molecular systems biology

netDx: Interpretable patient classification using integrated patient similarity networks

Shraddha Pai, Shirley Hui, Ruth Isserlin, Muhammad A Shah, Hussam Kaka and Gary D. Bader.

Review timeline:

	Submission date:	12th June 2018
	Editorial Decision:	23rd July 2018
	Revision received:	29th November 2018
	Editorial Decision:	5th February 2019
	Revision received:	12th February 2019
	Accepted:	13th February 2019

Editor: Maria Polychronidou

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 23rd July 2018

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the three referees who agreed to evaluate your study. As you will see below, the reviewers think that the presented approach seems interesting. They raise however a series of concerns, which we would ask you to address in a revision.

The reviewers' recommendations are rather clear so I think that there is no need to repeat the points listed below. Importantly, methodological details and software implementation should be made available to the reviewers. Please feel free to contact me in case you would like to discuss in further detail any of the issues raised by the reviewers.

REFEREE REPORTS

Reviewer #1:

The authors describe netDx, a procedure for performing supervised machine learning using similarity metrics. In this case, features are collapsed to similarity scores (usually just a Pearson correlation, going by Supplementary Table 1) before being fed into a machine learning algorithm.

Though the performance appears to be good, I also have some concerns related to the presentation. I may also have concerns about the underlying work, but I don't know that because the methods section appears to be relatively incomplete. It would be impossible for me to provide a substantive peer review on the methods/results given what is currently available.

The authors only lightly describe which exact performance metric goes into Figure 2 for netDx and how that performance is obtained. Is the variability from performance over splits, over models, etc? I need this information to make any judgement on the manuscript's validity.

Though the limited description makes it hard to assess the methods and results, I can provide some feedback on the framing. The authors describe this approach as being used for precision medicine. I have considerable numbers of concerns around this point:
* Similarity-based methods by design leak the information of participants. It's unclear how this can be mitigated. A classifier can at least be distributed, but for similarity you need to distribute enough information for someone else to calculate similarity + the clinical information.
* "netDx provides a complete framework for precision medicine." It does not appear to actually provide this. It would be better to more accurately define the scope of what netDx provides.

The checklist is both too specific to netDx and too incomplete all at the same time. It might be worth considering whether or not this would be better handled via a separate review or perspective paper. If it's going to be handled here, please consider the extent to which it is overly tuned for this context. For example "... in the integrated similarity network" appears to apply, in all the copious literature on the topic, to only this paper. Also, steps that seem relevant like independent validation in a separate cohort and prospective trials are both missing. Finally, the term orthogonal appears to be applied to something that does not appear to be orthogonal: "in the context of survival prediction in cancer, a predictor should result in significantly separable survival curves for the two predicted patient sets."

Finally, the largest problems associated with using similarity-based methods for precision medicine appear to be unaddressed. It requires a large bank of samples to be able to begin to make predictions, and it's unclear how similarity scores will generalize across all platforms. Since every value from every data type can contribute to the predictor, it may be substantially harder to deploy these methods in a production setting. This doesn't need methodological work within this exact contribution, but should be acknowledged as a limitation. As framed though, this appears to be a glaring weakness of both the method and the lack of independent test sets measured by different groups.

Minor:

* Remove the discussion of "borderline significant." Since the decision has been made to use this approach to assessing the robustness of results, either it is or it is not significant depending on the selected alpha.

Reviewer #2:

In this manuscript, the authors introduce a new method for survival-classification of patient by integration of omics and clinical data. The methodology is novel and generally well described. The evaluation adequately demonstrates that the method provides an advantage over conventional machine learning methods, and provides novel biological insights when integrated with molecular pathway information. The manuscript suffers from a number of unclear/imprecise formulations, and is missing some detailed information here and there that should be provided in the supplements. If these shortcomings are addressed, I believe the manuscript is excellent and suitable for publication.

Page 6, line 4 from bottom: You write that feature networks are integrated using "an established association network integration algorithm" and cite two papers (ref. 12 and 13), the first one being the 3Prop, and the second being an update paper for GeneMANIA. It appears the 3Prop algorithm is used, but this is not perfectly clear. I suggest mentioning 3Prop by name and/or removing ref. 13, as it is not relevant here.

The performance of NetDx is described as "excellent" in several places. While this is up to interpretation, the mean AUROCs generally vary between 0.6 and 0.8, and judging from Supp. table 1, the increase in performance from data integration seems to come mostly from the clinical data. While NetDx does perform comparatively well to the other methods, "excellent" is, to me, misleading, especially when used in the abstract without context.

Page 22, line 4 from bottom (Figure 2 caption): You write that you show a "representative split" from the survival models trained on KIRC. Please specify what constitutes representative in this context. Given the high variance in performance of the different splits, this is a significant detail.

