Predicting High-Impact Pharmacological Targets by Integrating Transcriptome and Text-Mining Features

Anatoly Mayburd[1], Ancha Baranova[1, 2, 3]

[1] School of Systems Biology, College of Science, George Mason University, Fairfax VA 22003. [2] Federal State Budgetary Institution "Research Centre for Medical Genetics," Moskvorechie str, 1 Moscow, Russia. [3] Atlas Biomed Group, Malaya Nikitskaya 31, Moscow, Russia.

Received, August 20, 2016; Revised, August 26, 2016; Accepted, November 14, 2016; Published, November 15, 2016.

ABSTRACT - Purpose: Novel, "outside of the box" approaches are needed for evaluating candidate molecules, especially in oncology. Throughout the years of 2000-2010, the efficiency of drug development fell to barely acceptable levels, and in the second decade of this century, levels have improved only marginally. This dismal condition continues despite unprecedented progress in the development of a variety of high-throughput tools, computational methods, aggregated databases, drug repurposing programs and innovative chemistries. Here we tested a hypothesis that the economic impact of targeting a particular gene product is predictable *a priori* by employing a combination of transcriptome profiles and quantitative metrics reflecting existing literature. **Methods:** To extract classification features, the gene expression patterns of *a posteriori* high-impact and low-impact anti-cancer target sets were compared. To minimize the possible bias of text-mining, the number of manuscripts published prior to the first clinical trial or relevant review paper, as well as its first derivative in this interval, were collected and used as quantitative metrics of public interest. **Results:** By combining the gene expression and literature mining features, a 4-fold enrichment in high-impact targets was produced, resulting in a favourable ROC curve analysis for the top impact targets. The dataset was enriched by the highest impact anti-cancer targets, while demonstrating drastic differences in economic value between high and low-impact targets. Known anti-cancer products of *EGFR, ERBB2, CYP19A1*/aromatase, *MTOR, PTGS2, tubulin, VEGFA, BRAF, PGR, PDGFRA, SRC, REN, CSF1R, CTLA4* and *HSP90AA1* genes received the highest scores for predicted impact, while microsomal steroid sulfatase, anticoagulant protein C, p53, CDKN2A, c-Jun, and TNSFS11 were highlighted as most promising research-stage targets. **Conclusions:** A significant cost reduction may be achieved by *a priori* impact assessment of targets and ligands before their development or repurposing. Expanding a suite of combinational treatments could also decrease the costs, while achieving a higher impact per developed ligand.

INTRODUCTION

Despite tremendous historical progress in anti-cancer research and the decrease in mortality rates associated with many forms of cancer (1), including tumors of the prostate, breast, testis, and colon, as well as many forms of leukemia, the outcomes for pancreatic, lung and brain tumors remain dismal. It is difficult to pinpoint individual factors contributing to the organ-specific treatment success rates. For some forms of cancer, including breast, ovary, prostate, uterus, leukemia, thyroid, and testis, there is a tremendous gap in outcomes for patients with treatment-sensitive and treatment-resistant tumors. For other cancers, the role of targeted/chemotherapy remains secondary, while long-term remission is being achieved through a combination of radiotherapy and early radical surgery.

However, on average, a ten-year survival rate for all cancer forms increased from 22% in 1971 to 45% in 2007, with the most significant contribution to this increment being novel therapeutics. Health-improvements in breast cancer that could be

Corresponding Author: Dr. Ancha Baranova, Associate Professor, Director for the Center for the Study of Chronic Metabolic and Rare Diseases, School of Systems Biology College of Science George Mason University, 4400 University Dr David King Hall, Fairfax VA USA. E-mail: abaranov@gmu.edu; aancha@gmail.com

attributed to chemotherapy were estimated at between 14% and 27% in 1998 (2). It is likely that the more recently developed regimens show even greater impact. The broadly cited assessment of De-Vita et al. attributes 50% of the increase in survival rates to drug-based therapies (3). In this manuscript, we discuss a novel cost-effective strategy for developing anti-cancer therapeutics and their combinations. This strategy is based on a premise that the spectrum of efficient physiological anti-cancer mechanisms is relatively limited, and that improvement in survival rates is primarily due to the targeting of so-called "super-targets". Examples of well-known "super-targets" that were drugged since long time ago include DNA (4), folate reductase (5), and microtubules (6); these "super-targets" are still commonly tackled by either combinational therapy (7) or adjuvant therapy (8).

Discovery of a novel high-impact anti-cancer "super-target" is a significant event, further increasing cancer survival rates by approximately 1%, in our estimate. However, the existing statistics pertaining to the introduction of novel anti-cancer therapies point to stagnation (9-12). This trend occurs in the backdrop of Food and Drug Administration (FDA)'s attempts to expedite the review process (12). Even more concerning is that the downturn in the number of New Drug Applications (NDA) takes place at the time when the enabling technologies appear to explode (13-15), including the newest investments in personalized genome projects, gene knock-out techniques, toolboxes for biological network modeling, RNAseq, genome-wide association studies along with centralization and dissemination of biomedical information by NCBI and other portals. Another alarming trend is the cost of drug design, reaching $5bn per an approved drug in 2013 (16). The combination of the NDA contraction and the exponentially rising costs of drug development points to a wall of resistance that has to be penetrated.

In this report we demonstrate a possibility to predict that anti-cancer molecule would tackle a super-target *a priori*. In other words, we describe the methodology of the "success mining" that attempts to identify the features of known "winner" molecules at the preclinical stage, then to prioritize current candidate molecules according to relative resemblance of a "winner" profile. This approach would aid in reallocating available funding to the most promising candidates and minimize costly attrition at later development stages.

Some attempts to evaluate the pharmacological promise of a given target or its ligand have been made before. Ma'ayan et al. introduced graph-theory methods to analyze the FDA-approved drugs and their known molecular targets (17). Zhu et al. in (18) explored multiple factors that collectively contribute to druggability of various targets, including its protein sequence, structural, physicochemical, and systems profiles. Importantly, the techniques to explore each of these profiles for target identification have been developed, but they have not been collectively used. Chen et al. in (19) proposed that a disease-independent property of proteins, "drug-target likeness", can be explored to facilitate the genomic scale target screening. Sakharkar et al. in (20) described quantitative characteristics of the currently explored (those that are not yet associated with any marketed drug) and successful (targeted by at least one marketed drug) biomolecules; these characteristics were translated into simple rules for selecting a target with larger possibility of success. These rules highlight target proteins with 5 or less homologs outside of their own family, proteins encoded by single-exon gene architecture and proteins interacting with more than 3 partners as more likely to be druggable. Bender et al. in (21) reported a success mining approach applied to ligands in the context of *in vitro* interaction profiles of their targets. According to Bender et al., Preclinical Safety Pharmacology (PSP) approach may anticipate adverse drug reactions (ADRs) during early phases of drug discovery by testing compounds in relatively simple *in vitro* binding assays.

All of these previously implemented methodologies attempt to discriminate the targets that have already acquired a ligand from the targets that are either in the process of ligand acquisition or would never acquire an approved ligand. We find that this approach needs supplementing due to a number of considerations. First, some targets may eventually acquire a somewhat beneficial and a relatively harmless ligand that would pass safety and efficiency criteria, if a sufficient investment is made. Secondly, the targets that have acquired the approved ligand may lose the association with the approval if the ligand is pulled from the market later. Finally, the practical impact of ligands is in proportion to the significance of the target for the pathophysiological mechanism that drives a given pathology. A still developing candidate with expected huge clinical (and market) niche, but without approved associated ligands, may be more valuable than a comparable target with an approved ligand of a more modest impact.

In this report we attempt to predict an impact of a pharmacological target candidate using its future market share as a proxy. The more clinical trials and especially advanced clinical trials are conducted around a certain ligand, the more likely the target would eventually be tackled by a high-impact therapeutics that would be inherently successful due to a combination of favorable biology and pharmacology. Aiming at the largest possible target impact significantly differs from using the criterion of being simply FDA approval; such studies have not been conducted yet. Incorporation of the forecasted future impacts in the decision criteria put forth by funding and regulatory agencies may aid in creating a policy instrumental in rolling back the escalating costs of the drug development.

METHODS

Overview of methodology

The proposed impact forecasting technique relies on the target impact predictors available at the preclinical stage. The independent predictors were combined into an Index enabling to gauge the commercial potential of a target candidate before the bulk of investment is committed.

Sources of data included in the study

The study used GEO Datasets at NCBI at http://www.ncbi.nlm.nih.gov/gds/; PubMed at http://www.ncbi.nlm.nih.gov/PubMed/ and Therapeutic Target Database (TTD) at http://bidd.nus.edu.sg/group/cjttd/. The GEO Datasets was searched for similarly normalized gene expression data, collected from normal and cancerous specimens of various tissue origins. Dataset GSE7307 includes 677 normal and diseased human tissues profiled for gene expression using the Affymetrix U133 plus 2.0 array. The target status information was extracted from TTD, where the ligand and clinical trial status for each target candidate is specified as either “successful” (approved ligand is associated with the gene's products) or “research” (no approved ligand is associated with the gene's products). The text-mining was performed using PubMed with the gene names and their synonyms extracted from TTD.

Definition of target impacts

Target impacts were approximated by the number of clinical trials associated with the gene's name. The number of clinical trials associated with the gene was extracted by applying the PubMed filters. Weights 1, 2, 5 and 10 were assigned to the ligands in Phase I, II, III clinical trials and marketed ligands, respectively. The values of these weights were selected in proportion to the attrition rate of the ligands at each trial level. By resources committed, the ligand at the Phase I stage is cheaper by than a ligand that have reached the Phase III stage, and the approximate differences in the costs are reflected in the weights, see Supplemental Table 1 for the data. The number of the ligands in each category was multiplied by weights producing proxy target impacts that reflect the prospective revenue of the target reached in case of successful development. These values were defined as “real-life impacts”, while the predicted impacts were derived from both microarray and text-mining data. The proxies for real-life impacts were designated Y for the purpose of deriving a prediction rule as a linear classifier, see below.

Table 1. Modeled relative impacts for the successful anti-cancer targets.

Gene ID	Synonyms	Target impacts
EGFR	ERBB, ERBB1, HER1, PIG61, mENA, EGFR, epidermal growth factor receptor, avian erythroblastic leukemia viral (v-erb-b) oncogene homolog|	1254
ERBB2	CD340, HER-2, HER-2/neu, HER2, MLN 19, NEU, NGL, TKR1, ERBB2 v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2 c-erb B2/neu protein|herstatin|	1202
CYP19A1/ aromatase	ARO, ARO1, CPV1, CYAR, CYP19, CYPXIX, P-450AROM, CYP19A1 cytochrome P450, family 19, subfamily A, polypeptide 1 aromatase, cytochrome P-450AROM, cytochrome P450 19A1	620
MTOR	FRAP, FRAP1, FRAP2, RAFT1, RAPT1, mechanistic target of rapamycin (serine/threonine kinase) FK506 binding protein 12-rapamycin associated protein 2	259
PTGS2	COX-2, COX2, GRIPGHS, PGG/HS, PGHS-2, PHS-2, hCox-2, prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	246

Table 1. Continued…..

