netDx: Interpretable patient classification using integrated patient similarity networks

Authors:

Shraddha Pai[1,2], Shirley Hui[1], Ruth Isserlin[1], Muhammad A Shah[1], Hussam Kaka[1], Gary D. Bader*[1,3,4,5]

Affiliations:

1. The Donnelly Centre, University of Toronto, Toronto, Canada
2. Affiliate Scientist, The Centre for Addiction and Mental Health, Toronto, Canada
3. Department of Molecular Genetics, University of Toronto, Toronto, Canada
4. Department of Computer Science, University of Toronto, Toronto, Canada
5. The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada

* gary.bader@utoronto.ca

Abstract

Patient classification has widespread biomedical and clinical applications, including diagnosis, prognosis and treatment response prediction. A clinically useful prediction algorithm should be accurate, generalizable, be able to integrate diverse data types, and handle sparse data. Importantly, the resulting model should be easily interpretable, as clinicians are unlikely to trust black box statistical models. We describe netDx, the first supervised patient classification framework based on patient similarity networks. netDx meets the above criteria and particularly excels at data integration and model interpretability. We demonstrate the features of this framework by integrating up to six heterogeneous datatypes, including clinical variables, DNA methylation, somatic mutations, mRNA, miRNA and protein expression profiles, for survival prediction in kidney, lung, ovarian and brain cancer. As a machine learning tool, netDx outperforms eight standard machine-learning methods in predicting binary survival in renal clear cell carcinoma, and performs at par with these methods in predicting ovarian carcinoma. In comparison to traditional machine learning-based patient classifiers, netDx results are more interpretable, visualizing the decision boundary in the context of patient similarity space and identifying biological pathways and other features important for prediction. Using pathway-level features in predicting kidney cancer survival from transcriptome data, netDx identified known and potentially novel kidney cancer pathways and biomarkers. Thus, netDx can serve both as a useful classifier and as a tool for discovery of biological features characteristic of disease. Upon publication, an open-source R/Java implementation of

netDx will be made publicly available along with sample files and automation workflows packaged as vignettes.

Introduction

The goal of precision medicine is to build quantitative models that guide clinical decision-making by predicting disease risk and response to treatment using data measured for an individual. Within the next five years, several countries will have general-purpose cohort databases with 10,000 to >1 million patients, with linked genetics, electronic health records, metabolite status, and detailed clinical phenotyping; examples of projects underway include the UK BioBank[1], the US NIH Precision Medicine Initiative (www.whitehouse.gov/precision-medicine), and the Million Veteran Program (http://www.research.va.gov/MVP/). Additionally, human disease specific research projects are profiling multiple data types across thousands of individuals, including genetic and genomic assays, brain imaging, behavioural testing and clinical history from integrated electronic medical records[2-4] (e.g. the Cancer Genome Atlas, http://cancergenome.nih.gov/). Computational methods to integrate these diverse patient data for analysis and prediction will aid understanding of disease architecture and promise to provide actionable clinical guidance.

Statistical models that predict disease risk or outcome are in routine clinical use in fields such as cardiology, metabolic disorders, and oncology[5-8]. Traditional clinical risk prediction models typically use generalized linear regression or survival analysis, in which individual measures are incorporated as terms (or features) of a single equation. Standard methods of this type have limitations analyzing large data from genomic assays. Machine learning methods can handle large data, but are often treated as black boxes that require substantial effort to interpret how specific features contribute to prediction. Black box methods are unlikely to be clinically successful, as physicians must understand the characteristic features of a disease to make a confident diagnosis[9]. Further, many existing methods do not natively handle missing data, requiring data pruning or imputation, and have difficulty integrating multiple different data types.

The patient similarity network framework can overcome these challenges and excels at integrating heterogeneous data and generating intuitive, interpretable models. In this framework, each input patient data feature (e.g. gene expression profile, age) is represented as a patient similarity network (PSN). Each PSN node is an individual patient and an edge between two patients corresponds to pairwise similarity for a given feature. For instance, two patients could be similar in age, mutation status or transcriptome. PSNs can be constructed based on any available data using a similarity measure. Because all data is converted to a single type of input (networks), integration across diverse data types is straightforward. Patient similarity networks (PSN) have been used successfully for unsupervised class discovery in cancer and type 2 diabetes[10,11].

We describe netDx, the first PSN-based approach for supervised patient classification. In this system, patients of unknown status can be classified based on their similarity to patients with known status. This process is clinically intuitive because it is analogous to clinical diagnosis, which often involves a physician relating a patient to a mental database of similar patients they have seen. As demonstrated below, netDx has strengths in classification performance, heterogeneous data integration, usability and interpretability.

Results

Algorithm Description

The overall netDx workflow is shown in Figure 1. This example conceptually shows how PSNs can be used to predict if a patient is at high or low risk of developing a disease based on a variety of patient-level data types. Similarity networks are computed for each patient pair and for each data type. In this example, high-risk patients are more strongly connected based on their clinical profile, which may capture age and smoking status, and metabolomics profile. Low-risk patients are more similar in their clinical and genomic profiles. The goal of netDx is to identify the input features predictive of high and low risk, and to accurately assign new patients to the correct class.

Input data design. Each patient similarity network (PSN) is a feature, similar to a variable in a regression model (we use the terms “input networks” and “features” interchangeably). A PSN can be generated from any kind of patient data, using a pairwise patient similarity measure (Figure 1A). For example, gene expression profile similarity can be measured using Pearson correlation, while patient age similarity can be measured by the normalized difference. A reasonable design is to define one similarity network per data type, such as a single network based on correlating the expression of all genes in the human genome, or a network based on similarity of responses to a clinical questionnaire. If a data type is multivariate, it is more interpretable to define a network for each individual variable. However, this approach may lead to too many features generated (e.g. millions of SNPs), which increases computational resource requirements and risk of overfitting. Thus, there is a trade-off between interpretability and overfitting/scalability, which is implicit in machine learning feature design. To help address this problem for 'omics data, we group gene-based measurements into biological pathways, which we assume capture relevant aspects of cellular and physiological processes underlying disease and normal phenotypes. This biological process-based design generates ~2,000 networks from gene expression profiles containing over 20,000 genes, with one network per pathway.

