Supplemental Information

for

Prognostic pan-cancer and single-cancer models: A large-scale analysis using a real-world clinico-genomic database

Sarah F. McGough, Svetlana Lyalina, Devin Incerti, Yunru Huang, Stefka Tyanova, Kieran Mace, Chris Harbron, Ryan Copping, Balasubramanian Narasimhan*, Robert Tibshirani*

* denotes equal contribution

| SI Tables & Figures | 2 |
|--|----------|
| Table S1. Summary of features included in prognostic models. | 2 |
| Table S2. Summary of prognostic model feature sets. | 3 |
| Figure S1-S5. Pan-cancer and single-cancer risk stratification plots for all cancer types | . 4 |
| Figure S6. Integrated Brier Score (IBS) for pan-cancer and single-cancer (A) Benchmark, (B) ROPRO-like, and (C) Full models. Lower IBS is indicative of better model calibration. Cancer types are arranged on the x-axis from largest to smallest sample size. | 9 |
| Figures S7-9. Top 25 clinico-genomic predictors in each single-cancer model. Predictor are ordered on the y-axis by descending coefficient (log hazard ratio). | rs 10 |
| Figure S10. Number of variables selected by each single-cancer model as a function of sample size. Cancer types are arranged on the x-axis from smallest to largest sample size. | f 13 |
| Figure S11. Comparison of left-truncated right-censored forest (LTRCF) model performance for single-cancer and pan-cancer training cohorts with respect to the (A) c-index and (B) integrated brier score (IBS). Cancer types are arranged on the x-axis from largest to smallest sample size. | 14 |
| Figure S12. Top 10 variants associated with node2vec dimensions. Displayed node2ve dimensions are a subset of the full 128 that were chosen at least once by any model. Association was determined by the BIOT method. | с 15 |
| Figure S13. Comparison of model performance between the stratified and non-stratified Cox models with respect to (A) c-index and (B) integrated Brier score (IBS). Cancer types are arranged on the x-axis from largest to smallest sample size. | ל 16 |
| SI Materials & Methods | 17 |
| Adaptation of the ROPRO prognostic model | 17 |
| Investigation of alternative model fitting approaches that support interactions. | 17 |

SI Tables & Figures

| | # Features | # Features | | | | | | |
|--------------------------|---------------|------------|----------------------------|--|--|--|--|--|
| Modality | (raw) | encoded) | Туре | Description | | | | |
| Clinical/dem ographic | 17 | 32 | Categorical, Continuous | Collection of baseline clinical/demographic information including: age, gender, race, BMI, ECOG, smoking status, cancer type, cancer stage at diagnosis, advanced or metastatic status at baseline, insurance type, socioeconomic status index, time from diagnosis to genomic test. | | | | |
| Labs and vital signs | 412 | 412 | Continuous | Time series summaries (e.g. median, min, max, proportion abnormal, slope, variance) of common labs and vital signs. | | | | |
| Treatment | 32 | 32 | Binary, Continuous | Indicators for each unique drug category received during frontline therapy and treatment at an academic center; continuous variable describing number of unique drugs received in frontline therapy, year of frontline therapy, time from diagnosis to first treatment. | | | | |
| Genomic | 1,511 | 1,513 | Categorical, Continuous | Binary alteration status (short variant "SV", copy number "CN", or rearrangement "RE") of genes; Raw variables are binary, while imputed ones are continuous due to the nature of KNN imputation (the resulting value reflects level of agreement between nearest neighbors). Variables consist of all combinations of HUGO gene symbol and variant type (SV, CN, RE) that were assayed by any Foundation Medicine test. Pathway affected status - biological pathways are designated as affected if any of their constituent genes have any kind of alteration. Node2Vec 128-dimensional embedding vector averages of all genes altered in a sample. Non-alteration related features derived from or associated with Foundation Medicine data: tissue tumor mutational burden (tTMB), tumor purity (computationally derived), PDL1 status, estimated ancestry (fractional assignment to 5 superpopulations) | | | | |
| Cancer-speci fic | 87 | 146 | Categorical, Continuous | Prognostic factors relevant to one or more cancer types, including: sites of metastases, | | | | |

| Table | S1. | Summary | of features | included in | prog | nostic models. |
|-------|-----|---------|-------------|-------------|------|----------------|
| | | | | | | |

| | | | | extranodal sites, disease subtypes, cancer-specific lab test results, disease-specific histologies. | | | |
|-------------------------------|---|-------|--|---|--|--|--|
| Total | 2,059 | 2,135 | | | | | |
| | | | | | | | |
| BMI: Body ma Organization. | Body mass index; ECOG: Eastern Cooperative Oncology Group; HUGO: HUman Genome nization. | | | | | | |

