RiskPath: Explainable deep learning for multistep biomedical prediction in longitudinal data

Nina de Lacy[1], Wai Yin Lam[2], Michael Ramshaw[1]

1: Department of Psychiatry, University of Utah, Salt Lake City, Utah
2: Scientific Computing Institute, University of Utah, Salt Lake City, Utah

Abstract

Predicting individual and population risk for disease outcomes and identifying persons at elevated risk is a key prerequisite for targeting interventions to improve health. However, current risk stratification tools for the common, chronic diseases that develop over the lifecourse and represent the majority of disease morbidity, mortality and healthcare costs are aging and achieve only moderate predictive performance. In some common, highly morbid conditions such as mental illness no risk stratification tools are yet available. There is an urgent need to improve predictive performance for chronic diseases and understand how cumulative, multifactorial risks aggregate over time so that intervention programs can be targeted earlier and more effectively in the disease course. Chronic diseases are the end outcomes of multifactorial risks that increment over years and represent cumulative, temporally-sensitive risk pathways. However, tools in current clinical use were constructed in older data and utilize inputs from a single data collection step. Here, we present RiskPath, a multistep deep learning method for temporally-sensitive biomedical risk prediction tailored for the constraints and demands of biomedical practice that achieves very strong performance and full translational explainability. RiskPath delineates and quantifies cumulative multifactorial risk pathways and allows the user to explore performance-complexity tradeoffs and constrain models as required by clinical use cases. Our results highlight the potential for developing a new generation of risk stratification tools and risk pathway mapping in time-dependent diseases and health outcomes by leveraging powerful timeseries deep learning methods in the wealth of biomedical data now appearing in large, longitudinal open science datasets.

Introduction

Risk stratification, the identification of individuals or populations at elevated risk for a particular disease or condition, is a cornerstone of clinical practice and public health. Appropriate risk stratification allows risk mitigation by targeting interventions, implementing preventative strategies or choosing among treatments. Continuous biomedical innovation has also spurred a growing focus on personalized medicine, where interventions are tailored to the individual or subpopulation based on their predicted response to a specific treatment. Effective risk stratification is built on a foundation of predictive modeling, where the overall objective is to explain outcome variance as well as possible and identify statistical regularities that generalize to other data with similar distributions. Importantly, effective predictive models do not necessarily include disease causes, and “cheap and easy to measure surrogates or biomarkers of causes” may be included and even preferred. [1] For instance, the ubiquitous Framingham Risk Score contains non-causal predictors of cardiac disease risk like High Density Lipoprotein or ‘Good’ Cholesterol. [2]

Because the goal of prediction is to explain outcome variance as well as possible, maximizing model accuracy and/or minimizing error is the desirable performance criterion. In biomedicine, positive predictive value (precision or PPV) sensitivity (recall) and specificity are also high priorities. Accordingly, there has been increasing interest in deep learning, given its empirical ability to leverage overparameterized regimes and achieve strong performance. For instance, convolutional neural networks (CNNs) applied to imaging data have built on advances in computer vision to produce highly-accurate risk predictions for melanoma, [3] tuberculosis, [4] lung cancer, [5] diabetic retinopathy[6] and macular edema[7] with clinically-deployable classifiers that are equal to or better in discriminating cases than board-certified specialists. [8-10] Performance is very strong, with accuracy of ~0.88-0.95 and good sensitivity and specificity. [7,10,11] Outside of image recognition, deep learning has achieved less purchase in the clinic. In particular, there is a prominent need for improved predictive risk stratification in chronic disease processes such as mental illness, cardiovascular disease, diabetes, chronic obstructive pulmonary disease (COPD), osteoarthritis, colorectal and breast cancer, which are highly prevalent and collectively represent >90% of morbidity, mortality and healthcare costs. [12] Most current risk stratification instruments for chronic conditions are 10-30 years old and were formulated using logistic regression or decision-tree algorithms (e.g. CART). They exhibit only relatively fair predictive performance. For instance, predictive accuracy varies from 0.51-0.83 in standard of care risk calculators in COPD, diabetes, colorectal and

breast cancer and cardiovascular conditions and PPV and specificity is frequently ≤0.15. [2,13-28] Of note, their constituent predictors were manually selected based on prior studies that used linear, groupwise inferential testing to identify factors with significant differences between cases and controls, i.e., associations. Such variables may not necessarily be optimal for prediction since "even statistically strong associations with very low *p* values often shed only modest light on their value for the goal of prediction" [29] and likely lack a uniform theoretical basis. [30-32]

Chronic diseases are the end outcomes of multifactorial risks that increment over years and represent cumulative, temporally-sensitive risk pathways. As such, there are likely multiple, sequential opportunities for mitigation and prevention if risk stratification could incorporate the time-dependent characteristics of these diseases and identified important precedent predictors and their requisite periods for intervention. However, tools in current clinical use were constructed and utilize inputs from a single data collection interval. In essence, improved risk stratification in chronic disease implies shifting from the current approach of contemporaneous or one-period prediction to multistep, sequence-based prediction. In turn, this implies the use of longitudinal data as *ex ante* input features and techniques that render such data tractable for prediction. A premier algorithmic class for deep learning in sequence prediction problems is the recurrent neural network (RNN), which has proven successful in climate science, financial forecasting and speech recognition. RNNs such as Long Short-Term Memory (LSTM) neural networks are typically applied to long timeseries data from one or a few sensors and conventionally ingest square data. However, the preponderance of data that is appropriate and available for large-scale, effective chronic disease risk stratification (e.g., demographic, clinical assays, cognitive testing or social determinants data) has the obverse characteristics, being non-square tabular data available from thousands of people ('sensors') over relatively few timepoints. Moreover, in order to be pragmatically operationalizable, models must be translationally explainable, precluding classical dimensionality reduction, since we must know the identity of the original predictors (in order to collect them anew in the clinic) and ideally understand the predictive cost-benefit tradeoff of streamlining the set of features to a compact representation -- since it is typically too costly or onerous to collect dozens or hundreds of non-imaging variables *de novo*.

To address these challenges, we present RiskPath, a multistep predictive pipeline for temporally-sensitive biomedical risk stratification that achieves very strong performance and is tailored to the constraints and demands of biomedical practice. The core algorithm is a LSTM network that we adapted to data with characteristics common in clinical practice (tabular; non-square; collected annually; ≤10 timepoints) and rendered translationally explainable by extending the Shapley method[33] of computing feature importances for timeseries data and embedding this into the model. We further provide data-driven approaches for streamlining features in timeseries data before and during model training and analyzing performance-complexity tradeoffs in model construction, showing that the inherent topologic complexity of deep learning provides a useful performance buffer to achieve the representational compactness demanded by clinical practice.