Selecting features informative of class prediction. Feature selection identifies the input networks with the highest generalizable predictive power, and is run once per patient class. netDx is trained on samples from the class of interest, using cross-validation (Figure S1) and an established association network integration algorithm[12,13]. The algorithm scores each network based on its value in the classification task. The ideal network is one connecting all patients of the same class without any connections to other classes. The least useful network is one that connects patients from one class to patients from other classes, without connecting any patients in the same class. In each cross-validation fold, regularized linear regression assigns network weights, reflecting the ability to discriminate query patients from others, and removes uninformative networks. netDx increases a network's score based on the frequency with which it is assigned a positive weight in multiple cross-validation folds. The classifier's sensitivity and specificity can be tuned by thresholding this score; a network with a higher score achieves greater specificity and lower sensitivity. The output of this feature selection step is a set of networks that can be integrated to produce a predictor for the patient class of interest.

Class prediction using selected features. After training and feature selection are separately run for each class, feature selected networks are combined by averaging their similarity scores to produce an integrated network. Test patients are ranked by similarity to each class using label propagation in the integrated network, and are assigned to the class with the highest rank[14,15] (Figure S2).

netDx output (Figure 1C-D). netDx returns predicted classes for all test patients and standard performance measures including the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPR), and accuracy. Scores for each feature are returned and if pathway features are used, they are visualized using an enrichment map (Figure 1D)[16]. The integrated patient network is visualized and used to assess the strength of class separation, and distance of one patient to others in the class, using network topology measures (Online Methods, Figure 1C).

Predictor checklist. Each netDx classifier should be assessed using a checklist of tests to gain confidence in the classification results (Figure 1C). Such a checklist should include:

1) traditional performance metrics, including the AUROC, AUPR, F1, and accuracy
2) the extent to which the predictor captures prior knowledge about the disease under study, such as known cellular pathways

3) an orthogonal measure of the validity of the predicted classes. For instance, in the context of survival prediction in cancer, a predictor should result in significantly separable survival curves for the two predicted patient sets
4) a measure of the strength of separation of the classes, such as the extent to which patient classes separately cluster in the integrated similarity network
5) if the results are better than random, measured using an appropriate set of negative controls

Each test results in a pass, fail or conditional pass. This simple categorization helps ease overall performance assessment (Figure 1C).

Predicting binarized survival in cancer

To demonstrate the utility of netDx, we predicted survival in four tumor types, using data from The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) via the TCGA PanCancer Survival Prediction project website of Yuan et al.[17], https://www.synapse.org/#!Synapse:syn1710282, Table 1). These tumor types were chosen because they have been rigorously analyzed using eight machine learning methods and thus provide an excellent performance benchmark[17]. Data were for renal clear cell carcinoma[18] (KIRC, N=150 patients), ovarian serous cystadenocarcinoma[19] (OV, N=252), glioblastoma multiforme[20] (GBM, N=155), and lung squamous cell carcinoma[21] (LUSC, N=77). Data for a given tumour type included: clinical variables (e.g. age, tumour grade); mRNA, miRNA and protein expression; DNA methylation; and somatic copy number aberrations. Binarization of survival and format of clinical variables followed Yuan et al.

Each data type was represented as a single patient similarity network. For each tumour type, we classified high and low survival using multiple combinations of input data: one data type at a time (one input network); clinical data plus one other data type (two input networks); or with all available data (five or six input networks, depending on the tumour type). Pairwise patient similarity was computed using Pearson correlation if five or more variables were present[22], or by average normalized similarity where data had less than five variables (Online Methods).

Informative features were identified during training using nested cross-validation. Patient samples were split 80:20 into a training and a blind test set. Using only the training samples, 10-fold cross validation was performed for each class (good survival; poor survival), generating for each network a score between 0 and 10. Networks that scored better than 8 out of 10 were used to classify blind test samples. This process was repeated 100 times for random splits of train and blind test (Figure 2A). Predictor performance was measured as the average of blind test classification across the 100 splits. A network that consistently scored well (10 out of 10) across all 100 splits was selected for the final predictor.

All data sets had some level of signal as indicated by AUROC; the best performing predictors had AUROC of 0.83 (SEM=0.01; mean over 100 splits +/- SEM) for KIRC, 0.65 (0.01) for GBM, 0.68 (0.01) for OV and 0.66 (0.01) for LUSC, respectively (Figures 2, S3-S5, Table S1). Integrating multiple data types was useful in some tumour types. KIRC shows the best such improvement (Figure 2B; one-way ANOVA of single vs. pair or all: $p < 4.3e-$

103). In contrast, integration does not improve performance in GBM compared to using solely clinical data (Figure 2C; mean AUROC for clinical=0.65 (0.01); Dunnett's test for clinical vs. other single data sources, $p < 1e\text{-}10$ for all; Dunnett's test for clinical vs. integrated data, $p > 0.1$). Results are similar for OV, where clinical data performs the best among all single sources, (Figure 2D; mean AUROC=0.68; Dunnett's test $p < 0.002$), but integration offers no improvement (Dunnett's test, $p > 0.1$). In LUSC, no condition significantly outperforms all the others but clinical and proteomic data perform equally best (Figure 2E).

netDx outperforms previously-published measures for binary survival prediction for KIRC (best mean AUROC for netDx=0.84, compared to 0.775 from Yuan et al.), and is similar for OV (netDx is 0.68 compared to Yuan et al. 0.684)[17]; it performs worse for the other two cancer types (GBM: netDx: 0.65 vs. Yuan et al. 0.71; LUSC: 0.66 vs. Yuan et al. 0.84). Performance statistics reported for Yuan et al. were the best scores determined from eight different machine learning methods: diagonal discriminant analysis; K-nearest neighbor; discriminant analysis; logistic regression; nearest centroid; partial least squares; random forest; and support vector machine. Thus, netDx is useful and complementary to other machine learning strategies as it can do better in certain circumstances.