Table S2. Summary of prognostic model feature sets.

| Model | # Features | Description | Feature set |
|------------|------------|--|--|
| Benchmark | 9 | Collection of variables commonly collected in clinical practice | Cancer type, Age, Race, Gender, Smoking status, Baseline ECOG, Cancer stage at diagnosis, Time from diagnosis to frontline treatment, Time from diagnosis to genomic test |
| ROPRO-like | 29 | All variables included in the ROPRO model (Becker et al. 2020) with <= 30% missing in the database, plus time from diagnosis to genomic test (to adjust for delayed entry). | Cancer type, Age, Gender, Smoking status, Baseline ECOG, Cancer stage at diagnosis, BMI, Body weight, Body height, Heart rate, Hemoglobin, Systolic blood pressure, Diastolic blood pressure, Urea nitrogen, ALP, ALT, AST, Calcium, Creatinine, Total protein, Bilirubin, Albumin, Hematocrit, Glucose, Platelet count, Lymphocyte count, Monocyte count, Neutrophil count, Time from diagnosis to genomic test |
| Full | 2,059 | All variables derived from the clinico-genomic database. | All clinical and genomic predictors |

Figure S1-S5. Pan-cancer and single-cancer risk stratification plots for all cancer types.



Figure S1. Pan-cancer and single-cancer risk stratification plots for Non-Small Cell Lung, Colorectal, and Breast cancers.



Figure S2. Pan-cancer and single-cancer risk stratification plots for Ovarian, Pancreatic, and Gastric cancers.

Figure S3. Pan-cancer and single-cancer risk stratification plots for Prostate, Urothelial, and Melanoma cancers.





Figure S4. Pan-cancer and single-cancer risk stratification plots for Renal, Small Cell, and Head and Neck cancers.





Figure S6. Integrated Brier Score (IBS) for pan-cancer and single-cancer (A) Benchmark, (B) ROPRO-like, and (C) Full models. Lower IBS is indicative of better model calibration. Cancer types are arranged on the x-axis from largest to smallest sample size.



Cancer type (largest to smallest)









Figure S10. Number of variables selected by each single-cancer "full" model as a function of sample size. Cancer types are arranged on the x-axis from smallest to largest sample size.



Cancer type (smallest to largest)

Figure S11. Comparison of left-truncated right-censored forest (LTRCF) model performance for single-cancer and pan-cancer training cohorts with respect to the (A) c-index and (B) integrated brier score (IBS). Cancer types are arranged on the x-axis from largest to smallest sample size.



Figure S12. Top 10 variants associated with node2vec dimensions. Displayed node2vec dimensions are a subset of the full 128 that were chosen at least once by any model. Association was determined by the BIOT method.