Here, we demonstrate RiskPath in major mental illnesses (MI) and hypertension, which are paradigmatic examples of chronic illness. The onset of clinical MI represents the accumulation of multifactorial risk factors over 8-10 years and treatments are currently palliative rather than curative. Because three-quarters of lifetime MI develops in the peri-adolescent lifestage (10-24 years of age), [34-36] patients are affected during the productive years of adulthood, which in the aggregate makes MI the most costly and disabling illness class. [37] However, no current clinical risk stratification instruments exist to reliably predict major MIs, which collectively affect ~20% of the population. [38] In particular, understanding the cumulative process of risk aggregation is a central research priority in the field. [39] Using data from the *Adolescent Brain Cognitive Development* (ABCD) cohort, the largest long-term study of child development in the US, we predicted cases of Anxiety, Depression, Attention Deficit Hyperactivity Disorder (ADHD), Disruptive Behaviors and the total burden of MI symptoms at age 14 -- the point by which 50% of lifetime illness has onset. We constructed holistic, temporally-sensitive disease prediction models that delineate cumulative risk trajectories from 9-14 years of age. Their strong performance in generalization testing (Accuracy and AUC: 0.93-0.98) represents a substantial increment over extant single-period prediction in this domain (~0.70-0.80 accuracy). [40-48] Moreover, RiskPath also delivers very strong precision, recall and specificity. While our primary use cases were the most common mental illnesses, RiskPath is applicable to any disease where it is desirable to provide modernized, time-sensitive foundation for earlier and more accurate risk stratification and longitudinal data is available. To illustrate this adaptability, we also provide an example of using RiskPath to predict hypertension in older adults using data from the

Cardiovascular Health Study (CHS), a longitudinal study of risk factors for development and progression of cardiovascular diseases and stroke in people aged 65 years and older. Here, we constructed prediction models that mapped out cumulative risk trajectories over 10 annual timepoints to achieve performance (Accuracy and AUC: 0.84) that was again superior to the performance of current single-period risk stratification instruments in the cardiovascular domain which is in the range 0.65-0.81 AUC with very low (~≤0.15) precision and specificity.[2,13,14,49]

Results

Overview of the RiskPath pipeline

Applying LSTMs in tabular timeseries big data for biomedical pathway prediction presents three major, interrelated challenges: feature selection, explainability and topologic optimization. LSTMs are customarily used with image, speech or text data and all available features are typically used. However, if a clinically operationalizable model is to be designed for tabular data, feature selection must occur. Typically, non-imaging risk stratification models used in clinical practice have ≤10 features since it is impractical and costly to collect more measures. Further, if new data is to be collected from a patient, the original identity of selected predictors must be exposed, even within a black box model. Hence, extant instruments were based on linear or tree-based models. LSTMs are innately highly parameterized given their complex topology: parameter size (N) quickly and easily exceeds 100,000 and can be in the millions or more. Few participant samples are sized such that N< *n* and therefore learning usually occurs within overparameterized regimes rather than the conventional bias-variance tradeoff zone. While participation in overparameterized learning has been demonstrated to account for the empirically observed strength of deep learning performance, [50] optimizing the topology of timeseries deep learning algorithms remains largely unexplored in the 'double descent' literature pertaining to overparameterization, particularly in the context of obligate feature selection.

Feature selection

The existing generation of risk prediction tools were often constructed with smaller datasets acquired for a specific disease class. Here, features were selected based on domain knowledge, i.e. metrics derived from prior group-level inferential studies. However, good scores in inferential metrics (e.g. *p*-value significance) do not guarantee performance in predictive analyses[29] and it is preferable to select features on the basis of predictive rather than inferential performance. Moreover, longitudinal population-level datasets are now available allowing biomedical discovery across disease classes such as the *UK Biobank* (*n*=500,000), *All of Us* (*n*=100,000) and ABCD (*n*=11,500) cohorts. Such studies collect thousands of variables about participants across time permitting fully data-driven approaches to risk prediction and novel discoveries. To accomplish principled, data-driven feature selection, the first step in our pipeline leverages two gold standard techniques that select features based on predictive performance: the LASSO CV (linear) and Boruta (nonlinear, ensemble) techniques. We adapted each algorithm for timeseries data to compute variable importances across all timepoints and extract a unique set of features that is important across the timeseries. A third, blended algorithm returns the union of the feature set extracted with both methods (linear + nonlinear). These three options are provided since it is not known whether linear or nonlinear feature selection is optimal for deep learning, a nonlinear technique. RiskPath offers users the ability to impose importance thresholds during selection to generate more parsimonious feature sets.

Explainability

Like most deep learning algorithms, LSTMs do not natively return the identity of predictors. To provide multistep prediction algorithms suitable for development into risk stratification tools, we extended the Shapley values technique to timeseries deep learning to render translationally explainable LSTMs. Shapley values are an approach from cooperative game theory widely used in machine learning to compute variable importances, though not historically RNNs. RiskPath integrates Shapley computations into LSTMs and extends the technique to return feature importance within each time period and to compute importances across all time periods. Collectively, this provides a complete picture of predictors and how risk aggregates across and within precedent time epochs to disease onset. This information can be used to guide assessment and intervention planning.

Topologic optimization for performance and utility

LSTMs are among the more complex deep learning architectures since they incorporate input, output and forget gates that tend to balloon the number of model parameters. Generically, the number of parameters in an LSTM is $4(hm + h^2 + h)$, where h = number of hidden units and m = the number of features. Commonly, practitioners use bidirectional layers which further increases the number of parameters. In a learning environment where we wish to constrain the number of features, m, the width of the LSTM h will therefore tend to assume greater prominence. There is no principled method to discover the optimal topology of a deep neural network and many practitioners use heuristics or trial and error. RiskPath embeds the option of an increasing range of model widths which may be controlled by the practitioner to identify a value for h and explore complexity-performance and time-complexity tradeoffs. RiskPath also adapts LSTMs for non-square data, since in both research and clinical practice it is commonly challenging to collect the same assessments in each time period.

Evaluation and performance of baseline models

We evaluated the performance of RiskPath in predicting cases of 4 leading MIs, the total burden of MI and hypertension and compared it with a standard deep learning feedforward (3-layer) network.

a	Training subjects	Number features	Accuracy	AUC	Precision	Recall	Specificity	F1	Loss	Time (secs)
Total MI	1263	176	0.961	0.961	0.945	0.958	0.964	0.961	0.370	1669
Depression	805	187	0.925	0.926	0.911	0.892	0.960	0.924	0.411	1140
Anxiety	967	232	0.938	0.939	0.923	0.918	0.960	0.938	0.386	1721
ADHD	762	149	0.961	0.961	0.950	0.950	0.974	0.961	0.644	1104
Disruptive	514	80	0.981	0.981	0.973	0.982	0.981	0.982	0.621	711
Hypertension	958	216	0.837	0.838	0.800	0.828	0.848	0.840	0.497	2568

b	Training subjects	Number features	Accuracy	AUC	Precision	Recall	Specificity	F1	Loss	Time (secs)
Total MI	1263	243	0.964	0.964	0.953	0.954	0.975	0.964	0.402	83
Depression	805	185	0.933	0.934	0.919	0.908	0.960	0.933	0.415	64
Anxiety	967	139	0.941	0.941	0.961	0.945	0.937	0.942	0.416	69
ADHD	762	196	0.967	0.970	0.962	0.949	0.987	0.968	0.402	63
Disruptive	514	142	0.981	0.981	0.964	1.000	0.961	0.982	0.664	56
Hypertension	958	216	0.821	0.821	0.775	0.820	0.822	0.826	0.540	52

Table 1: Performance of RiskPath compared with standard deep learning in predicting mental illness cases
Performance statistics are shown for the best-performing model in generalization testing in held-out, unseen data for **a** RiskPath with three-dimensional feature inputs (participants * time * features) versus **b** the same feature inputs flattened as two-dimensional inputs to a standard feedforward. In both cases 10-fold cross-validation is performed during training with learning rate=1.00×10^{-5}, decay=0.1 and AdamW optimizer.