Assessing predictor reliability through a checklist

We evaluated each predictor using a set of tests (Figure 3, Table 1). First, a predictor needed to perform better than chance, as measured by average AUROC and AUPR across the 100 splits (Figure 3A). To validate class assignments, we compared the survival profiles for netDx-predicted good and poor survivors using a log-rank test, and measured the fraction of the 100 splits with $p < 0.05$ (Figure 3B). We categorized a predictor has having fully passed the test (>75% splits with $p < 0.05$), conditionally passed the test (50-75% splits with $p < 0.05$) or failed the test (<50% splits with $p < 0.05$). Figure S6 shows all results for this test. KIRC passes this test for any predictor that includes clinical data (Figure 3B), and conditionally passes when proteomic data are included alone. The best result for OV is a conditional pass with only clinical data used (Figure 3B). LUSC and GBM fail this test in all conditions. We next compared the weighted shortest-path distances of same-class node pairs to different-class node pairs in the integrated patient network (Online Methods). A configuration passed if both classes, good and poor survival, grouped closer together than to nodes of the opposite class. The KIRC predictor using clinical data passes this test (Figure 3C, S7A; $p < 0.001$). By contrast, the OV predictor based on clinical data passed for good ($p <3.3e-25$) but not poor survivors ($p>0.9$); we define this as a conditional pass.

The comparison of checklist performance for the best scenario for KIRC and OV shows that KIRC consistently passes the checklist tests, making it a more reliable predictor (Figure 3). The KIRC survival predictor, particularly when used with clinical data, was the only consistently reliable predictor of all configurations and tumour types we tested. Predictors for OV survival are less reliable (Figure S6). None of the conditions tested for LUSC or GBM were reliable in our tests (Figure 2C, E; Figure S6). The checklist framework helps identify reliable predictors among all tested configurations.

Evaluating the predictive performance of individual clinical variables

Our first survival predictor, described above, used one network for each data type. Clinical data is composed of relatively few variables and separating these could result in a more interpretable predictor. To test the effect of coding each variable as a separate feature, we created a KIRC predictor splitting clinical data into three networks, one each for tumour stage, grade, and patient age. This predictor shows a significant improvement in AUROC score compared to our original classifier (Figure S8A; mean AUROC+/-SEM=0.85+/-0.01, p<0.036), though was not significantly improved in AUPR (Figure S7A, p>0.1). Individual features had different network scores, reflecting their variable importance. Networks representing tumour stage and grade were highest scored for both classes, while age was only highly scored for predicting low survival. Consistent with this, increased tumour stage and grade are significantly associated with low survival by a univariate test (Figure S8B; p <2e-5), and this agrees with previous literature[23]. Thus, using individual variables may improve performance and improves interpretability by measuring the predictive power of each variable.

Evaluating the value of pathway-level features

To examine how well netDx can provide biological insight into the classes of interest, we ran a predictor for KIRC survival where gene expression was split into pathways (Figure 4, Online Methods). No network scored 10 out of 10 in all 100 splits for the "good survival" class, thus we relaxed the feature selection threshold to 10 out of 10 in >= 70% of splits. Feature-selected pathways for good survival include "reactions specific to the complex N-glycan synthesis pathway" and "thyroxine biosynthesis", both related to glycoprotein hormones. Six pathways were feature-selected for poor survival, with themes including salt transport, vitamin and co-factor metabolism and cell adhesion (Figure 4; scores of top networks in Table S2, 3). The pathways identified are consistent with processes previously known to be altered in KIRC tumours (30% of themes for good and poor survival), including metabolic pathways for cholesterol biosynthesis, the regulation of the pyruvate dehydrogenase complex, and co-factor metabolism. Acetyl co-A carboxylase alpha (ACACA), frequently mutated in KIRC, is a member of these pathways[18]. Aquaporin-1 (AQP1), a member of pathways related to renal water homeostasis, is a urinary biomarker for early detection of renal clear cell carcinoma[24]. N-linked glycan expression is associated with features of metastatic cancer progression, such as neoplastic transformation, angiogenesis, tumour survival, and loss of contact inhibition in multiple cancer types[25]. N-glycans are potential biomarkers for detection of renal clear cell carcinoma[26,27].

Using pathway-based features in netDx results in a predictor that outperforms one where mRNA data is treated as a single feature (mean AUROC=0.73 for pathways; 0.66 for single; one-sided WMW $p < 3.1e-9$; Figure S9A). Additionally, the pathway-based netDx predictor results in an improved separation of survival curves for classified samples (one-sided WMW; $p <1.4e-3$, Figure S9B). The integrated patient network for single-feature and pathway-based features is similar, in that good survivors group closer together than opposite-class pairs, but poor survivors do not (Figure S9C). Therefore, pathway-level features perform better than the single-network configuration on the predictor checklist.

To examine how strongly each pathway feature correlates with outcome, we performed principal component analysis on the input gene expression matrix using the set of genes for each selected pathway, and correlated the projections of the first three principal components with clinical outcome. All features individually showed significant correlation with outcome (e.g. correlation for "Thyroxine biosynthesis" = -0.41, p<2.7e-7), and the class boundary is visually evident in these features (Figure 4B-C).