Tubuline	DAAP-285E11.4, M40, OK/SW-cl.56, TUBB1, TUBB5, tubulin, beta class I beta 5-tubulin|beta Ib tubulin|beta-4	227
VEGFA	RP1-261G23.1, MVCD1, VEGF, VPF, vascular endothelial growth factor A	220
BRAF	B-RAF1, BRAF1, NS7, RAFB1, BRAF v-raf murine sarcoma viral oncogene homolog B	162
PGR	NR3C3, PR, progesterone receptor nuclear receptor subfamily 3 group C member 3	123
PDGFRA	PDGFRA CD140A, PDGFR-2, PDGFR2, RHEPDGFRA, platelet-derived growth factor receptor, alpha polypeptide	95
SRC	RP5-823N20.1, ASV, SRC1, c-SRC, p60-Src, v-src avian sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog proto-oncogene c-Src	87
REN	HNFJ2, renin angiotensin-forming enzyme, angiotensinogenase, renin precursor, renal	84
CSF1R	CSF1R C-FMS, CD115, CSF-1R, CSFR, FIM2, FMS, HDLS, M-CSF-R, CSF1R colony stimulating factor 1 receptor CD115 antigen|CSF-1 receptor|FMS proto-oncogene	70
CTLA4	CD152, CELIAC3, CTLA-4, GRD4, GSE, IDDM12, CTLA4 cytotoxic T-lymphocyte-associated protein 4 CD152 isoform|celiac disease 3|cytotoxic T lymphocyte associated antigen 4 short spliced form|cytotoxic T-lymphocyte antigen 4	58
HSP90AA1	EL52, HSP86, HSP89A, HSP90A, HSP90N, HSPC1, HSPCA, HSPCAL1, HSPCAL4, HSPN, Hsp89, Hsp90, LAP2, HSP90AA1 heat shock protein 90kDa alpha (cytosolic), class A member 1	53
IGF1R	CD221, IGFIR, IGFR, JTK13, IGF1R insulin-like growth factor 1 receptor	51
RET	CDHF12, CDHR16, HSCR1, MEN2A, MEN2B, MTC1, PTC, RET-ELE1, RET51, ret proto-oncogene	47
CD52	CAMPATH-1 antigen, CDW52 antigen, HEL-S-171mP, cambridge pathology 1 antigen, epididymal secretory protein E5, epididymis secretory sperm binding protein Li 171mP, human epididymis-specific protein 5	39
IL2RA	RP1-261G23.1, MVCD1, VEGF, VPF, VEGFA, vascular endothelial growth factor A	39
RRM1	R1, RIR1, RR1, ribonucleotide reductase M1, ribonucleoside-diphosphate reductase large subunit	37
ADA	adenosine deaminase, adenosine aminohydrolase	33
STAT3	APRF, HIES,signal transducer and activator of transcription 3 (acute-phase response factor)	31
TOP2A	TOP2, TP2A, topoisomerase (DNA) II alpha, DNA gyrase, DNA topoisomerase (ATP-hydrolyzing, DNA topoisomerase 2-alpha, DNA topoisomerase II	30
TYMS	OK/SW-cl.29, HST422, TMS, TS, thymidylate synthetase, TSase	25
CXCR4	CD184, D2S201E, FB22, HM89, HSY3RR, LAP3, LCR1, LESTR, NPY3R, NPYR, NPYRL, NPYY3R, WHIM, CXCR4 chemokine (C-X-C motif) receptor 4	23
KDR	CD309, FLK1, VEGFR, VEGFR2, KDR kinase insert domain receptor	21
PPARG	CIMT1, GLM1, NR1C3, PPARG1, PPARG2, PPARgamma, peroxisome proliferator-activated receptor gamma	21
ABL1	RP11-83J21.1, ABL, JTK7, bcr/abl, c-ABL, c-ABL1, p150, v-abl, ABL1 c-abl oncogene 1, non-receptor tyrosine kinase Abelson tyrosine-protein kinase 1|bcr/c-abl oncogene	18
TLR7	RP23-139P21.3, toll-like receptor 7	14
LCK	RP4-675E8.4, LSK, YT16, p56lck, pp58lck, lymphocyte-specific protein tyrosine kinase	8
CCKBR	CCK-B, CCK2R, GASR, CCKBR , cholecystokinin B receptor, CCK-B receptor, CCK-BR, CCK2 receptor, CCK2-R, cholecystokinin-2 receptor, gastrin receptor	6
RXRA	NR2B1, RXRA retinoid X receptor, alpha nuclear receptor subfamily 2 group B member 1, retinoic acid receptor RXR-alpha, retinoid X nuclear receptor alpha	5
RRM2	R2, RR2, RR2M, ribonucleotide reductase M2, ribonucleoside-diphosphate reductase subunit M2	4
FYN	RP1-66H14.1, SLK, SYN, p59-FYN	3
ESR1	RP1-130E4.1, ER, ESR, ESRA, ESTRR, Era, NR3A1, ESR1 , estrogen receptor 1, ER-alpha, estradiol receptor, estrogen nuclear receptor alpha	2

Table 1. Continued…..

GNRH1	Gnrh, Gnrh2, LHRH, Lhrh1, Lnrh, hpg, Gnrh1, gonadotropin releasing hormone 1	2
SSTR2	somatostatin receptor 2 SRIF-1, SS2R, somatostatin receptor type 2	2
VDR	NR11, VDR vitamin D (1,25- dihydroxyvitamin D3), receptor 1,25-dihydroxyvitamin D3 receptor|nuclear receptor subfamily 1 group I member 1|	2
BDKRB2	B2R, BK-2, BK2, BKR2, BRB2, BDKRB2, bradykinin receptor B2 , BK-2 receptor	1
NTRK2	GP145-TrkB/GP95-TrkB, Tkrb, trk-B, trkB, neurotrophic tyrosine kinase, receptor, type 2 BDNF/NT-3 growth factors receptor	1
TOP1	RP3-511B24.1, TOPI, topoisomerase (DNA) I DNA topoisomerase 1	1
ALPL	AP-TNAP, APTNAP, HOPS, TNAP, TNSALP, ALPL alkaline phosphatase	0
CALM1	CALML2, CAMI, CPVT4, DD132, PHKD, caM, CALM1 calmodulin 1 (phosphorylase kinase, delta)	0
CSF2RA	CSF2RA CD116, CDw116, CSF2R, CSF2RAX, CSF2RAY, CSF2RX, CSF2RY, GM-CSF-R-alpha, GMCSFR, GMR, SMDP4, CSF2RA colony stimulating factor 2 receptor, alpha, low-affinity (granulocyte-macrophage)	0
DNMT1	ADCADN, AIM, CXXC9, DNMT, HSN1E, MCMT, DNMT1, DNA (cytosine-5-)-methyltransferase 1, CXXC-type zinc finger protein 9	0
EDNRA	ET-A, ETA, ETA-R, ETAR, ETRA, hET-AR, EDNRA endothelin receptor type A, G protein-coupled receptor, endothelin receptor subtype A, endothelin-1 receptor	0
EDNRB	RP11-318G21.1, ABCDS, ET-B, ET-BR, ETB, ETBR, ETRB, HSCR, HSCR2, WS4A, EDNRB endothelin receptor type B	0
FASN	FAS, OA-519, SDR27X1, FASN fatty acid synthase short chain dehydrogenase/reductase family 27X, member 1	0
FECH	EPP, FCE, FECH ferrochelatase	0
HDAC1	RP4-811H24.2, GON-10, HD1, RPD3, RPD3L1, HDAC1, histone deacetylase 1	0
HPSE	HPA, HPA1, HPR1, HPSE1, HSE1, heparanase, endo-glucoronidase, heparanase-1	0
IFNAR1	CD118, Ifar, Ifnar, Ifrc, Infar, Ifnar1 interferon (alpha and beta) receptor 1, IFN-R-1, IFN-alpha/beta receptor 1	0
IMPDH1	IMPD, IMPD1, LCA11, RP10, sWSS2608, IMP (inosine 5'-monophosphate) dehydrogenase 1	0
IMPDH2	hCG_2002013, IMPD2, IMPDH-II, IMPDH2 IMP (inosine 5'-monophosphate) dehydrogenase 2	0
ITGA2B	BDPLT16, BDPLT2, CD41, CD41B, GP2B, GPIIb, GT, GTA, HPA3, ITGA2B integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD4	0
LHCGR	HHG, LCGR, LGR2, LH/CG-R, LH/CGR, LHR, LHRHR, LSH-R, ULG5, LHCGR luteinizing hormone/choriogonadotropin receptor	0
MME	CALLA, CD10, NEP, SFE, MME, membrane metallo-endopeptidase atriopeptidase, common acute lymphocytic leukemia antigen	0
OXTR	OT-R, OXTR oxytocin receptor	0
PARP1	RP11-125A15.2, ADPRT, ADPRT1, ARTD1, PPOL, pADPRT-1, poly (ADP-ribose) polymerase 1 ADP-ribosyltransferase (NAD+);	0
PDE4A	PDE4A DPDE2, PDE4, PDE46, PDE4A, cAMP-specific 3',5'-cyclic phosphodiesterase 4A	0
PTH1R	PFE, PTHR, PTHR1, PTH1R parathyroid hormone 1 receptor, PTH/PTHr receptor|PTH/PTHrP type I receptor|PTH1 receptor|parathyroid hormone receptor 1	0
RARA	NR1B1, RAR, retinoic acid receptor, alpha	0
TSPO	RP3-526I14.4, BPBS, BZRP, DBI, IBP, MBR, PBR, PBS, PKBS, PTBR, mDRC, pk18, TSPO translocator protein (18kDa) benzodiazepine peripheral binding site	0
TXNRD1	GRIM-12, TR, TR1, TRXR1, TXNR, TXNRD1 thioredoxin reductase 1, KM-102-derived reductase-like factor	0

Microarray-based predictive features

The microarray data were retrieved by GEO NCBI at http://www.ncbi.nlm.nih.gov/gds/. The dataset GSE7307 is described above. Using this dataset, normal and cancer specimen pairs were formed for skin, lung, prostate, liver, uterus and ovary tissue environments. For each paired environment (cancer-norm), differential expression values were computed as:

$$D1 = [\text{cancer}]/[\text{norm in the same environment}] \quad (1)$$

$$D2 = [\text{cancer }]/[\text{average norm in all environments}] \quad (2)$$

Differential expression consistency metric was derived based on these primary data points. The direct data D1 and D2 were transformed in bonus-penalty data using the following rule:

$$\text{If } D1 > \text{cutoff1},\ D'1\ (\text{cutoff1}) = 2;\ \text{ If } D1 < \text{cutoff1},\ D'1\ (\text{cutoff1}) = 0 \quad (3)$$

$$\text{If } D2 > \text{cutoff2},\ D'2\ (\text{cutoff2}) = 2;\ \text{ If } D2 < \text{cutoff2},\ D'2\ (\text{cutoff2}) = 0$$

The transformed indirect values D'1 and D'2 were multiplied for each environment and the products were summed to produce the Differential Expression Consistency Score (DEXCON).