The same samples and hyperparameter settings were used and hidden units allowed to vary in the range [5,1200]. Features sets were determined as described above with default settings (non-zero LASSO correlations and $p \leq 0.05$ in the Boruta) and were input to LSTMs as three-dimensional (3D) data and as flattened two-dimensional (2D) data to feedforwards. Performance in testing in held-out, unseen data was very strong with RiskPath across all models and well above 0.85 for MIs (the commonly accepted threshold for clinical utility) across all metrics of interest. Performance was also very similar between 3D and 2D models, with fractional differences observed across standard metrics. We did not detect any directional relationships between feature selection with the LASSO, Boruta or LASSO+Boruta methods with best-performing models arising from all methods. On average, RiskPath LSTM models took 1486 seconds (~25 minutes) to compute versus 65 seconds (~1 minute) for the feedforwards.

Performance sensitivity to feature ablation

To generate models that are more clinically tractable, we explored the sensitivity of results to feature ablation. Because RiskPath computes feature importances across all time periods in a holistic manner, the relative importance of individual features to the model can be exposed and examined. **Figure 1** demonstrates that

feature importance drops off quickly after the first ~10 features and asymptotes over less important (weaker) features. The latter - collectively representing a substantial portion of the entire feature set - were therefore excellent candidates for ablation.

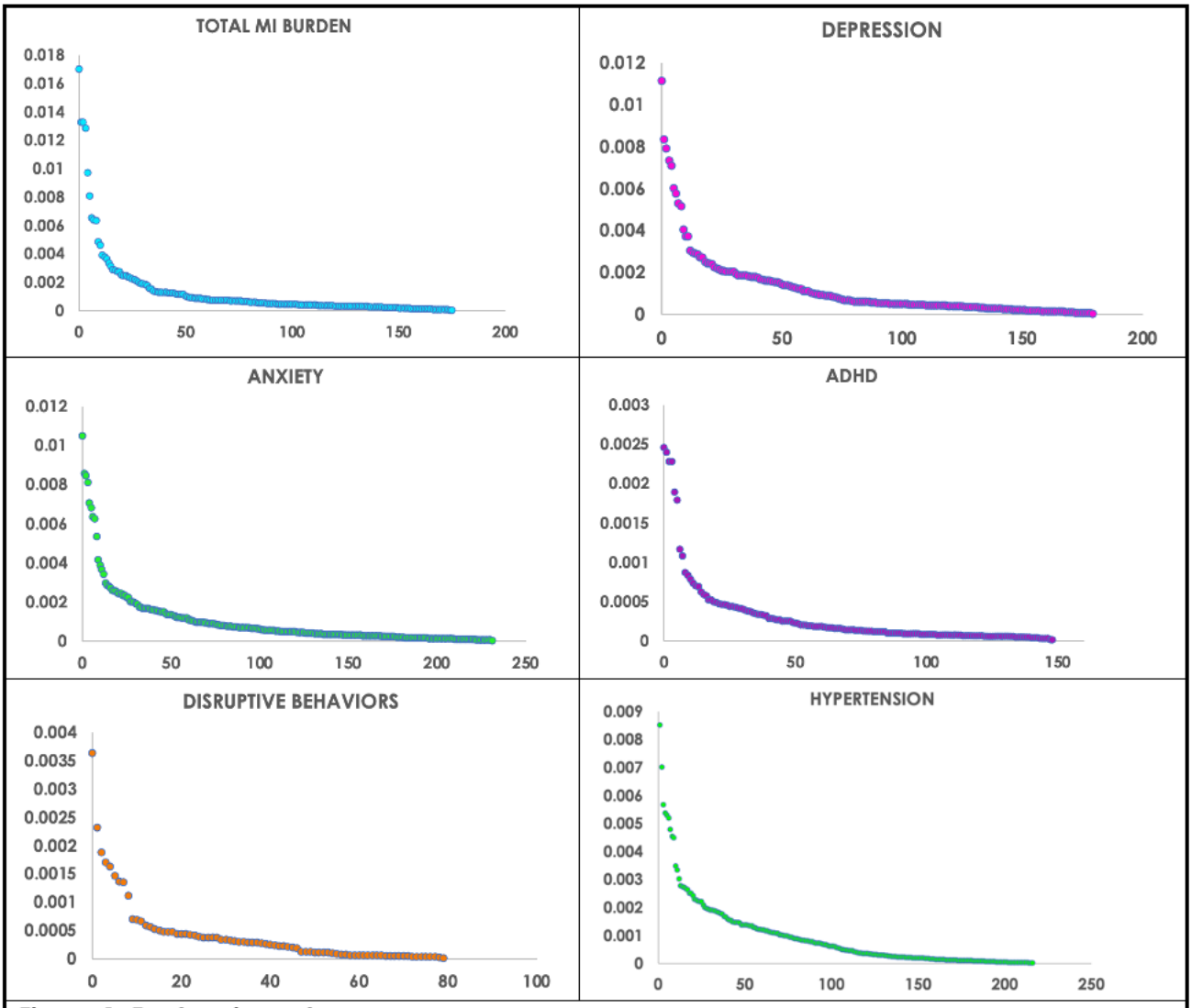

Figure 1: Feature importances
Feature importances are shown for each best-performing RiskPath model corresponding to **Table 1**. Vertical axes display mean Shapley values across time periods. Horizontal axes refer to the relative rank of importances. In all 5 models, feature importance declines quickly over the top 10 ranked features to asymptote.

Accordingly, we performed additional experiments using the feature ablation options offered in RiskPath and re-fit the best performing models. In the first, the number of features was determined at the elbow of the curve shown in **Figure 1** after model training. For example, the 23 most important features in the Anxiety model. In the second, the top 10 most important features, a convenient heuristic for clinical operationalizability, were selected for each condition based on their importances determined after selection with the LASSO CV method but before training any models.