| | 1 - | SMO SV | TMPRSS2 RE | FGF10 RE | PTCH1 SV | APC SV | TERT RE | EPHB4 SV | CD36 RE | ARFRP1 CN | HSD3B1 SV | | |
|--------|-------|-------------|-------------|------------|-----------|------------|-----------|------------|------------|------------|-----------|-----|--------------------------|
| | 3 - | APC SV | ALOX12B SV | ALK RE | ALK SV | SMARCA4 SV | CD22 SV | FGF14 CN | CYP2D6 SV | AR SV | RAD21 CN | | |
| | 4 - | SMAD4 SV | PDGFRA SV | ERBB3 SV | TAF1 SV | PDGFRB SV | TRRAP SV | IGF2R SV | SMARCA4 SV | PTCH1 SV | NF2 SV | | |
| | 5 - | TSC2 SV | GNAS SV | TET2 SV | KEL SV | RAD54L SV | BRAF SV | SOX9 SV | RUNX1T1 CN | RUNX1T1 SV | GNAS CN | | |
| | 6 - | TBX3 SV | NBN CN | RUNX1T1 CN | BCORL1 SV | CCNE1 CN | CDKN2A SV | NBN SV | ATR SV | RICTOR CN | SMAD4 SV | | |
| | 7- | KMT2D SV | KMT2C SV | KDM6A SV | FAT3 SV | APC SV | KMT2A SV | IDH1 SV | AR SV | TET2 SV | KRAS SV | | |
| | 9 - | FGFR2 SV | EGE19 CN | FGF4 CN | FGFR1 SV | EGER1 CN | FGF23 CN | EGER3 SV | FGF23 SV | EGE6 SV | FGE10 CN | | |
| | 13 | CIC SV | KELSV | CREBBP SV | PAK3 SV | MAP3K13 SV | TAE1 SV | SDHA CN | TMPRSS2 RE | NRASSV | ATR SV | | |
| | 1/1 | CDKN2A SV | CDKN24 CN | KMT2D SV | TRRAP SV | KMT2C SV | DOT1L SV | KMT24 SV | MUTVH SV | KAT6A SV | TYRO3 SV | | |
| | 15 | DTEN OV | | CDEN OV | RADOL CN | MED12 SV | IDEO EV | DIK2C2D OV | DTEN DE | PCOR SV | CATA2 SV | | |
| | 10 | PTEN SV | PTEN CN | OPICIND ON | RAD21 CN | TETR OV | 162 SV | PIK3C2B SV | | DECON SV | GATAS SV | | |
| | | DINITISK SV | EGFH SV | CDKN2B CN | | TET2 SV | HUST SV | JAK3 SV | PIK3C2G SV | PIPINIT SV | HBMITO SV | | |
| | | CUL4A CIN | DINMITSA SV | COL4A SV | CUL4D SV | CTP2D6 SV | INFI SV | IGFDH2 SV | MED 12 SV | PDGFRD SV | NNT2A SV | | |
| | 23 - | PIK3CA SV | FGF10 RE | TP53 SV | DDR2 SV | KLHL6 SV | KLHL6 CN | DDR2 CN | KRAS SV | KEAP1 SV | RETSV | | |
| | 25 - | NF1 SV | BRCA2 SV | PTEN SV | BAP1 SV | STAG2 SV | TSC2 SV | ATR SV | RAD21 CN | TERT SV | POLE SV | | |
| | 26 - | ATM SV | BCORL1 SV | CIC SV | DIS3 CN | DIS3 SV | BCOR SV | SOX9 SV | GNAS SV | SLIT2 SV | EMSY CN | | |
| | 31 - | KRAS SV | BRAF SV | NRAS SV | PIK3CA SV | ARAF SV | HRAS SV | MAP2K1 SV | KRAS CN | MUTYH SV | SMAD4 SV | | |
| | 34 - | FLT1 SV | VEGFA SV | PARP4 SV | VEGFA CN | KDR SV | FLT1 CN | PIK3CA SV | KMT2D SV | PARP4 CN | TEK SV | | |
| | 38 - | CDH1 SV | KRAS SV | PIK3CA SV | RAD21 CN | ARID1A SV | VHL SV | STAG2 SV | IGF2R SV | CDKN2A SV | AR SV | | |
| | 41 - | LRP1B SV | EGFR SV | ERBB2 SV | EPHB1 SV | ERBB2 CN | EGFR CN | TNKS SV | CYP2D6 SV | RNF43 SV | LRP1B RE | | |
| | 44 - | GABRA6 SV | CONE1 ON | MAGI2 SV | FANCA SV | PRSS1 SV | EPHA3 SV | TET2 SV | BCORL1 SV | ERBB3 SV | ZNRF3 SV | | |
| | 45 - | GNAS SV | GNAS CN | KMT2D SV | ATRX SV | INPP4B SV | BRAF SV | JAK3 SV | BCORL1 SV | NTRK3 SV | GATA3 SV | | |
| Б | 49 - | RNF43 SV | TYRO3 SV | VHL SV | EPHA6 SV | CHEK2 SV | IDH1 SV | SF3B1 SV | CIC SV | NOTCH1 SV | GRIN2A SV | | |
| lsic | 50 - | STK11 SV | BRAF SV | TBX3 SV | RNF43 SV | KEL SV | MPL SV | SPEN SV | MAP2K2 SV | STK11 CN | STK11 RE | | |
| Jer | 54 - | KMT2C SV | KMT2D SV | APC SV | JAK1 SV | JAK2 SV | KDM6A SV | PTEN SV | JAK3 SV | MPL SV | AR SV | Co | efficient of association |
| ng din | 55 - | SPTA1 SV | BRAF SV | ZNF217 CN | KDM5C SV | ERRFI1 SV | ABL1 SV | ZNF217 SV | BRIP1 SV | DOT1L SV | KEAP1 SV | fro | om BIOT W matrix |
| | 59 - | MYD88 SV | CARD11 SV | ERBB2 SV | MET SV | IRF4 SV | MST1R SV | TNFAIP3 SV | INPP4B SV | ERBB2 CN | ROS1 SV | | 0.025 |
| gi | 61 - | MAGI2 SV | APC SV | GRIN2A SV | ERBB4 SV | MTOR SV | PTEN SV | BCOR SV | GNAS SV | TET2 \$V | ATM SV | | 0.025 |