Finally, we asked if combining clinical data with feature-selected pathways would improve KIRC survival predictor performance. A classifier using feature-selected pathways and individual clinical variables shows a significant improvement over using clinical variables only, when measured with AUPR (0.82; one-sided WMW p < 8.35e-4) but not with other metrics (Figure S10, S7D).

Altogether, our results show that using pathway features provides performance improvements and substantially improves biological interpretability and insight into disease mechanisms.

Discussion

We describe netDx, the first supervised clinical sample classification system based on a patient similarity network framework, and demonstrate its features using multimodal data from four different tumour types. This framework can be used to create accurate, generalizable predictors, and has particular strengths in data integration and interpretation. netDx is targeted at clinical researchers who are interested to see if their data can answer a specific patient classification question. netDx provides a standard workflow that can quickly determine if the classification question can be answered based on a training set and if so, provides a set of relevant features, a reliability report card and a tool to classify new patients.

netDx is flexible and general. Heterogeneous datatypes are converted into a common “patient similarity” space, easing their integration. We integrated up to six major data types, including five ‘omic layers (Figure 2). netDx can accept input with missing values, because the network association algorithm it uses for feature-scoring has this capability[15]. In this case, the patient is not represented in the particular network where its value is missing, but it will be in other networks that have data for that patient. netDx also includes support for feature grouping to improve interpretability while keeping feature number low, to mitigate the risk of overfitting and improve signal detection with sparse data. This may improve prediction performance. We demonstrate this functionality by grouping gene-level expression measures into pathway-level features (Figure 4). Users may construct groupings for any patient data type, though groupings with clear clinical or mechanistic interpretation will aid class interpretation.

We implemented a predictor checklist or “report card” to evaluate predictors based on classification performance, model interpretability and consistency (Figure 1,3; Tables 1,2). While netDx provides output useful to construct a checklist, some checklist items are specific for a classification task (e.g. log-rank test for survival). We hope that including diverse performance measures will contribute to greater insight into possible cause and effect relationships in the model (e.g. Bradford Hill criteria for inferring causation[28]).

In the PanCancer survival prediction example, the renal clear cell carcinoma (KIRC) survival predictor consistently outperformed predictors for the other tumor types (Figure 2,3). The predictors for glioblastoma multiforme (GBM) and ovarian carcinoma perform worse than those for KIRC, despite having comparable or substantially more samples. Thus, increased sample size is not a guarantee of improved performance. This variation may be because survival was dichotomized to balance sample sizes in the two groups, rather than based on some clinical or biological criterion; the threshold for good survival is one year for GBM, while that for KIRC is the longest, at four years. Another possibility is that the 'omic data may not contain detectable signal for survival time in GBM. Also, in all instances, clinical data outperforms 'omic data in predicting survival. This seems surprising because the 'omics data is much larger than the clinical data. However, tumour stage and grade, which measure the spread and size of a tumour, are well known to negatively impact survival time (Figure S6B). Genomic data is still valuable, as it enables netDx to provide mechanistic insight into disease via use of pathway features. Thus, it is useful to analyze all available data to support both prediction performance and biological discovery.

While the pathway-based predictor performs better than the one where mRNA data is presented as a single similarity network, we observed that we can create a predictor with performance similar to that of the pathway-based predictor by using randomly generated "pseudo" pathways containing non-pathway genes (Figure S9). We interpret this to mean that grouping genes reduces noise and improves good vs. poor survival signal in the gene expression data. Consistent with this idea, destroying the correlation structure of the gene expression matrix drops performance to random (Figure S11). Use of real pathways selects those that match known biology of each cancer type, which we interpret to mean that netDx can select appropriate, interpretable pathways. Thus, pathways may improve performance by providing both biological signal and general noise reduction.

netDx provides a complete framework for precision medicine, however the ultimate vision is to enable clinical researchers to assess classification performance for questions of interest, such as 'will a patient respond to one therapy or another?' based on patient measurements and outcomes present in large electronic medical record databases. Output would include model performance and generalizability estimates on independent cohorts, feature interpretation, an interactive integrated patient similarity network visualization, a predictor checklist, and a ready-to-run classifier for new patients. Ideally this would be provided as a report that is easily interpretable by clinical researchers who would gain confidence in classification performance for further research or safe use with patients.

netDx is implemented as an open-source R software package available at http://netdx.org, with worked examples. We also propose that users store and publicly share patient similarity networks, useful as features for netDx and other PSN methods, in the NDEx network exchange system[29]. Patient similarity networks shown in this manuscript are publicly available in NDEx, under the UUID numbers 2f24606b-a217-11e7-a10d-

0ac135e8bacf, f9fab009-a218-11e7-a10d-0ac135e8bacf, 511ded80-a218-11e7-a10d-0ac135e8bacf, and a0c529c5-a218-11e7-a10d-0ac135e8bacf.

Acknowledgements

We thank Quaid Morris and Daniele Merico for discussions on method development. We also thank Han Liang for discussion on implementation details for the machine learning used in Yuan et al. (2014).

Author contributions

All authors contributed to netDx method development. S.P. and M.A.S. analyzed the datasets in this manuscript. S.P. wrote the netDx software package with contributions from M.A.S. S.H., H.K. and R.I. developed initial versions of netDx. S.P. and G.D.B. wrote the paper.

Competing financial interests

The authors declare no competing financial interests.

References

1. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
2. Calkins, M.E. et al. The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. *J Child Psychol Psychiatry* **56**, 1356-69 (2015).
3. Collins, F.S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med* **372**, 793-5 (2015).
4. Hudson, T.J. et al. International network of cancer genome projects. *Nature* **464**, 993-8 (2010).
5. Gail, M.H., Anderson, W.F., Garcia-Closas, M. & Sherman, M.E. Absolute risk models for subtypes of breast cancer. *J Natl Cancer Inst* **99**, 1657-9 (2007).