Comparing the metrics presented in **Tables 1** and **2** enables quantification of the impact of feature ablation on performance. We found that while the elbow method ablated an average of >85% of the feature sets across the five conditions, it resulted in only a 2.0% decrement to absolute average accuracy. Preserving just the top 10 features after feature selection prior to training was similarly unobtrusive, resulting in only a 2.2% lowering of average accuracy. Other performance metrics were also minimally impacted. In essence, this demonstrates that the majority of features can be safely ablated without a substantial impact to performance. Notably, in every condition all performance metrics remained >0.85 in MI and ~0.80 in hypertension, further indicating that

clinically important metrics such as PPV, specificity and sensitivity are similarly minimally affected by feature ablation.

a	Training subjects	Number features	Accuracy	AUC	Precision	Recall	Specificity	F1	Loss	Time (secs)
Total MI	1263	27	0.941	0.941	0.913	0.950	0.931	0.942	0.683	2105
Depression	805	21	0.911	0.911	0.875	0.914	0.910	0.914	0.669	1506
Anxiety	967	23	0.905	0.904	0.866	0.918	0.891	0.908	0.683	1696
ADHD	762	16	0.948	0.948	0.921	0.962	0.934	0.950	0.667	1442
Disruptive	514	10	0.957	0.957	0.938	0.963	0.951	0.959	0.679	1082
Hypertension	958	24	0.804	0.825	0.800	0.863	0.833	0.854	0.489	669

b	Training subjects	Number features	Accuracy	AUC	Precision	Recall	Specificity	F1	Loss	Time (secs)
Total MI	1263	10	0.939	0.939	0.913	0.933	0.946	0.940	0.617	2102
Depression	805	10	0.920	0.919	0.881	0.950	0.892	0.923	0.687	1510
Anxiety	967	10	0.910	0.911	0.898	0.857	0.966	0.907	0.552	1701
ADHD	762	10	0.939	0.939	0.917	0.930	0.947	0.939	0.640	1451
Disruptive	514	10	0.962	0.962	0.946	0.963	0.961	0.963	0.698	1087
Hypertension	958	10	0.803	0.804	0.800	0.783	0.825	0.804	0.533	2391

Table 2: Predictive performance of RiskPath after feature ablation
Performance statistics are shown for the best-performing model in generalization testing in held-out, unseen data using RiskPath with a) number of features selected at the elbow of importances ranked after model training and b) the top 10 features ranked after feature selection with the LASSO. In both cases 10-fold cross-validation is performed during training with learning rate=1.00 x 10^{-5}, decay=0.1 and AdamW optimizer. Performance metrics presented correspond to the same individual models as **Table 1** after feature ablation.

Risk pathway mapping

RiskPath leverages timeseries deep learning to allow full mapping of multistep risk pathways that predict disease cases with model complexity controlled by the practitioner. Here, we demonstrate how risk predictors for the total burden of MI aggregate over time in the 10-predictor model reported in **Table 2b.** The relative importance of each predictor is computed within each time period and mapped across all time periods to provide an integrated picture of cumulative predictor pathways over successive time periods (**Figure 2a**). Alternatively, the more conventional metric of mean predictor importance over the entire model may be computed (**Figure 2b**). The additional granularity offered by a multistep map shows how individual predictor importance may vary substantially from one period or lifestage to the next. For example, a deficit in executive function (the higher-level cognitive skills used to control other cognitive functions) is only an important predictor in late childhood (age 9-10) and thereafter recedes from the model. In contrast, sleep problems are a less important predictor in late childhood but accelerate over time to become one of the most important predictors in subsequent lifestages. The cumulative risk pathways delineated by RiskPath can therefore serve as prototype clinical intervention maps for clinicians to prioritize treatments and target specific lifestages for risk mitigation.

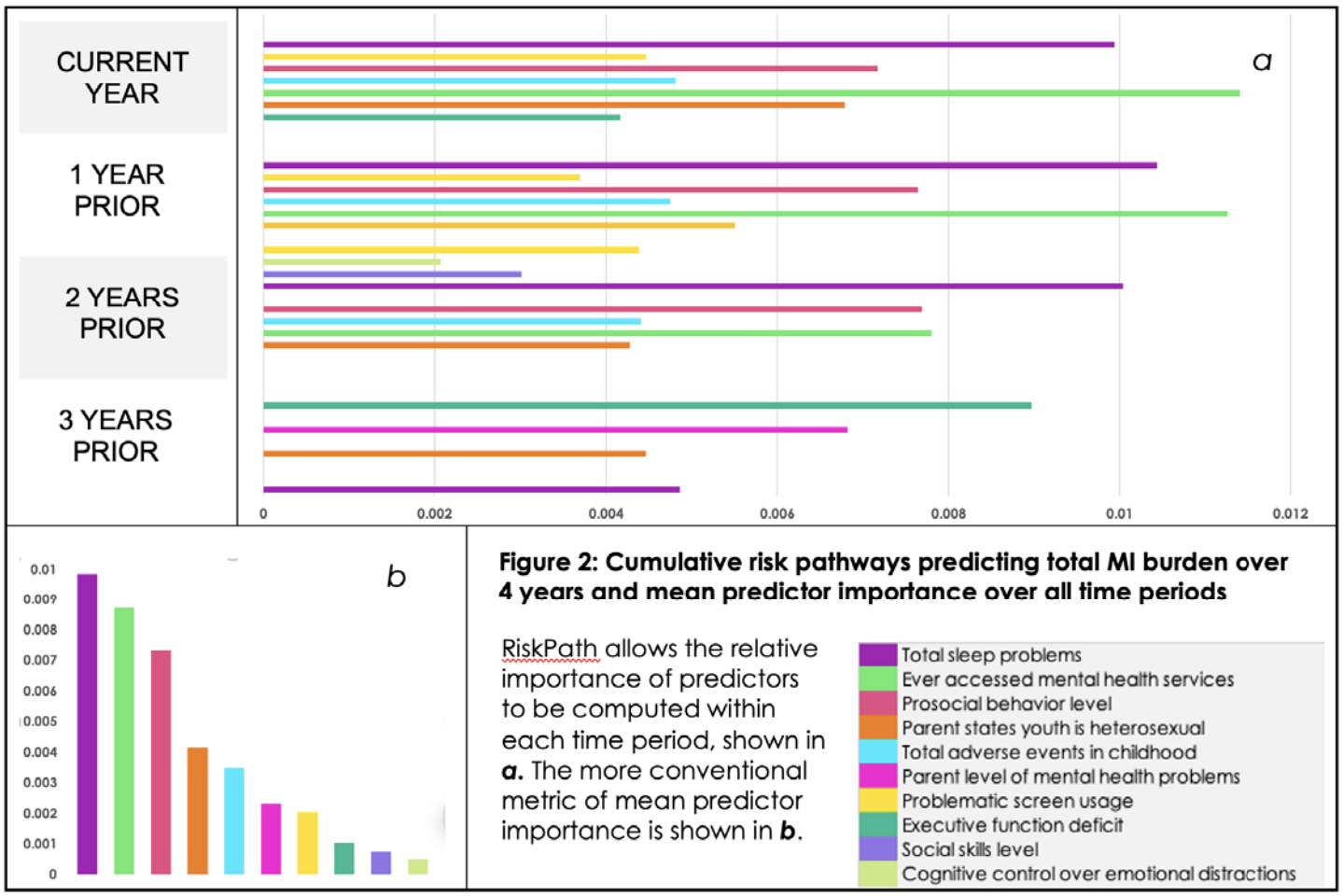

Figure 2: Cumulative risk pathways predicting total MI burden over 4 years and mean predictor importance over all time periods