Page 22, Figure 2: A one-sided Mann-Whitney U-test is used for comparing netDx performance to all other methods. Please provide a justification for using a one one-tailed test instead of a two-tailed test.
It would also be beneficial to include significance tests between netDx and each method individually (in supplementary table.)
Supplementary table 1: Please include variance of AUROC values in addition to mean. Also, please include AUROC values for methods compared to as well for reference.

Online methods, page 4, page mid: "an edge weight at floating-point precision". The term "floating-point precision" is unconventional, and may be confusing to less technically-oriented readers. More common terms would be "machine epsilon" or "unit roundoff". Consider writing "an edge with the smallest possible non-zero weight" or something similar instead.

Reviewer #3:

In this manuscript, the authors report a network-based classification model with applications in multi-omics integration and personalised medicine. The mathematical framework appears to be solid and builds upon Patient Similarity Networks, an approach that has previously been applied for unsupervised clustering (Wang et al, 2014 and Lie et al, 2015). The authors extend this framework to a supervised setting, with the aim to classify patients while identifying relevant discriminatory features.
Important advantages of this method compared to conventional classification algorithms are the natural integration of different data modalities, as well as gains in interpretability by combining features into predefined gene sets that reflect biological knowledge.

Overall, the manuscript is clear, well-structured and the specific applications are of broad interest. However, we have technical concerns that would need to be addressed. In addition, we request access to the software implementation for review purposes.

Major comments:
When using netDx using single features, rather than aggregated gene sets, the model appears to be conceptually related to conventional approaches, such as logistic regression of random forest classifical. Yet, the authors show in Figure 2, that the model offers practical performance gains. Can the authors provide further details where these gains come from?
The method builds on a series of filtering- and selection steps to identify the most relevant features. Given these operations, we feel that there is a risk of an implicit multiple testing, that would need to be controlled for. We would request additional controls, for example using permutations to define empirical null relationships, to assess the robustness of the approach.
In the quantitative assessment of the method (Figure 2), are the described steps to filter specific features only applied to netDx? Differences in data preprocessing and filtering hampers an objective comparison with alternative methods. It would seem important to tease apart performance differences of the actual methods versus differences in pre-processing of the data.
The authors show that the method yields interpretable classifications (Figure 3,4). However, these results are not compared to alternative methods. We would request that the author contrast these results to gene set enrichment analysis applied to the output of classical univariate methods (i.e. random forests or logistic regression), as well as methods that can cope with multi-omics data (e.g. DIABLO, implemented in the mixOmics package, Rohart et al 2017).
Specifically in the multi-omics settings, it is common that entire assays are missing for a subset of the samples. Can the model cope with this structure of missing data?

Minor comments:
Is there a stochastic component on the model? If so, can the authors show consistency statistics across multiple applications of the method?
An explanation for the asterisks in Figure 2 is missing.

1st Revision - authors' response 29th November 2018

netDx methods paper: MSB-18-8497

Responses are in bold under individual points raised by the reviewers.

Reviewer #1:

The authors describe netDx, a procedure for performing supervised machine learning using similarity metrics. In this case, features are collapsed to similarity scores (usually just a Pearson correlation, going by Supplementary Table 1) before being fed into a machine learning algorithm.

Though the performance appears to be good, I also have some concerns related to the presentation. I may also have concerns about the underlying work, but I don't know that because the methods section appears to be relatively incomplete. It would be impossible for me to provide a substantive peer review on the methods/results given what is currently available.

The authors only lightly describe which exact performance metric goes into Figure 2 for netDx and how that performance is obtained. Is the variability from performance over splits, over models, etc? I need this information to make any judgement on the manuscript's validity.

Thank you for this important feedback. In response to this and the feedback of other reviewers, we have added substantially more detail to the manuscript in the form of supplementary tables, figures, new sections and expanded text in the Results section. In addition, all the code and data to reproduce results in this study have been packaged as a Docker container to make it easier to see exactly how the results were generated.

In Figure 2, a box plot shows performance variation over models, where each model uses a different combination of input data types (e.g. clinical + mRNA). For each model, performance is measured using average AUROC across 20 train/test splits run on the full netDx pipeline (including feature selection). We clarified this in the caption text. Also:

1. **We added a new Figure EV1 which shows a detailed workflow of the netDx classifier algorithm and highlights stochastic components of the method.**
2. **To detail the comparison of the workflow of netDx and PanCancer survival analysis (processing, predictor design, univariate filtering choices), we added a new Figure EV3 and Appendix Table S1.**
3. **To support reproducibility and review, we provide reviewers with a Docker image with netDx software installed, code for all analyses reported in this manuscript, data, pre-generated results and documentation so that figures can be reproduced (Available at: https://tinyurl.com/y9b4alfl). This will be part of the software distribution once netDx is publicly released.**

Though the limited description makes it hard to assess the methods and results, I can provide some feedback on the framing. The authors describe this approach as being used for precision medicine. I have considerable numbers of concerns around this point:
* Similarity-based methods by design leak the information of participants. It's unclear how this can be mitigated. A classifier can at least be distributed, but for similarity you need to distribute enough information for someone else to calculate similarity + the clinical information.

