Levels 3A, 3B and 4). As of January 2024, more than 7,000 variants across 136 cancer types have been curated in OncoKB, with ~1,600 variants (~20%) associated with a level of actionability[73]. OncoKB's rigorous variant curation process and clinical expertise renders its variant classification

highly accurate and specific; however, the manual effort required to curate each variant results in a limited number of annotated variants in the knowledge base as well as a reporting bias towards known, clinically actionable driver genes[104].

Computational prediction of mutational effects

Multiple computational approaches exist to annotate variants, which help bridge the large gaps between variant detection in cancer sequencing and variant functional studies. One approach is to leverage the mutational patterns and statistical properties of driver mutations, such as recurrence, to identify mutations that follow these expected patterns. Cancer Hotspots[79,105] identifies mutational hotspots, which are residues mutated more frequently than would be expected in the absence of selection and thus potentially associated with driver activity in cancer, using a statistical model that incorporates gene-specific background mutation rates. 3D Hotspots extends the idea of Cancer Hotspots by identifying hotspot mutations that cluster together for proteins with available 3D structures[106]. Mutational recurrence identified by algorithms such as Cancer Hotspots is considered to be strong evidence supporting the oncogenicity of a variant: for example, OncoKB considers all ‘hotspot’ mutations to be likely oncogenic[103]. However, methods that rely on statistical properties of mutations can miss out on rare driver mutations, which are less frequently detected in sequencing samples and thus may not be able to achieve statistical significance.

Another approach to annotating mutations is to predict the function of single variants using variant effect predictors (VEPs), which does not rely on the statistical properties of a variant of interest and thus can annotate rare variants.

Multiple VEPs have been developed to leverage information regarding evolutionary conservation, protein structures and physical and chemical properties of the original vs. substituted residues to predict whether a variant is functionally damaging or benign[95].

VEP approaches

VEPs' approaches differ by the features that they incorporate in predicting variant pathogenicity and the algorithm they use to predict pathogenicity from the selected features. Each approach comes with its own strengths and drawbacks, contributing to their varying performance when evaluated on the same dataset[94,95]. Here, we provide a broad overview of available VEPs grouped by the features and algorithms they use, as well as discuss the strengths and limitations of these approaches.

VEPs leverage two complementary sets of features to infer variant effects: evolutionary conservation and protein structure information[96]. Evolutionary conservation can indicate functional significance, as highly conserved sites may be crucial for protein function and more likely to cause damage when mutated[107,108]. On the other hand, protein structure information and other physicochemical properties are useful for understanding the structural context of a modified amino acid. This helps identify variants that may induce significant changes to the protein structure, potentially disrupting normal protein functions[109]. Different VEPs model variant pathogenicity based on varying combinations of features from these two primary feature types and assign varying weights to them when more than one feature type is used. The difference in input features

of these methods partially explains the differences in their predictions and performance[94,95,110].

The first group of methods considers mainly evolutionary conservation information under the assumption that well-conserved sites are important for protein functions; therefore, mutations occurring at these sites are likely to be more pathogenic than those occurring in less well-conserved sites. These methods, including SIFT[108], MutationAssessor[107], FATHMM[111] and others, rely on homologous sequence searching across species and multiple sequence alignments (MSAs) to identify well-conserved sites. SIFT additionally considers the type of amino acid change in predicting pathogenicity[108]. Beyond constructing MSAs, MutationAssessor also groups sequences into protein families and subfamilies to identify residues specific to protein families, thus adding a layer of functional information to predict variant effects[107]. FATHMM uses hidden Markov models to represent the alignment of homologous sequences and conserved protein domains, and predicts the effect of a mutation on protein function by examining changes in amino acid probabilities and weighting with with inherited disease-specific or cancer-specific pathogenicity weights[111,112]. Besides homology search and MSAs, a new generation of deep learning-based approaches such as EVE[113] uses a generative language model to learn the distribution of sequence variation across species for each human protein and estimate the likelihood of a particular variant using the learned distribution.

Few VEPs are based solely on protein structure information, as this approach relies on, and is limited by, the availability of 3D protein structures.

Within this category, methods focus on predicting variant effects based on changes in protein folding stability[114,115] and protein binding[116]. A more recent approach to using protein information in predicting variant effects that does not rely on 3D structures is deep protein language modeling, as seen in ESM1b[117]. Protein language modeling uses language models, similar to those used to analyze texts, to understand and analyze sequences of amino acids in proteins[118]. These models learn from all known protein sequences, capturing complex relationships between amino acids and helping to predict how changes in the sequence might affect the protein's function and can incorporate existing structural and functional knowledge[117,118]. Built on a large protein language model trained on 250 million protein sequences[119], ESM1b can predict variant pathogenicity by comparing the likelihood of observing an alternate amino acid with the likelihood of the actual amino acid in a sequence given its context and assigning higher pathogenicity to amino acids with lower likelihood of being observed[117].

More commonly seen in VEPs is the integration of both evolution and protein structure features in predicting variant effects. An early implementation of this approach is PolyPhen2[120], which combines structural features such as surface area changes and crystallographic beta factor with evolution metrics obtained from MSAs. More recent deep learning methods circumvent the problem of crystal protein availability by training models to predict structural features or folding from nucleotide or amino acid sequences. PrimateAI[121] is a deep learning model that uses an amino acid sequence flanking a variant and

orthologous sequence alignments in other species to extract evolutionary features as well as predict the protein's secondary structure and solvent accessibility, before incorporating all these elements into variant pathogenicity prediction. AlphaMissense[109] leverages the protein-folding prediction tool AlphaFold2[122] to predict variant effects. The architecture of AlphaMissense involves pre-training an AlphaFold2 model to predict the folded structure of an amino acid sequence, then fine-tuning the trained models on weakly-labeled "pathogenic" (common variants found in human and primates) and "benign" (simulated variants not seen in human and primates) variants to predict variant pathogenicity[109].

A final class of VEPs, empowered by machine learning models that can learn from a large number of features and leveraging the increasing numbers of VEPs, is ensemble methods, which incorporate predictions from other methods to predict variant pathogenicity. Examples of methods in this class include REVEL[123], CADD[124,125] and VARITY[126]. REVEL uses as input features predictions from 13 different VEPs and conservation metrics, including evolution-based methods, e.g. SIFT[108] and MutationAssessor[107], conservation scores, e.g. phastCons[127], as well as two homology-and-structure-based methods, PolyPhen2[120] and MutPred[128]. Similarly, CADD also incorporates predictions of the aforementioned methods, along with other functional genomics features such as DNase hypersensitivity, transcription factor binding, splicing effect[124] and most recently regulatory sequence effect[125]. VARITY[126] only includes predictions from unsupervised VEPs that do not rely on human-labeled training labels, e.g. SIFT,

in order to avoid data circularity (further discussed below), as well as protein-protein interaction and protein structure information, e.g. accessible surface area.

Besides differences in the types of included features, VEPs also differ on the approach to predict pathogenicity from their selected features. There are two main approaches - statistical estimations or machine learning/deep learning models. Some VEPs, including SIFT[108] and MutationAssessor[107], directly estimate the pathogenicity scores for variants using statistics and probability estimations disciplined by specific features. These VEPs do not require training, and their performances are benchmarked on datasets of annotated variants such as ClinVar and dbSNP[129].

Other VEPs rely on machine learning and deep learning models, including naïve Bayesian classifiers, logistic regression, support vector machines, random forest and deep neural network, to classify variants. These methods are mostly supervised and are commonly trained using annotated variants from databases such as ClinVar[98], the Human Gene Mutation Database (HGMD)[130], UniprotKB[120] and VariBench[131]. PolyPhen2 uses a naïve Bayesian classifier, CADD implements both a support vector machine and logistic regression in different versions, REVEL uses random forest, while AlphaMissense, PrimateAI, ESM1b and EVE are based on deep neural networks.

Training on human-curated labels results in inherent bias, as explored below. To overcome these biases, CADD, PrimateAI and AlphaMissense leverage large population germline data to create proxy labels for pathogenicity.

Specifically, variants frequently observed in the human and primate populations (minor allele frequency > 1e-5) are designated as benign, and simulated variants absent from human and primate populations are labeled as pathogenic, based on the assumption that pathogenic variants are selected against in populations and therefore are not commonly observed. The resulting dataset is noisy, but AlphaMissense introduces multiple rounds of fine-tuning in which the training dataset is updated with high-confidence predictions from previous training rounds ("self-distillation"), thus gradually improving the specificity of the training data[109]. The proxy label approach not only circumvents the biases associated with human-labeled labels but also creates a much larger training and evaluation set by orders of magnitude compared to ClinVar and others. For example, AlphaMissense was trained on 2.4 millions proxy-labeled variants, 100 times larger than the set of ~200,000 somatic variants in ClinVar[109].

There are a few unsupervised machine learning VEPs that do not rely on labeled training and test sets, such as EVE[113] and ESM1b[117]. Since both are generative neural networks, they can predict pathogenicity directly as the difference in predicted log-likelihood between the reference and alternate amino acid at a given position.

Challenges with computational VEPs

VEPs have been shown to exhibit varying performance in systematic comparisons that benchmark them on the same sets of clinical labels or functional data[94,95,110]. Performances vary depending on the data they are being evaluated on, and there is no consensus of the best performing method across

the board[94,95,110]. This is likely a result of the different choices regarding input features, model architectures, as well as training and testing data resulting in different strengths and weaknesses of each method. Some of these design choices result in significant caveats that should be carefully considered in interpreting predictions from these methods.

First, the type of input features can restrict method performance. Homology-based methods do not take into account structural consequences of variants, and rely heavily on good MSAs to infer variant effects. Therefore, these methods tend to perform worse in regions with low MSA coverage[117]. Meanwhile, methods that incorporate structural information, including PolyPhen2 and VARITY, are limited by the availability of 3D structures.

Second, the type of data used to train, test and validate machine learning-based methods can also result in multiple types of circularity and bias that affects the methods' performance on new variants[132]. The first type of circularity comes from data leakage, coming from the same data, either the same variant or different variants at the same residues, being used in the training and evaluation of a method, leading to over-optimistic performance metrics that do not generalize to new variants. This is an issue with ensemble methods such as REVEL and VARITY, as they cannot adequately ensure that all variants used in the training of the 13 individual predictors are excluded from its evaluation set[95,123]. Indeed, when possible sources of leakage are removed from the training set, the performance of VARITY[126] as measured by the area under the receiver operating curve (AUROC) reduced from 0.949 to 0.935[109]. The second type of

circularity is label bias, coming from the overrepresentation of variants in a few genes (50% of labels in ClinVar are found in only 7% of the genes[113]), as well as the predominantly pathogenic or non-pathogenic variants in certain genes within curated training datasets (e.g. most variants in *TP53* are annotated pathogenic). This imbalanced representation results in models overfitting on labels from a few genes and learning to predict all variants within certain genes as pathogenic. A simple model to predict variant pathogenicity in a gene using only the ratio of pathogenic variants of that same gene as a predictor achieves an AUROC of 0.914 on a test set comprising 40% ClinVar curated variants[109]. These models can demonstrate excellent, albeit conflated, performance on the evaluation set, even though this does not translate into generalized ability to predict pathogenic variants in less commonly mutated genes. Finally, the predictions of many commonly-used VEPs such as SIFT and PolyPhen2 are used as a criteria to curate variant effects in ClinVar; therefore, using ClinVar labels to train and validate a model that incorporates SIFT and PolyPhen2 scores as predictive features, as many ensemble methods do, introduces additional circularity. All of these biases caution against using human-annotated labels as training data for VEPs, as it leads to inherited biases from human curators and previous predictors.

The development of unsupervised VEPs, such as EVE and ESM1b, and methods trained on proxy-label data, such as AlphaMissense, PrimateAI and CADD, can alleviate the bias associated with human-curated training data. Furthermore, the advances in deep mutational scanning or multiplexed assays of

variant effect (MAVE)[93], which can measure molecular and cellular phenotypes across thousands of variants simultaneously, have resulted in large functional databases that can be used to train, fine-tune and benchmark VEPs without the concern about correlation with clinical labels[110,133]. VARITY[126] incorporates MAVE variants into their training data, while AlphaMissense[109], EVE[113] and ESM1b[117] assess their performance using MAVE datasets besides curated clinical datasets, demonstrating comparable metrics to evaluations based on ClinVar labels. Therefore, newer VEPs have great potential to provide more accurate and unbiased predictions for variant effects that can contribute to more effective discovery of pathogenic drivers of diseases, especially in cancer.

In summary, large-scale cancer genomics projects have significantly advanced our understanding of cancer genetics. The identification of driver mutations has become crucial for assessing clinical phenotypes, prognosis, and treatment responses. However, the majority of mutations in cancer are often classified as VUSs, with emerging evidence suggesting nuanced roles for even seemingly neutral mutations. VEPs play a vital role in systematically characterizing variants, although they exhibit varying performance and face challenges such as input feature restrictions and biases associated with human-curated labels. Databases like ClinVar and oncology-specific knowledge bases like OncoKB consolidate variant information but may have limitations in number of curated variants and reporting bias. Recent advancements in unsupervised VEPs and methods trained on proxy-label data show promise in overcoming biases and enhancing the discovery of pathogenic drivers, which can

potentially contribute to a more effective understanding of genetic variation and its implications for human health, particularly in the context of cancer.

1.3. Thesis objectives

In this thesis, we aimed to develop methods to address the two aforementioned challenges in utilizing clinicogenomic RWD in oncology research, specifically focusing on issues with unstructured clinical information and genomic variant interpretation. Antineoplastic therapy is an important variable in cancer studies, yet information about whether patients receive prior treatment is only available in free-text clinical notes, requiring manual curation. To address this bottleneck, we developed and validated an NLP pipeline on 48,447 patients at Memorial Sloan Kettering Cancer Center. This pipeline identified 25% of patients with putative prior treatments, linking them to resistance mutations and worse overall survival (Chapter 2). While computational methods for variant effect prediction have advanced, their applicability to identifying cancer drivers remains underexplored. Our assessment of VEPs in annotating variants in real-world cancer genomic data revealed that methods incorporating protein structure or functional genomic information outperformed those relying solely on evolutionary data. Validation in two real-world NSCLC cohorts (N=8,690 and 977) underscored the biological validity of VEP predictions, emphasizing their potential in studying patient outcomes and biology in cancer (Chapter 3). Overall, our work demonstrates the efficacy of machine learning models in streamlining data annotations for more efficient use of RWD in oncology research.

CHAPTER 2

AUTOMATED EXTRACTION OF EXTERNAL TREATMENT STATUS USING NATURAL LANGUAGE PROCESSING

Summary

Anticancer therapy changes tumor physiology and genomics, making it a key variable in cancer studies. Although antineoplastics given at a single institution may be available in research-ready format, treatment at external institutions prior to receiving care at academic medical centers, common among patients at these centers, is often only described in free-text clinical notes, necessitating manual curation for downstream analysis. To overcome this bottleneck, we trained and validated natural language processing (NLP) models using initial consult notes to identify whether patients had received treatment at external institutions and studied the impact of these putative treatments on tumor genomics. We deployed the model on 48,447 patients at Memorial Sloan Kettering Cancer Center (MSK) and identified 25% of patients with putative prior treatments, linking to resistance mutations and worse overall survival. These results demonstrate that NLP can abstract external treatment status from clinical notes. When applied at scale, our model could help mitigate confounding variables and identify relationships between clinicogenomic variables and anticancer therapy.

Results

Treatment history is an important variable in clinical and genomics studies, as cancer treatment influences patient outcomes and provides the evolutionary pressure that can lead to resistance through diverse mechanisms, such as acquired resistance mutations and histology transformation[134–137]. Patients seen at tertiary referral cancer centers such as MSK are often referred from their local oncology centers in advanced disease stages, after having received multiple lines of antineoplastics[138,139]. Information about treatment received externally is often only recorded in free-text clinical notes during patients' first visit and not easily accessible for downstream research use, which could lead to analyses suffering from the confounding effect of treatment status. To overcome the need for manual curation in making this variable available in structured format, we sought to automatically abstract the status of whether patients had received treatment outside of MSK from their initial consult notes.

Model performance

We trained and evaluated the performance of two NLP models, a logistic regression (LR) model and a Transformer-based Clinical-Longformer ("Longformer")[58], to predict the probability that patients received treatment prior to their first visit at MSK from their initial consult (IC) notes. Both models were trained and evaluated on 3,570 patients in five GENIE-BPC cohorts, who were contributed by MSK and therefore had clinical notes available in MSK's EHR system. For each patient, we selected one IC note from their first visit with a medical or radiation oncologist at MSK as input to the model (see **Methods** for

more details on note selection). The Longformer took sequences up to 4,096 tokens in length as input, while the LR model required pre-processing of IC notes into binary feature matrices comprising 10,000 common uni- and bigrams, along with an indicator of whether a medication name or adjacent terms such as 'chemotherapy' was found in the note **(Figure 1)**.

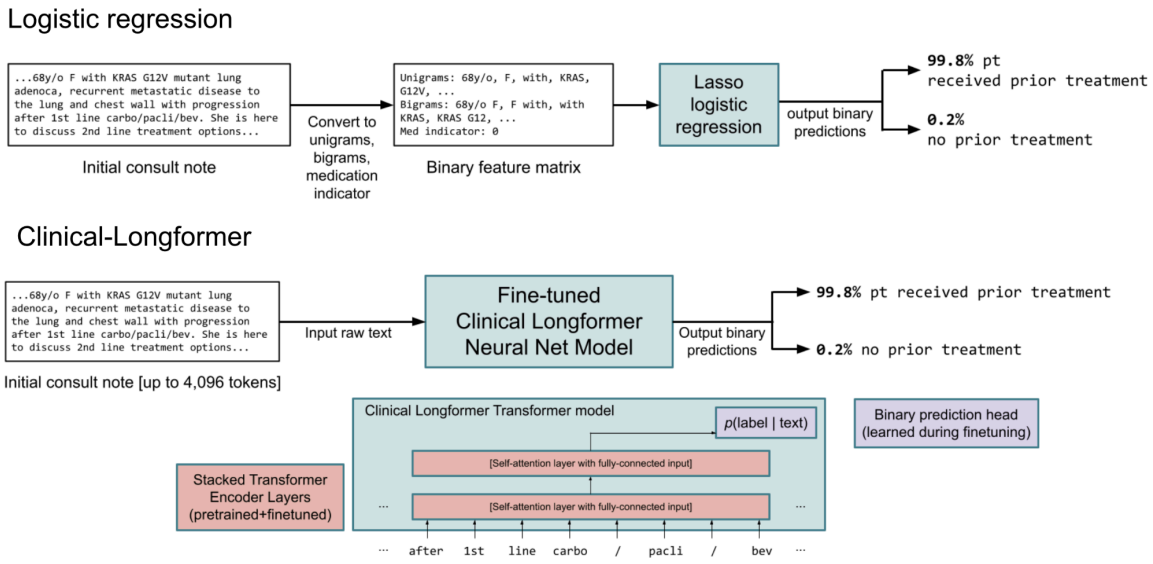

Figure 1. Architecture of NLP models trained to infer external treatment status.
Two models were trained using IC notes. The lasso logistic regression model takes a binary feature matrix, constructed from common unigrams and bigrams in IC notes, as input, while the Clinical-Longformer model takes raw IC notes up to 4,096 tokens in length as input. Both models output the probability of having received external treatment.

We used cross validation (CV) to evaluate model performance in two settings: 5-fold CV with random data splits, and cancer-holdout CV, where patients from one cancer type were held out for evaluation and the models were trained on patients from the other four cancer types. Performance metrics were macro-averaged from all five folds in both settings.

In evaluating the LR model, we observed that its performance varied across evaluation scenarios. In 5-fold CV, LR achieved a precision of 0.643, recall of 0.657 and an AUROC of 0.719 **(Table 1)**. When specifically assessing its ability to generalize to a new cancer type when trained on others in the cancer-holdout CV, LR showed improvement across all metrics with a precision of 0.822, recall of 0.768, and an AUROC of 0.857 **(Table 1)**. Comparison of two AUCs between the CV settings using DeLong's test for significant changes in AUROC[140] showed that the performance of the LR model in the cancer-holdout setting was significantly better compared to that of the model in the CV setting (p-value = 0.05).

The Longformer model, on the other hand, performed well in both validation settings. In the 5-fold CV, the Longformer achieves a high AUROC of 0.977 and performance metrics **(Table 1)**. In the cancer-holdout evaluation, the Longformer's performance dropped slightly across all metrics, with an AUROC of 0.965; however, this decrease in AUROC was not significant in a DeLong's test (p-value = 0.5). These results underscore the Longformer's effectiveness in providing accurate predictions and its ability to generalize to new cancer types not present in the training set.

We compared the performance of the LR and Longformer using DeLong's test. The AUROCs achieved by the Longformer model in both CV settings were significantly higher than those of the LR models (p-value < 0.01, **Figure 2**). Considering both models together, the comparison suggests a clear advantage for the Longformer across all metrics and evaluation scenarios. Its consistently

higher precision, recall, F1 score, and AUROC emphasize its robust performance and potential for accurate predictions.

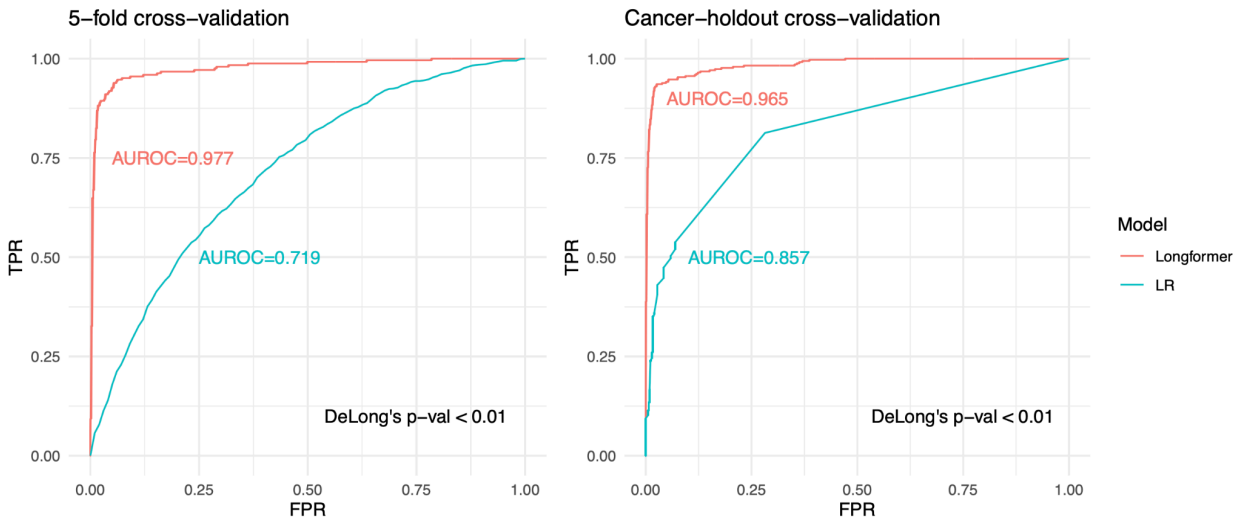

Figure 2. ROCs of Clinical-Longformer and LR models in two CV settings
Significant differences between models' AUROCs were calculated using DeLong's test.

Table 1. Macro-averaged performance metrics for NLP models for external status abstraction

Metrics	**LR 5-fold CV**	**LR - Cancer CV**	**Longformer - 5-fold CV**	**Longformer - Cancer CV**
Precision	0.643	0.822	0.907	0.889
Recall	0.657	0.768	0.909	0.892
F1 score	0.642	0.764	0.908	0.884
AUROC	0.719	0.857	0.977	0.965

To better understand the errors of the Longformer model, we collaborated with a medical oncologist to manually annotate high-confidence false positive (FP, predicted probability of receiving external treatment > 0.8) and false negative (FN, predicted probability of receiving external treatment < 0.2) cases from the 5-fold cross-validated model. A total of 42 cases were manually reviewed,

accounting for 40% of all discrepant cases. The review involved a medical oncologist looking at the same IC notes used to train the model and, without knowing the model's predictions, determining whether the patients had received prior treatment. Out of 21 FP cases, 17 (80%) had received prior antineoplastic treatment according to the medical oncologist's interpretation of the IC notes, which agreed with the model's predictions **(Figure 3)**. The most common reason for the medical oncologist's agreement with the model's predictions instead of GENIE BPC's labels was due to the identification of treatment information in IC notes that had been overlooked during the curation process. This result suggests that there may be curation errors in the GENIE BPC treatment data and highlights how automated feature abstraction with NLP can supplement the manual curation process of clinical variables.