Performance-complexity tradeoffs

Sample collection is at a premium in biomedicine: most practitioners will use all available observations since thousands of observations (participants) is typically considered a large dataset. Further, for translational tractability it is often preferred to limit the size of the feature set. Thus, the likeliest deep learning environment is one where $N >> n$. In the present study, the smallest number of parameters, N, in our baseline LSTM-based models (**Figure 1a**) ranged in [4182, 46,638,322] but in all cases was larger than n, the sample size (**Tables 1, 2**). LSTMs are one of the more topologically complex and highly parameterized artificial neural network types, an algorithmic class where performance is quite sensitive to parameterization. Practitioners control h, the number of hidden units (network width) but there is no principled method to optimize this setting and its value implies tradeoffs with compute demand, since increasing h typically increases runtime. Accordingly, RiskPath allows the practitioner to directly specify the number of hidden units or explore performance across a range of h. We performed successive experiments varying LSTM topology with hidden units in the range [5,1200].

In **Figure 3**, we depict these performance-complexity tradeoffs for our baseline models reported in **Table 1**. In all cases, we observed a suboptimal learning zone with lower parameterization and hidden unit size as models initially entered the overparameterized regime. Typically, this corresponds to the end of the conventional bias-variance tradeoff zone. Here, test metrics were higher than training metrics suggesting less robust fit, with overall higher loss and lower accuracy. However, as network width increased further into the overparameterized regime, the observed train/test curves smoothed out. A clear inflection point was reached at ~100-500 hidden units into a learning regime where test metrics dipped slightly below training metrics on a sustained basis as hidden unit size was further increased. Consistent performance after this point suggests that it characterizes an observable transition to a learning regime with desirable properties of robust fit and high performance. While the very best models in some cases were obtained with unit sizes >1000, very strong performance was obtainable anywhere in the learning regime after this transition point.

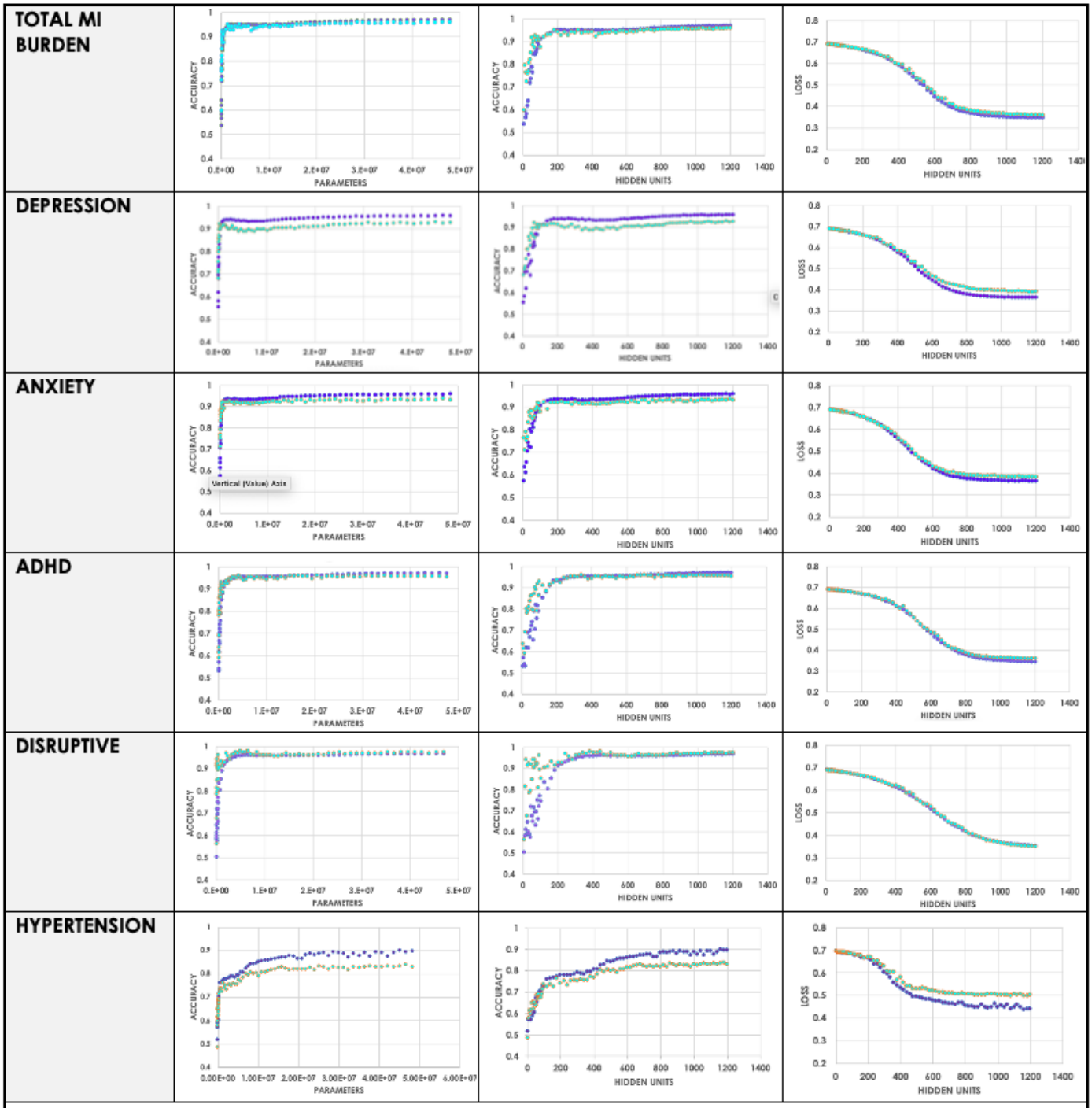

Figure 3: Performance-Complexity Tradeoffs in RiskPath Prediction with LSTMs
The relationship between increasing the number of LSTM parameters and accuracy is shown (left) as well as that between increasing hidden unit size and accuracy (middle) and (loss) in classification of 5 mental illnesses in mid-adolescence using features from 4 time periods. Metrics obtained from model training are shown in purple and those computed in testing in held-out, unseen data are shown in blue. Accuracy is computed as the average accuracy over 10 folds in *k*-fold validation.