In this system, the leading scores were assigned to the genes that show high expression levels in cancer, but low expression levels in the normal environment related to the tumor as well as low expression levels in all unrelated normal environments. It is apparent that the potential targets with such parameters would possess wider therapeutic windows under all other circumstances being equal.

The absolute levels of gene expression were measured across the panel of 90 tissue environments and this feature was termed INTENSITY, to reflect intensity of absolute mRNA transcript expression.

The parameters DEXCON and INTENSITY became the features embed in an integrated classifier and were designated as X1 and X2 for the future reference. See the specific values of the parameters and corresponding bonus-penalty points in the Supplemental Table 2.

Text-mining predictive features

The future target impacts were anticipated by extracting the levels of early scientific interest as measured by the number of non-review research publications available prior to the first review published and by first derivative of the research interest. This extraction was accomplished by querying the PubMed with gene name and all its synonyms followed by manual review of the result to ensure that all selected articles are relevant to the biology of the target and the resultant therapeutic avenues.

The derivative was measured as the ratio of the

$$ABS\ (NT – NR)/(\text{Spacing}) \quad (5)$$

where NT – is the number of non-review publications addressing the role of the gene in the disease of interest prior to the date of the first clinical trial inception; NR – is the number of non-review publications addressing the role of the gene in the disease of interest prior to the date of the first review published; Spacing – is the number of years between first clinical trial and first review. The function ABS is the absolute value operator and it accounts for the fact that the first review and the first clinical trial may follow in any order.

$$\text{The average } N = (NT + NR)/2 \quad (6)$$

The average measures the absolute number of peer-reviewed research publications related to the gene. All numbers were normalized for the natural growth of PubMed population in time, by the formula:

$$N\ (T2)/N(T1) = 1.045\text{^}(T2\text{-}T1) \quad (7)$$

The features (5) and (6) were designated as X3 and X4 for combining them within the integrated classifier.

Classifier design

To design a linear classifier locating high-impact targets, the values of Y were transformed as

$$Y' = \log\ (Y + 1) \quad (8)$$

The purpose of transform was to smooth the data-set by relatively diminishing the effects of a few very high Y values, dominating the numerical structure. The smoothing allows effective increase of diversity in the training set and is equivalent to the increased training set size, facilitating a more objective training process. The correction term + 1 in (8) accounts for zero Y values not amenable to log transform. The distortion introduced by + 1 correction is minimal and does not outweigh the benefit of the smoothing procedure. The features X1-X4 were ranked. The Y' was related to the ranked X1-X4 by a linear regression:

$$YP' = W1X1 + W2X2 + W3X3 + W4X4 + A \qquad (9)$$

where W1-W4 are the corresponding weight coefficients to be determined by an error minimization procedure:

$$\text{Sum } (Y' - YP')\^{}2 = \text{MIN } (10)$$

where YP' are the predicted impacts for the entire population of the training set, Y' are the above defined “true impacts” for the entire population of the training set. The set of training weights [W1-W4, A] was determined by minimizing (10) using the Least Square Method.

Further steps taken to improve the resolution of the method included the following. The ranked values of X1-X4 were each scored in the following manner: the top rank [0.9-1.0] received a score of 2, the next bin [0.75-0.89] received a score of 1 and all bins below 0.75 received score of 0.

The direct column vectors of features were replaced with the bonus-penalty values as defined here. The error minimization procedure was repeated and the proportions between the weight coefficients W remained practically unchanged.

The meaning of the bonus-penalty transform is to emphasize the role of informative outliers on the side of the model factors X (as opposed to the outputs Y) and to smooth the effects of the random noise, distributed equally among the members of the training set. To re-phrase, the bonus-penalty system selects only the most informative genes for contributions in the prediction rule and maximizes the signal-to-noise ratio at the given size of the training set.

Validation of selected approach

The list of targets was randomly divided in the training and testing sets of equal size. The testing set was set apart until the prediction rule was derived in the training set. Based on the derived prediction rule, the residual errors were computed in the training and test sets by comparing the predicted and real impacts. The populations of the residual errors (10) were compared for the training and testing sets to assess generalization by the prediction rule. We tested for the statistical equality of the error populations in the absence of over-fit. To select a proper T-test form (equal or unequal variance, 2-tail), a preliminary F-test was run to compare variances. The F-test reported equal variances between the error populations and based on these data equal variance T-test was applied for population comparison. The populations of errors were identical, with no over-fitting detected. Based on this conclusion, the training and testing populations were merged for the plotting of Receiver Operating Characteristic (ROC) curve. Ranking of real impacts were used for defining high and low real impact categories and respective labeling of the targets. At the next step, predicted target scores were ranked, and the distribution of real score labels was traced as a function of the predicted score. The bin with high predicted score on the ROC curve provided significant enrichment to the real high impact labels, thus, validating our approach.

ROC curve plotting and its use for computing relative enrichment

The true impacts Y were subdivided based on rank in the “high-impact” bin with the ranks [0.75-1.0] and “low-impact” bin with the ranks [0 – 0.75]. The members of these groups acquired the positive and negative labels respectively. The predicted scores PY' were ranked as well, and the population of true impacts followed the rank of PY', producing a non-ideal, but a generally correlating pattern. The true “high-impact” labels were predominantly concentrated in the higher regions of PY' rank. The predicted score ranks were explored from the top (1.0) to the bottom (0.0) values.

The fraction of high-impacts f1 was computed by summarizing the positive labels as defined above. The fraction of low-impacts f2 was computed in a similar fashion.

The fraction f1 of “high-impact” Y and the fraction f2 of “low-impact” Y were forming the Y-axis and the X-axis of the plot, respectively. Per each 0.1 (10%) increment of “low-impact” count, the fractional increment of “high-impact” targets was also computed. Every point on ROC curve can be represented in the coordinates [summary fraction of “low-impact” values; summary fraction of “high-impact” values], the summary fraction is the sum of all increments over the previous intervals. To exemplify, the “low-impact” summary fraction 0.1 + 0.1 + 0.1 = 0.3; the matching value of the “high-impact” summary fraction becomes 0.5 + 0.1 + 0.05 = 0.65.

The ratio of the summary fractions characterizes the relative enrichment in the true “high-impact” values as a function of the predicted impact rank. To exemplify, in the highest 0.1 fraction of the predicted impact rank corresponding to the left-most part of the ROC curve and [1.0-0.9] bin of the PY' rank, the summary fraction of the “high-impact” values is 0.5, therefore the relative enrichment is 0.5:0.1 = 5. Considering a predicted impact bin of rank [1.0-0.8], the summary fraction of the “high-impact” values is 0.5 + 0. 1 = 0.6, while the summary fraction of the “low-impact” values is 0.1 + 0.1 = 0.2, therefore the relative enrichment is 0.6:0.2 = 3. On comparative basis, using the top rank bin of the PY' rank produces 5-fold higher chance to encounter a true “high-impact” label than using the population before the computational filter was applied.

RESULTS

A priori evaluation of future economic impact of a putative anti-cancer targets

Figure 1 illustrates the attempt to predict target impacts based on the combination of DEXCON (X1) and INTENSITY (X2) gene expression features, extracted from the microarray data by the above-described methodology. Figure 2 illustrates the attempt to predict target impacts based on text-mining features. Figure 3 illustrates the attempt to predict target impacts based on the combination of DEXCON and INTENSITY gene expression features as well as the text-mining features. The ROC curves demonstrate non-zero area between the diagonal baseline, which reflects the ratio of false positive to false positive summary functions in each bin, and the thicker upper line which reflects the ratio of true positive and false positive summary functions in each bin. The ratio of true positives to the false positives was significantly higher than the baseline, as it is especially evident for the left corner of the ROC plot that describes the highest range of the predicted scores. The regions with higher predicted scores embed the majority of real-life high impact targets, while the regions with lower predicted scores are depleted in real-life high impact targets. These Figures point to the possibility of predicting high-impact target category *a priori,* already at the stage of preclinical development and before the onset of the most expensive clinical trial phase. It is very unlikely that the inherent biological mechanism determining the target's future impact at the preclinical development stages remain obscure. This mechanism leaves its signature in a variety of large-scale high-throughput studies as well as in collective research activity patterns. The more diverse sources of information are incorporated and the more the prediction point is shifted away from an onset of active clinical trial stage, the lesser the role of “me-too” bias factor in the emergence of the detected patterns.

The Tables 1 and 2, respectively, show the sets of FDA-approved and “still-in a-pipeline” anti-cancer targets with the predicted impacts identified on the above-described basis. The impact leaders on the side of the targets with the approved ligands are *EGFR, ERBB2, CYP19A1, MTOR, PTGS2, tubulin, VEGFA, BRAF, PGR* and *PDGFRA.* The functions and cancer-related status of the genes were explored using database Genes at NCBI. The predicted impact leaders demonstrate favorable biological anti-cancer features due to their central roles in more universal pathophysiological mechanisms. Thus, many forms of cancer depend on overexpression of EGFR and ERBB2 for their survival. Blockade of these kinases synergizes with cytotoxic therapeutics. In normal cells, such dependence is rare or absent; therefore, the combinational regimens based on EGFR and ERBB2 have an access to a broad therapeutic window. The aromatase CYP19A1 is a key enzyme in estrogen synthetic pathway and is selectively important for the cancer subtypes that rely on estrogen stimulation for growth and survival. Mammalian target of rapamycin (mTOR) regulates the functions of cell survival, motility, proliferation, protein synthesis and transcription, making this target extremely important for rapidly propagating cells characterized by destabilized metabolism. PTGS2 (cyclooxygenase-2) is among the most important mediators of inflammation. Consequently, this target is indispensable for the growth stimulation produced by stromal immune cells and for metastasis. Many tumor types directly depend on prostaglandin stimulation. Tubulins are selectively more important for rapidly dividing cells undergoing mitotic process. VEGFA pathway is necessary for neovascularization of tumor foci. VEGFA also produces autocrine stimulation of multiple pro-cancer survival pathways, thus, its inhibition selectively affects almost all cancers. BRAF (serine-threonine kinase B-Raf) is a proto-oncogene with a broad transforming potential and, therefore, the blockade of its product is selectively more important for the cells where BRAF is constitutively activated. Progesterone receptor (PGR) is selectively more important for ovarian, mammary and endometrium cell populations that depend on progesterone for their growth and survival. Finally, PDGRFA (platelet-derived growth factor receptor, alpha polypeptide) is the powerful mitogen indispensable in certain tissues. To summarize, all genes demonstrating high predicted impacts also demonstrate highly selective involvement in specific cancer types and lack of such involvement in a majority of normal tissues. These differential roles and the ability to bind relatively non-toxic ligands explain their observed ranks presented in Table 1.