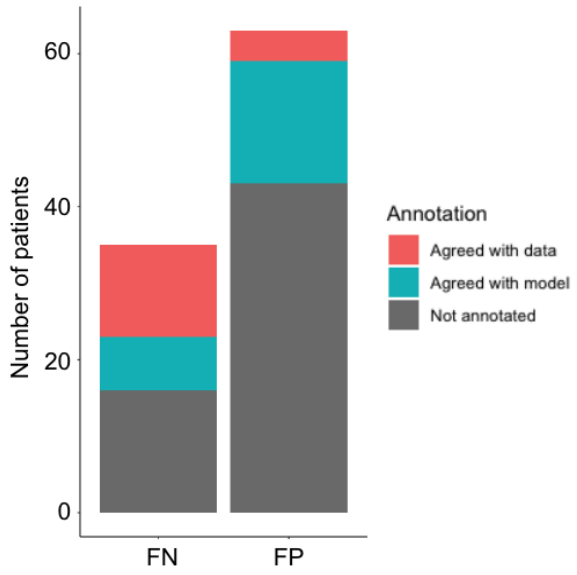

Figure 3. Manual review of discrepant cases from Longformer model
IC notes from 42 high-confidence FN and FP cases were reviewed by a medical oncologist, and external treatment annotation from manual review was compared with the Longformer's predictions as well as GENIE BPC's labels.

External treatment status in MSK-IMPACT cohorts

We deployed the fine-tuned 5-fold cross-validated Longformer model to infer external treatment status for 48,447 patients at MSK with IC notes. We defined externally treated patients as those with predicted probability of receiving external treatment larger than or equal to 0.6 and with confidence intervals not overlapping with 0.5. This cutoff was chosen by looking at the distribution of predicted positive probability **(Figure 4)** and opting for a more lenient threshold to boost our recall rate.

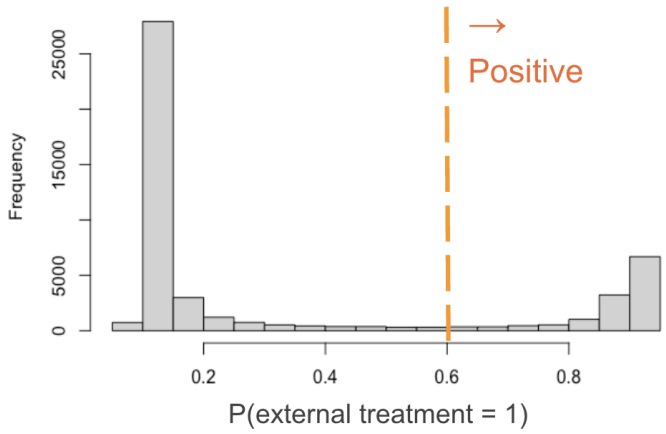

Figure 4. Histogram of the predicted probability of receiving external treatment for 48,447 patients at MSK
Dotted orange line indicates the threshold to identify cases predicted to have received external treatment.

Using this cutoff, about 25% of the MSK-IMPACT cohort (11,900 patients) was predicted to receive antineoplastics before coming to MSK. The proportion of putative externally treated patients varied by cancer types, ranging from ~70% in tumors in the peripheral nervous system to 5% in renal cell carcinoma **(Figure 5)**. Among the most common cancer types with >= 2,000 patients, the proportion of externally treated patients also varied, even though with much less variance,

with patients with breast cancers having the highest rate of putative external treatment of 30% and NSCLC having the lowest rate with 15%. The large difference in putative external treatment status between cancer types could be due to multiple reasons. First, patients with different cancer types may seek care at MSK at different stages, resulting in variation in their treatment history. For cancer types that were more aggressive, commonly diagnosed in advanced stages or without any available first-line antineoplastics, treatment-naïve patients may be referred to MSK right away, resulting in lower fractions of externally treated patients, while cancer types that were diagnosed earlier or with well-established treatment protocols may be managed at local oncology centers first before referral to MSK, thus resulting in higher fractions of externally treated patients. Another possible contribution to this large variation in external treatment rate could be the types of IC notes available for patients in each cancer type. Various types of IC notes differed in the amount of information they provided about previous treatment (see **Methods** for more details on note type prioritization). Patients with different types of cancer may be more likely to have their initial visits with healthcare providers who gave relatively less emphasis to their treatment history, leading to less informative corpora of notes for the model to base its prediction on. Finally, cancer type-specific model performance may account for some of the observed large variation in external treatment rate. Even though the Longformer performed well when evaluated on a held-out cancer type not seen during training, it was only trained and evaluated on five common cancer types and thus may not generalize well on rarer cancer types.

We further annotated the treatment status of first MSK-IMPACT samples by combining NLP-derived external treatment labels and prior treatment labels, derived by aligning the sample acquisition dates of first MSK-IMPACT samples with treatment records in MSK's institutional treatment database to identify whether a sample was treatment-naïve or pre-treated. Comparing these two labels allowed us to identify two groups of samples: treatment-naïve samples, with no treatment before sample acquisition in MSK database and negative external treatment prediction (MSK treatment (–) NLP (–)) and pre-treated samples, with either recorded treatment before sample acquisition in MSK database (MSK treatment (+)), positive external treatment prediction (NLP (+)), or both (MSK treatment (+) NLP (+)). Within the pre-treated samples, there existed samples with no recorded treatment before sample acquisition in the MSK database, yet came from patients predicted to have received external treatment. These samples, which made up about 30% of all samples with no known treatment before sample acquisition, would have been labeled "treatment-naïve" without additional information about external treatment **(Figure 5)**. Since patients may have received treatment at MSK before tumor biopsy, the samples with MSK treatment (+) NLP (–) label did not necessarily reflect discrepancy in model performance as these MSK treatments would not have been recorded in IC notes.

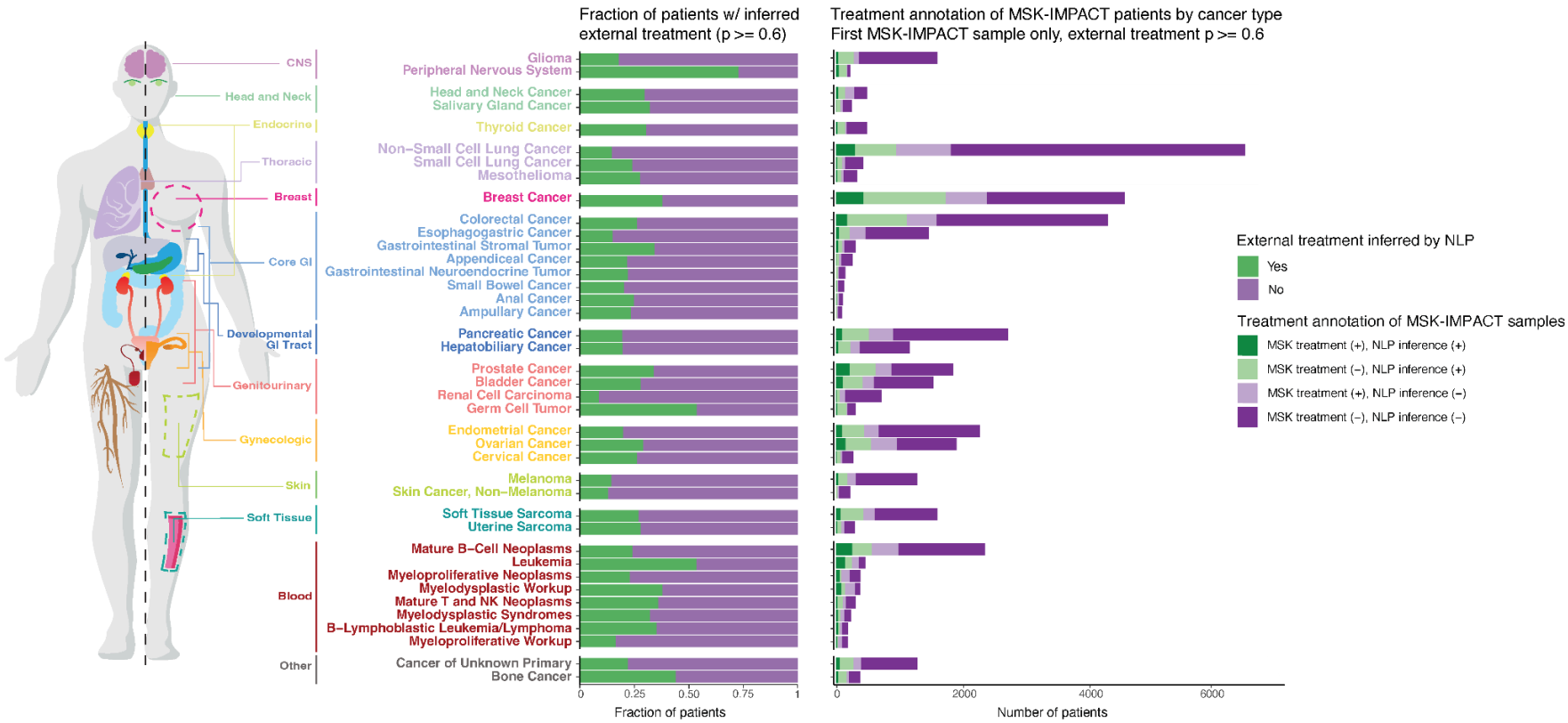

Figure 5. Inferred external treatment status for 48,447 MSK patients

Prior treatment probabilities for MSK patients were predicted using the fine-tuned Longformer model and patients with predicted probabilities of >= 0.6 were labeled as putatively externally treated. **Left:** Proportion of patients predicted to receive prior external treatment. **Right:** Known and prior treatment annotation of first MSK-IMPACT samples for patients with external treatment inference.

Clinicogenomic characteristics of externally treated patients

Antineoplastic treatment alters tumor genomics and physiology. Genomic sequencing of paired longitudinal samples before and after treatment have revealed multiple acquired genomic changes, some of which are responsible for drug resistance[134–136,141]. We therefore hypothesized that samples with putative external treatment would have similar genomic profiles to samples with known treatment at MSK, with both exhibiting more treatment-related genomic changes compared to treatment-naïve samples.

To test this hypothesis, we compared the frequencies and statistical enrichment of oncogenic alterations, including point mutations, copy number alterations and fusions, at the gene and residue levels in the first MSK-IMPACT sample of each patient using logistic regressions, adjusting for sample type. We found significant enrichment of alterations in resistance-related genes in both the inferred externally treated and pre-treated groups, including *AR* and *TP53* in prostate[141,142] and *ESR1* in breast cancer[134] **(Figure 6)**. Additionally, pre-treated samples, both known and putative, from patients with breast cancer showed enrichment in multiple genes, mostly due to amplification, aligning with previous observations about genomic changes following endocrine therapy[134]. Similarly, for *EGFR*-mutated patients with non-small cell lung cancer, *EGFR* T790M, a well-known resistance mutation, was enriched in both the inferred externally treated and pre-treated groups compared to treatment-naive[143] **(Figure 6)**. The alteration frequencies of resistance-related genes were highest in samples with MSK treatment, followed by samples with putative prior treatment, whereas

treatment-naïve samples had low alteration frequencies of these genes and mutations. The low frequency of resistance mutations e.g. *EGFR* T790M in treatment-naïve samples aligns with our expectation that this mutation was primarily acquired after treatment and rarely detected at baseline[144,145]. Furthermore, the intermediate alteration frequencies of resistance genes and mutations in the putative prior treatment samples suggests a treatment dosage effect, in which the amount of treatment patients have undergone correlates with the likelihood of acquiring genetic resistance. Together, these results suggest that externally treated patients defined a patient cohort with distinct genomic characteristics from treatment-naïve patients, and thus should be analyzed separately in downstream genomic analysis to reduce confounding effect.

Finally, we looked at overall survival (OS) of patients and whether it is associated with treatment status. Patients with putative external treatment have shorter OS compared to treatment-naive patients and in most cases have comparable OS compared to known pre-treated patients, likely reflecting the more advanced stages of disease for pre-treated patients. These differences in survival, along with the enrichment in resistance mutations, demonstrate that patients with putative external treatment should be analyzed as a separate group or as part of the pre-treated group **(Figure 7)**.

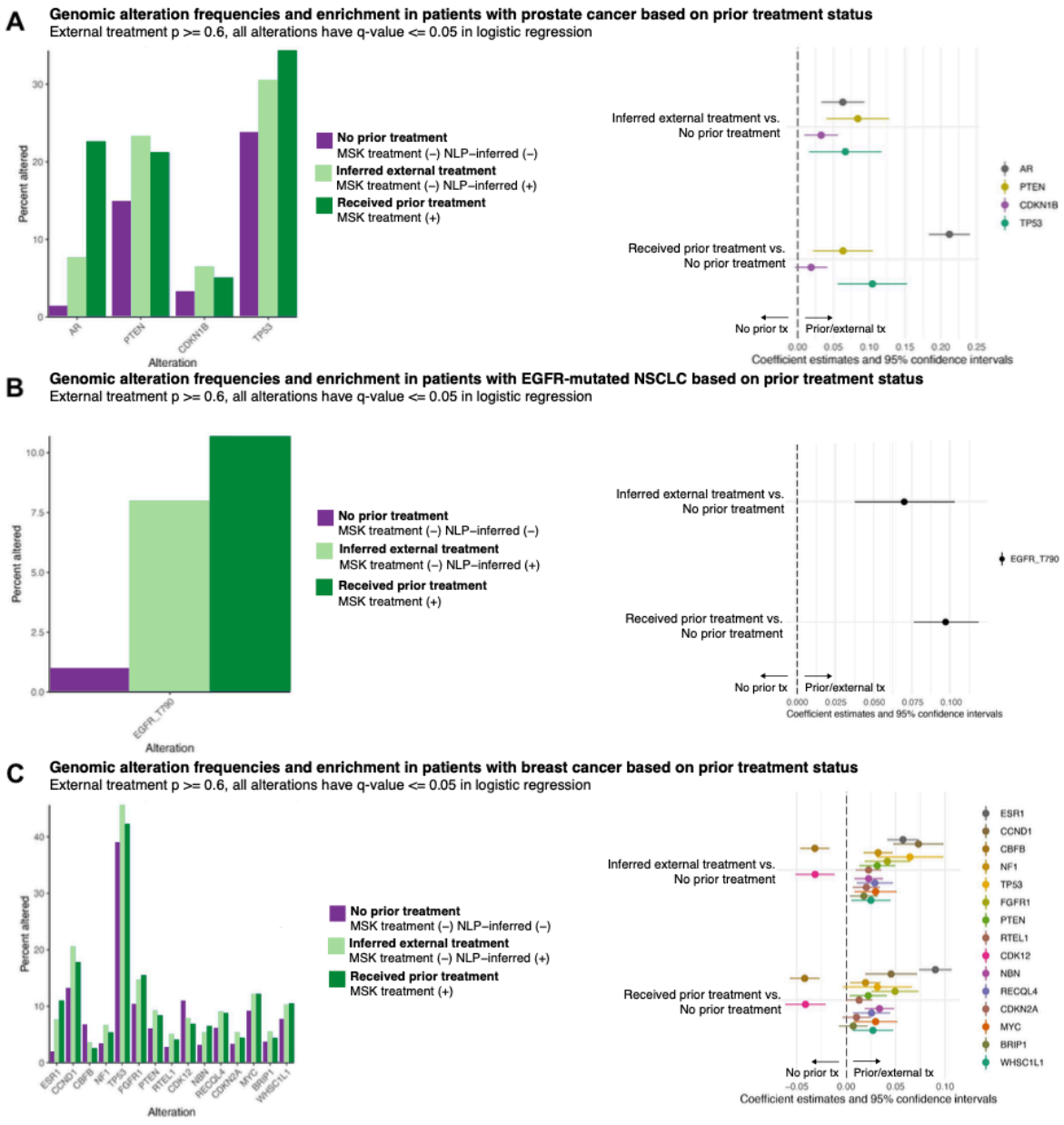

Figure 6. Enrichment of resistance-related genes and alterations in putative externally treated MSK patients

Frequencies (left) and logistic regression coefficients showing enrichment (right) of genomic alterations in patients with known prior or putative treatment status compared to treatment naïve patients with **A.** prostate cancer, **B.** *EGFR*-mutated NSCLC and **C.** breast cancer.

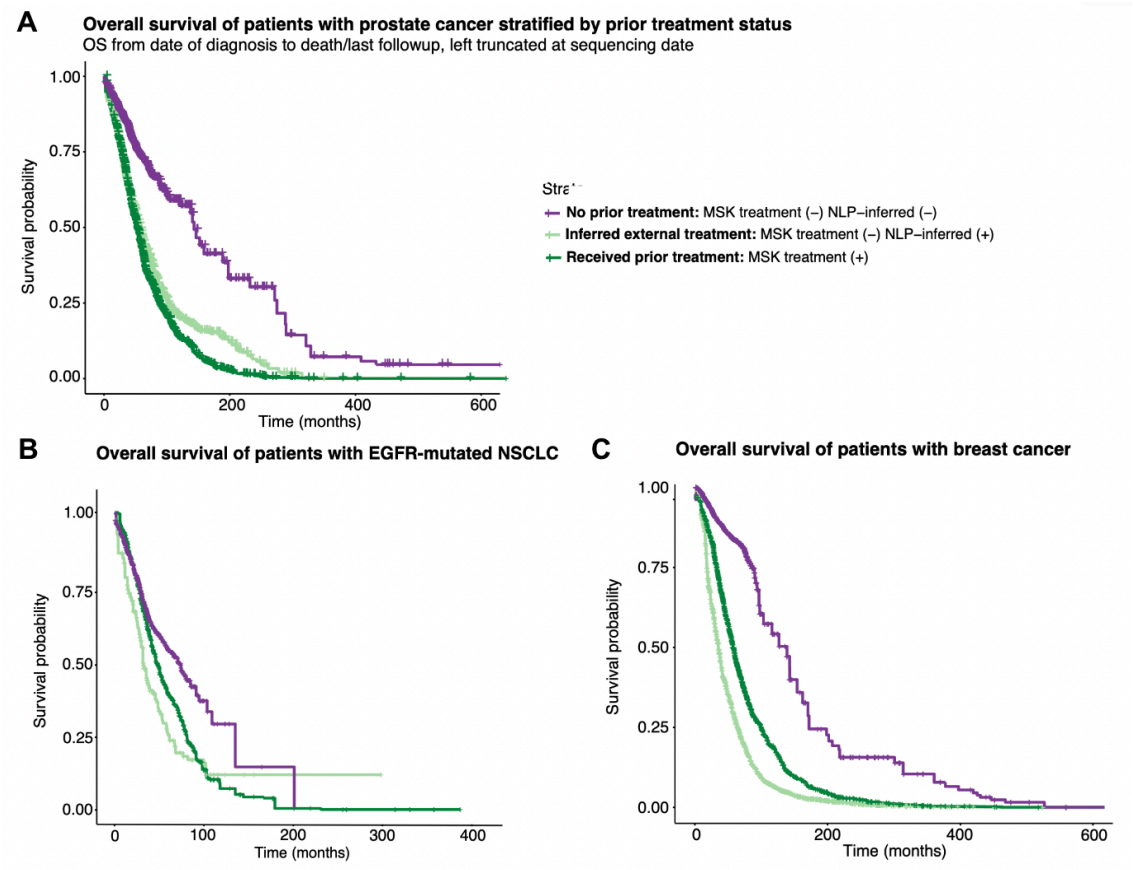

Figure 7. Overall survival of putative externally treated MSK patients
Kaplan-Meier (KM) curves showing OS from date of diagnosis stratified by treatment status of MSK-IMPACT patients with **A.** prostate cancer, **B.** *EGFR*-mutated NSCLC and **C.** breast cancer.

Discussions

In this work, we trained and evaluated two models with different architectural complexities, a ngram-and-medication-token-based LR model and a Clinical-Longformer model, on IC notes to predict whether patients had received external treatment before coming to MSK. In certain NLP tasks, such as identifying smoking status or disease state[146], where inference relies on identifying keywords, regular expressions matching may be sufficient. This approach does not require training or fine-tuning, and has the added advantage

of being interpretable. Even though the presence of drug names can serve as a proxy for prior treatment status, clinicians often refer to medications by their abbreviations, brand names or regimen names in clinical notes. These non-standard nomenclature, along with misspellings, render rule-based approaches difficult and thus require more sophisticated models. In addition to handling cases without standard vocabularies effectively, machine learning models also demand less domain knowledge for implementation compared to rule-based approaches, which enables these models to be scaled more easily to new cancer types.

We tested both the LR and Longformer's ability to generalize to new cancer types using two CV schemes, one involving testing on a held-out cancer type and another using random splits. The LR model performed better in the cancer-holdout CV compared to random split 5-fold CV (AUROC of 0.857 compared to 0.719, p-value < 0.05). This difference may be due to the similarities in features which were predictive of receiving prior treatment between cancer types, allowing the cancer-holdout model to generalize more efficiently to a new cancer type. Furthermore, the random split CV might fail to capture important interactions between variables across different cancer types. The cancer holdout model, by focusing on specific cancer types during training, may better control for confounding factors and capture interactions that are crucial for accurate predictions in the held-out cancer type.

On the other hand, the Longformer consistently performed well in both validation strategies, achieving AUROCs between 0.965-0.977. The stability in

the Longformer's performance in different CV strategies compared to the LR's highlights the strengths of Transformer-based models in handling complex language tasks. Specifically, the Longformer's ability to handle long input sequences, up to 4,096 subtokens, allowed it to process the complex contexts related to treatment in IC notes. Since the Longformer does not require pre-processing, it is also not sensitive to the feature engineering choices that need to be made before training a LR model. Should a better alternative model such as the Longformer not be available, other feature engineering strategies exist that can potentially improve the LR's performance, such as selecting n-grams with high tf-idf[147], which weighted words by their importance in a document, or using terms from word2vec-augmented vocabularies exclusively.

The selection of notes during training and prediction is a crucial factor in ensuring the success of the models. In particular, IC notes encompassed various sections covering extensive details such as medical history, family history, present illness, and future treatment plans. In the earlier training phases of both models, when entire notes were utilized, it became evident that a significant number of false positives originated from references to treatment plans within the Future Treatment section. This observation underscores the importance of refining note selection strategies to enhance the models' accuracy and reliability in distinguishing relevant information.

Even though real-world cohorts of patients with fully curated clinical data, such as GENIE BPC, remain small, they serve as invaluable resources to train and validate NLP models. In turn, NLP efforts can help provide assistance and

feedback for the curation process. Manual review of high-confidence FP cases from our NLP model revealed a non-trivial number of curation errors, suggesting that NLP models can learn and predict patterns in the identification and classification of clinical data. This collaborative approach, combining the efficiency of NLP models with the expertise of human curators, enhances the overall accuracy and reliability of curated clinical datasets. Additionally, the iterative feedback loop between NLP models and manual curation enables continuous refinement and improvement in the performance of the models over time. Furthermore, the scalability of NLP allows for the analysis of vast amounts of unstructured clinical text, contributing to the enrichment of curated datasets with a diverse range of patient cases. This approach not only facilitates a more comprehensive understanding of disease patterns and treatment outcomes but also opens avenues for exploring novel insights that may have been overlooked in smaller, manually curated cohorts.

This study focused on predicting a binary indicator of prior treatment, which can be sufficient to build cohorts of patients based on treatment status or control for treatment in outcome analyses. Case-by-case annotations of the prior treatment type that patients received have shown that patients generally received the standard of care for their cancer type. For example, patients with *EGFR*-mutated NSCLC received first-line EGFR tyrosine kinase inhibitors for their prior treatment, as evidenced by the enrichment of *EGFR* resistance mutations in putatively pre-treated samples **(Figure 6)**. However, as the treatment landscape is changing, patients will have received more

heterogeneous therapies, thus necessitating further manual annotations for use cases that involve knowing the exact prior treatment type. NLP models could be trained to identify not only whether a patient has received treatment, but also the type of treatment they received and the duration of treatment. Multiple models have shown promising performance in contextualized medication event extraction from longitudinal clinical notes during the 2022 National NLP Clinical Challenges[148]. With the appropriate training data, these models could be fine-tuned and deployed for medication name and duration abstraction from IC notes, which would further enhance the utility of our real-world clinical data.