| pe | 63 - | RB1 SV | FOXL2 SV | NBN CN | RB1 CN | ATR SV | BRCA2 SV | FGF12 CN | TEK SV | MITF SV | TAF1 SV | | 0.000 |
| E | 73 - | BRCA2 SV | KEAP1 SV | BRIP1 SV | KDM5C SV | PALB2 SV | PMS2 SV | BRIP1 CN | CARD11 SV | FOXL2 SV | SOX9 SV | | -0.025 |
| ő | 74 - | BRCA2 SV | SPTA1 SV | FANCA SV | PDCD1 RE | FANCM SV | MUTYH SV | NBN CN | ABCC4 SV | BRIP1 SV | FANCD2 SV | | -0.050 |
| Š | 76 - | KEAP1 SV | PTPRD SV | KRAS SV | MTAP CN | ALK SV | ALK RE | TET2 SV | PALB2 SV | KEAP1 CN | CTNNB1 SV | | 0.000 |
| de | 79- | APC SV | CDH1 SV | ZNE217 CN | MAP2K4 SV | FBBB4 SV | SMO SV | MAP2K4 CN | DNMT3A SV | KEAP1 SV | EBBB3 SV | | -0.075 |
| ş | 83 | ZNEZ03 CN | ALK RE | ALKSV | CDH1 SV | MAP2KA SV | 7NE703 SV | EPHA6 SV | MAP2K4 CN | MAP3K1 SV | IDH1 SV | | |
| _ | 0.1 | CDKN2P CN | TERT SV | CTK11 CV | LIGE SV | EGE10 CN | APC SV | CONE1 ON | MET SV | EGERA SV | CDKECN | | |
| | 04 | CUEKO PV | TERT OF | BBCAD SV | FOED OV | | FOE10 PE | CONCTON ON | CDCN CV | FANIOE ON | CIDICO ON | | |
| | 00] | DDD LOW | DOTAL OV | BRUAZ SV | EGFN SV | BOOH SV | KDD OV | DODVO OV | SPEN SV | | OF OF OV | | |
| | 89 - | BRD4 SV | DOTIL SV | KAIBA SV | KM12A SV | KAIBA CN | KDH SV | P2H18 SV | FGFR2 SV | JAK3 SV | SF3BT SV | | |
| | 90- | ESR1 SV | MED12 SV | TRRAP SV | FBXW7 SV | CDK8 CN | CDK8 SV | LZTR1 SV | DNMT3A SV | MYCN SV | BRD4 SV | | |
| | 91 - | MTOR SV | RICTOR CN | RICTOR SV | NF1 SV | SPEN SV | RPTOR SV | TSC2 SV | TERT SV | BRAF SV | IRS2 SV | | |
| | 94 - | STAG2 SV | NF1 SV | KMT2D SV | RAD21 CN | ACVR1B SV | MYD88 SV | INHBA SV | MAP2K4 SV | RET SV | MAP3K1 SV | | |
| | 96 - | CHEK2 SV | MYC CN | FGFR2 SV | MYC SV | PREX2 SV | BAP1 SV | KDM5C SV | PREX2 CN | ARAF SV | TRRAP SV | | |
| | 98 - | RNF43 SV | KDM5C SV | RNF43 CN | DOT1L SV | ATRX SV | ROS1 SV | FBXW7 SV | STAG2 SV | DICER1 SV | TRRAP SV | | |
| | 99 - | NF1 SV | NF1 CN | NF1 RE | PRKDC SV | JAK3 SV | FAT1 SV | PTPN11 SV | PRDM1 SV | SMAD4 SV | RPTOR SV | | |
| | 102 - | ATM SV | RB1 SV | SF3B1 SV | CDH1 SV | MERTK SV | VHL SV | TET2 SV | GATA6 CN | RB1 CN | NFKBIA CN | | |
| | 105 - | DNMT3A SV | JAK3 SV | FLT4 SV | ROS1 SV | FGFR2 SV | PTPN11 SV | CCND3 CN | POLD1 SV | EP300 SV | WT1 SV | | |
| | 111 - | SDHA CN | SDHA SV | CTNNB1 SV | FLT4 SV | KDM6A SV | WT1 SV | KAT6A SV | GABRA6 SV | RANBP2 SV | MUTYH SV | | |
| | 112 - | CDKN2A CN | CDKN2A SV | APC SV | MYCN SV | MTAP CN | CCND1 CN | ARAF SV | SOX2 CN | WT1 SV | GRM3 SV | | |
| | 114 - | NOTCH1 SV | BRD4 SV | BCOR SV | MERTK SV | ESR1 SV | CARD11 SV | TGFBR2 SV | TYRO3 SV | PRKDC SV | MTAP CN | | |
| | 116 - | TP53 SV | NF1 SV | RAD21 CN | GRM3 SV | MDM2 CN | FGF10 RE | DNMT3A SV | NTRK3 SV | STAG2 SV | CDKN2A CN | | |
| | 117 - | BCL2L1 CN | JAK3 SV | KMT2A SV | MCL1 CN | CASP8 SV | MYC CN | INPP4B SV | RAD21 CN | KDM5C SV | BCL2 SV | | |
| | 119- | LRP1B SV | EGFR SV | CDK12 SV | PIK3CA SV | CHEK2 SV | ERRFI1 SV | CDK12 CN | RUNX1T1 CN | RUNX1T1 SV | NRAS SV | | |
| | 124 - | TMPRSS2 RE | CTNNB1 SV | TBX3 SV | MYC CN | ATR SV | AXIN1 SV | LRP6 SV | APC SV | PREX2 SV | PREX2 CN | | |
| | 125 - | MCL1 CN | EGFR SV | MYC CN | RNF43 SV | ASXL1 SV | AR SV | NOTCH4 SV | RAD54L SV | TNKS2 SV | MCL1 SV | | |
| | 126 - | KRAS SV | ROS1 SV | INPP4B SV | CARD11 SV | ERBB2 SV | ESR1 SV | ROS1 RE | INPP4B CN | ERBB2 CN | BRAF SV | | |
| | 127 | PREX2 SV | APC SV | MTOR SV | PREX2 CN | VHL SV | GNAS SV | TSC2 SV | ZNE703 CN | NOTCH2 SV | TERT SV | | |
| | 12/ 1 | CHEAZ SV | AFGSV | MICHSV | FREAZ ON | VIL SV | 01043-54 | 1802.50 | ZINF703 GN | NOTOH2 SV | TENT SV | | |