The sets of targets presented in the Table 2 are not yet drugged by suitable ligands. The target STS (steroid sulfatase, EC 3.1.6.2) is a member of a steroid synthesis pathway that is capable to produce in selected cancer cell populations the same dependence as other steroid pathway mediators, explaining its relatively high predicted impact. PROC (protein C) is actively involved in blood anti-coagulation pathways, stimulates cell migration and influences the secretory behavior of tumor cells, while suppressing NK killers and T helper cells of TH2, TH17 and TH21 subtypes. Therapeutic activation of TP53 is intended to restore the powerful tumor-suppressor phenotype mediated by protein, explaining its high predicted impact.

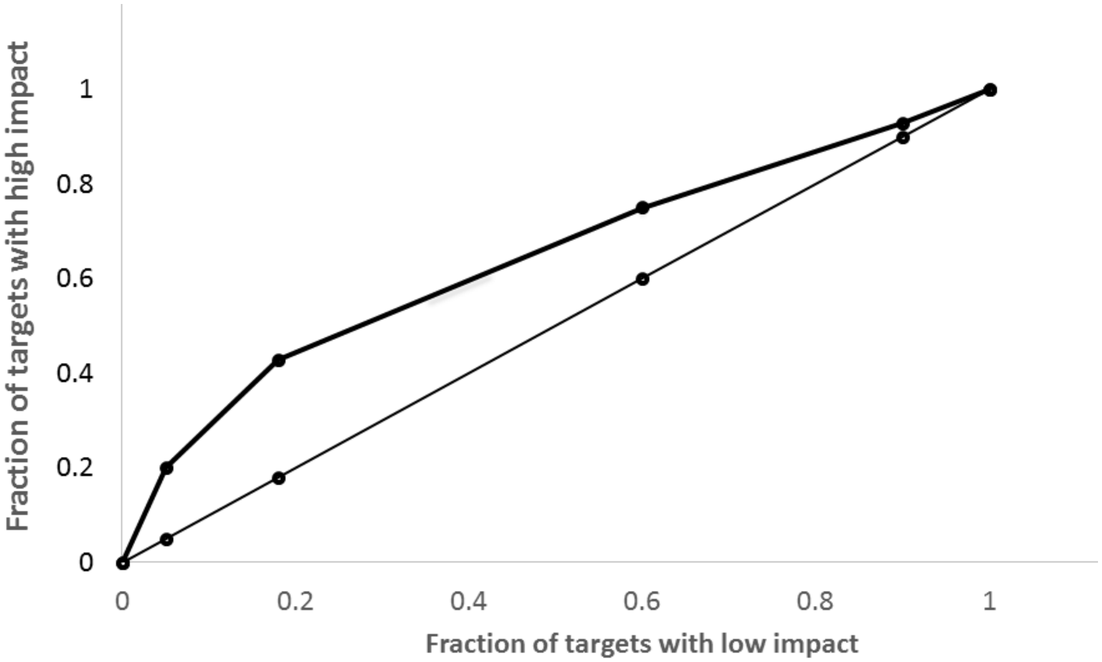

Figure 1. ROC curve of high-impact targets vs low-impact targets. The ROC curve was built using DEXCON and INTENSITY parameters, using the following procedure: 1) divide the combined list of targets (successful and research) by impact categories, the top 25% forming the "high-impact" class and the rest forming "low-impact" class. 2) apply bonus-penalty scoring approach to the DEXCON and INTENSITY values for the list of targets and combining the indirect bonus-penalty scores with the optimized weight in a 2-feature classifier. 3) rank the target list by the values of the 2-feature classifier. 4) determine the fractions of the "high-impact" and "low-impact" categories in each 0.2 bin of rank by the 2-feature classifier. 5) summarize the differential fractions for each category accrued on the range from 0 (highest ranked 2-feature scores) to the given point of rank for the 2-feature score. 6) the sum of differential fractions for "low-impact" categories forms a X-axis coordinate; the sum of differential fractions for "high-impact" categories forms a Y-axis coordinate. The diagonal thin baseline at 45 degrees across the plot reflects the ratio of false positives to false positives for each new point in the form of summary functions, while thicker line reflects the ratio of true positive summary function to the false positive summary function in ROC analysis. At the far right corner, the summary function for the true and false positives are both equal to 1, and the lines cross. The area between the lines is proportional to the resolution quality at multiple possible cut-offs.

Attempts to reactivate CDKN2A (cyclin-dependent kinase inhibitor 2A, multiple tumor suppressor 1) are performed within the same therapeutic paradigm as for TP53. Proto-oncogene c-JUN encodes transcription factor that mediates apoptosis resistance, with good potential of pharmacological inhibition for a significant impact. TNFSF11 (Tumor necrosis factor (ligand) superfamily, member 11) is involved in metastasis and bone-remodeling. Remarkably, the high predicted impact was assigned to TNFSF11 based on Therapeutic Target Database definition of the candidate as not yet matching an FDA-approved ligand. However, we found that TNFSF11 ligand denosumab was approved in 2010 under the names of Xgeva and Prolia, thus, validating our approach. CD40 (TNF receptor superfamily member 5) is a co-stimulatory protein found on antigen presenting cells and as such is a key molecule in establishing an immune response. Alterations of CD40 function determine probability of cancer emergence and metastasis. Proto-oncogene c-MET is involved in *de novo* angiogenesis and metastasis; its activation in tumors is correlated with poor prognosis. Being a receptor component adds up to its potential for higher impact upon drugging. Hepatocyte growth factor/scatter factor HGF is activating a tyrosine kinase signaling cascade of c-Met, thus, contributing to the same metastasis-related pathway. In leukemia patients, JAK2 (Janus kinase 2) forms fusions with the TEL(ETV6) (TEL-JAK2) and PCM1 genes providing the targets that do not exist in normal cells. These targets are druggable in the same manner as by well-known Gleevec. To summarize, the genes pinpointed as promising tend to participate in the most central mechanism of tumor cells survival and propagation.

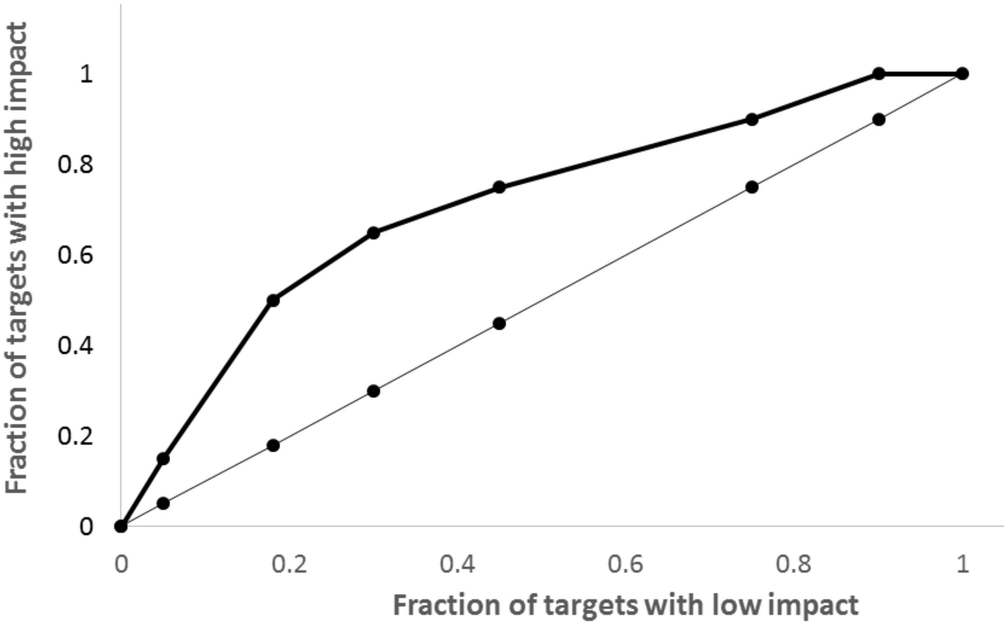

Figure 2: ROC curve of high-impact targets vs. low-impact targets. The ROC curve was built using text-mining parameters, using the following procedure: 1) divide the combined list of targets (successful and research) by impact categories, the top 25% forming the “high-impact” class and the rest forming “low-impact” class. 2) apply bonus-penalty scoring approach to the N (Number of publications between first review and first clinical) and time-derivative of N for the list of targets and combining the indirect bonus-penalty scores with the optimized weight in a 2-feature classifier. 3) rank the target list by the values of the 2-feature classifier. 4) determine the fractions of the “high-impact” and “low-impact” categories in each 0.2 bin of rank by the 2-feature classifier. 5) summarize the differential fractions for each category accrued on the range from 0 (highest ranked 2-feature scores) to the given point of rank for the 2-feature score. 6) the sum of differential fractions for “low-impact” categories forms a X-axis coordinate; the sum of differential fractions for “high-impact” categories forms a Y-axis coordinate. The diagonal thin baseline at 45 degrees across the plot reflects the ratio of false positives to false positives for each new point in the form of summary functions, while thicker line reflects the ratio of true positive summary function to the false positive summary function in ROC analysis. At the far right corner, the summary functions for the true and false positives are both equal to 1, and the lines cross. The area between the lines is proportional to the resolution quality at multiple possible cut-offs.