Discussion

Our understanding of the cumulative risk pathways of evolving conditions like chronic diseases has been impeded by a paucity of effective methods to perform multistep prediction in the biomedical context. In this paper, we present a new method that captures a fine-grained picture of risk aggregation over time via multistep case prediction from longitudinal data in a holistic model. Unlike existing predictive and risk stratification techniques in biomedicine, typically based on single-period prediction, this allows the delineation of risk

trajectories and the quantification of relative predictor importance both across and within time periods. Accordingly, findings obtained from RiskPath can serve as a useful substrate for improved disease mitigation, enabling practitioners to identify, evaluate and set priorities for early intervention and prevention strategies. For example, targeting an intervention to a specific time period or lifestage or performing simulations on the risk mitigation profile of an intervention. RiskPath delivers performance that comfortably exceeds typical domain performance in one-step learning and notably achieves similarly strong and consistent performance in metrics that are usually more challenging in biomedical prediction such as PPV, sensitivity and specificity. For context, typical classification performance in one-step prediction in the mental health domain is 0.7-0.8 accuracy and AUC and in cardiovascular medicine is 0.65-0.81 AUC with very low (~≤0.15) precision and specificity. LSTMs and timeseries deep learning have very occasionally been used in biomedical risk prediction with tabular data but to our knowledge only as black box predictive models, where they exhibited more uneven performance and lacked the embedding and exposition of risk trajectories offered by RiskPath. [51,52]

RiskPath exploits the properties of LSTMs for temporal risk stratification but also delivers very strong performance. However, the latter is not attributable to an RNN-based structure *per se* but rather to its inbuilt topologic optimization. As we show, building similar topologic optimization into standard feedforwards using the same data as flattened, 2D inputs yields similarly strong performance. Recently, a rich literature has emerged on the topic of performance-complexity tradeoffs in machine learning, particularly with respect to deep learning. This shows that the empirically-observed strong performance of these methods is likely due to their participation in overparameterized learning regimes. Here, 'double-descent' is initially observed in the conventional bias-variance tradeoff zone (U-shaped test loss), but beyond the interpolation point where N = *n*, training loss continues to fall and test loss reverses itself to also descend to a performance level better than that of the nadir observed in the conventional learning regime. [50,53] Effectively, better performance is achieved via overparameterization, i.e. where N>*n.* However, pragmatic feature ablation has only been thinly explored in this literature. Mindful of the need for parsimony in clinical contexts, we add to this literature by demonstrating that using the tabular data common in biomedicine, the vast majority of features can be safely ablated in such overparameterized learning environments without unduly affecting performance. Moreover, this large-scale ablation may be performed without the need to train a model first if principled feature selection based on feature importance is conducted, as we offer in RiskPath. In the present study across multiple chronic mental health conditions, we show that >95% of input features may be ablated with only a ~2% absolute decrement in most performance statistics. The overall very strong performance of this topologically-optimized deep learning approach provides a buffer for feature ablation and maintains consistently strong performance statistics to promote clinical utility while achieving the parsimony required by pragmatic data collection concerns in the clinic.

Since LSTMs are inherently complex algorithms, they consume more runtime resources. In the dataset used here, training time was ~25 minutes for RiskPath versus ~1 minute for feedforwards, an expected result. If a practitioner is interested in training a full RiskPath model in new data, this longer runtime makes the algorithm more suitable for intervention and prevention planning or for development into a temporally-sensitive risk stratification tool. Should very fast turnaround be required (for example, within a 20-30 minute clinic visit), we encourage pre-training a model that is implemented in later deployment. RiskPath is implemented in Python and PyTorch with a modular architecture to allow practitioners to select among its features. Future work may apply this method in other datasets to validate our findings and discern multistep risk trajectories in other diseases.

Experimental Procedures

Resource Availability

Lead contact

Further information and requests will be fulfilled by the lead contact, Nina de Lacy (nina.delacy@utah.edu)

Data and code availability

ABCD data used in this study may be obtained by applying to the ABCD Repository at the National Institute of Mental Health Data Archive (https://nda.nih.gov/abcd): DOI: 10.15154/z563-zd24

CHS data used in this study may be obtained by applying to the BioLINCC repository at the National Institutes of Health (https://biolincc.nhlbi.nih.gov/home/): Accession Number: HLB00040019a

RiskPath is an open source code repository in Python and PyTorch that is available at our GitHub repository (https://github.com/delacylab/RiskPath).

Data and Data Pre-Processing

This study has been deemed not human subjects research by the University of Utah Institutional Review Board.

ABCD Data

The ABCD data used in this study comes from the ABCD open science repository. ABCD is an epidemiologically-informed study launched in 2017 which is collecting data over 10 years from a 21-site cohort of adolescents across the US. Participants (52% male; 48% female) were enrolled at age 9-10 and are currently 13-14 years old. This is a naturalistic, unstratified cohort. Further descriptions of the overall design as well as recruitment procedures and the participant sample may be found in Jernigan et al; Garavan et al; and Volkow et al[54-56] and the study website at abcdstudy.org. ABCD collects rich multimodal data youth participants and their parents. Here, we utilize variables from assessments of physical and mental health, substance use, neurocognition, school performance and quality, culture, community and environment contributed by youth and their parents as well as biospecimens (e.g. pubertal hormone levels) and environmental toxin exposure. [57,58] A fuller description of phenotypic assessments and variables that we analyzed may be inspected in **Supplementary Table 1**.

ABCD is a longitudinal study where the cohort contains ~800 twin pairs and non-twin siblings may be enrolled. General inclusion criteria for the present study were a) participants enrolled in the study at baseline (9-10 yrs) who were still enrolled in the ABCD study through 13-14 yrs (*n*=10,093) who were b) youth participants unrelated to any other youth participant in the study (*n*=8,363). If a youth had sibling(s) present in the cohort, we selected the oldest sibling for inclusion. Resultant data were randomly partitioned into a baseline training dataset (*n*=5,854) and test dataset (n=2,509). The pre-processing pipeline described below was then performed separately for these training and test data partitions.

The baseline feature set comprised the majority of phenotypic variables available from the ABCD study, including data collection site, a proxy for geographic location. Variables related to mental health symptoms were not included as features to avoid bias and redundancy. For continuous features, subscale or total scores were used. Nominal or ordinal variables were one-hot encoded to transform them into discrete variables. Features with >35% missing values were discarded, where prior research shows that good results may be obtained with ML methods with imputation up to 50% missing data**.** [59] Continuous variables were trimmed to [mean +/- 3] standard deviations to remove outliers and scaled in the interval [0,1] with MinMaxScaler. Missing values were imputed within each time period of data using non-negative matrix factorization, an imputation method that is particularly suitable for large-scale multimodal data since it performs well regardless of the underlying pattern of missingness. [60-62] Subsequently, phenotypic variables lacking summary scores were reduced to a summary metric using feature agglomeration. Given the longitudinal nature of the data, it is possible that a participant is still enrolled in the study at 13-14 years of age but did not participate in assessment at certain *ex ante* timepoints. Such participants were identified and excluded after preprocessing, resulting in an average 4% and 5% of eliminations from training and test sets, respectively, across individual predictive targets. In total, feature sets included 767, 322, 572 and 403 features at timepoints 1 (enrollment); 2; 3 and 4 respectively.