Top 10 original mutation variables used to explain embedding value

15

Figure S13. Comparison of model performance between the stratified and non-stratified Cox models with respect to (A) c-index and (B) integrated Brier score (IBS). Cancer types are arranged on the x-axis from largest to smallest sample size.



SI Materials & Methods

Adaptation of the ROPRO prognostic model

The ROPRO (Real wOrld PROgnostic score) was developed by by Becker et al⁹ and is composed of 27 clinical and demographic variables derived from the Flatiron Health electronic health record de-identified database.

We developed a "ROPRO-like" model, intended to adapt the ROPRO model using the available data in the clinico-genomic database (Table S2). However, our approach deviates in several ways:

- To remain consistent with the rest of our benchmark models, we allowed missingness in the selected variables of up to 30% whereas the ROPRO model allowed missingness up to 75%. As a result, our "ROPRO-like" model excludes several ROPRO variables with high rates of missingness including lactate dehydrogenase (LDH), chloride, oxygen, and eosinophils.
- 2. Where variables like "AST-to-ALT ratio" were not available, we included the individual laboratory components, AST and ALT.
- 3. The ROPRO model imputed missing data using a tree-based approach, whereas we used multiple imputation to remain consistent across our benchmark models.

Investigation of alternative model fitting approaches that support interactions.

The left-truncated right censored forests (LTRCF) R package^{73,74} was used within the same training and testing setup described for the Cox PH and Cox lasso models in the main results. The 'mtry' parameter was tuned via the LTRCforests::tune.ltrcrrf function (all arguments to which were set to their default values except time.eval, which was a grid from 0 to max observed time with granularity of 30 days. Default parameters: starting mtry = sqrt(number of variables), stepFactor = 2, ntreeTry = 100, bootstrap = "by.root", samptype="swor", sampfrac=0.632, nsplit = 10, nodesizeTry = max of either sqrt(number of observations) or 15. Values of the linear predictor were obtained from the resulting fit using the LTRCforests::predict.ltrcrfsrc method, and survival probabilities were calculated with the LTRCforests::predictProb method. These values were then used to calculate the final performance metrics: the concordance index and integrated Brier score.