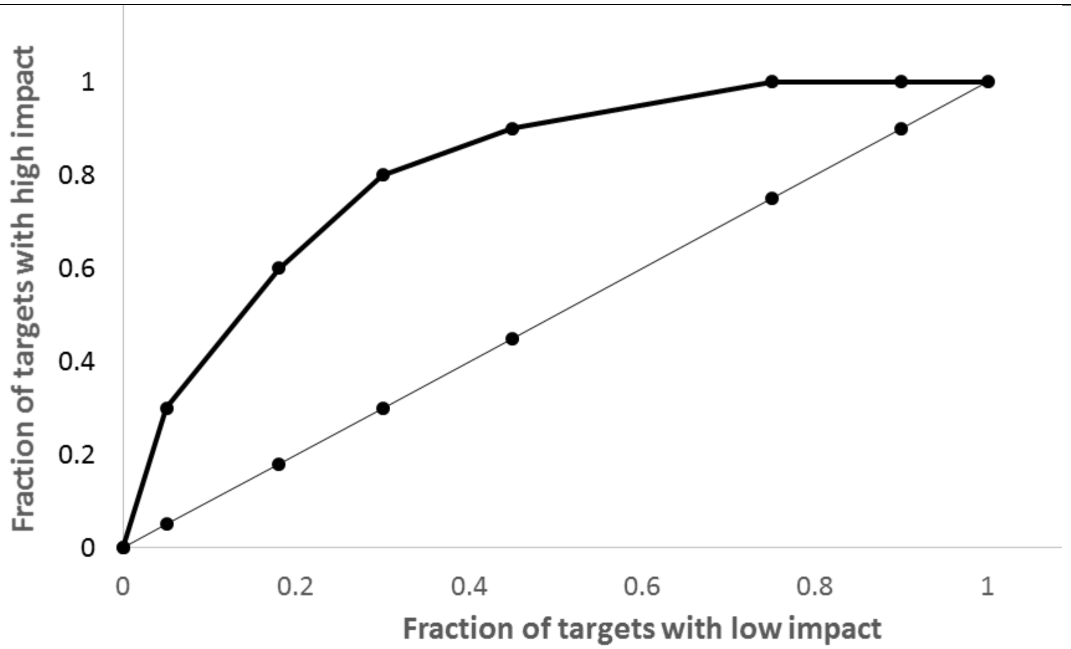

Figure 3: ROC curve of high-impact targets vs low-impact targets. The ROC curve was built using DEXCON, INTENSITY and text-mining parameters, using the following procedure: 1) divide the combined list of targets (successful and research) by impact categories, the top 25% forming the "high-impact" class and the rest forming "low-impact" class. 2) apply bonus-penalty scoring approach to the DEXCON and INTENSITY values for the list of targets. 3) apply bonus-penalty scoring approach to the ranked derivative of early research interest and to the volume N of early research interest. 4) combine the secondary bonus-penalty values for all features with the optimized training weights in a 4-feature classifier. 5) rank the target list by the values of the 4-feature classifier. 6) determine the fractions of the "high-impact" and "low-impact" categories in each 0.2 bin of rank by the 4-feature classifier. 7) summarize the differential fractions for each category accrued on the range from 0 (highest ranked 4-feature scores) to the given point of rank for the 4-feature score. 8) the sum of differential fractions for "low-impact" categories forms a X-axis coordinate; the sum of differential fractions for "high-impact" categories forms a Y-axis coordinate. The diagonal thin baseline at 45 degrees across the plot reflects the ratio of false positives to false positives for each new point in the form of summary functions, while thicker line reflects the ratio of true positive summary function to the false positive summary function in ROC analysis. At the far right corner, the summary functions for the true and false positives are both equal to 1, and the lines cross. The area between the lines is proportional to the resolution quality at multiple possible cut-offs.

The genes with the lowest rank in the Table 1 were analyzed in a similar fashion. IMPDH1 (IMP (inosine 5'-monophosphate) dehydrogenase 1) is mostly involved in transplant rejection and retinitis pigmentosa, its link to the cancer is tenuous. IMPDH2 (inosine 5'-monophosphate dehydrogenase 2) is also involved in autograft rejection, and its connections to cancer are not apparent. ITGA2B (integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41) is mostly involved in fibrinogen activation and coagulopathies, however, its connections to metastasizing are proven. LHCGR (luteinizing hormone/choriogonadotropin receptor) is involved in a broad diversity of pathways and its levels correlate with survival in ovarian epithelial cancer patients. MME (MME membrane metallo-endopeptidase) shows a strong link to cancer, to both prognosis and metastasis, however targeting of metallo-endopeptidases (MMPs) was historically not successful, despite an investment of a lot of efforts. OXTR (oxytocin receptor) is mostly involved in behavior and social adaptation and its involvement in tumorigenesis is a stretch. PARP1 (poly (ADP-ribose) polymerase is strongly related to malignancy, however its expression is ubiquitous, and only a few clinical trials have been published for the targeting of this gene. While PDE4A (phosphodiesterase 4A, cAMP-specific) is involved in cardiac muscle activity and fibroblast proliferation, the links to malignancy are indirect. PTH1R (parathyroid hormone 1 receptor) functions are pleiotropic, with known roles in transplant rejection, organ development and bone maturation, with some evidence of its contribution to certain breast cancers. RARA (retinoic acid receptor, alpha) is vital for differentiation of hematopoietic lineages and respective malignancies. However, the utility of RARA agonists is confined to leukemia field, and the number of clinical trials in this area is limited. TSPO (translocator protein) is involved in microglia and retinal inflammation, HIV-1 virus maturation, while overexpression of TSPO correlates with the progress of breast cancer. However, the number of clinical trials for drugs that target this molecule is small, and its expression is not highly specific to cancer samples. TXNRD1 (thioredoxin reductase 1) participates in redox processes, apoptosis, membrane raft formation, while its overexpression correlates with glioblastoma multiforme progression. Speaking generally, comparison of high and low-impact score bins indicates that the higher ranking score gene set consistently ignites substantially higher interests of members of research community. The entire list of the high score bin members is associated with prominent cancer-related finding and produces relevant DEXCON signals, exceeding the threshold of interest. The drugging of candidates with highest impact ranks would be the most influential on cancer outcomes and deserves prioritization.

Table 2. Modeled impacts for anti-cancer targets at the "research" stage

Gene ID	**Synonyms**	**Target impacts**
STS	ARSC, ARSC1, ASC, ES, SSDD, XLI, STS steroid sulfatase (microsomal), isozyme S arylsulfatase C, estrone sulfatase, steryl-sulfatase, steryl-sulfate sulfohydrolase	155
PROC	APC, PC, PROC1, THPH3, THPH4, protein C (inactivator of coagulation factors Va and VIIIa), anticoagulant protein C, autoprothrombin IIA, blood coagulation factor XIV	108
TP53	BCC7, LFS1, P53, TRP53, tumor protein p53	103
CDKN2A	ARF, CDK4I, CDKN2, CMM2, INK4, INK4A, MLM, MTS-1, MTS1, P14, P14ARF, P16, P16-INK4A, P16INK4, P16INK4A, P19, P19ARF, TP16	80
c-JUN	AP1, V-jun avian sarcoma virus 17 oncogene homolog, activator protein 1, proto-oncogene c-jun, transcription factor AP-1	72
TNFSF11	RP11-86N24.2, CD254, ODF, OPGL, OPTB2, RANKL, TRANCE, hRANKL2, sOdf, TNFSF11, tumor necrosis factor (ligand) superfamily, member 11 , TNF-related activation-induced cytokine, osteoclast differentiation factor, osteoprotegerin ligand	66
CD40	Bp50, CDW40, TNFRSF5, p50, TNF receptor superfamily member 5, B cell surface antigen CD40	41
c-MET	AUTS9, HGFR, RCCP2, c-Met, MET met proto-oncogene HGF receptor, HGF/SF receptor, SF receptor, hepatocyte growth factor receptor, met proto-oncogene tyrosine kinase	41
JAK2	JTK10, THCYT3, Janus kinase 2 JAK-2, Janus kinase 2 (a protein tyrosine kinase), tyrosine-protein kinase JAK2	40
HGF	DFNB39, F-TCF, HGFB, HPTA, SF, hepatocyte growth factor (hepapoietin A; scatter factor), fibroblast-derived tumor cytotoxic factor	29
PRKCA	AAG6, PKC-alpha, PKCA, PRKACA, PRKCA protein kinase C, alpha PKC-A, aging-associated gene 6, protein kinase C alpha type	27
PRKCB	PKC-beta, PKCB, PRKCB1, PRKCB2, protein kinase C, beta PKC-B, protein kinase C beta type	27
CXCR2	CD182, CDw128b, CMKAR2, IL8R2, IL8RA, IL8RB, chemokine (C-X-C motif) receptor 2	23
TGFB1	CED, DPD1, LAP, TGFB, TGFbeta, transforming growth factor, beta 1 , TGF-beta-1	22
TLR9	UNQ5798/PRO19605, CD289, toll-like receptor 9	22

Table 2 Continued…..

AKT1	AKT, CWS6, PKB, PKB-ALPHA, PRKBA, RAC, RAC-ALPHA, v-akt murine thymoma viral oncogene homolog 1 PKB alpha, RAC-PK-alpha, RAC-alpha serine, threonine-protein kinase	20
BCL2	Bcl-2, PPP1R50, B-cell CLL, lymphoma 2 apoptosis regulator Bcl-2, protein phosphatase 1, regulatory subunit 50	18
CLU	AAG4, APO-J, APOJ, CLI, CLU1, CLU2, KUB1, NA1/NA2, SGP-2, SGP2, SP-40, TRPM-2, TRPM2, clusterin, aging-associated protein 4, apolipoprotein J|complement cytolysis inhibitor	18
HIF1A	HIF-1A, HIF-1alpha, HIF1, HIF1-ALPHA, MOP1, PASD8, bHLHe78, hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)	18
IL6	BSF2, HGF, HSF, IFNB2, IL-6, interleukin 6 B-cell differentiation factor, B-cell stimulatory factor 2, BSF-2, CDF, CTL differentiation factor	18
KIF11	EG5, HKSP, KNSL1, MCLMR, TRIP5, kinesin family member 11, TR-interacting protein 5,TRIP-5, kinesin-like protein 1|kinesin-like protein KIF11	17
TNFRSF10A	APO2, CD261, DR4, TRAILR-1, TRAILR1, TNFRSF10A tumor necrosis factor receptor superfamily, member 10a TNF-related apoptosis-inducing ligand receptor 1	16
TNFRSF10B	UNQ160/PRO186, CD262, DR5, KILLER, KILLER/DR5, TRAIL-R2, TRAILR2, TRICK2, TRICK2A, TRICK2B, TRICKB, ZTNFR9, tumor necrosis factor receptor superfamily	16
IGF1	IGF-I, IGF1A, IGFI, insulin-like growth factor 1 (somatomedin C) IGF-IA, IGF-IB, MGF, insulin-like growth factor I, insulin-like growth factor IA	15
MYC	MRTL, MYCC, bHLHe39, c-Myc, MYC v-myc avian myelocytomatosis viral oncogene homolog	15
ANGPT2	AGPT2, ANG2, ANGPT2, angiopoietin 2, ANG-2, Tie2-ligand, angiopoietin-2, angiopoietin-2B, angiopoietin-2a	14
MAP2K1	CFC3, MAPKK1, MEK1, MKK1, PRKMK1, mitogen-activated protein kinase kinase 1, ERK activator kinase 1, MAPK/ERK kinase 1	14
TERT	CMM9, DKCA2, DKCB4, EST2, PFBMFT1, TCS1, TP2, TRT, hEST2, hTRT, TERT telomerase reverse transcriptase	13
CDK4	CMM3, PSK-J3, cyclin-dependent kinase 4 cell division protein kinase 4	10
MDM2	ACTFS, HDMX, hdm2, MDM2 oncogene, E3 ubiquitin protein ligase	9
NQO1	DHQU, DIA4, DTD, NMOR1, QR1, NQO1, NAD(P)H dehydrogenase, quinone 1, DT-diaphorase, NAD(P)H dehydrogenase [quinone] 1	9
MMP2	CLG4, CLG4A, MMP-II, MONA, TBE-1, MMP2 matrix metallopeptidase 2, gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase	8
MMP9	CLG4B, GELB, MANDP2, MMP-9, matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)	7
MMP3	CHDS6, MMP-3, SL-1, STMY, STMY1, STR1, matrix metallopeptidase 3 (stromelysin 1, progelatinase)	6
ERBB3	ErbB-3, HER3, LCCS2, MDA-BF-1, c-erbB-3, erbB3-S, p180-ErbB3, p45-sErbB3, p85-sErbB3, v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 3 proto-oncogene-like protein	5
IL4R	582J2.1, CD124, IL-4RA, IL4RA, interleukin 4 receptor IL-4 receptor subunit alpha, IL4R nirs variant 1, interleukin-4 receptor alpha chain	5
PLK1	PLK, STPK13, polo-like kinase 1 , cell cycle regulated protein kinase, polo (Drosophia)-like kinase, polo like kinase	5
PPARD	FAAR, NR1C2, NUC1, NUCI, NUCII, PPARB, PPARD peroxisome proliferator-activated receptor delta	5

Table 2 Continued…..