This is an excellent point. The scenario we had envisaged in terms of data sharing is that a classifier could be created as a service where Party A (classification user and patient data provider) shares a set of patient similarities and anonymous patient labels with Party B (the

classification service) and the results can be returned in terms of anonymized patient identifiers. In this case, the patient data remains private with Party A and only the similarities need to be shared with Party B. The similarity computing code would be available for Party A to use to convert their data into the similarity matrix form that Party B needs. Classification results are sent back to Party A. Party B never needs to see private patient information. This scenario is relevant to research teams who are analyzing their own patient cohorts.

However, in light of the reviewer comment, this is not relevant in a production classifier scenario, where a generally useful classifier is shared with the world while also maintaining the privacy of the training data. In this scenario, Party A creates a classifier and wants to share the classifier with Party B, but without sharing any of the patient information used to create the classifier. This of course is not possible with our current system, as pointed out, because classification would require computing similarities between new patients and those in the training data. A traditional classifier would not have this problem, as it takes as input either the raw data or a selected set of informative variables that could be used to classify new data without sharing the original training data.

We may be able to address this in a future system in a few ways:
1) We could develop a new method that computes similarities to the training set samples given a set of public training features (e.g. gene expression values from a publicly available data set such as The Cancer Genome Atlas). A subset of data features could be selected to reduce the need to access all the raw data.
2) Required training data could be shared under standard secure and ethical mechanisms, including encrypted data transfer and data access and use agreements, similar to how dbGAP or EGA works with many existing data sets.
3) We could use differential privacy techniques to anonymize the training data in a way that enables us to share it without enabling re-identification, but still enables using it to compute similarities (e.g. by using noise to mask training data signal and prevent identification, but still maintaining enough signal to compute similarities).
4) We could use distance preserving homomorphic encryption from the field of privacy-preserving data mining (e.g. to enable secure clustering - http://insticc.org/node/TechnicalProgram/ic3k/presentationDetails/68908).

In any case, netDx does not require data privacy issues to be solved to be useful, thus we have removed claims related to this point and instead briefly discuss the need for future work in this area.

* "netDx provides a complete framework for precision medicine." It does not appear to actually provide this. It would be better to more accurately define the scope of what netDx provides.

We have toned down this claim and now claim that it can be a useful tool for precision medicine research.

The checklist is both too specific to netDx and too incomplete all at the same time. It might be worth considering whether or not this would be better handled via a separate review or perspective paper. If it's going to be handled here, please consider the extent to which it is overly tuned for this context. For example "... in the integrated similarity network" appears to apply, in all the copious literature on the topic, to only this paper. Also, steps that seem relevant like independent validation in a separate cohort and prospective trials are both missing. Finally, the term orthogonal appears to be applied to something that does not appear to be orthogonal: "in the context of survival prediction in cancer, a predictor should result in significantly separable survival curves for the two predicted patient sets."

We have removed the predictor checklist idea from this manuscript. We agree that it could be part of a review paper.

Finally, the largest problems associated with using similarity-based methods for precision medicine appear to be unaddressed. It requires a large bank of samples to be able to begin to make predictions, and it's unclear how similarity scores will generalize across all platforms. Since every value from every data type can contribute to the predictor, it may be substantially harder to deploy these methods in a production setting. This doesn't need methodological work within this exact contribution, but should be acknowledged as a limitation. As framed though, this appears to be a glaring weakness of both the method and the lack of independent test sets measured by different groups.

This is true and also a limitation for most machine learning methods. We have acknowledged this limitation and changed the framing to clarify the more realistic scope and utility of the method.

Minor:

* Remove the discussion of "borderline significant." Since the decision has been made to use this approach to assessing the robustness of results, either it is or it is not significant depending on the selected alpha.

Done.

Reviewer #2:

In this manuscript, the authors introduce a new method for survival-classification of patient by integration of omics and clinical data. The methodology is novel and generally well described. The evaluation adequately demonstrates that the method provides an advantage over conventional machine learning methods, and provides novel biological insights when integrated with molecular pathway information. The manuscript suffers from a number of unclear/imprecise formulations, and is missing some detailed information here and there that should be provided in the supplements. If these shortcomings are addressed, I believe the manuscript is excellent and suitable for publication.

Page 6, line 4 from bottom: You write that feature networks are integrated using "an established association network integration algorithm" and cite two papers (ref. 12 and 13), the first one being the 3Prop, and the second being an update paper for GeneMANIA. It appears the 3Prop algorithm is used, but this is not perfectly clear. I suggest mentioning 3Prop by name and/or removing ref. 13, as it is not relevant here.