6. Lee, A.J. et al. BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. *Br J Cancer* **110**, 535-45 (2014).
7. Schmidt, M.I. et al. Identifying individuals at high risk for diabetes: The Atherosclerosis Risk in Communities study. *Diabetes Care* **28**, 2013-8 (2005).
8. Wilson, P.W. et al. Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-47 (1998).
9. Castelvecchi, D. Can we open the black box of AI? *Nature* **538**, 20-23 (2016).
10. Li, L. et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med* **7**, 311ra174 (2015).
11. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* **11**, 333-7 (2014).
12. Mostafavi, S., Goldenberg, A. & Morris, Q. Labeling nodes using three degrees of propagation. *PLoS One* **7**, e51947 (2012).
13. Zuberi, K. et al. GeneMANIA prediction server 2013 update. *Nucleic Acids Res* **41**, W115-22 (2013).
14. Mostafavi, S. & Morris, Q. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics* **26**, 1759-65 (2010).
15. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* **9 Suppl 1**, S4 (2008).
16. Merico, D., Isserlin, R. & Bader, G.D. Visualizing gene-set enrichment results using the Cytoscape plug-in enrichment map. *Methods Mol Biol* **781**, 257-77 (2011).
17. Yuan, Y. et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* **32**, 644-52 (2014).
18. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-9 (2013).
19. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-15 (2011).
20. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8 (2008).
21. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-25 (2012).
22. Abdel-Megeed, S.M. Accuracy of Correlation Coefficient with Limited Number of Points. *The Journal of Experimental Education* **52**, 188-191 (1984).
23. Zisman, A. et al. Improved prognostication of renal cell carcinoma using an integrated staging system. *J Clin Oncol* **19**, 1649-57 (2001).
24. Morrissey, J.J. et al. Evaluation of Urine Aquaporin-1 and Perilipin-2 Concentrations as Biomarkers to Screen for Renal Cell Carcinoma: A Prospective Cohort Study. *JAMA Oncol* **1**, 204-12 (2015).
25. Pinho, S.S. & Reis, C.A. Glycosylation in cancer: mechanisms and clinical implications. *Nat Rev Cancer* **15**, 540-55 (2015).
26. Gbormittah, F.O. et al. Clusterin glycopeptide variant characterization reveals significant site-specific glycan changes in the plasma of clear cell renal cell carcinoma. *J Proteome Res* **14**, 2425-36 (2015).
27. Hatakeyama, S. et al. Serum N-glycan alteration associated with renal cell carcinoma detected by high throughput glycan analysis. *J Urol* **191**, 805-13 (2014).

28. Hill, A.B. THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION? *Proc R Soc Med* **58**, 295-300 (1965).

29. Pratt, D. et al. NDEx, the Network Data Exchange. *Cell Syst* **1**, 302-305 (2015).

Figures

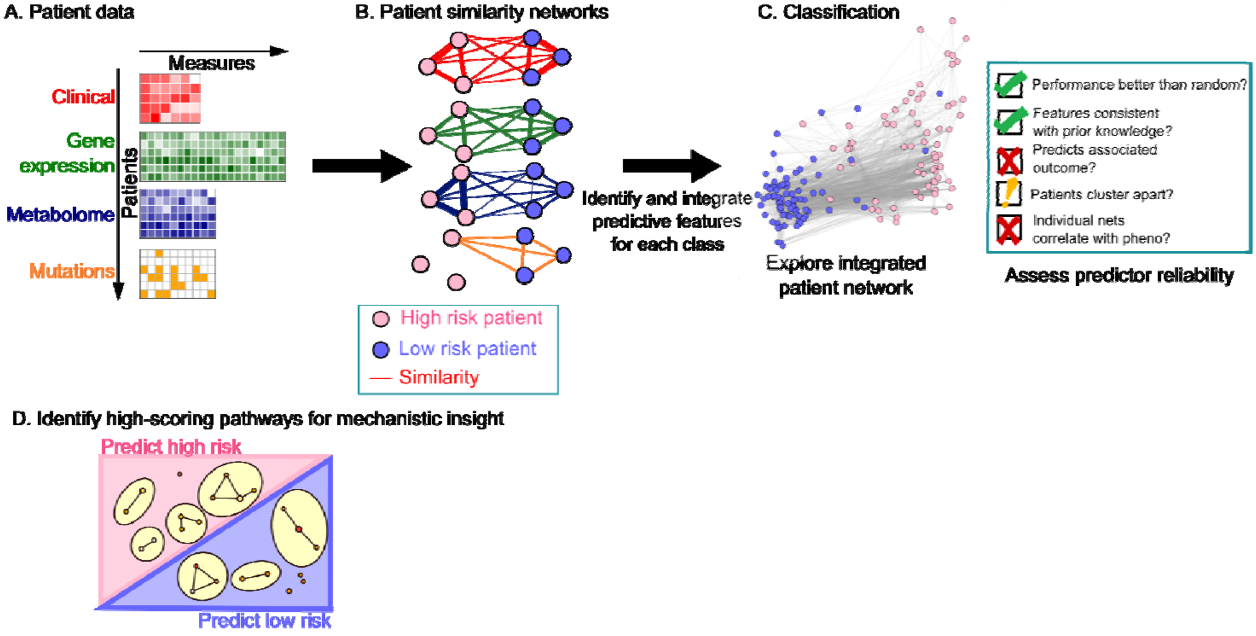
A. Patient data
Measures
Clinical
Gene expression
Metabolome
Mutations
Patients
B. Patient similarity networks
High risk patient
Low risk patient
Similarity
Identify and integrate predictive features for each class
C. Classification
Explore integrated patient network
Performance better than random?
Features consistent with prior knowledge?
Predicts associated outcome?
Patients cluster apart?
Individual nets correlate with pheno?
Assess predictor reliability
D. Identify high-scoring pathways for mechanistic insight
Predict high risk
Predict low risk

Figure 1. The netDx method.

netDx converts patient data (A) into a set of patient similarity networks (PSN), with patients as nodes and a user-provided similarity measure as weighted edges (B). The simple example for predicting low/high risk for disease uses clinical, genomic, metabolomic and genetic data. netDx identifies which networks strongly relate high-risk patients (here, clinical and metabolomic data) and which relate low-risk patients (clinical and gene expression data). Cross-validation is used to score each input network by its ability to predict patient class; details in Figure S1.