EPHB4	HTK, MYK1, TYRO11, EPH receptor B4 ephrin receptor EphB4, ephrin type-B receptor 4, hepatoma transmembrane kinase	4
PLAUR	CD87, U-PAR, UPAR, URKR, plasminogen activator, urokinase receptor monocyte activation antigen Mo3, u-plasminogen activator receptor form 2	4
TXN	RP11-427L11.1, TRDX, TRX, TRX1, thioredoxin ADF, ATL-derived factor, SASP, TXN delta 3, surface-associated sulphydryl protein, thioredoxin delta 3	4
ERBB4	ALS19, HER4, p180erbB4, v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 4	3
FGFR3	ACH, CD333, CEK2, HSFGFR3EX, JTK4, fibroblast growth factor receptor 3	3
MCL1	BCL2L3, EAT, MCL1-ES, MCL1L, MCL1S, Mcl-1, TM, bcl2-L-3, mcl1/EAT, myeloid cell leukemia 1 bcl-2-like protein 3	3
NRP1	C530029I03, NP-1, NPN-1, Npn1, Nrp, neuropilin 1 A5 protein, Neuropilin-1 precursor (A5 protein), neuropilin-1	3
CDK9	RP11-228B15.5, C-2k, CDC2L4, CTK1, PITALRE, TAK, cyclin-dependent kinase 9, CDC2-related kinase, cell division cycle 2-like protein kinase 4	2
CENPE	CENP-E, KIF10, PPP1R61, CENPE centromere protein E, Centromere autoantigen E	2
CHEK2	RP11-436C9.1, CDS1, CHK2, HuCds1, LFS2, PP1425, RAD53, hCds1, CHEK2 , checkpoint homolog	2
FGFR1	BFGFR, CD331, CEK, FGFBR, FGFR-1, FLG, FLT-2, FLT2, HBGFR, HH2, HRTFDS, KAL2, N-SAM, OGD, bFGF-R-1, fibroblast growth factor receptor 1	2
FN1	CIG, ED-B, FINC, FN, FNZ, GFND, GFND2, LETS, MSF, FN1 fibronectin 1 cold-insoluble globulin, fibronectin, migration-stimulating factor	2
MMP1	MMP1, CLG, CLGN, MMP1 matrix metallopeptidase 1 (interstitial collagenase), fibroblast collagenase, interstitial collagenase, matrix metalloprotease 1	2
SMO	E130215L21Rik, Smoh, bnb, smoothened, Smo smoothened homolog (Drosophila) bent body|smoothened homolog	2
TEK	CD202B, TIE-2, TIE2, VMCM, VMCM1, TEK tyrosine kinase, endothelial angiopoietin-1 receptor	2
CCR2	hCG_14621, CC-CKR-2, CCR-2, CCR2A, CCR2B, CD192, CKR2, CKR2A, CKR2B, CMKBR2, MCP-1-R, CCR2 chemokine (C-C motif) receptor 2	1
CDK6	PLSTIRE, CDK6 cyclin-dependent kinase 6, cell division protein kinase 6, serine/threonine-protein kinase PLSTIRE	1
CDK7	CAK1, CDKN7, HCAK, MO15, STK1, p39MO15, CDK7 cyclin-dependent kinase 7, 39 KDa protein kinase, CAK, CDK-activating kinase 1, TFIIH basal transcription factor complex kinase subunit	1
IKBKB	IKK-beta, IKK2, IKKB, IMD15, NFKBIKB, IKBKB inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta	1
MAPK14	RP1-179N16.5, CSBP, CSBP1, CSBP2, CSPB1, EXIP, Mxi2, PRKM14, PRKM15, RK, SAPK2A, p38, p38ALPHA, mitogen-activated protein kinase 14	1
MPO	myeloperoxidase	1
NGFR	RP23-67E18.6, LNGFR, Tnfrsf16, p75, p75NGFR, p75NTR, nerve growth factor receptor (TNFR superfamily, member 16)	1
PRKCD	CVID9, MAY1, PKCD, nPKC-delta, protein kinase C, delta protein kinase C delta VIII, protein kinase C delta type, tyrosine-protein kinase PRKCD	1
WEE1	WEE1A, WEE1hu, WEE1 G2 checkpoint kinase, WEE1+ homolog, wee1-like protein kinase, wee1A kinase	1
MMP14	MMP-14, MMP-X1, MT-MMP, MT-MMP 1, MT1-MMP, MT1MMP, MTMMP1, WNCHRS, matrix metallopeptidase 14 (membrane-inserted)	0.333333333

Table 2 Continued…..

ACVRL1	ACVRL1, ACVRLK1, ALK-1, ALK1, HHT, HHT2, ORW2, SKR3, TSR-I, ACVRL1 activin A receptor type II-like 1, TGF-B superfamily receptor type I, activin A receptor	0
BCL2L2	BCL-W, BCL2-L-2, BCLW, PPP1R51, BCL2-like 2 apoptosis regulator BCL-W, bcl-2-like protein 2, protein phosphatase 1, regulatory subunit 51	0
CCR1	Cmkbr1, Mip-1a-R, Ccr1 chemokine (C-C motif) receptor 1, C-C CKR-1, C-C chemokine receptor type 1, CC-CKR-1, MIP-1 alpha R, MIP-1 alphaR	0
CDC42	CDC42 RP1-224A6.5, CDC42Hs, G25K, cell division cycle 42 G25K GTP-binding protein, GTP binding protein, 25kDa	0
CDH2	CDHN, N-cadherin, Ncad, Cdh2, cadherin 2,neural cadherin	0
CSF1	RP11-195M16.2, CSF-1, MCSF, colony stimulating factor 1 (macrophage), lanimostim, macrophage colony-stimulating factor 1	0
CTSK	RP11-363I22.4, CTS02, CTSO, CTSO1, CTSO2, PKND, PYCD, cathepsin K, cathepsin O, cathepsin O1, cathepsin O2, cathepsin X	0
FNTB	FPTB, farnesyltransferase, CAAX box, beta, CAAX farnesyltransferase subunit beta, FTase-beta protein farnesyltransferase subunit	0
GHSR	growth hormone secretagogue receptor, GH-releasing peptide receptor, GHRP, GHS-R, ghrelin receptor, growth hormone secretagogue receptor type 1	0
GRPR	BB2, gastrin-releasing peptide receptor GRP-R|GRP-preferring bombesin receptor|bombesin receptor 2	0
GSK3B	glycogen synthase kinase 3 beta GSK-3 beta, GSK3beta isoform, glycogen synthase kinase-3 beta, serine/threonine-protein kinase GSK3B	0
GUCY2C	GUCY2C DIAR6, GUC2C, MECIL, MUCIL, STAR, GUCY2C , guanylate cyclase 2C , heat stable enterotoxin receptor, GC-C STA receptor, guanylyl cyclase C, hSTAR	0
HDAC4	AHO3, BDMR, HA6116, HD4, HDAC-4, HDAC-A, HDACA, histone deacetylase 4, histone deacetylase A	0
IL7R	CD127, CDW127, IL-7R-alpha, IL7RA, ILRA, IL7R interleukin 7 receptor, CD127 antigen, IL-7 receptor subunit alpha, IL-7R subunit alpha, IL-7RA, interleukin 7 receptor alpha chain	0
LTA4H	leukotriene A4 hydrolase, LTA-4 hydrolase, leukotriene A-4 hydrolase	0
LTB4R	BLT1, BLTR, CMKRL1, GPR16, LTB4R1, LTBR1, P2RY7, P2Y7, leukotriene B4 receptor G protein-coupled receptor 16, G-protein coupled receptor 16, LTB4-R 1|LTB4-R1|P2Y purinoceptor 7, chemoattractant receptor-like 1	0
MAPK12	ERK3, ERK6, P38GAMMA, PRKM12, SAPK-3, SAPK3, mitogen-activated protein kinase 12, ERK-6, MAP kinase 12, MAP kinase p38 gamma	0
MAPK6	ERK3, HsT17250, PRKM6, p97MAPK, MAPK6 mitogen-activated protein kinase 6, ERK-3, MAP kinase 6, MAP kinase isoform p97	0
MAPK8	JNK, JNK-46, JNK1, JNK1A2, JNK21B1/2, PRKM8, SAPK1, SAPK1c, MAPK8 mitogen-activated protein kinase 8 , JUN N-terminal kinase	0
METAP2	METAP2 MAP2, MNPEP, p67, p67eIF2, methionyl aminopeptidase 2, eIF-2-associated p67 homolog, initiation factor 2-associated 67 kDa glycoprotein	0
MMP12	HME, ME, MME, MMP-12, matrix metallopeptidase 12 (macrophage elastase)	0
MMP13	CLG3, MANDP1, MMP-13, matrix metallopeptidase 13 (collagenase 3) collagenase 3|matrix metalloproteinase 13 (collagenase 3)	0
MMP7	MPSL1, PUMP-1, matrix metallopeptidase 7 (matrilysin, uterine) matrilysin, matrin, matrix metalloproteinase 7 (matrilysin, uterine)	0
NFKB1	EBP-1, KBF1, NF-kB1, NF-kappa-B, NF-kappaB, NFKB-p105, NFKB-p50, NFkappaB, p105, p50	0

Table 2 Continued…..