This work has limitations. The training cohort, though well-curated, was small and only included patients from five cancer types, which could limit the model's ability to learn more heterogeneous patterns in other rare cancer types. Additionally, the lack of abstraction for precise treatment information or associated cancer types posed a challenge in distinguishing prior treatments for different primary cancers in cases where a patient has multiple primaries. To address this issue for patients with multiple primaries in our cohort, we opted to utilize IC notes from a narrow timeframe (within 90 days) during their first visits with a specific disease management team at MSK. However, these notes may not be the most informative regarding external treatment. More diverse training data and additional labels for the cancer types associated with prior treatment could allow the model to be applied on more IC note types and generalize better to more cancer types.

Methods

Patients & Data

This study analyzed patients with tumor genomic sequencing from two sources: The MSK-IMPACT cohort and the AACR Project GENIE Biopharma Collaborative Cohort, which includes patients from the MSK-IMPACT cohort. Details regarding the AACR GENIE Biopharma Collaborative (BPC) have been published previously[37]. Here we included patients in the BPC with single-primary NSCLC, breast, colorectal, prostate, or pancreatic cancer. The MSK-IMPACT cohort comprises patients at Memorial Sloan Kettering (New York, NY), an academic cancer hospital with tumor genomic sequencing using MSK-IMPACT, an FDA-authorized tumor genomic profiling assay, which uses matched white blood cell sequencing to filter clonal hematopoietic and germline variants. All MSK patients were enrolled as part of a prospective sequencing protocol (NCT01775072). The study was independently approved by the Institutional Review Board of each site. Patients provided written, informed consent and were enrolled in a continuous, nonrandom fashion. Data here is from a September 22, 2022 snapshot. Only patients with completed tumor registry entries were included in analysis.

Our training cohort included 3,570 patients across five GENIE BPC cohorts, including breast (BRCA, N = 463), colorectal (CRC, N = 885), lung (NSCLC, N = 1293), pancreatic (PANC, N = 500) and prostate (PRAD, N = 429). GENIE BPC patients were included in the training data if they were contributed by MSK and thus had clinical notes available in MSK's EHR system.

We deployed our model on a cohort of 48,447 MSK-IMPACT patients with records of their first tumors in MSK's system. We did not exclude any patients based on visit date, even though further note filtering excluded some MSK-IMPACT patients from the final deployment cohort.

Note selection and prioritization

We developed our model based on IC notes, which included notes from medical oncologists, radiation oncologists, surgery, inpatient services and others, as they were most likely to contain information about external treatments prior to MSK. IC notes, defined as all notes with prefix "initial consult" in ClinDoc, were obtained for all patients in Tumor Registry. For all IC notes, we included sections relevant to external treatments, including past medical history, history of present illness and chief complaint, while excluding sections mentioning future treatment plans to avoid possible confounding language.

In both our training and testing cohort, a small number (~2%) of patients were excluded due to the lack of initial visit notes **(Figure 8)**. Most of these excluded patients have non-digital IC notes, commonly before 2012 where digitization efforts started at MSK.

We further developed a prioritized scheme for IC note types based on *a priori* knowledge of how likely each note type was to contain information about external treatments based on our understanding of the notes' content. Medical oncology notes from disease management teams were most likely to include information regarding external treatment, while radiation oncology notes were less likely to do so. Both of these note types are included in training and

validating the model ('allowable note types'), while other note types unlikely to mention external treatment, including inpatient IC notes, were excluded. Patients with no IC notes in the allowable note categories (about 5% of the total cohort) were excluded from the training and validation set. Finally, to avoid patients with notes from later procedures wrongly attributed to their initial visit and thus unlikely to contain external treatment information, we excluded patients with IC notes dated more than 90 days after their initial visit date if they had multiple primaries in Tumor Registry. About 10% of the initial training and 15% of the deployment cohort were excluded due to this filtering **(Figure 8)**.

For the final cohort, we selected one note per patient to analyze. If a patient had multiple notes, the final note was selected based on higher priority, then by earlier creation time should all eligible notes have similar priority.

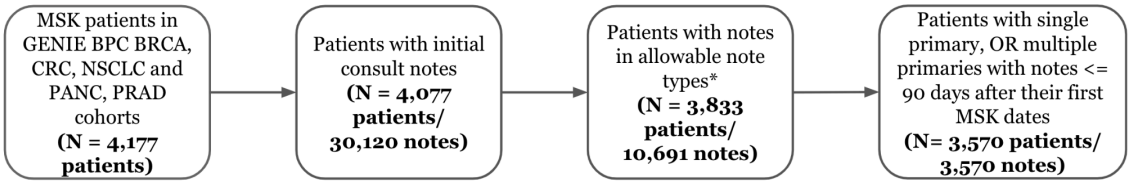

Figure 8. Consort diagram of patient and note selection schemes

Model training and evaluation

Logistic regression model

We trained logistic regression models on 10,000 common unigrams and bigrams from the training corpus and an indicator of whether medication names or adjacent words were mentioned in the text. L1 regularization was applied to all models.

To create a medication name indicator, we first built a vocabulary of medication names by including all drug generic names (column 'DRUG_GENERIC_NAME') curated in RedCap for all five GENIE cohorts. For cancer-holdout CV, we created five vocabularies, each containing medication names from the four cancer types included in the training sets, and used the same vocabulary for training and testing. Since medications can be misspelled, abbreviated, or referred to by their brand names (e.g. Tarceva for erlotinib, 'carbo' for carboplatin) or by the regimens that they are a part of (FOLFOX for leucovorin, 5-FU and oxaliplatin) in clinical notes, we further expanded the generic name medication vocabulary using word2vec[50], which identified words that appear in similar context. We used the list of generic drug names along with terms commonly used to refer to treatment, including 'chemo', 'chemotherapy', and 'rt', as the starting words and augmented our existing drug vocabulary with word2vec. The binary medication indicator was then created for each note by finding matches between the word2vec-augmented medication vocabularies and the IC note corpora.

To evaluate model performance on the GENIE-BPC cohort comprising the training set, we used 5-fold CV. Folds were constructed randomly at the patient level, so that a fold's validation set consisted of reports from patients not present in the train portion of the fold. We used two CV schemes to evaluate model performance. In the 5-fold CV scheme, we splitted the cohort into five folds of equal sizes, then trained models on 80% of the data and tested on the 20% that

is left out in each fold. In the leave-one-cancer-out CV scheme, we trained models on four cancer types and tested on the held-out cancer type.

Clinical-Longformer model

Full IC reports are generally long, and any mention of prior antineoplastic treatments occurs within an expansive textual context. We therefore used a transformer model engineered to have subquadratic memory requirements; in particular, we fine-tuned a Clinical-Longformer model[52], which is itself a Longformer model[53] fine-tuned on the MIMIC-III v1.4 database[51]. This model has a maximum input sequence length of 4,096 subtokens.

The Clinical-Longformer model is implemented in PyTorch;[46] we used the model and pre-trained model weights in the HuggingFace library and model hub.[47] We pooled transformer model output using the default method of conditioning on the first [CLS] pseudo-token prepended to the sequence comprising the impression section. We trained using AdamW[48] using a batch size of 64 and a learning rate of 1×10^{-6}, training for 20 epochs with a warm-up period of 2 epochs. We upsampled minority-class examples uniformly with replacement to achieve class balance during training. Extrinsic results (i.e. main results incorporating model predictions) were presented on models trained on the full GENIE BPC cohort.

Genomic and survival analyses of patients receiving external treatment

Genomic analysis

For each alteration with absolute count higher than 5 or gene with alteration frequency of at least 3% in a cancer type, we used logistic regression

models, regressing alteration status on treatment status and accounting for the sample type at sequencing:

```
alteration status (0/1) ~ treatment status + sample_type
```

The Benjamini-Hochberg procedure was used to adjust the treatment status p-values for multiple hypothesis testing where applicable.

Survival analysis

KM curves were constructed to examine the relationship between treatment status and overall survival for genes. KM curves were calculated from time of diagnosis to time of death or last follow-up, left truncated at time of cohort entry (tissue sequencing). Patients were stratified based on their prior treatment status, either as “no prior treatment” (no treatment record before IC note date in MSK internal database or positive NLP prediction), “inferred prior treatment” (no treatment record before IC note date in MSK internal database but positive NLP prediction), and “received prior treatment” (at least one treatment record before IC note date existed in MSK internal database). Only strata with ≥ 10 patients were included.

CHAPTER 3

MACHINE LEARNING PREDICTIONS IMPROVE IDENTIFICATION OF REAL-WORLD CANCER DRIVER MUTATIONS

Summary

Characterizing and validating which mutations influence development of cancer is challenging. Computational variant effect predictors (VEPs) can predict the effect of variants at scale by leveraging evolutionary, biological and structural features of proteins; however, their utility for identifying cancer drivers is less explored[107–109,120,123,124]. We evaluated 11 computational methods for identifying cancer driver alterations. For identifying known drivers, deep-learning methods incorporating protein structure or functional genomics outperformed methods trained only on evolutionary data. We further validated VUSs annotated as pathogenic by testing their association with overall survival in two cohorts of patients with non-small cell lung cancer (NSCLC, N=8,690 and 977). "Pathogenic" VUSs in *KEAP1* and *SMARCA4* identified by several methods were associated with worse survival, unlike "benign" VUSs. "Pathogenic" VUSs exhibited mutual exclusivity with known oncogenic alterations at the pathway level, further suggesting biological validity. Despite training primarily on germline, rather than somatic, mutation data, computational VEP tools contribute to a more comprehensive understanding of tumor genetics as validated by real-world data.

Results

We sought to use real-world cancer genomic and outcome data to assess the utility of computational methods for annotating VUSs. Specifically, we looked at 11 VEPs, including AlphaMissense[109], CADD[124], ESM1b[117], EVE[113], FATHMM[111], MutationAssessor[107], PolyPhen2, which includes two models trained on different datasets PolyPhen2-HDIV and PolyPhen2-HVAR[120], PrimateAI[121], REVEL[123], SIFT[108] and VARITY, specifically the VARITY_R_LOO model trained on rare ClinVar variants (MAF < 0.5%) and cross-validated using a leave-one-variant-out strategy[126], as well as a database of variants effect, ClinVar[98]. Methods were chosen to be included based on their novelty, demonstrated superior performance compared to other methods in the same class[94,95], relevance to cancer and popularity (see **Methods** for more detail). We evaluated the utility of these VEPs and database (henceforth “methods”) for 1) annotating known pathogenic somatic cancer variants, 2) identifying VUSs associated with OS in patients with cancer and 3) identifying VUSs associated in tumors without other drivers in the same oncogenic pathway. A summary of each method’s approach is presented in **Table 2**.

Table 2. Overview of evaluated VEPs and databases
Eleven variant function prediction methods and two databases were included in this study.

Method	**Description**	**Required human-curated data**
AlphaMissense	Deep neural network based on the protein structure predictor AlphaFold2, trained on weak labels derived from human and primate germline variants	No

CADD	Support vector machine model combining conservation metrics and functional genomic data, trained on weak labels derived from human germline data variants	No
ClinVar	Database of human variants and supporting evidences for their functions	Yes
ESM1b	Unsupervised deep protein language model	No
EVE	Unsupervised deep learning model based on distribution of sequence variation across organisms	No
FATHMM	Variant effect predictor based on sequence conservation within hidden Markov models, weighted for cancer-associated mutations	Yes
MutationAssessor	Variant functional impact predictor based on evolutionary conservation patterns across protein families and subfamilies	No
OncoKB	FDA-approved, manually-curated somatic mutation database	Yes
PolyPhen2-HDIV	Damaging mutation classifier based on sequence, homology and structural features, trained on Mendelian disease variants and closely related homologs	Yes
PolyPhen2-HVAR	Damaging mutation classifier based on sequence, homology and structural features, trained on disease-causing mutations and non-disease-related nsSNPs	Yes
PrimateAI	Deep learning network, incorporating information from two sub-networks that predict secondary structure and solvent accessibility, trained on common human variants and primate variation	No

REVEL	Random forest-based ensemble method trained on other prediction methods and curated variants	Yes
SIFT	Sequence homology-based prediction of amino acid substitution	No
VARITY (VARITY_R_LOO score)	XGBoost model incorporating scores from unsupervised VEPs as well as information about protein-protein interaction and protein structure. VARITY_R_LOO is the VARITY model trained on rare ClinVar variants (MAF < 0.5%) and cross-validated using a leave-one-variant-out strategy	Yes

Spectrum of VUSs across cancer genes

We first characterized the spectrum of VUSs across cancer genes in the large pan-cancer, multi-institutional AACR GENIE v.14-public cohort ("GENIE")[74]. Missense mutations in 183,302 tumors from 160,965 patients were annotated using OncoKB, an FDA-recognized molecular knowledge database[99]. Approximately 80% of somatic missense mutations detected were VUSs **(Table 4)**, even though the percentages of VUSs varied by gene **(Figure 9)**. Almost all mutations in *TP53* and *KRAS* were annotated to be oncogenic, suggesting that these genes are not only extensively studied but also well-conserved and exhibit a high level of intolerance towards protein modifications[149,150]. On the other end of the spectrum were genes with high proportion of VUSs, including those commonly altered by other types of variants e.g. fusions instead of mutations, such as *ALK*[151] and *ABL*[152], as well as those commonly mutated but less well-studied, such as *KEAP1*[153].

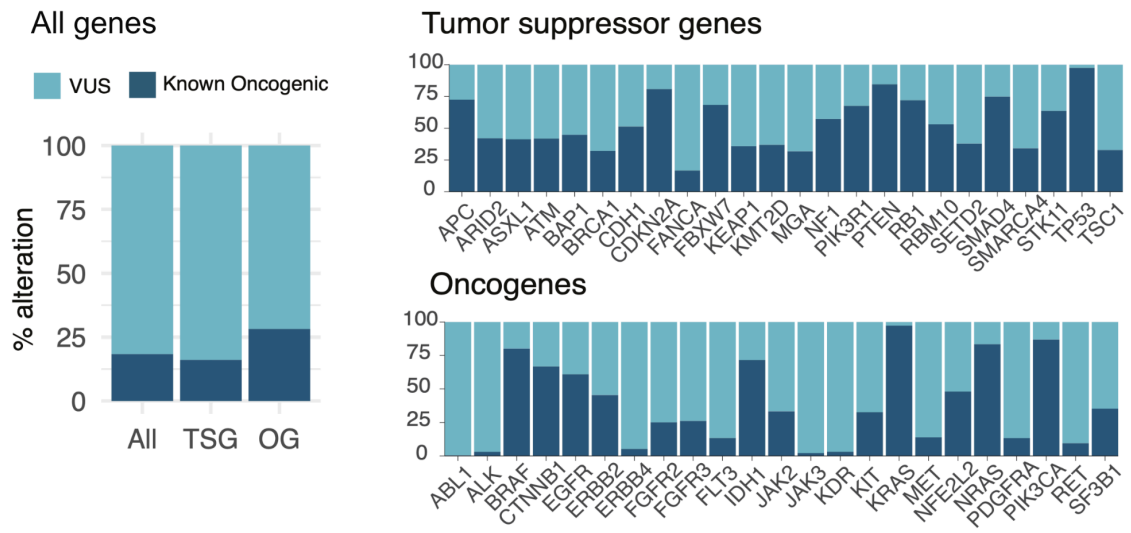

Figure 9. Frequency of known oncogenic mutations and variants of unknown significance in GENIE v14-public cohort as annotated by OncoKB

We then tested the utility of the selected methods in discriminating known pathogenic somatic mutations from benign ones. We only focused on missense mutations as most methods were designed to work on them. To this end, we leveraged missense mutations detected in GENIE and supplemented this data with 10,000 randomly selected missense mutations from the dbSNP Human Variation Sets labeled as having no known medical impact, which served as our negative controls. Prediction scores for all VEPs, representing the likelihood of variants being pathogenic, were obtained from dbNSFP[154]. Since many downstream analyses required discrete labels for variants, we also obtained off-the-shelf mutation classifications, which resulted from imposing predetermined thresholds on the predicted functional scores to stratify mutations into broad categories of pathogenic, non-pathogenic or uncertain, for AlphaMissense, FATHMM, MutationAssessor, PolyPhen2 and SIFT, as well as annotation from the ClinVar database. The rest of the evaluated VEPs required

additional steps to determine the appropriate variant classification from predicted scores. First, EVE provides variant classifications at different uncertainty levels, aiming to enhance accuracy by excluding uncertain ones[113]. For instance, Class25 EVE means 25% of the most uncertain variants were omitted. We determined the optimal uncertainty threshold for accuracy by comparing AUROCs in classifying oncogenic mutations in GENIE data, selecting the threshold with the highest AUROC **(Figure 10A)**. Additionally, for VEPs that offered prediction scores without predefined classifications or recommended score cutoffs, including CADD, ESM1b, PrimateAI, REVEL and VARITY, we manually determined cutoff thresholds by optimizing the sum of sensitivity and specificity in distinguishing known oncogenic mutations from benign dbSNPs **(Figure 10B-F)**. We then applied these thresholds to derive discrete classifications of variants from continuous prediction scores (see **Methods** for more details).

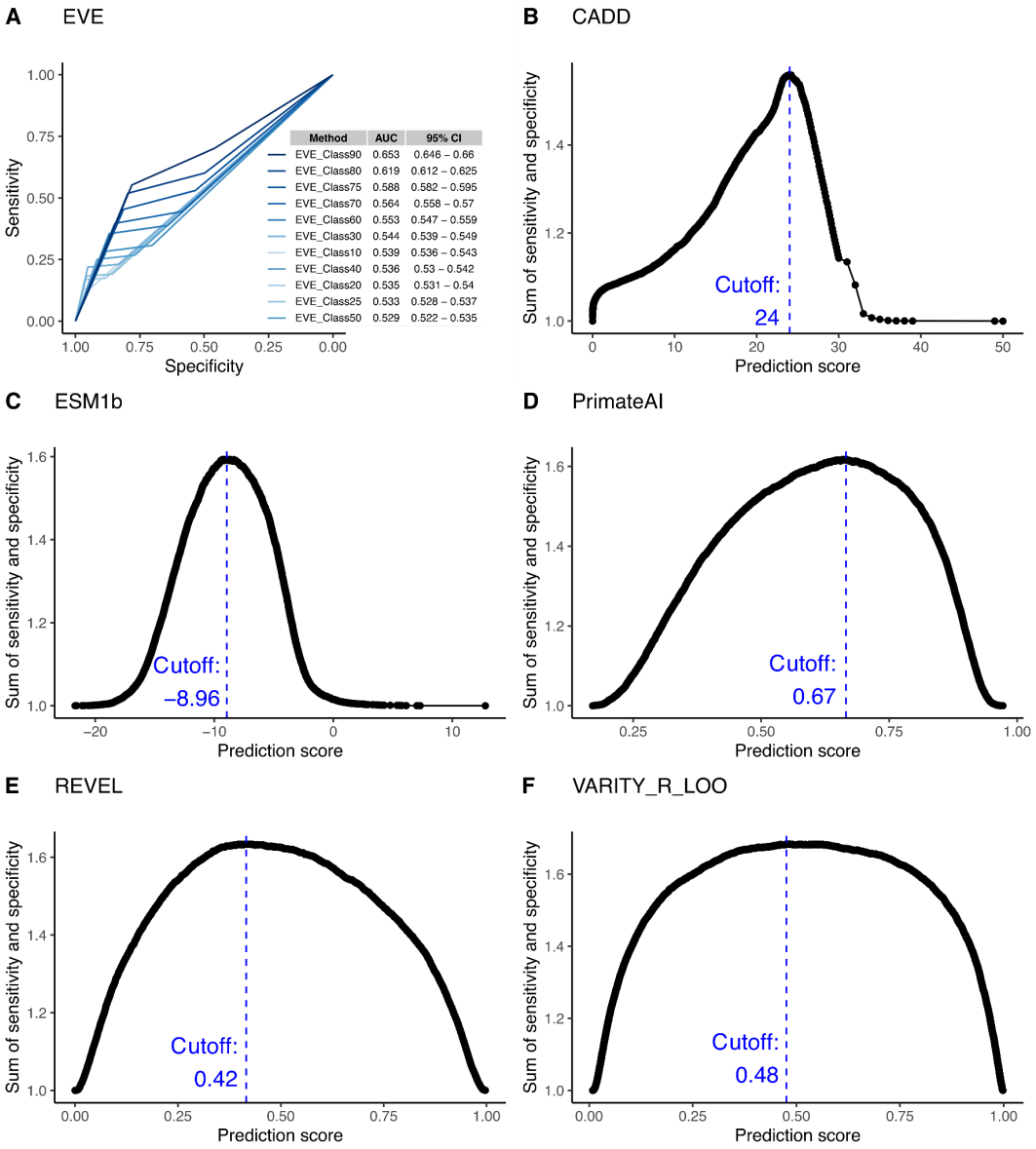

Figure 10. Determination of variant classification thresholds.
A. EVE uncertainty threshold by maximizing AUROC and **B-F.** prediction score cutoff thresholds for **B.** CADD, **C.** ESM1b, **D.** PrimateAI, **E.** REVEL and **F.** VARITY by maximizing the sum of sensitivity and specificity in classifying known oncogenic mutations from benign dbSNPs.

To identify whether predictions from different methods agreed with each other, we calculated Spearman's correlation coefficients between predictions from all methods for variants with OncoKB annotations, including known pathogenic and known neutral mutations, as well as VUSs. Correlations were calculated at the mutation level, where each unique mutation was counted once, and at the population level, where all occurrences of mutations were counted to reflect actual population frequencies of each mutation. Among the OncoKB-annotated variants, OncoKB annotations had overall positive but low correlations with all other methods at both the mutation and population level **(Figure 11A-B)**. Methods with the strongest correlation with OncoKB included VARITY, ESM1b, AlphaMissense and both PolyPhen2 models. Since OncoKB-annotated variants include mostly oncogenic variants **(Tables 3-4)**, the low correlation coefficients suggest that even though most methods were able to identify some oncogenic mutations as pathogenic, their definition of pathogenicity did not fully recapitulate oncogenicity, which included both loss-of-function mutations in TSGs and gain-of-function mutations in OGs. Furthermore, many methods could not properly annotate the neutral variants as non-pathogenic, with some methods such as SIFT and PolyPhen2 annotating ~60% of the neutral mutations as pathogenic **(Tables 3-4)**.

When comparing VEP predictions among themselves, the highest correlations were observed between methods that incorporated evolutionary features, including both PolyPhen2 models, SIFT, MutationAssessor, FATHMM and EVE **(Figure 11A-B)**. ESM1b, a protein language model that does not

explicitly look at homology, as well as two ensemble methods VARITY and CADD also exhibited high correlation in predictions with evolution-based methods. Among deep learning-based methods, AlphaMissense and PrimateAI's predictions were highly correlated at the mutation level, which reflected similarities in their approaches of predicting protein structure and training on weak labels derived from primates **(Figure 11A)**. However, at the population level, PrimateAI's predictions were anti-correlated with those of many VEPs, especially evolution-based VEPs including PolyPhen2, FATHMM and MutationAssessor **(Figure 11B)**, because PrimateAI mis-annotated common oncogenic variants, such as *TP53* R248Q, as non-pathogenic, while other VEPs correctly annotated these variants (not shown). REVEL's predictions were weakly correlated with other methods at both levels, which was surprising considering how its approach incorporated predictions from many methods evaluated here **(Figure 11A-B)**.

Among variant effect predictions for VUSs, strong correlations were observed between the majority of VEPs except for FATHMM, EVE and REVEL **(Figure 11C-D)**. At the population level, the correlation coefficients were stronger across all methods, suggesting that VUSs with similar predicted effects were more common in the population. Since ClinVar relied on human-curated knowledge and is unlikely to include annotations for VUSs, it had weak or negative correlations with other VEPs.