Apologies, the link to 3Prop was incorrect. We use the original GeneMANIA algorithm. The text has been updated to reflect this and the citation is now for the 2008 GeneMANIA algorithm paper.

The performance of NetDx is described as "excellent" in several places. While this is up to interpretation, the mean AUROCs generally vary between 0.6 and 0.8, and judging from Supp. table 1, the increase in performance from data integration seems to come mostly from the clinical data. While NetDx does perform comparatively well to the other methods, "excellent" is, to me, misleading, especially when used in the abstract without context.

The word "excellent" has now either been replaced in all the locations where it is used in the abstract and text.

Page 22, line 4 from bottom (Figure 2 caption): You write that you show a "representative split" from the survival models trained on KIRC. Please specify what constitutes representative in this context. Given the high variance in performance of the different splits, this is a significant detail.

We used a split with AUROC closest to the average AUROC for all splits. We updated Figure 2 caption accordingly.

Page 22, Figure 2: A one-sided Mann-Whitney U-test is used for comparing netDx performance to all other methods. Please provide a justification for using a one one-tailed test instead of a two-tailed test.

We used the one-tailed test to ascertain strictly higher performance of netDx relative to other methods. This has been clarified in the manuscript.

It would also be beneficial to include significance tests between netDx and each method individually (in supplementary table.)

Appendix Table S3 now shows this for each compared machine learning method. The median performance by netDx (AUROC of 0.67) exceeds that of all other methods (AUROC ranging from 0.615 for random forests to 0.655 for Nearest Centroids). Nominal p values are less than 0.05 for 6 out of 8 comparisons.

Supplementary table 1: Please include variance of AUROC values in addition to mean. Also, please include AUROC values for methods compared to as well for reference.

We have updated Appendix Table S2 (previously Supplementary Table 1) to include the variance of AUROC values.

Appendix Table S4 now has the values reproduced from the original PanCancer survival paper, including the various methods to which netDx was compared.

Online methods, page 4, page mid: "an edge weight at floating-point precision". The term "floating-point precision" is unconventional, and may be confusing to less technically-oriented readers. More common terms would be "machine epsilon" or "unit roundoff". Consider writing "an edge with the smallest possible non-zero weight" or something similar instead.

We agree and updated the relevant sentences to use "smallest possible non-zero weight".

Reviewer #3:

In this manuscript, the authors report a network-based classification model with applications in multi-omics integration and personalised medicine. The mathematical framework appears to be solid and builds upon Patient Similarity Networks, an approach that has previously been applied for unsupervised clustering (Wang et al, 2014 and Lie et al, 2015). The authors extend this framework to a supervised setting, with the aim to classify patients while identifying relevant discriminatory features.

Important advantages of this method compared to conventional classification algorithms are the natural integration of different data modalities, as well as gains in interpretability by combining features into predefined gene sets that reflect biological knowledge.

Overall, the manuscript is clear, well-structured and the specific applications are of broad interest. However, we have technical concerns that would need to be addressed. In addition, we request access to the software implementation for review purposes.

Thanks for the positive comments. We apologize that you were not able to access the software implementation. Details for access were in the abstract of the original submission, but may have been missed because of a page break at the end of the Abstract. "(Reviewer note: download at http://netdx.org/index.php/netdx-reviewer-page/ (password psn123). Upon publication, netDx will be made publicly available via GitHub.)"

For easier reproducibility, we have now created a Docker container with all the software and data needed to reproduce the analyses in this manuscript. The container comes with netDx installed, and a README file details the code needed to run every analysis, or plot results from pre-generated runs. These files have been made available as part of the resubmission in a shared Dropbox folder at: https://tinyurl.com/y9b4alfl .

Major comments:
When using netDx using single features, rather than aggregated gene sets, the model appears to be conceptually related to conventional approaches, such as logistic regression of random forest classifical. Yet, the authors show in Figure 2, that the model offers practical performance gains. Can the authors provide further details where these gains come from?

Where the univariate (one network per gene) feature model performs the best, netDx does not, in fact, outperform other methods on average. The table below shows all instances where the best-performing netDx model involves univariate features; the average AUROC for netDx is 0.57 (variance=0.02), and average AUROC for other methods is 0.60 (variance=0.015), though these are not significantly different from each other (p=0.3 from two-tailed t-test).

Tumour type	Datatype	netDx best (univariate model)	Other methods. Yuan et al. (2014)
GBM	cnv	0.56	0.51
GBM	DNAm	0.58	0.59
GBM	miRNA	0.52	0.57
GBM	rna	0.59	0.59
LUSC	clinical, rna	0.66	0.67
LUSC	miRNA	0.44	0.57
LUSC	rna	0.60	0.67
OV	cnv	0.66	0.61
OV	DNAm	0.57	0.61
OV	miRNA	0.55	0.62
average		**0.57**	**0.60**

However, we can comment on the performance gain for models with aggregate features (e.g. a single feature based on all genes in transcriptomic data). For instance, for KIRC, the basic netDx model outperformed other methods without any tuning (e.g. average AUROC across all models was 0.75 for netDx and 0.70 for other methods combined). Some performance gain may be due to the use of regularized regression in GeneMANIA; the original GeneMANIA algorithm benchmarking in the context of gene function prediction found that this reduced overfitting and resulted in superior performance relative to the use of unregularized regression.