C. netDx returns several types of output. Top-scoring features are combined into a single view of overall patient similarity. This integrated network view can be used to classify new patients based on relative similarity to known patient classes. It provides traditional performance metrics for the classifier, such as AUROC. netDx also provides scores for the predictive value of individual features. A performance checklist including a set of metrics, tests and controls helps evaluate the predictor.

D. If pathway features are used, netDx provides visualization of the relationship between top-scoring pathways.

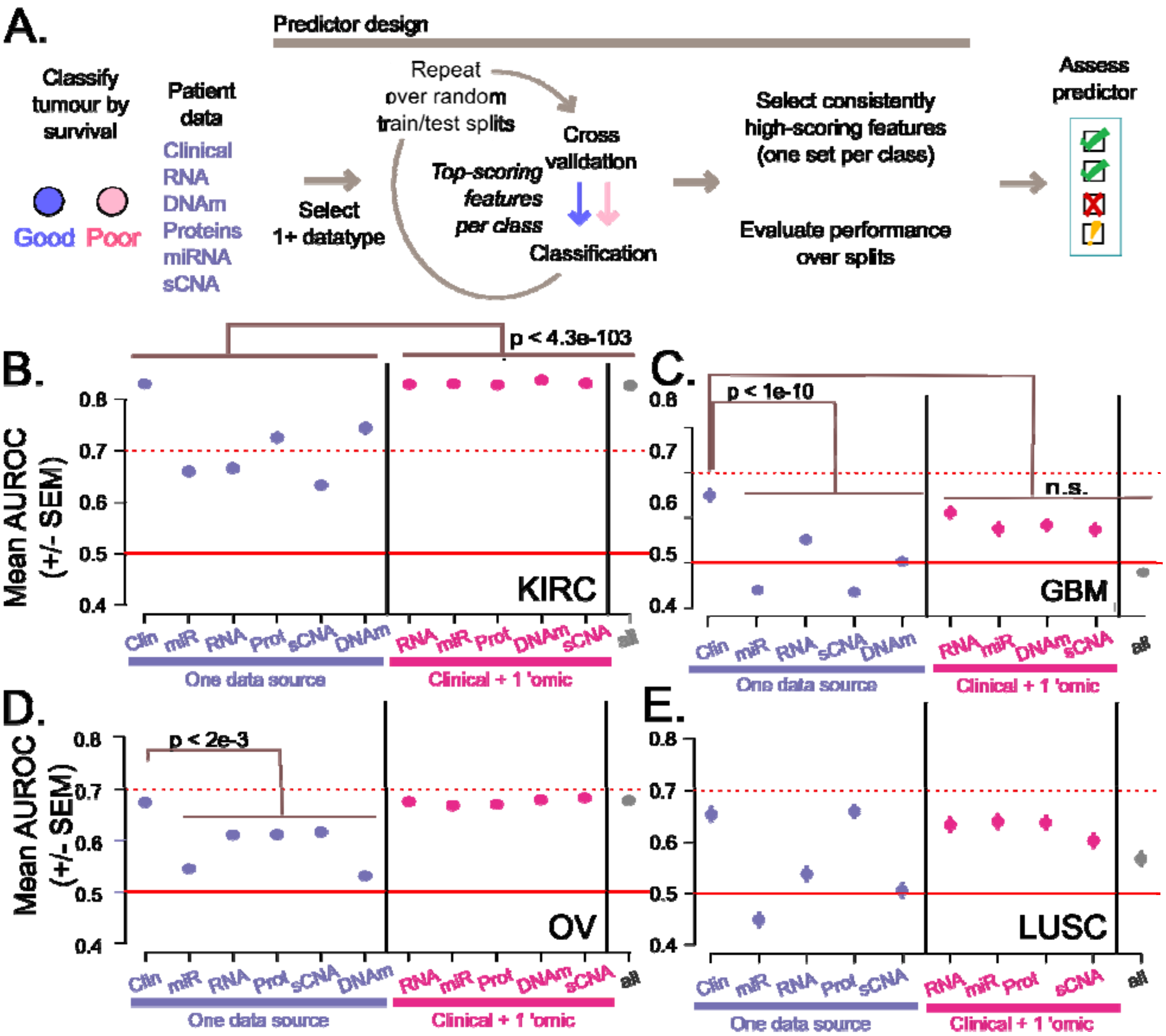

Figure 2. PanCancer predictor design and results.

A. Method. Binary survival was predicted using different combinations of input networks. Predictors were built by running 10-fold cross validation over 100 different train/blind test splits. This step resulted in 100 performance measures, which are plotted in panels B-E. Selected features were those that scored 10 in all 100 splits. Predictor reliability was assessed using a checklist of tests (Table 1).

B-E. Average AUROC over 100 train/blind test splits for each predictor scenario for (B) KIRC, (C) GBM, (D) OV and (E) GBM. Each panel is divided based on whether a data source is used singly (purple), is a genomic source paired with clinical data (pink) or whether all data sources were included as input (all). KIRC: p-value from one-way ANOVA; GBM and OV: p-values from Dunnett's test. The solid reference line indicates random performance and dashed line indicates AUROC=0.70.