Overall, we showed that VEPs using different approaches agreed in their predictions, especially for VUSs. Among variants with OncoKB annotations,

which were dominantly oncogenic mutations, most methods have positive but weak correlations with OncoKB, suggesting that they did not annotate all oncogenic variants accurately.

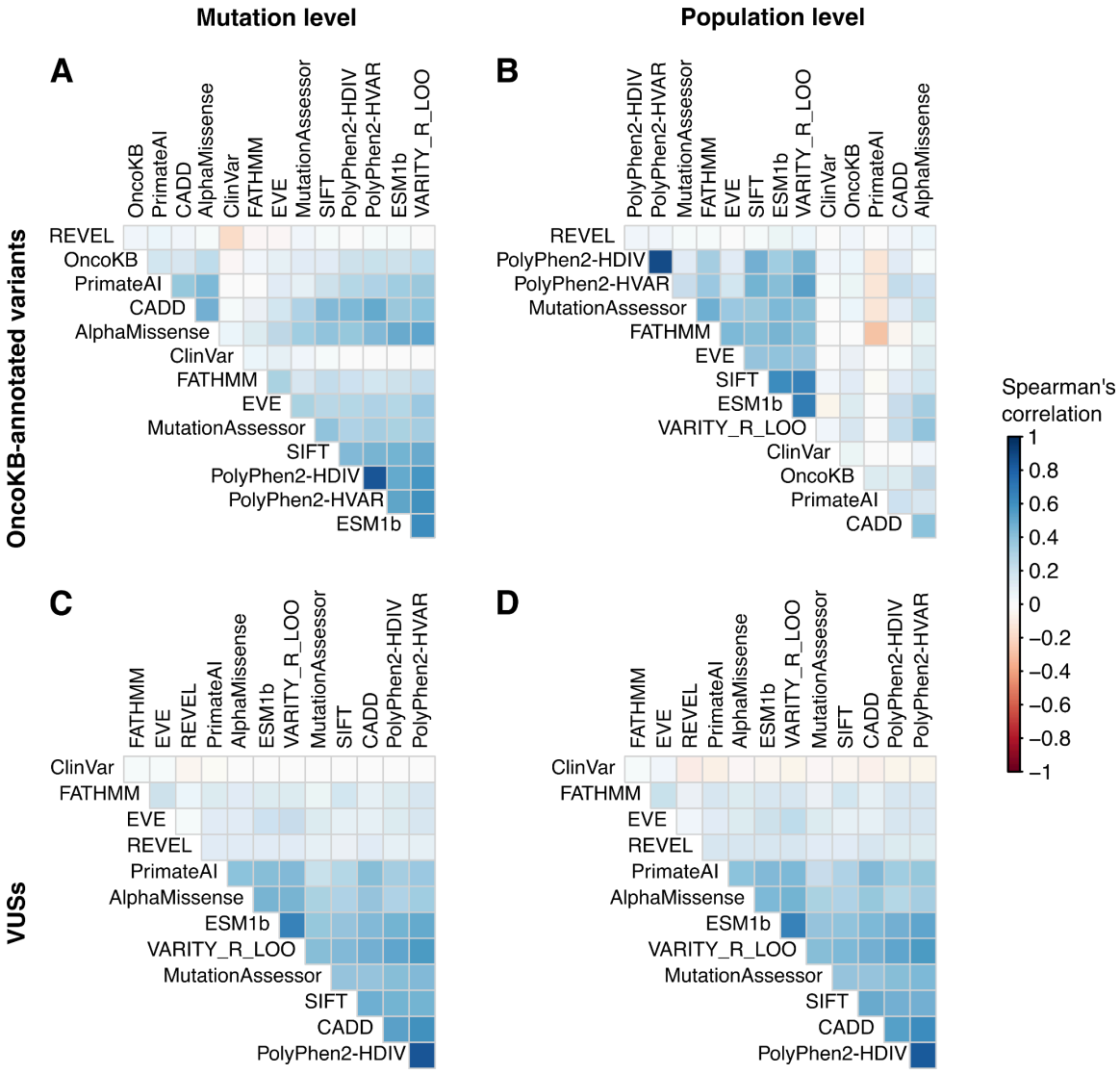

Figure 11. Correlation between predictions of variant effects from VEPs and databases

Spearman's correlation coefficients were calculated for computational predictions of **A-B.** OncoKB-annotated variants, including oncogenic and neutral variants, and **C-D.** VUSs. Correlation coefficients were calculated at the mutation level (A and C) and population level (B and D).

Table 3. Mutation-level count and proportion of mutations in GENIE v14-public by predicted and annotated functions

Missense mutations in GENIE v14-public were annotated with OncoKB and other methods evaluated in this study. Counts and proportions were calculated at the mutation level, where each unique mutation was counted once.

Method/ Database	Annotation	OncoKB: Unknown		OncoKB: Neutral		OncoKB: Oncogenic		OncoKB: Resistance	
		# of mutations	% of total	# of mutations	% of total	# of mutations	% of total	# of mutations	% of total
Alpha Missense	Non-pathogenic	275,229	52.85	336	0.06	1,015	0.19	6	0
	Pathogenic	166,463	31.97	269	0.05	5,944	1.14	72	0.01
	Unknown	70,234	13.49	87	0.02	1,074	0.21	3	0
CADD	Non-pathogenic	249,836	47.98	275	0	1,332	0.26	7	0
	Pathogenic	252,528	48.49	401	0	6,110	1.17	73	0.01
	Unknown	9,562	1.84	16	0	591	0.11	1	0
ClinVar	Non-pathogenic	2,998	0.58	75	0.01	28	0.01	0	0
	Pathogenic	2,504	0.48	110	0.02	1,425	0.27	12	0
	Unknown	506,424	97.25	507	0.1	6,580	1.26	69	0.01
ESM1b	Non-pathogenic	312,520	61.07	293	0.06	1,742	0.34	17	0
	Pathogenic	16,3705	31.99	267	0.05	5,410	1.06	61	0.01
	Unknown	27,305	5.34	117	0.02	314	0.06	3	0
EVE	Non-pathogenic	104,315	20.38	204	0.04	1,143	0.22	10	0
	Pathogenic	65,599	12.82	155	0.03	3,554	0.69	31	0.01
	Unknown	333,616	65.19	318	0.06	2,769	0.54	40	0.01
FATHMM	Non-pathogenic	332,501	64.97	337	0.07	2,973	0.58	28	0.01
	Pathogenic	152,804	29.86	327	0.06	4,315	0.84	51	0.01
	Unknown	18,225	3.56	13	0	178	0.03	2	0

Mutation Assessor	Non-pathogenic	269,492	52.66	326	0.06	1,978	0.39	51	0.01
	Pathogenic	213,919	41.8	341	0.07	5,289	1.03	30	0.01
	Unknown	20,119	3.93	10	0	199	0.04	0	0
PolyPhen2-HDIV	Non-pathogenic	173,569	33.92	176	0.03	877	0.17	9	0
	Pathogenic	294,375	57.52	380	0.07	6,248	1.22	69	0.01
	Unknown	35,586	6.95	121	0.02	341	0.07	3	0
PolyPhen2-HVAR	Non-pathogenic	222,851	43.55	242	0.05	1,247	0.24	10	0
	Pathogenic	245,093	47.89	314	0.06	5,878	1.15	68	0.01
	Unknown	35,586	6.95	121	0.02	341	0.07	3	0
PrimateAI	Non-pathogenic	254,821	49.79	338	0.07	1,703	0.33	4	0
	Pathogenic	236,114	46.14	337	0.07	5,733	1.12	77	0.02
	Unknown	12,595	2.46	2	0	30	0.01	0	0
REVEL	Non-pathogenic	132,056	25.8	66	0.01	654	0.13	5	0
	Pathogenic	58,637	11.46	88	0.02	1,481	0.29	27	0.01
	Unknown	312,837	61.13	523	0.1	5,331	1.04	49	0.01
SIFT	Non-pathogenic	202,790	39.63	243	0.05	1,151	0.22	17	0
	Pathogenic	283,145	55.33	419	0.08	6,066	1.19	62	0.01
	Unknown	17,595	3.44	15	0	249	0.05	2	0
VARITY_R_LOO	Non-pathogenic	290,183	56.7	252	0.05	1,009	0.2	9	0
	Pathogenic	177,613	34.71	308	0.06	6,110	1.19	69	0.01
	Unknown	3,5734	6.98	117	0.02	347	0.07	3	0
Total # of mutations		503,530		677		7,466		81	

Table 4. Population-level count and proportion of mutations in GENIE v14-public by predicted and annotated functions

Missense mutations in GENIE v14-public were annotated with OncoKB and other methods evaluated. Counts and proportions were calculated at the population level, where each unique occurrence of mutation was counted once.

Method/ Database	Annotation	OncoKB: Unknown		OncoKB: Neutral		OncoKB: Oncogenic		OncoKB: Resistance	
		# of mutations	% of total	# of mutations	% of total	# of mutations	% of total	# of mutations	% of total
Alpha Missense	Non-pathogenic	531,225	44.63	2,277	0.19	9,105	0.76	23	0
	Pathogenic	285,513	23.99	1,493	0.13	223,898	18.81	675	0.06
	Unknown	122,141	10.26	477	0.04	13,398	1.13	9	0
CADD	Non-pathogenic	464,410	39.02	1,951	0.16	26,116	2.19	16	0
	Pathogenic	460,696	38.71	2,249	0.19	216,378	18.18	685	0.06
	Unknown	13,773	1.16	47	0	3,907	0.33	6	0
ClinVar	Non-pathogenic	30,103	2.53	717	0.06	323	0.03	0	0
	Pathogenic	18,643	1.57	1,071	0.09	145,246	12.2	241	0.02
	Unknown	890,133	74.79	2,459	0.21	100,832	8.47	466	0.04
ESM1b	Non-pathogenic	574,574	52.41	2,062	0.19	30,278	2.76	96	0.01
	Pathogenic	286,721	26.15	1,426	0.13	141,208	12.88	403	0.04
	Unknown	49,818	4.54	614	0.06	9,054	0.83	16	0
EVE	Non-pathogenic	198,217	18.08	1,317	0.12	26,819	2.45	153	0.01
	Pathogenic	119,842	10.93	899	0.08	99,835	9.11	208	0.02
	Unknown	593,054	54.1	1,886	0.17	53,886	4.92	154	0.01
FATHMM	Non-pathogenic	597,977	54.55	2,247	0.2	77,068	7.03	178	0.02
	Pathogenic	280,271	25.57	1,793	0.16	94,545	8.62	327	0.03
	Unknown	32,865	3	62	0.01	8,927	0.81	10	0

Mutation Assessor	Non-pathogenic	496,686	45.31	2,130	0.19	46,565	4.25	368	0.03
	Pathogenic	377,468	34.43	1,924	0.18	130,641	11.92	147	0.01
	Unknown	36,959	3.37	48	0	3,334	0.3	0	0
PolyPhen2-HDIV	Non-pathogenic	317,491	28.96	1,179	0.11	24030	2.19	23	0
	Pathogenic	530,150	48.36	2,309	0.21	147,294	13.44	476	0.04
	Unknown	63,472	5.79	614	0.06	9216	0.84	16	0
PolyPhen2-HVAR	Non-pathogenic	412,450	37.62	1,590	0.15	33,303	3.04	28	0
	Pathogenic	435,191	39.7	1,898	0.17	138,021	12.59	471	0.04
	Unknown	63,472	5.79	614	0.06	9,216	0.84	16	0
PrimateAI	Non-pathogenic	476,039	43.42	2,324	0.21	31,795	2.9	13	0
	Pathogenic	411,014	37.49	1,766	0.16	148,620	13.56	502	0.05
	Unknown	24,060	2.19	12	0	125	0.01	0	0
REVEL	Non-pathogenic	231,694	21.13	592	0.05	13,135	1.2	27	0
	Pathogenic	110,298	10.06	757	0.07	69,650	6.35	282	0.03
	Unknown	569,121	51.91	2,753	0.25	97,755	8.92	206	0.02
SIFT	Non-pathogenic	37,6741	34.37	1,626	0.15	10,836	0.99	172	0.02
	Pathogenic	501,043	45.7	2,411	0.22	160,597	14.65	333	0.03
	Unknown	33,329	3.04	65	0.01	9,107	0.83	10	0
VARITY_R_LOO	Non-pathogenic	543,488	49.58	1,834	0.17	12,415	1.13	29	0
	Pathogenic	303,799	27.71	1,654	0.15	157,370	14.36	470	0.04
	Unknown	63,826	5.82	614	0.06	10,755	0.98	16	0
Total # of mutations		911,113		4,102		180,540		515	

VEPs can identify pathogenic cancer mutations

As proof-of-concept that VEPs were able to identify pathogenic cancer mutations, we first compared the distribution of prediction scores outputted by each method for each subgroup of mutations as annotated by OncoKB. We conducted this analysis at both the mutation and population levels, scaling prediction scores from all VEPs to 0-1 and standardizing so that higher scores indicated higher predicted pathogenicity. At the mutation level, all methods predicted that benign dbSNPs had significantly lower scores compared to oncogenic mutations annotated by OncoKB, demonstrating their ability to distinguish benign versus pathogenic mutations in cancer (Tukey's test p-val < 0.001, **Figure 12A**). Interestingly, mutations in GENIE annotated as neutral or unknown/VUSs by OncoKB had significantly higher scores than known benign mutations, suggesting that there were potential unannotated drivers **(Figure 12A)** and helping to explain the difficulty most VEPs had in annotating neutral mutations **(Tables 3-4)**. Across all methods, oncogenic mutations in TSGs had higher predicted scores compared to those in OGs. This observation aligns with previous study showing that activating OG mutations are significantly less well-predicted than inactivating mutations in TSGs[155,156] and partially explains the weak correlation between VEPs' predictions and OncoKB annotations **(Figure 11)**. Similar results were observed at the population level **(Figure 12B).**

We noted that at both mutation and population levels, certain methods, such as AlphaMissense and VARITY, achieved larger separations in predicted score means between known benign and known oncogenic mutations compared

to other methods, potentially indicating better performance (**Figure 12**). To quantify and compare performance of VEPs in annotating known oncogenic and non-oncogenic dbSNP variants, we calculated AUROC for each method and compared them using DeLong's test.

We found that across four main approaches, at both mutation and population levels, the ensemble and deep learning-based methods outperformed the evolution-based methods in predicting both TSG and OG oncogenic mutations **(Figure 13)**. Notably, AlphaMissense significantly outperformed other deep learning-based methods as well as other best-in-class methods in predicting pathogenic mutations, achieving best AUROC of 0.979 and 0.980 for annotating oncogenic mutations in OGs and TSGs at the population level respectively **(Figure 13B)**. Among the ensemble methods, VARITY and REVEL, both trained on human-curated data outperformed CADD, which was trained on weak population-derived labels **(Figure 13)**. Among the evolution-based methods, EVE, the only unsupervised deep learning method in this class, outperformed others at the population level (AUROC of 0.833 and 0.919 for OGs and TSGs respectively); however, SIFT, the oldest method in this class, did not fall far behind (AUROC of 0.815 and 0.905 for OGs and TSGs, respectively), attesting to the merit of SIFT's evolutionary conservation approach in predicting pathogenicity **(Figure 13)**. Among two PolyPhen2 models, which integrated evolution and protein structure information, PolyPhen2-HVAR outperformed PolyPhen2-HDIV in all evaluation contexts (best AUROC of 0.931 compared to 0.909 in predicting oncogenic mutations in TSGs at the population level, **Figure**

13B) even though the predictions of these two models were highly correlated **(Figure 11)**. The superior performance of PolyPhen2-HVAR in annotating cancer mutations could be due to its training on disease-causing mutations and non-disease-related nsSNPs, many of which may be related to cancer, whereas PolyPhen2-HDIV was trained on Mendelian diseases-related mutations that were less relevant to cancer.

Finally, this analysis highlights the discrepancy in method performance in predicting TSG and OG pathogenic mutations. As predicted by the higher predicted score means in TSG oncogenic mutations versus OG **(Figure 12)**, all methods had higher AUROC when predicting TSG mutations. This finding was further corroborated by the higher sensitivity in TSG compared to OG across all methods **(Figure 14)**. However, within each gene group, sensitivity further varied at the gene level **(Table 5)**, which can potentially result from overrepresentation of certain genes in human-curated training data.

Taken together, these results demonstrate that VEPs were able to identify pathogenic mutations in cancer, with deep learning-based methods, especially AlphaMissense, outperforming other methods. The use of manually-curated data in training ensemble methods such as VARITY and REVEL did result in superior performance compared to methods trained on weak population-derived labels, even though concerns about generalizability remained. Furthermore, oncogenic mutations in TSGs were predicted more accurately compared to those in OGs.

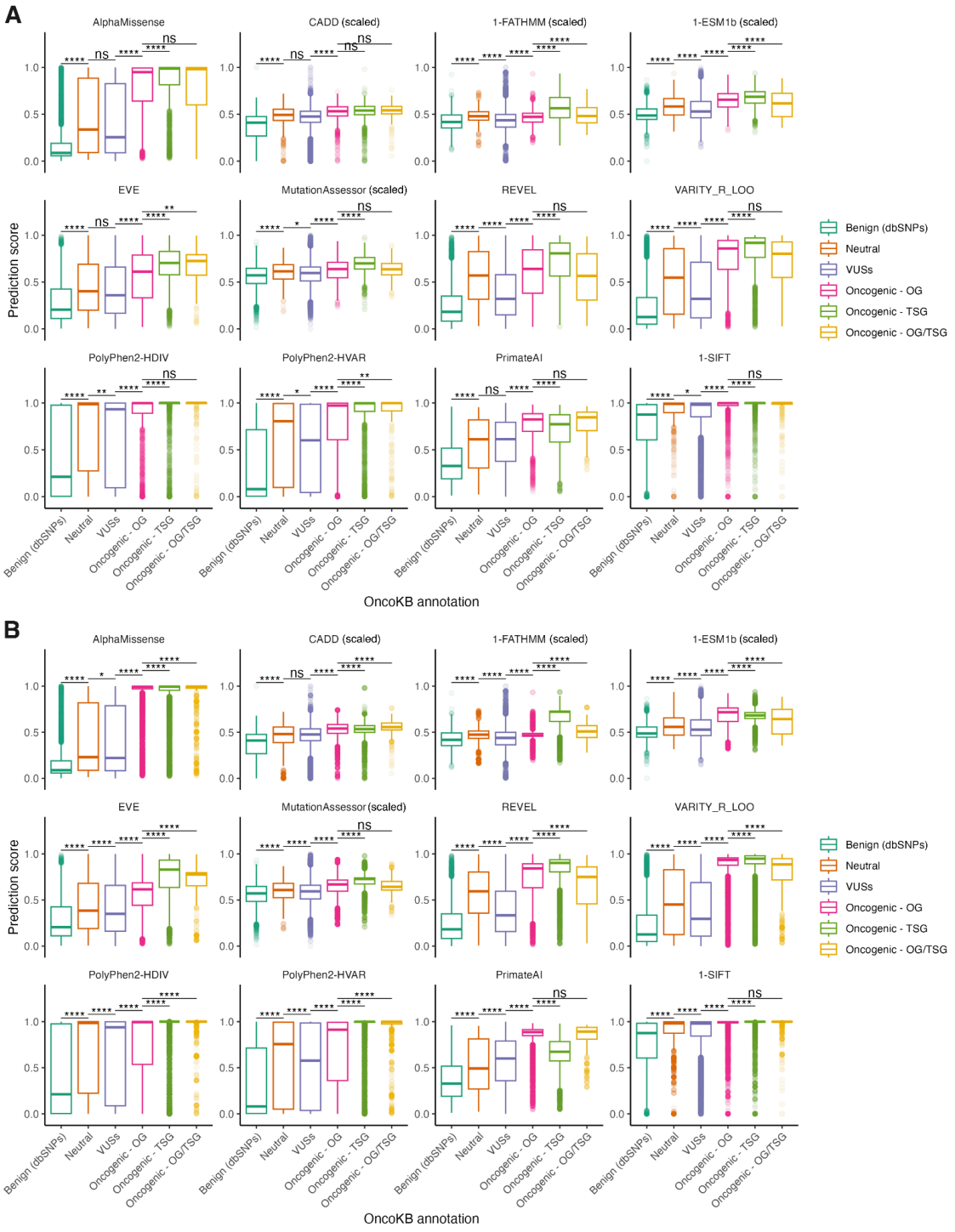

Figure 12. Prediction scores of missense mutations and benign dbSNPs
Distribution of prediction scores from benign dbSNPs and missense mutations in GENIE v.14-public, broken down by their occurrence in oncogenes (OG), tumor suppressor genes (TSG) or genes that act as both (OG/TSG). For most VEPs, scores are in the range of 0-1, with higher scores suggest higher predicted

pathogenicity. For ease of comparison, scores from methods with different ranges, including CADD, MutationAssessor, FATHMM and ESM1b, were scaled to the 0-1 range. For methods where lower scores indicated higher pathogenicity, the difference between 1 and the (scaled) prediction scores was plotted. Boxplots depict means +- interquartile ranges.
Pairwise differences in means between groups were tested for significance using Tukey's range test and p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method. Asterisks on brackets depict significance levels of q-value.

A. Score distributions at the mutation level, in which each unique mutation is included once.

B. Score distributions at the population level, in which all occurrences of missense mutations are included.

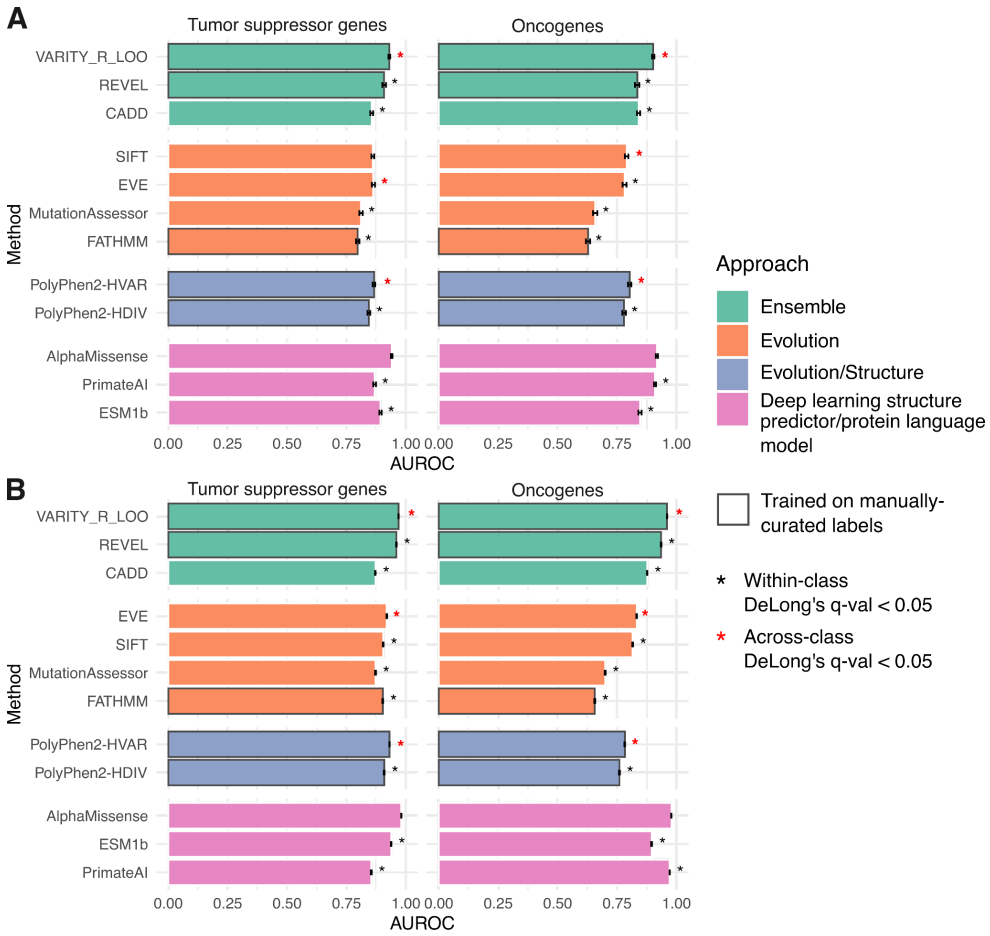

Figure 13. Areas under receiver operating curves for VEPs
Significant differences in AUCs were tested using DeLong's tests and p-values were corrected with the Benjamini-Hochberg procedure. Within each class of approach, pairwise tests were performed between the best-in-class method and

all other methods, with significant differences denoted with black asterisks. Additionally, across-class pairwise tests were performed between each best-in-class method and the overall best method, with significant differences denoted with red asterisks.