In other instances (LUSC and OV, and some GBM models), tuning the sparsification level of the input networks - e.g. keeping the strongest X% of edges per patient - changed netDx's performance sufficiently to outperform other methods. Sparsification parameters include:

1) **How many of each patient's strongest edges to retain (i.e. top 30-50 per patient); the optimal parameter choice differed depending on the data set (described in the Methods section, "Sparsification of Input Networks")**
2) **Applying an exponential scaling filter on edge weights (described in Methods section, "PanCancer Survival Benchmark Models: Variable prefiltering and Scaled Euclidean / Pearson")**

These types of parameter changes were also found to positively influence the performance of GeneMANIA in the context of gene function prediction and presumably achieves performance gains by accentuating signal and reducing noise.

Therefore, we attribute the improved performance of netDx to the fine-tuning of the input networks through sparsification and to the use of regularized regression in the GeneMANIA algorithm used by netDx.

We have added a paragraph to the discussion section of the manuscript, discussing the value of regularization and sparsification to improved classifier performance (Page 19, bottom).

The method builds on a series of filtering- and selection steps to identify the most relevant features. Given these operations, we feel that there is a risk of an implicit multiple testing, that would need to be controlled for. We would request additional controls, for example using permutations to define empirical null relationships, to assess the robustness of the approach.

All of the filtering and selection steps are part of the feature selection process, which is solely concerned with ranking features, not testing their significance for outcome prediction. Standard multiple testing correction procedures, such as FDR, would not change this ranking. We are careful that the feature selection process is applied only to training and not test samples. Additionally, we repeat this feature selection step on multiple splits of the training data, which helps assess robustness.

We performed 20 train/test data splits for each of the 40 netDx prediction models we report (performance measures stabilized at 20 splits). For models with pathway-level features (breast cancer and asthma), we chose 100 splits to ensure variance in pathway-level feature scores stabilized (Appendix Figure S2). We have clarified these points in the manuscript:
1) The role of prefiltering in ranking features (Page 20, "PanCancer Survival benchmark models")
2) Multiple train/test splits to stabilize performance on blind test samples (Page 11, "Benchmarking performance by predicting binarized survival in cancer"), and to achieve stable pathway-level scores (Page 13, "Pathway-level feature selection identifies cellular processes predictive of clinical condition")

In the quantitative assessment of the method (Figure 2), are the described steps to filter specific features only applied to netDx? Differences in data preprocessing and filtering hampers an objective comparison with alternative methods. It would seem important to tease apart performance differences of the actual methods versus differences in pre-processing of the data.

We now include a more detailed description of the various data preprocessing steps and differences in predictor design, and compare netDx to the PanCancer survival analysis methods in Appendix Table S1, as well as in Figure EV3. In summary:

1) **We downloaded already processed data from the PanCancer survival project (Yuan et al.) (https://www.synapse.org/#!Synapse:syn1710282/wiki/27303)**
2) **We used identical variable coding to Yuan et al.**
3) **Similar to Yuan et al. we use prefiltering within a resampling feature selection loop, except that we use lasso regression (because it is an established method known to have good performance) while they use ANOVA and shrunken centroids.**
4) **Similar to Yuan et al. we use imputation within the training/resampling loop, except that we use it only for GBM and only use imputation by median (not by mode), because we found this choice sufficient to provide a performance improvement.**

The authors show that the method yields interpretable classifications (Figure 3,4). However, these results are not compared to alternative methods. We would request that the author contrast these results to gene set enrichment analysis applied to the output of classical univariate methods (i.e. random forests or logistic regression), as well as methods that can cope with multi-omics data (e.g. DIABLO, implemented in the mixOmics package, Rohart et al 2017).
Specifically in the multi-omics settings, it is common that entire assays are missing for a subset of the samples. Can the model cope with this structure of missing data?