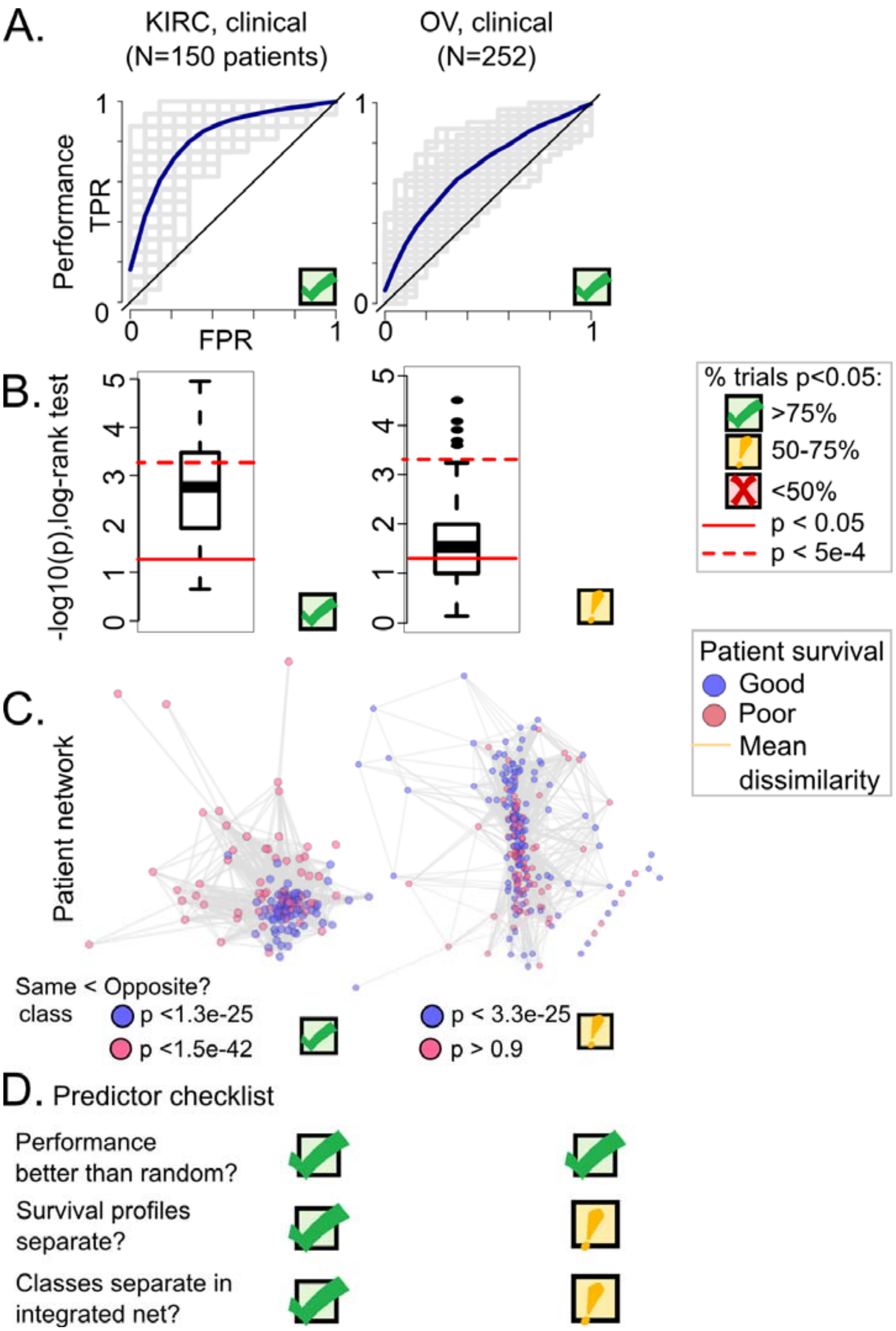

Figure 3. A predictor checklist is useful to compare reliability of different netDx predictors. Here we compare the performance of survival prediction in kidney (KIRC; left) and ovarian (OV; right) carcinoma. The predictor was built using nested cross-validation with 100 train and blind test data splits. A) Test of predictor performance, where both predictors pass

with performance above chance. B) Test of the difference of survival profiles of predicted classes. Passing depends on the fraction of splits with a significant p-value in the log-rank test. C) Test to measure the closeness of samples within the same class in the integrated patient network relative to opposite-class pairs. D) Compilation of test results in a tabular format, where it is clear that the KIRC predictor consistently performs well and is likely more reliable than the OV predictor.