NPEPPS	AAP-S, MP100, PSA, aminopeptidase puromycin sensitive cytosol alanyl aminopeptidase	0
NTSR1	NTR, NTSR1 neurotensin receptor 1 (high affinity), NTRH, high-affinity levocabastine-insensitive neurotensin receptor, neurotensin receptor type 1	0
P2RY1	P2Y1, purinergic receptor P2Y, G-protein coupled, 1 ATP receptor, P2 purinoceptor subtype Y1, P2Y purinoceptor 1, platelet ADP receptor	0
PPIA	CYPA, CYPH, HEL-S-69p, peptidylprolyl isomerase A, (cyclophilin A), PPIase A, T cell cyclophilin, cyclophilin, cyclosporin A-binding protein	0
PRKCG	PKC-gamma, PKCC, PKCG, SCA14, protein kinase C, gamma protein kinase C gamma type	0
PRLR	HPRL, MFAB, hPRLrI, prolactin receptor PRL-R, hPRL receptor, secreted prolactin binding protein	0
PTPN11	BPTP3, CFC, NS1, PTP-1D, PTP2C, SH-PTP2, SH-PTP3, SHP2, protein tyrosine phosphatase, non-receptor type 11 PTP-2C, protein-tyrosine phosphatase 1D	0
PTPN22	LYP, LYP1, LYP2, PEP, PTPN8, protein tyrosine phosphatase, non-receptor type 22 (lymphoid) PEST-domain phosphatase|hematopoietic cell protein-tyrosine phosphatase	0
RPS6KB1	PS6K, S6K, S6K-beta-1, S6K1, STK14A, p70 S6KA, p70(S6K)-alpha, p70-S6K, p70-alpha,ribosomal protein S6 kinase, 70kDa, polypeptide 1 p70 S6 kinase	0
SLC44A4	SLC44A4 DAAP-66K18.1, C6orf29, CTL4, NG22, TPPT, solute carrier family 44, member 4, choline transporter-like protein 4	0
SPHK1	SPHK, sphingosine kinase 1 , SK 1, SPK 1	0
TEP1	TLP1, TP1, TROVE1, VAULT2, p240, telomerase-associated protein 1, TROVE domain family, member 1, p80 telomerase homolog, telomerase protein 1	0
YES1	HsT441, P61-YES, Yes, c-yes, v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1	0

By contrast, the lower scoring members not always show well-established association with cancer phenotypes as evident from the functional description entries that passed manual curation before linking to respective gene names in “Genes” subdivision of NCBI. For this group of genes, a majority of evidence is based on correlation of mRNA or protein expression levels to cancer phenotypes or outcomes, however, an independent evaluation of the consistency of overexpression findings does not confirm uniformity of this observation. However, retrospective post-prediction analysis revealed an important caveat which is likely to result in future improvements of the impact score concept. Some genes, like RARA or MME, demonstrate strong links to cancer, but produce low scores due to relatively narrow utility, i.e. limiting the applicability of developed ligands to limited spectrum of tumors. As these diseases are generally considered “orphan”, it is important that the development of the ligands aimed at the treatment of these pathologies should continue unimpeded despite low scores for respective targets. Hence, the model that we propose should be further optimized by introducing an “orphan” disease coefficient that would preclude an attrition of the targets that are highly specific to certain malignancies that ail relatively low number of patients.

An impact of targeting “super-targets” as a component of a combinatorial treatment

Typically, anti-cancer therapies are combinatorial as they include at least 2 components. Assuming typical three-component therapy approach, n-fold increase in the number of available high-impact targets would result in n3 increase in the number of available drug combinations, prompting further progress in their evaluation and testing. Based on the data in Table 1 and different evaluations (2- 3), the effect of the current pool of therapies can be substantially magnified by this combinatorial expansion. Above we discussed that a conservative estimate of the contribution of therapeutics to the observed doubling of cancer survival is at 40%. Designating this increment as SIDT (Survival Increment Due To Therapies), being equal of 40%, one can draw a model:

$$\mathrm{SIDT}(2) = \mathrm{SIDT}\ (1)\ N^{nm} \qquad (12)$$

where SIDT (1) is the survival increment due to therapies at the current level, SIDT(2) is the survival increment at the projected level, N is the number of available therapies at the current level, m is the increment in the number of high-impact targets applied in cancer field. The exponential coefficient 2-3 assumes a formation of two or three-component drug cocktails, however this number may be greater or lesser in the future.

Based on the model (12), at n = 2 and m = 2, the SIDT(2) would increase 4-fold, which would be coming very close to curing at least some types of cancer (see Table 1), and improving ten year survival rates for lung, brain and pancreatic cancers by at least 30%. At n = 3 and m = 2, the SIDT (2) would increase 9-fold, making many now lethal types of cancer eradicated, producing >60% improvement of the survival rates after 10 years observation for many other malignant disorders.

DISCUSSION

A significant progress is achieved for treatment of the majority of cancer form, with the average survival rate practically doubling over the last 45 years (Table 3 compiled according to the UK data presented at (1). The rate of progress appears to be constant for most of individual sub-ranges of the plot, however a certain recent acceleration is observed. Likely, increase in the number of available treatments contributes to this improvement, although not as a single factor. Assuming that the level of investment, and, therefore, projected impacts correctly reflect the level of future revenues/sales both for the ‘successful” and for the “research” sets of targets and also assuming that the level of sales correctly reflects the benefit to society, one can argue that the top 10% of the successful targets, or just seven of them, produce 75% of all anti-cancer effects. The top impact target alone, EGFR, mediates 23% of total anti-cancer effect, while the second best target, ERBB2, mediates 22% of total effect. While from purely clinical point of view these numbers seem disproportional, many novel therapies rely on a combinatorial synergy with EGRF and ERBB2 ligands (22-23). Hence, introduction of novel high-impact targets could alter the survival dynamics even further for at least some of the cancer forms.

The technique presented in this paper allowed us to evaluate the potential impact of the targets which are currently at the research stage. Our analysis highlighted microsomal steroid sulfatase (estrone sulfatase) at the top of the list which, after some score gap, was followed by anticoagulant protein C, p53, CDKN2A, c-Jun, TNSFS11, CD40, c-MET and JAK2, all of which were highlighted as the most promising research-stage targets. Accoring to our calculations, the relative importance of microsomal steroid sulfatase is at approximately the same level as that of well-known anticancer targets BRAF kinase and progesterone receptor. A PubMed search using term ”estrone sulfatase” or “estrone sulfatase” AND “cancer” returns 193 and 125 manuscripts, respectively. Recent years resulted in the development of a number of potent estrone sulfatase inhibitors aimed at the suppression of the formation of both E1 and breast carcinoma-promoting steroid dehydroepiandrosterone (DHEA) from DHEA-sulfate (DHEAS) (24-26). As approximately 40% of breast tumors are estrogen-dependent, successful advancement of estrone sulfatase inhibitors into clinical practice could potentially lead to sizable global effects. On the other hand, many potential targets were predicted to have minimal impact; deprioritization of these targets may lead to substantial savings and subsequent shift of clinical development efforts toward the most promicing drug candidates.

Our impact scoring technique is, in a nutshell, a 4-feature bonus-penalty classifier which comprises two components, the microarray and the literature mining. The INTENSITY feature is the level of absolute transcript expression demonstrated by the target candidate. With all other factors being equal, the candidates with more intensive expression would influence biological signal transduction events more robustly as they produce higher amounts of mRNA, and, therefore, the protein. While the detected correlation of the impacts and the transcript levels is relatively weak (r = 0.2), it is sufficient to boost performance of a classifier of the bonus-penalty type. The DEXCON feature reflects the stability of the differential expression signal in tumors of various tissue origins, and in the tumors of same origin. As it was shown earlier, the

DEXCON score is superior to typical t-test based evaluations of the significance of observed differential expression patterns, as it takes into consideration a consistency of evidence (27). It is important to note that microarray-derived features are capable of serving as predictors even when completely novel target candidates comes into the scope of study; hence, their value is higher than that of text-mining features. The text-mining features rely on pre-existing information regarding a potential action mechanism and perceived value of a target candidate. When a majority of scientific data is collected prior to the major information disseminating event, i.e. publishing of influential review and/or result of a clinical trial, and the "me-too" bias is minimized by pre-dissemination choice of cut-off, the literature mining features become a less-than-obvious predictor, although still inferior to the experimental measurements such as microarrays. It is obvious that the bibliometric aspect of the study may be improved by taking into account the impact factors of the journals, number of patents, relative sizes of each study, total amount of grant support etc. In this initial report, the bibliometric aspects were limited to the number of publications and to the rate of accumulation prior to the critical bias-producing events.

Table 3: 10 year survival rates post-diagnosis based on the data collected in UK during 1971-2007 period, survival increments are computed over entire period.

Cancer localizations		**Period of Diagnosis**				**Survival increment % from 1971 to 2007**
		1971-1972	**1980-1981**	**1990-1991**	**2007(2)**	
Bladder		34.6	52.2	50.3	48.9	+14.4
Bowel	**Colon**	22.6	33.3	38.3	50.4	+27.8
	Rectum (1)	23.9	29.9	33.4	49.3	+25.4
Brain		5.7	7.4	8.7	9.4	+2.9
Breast (Female)		38.9	49.3	61.1	77.0	+38.1
Cervix		48.4	52.6	59.5	63.0	+14.6
Hodgkin Lymphoma		49.0	57.8	68.7	77.9	+28.9
Kidney		22.2	27.1	31.7	43.5	+21.3
Larynx (Male)		50.5	56.9	55.1	59.6	+9.1
Leukemia		8.1	13.8	23.5	33.2	+25.2
Lung		3.2	3.9	3.7	5.3	+2.1
Malignant Melanoma		49.3	57.5	69.5	83.2	+32.1
Myeloma		5.3	8.8	9.3	17.1	+11.8
Non-Hodgkin Lymphoma (1)		21.8	29.7	35.0	50.8	+29
Oesophagus		3.6	4.5	4.6	10.0	+6.4
Ovary		18.0	21.9	25.3	35.4	+17.4
Pancreas		1.9	2.1	1.7	2.8	+0.9
Prostate		20.4	30.1	30.9	68.5	+48.1
Stomach		4.6	7.2	8.5	13.5	+8.9
Testis		67.4	84.7	91.7	96.5	+28.6
Uterus		55.2	62.9	64.4	74.5	+18.7
Other Cancers		34.0	31.4	30.1	36.3	+2.3
All Cancers		23.7	30.0	33.7	45.2	+21.5

The rationale behind our approach is that both the total number of publications and their accelerating deposition into the PubMed are proxies to research interest, which, in turn, correlates with the objective value of the target. The research interest fundamental also drives publishing in higher impact journals and determines awarded grant support, which is instrumental to perform studies in larger groups of animals or patients cohorts. Hence, introduction of additional bibliometric parameters will also introduce co-correlating variables. Relative weights for each of these parameters should be evaluated by experimenting *in silico*. On the other hand, target gene expression related features are *a-priori* independent of the bibliometrics and, therefore, more likely to add an input to the model.

The linear classifier for YP' was selected due to its robustness which aids in prevention of model overfitting. Since the training set was as small as ~ 30 successful anticancer targets, this precaution appears to be warranted, especially if the number of features would be increasing due to inclusion of other information sources, for example, the networks of biomolecules. The least square method was selected for regression modeling, with the ranking-based cutoffs for high and low real impact. High rank was defined as the highest quartile and the low rank was defined as two lowest quartiles. These cut-offs are somewhat arbitrary and, therefore, the results of the study are qualitative rather than quantitative. However, the proposed technique for estimating the commercial promise of still potential targets, which are costly to develop, is inexpensive, and, therefore, of value. In this study, selected cut-offs clearly separated the high and low impact groups of targets while preserving sufficient number of targets in each group and, by that, allowing for ROC plotting.