1) We added a new paragraph in the Discussion section, comparing netDx to both GSEA and DIABLO, and two accompanying supplementary figures (Appendix Figure S3 and S4), one per comparison. In summary:

- **A. netDx selects qualitatively similar pathway themes as GSEA but is more conservative (selecting fewer pathways). We attribute this stricter selection of pathways to redundancy-reduction by regularization in the feature-scoring step.**
- **B. We compare netDx and DIABLO by integrating RNA and miRNA for binary classification of breast tumours. Both have similar accuracy (91% for netDx and 90% for DIABLO). netDx has built-in support for pathway selection, which provides insight into cellular processes that discriminate between classes. DIABLO selects individual data variables (e.g. microRNA or mRNA expression levels) within each data type, thus is better suited for e.g. developing a panel of biomarker molecules. Unfortunately DIABLO currently has no built in functionality to easily generate pathway-level features and visualize the resulting selected features (e.g. there is nothing similar to netDx's EnrichmentMap view). We therefore conclude that DIABLO and netDx provide complementary views (i.e. pathway-level in netDx vs. molecule-level in DIABLO) of the system and could be useful in tandem.**

2) Yes, netDx can handle entire assays missing for a subset of samples. Consider a dataset with RNA, DNAm, and miRNA data. Suppose patients {A,B,C,D,E} all have RNAseq measurements. E does not have DNAm, and C,D,E don't have miRNA. During patient similarity network construction, all five patients would be used for the RNA network; only A-D would be used for the DNAmeth network; and only A-B would be used for the miRNA network. All five patients are still represented in one or more features. The GeneMANIA algorithm is still able to integrate networks, based on the available data.

Minor comments:
Is there a stochastic component on the model? If so, can the authors show consistency statistics across multiple applications of the method?

Yes, there are three stochastic components in the model; the table below lists these and how stochasticity is handled in the model as well as which consistency statistics are used. To make the stochastic components transparent to the reader, we have added a new supplementary figure that shows the detailed workflow of netDx, highlighting components with stochasticity (Figure EV1).

Stochastic component	**Consistency statistics**
Splitting of samples into train and blind test	We report performance over multiple train/test splits. Variability of AUROC scores are provided (reported in Appendix Supplementary Table S2).
Univariate filtering (optional). We use lasso regression, which uses stochastic resampling.	The netDx pipeline sets the random number generator seed once for each train/test split, thus univariate filtering varies per split. We have added a supplementary figure showing the frequency with which individual variables are selected by regularized

	regression, using an ovarian cancer predictor as an example (Appendix Figure S1).
Feature selection uses stochastic resampling to evaluate stability.	Stability is reflected in the scores of individual features, measured by how frequently they are selected across resamplings. Appendix Table S5 and 6 show feature scores for the breast cancer and asthma example.

An explanation for the asterisks in Figure 2 is missing.

Thanks for catching this. The asterisks had indicated instances where netDx performance was significantly better than other methods ($p < 0.05$ from a one-tailed Wilcoxon-Mann-Whitney). The asterisk has been replaced by p-values.

2nd Editorial Decision 5th February 2019

Thank you for sending us your revised manuscript. I apologize for the delay in sending you a decision, which was due to the fact that we never received a report from reviewer #3 after a series of reminders. In order to not delay the process further, we have decided to proceed with making a decision based on our evaluation of your responses to reviewers #2 and #3 and on the comments of reviewer #1. We consider the issues raised by reviewers #2 and #3 resolved by the preformed revisions and clarifications. As you will see below, reviewer #1 also thinks that the study is now suitable for publication, pending some minor text modifications, which we would ask you to perform in a minor revision.

Before we formally accept the paper for publication we would also ask you to address the following editorial issues.

Reviewer #1:

The priority claim here seems not well justified: "It is the first classifier to apply the idea of recommender systems, similar to those used in Amazon or Netflix ("find movies like this one"), to precision medicine ("find patients who don't respond to therapy")." A quick search finds the term recommender system applied to treatment prioritization from at least as early as 2014.

My other concerns around the framing have been addressed. Some revisions would make this easier to read, like picking a term and sticking with it. For example, "(we use the terms "input networks" and "features" interchangeably)" just makes the manuscript a bit harder to read. A careful editing for clarity and ease of reading would be helpful to the reader, but wouldn't require re-review.

2nd Revision - authors' response 12th February 2019

Responses to Reviewer #1

The priority claim here seems not well justified: "It is the first classifier to apply the idea of recommender systems, similar to those used in Amazon or Netflix ("find movies like this one"), to precision medicine ("find patients who don't respond to therapy")." A quick search finds the term recommender system applied to treatment prioritization from at least as early as 2014.

Response: We thank the reviewer for re-reviewing our manuscript. The claim above has been toned down. The text now reads: "[netDx] applies the idea of recommender systems, similar to those used in Amazon or Netflix (“find movies like this one”), to precision medicine (“find patients who don’t respond to therapy”)."

My other concerns around the framing have been addressed. Some revisions would make this easier to read, like picking a term and sticking with it. For example, "(we use the terms "input networks" and "features" interchangeably)" just makes the manuscript a bit harder to read. A careful editing for clarity and ease of reading would be helpful to the reader, but wouldn't require re-review.