A. Mutation-level AUROC for each method were calculated using all 8,033 missense mutations in either OGs or TSGs as the positive class and 10,000 dbSNPs as the negative class.

B. Population-level AUROC for each method were calculated using all occurrences of missense mutations in either OGs or TSGs as the positive class and up-sampled 10,000 dbSNPs as the negative class.

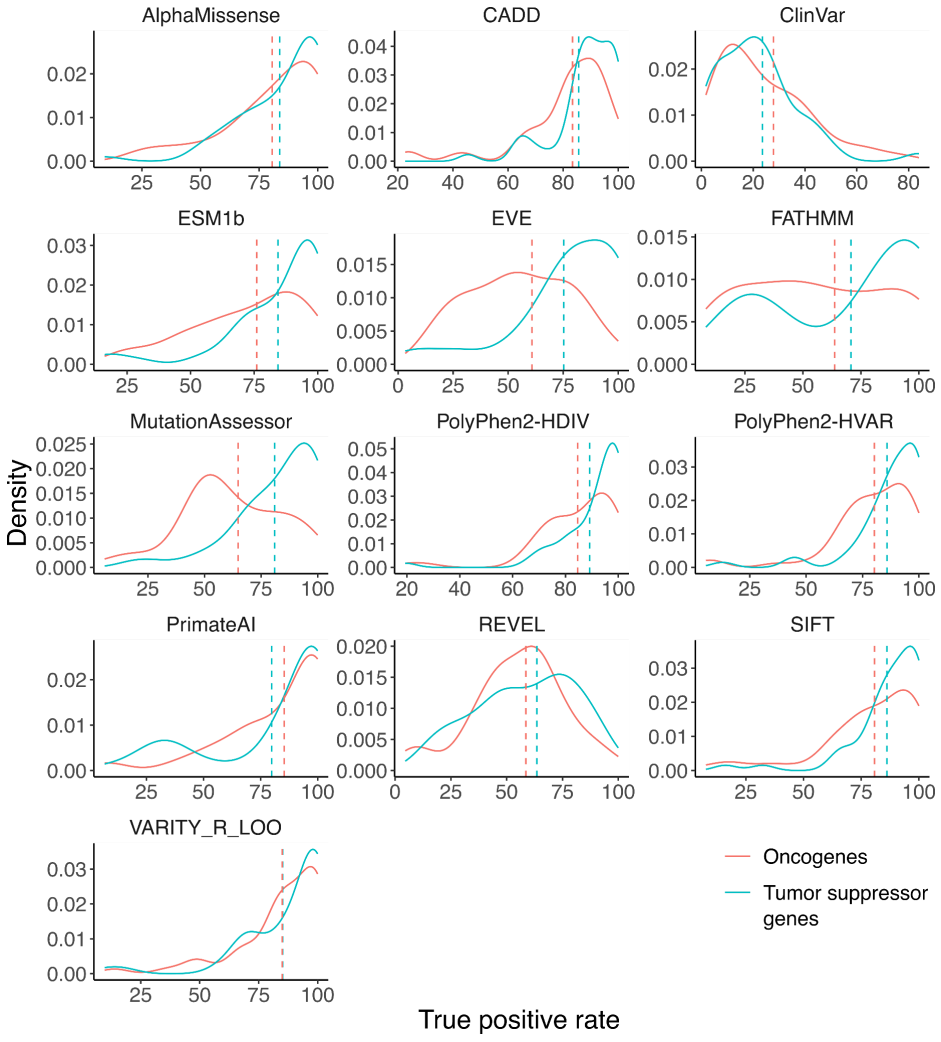

Figure 14. True positive rates (TPR) of VEPs in all GENIE-sequenced genes TPR is defined as the number of known oncogenic mutations accurately annotated as pathogenic by each method divided by the total number of known oncogenic mutations.

Table 5. TPR of methods at gene level for genes with more than 100 unique oncogenic mutations in the GENIE v14-public cohort

Gene	TP53	DNMT3A	PTEN	EGFR	PIK3CA	SMAD4	CDKN2A	ATM	VHL	PDGFRA	BRAF	KEAP1	KRAS	BRCA1
Type	TSG	TSG	TSG	OG	OG	TSG	TSG	TSG	TSG	OG	OG	TSG	OG	TSG
# Oncogenic mutations	807	460	387	207	199	153	143	141	130	127	122	112	109	104
Alpha Missense	73.1	93.0	95.4	66.2	82.4	99.4	72.7	57.5	87.7	85.8	91.8	94.6	93.6	65.4
CADD	64.1	94.8	91.7	71.0	72.4	96.1	79.7	83.0	88.5	90.6	88.5	92.0	91.7	88.5
ClinVar	41.0	4.6	28.4	14.0	33.7	15.0	25.9	20.6	50.0	7.1	41.0	NA	37.6	2.9
ESM1b	73.5	89.6	95.1	65.2	38.7	99.4	77.6	53.2	29.2	86.6	0	75.9	94.5	66.4
EVE	73.9	64.8	74.9	49.3	17.6	97.4	NA	68.8	53.1	NA	NA	NA	51.4	82.7
FATHMM	100	75.0	99.7	34.3	10.6	89.5	34.3	31.2	100	84.3	19.7	28.6	57.8	82.7
Mutation Assessor	84.4	77.0	91.0	50.2	47.2	98.7	27.3	86.5	77.7	50.4	51.6	67.9	72.5	76.0
PolyPhen2-HDIV	86.5	93.0	88.9	84.1	82.4	94.8	93.0	90.8	95.4	96.9	NA	95.5	86.2	72.1
PolyPhen2-HVAR	82.4	88.0	84.0	75.4	68.3	92.8	85.3	85.1	90.0	95.3	NA	92.9	78.0	44.2
PrimateAI	30.4	97.6	99.7	69.1	99.5	100	35.7	49.7	88.5	92.1	95.9	86.6	100	31.7
REVEL	77.5	49.1	71.3	39.1	49.3	77.8	54.6	46.8	66.2	44.9	66.4	76.8	63.3	45.2
SIFT	85.9	88.0	92.5	71.5	61.8	98.7	79.7	83.7	32.3	92.1	18.9	85.7	92.7	90.4
VARITY_R_LOO	90.5	94.4	95.9	89.4	84.9	100	69.9	68.8	100	87.4	NA	93.8	100	81.7

VEPs can identify unannotated driver mutations

Reclassified pathogenic variants occur at binding residues

We next investigated the ability of VEPs to identify unannotated driver mutations from the large pool of detected VUSs. We used multiple angles to evaluate whether VUSs classified as pathogenic by VEPs had potential functional effect in cancer, including whether they affected protein binding sites, correlated with patient outcomes and followed expected patterns of driver mutual exclusivity.

Pathogenic mutations can alter protein function by disrupting interactions with other proteins and ligands[157]. We sought to provide structure-based evidence for the function of reclassified pathogenic variants by probing whether these variants may lead to disruption of ligand or protein-protein binding. To that end, we identified and annotated mutations in GENIE that occurred at residues known to be involved in ligand binding and protein-protein interactions ("binding residues") for proteins with available crystal structures.

Mutations affecting binding residues in all genes were significantly more likely to be annotated as oncogenic by OncoKB (Fisher's test, p-value < 0.001, **Figure 15A**). Among VUSs, all VEPs predicted significantly more binding residue mutations to be pathogenic compared to non-binding ones. In genes with the highest number of binding residues, including *TP53* and *KEAP1*, binding residue mutations were more likely to be reclassified as pathogenic, whereas non-binding residue mutations were more likely to be reclassified as benign **(Figure 15A)**. This result suggests that the disruption of ligand or protein-protein interactions at

these critical binding residues may contribute to the pathogenic nature of these reclassified pathogenic variants.

Figure 15. Reclassified pathogenic mutations occur at ligand-binding residues and protein-protein interaction (PPI) hotspots

Frequency and annotation of missense mutations occurring at binding residues (either ligand binding or PPI hotspots, see Supplemental Appendix) or non-binding residues of genes with high number of binding residues in

A. GENIE v14-public pan-cancer cohort,
B. MSK-IMPACT NSCLC cohort.

Asterisks denote significance (q-value <= 0.1) in Fisher's exact tests. Numbers at

the bottom of barcharts indicate the number of unique mutations represented in each bar. OncoKB groups include all missense mutations, whereas variant effect predictor groups only include VUSs.

Association between reclassified pathogenic variants and overall survival in non-small cell lung cancer

Alterations in cancer genes can serve as prognostic markers across different cancer types[4]. To study the impact of VUS by classification scheme on OS, we focused on non-small cell lung cancer (NSCLC), the world's leading cause of cancer mortality[158] and the cancer type most common in our tumor genomic sequencing cohort. In patients with NSCLC, mutations in multiple genes, including *KRAS*, *STK11* and *KEAP1*, have been associated with worse OS[159–162]. We investigated whether reclassified pathogenic variants were associated with overall survival in NSCLC using two cohorts of patients: 8,690 patients with MSK-IMPACT clinical sequencing and 977 non-MSK patients from the GENIE BPC NSCLC cohort.

To identify the association between reclassified pathogenic variants and outcome, we stratified patients based on gene-level pathogenicity annotations and compared overall survival between groups using Cox's proportional hazard (PH) models **(Methods)**. To balance the differences in demographic, clinical and genomic covariates between comparison groups, we calculated inverse probability of treatment weights (IPTW) based on variables that could influence OS, including sex, age, stage, NLP-derived metastasis and treatment status, as well as TMB and MSI status **(Figure 16)**. We weighted each case using the calculated weights before fitting Cox PH models or KM curves.

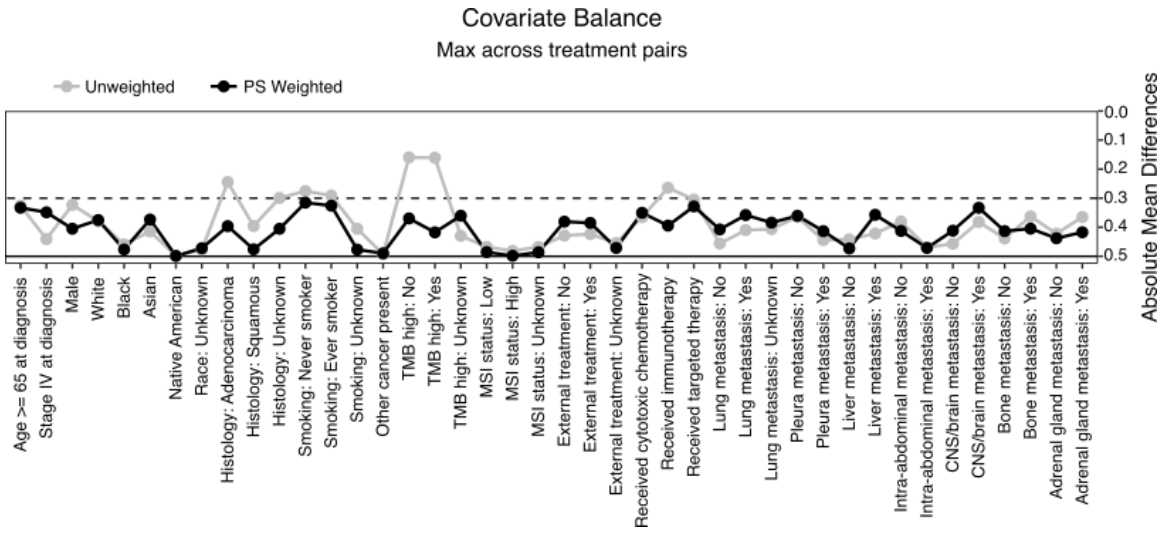

Figure 16. Balance of covariates between patient strata using IPTW
Absolute mean differences between clinical and demographic covariates in four groups of patients stratified by *KEAP1* mutational status before and after IPTW. IPTW helped achieve balance in variables with large imbalance between groups, such as TMB high status.

Known oncogenic variants[3] in several genes were associated with worse OS, controlling for tumor burden, treatment, and other features. VUSs in *KEAP1* and *SMARCA4* annotated pathogenic by multiple methods were also associated with worse OS, while those annotated as likely benign were associated with better outcome, suggesting meaningful discrimination among VUSs by computational methods **(Figure 17)**. These findings were consistent in both the MSK-IMPACT and non-MSK GENIE BPC cohorts **(Figure 17)**. The higher OS risks in patients with reclassified pathogenic mutations in these two genes were comparable with the risks associated with known oncogenic mutations **(Figure 18A)**, while patients with reclassified benign mutations from most VEPs in these genes had similar OS risk compared to those with no mutation **(Figure 18B)**. However, patients with *KEAP1* reclassified benign mutations identified by CADD,

FATHMM and PrimateAI had significantly worse OS compared to patients without *KEAP1* mutations, suggesting that these methods may not be as sensitive in identifying pathogenic VUSs compared to other VEPs **(Figure 18B)**.

Looking closely at the distributions of mutations, we observed that reclassified pathogenic mutations in *KEAP1* occurred throughout the four domains of the protein, whereas VUSs were observed in the N-terminus **(Figure 19A)**. In *SMARCA4*, reclassified pathogenic mutations were concentrated in the functional SNF2 family N-terminal and helicase domains of the protein, where there were few reclassified benign mutations **(Figure 19B)**. NSCLC-specific reclassified pathogenic mutations in *KEAP1* and *SMARCA4* also occurred at binding residues, suggesting a potential mechanism for their pathogenicity compared to reclassified benign mutations **(Figure 15B)**.

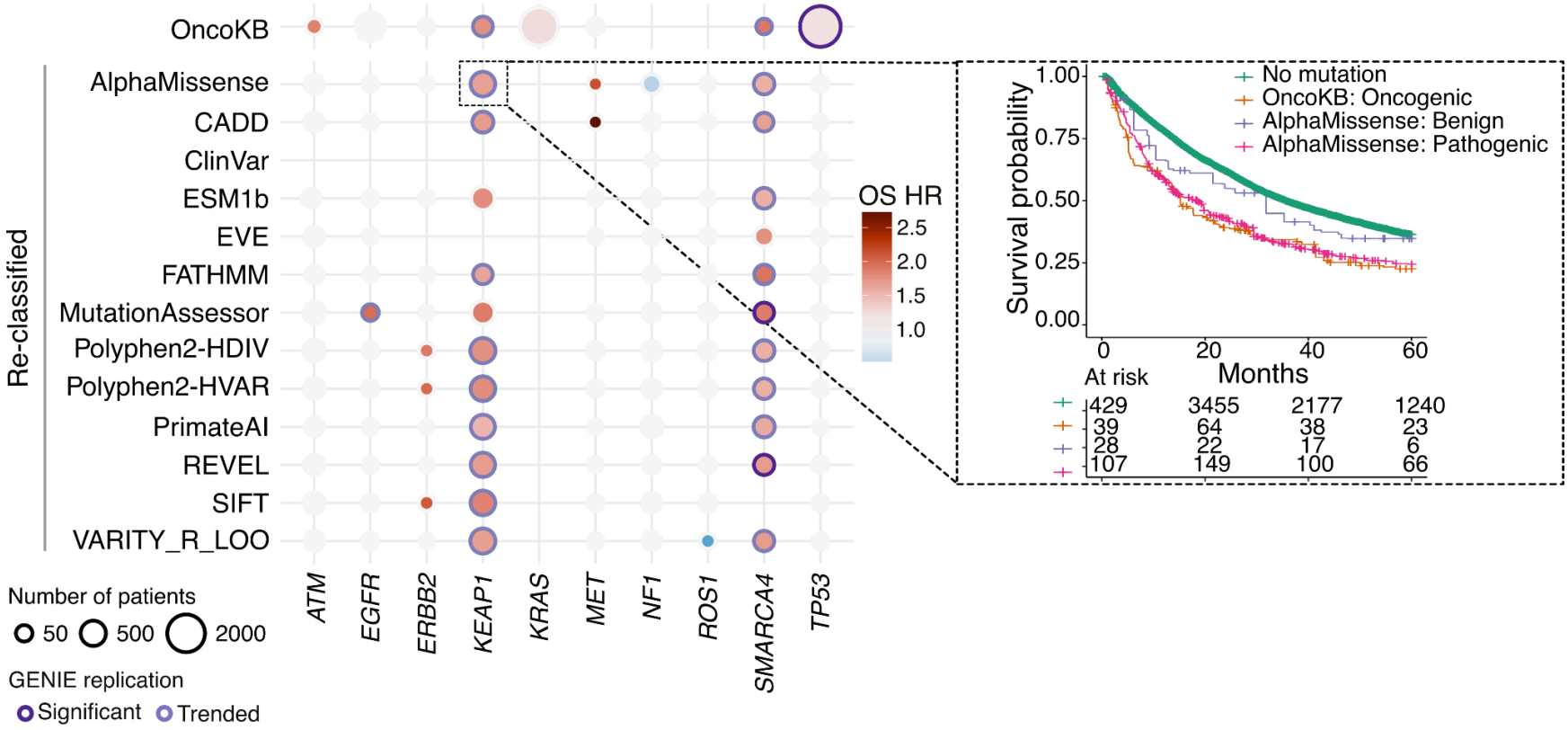

Figure 17. OS stratified by reclassified mutation status

Weighted OS hazard ratios (from time of diagnosis left truncated at time of sequencing) of patients harboring reclassified oncogenic mutations compared to patients without mutation in commonly mutated genes in non-small cell lung cancer (NSCLC). Patients were from MSK-IMPACT NSCLC (N=8,690) and AACR GENIE Biopharma Collaborative NSCLC (N=977) cohorts.

Inset: Weighted KM curves comparing OS from time of diagnosis left-truncated at time of sequencing of patients based on *KEAP1* mutations annotation in the MSK-IMPACT NSCLC cohort. KM curves were truncated at 5 years.

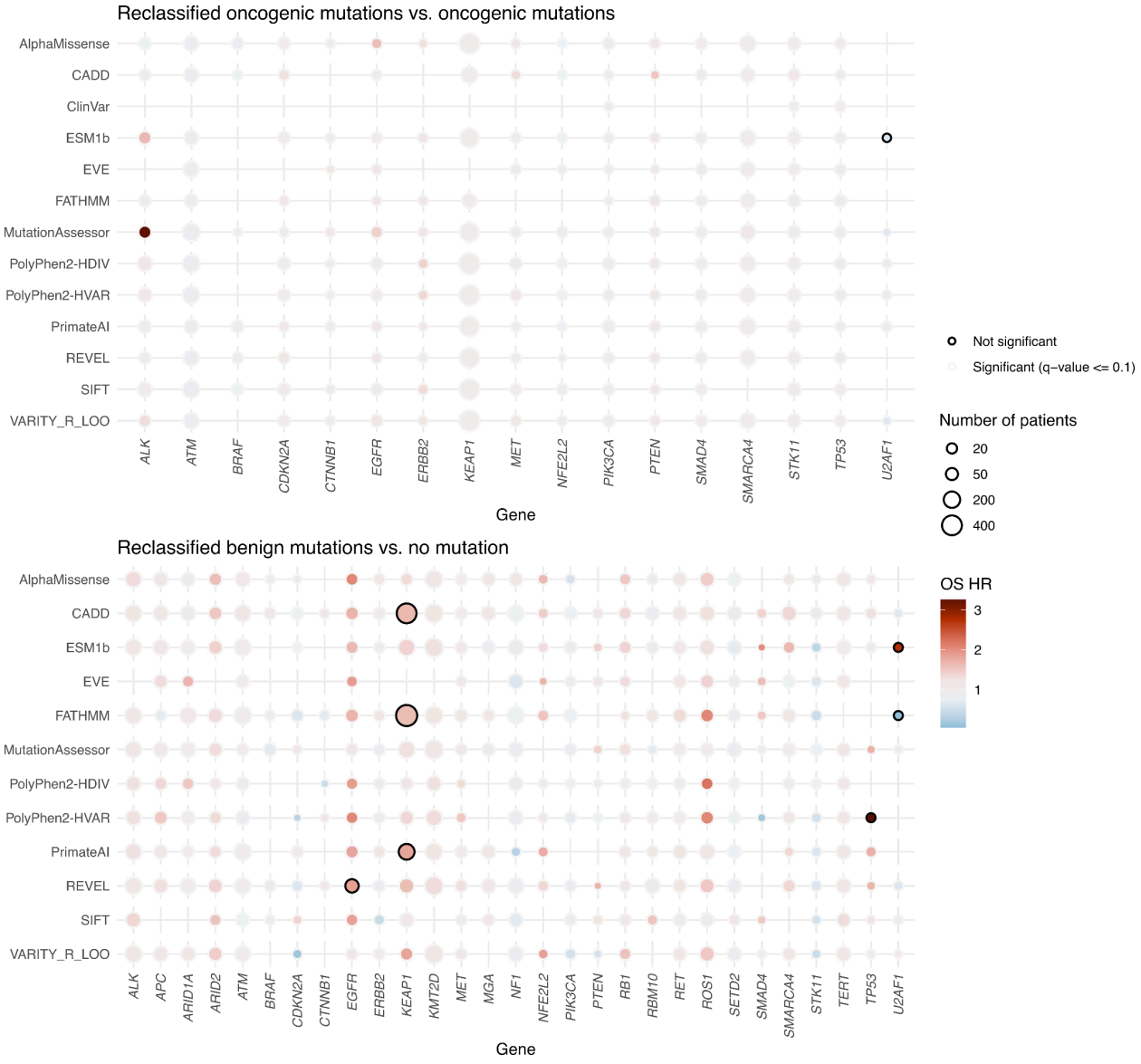

Figure 18. Cox PH coefficients of reclassified mutations versus known oncogenic mutations or no mutation

A. OS hazard ratios of patients with reclassified pathogenic mutations compared to patients with oncogenic mutations were obtained from weighted Cox PH models.

B. OS hazard ratios of patients with reclassified benign mutations compared to patients with no mutation in a given gene were obtained from weighted Cox PH models.