Predictive targets were formed from participant scores in the ASEBA Child Behavior Checklist (CBCL), a standardized assessment of mental health in widespread clinical and research use. [63] Parents rate their child on a 0-1-2 scale on 118 specific problem items which are then used to form continuous subscale scores in

clinical dimensions of interest such as Anxiety or Depression. To form binary classification targets, we thresholded and discretized CBCL subscale T scores using cutpoints established by ASEBA for clinical practice by deeming every individual with a T score ≥ 65 as a 'case' or [1] and every individual with a score <65 as a 'not case' or [0] for Anxiety, Attention, Depression and a score ≥60 for Total Problems. Disruptive Behavior cases are established by meeting these case criteria for either Aggression or Rulebreaking, corresponding to the clinical definition of Disruptive Behavior Disorder. These cases were matched for age and gender with participants with the lowest possible scores in the subscale and for Total MI Problems to form a balanced case-control sample for each disorder. Age was separated into two categories for matching purposes (9-10 and 11-12) due to unevenly distributed ages in participants.

CHS Data

The CHS data that we used in this study come from NIH's BioLINCC open science repository. The CHS study (https://chs-nhlbi.org/) was launched in 1987 to identify risk factors for cardiovascular disease related to coronary heart disease in adults aged 65 or older and the investigation of pulmonary disorders, diabetes, kidney disease, vascular dementia, and frailty. CHS collected data annually from 1989-1999 via extensive annual clinical examinations. Measurements included traditional risk factors such as blood pressure and lipids as well as measures of subclinical disease, including echocardiography of the heart, carotid ultrasound, and cranial magnetic-resonance imaging (MRI). This is an observational cohort. Further descriptions of the overall design as well as recruitment procedures and the participant sample may be found in Fried et al. [64] Participants were contacted by phone to ascertain their health status every 6 months. Clinical outcomes were recorded in a binary fashion and included coronary heart disease, angina, heart failure, stroke, transient ischemic attack, claudication, diabetes and hypertension. Here, we utilize variables collected from patients spanning laboratory assays (e.g. thyroid hormone, glucost and lipid levels), health, medication and mental health history, behavioral, diet, exercise and health habits and clinical testing (e.g. blood pressure, heart auscultation, brain imaging). The number (>4,000) of variables collected by CHS preclude enumeration variables used in this study may be viewed via the CHS data dictionary at https://biolincc.nhlbi.nih.gov/media/studies/chs/data_dictionary/CHS_v2019a.pdf. In total, resultant baseline feature sets included 1523; 581; 589; 734; 841; 639; 589; 710; 598; and 891 features at timepoints 1 through 10 respectively.

Inclusion criteria for the present study are participants who were enrolled and remained alive in the CHS study for 10 years. This comprised a total of 3,215 participants. Resultant data were randomly partitioned into a baseline training dataset (*n*=2,250) and test dataset (n=965). Of these, 1000 participants in the training set and 776 in the test set were matched for age and natal sex. The same pre-processing pipeline described for ABCD above was then performed separately for these training and test data partitions. Participants who were still enrolled in the study in the tenth year but who did not participate in assessment at certain *ex ante* timepoints were identified and excluded after preprocessing, resulting in an average 4% and 5.6% of eliminations from training and test sets, respectively.

The predictive target in CHS is supplied in the original data as a clinical variable where original data had 1=Normal; 2=Borderline; 3=Hypertension. We omitted Borderline subjects and then relabeled subjects that had Hypertension as 1 and Normal subjects as 0.

RiskPath: Tunable linear and nonlinear feature selection to constrain feature set sizes

A basic filtering process was first performed for each target where features with <5% variance with each target were discarded. Subsequently, feature selection was performed within RiskPath, which currently offers the options of the LASSO CV (linear) or Boruta (nonlinear) techniques and a combination of both methods (union of linear + nonlinear). RiskPath implements the *scikit-learn* LASSO CV and *BorutaPy* packages within Python wrappers to a) adapt these methods for timeseries data and b) give users explicit control of settings that constrain the number of features obtained by each technique. This provides principled linear and nonlinear method for feature selection in timeseries data determined by feature importance to control the size of the feature set and obtain the parsimony often required in biomedical prediction. In the present study, linear, nonlinear and linear+nonlinear feature sets were generated for each experiment and performance compared across these feature sets.

The LASSO CV is a gold standard linear feature selection technique for ML that identifies a set of linear coefficients minimizing the Mean Squared Error (MSE) of the prediction under L1-regularization. The key

hyperparameter in the LASSO algorithm is the alpha, which controls the amount of L1 penalization. LASSO CV has built-in cross-validation capability that supports choosing a setting for the alpha using a grid search along the regularization path computed from the input data. In RiskPath, the LASSO CV subroutine is adapted for timeseries data to separately perform feature selection at each timepoint and the union of features across timepoints is then obtained to determine an overall linear feature set for each target and experiment. RiskPath users are offered the additional option of thresholding coefficients output by the LASSO CV to further constrain the size of feature sets as may be required in translational or clinical use cases. In this study we provide examples of these options. Our baseline analyses left results unthresholded where linear models reported accepted all features with a non-zero coefficient and results from these larger feature sets may be seen in **Supplementary Table 2**. However, a user is also offered the option of imposing a non-zero threshold on coefficients to futher reduce the number of features selected.

The Boruta[65] is an ensemble-based technique that is a gold standard nonlinear method for predictive feature selection. Constructed around a random forest algorithm, it compares real features with their 'shadow copies' (real features with values shuffled) such that a feature is important only when it has an importance score higher than the maximal importance score in the set of shadow features i.e. if it has a significantly higher importance over N random forest runs than the expected value 0.5N (as defined by a binomial distribution with $p = 0.5$). Significance is measured by multiple hypothesis testing. The RiskPath implementation of the Boruta offers several adaptations tailored for analysis of timeseries data. Firstly, the RiskPath Boruta subroutine is separately performed at each timepoint to generate timepoint-specific feature sets and then the union of features across timepoints is obtained to determine an overall nonlinear feature set for the experiment. Secondly, RiskPath offers users the ability to control two tunable hyperparameters that determine whether a feature is deemed important in practice and thereby control the size of the feature set obtained. The first tunable hyperparamter, level of significance (alpha_b), can be adjusted to expand or shrink the rejection region and concomitantly the size of the feature set. The second tunable hyperparameter, the percentile (perc), may be controlled such that each real feature is compared to a specific percentile of the importance scores of the shadow feature set (instead of the maximal case where perc=100). Intuitively, setting a higher alpha_b or a lower perc will return a larger number of important features at a higher risk of type-1 error. In the present study, we implemented Boruta nonlinear selection with default hyperparameter settings and a maximum tree depth of 7. The results of this selection may be seen in **Supplementary Table 3**.

Baseline feature sets produced by combining the results of linear+nonlinear feature selection with the LASSO CV and Boruta may be seen in **Supplementary Table 4**.