Response: As per the request, we have read through the manuscript and revised the text. We now consistently use the term "feature", as it is standard machine-learning terminology and clarifies the use of networks in the context of the machine-learning algorithm. We reserve the use of the term "network" for those instances where the network properties are relevant to the context of the sentence.

Accepted 13th February 2019

Thank you again for sending us your revised manuscript. We are now satisfied with the modifications made and I am pleased to inform you that your paper has been accepted for publication.

EMBO PRESS

YOU MUST COMPLETE ALL CELLS WITH A PINK BACKGROUND ↓

PLEASE NOTE THAT THIS CHECKLIST WILL BE PUBLISHED ALONGSIDE YOUR PAPER

Corresponding Author Name: Gary Bader
Journal Submitted to: Molecular Systems Biology
Manuscript Number: MSB-18-8497

Reporting Checklist For Life Sciences Articles (Rev. July 2015)

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. These guidelines are consistent with the Principles and Guidelines for Reporting Preclinical Research issued by the NIH in 2014. Please follow the journal's authorship guidelines in preparing your manuscript.

A- Figures

1. Data

The data shown in figures should satisfy the following conditions:

➔ the data were obtained and processed according to the field's best practice and are presented to reflect the results of the experiments in an accurate and unbiased manner.
➔ figure panels include only data points, measurements or observations that can be compared to each other in a scientifically meaningful way.
➔ graphs include clearly labeled error bars for independent experiments and sample sizes. Unless justified, error bars should not be shown for technical replicates.
➔ if n< 5, the individual data points from each experiment should be plotted and any statistical test employed should be justified
➔ Source Data should be included to report the data underlying graphs. Please follow the guidelines set out in the author ship guidelines on Data Presentation.

2. Captions

Each figure caption should contain the following information, for each panel where they are relevant:

➔ a specification of the experimental system investigated (eg cell line, species name).
➔ the assay(s) and method(s) used to carry out the reported observations and measurements
➔ an explicit mention of the biological and chemical entity(ies) that are being measured.
➔ an explicit mention of the biological and chemical entity(ies) that are altered/varied/perturbed in a controlled manner.

➔ the exact sample size (n) for each experimental group/condition, given as a number, not a range;
➔ a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates (including how many animals, litters, cultures, etc.).
➔ a statement of how many times the experiment shown was independently replicated in the laboratory.
➔ definitions of statistical methods and measures:
- common tests, such as t-test (please specify whether paired vs. unpaired), simple χ2 tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
- are tests one-sided or two-sided?
- are there adjustments for multiple comparisons?
- exact statistical test results, e.g., P values = x but not P values < x;
- definition of 'center values' as median or average;
- definition of error bars as s.d. or s.e.m.

Any descriptions too long for the figure legend should be included in the methods section and/or with the source data.

Please ensure that the answers to the following questions are reported in the manuscript itself. We encourage you to include a specific subsection in the methods section for statistics, reagents, animal models and human subjects.

In the pink boxes below, provide the page number(s) of the manuscript draft or figure legend(s) where the information can be located. Every question should be answered. If the question is not relevant to your research, please write NA (non applicable).

USEFUL LINKS FOR COMPLETING THIS FORM

http://www.antibodypedia.com
http://1degreebio.org
http://www.equator-network.org/reporting-guidelines/improving-bioscience-research-repo

http://grants.nih.gov/grants/olaw/olaw.htm
http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Useofanimals/index.htm
http://ClinicalTrials.gov
http://www.consort-statement.org
http://www.consort-statement.org/checklists/view/32-consort/66-title

http://www.equator-network.org/reporting-guidelines/reporting-recommendations-for-tum

http://datadryad.org

http://figshare.com

http://www.ncbi.nlm.nih.gov/gap

http://www.ebi.ac.uk/ega

http://biomodels.net/

http://biomodels.net/miriam/
http://jjj.biochem.sun.ac.za
http://oba.od.nih.gov/biosecurity/biosecurity_documents.html
http://www.selectagents.gov/

B- Statistics and general methods

Please fill out these boxes ↓ (Do not worry if you cannot see all your text once you press return)

1.a. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?	No data was generated in this study. Sample sizes were determined by the data downloaded for method comparison. In all instances the full dataset was used, without further downsampling.
1.b. For animal studies, include a statement about sample size estimate even if no statistical methods were used.	NA
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?	NA
3. Were any steps taken to minimize the effects of subjective bias when allocating animals/samples to treatment (e.g. randomization procedure)? If yes, please describe.	NA
For animal studies, include a statement about randomization even if no randomization was used.	NA
4.a. Were any steps taken to minimize the effects of subjective bias during group allocation or/and when assessing results (e.g. blinding of the investigator)? If yes please describe.	Software was used to randomly allocate samples to training or test groups. This process was repeated multiple times (10 or 20 times, depending on the test) to ensure adequate cross-validation. All sampling was performed using pseudo-random number generator software , therefore no experimenter bias was involved.
4.b. For animal studies, include a statement about blinding even if no blinding was done	NA
5. For every figure, are statistical tests justified as appropriate?	Yes
Do the data meet the assumptions of the tests (e.g., normal distribution)? Describe any methods used to assess it.	Where centers of two distributions were compared, we always used the nonparametic Wilcoxon Mann-Whitney test. This test does not require assumption of normality.
Is there an estimate of variation within each group of data?	Yes, we provide standard errors of the mean or variance in all instances.
Is the variance similar between the groups that are being statistically compared?	Yes