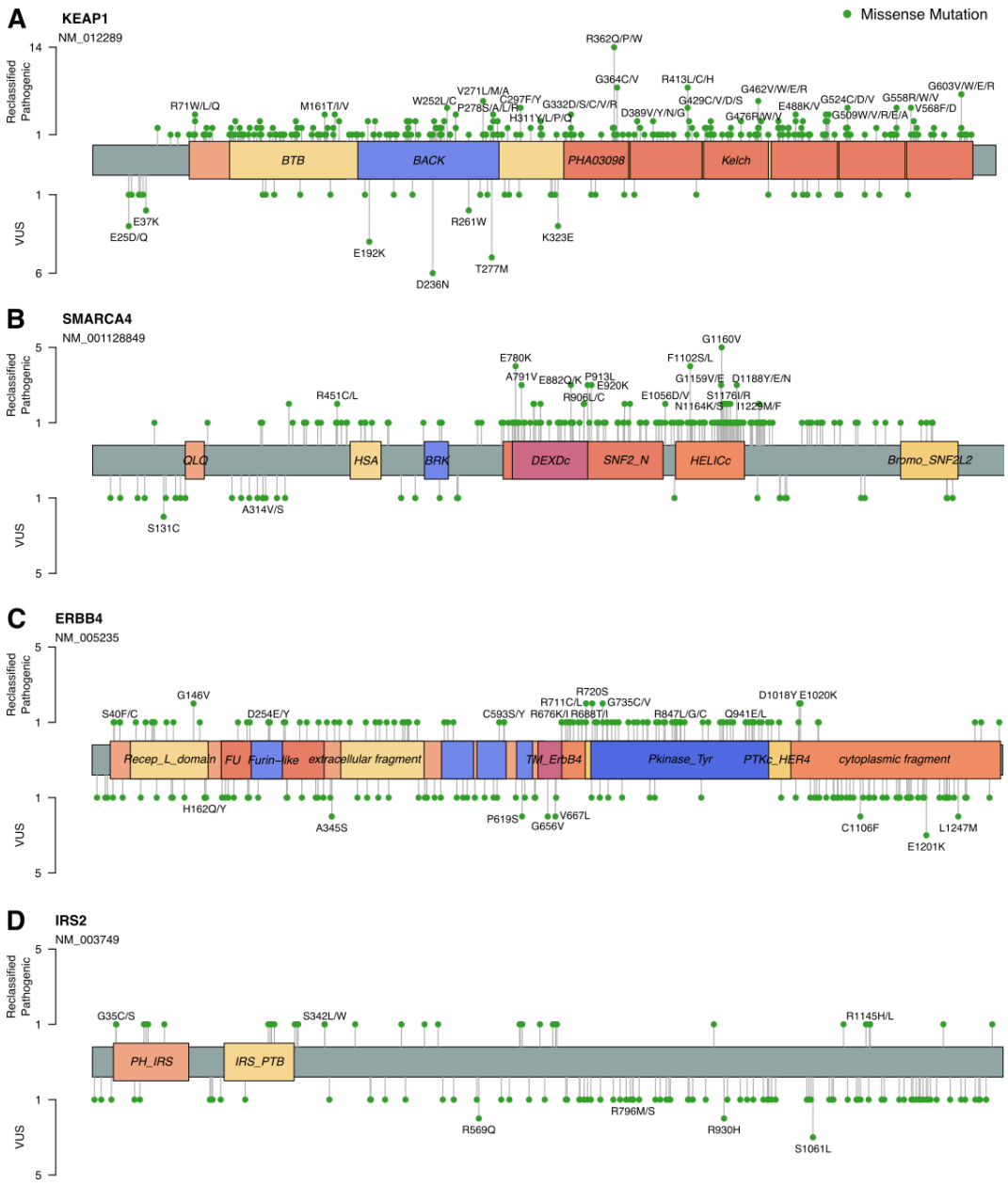

Figure 19. Lollipop plots of VUSs reclassified as pathogenic or benign by AlphaMissense

Amino acid-level changes resulting from AlphaMissense-reclassified pathogenic mutations in the MSK-IMPACT NSCLC cohort are plotted for **A.** *KEAP1,* **B.** *SMARCA4,* **C.** *ERBB4* and **D.** *IRS2*. The most commonly mutated residues are labeled.

Concurrent mutations in certain gene combinations may worsen survival in an additive manner, as is known for *STK11* and *KEAP1* in NSCLC[160]. To test whether AlphaMissense reclassified pathogenic variants in *STK11* and *KEAP1* were similarly associated with worse survival, we compared OS of patients with double *KEAP1* and *STK11* reclassified mutations with patients with single reclassified mutation and patients without any mutation in these two genes. We found that patients harboring both *KEAP1* and *STK11* AlphaMissense mutations had worse OS compared to those with reclassified pathogenic mutations in either genes, as well as those with no mutation in these two genes **(Figure 20A)**. This result further suggests that the *KEAP1* and *STK11* reclassified variants from AlphaMissense followed expected patterns of additive prognostic significance, suggesting biologic validity. A similar pattern was observed in patients carrying concurrent *KEAP1* and *STK11* known oncogenic mutations **(Figure 20B)**.

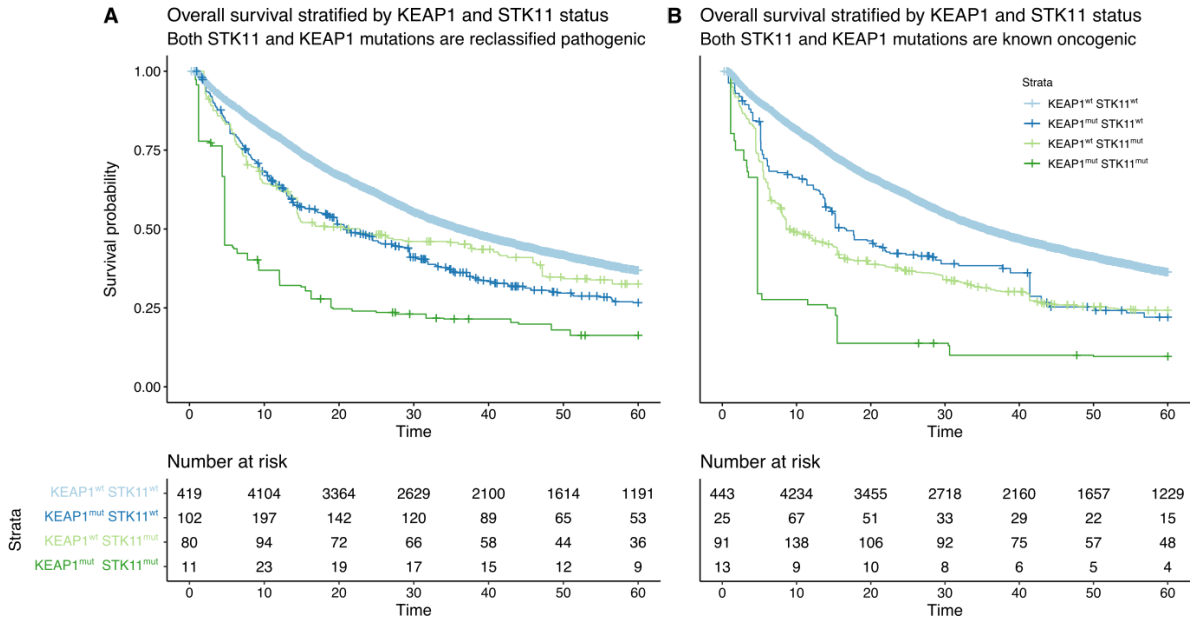

Figure 20. OS of patients with concurrent *STK11* and *KEAP1* mutations

A. KM curves comparing patients with concurrent AlphaMissense-reclassified

pathogenic *KEAP1* and *STK11* mutations versus reclassified pathogenic mutations in either gene or without any mutations.

B. KM curves comparing patients with concurrent OncoKB oncogenic *KEAP1* and *STK11* mutations versus oncogenic mutations in either gene or without any mutations.

Beyond *KEAP1* and *SMARCA4*, mutations reclassified as pathogenic by AlphaMissense in *MET* were also associated with worse OS, while reclassified pathogenic mutations in *NF1* were associated with better OS compared to no mutations; however, these findings were not replicated in the non-MSK GENIE BPC cohort likely due to low numbers **(Figure 17)**.

Reclassified pathogenic variants follow mutual exclusivity pattern within oncogenic pathways

Oncogenic mutations in genes within the same oncogenic signaling pathway tend to not co-occur in the same patient due to functional redundancy[163,164]. In NSCLC, oncogenic mutations in the RTK/RAS, NRF1 and TP53 pathways have been shown to exhibit mutual exclusivity[164]. To demonstrate that reclassified pathogenic mutations have comparably pathogenic effect on a pathway as known oncogenic mutations, we aimed to identify whether reclassified pathogenic mutations were mutually exclusive with other known oncogenic mutations in these pathways within the MSK-IMPACT NSCLC data.

To this end, we identified genes in a given pathway whose reclassified pathogenic mutations were mutually exclusive with other oncogenic mutations in the same pathway using two-sided Fisher's exact tests. Then, for each method and each pathway, we divided the number of genes with mutually exclusive reclassified pathogenic mutations by the total number of genes in the pathway,

resulting in the percentage of one-versus-all mutual exclusivity for each pathway and method combination **(Figure 21A)**. All methods were able to annotate mutations that exhibit mutual exclusivity with other oncogenic mutations in all three pathways **(Figure 21A)**.

Notably, within the NRF2 pathway, *KEAP1* VUSs reclassified as pathogenic by any methods were mutually exclusive with oncogenic mutations in *KEAP1*, *NFE2L2* and *CUL3*, whereas *KEAP1* VUSs reclassified as benign by most VEPs tended to co-occur with other oncogenic mutations in the pathway **(Figure 21C)**. It is noteworthy that *KEAP1* reclassified benign mutations as identified by CADD, ESM1b, FATHMM and MutationAssessor were also mutually exclusive with other NRF2 oncogenic mutations, which supported the previous observation that patients with *KEAP1* reclassified benign mutations identified by CADD and FATHMM had significantly worse OS than patients without *KEAP1* mutations **(Figure 18B)** in highlighting the lower sensitivities of these methods in annotating *KEAP1* variants.

Similarly, all VEPs except for MutationAssessor were able to identify reclassified pathogenic mutations in *ATM* and *TP53* that were mutually exclusive with other oncogenic mutations in the TP53 pathway **(Figure 21A)**. The pattern of mutual exclusivity in *TP53* and *ATM* reclassified pathogenic mutations was not observed in reclassified benign mutations in the same genes **(Figure 21B)**.

Within the RTK/RAS pathway, the number of genes with mutually exclusive reclassified pathogenic mutations varied by methods. AlphaMissense identified reclassified pathogenic mutations in *NF1*, *ERBB4* and *IRS2* as mutually

exclusive with other oncogenic mutations in the pathway **(Figure 21A)**. However, reclassified benign mutations in other RTK/RAS pathway genes, including *ERBB4* and *IRS2*, were also mutually exclusive with known oncogenic mutations **(Figure 21B)**. This finding suggests that there are potential drivers in RTK/RAS pathway genes that require more thorough annotation to identify, potentially by integrating biological knowledge of the pathway with other computational approaches.

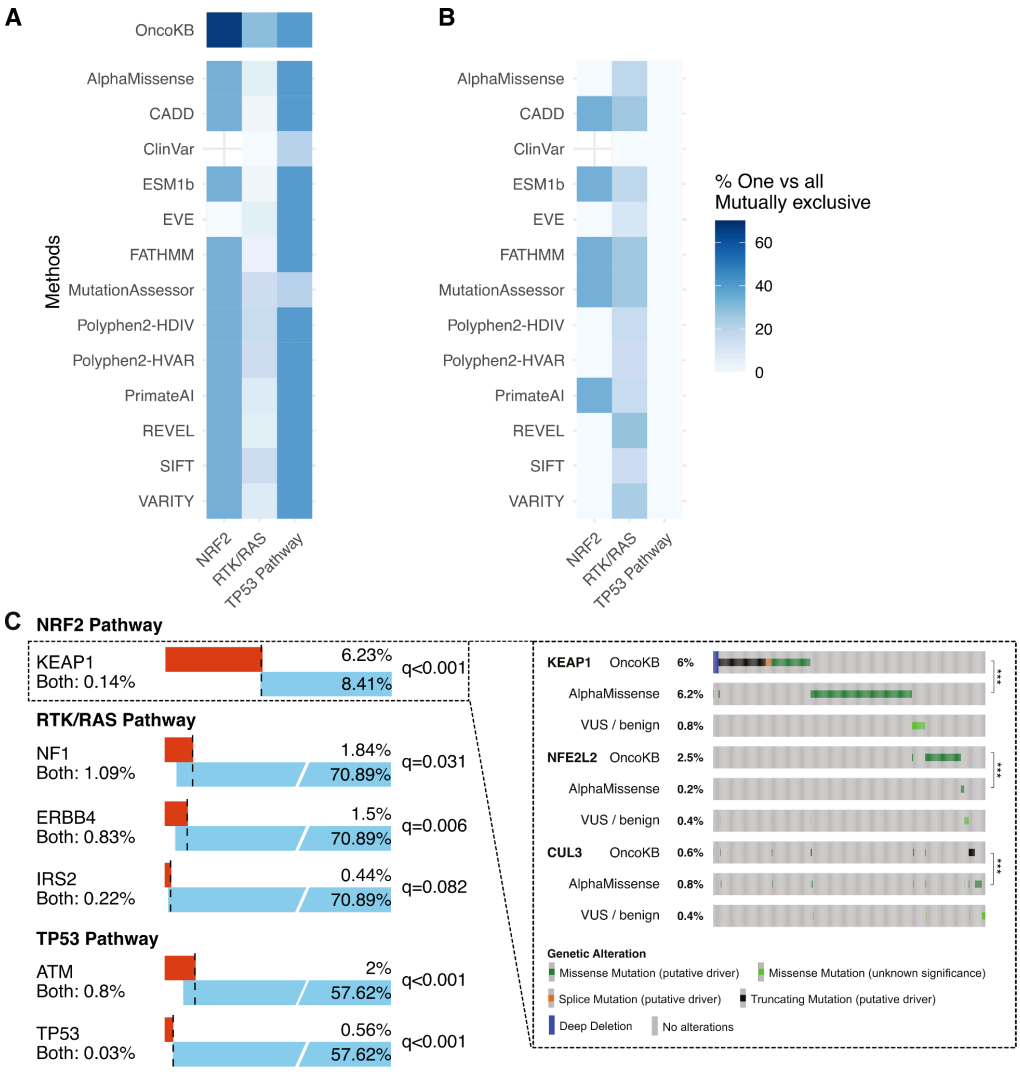

Figure 21. Mutual exclusivity of mutations in oncogenic signaling pathways

A-B. One-versus-all mutual exclusivity for each gene in a pathway is determined using a two-sided Fisher's exact test. Mutual exclusivity percent for each pathway is then calculated as the number of FDR-corrected significant tests divided by the total number of tests.

- **A.** One-versus-all mutual exclusivity between reclassified pathogenic mutations of each gene and known oncogenic mutations in all other genes in a given pathway.
- **B.** One-versus-all mutual exclusivity between reclassified benign mutations/VUSs of each gene and known oncogenic mutations in all other genes in a given pathway.
- **C.** Alteration frequencies and overlap of AlphaMissense reclassified pathogenic mutations (red) with oncogenic alterations of genes in the same pathway (blue) in MSK-IMPACT NSCLC cohort.
  Inset: Oncoprint of genes in the NRF2 pathway. KEAP1 reclassified mutations, similar to *KEAP1* oncogenic mutations, are mutually exclusive with other oncogenic mutations in *NFE2L2* and *CUL3*. See Supplemental Appendix for details.

We observed that VUSs in *ERBB4* and *IRS2* that were classified as pathogenic by AlphaMissense occurred frequently in samples with high TMB, whereas known oncogenic mutations in the RTK/RAS pathway occurred more frequently in TMB-low samples **(Figure 22A)**. To extricate the influence of TMB status on the observed mutual exclusivity, we used logistic regressions to describe the relationship between *ERBB4* or *IRS2* reclassified pathogenic mutations and other RTK/RAS oncogenic mutations while controlling for TMB status. We found that RTK/RAS oncogenic mutations were independently mutually exclusive with both reclassified pathogenic mutations in *ERBB4* and TMB-H status **(Figure 22B)**. *IRS2* reclassified pathogenic mutations were negatively associated with RTK/RAS oncogenic mutations when controlling for TMB-H status; however, this negative association was not significant, likely due to the small number of samples harboring these mutations. A similar pattern was observed with *ERBB4* or *IRS2* reclassified benign mutations **(Figure 22C)**,

further suggesting that there remained potential, unannotated drivers within these two genes.

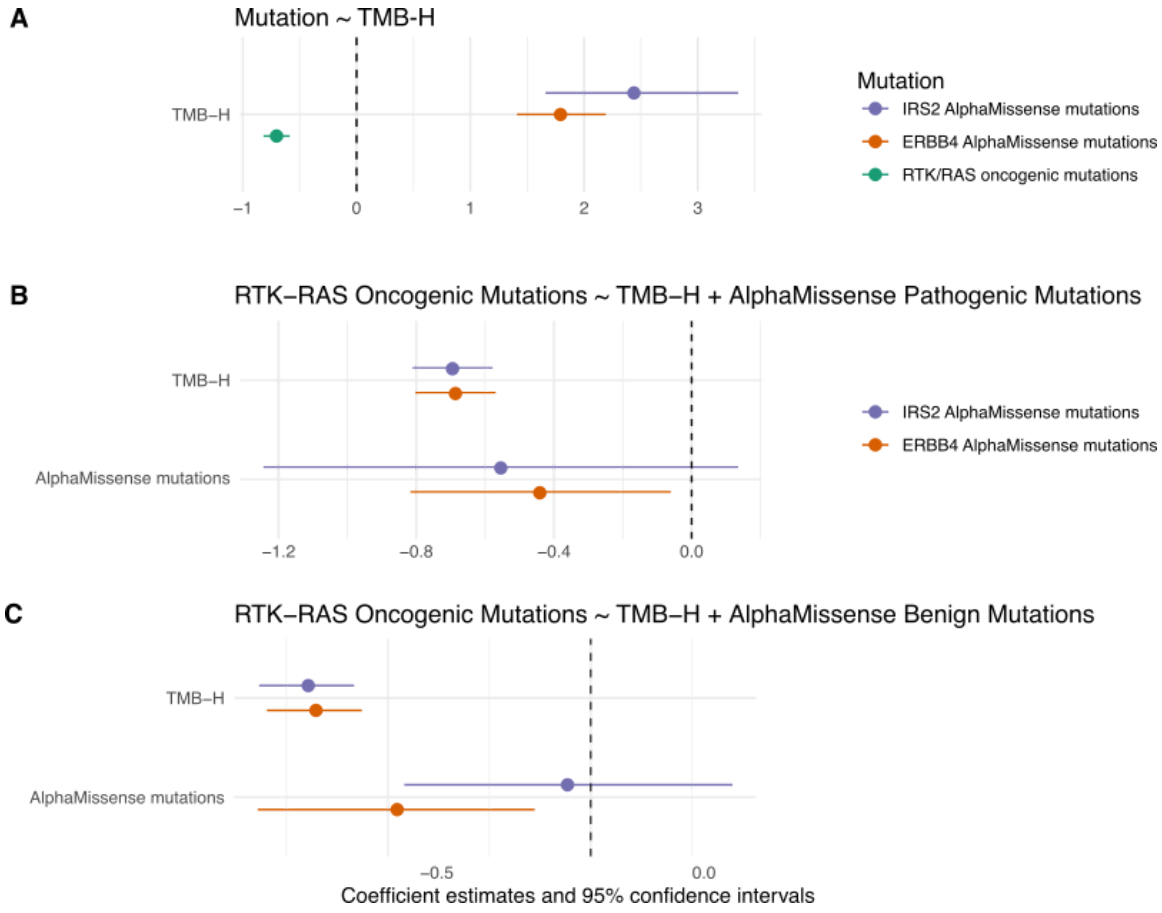

Figure 22. Mutations in *ERBB4* and *IRS2* are mutually exclusive with RTK/RAS oncogenic mutations independent of TMB-H status

- **A.** Coefficients for TMB-H status regressed on either RTK/RAS oncogenic mutations or AlphaMissense reclassified mutations in *ERBB4* and *IRS2*. AlphaMissense reclassified pathogenic mutations in *ERBB4* and *IRS2* are positively correlated with TMB-H status, suggesting they occur more frequently in TMB-H samples, whereas RTK/RAS known oncogenic mutations are observed frequently in TMB-low samples.
- **B.** Coefficients for TMB-H status and AlphaMissense pathogenic mutations in either *IRS2* or *ERBB4* regressed on RTK/RAS oncogenic mutations.
- **C.** Coefficients for TMB-H status and AlphaMissense benign mutations in either *IRS2* or *ERBB4* regressed on RTK/RAS oncogenic mutations.

To summarize, *KEAP1*, *ATM* and *TP53* reclassified pathogenic mutations, and not reclassified benign mutations, followed expected patterns of mutual exclusivity within their respective oncogenic pathways, which aligns with previous knowledge of these pathways[164] and supports their biological validity. The observed pattern of mutual exclusivity between RTK/RAS oncogenic mutations and *ERBB4* or *IRS2* reclassified pathogenic mutations may reflect the driver potentials of these reclassified mutations, but further investigations with larger sample sizes are warranted to confirm its significance. Furthermore, the mutual exclusivity pattern observed in reclassified benign mutations by some methods suggest that there are false negatives, suggesting a potential avenue for improvement. Taken together, we demonstrated that pathway mutual exclusivity can serve as an additional benchmark for VEP performance in predicting cancer driver variants.

Discussions

This study evaluated VEPs in annotating somatic cancer variants and characterized reclassified pathogenic mutations for their potential oncogenicity using multiple approaches, including their potential effects on ligand binding and protein-protein interactions, mutual exclusivity pattern with other oncogenic mutations in the same pathway, and prognostic effect on OS.

Leveraging a large pan-cancer multi-institutional cohort and known benign dbSNPs, we found that VEPs could effectively distinguish between benign and pathogenic mutations in cancer, with better accuracy for TSGs than OGs. Among four main classes of VEP approaches, ensemble and deep-learning methods had

consistently better performance compared to evolution-based approaches. The large number of features that ensemble and deep learning methods included in their models likely contributed to their superior performance and allowed them to predict pathogenicity of variants occurring in regions with low sequence homology and/or in proteins without known crystal structures. Among the ensemble methods, the two supervised VEPs trained on ClinVar or other human-curated labels outperformed CADD, which was trained on weak labels derived from population frequency. The improved performance of VARITY and REVEL compared to CADD could reflect upward bias due to data leakage, as their performances were evaluated on well-annotated variants that they were likely trained on. Despite the concern about data leakage bias affecting method evaluation, the superior performance of deep learning unsupervised methods or methods supervised by weak labels, including AlphaMissense, ESM1b, EVE and PrimateAI, suggests other viable and less biased alternatives for annotating novel variants. In particular, deep learning methods that explicitly incorporate both evolutionary and protein structure information, including AlphaMissense and PrimateAI, performed better than deep learning methods that only incorporate one modality of feature, suggesting that evolution and protein structure offer orthogonal information that is helpful for inferring variant pathogenicity.

We explored three different prongs to validate VEP predictions of VUS pathogenicity in cancer. First, structural analysis for proteins with available crystal structures revealed that mutations at binding residues were more likely to be pathogenic, offering functional evidence for certain reclassified pathogenic VUSs.

Second, we identified that reclassified pathogenic variants in *KEAP1* and *SMARCA4* were associated with worse OS in NSCLC. The hazard ratios of patients harboring reclassified pathogenic mutations relative to patients with no mutation were comparable to those of known oncogenic mutations and much higher than those of reclassified benign mutations in these genes, thus suggesting that VEPs were able to meaningfully distinguish between functionally different VUSs. These findings further contribute to the growing literature on the mutational landscapes and prognostic roles of *KEAP1*[160,161,165] and *SMARCA4*[166] in NSCLC, as well as emphasize the potential clinical relevance of VUSs in annotating variants in cancer. Finally, the study explored mutual exclusivity patterns, demonstrating that reclassified pathogenic mutations followed expected patterns within oncogenic pathways. The mutual exclusivity pattern followed by *KEAP1* reclassified pathogenic mutations, but not reclassified benign mutations, with other known oncogenic mutations in the NRF2 pathway further suggest the biological validity of *KEAP1* VEP annotations. Furthermore, we found that reclassified pathogenic mutations in *ERBB4* and *IRS2* were mutually exclusive with other known oncogenic mutations in the RTK/RAS pathway. However, multiple issues may complicate the interpretation of this finding. First, reclassified benign mutations in *ERBB4* and *IRS2* were also mutually exclusive with other RTK/RAS oncogenic mutations, suggesting that the VEPs had low sensitivity in annotating activating mutations in these two genes in particular, and in OGs in general. Second, *ERBB4* and *IRS2* VUSs occurred more frequently in samples with high TMB, and even though their mutual exclusivity with other RTK/RAS

oncogenic mutation remained significant when controlling for TMB status, the small number of samples with these mutations makes it difficult to conclude how TMB influenced the mutational patterns. Finally, few oncogenic *ERBB4* and *IRS2* mutations have been reported in NSCLC[167–169], making the interpretation of these novel variants less certain. However, the mutual exclusivity pattern observed in *ERBB4* and *IRS2* suggests that there are potentially pathogenic variants in these genes and thus helps prioritize reclassified pathogenic mutations occurring in samples without obvious oncogenic drivers for functional validation. Taken together, these findings underscore the potential of VEPs in identifying pathogenic mutations in cancer and propose new benchmarks for VEP performance in predicting cancer driver variants.