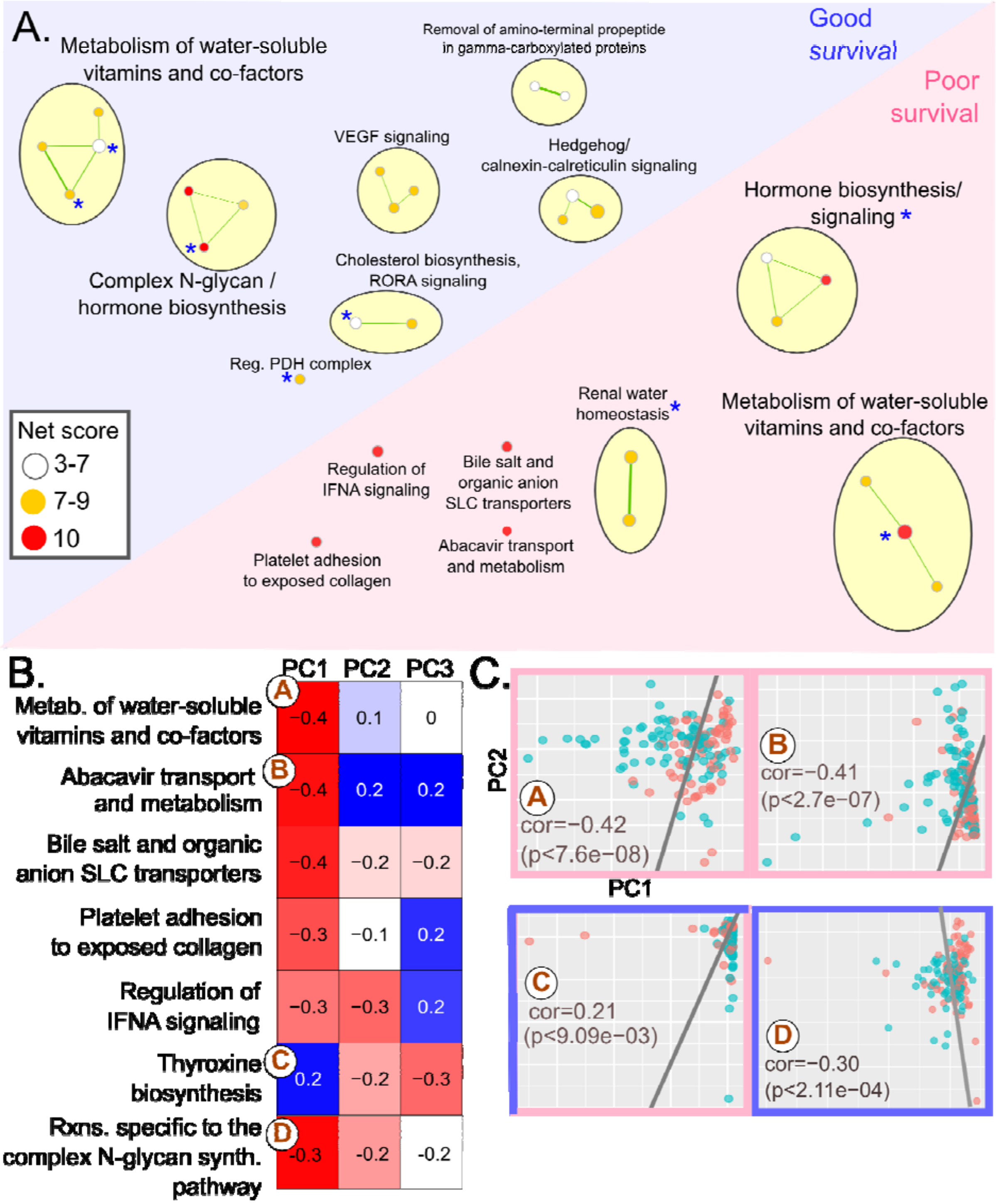

Figure 4. Top-scoring pathways for predicting survival in clear cell renal carcinoma, using only pathway features.

A. Enrichment map of top-scoring pathways (nodes), where edges represent genes shared by two pathways. Node colour indicates the maximum score achieved by a feature in >70% of trials; nodes in red are selected features. Nodes are thematically clustered[1]. Blue asterisks indicate pathways/genes known to be altered in this tumour. Single pathways scoring less than 10 are indicated in Table S2 and S3.

B. Correlation of top-scoring features (represented as the first three principal components of pathway-specific gene expression) with survival (Spearman's correlation). Table cells are colored by sign and magnitude of correlation (blue: Spearman corr. **>**0; red, corr. <0). Circled letters correspond to detailed panels in C.

C. Projections of patient-level gene expression in feature-selected pathways onto first two principal components (individual dots indicate patients). Points are colored by survival class. Panel border colour indicates whether a feature was selected for low survival (pink), good survival (blue) or both (mixed). Decision boundaries were calculated using logistic regression on scatterplot data.

Tables

A. PanCancer performance checklist for one-net-per-datatype

Tumour type (num good; poor survival)	**Data type**	**Clinical variables**	**Feature selected (scored 10 in 100 runs)**	**Perf. better than random?**	**Are multiple data sources better than single?**	**Do survival curves of predicted classes separate?**	**In the overall PSN, are same-class pairs closer than opposite-class pairs?**
KIRC (N=80 good; 70 poor)	**C,G,R,m,P, D**	Age, tumour stage, tumour grade	**Good: P** **Poor: C,P,D**				
OV (153 good, 99 poor)	**C,G,R,m,P, D**	Age	**Good: C** **Poor: C**			!	!
LUSC (49 good; 28 poor)	**C,G,R,m,P**	Age, stage	*None*				
GBM (88 good; 67 poor)	**C,G,R,m,D**	Age,Sex, Karnofsky score	**Good: -** **Poor: C**				

B. Test criteria for pass/fail

Test	Pass	! Conditional pass	Fail
Are multiple data sources better than single? One-way ANOVA comparing performance of predictor with two or more data sources to that with single data sources.	one-sided WMW or Dunnett's test, p < 0.05	n/a	One-sided WMW or Dunnett's test, p > 0.05
Do survival curves of predicted classes separate? % train/test splits with p < 0.05 for log-rank test (must be true for at least one combination of input datatypes tested)	>75%	50-75%	<50%
In the overall PSN, are same-class pairs closer than opposite-class pairs? In the integrated patient similarity network, distribution of average weighted shortest path distance between same-class pairs, compared to opposite-class pairs, p < 0.05 for one-sided WMW	p < 0.05 for all classes	p <0.05 for at least one class	p >0.05 for all classes

Table 1. A. Summary for PanCancer binary survival prediction and reliability checklist results *Data type:* **C** Clinical; **C*** Clinical represented as separate variables; **G** somatic copy number alterations; **R** RNA; **m** miRNA, **P** Proteomic (RPPA); **D** DNA methylation.

B. Description of tests for predictor checklist and criteria for test assessment