C- Reagents

6. To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile. e.g., Antibodypedia (see link list at top right), 1DegreeBio (see link list at top right).	NA
7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination.	NA

* for all hyperlinks, please see the table at the top right of the document

D- Animal Models

8. Report species, strain, gender, age of animals and genetic modification status where applicable. Please detail housing and husbandry conditions and the source of animals.	NA
9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.	NA
10. We recommend consulting the ARRIVE guidelines (see link list at top right) (PLoS Biol. 8(6), e1000412, 2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under 'Reporting Guidelines'. See also: NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance.	NA

E- Human Subjects

11. Identify the committee(s) approving the study protocol.	NA ; All data used in this manuscript was previously generated and is now publicly available. As a consequence, no ethics approval was required and there was never direct interaction with human subjects.
12. Include a statement confirming that informed consent was obtained from all subjects and that the experiments conformed to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report.	NA ;All data used in this manuscript was previously generated and is now publicly available. As a consequence, no ethics approval was required and there was never direct interaction with human subjects.
13. For publication of patient photos, include a statement confirming that consent to publish was obtained.	NA ;All data used in this manuscript was previously generated and is now publicly available. As a consequence, no ethics approval was required and there was never direct interaction with human subjects.
14. Report any restrictions on the availability (and/or on the use) of human data or samples.	NA ;All data used in this manuscript was previously generated and is now publicly available. As a consequence, no ethics approval was required and there was never direct interaction with human subjects.
15. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable.	NA
16. For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under 'Reporting Guidelines'. Please confirm you have submitted this list.	NA
17. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under 'Reporting Guidelines'. Please confirm you have followed these guidelines.	NA

F- Data Accessibility

18. Provide accession codes for deposited data. See author guidelines, under 'Data Deposition'. Data deposition in a public repository is mandatory for: a. Protein, DNA and RNA sequences b. Macromolecular structures c. Crystallographic data for small molecules d. Functional genomics data e. Proteomics and molecular interactions	No data were generated for this study. All referenced data has been cited in the Data Availability section of the manuscript.
19. Deposition is strongly recommended for any datasets that are central and integral to the study; please consider the journal's data policy. If no structured public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under 'Expanded View' or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).	All data used in this manuscript was previously generated and is now publicly available. This study simply downloaded and analyzed these data. In all instances, the public data accession codes and online location has been mentioned in the methods section of the report.
20. Access to human clinical and genomic datasets should be provided with as few restrictions as possible while respecting ethical obligations to the patients and relevant medical and legal issues. If practically possible and compatible with the individual consent agreement used in the study, such data should be deposited in one of the major public access-controlled repositories such as dbGAP (see link list at top right) or EGA (see link list at top right).	All data used in this manuscript was previously generated and is now publicly available. This study simply downloaded and analyzed these data. Therefore this point is not applicable.
21. As far as possible, primary and referenced data should be formally cited in a Data Availability section. Please state whether you have included this section. Examples: **Primary Data** Wetmore KM, Deutschbauer AM, Price MN, Arkin AP (2012). Comparison of gene expression and mutant fitness in Shewanella oneidensis MR-1. Gene Expression Omnibus GSE39462 **Referenced Data** Huang J, Brown AF, Lei M (2012). Crystal structure of the TRBD domain of TERT and the CR4/5 of TR. Protein Data Bank 4O26 AP-MS analysis of human histone deacetylase interactions in CEM-T cells (2013). PRIDE PXD000208	Yes we have included this section
22. Computational models that are central and integral to a study should be shared without restrictions and provided in a machine-readable form. The relevant accession numbers or links should be provided. When possible, standardized format (SBML, CellML) should be used instead of scripts (e.g. MATLAB). Authors are strongly encouraged to follow the MIRIAM guidelines (see link list at top right) and deposit their model in a public database such as Biomodels (see link list at top right) or JWS Online (see link list at top right). If computer source code is provided with the paper, it should be deposited in a public repository or included in supplementary information.	The software implementation of our method and all code to generate results in the paper have been made available via a public GitHub repository (upon publication).

G- Dual use research of concern

23. Could your study fall under dual use research restrictions? Please check biosecurity documents (see link list at top right) and list of select agents and toxins (APHIS/CDC) (see link list at top right). According to our biosecurity guidelines, provide a statement only if it could.	NA