Our study has limitations. The genomic data from GENIE came from multiple contributing cancers with their own sequencing pipelines, the majority of which were tumor-only sequencing. Even though all data went through a rigorous germline SNP filtering pipeline before public release, it is possible that there remained private SNPs in the data, which may have artificially inflated the number of more easily characterizable VUSs in a given dataset, although this would reflect a clinical reality of such SNPs appearing in tumor-only sequencing assays. Furthermore, the different mutation calling pipelines used by each center can result in heterogeneity in which mutations were called and therefore analyzed in this study. The MSK-IMPACT and GENIE BPC cohorts, though richly annotated, remain small, and therefore may not be sufficiently powered to discover rare driver variants in less commonly mutated genes or assert the

association between these putative drivers and outcomes. Finally, the approaches we proposed to validate VEPs' predictions may not be applicable to all cancer types. Future work will focus on validating high-confidence reclassified pathogenic VUSs and identifying other potential benchmarks for VEP predictions, such as biallelic inactivation[170], which can further improve the utility of VEPs in studying cancer biology and discovering novel targets.

Methods

Patients and data collection

This study analyzed patients with tumor genomic sequencing from two sources: The MSK-IMPACT cohort and the AACR Project GENIE cohort, which includes patients from the MSK-IMPACT cohort.

GENIE

Details of the AACR Project GENIE cohort have been published previously[74]. In short, the pan-cancer registry contains genomic and clinical data from 11 international institutions. In this study, we analyzed data from the v.14-public release, which consists of genomic data for 183,302 tumors from 160,965 patients. For genomic landscape analyses involving gene-level counts, only patients with tumor sequencing panels including a given gene of interest were included in the respective analysis.

MSK-IMPACT

The MSK-IMPACT cohort comprised patients at Memorial Sloan Kettering (New York, NY), an academic cancer hospital with tumor genomic sequencing using MSK-IMPACT, an FDA-authorized tumor genomic profiling assay, which

uses matched white blood cell sequencing to filter clonal hematopoietic and germline variants. All MSK patients were enrolled as part of a prospective sequencing protocol (NCT01775072). The study was independently approved by the Institutional Review Board of each site. Patients provided written, informed consent and were enrolled in a continuous, nonrandom fashion. Data here is from a February 15, 2023 snapshot, consisting of 11,649 samples from 8,690 patients with NSCLC.

For patients in the MSK-IMPACT cohort, demographic and clinical information, including tumor stage, age, sex, race and histology were retrieved from the electronic health records database. Smoking history, prior treatment and metastatic events were abstracted from clinical notes using previously validated NLP methods[171–174]. Tumor mutation burden per sample was calculated as the total number of nonsynonymous mutations divided by the actual number of bases analyzed, and samples with TMB >= 10 mut/Mb were defined as TMB-high. MSI-H status was defined for each sample by an MSIsensor score >10 [175].

GENIE BPC

Genomic and clinical data for a subset of patients with non-small cell lung cancer in the AACR GENIE cohort have been recently published as part of the AACR GENIE Biopharma Collaborative (BPC)[37]. Out of 1,846 patients from four contributing institutions in the cohort, we included all 977 patients from Dana-Farber Cancer Institute, Vanderbilt-Ingram Cancer Center and Princess Margaret Cancer Centre-University Health Network with single-primary NSCLC in our analysis cohort. All MSK patients were included in the MSK-IMPACT cohort.

For survival analyses involving gene-level cohorts, only patients with tumor sequencing panels including a given gene of interest were included in the respective analysis.

Genomic landscape

Genomic data, including mutational calls, copy number alteration and structural variant data, for the GENIE v14-public cohort were obtained from Synapse. All genomic alterations were annotated with OncoKB version 4.2 (release date February 10, 2023). Genes were labeled as OGs or TSGs using the OncoKB Cancer Gene List (updated October 2, 2023). For genomic landscape analyses involving gene-level counts, only patients with tumor sequencing panels including a given gene of interest were included in the respective analysis.

Predicting known pathogenic variants

We evaluated the performance of 11 variant function prediction (VEP) methods and one human variant archive for recapitulating known oncogenic cancer variants as annotated by OncoKB, the first FDA-recognized somatic molecular knowledge database for this purpose. Methods were chosen to be included based on their novelty (including methods with recent release dates e.g. AlphaMissense, PrimateAI and ESM1b, as well as novel methodology to an old approach e.g. EVE's deep generative model to predict pathogenicity based on evolutionary conservation), superior performance compared to other methods in the same class (e.g. REVEL and CADD outperformed other VEPs in multiple comparison studies[94,95], while VARITY_R_LOO's performance was better than 12

other VEPs and comparable to AlphaMissense in certain evaluation tasks[109]), specific relevance to cancer (FATHMM implements cancer-specific pathogenicity weight[112] and MutationAssessor previously demonstrated utility in annotating cancer variants[107]), as well as historical significance and popularity (SIFT and PolyPhen2 are among the earliest VEPs and have the highest number of citations to date[96]). A brief description of the methods/database is presented in **Table 2**.

Non-oncogenic Variants

Non-oncogenic missense mutations were randomly selected from the dbSNP Human Variation Sets build 150 (April 2017) labeled as having “no known medical impact” (retrieved from https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive_1.0/2017/). This dataset includes variants with germline minor allele frequency of ≥ 0.01 and no records of clinical phenotypes in ClinVar. We randomly selected 10,000 variants and annotated them with the same methods as described below. These non-oncogenic variants served as the negative class in subsequent analyses.

Annotation schema

OncoKB annotations were performed with version 4.2 (released February 10, 2023). Prediction scores for all VEPs and ClinVar were obtained from dbNSFP[154] v4.6 (released February 18, 2024). Categorization of variants into pathogenic, non-pathogenic or uncertain, obtained by imposing predetermined cutoff thresholds on the predicted functional scores, was also provided by AlphaMissense, FATHMM, MutationAssessor, PolyPhen2 and SIFT. For VEPs in

this category, we used off-the-self classifications provided by these methods. The rest of the evaluated VEPs required additional steps to determine the appropriate variant classification from predicted scores.

First, EVE provides variant classifications at different degrees of uncertainty, aiming to maximize accuracy by excluding variants that the model is uncertain about. For example, a Class25 EVE classification means that 25% of the most uncertain variants were excluded when making predictions[113]. To identify the degree of uncertainty that would maximize accuracy in our data, we compared the AUROCs for classifying known oncogenic mutations from benign dbSNPs and selected the uncertainty threshold that resulted in the highest AUROC. Finally, for VEPs that provided prediction scores but not off-the-shelf classification or recommended score cutoffs for classifications, including CADD, ESM1b, PrimateAI, REVEL and VARITY, we manually identified cutoff thresholds to classify pathogenic and non-pathogenic mutations. To identify cutoff, for each method, we set up a binary classification task in which known oncogenic mutations from the dataset represented the positive class, and the benign dbSNPs represented the negative class. We then calculated the sum of sensitivity and specificity of each method at different thresholds and identified the optimal cutpoint as the threshold where this sum was maximized. In cases where more than one cutpoint were found, we used their median as the final threshold.

After cutoffs were determined, we applied them to the VEPs' predicted scores to stratify variants into categories. For CADD, PrimateAI, REVEL and VARITY_R_LOO, variants with scores higher than or equal to the determined

thresholds were classified as pathogenic and vice versa. For ESM1b, variants with scores smaller than or equal to the determined thresholds were classified as pathogenic.

We repeated this process for all datasets, resulting in data-specific cutoff and EVE uncertainty thresholds. The cutoffs used are summarized in **Table 6**.

Table 6. VEP prediction score cutoffs for variant classifications

	GENIE v14	**MSK-IMPACT NSCLC**	**Non-MSK GENIE BPC NSCLC**
EVE	Class90	Class90	Class90
CADD	24	24	23.6
ESM1b	-8.96	-9.11	-4.3
PrimateAI	0.67	0.67	0.6
REVEL	0.42	0.42	0.44
VARITY_R_LOO	0.48	0.48	0.66

Receiver operating curves (ROCs)

We performed two ROC analyses, 1. Weighting all positive variants equally and 2. sampling positive variants from the GENIE cohort, i.e. appearing proportionally to their frequency in a real-world population, to better understand method performance in a manner that reflects actual population levels of a given mutation. In the first analysis, mutation-level ROC for each method was constructed using 8,033 missense mutations as the positive class and 10,000 dbSNPs as the negative class. In the second analysis, population-level ROCs were constructed using all occurrences of missense mutations in the GENIE v14-public cohort as the positive class (N=246,401) and 10,000 dbSNPs as the negative class. The non-oncogenic mutations were upsampled to match the

number of oncogenic mutations in the positive class, resulting in a balanced N=246,401 for each class. Area under the curve and 95% confidence interval were calculated for each curve.

Ligand binding and protein-protein interaction residues analysis

Residues involved in binding ligands, including small molecules, peptides, DNA and RNA, were retrieved from BioLiP2[176], a curated database of biologically relevant protein-ligand interactions. Residues important in protein-protein interactions (PPI), termed PPI hotspots and defined as residues whose replacements decrease the binding free energy significantly, were retrieved from PPI-HotspotDB[177]. We grouped missense mutations in either the GENIE v14-public or MSK-IMPACT NSCLC cohort into binding residues (including ligand-binding residues and PPI hotspots) versus non-binding residues. Fisher's exact tests were performed to test the enrichment of mutations occurring at binding residues for being reclassified as pathogenic by different methods.

Pathway analysis

Mutual exclusivity test

Canonical oncogenic pathway level analyses were conducted using curated pathway templates as previously reported[164].

We aimed to identify whether reclassified pathogenic mutations are mutually exclusive with other known oncogenic mutations, which demonstrates that reclassified pathogenic mutations have comparably pathogenic effect on a pathway as known oncogenic mutations. To this end, we calculated one versus all mutual exclusivity for each gene for patients with NSCLC in the MSK-IMPACT

cohort. For each test, we first set up a 2x2 contingency table with two variables: the number of patients carrying reclassified pathogenic mutations in that gene, and the number of patients carrying oncogenic mutations in all genes within the same pathway. A two-sided Fisher's exact test was applied to the contingency table to test for mutual exclusivity.

An example contingency table used to test for pathway mutual exclusivity between *KEAP1* and all genes in the NRF2 pathway, including *KEAP1*, *CUL3* and *NFE2L2*, is below:

		Mutations in *KEAP1*	
		Reclassified pathogenic	VUSs/no mutation
Mutations in *KEAP1*, *CUL3* and *NFE2L2*	Known oncogenic		
	VUSs/no mutation		

In particular, this table was used in a Fisher's exact test for mutual exclusivity between *KEAP1* and all genes in the *NRF2* pathway. Two other tests were set up to test for mutual exclusivity of *NFE2L2* and *CUL3* with oncogenic mutations in NRF2 pathway genes.

The procedure was repeated for all genes present in a given pathway, and for all eight methods. P-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure. Tests with a logOR < 0 and adjusted p-value <= 0.1 were considered significant for mutual exclusivity. The pathway-level one versus all mutual exclusivity rate was then calculated for each

pathway by dividing the total number of significant mutually exclusive tests by the number of genes in the pathway.

As negative controls, we tested for mutual exclusivity between reclassified benign mutations and known oncogenic mutations in all genes in a given pathway using the same procedure.

The role of tumor mutational burden in observed mutual exclusivity

To identify whether *ERBB4* and *IRS2* AlphaMissense reclassified pathogenic mutations are mutually exclusive with RTK/RAS oncogenic mutations independent of TMB-high status, we performed a logistic regression:

`RTK/RAS oncogenic mutations ~ AlphaMissense mutations + TMB-H`

Where

`RTK/RAS oncogenic mutations` is 1 if a sample has any oncogenic mutations in RTK/RAS pathway genes, 0 otherwise

`AlphaMissense mutations` is 1 if a sample has an AlphaMissense reclassified pathogenic mutation in a gene of interest (*ERBB4* or *IRS2*), 0 otherwise

`TMB-H` is 1 if a sample has TMB >= 10 mut/Mb, 0 otherwise

Two independent regressions were run for *ERBB4* and *IRS2*.

Survival analysis

To test the association of gene-level pathogenicity annotations with overall survival we performed a series of Cox PH models from time of diagnosis to time of death or last follow-up, left truncated at time of cohort entry (tissue sequencing). For patients with multiple sequencing events, the first was used as the time of cohort entry. To adjust for confounding variables between comparison

groups, inverse probability of treatment weights (IPTW) were calculated using covariate values at baseline, including tumor stage, age, sex, race, histology, smoking history, tumor mutational burden (TMB), microsatellite instability status (MSI), prior treatment and metastatic sites if any, before fitting Cox PH models.

Hazard ratios, 95% confidence interval and p-values for gene-level associations between a "pathogenic" alteration vs no alteration were computed. In the MSK-IMPACT cohort, all genes altered in >2% of the cohort were considered. For each gene of interest, only patients with tumor sequencing panels including a given gene in the target region were included in the analysis. The following gene mutation annotation schema was used: For OncoKB, any alterations annotated as oncogenic or likely oncogenic were used. For all other databases, any pathogenic alteration considered a variant of unknown significance in OncoKB (i.e. a "reclassified" pathogenic alteration) was used. q-values were computed using the Benjamini-Hochberg method; a false discovery rate 0.1 across all comparisons described in this analysis was used to determine statistical significance according to a prespecified statistical analysis plan (see 12-245 Appendix C Project Plan). The GENIE BPC cohort was used as a confirmation dataset in which only significant associations from the MSK-IMPACT analysis were tested; q-values were computed similarly but for only the number of hypotheses tested in the BPC.

Kaplan-Meier curves

Weighted KM curves were constructed to further examine the relationship between gene-level pathogenicity annotations and overall survival for genes with

significant hazard ratios in univariate weighted Cox PH models. Similar to the Cox PH regressions, KM curves were calculated from time of diagnosis to time of death or last follow-up, left truncated at time of cohort entry (tissue sequencing). Patients were stratified based on the presence of OncoKB oncogenic, 'reclassified' pathogenic, 'reclassified' benign mutations or without any mutation in each gene of interest. Only strata with ≥ 10 patients were included. KM curves were weighted using the same IPTWs used for the corresponding Cox PH regression.

STK11/KEAP1 concurrent mutation analysis

To test whether AlphaMissense reclassified pathogenic variants in *STK11* and *KEAP1* were similarly associated with worse survival, we compared overall survival of patients with double *KEAP1* and *STK11* reclassified mutations with patients with single reclassified mutation and patients without any mutation in these two genes. Patients with reclassified pathogenic mutations in both *KEAP1* and *STK11* are labeled as *KEAP1*/*STK11* double mutants, while patients carrying only reclassified pathogenic mutations in either gene are labeled as single mutants for the respective genes. Patients carrying oncogenic mutations in either genes are excluded from this analysis. Weighted KM curves were constructed as described above to compare overall survival of patients in these four strata.

All statistical analyses were performed in R version 4.2.2 (2022-10-21).

CHAPTER 4

DISCUSSIONS & FUTURE WORK

Real-world data (RWD) has the potential to empower studies on cancer genomics and outcomes due to their feature and temporal complexities, but multiple challenges hinder their utility in research. In this work, we addressed two particular challenges that affect the use of clinical and genomic RWD: unstructured data fields and variant annotation, by developing and benchmarking the utility of machine learning models in automating data abstraction and predicting variant effects.

The lack of annotated RWD is a bottleneck in many studies, as it prevents the studies of important clinical variables. The amount of available genomic data is growing rapidly due to routine sequencing in clinical care, but the clinical data from these patients with available genomic data are not growing at the same rate due to the need for manual curation. Cohorts with both genomic and clinical data available remain small, with the AACR GENIE BPC projects exemplifying one effort to make harmonized clinicogenomic datasets publicly available for the research community. Many clinical variables, such as treatment, metastasis status and smoking history, provide context for the genomic alterations observed in sequenced samples and help explain much of the heterogeneity observed in cancer genomics. Therefore, without complete clinical and demographic information to accompany the genomic data, discoveries in cancer genomics could also remain limited. Furthermore, late fusion models incorporating multiple data modalities have shown significantly improved accuracy in stratifying patients

and predicting outcomes[178,179], further underscoring the need for automated curation of clinical data. In this work, we demonstrated that the Clinical-Longformer model fine-tuned on initial visit notes performed well in inferring whether patients had received treatment outside of MSK, achieving the AUC of 0.97. When deployed on a large cohort of patients at MSK, the model identified patients with putative external treatment who had similar genomic profiles and outcomes compared to patients with known treatment, further highlighting the importance of properly controlling for treatment status in analyses to avoid possible confounding effects.

Together with this work, other NLP models have been trained using GENIE BPC data and deployed on MSK patients to abstract a multitude of clinical variables, from metastatic sites to smoking status and progression, thus generating a richly annotated, harmonized dataset including RWD from patients with multiple cancer types that is many times larger than the original BPC cohorts[179]. We demonstrated that models incorporating genomic features and NLP-derived clinical variables, including external treatment, outperformed models trained on single classes of data in predicting overall survival[179]. Future studies will explore whether addition of more data modalities, such as liquid biopsy and laboratory data, which are known to have prognostic value[180] can further improve outcome prediction.

The advances in machine learning-based VEPs, particularly in ensemble models that integrate predictions from multiple methods and deep learning models that leverage latent information about evolutionary and structural context

embedded in nucleotide or amino acid sequences, have demonstrated to be fruitful in identifying cancer driver mutations. We showed that ensemble models and deep learning models, in particular AlphaMissense, outperformed homology-based models and supervised models in identifying known cancer driver mutations, achieving AUROCs as high as 0.977 in predicting population-level driver mutations in TSGs. Beyond identifying known drivers, VEPs can meaningfully discriminate putative drivers among other VUSs as demonstrated by enrichment of reclassified pathogenic VUSs in binding residues of proteins compared to benign ones. Furthermore, reclassified pathogenic VUSs in *KEAP1* and *SMARCA4* identified by several methods were associated with worse survival in two cohorts of patients with NSCLC, unlike reclassified benign VUSs. Finally, reclassified pathogenic VUSs exhibited mutual exclusivity with known oncogenic alterations at the pathway level, further suggesting biological validity.

VEPs have the potential to accelerate variant interpretation and prioritization for further functional studies. Future advances in predicting cancer driver mutations using VEPs need to focus on two main prongs: 1) improving prediction of driver mutations in OGs, and 2) validating predictions in the cancer context. Most VEPs do not distinguish between gain-of-function and loss-of-function mutations in their pathogenicity prediction and often struggle to predict gain-of-function driver mutations in OGs. We observed that all VEPs have lower AUROCs in predicting known driver mutations in OGs compared to in TSGs, and that despite VEPs' ability to identify *KEAP1* VUSs that were mutually

exclusive with known driver mutations in the NRF2 pathway, reclassified pathogenic VUSs in genes within the RTK/RAS pathway, which comprise mostly OGs, exhibited higher co-occurrence with known drivers. This result suggests that they may be weak drivers, as is the case with *NF1*[181], or inaccurately predicted to be pathogenic. The difficulty in predicting pathogenic mutations in OGs is likely due to the fact that the expected effect of activating mutations on protein structure may be milder than that of inactivating mutations and therefore harder to predict using just structural features[156]. For example, the mutation L858R activates the kinase activity of *EGFR* by disrupting interactions that stabilize its inactive conformation[182], which may be difficult to predict without knowledge of the different conformations of the protein. Therefore, predicting activating mutations would require better understanding of each protein's function, key regulatory residues within these proteins, their interaction partners and their roles within oncogenic signaling pathways. Initial work incorporating protein folding information from AlphaFold2 and protein-protein interactions from the human protein interactome shows promising performance in classifying gain-of-function and loss-of-function variants, but larger benchmarking sets and non-human-curated training data are needed to ensure that the method can generalize to new variants[183].

Improvement in computational approaches to predicting variant effects alone is not sufficient in accelerating cancer driver discovery, as each method will have its own drawbacks and predictions across methods may not agree, leading to difficulty and inconsistency in annotating VUSs based solely on computational

predictions. Therefore, the second prong in realizing VEPs' utility in cancer studies is devising a unified framework to validate high-confidence predictions in the context of cancer. In this work, we proposed that correlation with outcomes and mutual exclusivity with other known drivers are two important benchmarks in cancer to consider when validating VEP predictions. These benchmarks have demonstrated to be useful in providing further evidence to support VEP predictions; for example, we observed that VUSs predicted to be pathogenic in certain genes, including *KEAP1* and *SMARCA4*, correlated with worse outcomes and exhibited mutual exclusivity with other driver genes in the same pathway, suggesting that they are functionally different from the VUSs predicted to be benign. However, these benchmarks may not be universally useful to study VUSs across cancer types. Genomic alterations are not always clear biomarkers of prognosis, for example in high-grade serous ovarian cancer where gene expression signature and HRD subtypes are more prognostic than any single alteration[184]. Similarly, even though driver mutations in the same oncogenic pathways are mutually exclusive in NSCLC, the mutual exclusivity patterns are not alway observed in all pathways and cancer types. For example, *PIK3CA* and *RAS* mutations, which both activate the PIK3CA signaling pathway, can co-occur in colorectal cancer, even though they are mutually exclusive in NSCLC[185]. Therefore, cancer-type specific benchmarks to validate VEP predictions are needed to ensure that VEP can be used successfully to prioritize candidates in different contexts; for example, in cancer types where gene expression patterns are more prognostic than mutations, methods to reconstruct gene expression

from genomic profiles could be applied to identify how different VUSs affect gene expressions and subsequently patient outcomes[186]. By combining available real-world clinicogenomic data and knowledge about the biology of specific cancer types, it is possible to develop rigorous, scalable benchmarks to validate VEP predictions using various lines of evidence and prioritize putative drivers for functional characterization.

The potentials for real-world validated VEP predictions extend beyond the analyses in NSCLC presented here. Our understanding of NSCLC biology and the large amount of available data from patients with NSCLC allowed for proof-of-concept demonstrations that VEPs could identify putative drivers among VUSs. These analyses, however, can be extended to other cancer types that are more rare and/or lacking known driver mutations, which have the potential to accelerate discovery of driver and targetable alterations. Furthermore, VUSs annotated as putative drivers in targetable genes could potentially be used to enroll patients without known drivers in basket clinical trials, thus broadening the subset of patients who could potentially benefit from targeted therapies. Finally, similar to how NLP inference could assist with manual curation effort of clinical variables, VEPs could help with the curation effort of variant effect knowledge bases such as OncoKB. *In silico* predictions have been incorporated in the curation process of OncoKB, but further validations of VEP predictions using real-world data could provide stronger evidence for certain variants to be annotated and thus treated as oncogenic mutations. Finally, even though we used OncoKB labels as ground truth in this study to validate VEP predictions,

VEP predictions could also be compared to OncoKB labels to identify any discrepancies, especially false positives, thus improving the accuracy of the database.

In conclusion, our study highlights the potential of NLP models and computational VEPs to enhance the utility of oncology RWD. These methods enable the automated extraction of clinical data elements on a large scale and provide putative annotations for variants of uncertain significance (VUSs). We further demonstrated the effectiveness of utilizing outcome association and mutational patterns from RWD as benchmarks to evaluate the predictions made by VEPs in the context of cancer. Taken together, these results underscore the importance of developing methods and benchmarking frameworks to assess and improve the performance of both NLP models and computational VEP methods when leveraging oncology RWD. Looking ahead, ongoing refinement and expansion of these methodologies are crucial, given the evolving landscape of cancer research and the growing complexity of genomic data. The establishment of robust benchmarking frameworks will play a pivotal role in ensuring the accuracy and reliability of these tools, contributing to their broader application and impact in advancing our understanding of cancer biology and ultimately improving patient outcomes.