To generate a high proxy impact Y, a target should demonstrate a consistent promise in multiple clinical trials. There is certainly an informational gap between early research interest and clinical trial results. An intriguing discovery of a novel pharmacological mechanism and its experimental confirmations at pre-clinical level may not even acknowledge possible inability of a target to acquire a non-toxic ligand, unfavorable patterns of expression or pharmacodynamics etc. From the point of information theory, a predictor is a function that contributes a quantity of information sufficient to measure the pattern of the future event or approach it.

$$I_2 = I_1 + \Delta I \qquad (11)$$

Where I_2 - is the final state, the completeness of information allows reliably describe the target pattern of the future; I_1 - is the initial state, the fragmentary or zero initial information concerning the target pattern is insufficient

ΔI – the predictor produces the increment of information rendering knowledge of the future pattern.

Based on the formula (11), the microarray setting corresponds to $I_1 \sim 0$ (little is known about any aspect of the target and its behavior prior to the experiment), while text-mining corresponds to $I_1 > 0$ (substantial knowledge about the target and its expected behavior prior to computation of metrics).

From these considerations it is apparent that a perfect microarray classifier that allows complete prediction of a future pattern produces a greater informational increment than an equally perfect text-mining classifier. At the same time, the contribution of the text-mining classifier is non-zero, unless the information gap between the present state and the future state is negligible. Thus, we argue that the proposed text-mining approach is at least partially objective and, therefore, provides an added value when coupled with the microarray data. Speaking generally, the fusion of orthogonal sets of features produces a greater summary □I to elucidate the final state more reliably than the component set of features in isolation. In that sense, the imperfect (biased) contribution of the text-mining features is still useful, due to its informative component permitting to bridge the information gap in the equation sooner (11). The method of extracting text-mining features employed in this report is analogous to consensus forecasting used in economic modeling. In most of cases the expert consensus is correct, but historical record attests that it never should be applied in isolation (28-29). Thus, the text-mining derived features and microarray features act synergistically, supporting each other, and provide an integrated predictor.

The current rates of attrition for the ligands and targets are discussed extensively (30, 31). As an example, Hutchinson et al reports an average attrition rate for the anti-cancer therapies as 95-96% (30). As the leads to the loss of all the costs accrued by the rejected ligands, the attrition of more advanced candidates is more damaging event. Implementation of targets evaluation by their potential for eventual success will lead to earlier elimination of some ligands off the development pipelines. The preference towards anti-cancer targets with the widest possible therapeutic window would contract the overall volume of the clinical trials. If we would treat a clinical trial as a test with a certain signal-to-noise ratio, we could apply known statistical observation that the size of the test is smaller if the signal-to-noise ratio is inherently higher. In other words, if the targets are selected on their favorable gene expression pattern with preferential expression in tumors rather that in normal cells, lesser toxicities are expected, and an enrollment of lower numbers of patients into dose escalation trials would be necessary.

CONCLUSION

The main result of the report is demonstration that the promising behavior of a pharmacological target is predictable early based on their expression signatures and text-mining of the pattern of the early research interest. Considering a very divergent levels of promise displayed by the currently approved targets, we conclude that most of survival increment attributed to anticancer therapies in general is achieved via the high-impact targets. Improvement in predicting the targets with inherently wide therapeutic window may result in clinical trials stage savings and, eventually, in explosion of therapeutic opportunities that would benefit the entire society.

CONFLICT OF INTEREST

There is no conflict of interest at all stages of the manuscript creation

ACKNOWLEDGEMENTS

The authors acknowledge the diligent effort by Neeraj Prasad in data mining based on the author’s instructions and a proofread by Jerome Glasser. This study has been funded by Ministry of Science and Education, Russia (Project no. RFMEFI60714X0098).

REFERENCES

1. Cancer survival for common cancers. [cited 10 August 2016]. Available from: http://www.cancerresearchuk.org/cancer-info/cancerstats/survival/common-cancers/#Cancer.
2. Esteva FJ, Hortobagyi GN. Integration of systemic chemotherapy in the management of primary breast cancer. The Oncologist. 1998 Oct 1;3(5):300-13.
3. DeVita VT, Chu E. A history of cancer chemotherapy. Cancer research. 2008 Nov 1;68(21):8643-53.
4. Goodman LS, Wintrobe MM, Dameshek W, Goodman MJ, Gilman A, McLennan MT. Nitrogen mustard therapy: Use of methyl-bis (beta-chloroethyl) amine hydrochloride and tris (beta-chloroethyl) amine hydrochloride for Hodgkin's disease, lymphosarcoma, leukemia and certain allied and miscellaneous disorders. Journal of the American Medical Association. 1946 Sep 21;132(3):126-32.
5. Wright JC, Prigot A, Wright BP, Weintraub S, Wright LT. An Evaluation of Folic Acid Antagonists in Adults with Neoplastic Diseases. Journal of the National Medical Association. 1951 Jul;43(4):211.
6. Nathanson L, Hall TC, Schilling A, Miller S. Concurrent combination chemotherapy of human solid tumors: experience with a three-drug regimen and review of the literature. Cancer Res. 1969 Feb;29(2):419-25.
7. Vassilakopoulos TP, Angelopoulou MK. Advanced and relapsed/refractory Hodgkin lymphoma: what has been achieved during the last 50 years. Semin Hematol. 2013;50(1):4-14.
8. Wilson PM, Danenberg PV, Johnston PG, Lenz HJ, Ladner RD. Standing the test of time: targeting thymidylate biosynthesis in cancer therapy. Nat Rev Clin Oncol. 2014;11(5):282-98. doi: 10.1038/nrclinonc.2014.51.
9. LoRusso PM. Phase 0 clinical trials: an answer to drug development stagnation? Journal of clinical oncology. 2009;27(16):2586-8.
10. Stewart DJ, Batist G. Redefining cancer: a new paradigm for better and faster treatment innovation. J Popul Ther Clin Pharmacol. 2014;21(1):e56-65.
11. Juliano RL. Pharmaceutical innovation and public policy: The case for a new strategy for drug discovery and development. Science and Public Policy. 2013: scs125.
12. FDA's shorter approval time for new drugs raises questions. [cited 10 August 2016]. Available from http://www.nbcnews.com/health/fdas-shorter-approval-time-new-drugs-raises-questions-8C11484613.
13. Zhang D, Surapaneni S, editors. ADME-enabling technologies in drug design and development. John Wiley & Sons, Incorporated; 2012 Apr 13.

14. Zhou W, Wang Y, Lu A, Zhang G. Systems Pharmacology in Small Molecular Drug Discovery. Int J Mol Sci. 2016 Feb 18;17(2):246. doi: 10.3390/ijms17020246.
15. Manis JP. Knock out, knock in, knock down—genetically manipulated mice and the Nobel Prize. New England Journal of Medicine. 2007 Dec 13;357(24):2426-9.
16. The Cost Of Creating A New Drug Now $5 Billion, Pushing Big Pharma To Change. [cited 10 October 2014]. Available from: http://www.forbes.com/sites/matthewherper/2013/08/11/how-the-staggering-cost-of-inventing-new-drugs-is-shaping-the-future-of-medicine/
17. Ma'ayan A, Jenkins SL, Goldfarb J, Iyengar R. Network analysis of FDA approved drugs and their targets. Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine. 2007 Apr 1;74(1):27-32.
18. Zhu F, Han L, Zheng C, Xie B, Tammi MT, Yang S, Wei Y, Chen Y. What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets. Journal of Pharmacology and Experimental Therapeutics. 2009 Jul 1;330(1):304-15.
19. Xu H, Fang Y, Yao L, Chen Y, Chen X. Does Drug-target Have a Likeness? Target. 2007;3000:49.
20. Sakharkar MK, Li P, Zhong Z, Sakharkar KR. Quantitative analysis on the characteristics of targets with FDA approved drugs. Int J Biol Sci. 2008 Jan 1;4(1):15-22.
21. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins JL. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. ChemMedChem. 2007 Jun 11;2(6):861-73.
22. Stinchcombe TE, Borghaei H, Barker SS, Treat JA, Obasaju C. Pemetrexed With Platinum Combination as a Backbone for Targeted Therapy in Non-Small-Cell Lung Cancer. Clin Lung Cancer. 2016; 17(1):1-9. doi: 10.1016/j.cllc.2015.07.002.
23. Singla S, Pippin JA, Drebin JA. Dual ErbB1 and ErbB2 receptor tyrosine kinase inhibition exerts synergistic effect with conventional chemotherapy in pancreatic cancer. Oncology reports. 2012 Dec 1;28(6):2211-6.
24. Hong Y, Chen S. Aromatase, estrone sulfatase, and 17β-hydroxysteroid dehydrogenase: structure-function studies and inhibitor development. Mol Cell Endocrinol. 2011 Jul 4;340(2):120-6.
25. Numazawa M, Tominaga T, Watari Y, Tada Y. Inhibition of estrone sulfatase by aromatase inhibitor-based estrogen 3-sulfamates. Steroids. 2006 May;71(5):371-9.
26. Aidoo-Gyamfi K, Cartledge T, Shah K, Ahmed S. Estrone sulfatase and its inhibitors. Anticancer Agents Med Chem. 2009 Jul;9(6):599-612.
27. Mayburd A, Baranova A. Knowledge-based compact disease models identify new molecular players contributing to early-stage Alzheimer's disease. BMC systems biology. 2013 Nov 7;7(1):1.
28. Batchelor R. The IMF and OECD versus consensus forecasts. City University Business School, London, August. 2000 Aug. http://www.consensuseconomics.com/
29. Golinelli R, Parigi G. Real-time squared: A real-time data set for real-time GDP forecasting. International Journal of Forecasting. 2008 Sep 30;24(3):368-85.
30. Hutchinson L, Kirk R. High drug attrition rates—where are we going wrong? Nature Reviews Clinical Oncology. 2011 Apr 1;8(4):189-90.
31. Stifling New Cures: The True Cost of Lengthy Clinical Drug Trials [cited 10 October 2014] Available from: http://www.manhattan-institute.org/html/fda_05.htm

Supplemental Table 1. Cost structure for different stages in drug development process. From A. Roy "Stifling new cures: the true cost of lengthy clinical trials, FDA Project reports, V. 5, 2012

Function	Share of Total (expenses)	Probability of FDA approval	Reciprocal of approval probability (tested ligands per a marketed ligand)
Preclinical	28%	8%	12.5
Phase I	9.2%	21%	4.9
Phase II	17.4%	28%	3.8
Phase III	39.8%	58%	1.6
Approval	5%	90%	1.1