BIBLIOGRAPHY

1. Penberthy, L. T., Rivera, D. R., Lund, J. L., Bruno, M. A. & Meyer, A.-M. An overview of real-world data sources for oncology and considerations for research. *CA Cancer J. Clin.* **72**, 287–300 (2022).
2. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
3. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
4. Liu, R. *et al.* Systematic pan-cancer analysis of mutation-treatment interactions using large real-world clinicogenomics data. *Nat. Med.* **28**, 1656–1661 (2022).
5. Feinberg, B. A. *et al.* Use of Real-World Evidence to Support FDA Approval of Oncology Drugs. *Value Health* **23**, 1358–1365 (2020).
6. Lewis, J. R. R., Kerridge, I. & Lipworth, W. Use of Real-World Data for the Research, Development, and Evaluation of Oncology Precision Medicines. *JCO Precis. Oncol.* **1**, 1–11 (2017).
7. Kim, H. & Comey, S. LATE-BREAKING ABSTRACT: Two year overall survival rate of all advanced melanoma patients treated with ipilimumab in Australia 2013-2014. *Eur. J. Cancer* **72**, S129 (2017).
8. FDA's Sentinel Initiative | FDA. https://www.fda.gov/safety/fdas-sentinel-initiative.
9. Giles, F. J. *et al.* Mylotarg™ (gemtuzumab ozogamicin) therapy is associated with hepatic venoocclusive disease in patients who have not received stem cell transplantation. *Cancer* (2001).
10. Selby, C., Yacko, L. R. & Glode, A. E. Gemtuzumab ozogamicin: back again. *J. Adv. Pract. Oncol.* **10**, 68–82 (2019).
11. Arondekar, B. *et al.* Real-World Evidence in Support of Oncology Product Registration: A Systematic Review of New Drug Application and Biologics License Application Approvals from 2015-2020. *Clin. Cancer Res.* **28**, 27–35 (2022).
12. Wedam, S. *et al.* FDA Approval Summary: Palbociclib for Male Patients with Metastatic Breast Cancer. *Clin. Cancer Res.* **26**, 1208–1212 (2020).
13. Huang, L.-W. & Wang, S. Cancer clinical trial enrollment in older vs younger adults. *JAMA Netw. Open* **5**, e2235718 (2022).
14. Feinberg, B. A., Kish, J. K., Garofalo, D. & Mujumdar, U. Evaluation of real-world electronic medical record (EMR) treatment outcome data compared to clinical trial data for an advanced breast cancer treatment. *JCO* **34**, e12506–e12506 (2016).
15. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing

multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **22**, 114–126 (2022).

16. Huang, Y., Li, J., Li, M. & Aparasu, R. R. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC Med. Res. Methodol.* **23**, 268 (2023).
17. SEER Incidence Data, 1975 - 2020. https://seer.cancer.gov/data/.
18. Clegg, L. X. *et al.* Impact of socioeconomic status on cancer incidence and stage at diagnosis: selected findings from the surveillance, epidemiology, and end results: National Longitudinal Mortality Study. *Cancer Causes Control* **20**, 417–435 (2009).
19. Woods, L. M., Rachet, B. & Coleman, M. P. Origins of socio-economic inequalities in cancer survival: a review. *Ann. Oncol.* **17**, 5–19 (2006).
20. Vaccarella, S. *et al.* Socioeconomic inequalities in cancer mortality between and within countries in Europe: a population-based study. *Lancet Reg. Health Eur.* **25**, 100551 (2023).
21. Callahan, A., Shah, N. H. & Chen, J. H. Research and reporting considerations for observational studies using electronic health record data. *Ann. Intern. Med.* **172**, S79–S84 (2020).
22. Olier, I., Zhan, Y., Liang, X. & Volovici, V. Causal inference and observational data. *BMC Med. Res. Methodol.* **23**, 227 (2023).
23. Saesen, R. *et al.* Defining the role of real-world data in cancer clinical research: The position of the European Organisation for Research and Treatment of Cancer. *Eur. J. Cancer* **186**, 52–61 (2023).
24. Castelo-Branco, L. *et al.* ESMO Guidance for Reporting Oncology real-World evidence (GROW). *Ann. Oncol.* **34**, 1097–1112 (2023).
25. Levenson, M. *et al.* Biostatistical considerations when using RWD and RWE in clinical studies for regulatory purposes: A landscape assessment. *Stat. Biopharm. Res.* **15**, 3–13 (2023).
26. Hariton, E. & Locascio, J. J. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG* **125**, 1716 (2018).
27. Nørgaard, M., Ehrenstein, V. & Vandenbroucke, J. P. Confounding in observational studies based on large health care databases: problems and potential solutions - a primer for the clinician. *Clin. Epidemiol.* **9**, 185–193 (2017).
28. Schuemie, M. J. *et al.* How Confident Are We about Observational Findings in Healthcare: A Benchmark Study. *Harvard Data Science Review* **2**, (2020).
29. Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav. Res.* **46**, 399–424 (2011).

30. George, M., Smith, A., Sabesan, S. & Ranmuthugala, G. Physical Comorbidities and Their Relationship with Cancer Treatment and Its Outcomes in Older Adult Populations: Systematic Review. *JMIR Cancer* **7**, e26425 (2021).

31. Dimick, J. B. & Ryan, A. M. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA* **312**, 2401–2402 (2014).

32. Callaway, B. & Sant'Anna, P. H. C. Difference-in-Differences with multiple time periods. *J. Econom.* (2020) doi:10.1016/j.jeconom.2020.12.001.

33. Baker, A. C., Larcker, D. F. & Wang, C. C. Y. How much should we trust staggered difference-in-differences estimates? *J. financ. econ.* **144**, 370–395 (2022).

34. Bertrand, M., Duflo, E. & Mullainathan, S. How much should we trust differences-in-differences estimates? *Q J Econ* **119**, 249–75 (2004).

35. Pellat, A. *et al.* 1689O Comprehensive mapping review of real-world evidence publications focusing on targeted therapies in solid tumors: A collaborative work from ESMO real-world data and Digital Health Working Group. *Ann. Oncol.* **34**, S925 (2023).

36. Medicaid: data completeness and accuracy have improved, though not all standards have been met : report to Congressional addressees - Digital Collections - National Library of Medicine. https://collections.nlm.nih.gov/catalog/nlm:nlmuid-9918250407506676-pdf.

37. Choudhury, N. J. *et al.* The GENIE BPC NSCLC Cohort: A Real-World Repository Integrating Standardized Clinical and Genomic Data for 1,846 Patients with Non-Small Cell Lung Cancer. *Clin. Cancer Res.* **29**, 3418–3428 (2023).

38. AACR Annual Meeting 2021. Advancing Cancer Research Through An International Cancer Registry: AACR Project GENIE Use Cases. https://www.abstractsonline.com/pp8/#!/9325/session/783.

39. Miller, A. R. & Tucker, C. Health information exchange, system size and information silos. *J. Health Econ.* **33**, 28–42 (2014).

40. Alberto, I. R. I. *et al.* The impact of commercial health datasets on medical research and health-care algorithms. *Lancet Digit. Health* **5**, e288–e294 (2023).

41. Kong, H.-J. Managing unstructured big data in healthcare system. *Healthc. Inform. Res.* **25**, 1–2 (2019).

42. Yim, W.-W., Yetisgen, M., Harris, W. P. & Kwan, S. W. Natural language processing in oncology: A review. *JAMA Oncol.* **2**, 797–804 (2016).

43. Hossain, E. *et al.* Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. *Comput. Biol. Med.* **155**, 106649 (2023).

44. Sohn, S. *et al.* MedXN: an open source medication extraction and normalization tool for clinical text. *J. Am. Med. Inform. Assoc.* **21**, 858–865 (2014).

45. Jouhet, V. *et al.* Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods Inf. Med.* **51**, 242–251 (2012).

46. Jain, N. L. & Friedman, C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *Proc. AMIA Annu. Fall Symp.* 829–833 (1997).

47. Koh, Y., Yap, C. W. & Li, S.-C. Development of a combined system for identification and classification of adverse drug reactions: Alerts Based on ADR Causality and Severity (ABACUS). *J. Am. Med. Inform. Assoc.* **17**, 720–722 (2010).

48. Wu, H. *et al.* A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *npj Digital Med.* **5**, 186 (2022).

49. Wu, S. *et al.* Deep learning in clinical natural language processing: a methodical review. *J. Am. Med. Inform. Assoc.* **27**, 457–470 (2020).

50. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv* (2013) doi:10.48550/arxiv.1301.3781.

51. Wilcox, A. B. & Hripcsak, G. The role of domain knowledge in automating medical text report classification. *J. Am. Med. Inform. Assoc.* **10**, 330–338 (2003).

52. De Mulder, W., Bethard, S. & Moens, M.-F. A survey on the application of recurrent neural networks to statistical language modeling. *Comput. Speech Lang.* **30**, 61–98 (2015).

53. Sivakumar, S. *et al.* Review on word2vec word embedding neural net. in *2020 International Conference on Smart Electronics and Communication (ICOSEC)* 282–290 (IEEE, 2020). doi:10.1109/ICOSEC49089.2020.9215319.

54. Vaswani, A. *et al.* Attention is all you need. *arXiv* (2017) doi:10.48550/arxiv.1706.03762.

55. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* (2018) doi:10.48550/arxiv.1810.04805.

56. Alsentzer, E. *et al.* Publicly Available Clinical BERT Embeddings. *arXiv* (2019) doi:10.48550/arxiv.1904.03323.

57. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

58. Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H. & Luo, Y. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. *arXiv* (2022)

doi:10.48550/arxiv.2201.11838.

59. Jiang, L. Y. *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).

60. Biswas, B., Pham, T.-H. & Zhang, P. [2104.10652] TransICD: Transformer Based Code-wise Attention Model for Explainable ICD Coding. *arXiv* (2021).

61. He, Y., Zhu, Z., Zhang, Y., Chen, Q. & Caverlee, J. [2010.03746] Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. *arXiv* (2020).

62. Schmidt, L., Weeds, J. & Higgins, J. Data mining in clinical trial text: transformers for classification and question answering tasks. in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies* 83–94 (SCITEPRESS - Science and Technology Publications, 2020). doi:10.5220/0008945700830094.

63. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

64. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).

65. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat. Med.* **10**, 789–799 (2004).

66. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).

67. Druker, B. J. *et al.* Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344**, 1031–1037 (2001).

68. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004).

69. Raphael, B. J., Dobson, J. R., Oesper, L. & Vandin, F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.* **6**, 5 (2014).

70. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).

71. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

72. Sosinsky, A. *et al.* Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nat. Med.* (2024) doi:10.1038/s41591-023-02682-0.

73. Suehnholz, S. P. *et al.* Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. *Cancer Discov.* **14**, 49–65 (2024).

74. Pugh, T. J. *et al.* AACR project GENIE: 100,000 cases and beyond. *Cancer Discov.* **12**, 2044–2057 (2022).

75. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).

76. Bailey, M. H. *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371-385.e18 (2018).

77. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).

78. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).

79. Chang, M. T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).

80. Williams, M. J. *et al.* Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dN/dS ratios. *eLife* **9**, (2020).

81. Marks, J. L. *et al.* Prognostic and therapeutic implications of EGFR and KRAS mutations in resected lung adenocarcinoma. *J. Thorac. Oncol.* **3**, 111–116 (2008).

82. Johnson, M. L. *et al.* Association of KRAS and EGFR mutations with survival in patients with advanced lung adenocarcinomas. *Cancer* **119**, 356–362 (2013).

83. Wang, M., Herbst, R. S. & Boshoff, C. Toward personalized treatment approaches for non-small-cell lung cancer. *Nat. Med.* **27**, 1345–1356 (2021).

84. McFarland, C. D. *et al.* The damaging effect of passenger mutations on cancer progression. *Cancer Res.* **77**, 4763–4772 (2017).

85. Nussinov, R. & Tsai, C.-J. “Latent drivers” expand the cancer mutational landscape. *Curr. Opin. Struct. Biol.* **32**, 25–32 (2015).

86. Yavuz, B. R., Tsai, C.-J., Nussinov, R. & Tuncbag, N. Pan-cancer clinical impact of latent drivers from double mutations. *Commun. Biol.* **6**, 202 (2023).

87. Kumar, S. *et al.* Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. *Cell* **180**, 915-927.e16 (2020).

88. McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R. & Mirny, L. A. Impact of deleterious passenger mutations on cancer progression. *Proc Natl Acad Sci USA* **110**, 2910–2915 (2013).

89. Johnson, A. *et al.* Actionability classification of variants of unknown significance correlates with functional effect. *NPJ Precis. Oncol.* **7**, 67 (2023).

90. Guidugli, L. *et al.* Functional assays for analysis of variants of uncertain significance in BRCA2. *Hum. Mutat.* **35**, 151–164 (2014).

91. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).

92. Federici, G. & Soddu, S. Variants of uncertain significance in the era of high-throughput genome sequencing: a lesson from breast and ovary cancers. *J. Exp. Clin. Cancer Res.* **39**, 46 (2020).

93. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).

94. Wang, D., Li, J., Wang, Y. & Wang, E. A comparison on predicting functional impact of genomic variants. *NAR Genom. Bioinform.* **4**, lqab122 (2022).

95. Mahmood, K. *et al.* Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum Genomics* **11**, 10 (2017).

96. Katsonis, P., Wilhelm, K., Williams, A. & Lichtarge, O. Genome interpretation using in silico predictors of variant impact. *Hum. Genet.* **141**, 1549–1577 (2022).

97. Ostroverkhova, D., Przytycka, T. M. & Panchenko, A. R. Cancer driver mutations: predictions and reality. *Trends Mol. Med.* **29**, 554–566 (2023).

98. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

99. Chakravarty, D. *et al.* OncoKB: A precision oncology knowledge base. *JCO Precis. Oncol.* **2017**, (2017).

100. Holt, M. E. *et al.* My cancer genome: coevolution of precision oncology and a molecular oncology knowledgebase. *JCO Clin. Cancer Inform.* **5**, 995–1004 (2021).

101. Patterson, S. E., Statz, C. M., Yin, T. & Mockus, S. M. Utility of the JAX Clinical Knowledgebase in capture and assessment of complex genomic cancer data. *NPJ Precis. Oncol.* **3**, 2 (2019).

102. Zeng, J. *et al.* Operationalization of Next-Generation Sequencing and Decision Support for Precision Oncology. *JCO Clin. Cancer Inform.* **3**, 1–12 (2019).

103. OncoKB Curation Standard Operating Procedure. https://sop.oncokb.org/.

104. Chen, H. *et al.* Comprehensive assessment of computational algorithms in predicting cancer driver mutations. *Genome Biol.* **21**, 43 (2020).

105. Chang, M. T. *et al.* Accelerating discovery of functional mutant alleles in cancer. *Cancer Discov.* **8**, 174–183 (2018).

106. Gao, J. *et al.* 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med.* **9**, 4 (2017).

107. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).

108. Pauline, C. N. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).

109. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

110. Livesey, B. J. & Marsh, J. A. Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol. Syst. Biol.* **19**, e11474 (2023).

111. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).

112. Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. M. & Gaunt, T. R. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* **29**, 1504–1510 (2013).

113. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

114. Fariselli, P., Martelli, P. L., Savojardo, C. & Casadio, R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics* **31**, 2816–2821 (2015).

115. Quan, L., Lv, Q. & Zhang, Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics* **32**, 2936–2946 (2016).

116. Berliner, N., Teyra, J., Colak, R., Garcia Lopez, S. & Kim, P. M. Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS ONE* **9**, e107353 (2014).

117. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).

118. Bepler, T. & Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst.* **12**, 654-669.e3 (2021).

119. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* **118**, (2021).

120. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

121. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).

122. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

123. Ioannidis, N. M. *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885

(2016).

124. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021).
125. Schubach, M., Maass, T., Nazaretyan, L., Röner, S. & Kircher, M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res.* **52**, D1143–D1154 (2024).
126. Wu, Y., Li, R., Sun, S., Weile, J. & Roth, F. P. Improved pathogenicity prediction for rare human missense variants. *Am. J. Hum. Genet.* **108**, 1891–1906 (2021).
127. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinformatics* **12**, 41–51 (2011).
128. Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* **11**, 5918 (2020).
129. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
130. Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **139**, 1197–1207 (2020).
131. Sasidharan Nair, P. & Vihinen, M. VariBench: a benchmark database for variations. *Hum. Mutat.* **34**, 42–49 (2013).
132. Grimm, D. G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).
133. Notin, P. *et al.* ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. *BioRxiv* (2023) doi:10.1101/2023.12.07.570727.
134. Razavi, P. *et al.* The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers. *Cancer Cell* **34**, 427-438.e6 (2018).
135. Janjigian, Y. Y. *et al.* Genetic predictors of response to systemic therapy in esophagogastric cancer. *Cancer Discov.* **8**, 49–58 (2018).
136. Schoenfeld, A. J. *et al.* Tumor Analyses Reveal Squamous Transformation and Off-Target Alterations As Early Resistance Mechanisms to First-line Osimertinib in EGFR-Mutant Lung Cancer. *Clin. Cancer Res.* **26**, 2654–2663 (2020).
137. Murtuza, A. *et al.* Novel Third-Generation EGFR Tyrosine Kinase Inhibitors and Strategies to Overcome Therapeutic Resistance in Lung Cancer. *Cancer Res.* **79**, 689–698 (2019).
138. Salgia, N. J. *et al.* Characterizing the relationships between tertiary and community cancer providers: Results from a survey of medical oncologists

in Southern California. *Cancer Med.* **10**, 5671–5680 (2021).

139. Wolfson, J. A., Sun, C.-L., Wyatt, L. P., Hurria, A. & Bhatia, S. Impact of care at comprehensive cancer centers on outcome: Results from a population-based study. *Cancer* **121**, 3885–3893 (2015).
140. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
141. Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
142. Chen, C. D. *et al.* Molecular determinants of resistance to antiandrogen therapy. *Nat. Med.* **10**, 33–39 (2004).
143. Suda, K., Onozato, R., Yatabe, Y. & Mitsudomi, T. EGFR T790M mutation: a double role in lung cancer cell survival? *J. Thorac. Oncol.* **4**, 1–4 (2009).
144. El Kadi, N. *et al.* The EGFR T790M Mutation Is Acquired through AICDA-Mediated Deamination of 5-Methylcytosine following TKI Treatment in Lung Cancer. *Cancer Res.* **78**, 6728–6735 (2018).
145. Yu, H. A. *et al.* Poor response to erlotinib in patients with tumors containing baseline EGFR T790M mutations found by routine clinical molecular testing. *Ann. Oncol.* **25**, 423–428 (2014).
146. Xu, H. *et al.* Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annu. Symp. Proc.* **2011**, 1564–1572 (2011).
147. Ramos, J. Using TF-IDF to determine word relevance in document queries. (2003).
148. Mahajan, D., Liang, J. J., Tsou, C.-H. & Uzuner, Ö. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *J. Biomed. Inform.* **144**, 104432 (2023).
149. Hainaut, P. & Pfeifer, G. P. Somatic TP53 mutations in the era of genome sequencing. *Cold Spring Harb. Perspect. Med.* **6**, (2016).
150. Huang, L., Guo, Z., Wang, F. & Fu, L. KRAS mutation: from undruggable to druggable in cancer. *Signal Transduct. Target. Ther.* **6**, 386 (2021).
151. Shreenivas, A. *et al.* ALK fusions in the pan-cancer setting: another tumor-agnostic target? *NPJ Precis. Oncol.* **7**, 101 (2023).
152. Greuber, E. K., Smith-Pearson, P., Wang, J. & Pendergast, A. M. Role of ABL family kinases in cancer: from leukaemia to solid tumours. *Nat. Rev. Cancer* **13**, 559–571 (2013).
153. Scalera, S. *et al.* KEAP1-Mutant NSCLC: The Catastrophic Failure of a Cell-Protecting Hub. *J. Thorac. Oncol.* **17**, 751–757 (2022).
154. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103

(2020).

155. Martelotto, L. G. *et al.* Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol.* **15**, 484 (2014).
156. Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat. Commun.* **13**, 3895 (2022).
157. Nishi, H. *et al.* Cancer missense mutations alter binding properties of proteins and their interaction networks. *PLoS ONE* **8**, e66273 (2013).
158. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
159. Wood, K., Hensing, T., Malik, R. & Salgia, R. Prognostic and Predictive Value in KRAS in Non-Small-Cell Lung Cancer: A Review. *JAMA Oncol.* **2**, 805–812 (2016).
160. Shen, R. *et al.* Harnessing Clinical Sequencing Data for Survival Stratification of Patients with Metastatic Lung Adenocarcinomas. *JCO Precis. Oncol.* **3**, (2019).
161. Papillon-Cavanagh, S., Doshi, P., Dobrin, R., Szustakowski, J. & Walsh, A. M. STK11 and KEAP1 mutations as prognostic biomarkers in an observational real-world lung adenocarcinoma cohort. *ESMO Open* **5**, (2020).
162. El Osta, B. *et al.* Characteristics and Outcomes of Patients With Metastatic KRAS-Mutant Lung Adenocarcinomas: The Lung Cancer Mutation Consortium Experience. *J. Thorac. Oncol.* **14**, 876–889 (2019).
163. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406 (2012).
164. Sanchez-Vega, F. *et al.* Oncogenic signaling pathways in the cancer genome atlas. *Cell* **173**, 321-337.e10 (2018).
165. Saleh, M. M. *et al.* Comprehensive Analysis of TP53 and KEAP1 Mutations and Their Impact on Survival in Localized- and Advanced-Stage NSCLC. *J. Thorac. Oncol.* **17**, 76–88 (2022).
166. Schoenfeld, A. J. *et al.* The Genomic Landscape of SMARCA4 Alterations and Associations with Outcomes in Patients with Lung Cancer. *Clin. Cancer Res.* **26**, 5701–5708 (2020).
167. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
168. Chakroborty, D. *et al.* An unbiased functional genetics screen identifies rare activating ERBB4 mutations. *Cancer Res. Commun.* **2**, 10–27 (2022).
169. Kurppa, K. J., Denessiouk, K., Johnson, M. S. & Elenius, K. Activating ERBB4 mutations in non-small cell lung cancer. *Oncogene* **35**, 1283–1291

(2016).

170. Bielski, C. M. *et al.* Widespread selection for oncogenic mutant allele imbalance in cancer. *Cancer Cell* **34**, 852-862.e4 (2018).
171. Tran, T. N. *et al.* Abstract 4259: Identification of anti-neoplastic therapy given before initial visit at a referral center using natural language processing applied to medical oncology initial consultation notes. *Cancer Res.* **83**, 4259–4259 (2023).
172. Jee, J. *et al.* Abstract 5721: Automated annotation for large-scale clinicogenomic models of lung cancer treatment response and overall survival. *Cancer Res.* **83**, 5721–5721 (2023).
173. Luthra, A. *et al.* Abstract 1158: A.I.-assisted clinical data curation to determine genomic biomarkers of cancer metastasis. *Cancer Res.* **82**, 1158–1158 (2022).
174. Do, R. K. G. *et al.* Patterns of Metastatic Disease in Patients with Cancer Derived from Natural Language Processing of Structured CT Radiology Reports over a 10-year Period. *Radiology* **301**, 115–122 (2021).
175. Middha, S. *et al.* Reliable Pan-Cancer Microsatellite Instability Assessment by Using Targeted Next-Generation Sequencing Data. *JCO Precis. Oncol.* **2017**, (2017).
176. Zhang, C., Zhang, X., Freddolino, P. L. & Zhang, Y. BioLiP2: an updated structure database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **52**, D404–D412 (2024).
177. Chen, Y. C., Chen, Y.-H., Wright, J. D. & Lim, C. PPI-HotspotDB: Database of Protein-Protein Interaction Hot Spots. *J. Chem. Inf. Model.* **62**, 1052–1060 (2022).
178. Boehm, K. M. *et al.* Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat. Cancer* **3**, 723–733 (2022).
179. Jee, J. *et al.* Automated multimodal integration improves prediction of cancer outcomes. (2024).
180. Jee, J. *et al.* Overall survival with circulating tumor DNA-guided therapy in advanced non-small cell lung cancer. *JCO* **39**, 9009–9009 (2021).
181. Skoulidis, F. & Heymach, J. V. Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. *Nat. Rev. Cancer* **19**, 495–509 (2019).
182. Yun, C.-H. *et al.* Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell* **11**, 217–227 (2007).
183. Stein, D. *et al.* Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of a diverse feature set. *Genome Med.* **15**, 103 (2023).

184. Garsed, D. W. *et al.* The genomic and immune landscape of long-term survivors of high-grade serous ovarian cancer. *Nat. Genet.* **54**, 1853–1864 (2022).

185. El Tekle, G. *et al.* Co-occurrence and mutual exclusivity: what cross-cancer mutation patterns can tell us. *Trends Cancer* **7**, 823–836 (2021).

186. Ramchandran, M. & Baron, M. Reconstructing gene expression and knockout effect scores from DNA mutation (Mut2Ex): methodology and application to cancer prediction problems. *arXiv* (2022) doi:10.48550/arxiv.2212.05170.