RiskPath: Multistep deep learning with topologic optimization

LSTMs in RiskPath are trained with two bidirectional layers and the tanh activation function followed by a softmax function. Input hidden weights are initialized with the Xavier uniform distribution and hidden-hidden weights with orthogonal matrices. In the present study the AdamW optimizer is used with weight decay=0.1, learning rate=$1*10^{-5}$ and models were trained for a maximum of 150 epochs with early stopping (patience=5 and metric=validation loss). RiskPath allows these settings to be modified by the practitioner for local heuristics or experimental preferences.

RiskPath offers practitioners an automated option to optimize over the number of hidden units (network width) where this refers to the units per individual layer (forward layer and backward layer) in each bidirectional layer. Users may specify a minimum and maximum number of hidden units where RiskPath's default behavior is to subsequently present a series of corresponding model fits that increment hidden units by 5 in the range [5,100] and increments of 20 in the range [100, *h*] hidden units. This allows the practitioner to explore performance-complexity tradeoffs or to constrain complexity for runtime speed or infrastructural limitations. In the present study, all experiments were performed with *h* in the range [5, 1200] and results compared. In each experiment, the best-performing model in terms of accuracy was selected for reporting.

The LSTM algorithm in RiskPath is modified in a number of ways to render it suitable for analyzing longitudinal biomedical data and delineating cumulative risk pathways. Firstly, SHAP is embedded within the algorithm to render the LSTMs translationally explainable using the GradientShap function for 3D data. In RiskPath, SHAP-

based feature importances are computed in the validation sets of the *k*-fold process with the training set used as background for every permutation of participant, time period and feature. Three types of importances are thereby computed and returned: mean importance over the participant sample; per time period importance score for each feature and an overall feature importance for each feature averaged over time periods. This enables users to compute and examine not only the mean SHAP value over the participant sample that is conventionally reported in the machine learning literature, but also to expose relative feature importance within and across timepoints. Secondly, LSTMs expect square data in a 3D matrix. However, it is common in biomedical protocols or datasets that a particular feature was only collected in a subset of timepoints, or that it was not selected during the feature selection process at a particular timepoint. The latter can reflect the scientifically important condition where a predictor is selectively important at different points in a disease course or lifestage. RiskPath users may select all timepoints in longitudinal data or specific timepoints and in the present analysis, we used all available timepoints in the ABCD dataset. To ensure that square data enters the timeseries prediction, RiskPath automatically completes the 3D matrix for the user, filling values that are not available with a 1. Thirdly, the user may want to compare 2D vs 3D data within the algorithm as we have done in the present study and RiskPath gives the option to use either timeseries or flattened data. Lastly, the practitioner may select the number of GPUs to run the program: RiskPath parallelizes over unit sizes. RiskPath is developed in PyTorch and Python.

Deep learning with Feedforwards

We trained standard feedforwards with 3 layers and the Relu activation function. The last output layer contained a softmax function. AdamW was used as the optimizer with weight decay 0.1, learning rate $1 * 10^{-5}$, training for a maximum of 150 epochs with early stopping (patience=5 and metric=validation loss). All models were trained with *k* fold cross-validation where $k = 10$. To recapitulate the topologic optimization offered in RiskPath and ensure head-to-head comparisons, feedforward models in each experiment were re-trained and tested with the number of hidden units (width) varying in the interval [5,1200]. Feedforwards were encoded with PyTorch.

Acknowledgements

This work was supported by the National Institute of Mental Health under award R00MH118359 to NdL.

ABCD data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development[SM] (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from 10.15154/z563-zd24). DOIs can be found at https://nda.nih.gov/abcd/abcd-annual-releases.

Author Contributions

Conceptualization, N.dL.; Methodology, N.dL., M.R., software, MR., WL.; formal analysis, NdL., MR.; validation, N.dL., M.R., investigation, N.dL.; writing – original draft, N.dL., W.L., M.R.; writing – review & editing, N.dL., W.L., M.R., funding acquisition, N.dL.; resources, N.dL.; supervision, N.dL.

Declaration of Interests

The authors declare no competing interests.

References

1 Schooling, C. M. & Jones, H. E. Clarifying questions about "risk factors": predictors versus explanation. *Emerg Themes Epidemiol* **15**, 10 (2018). https://doi.org:10.1186/s12982-018-0080-z

2 D'Agostino, R. B., Sr. *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743-753 (2008). https://doi.org:10.1161/CIRCULATIONAHA.107.699579

3 Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118 (2017). https://doi.org:10.1038/nature21056

4 Lakhani, P. & Sundaram, B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* **284**, 574-582 (2017). https://doi.org:10.1148/radiol.2017162326

5 Yu, K. H. *et al.* Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun* **7**, 12474 (2016). https://doi.org:10.1038/ncomms12474

6 Ting, D. S. W. *et al.* Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* **318**, 2211-2223 (2017). https://doi.org:10.1001/jama.2017.18152

7 Cheung, C. Y., Tang, F., Ting, D. S. W., Tan, G. S. W. & Wong, T. Y. Artificial Intelligence in Diabetic Eye Disease Screening. *Asia Pac J Ophthalmol (Phila)* **8**, 158-164 (2019). https://doi.org:10.22608/APO.201976

8 Kanagasingam, Y. *et al.* Evaluation of Artificial Intelligence-Based Grading of Diabetic Retinopathy in Primary Care. *JAMA Netw Open* **1**, e182665 (2018). https://doi.org:10.1001/jamanetworkopen.2018.2665

9 Abramoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* **1**, 39 (2018). https://doi.org:10.1038/s41746-018-0040-6

10 Ladbury, C. *et al.* Integration of artificial intelligence in lung cancer: Rise of the machine. *Cell Rep Med* **4**, 100933 (2023). https://doi.org:10.1016/j.xcrm.2023.100933

11 Gandhi, Z. *et al.* Artificial Intelligence and Lung Cancer: Impact on Improving Patient Outcomes. *Cancers (Basel)* **15** (2023). https://doi.org:10.3390/cancers15215236

12 Dieleman, J. L. *et al.* US Health Care Spending by Payer and Health Condition, 1996-2016. *JAMA* **323**, 863-884 (2020). https://doi.org:10.1001/jama.2020.0734

13 Kist, J. M. *et al.* SCORE2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the Netherlands: an external validation study. *EClinicalMedicine* **57**, 101862 (2023). https://doi.org:10.1016/j.eclinm.2023.101862

14 Muntner, P. *et al.* Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA* **311**, 1406-1415 (2014). https://doi.org:10.1001/jama.2014.2630

15 Criner, R. N. *et al.* Mortality and Exacerbations by Global Initiative for Chronic Obstructive Lung Disease Groups ABCD: 2011 Versus 2017 in the COPDGene(R) Cohort. *Chronic Obstr Pulm Dis* **6**, 64-73 (2019). https://doi.org:10.15326/jcopdf.6.1.2018.0130

16 Celli, B. R. *et al.* The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *N Engl J Med* **350**, 1005-1012 (2004). https://doi.org:10.1056/NEJMoa021322

17 Athlin, A. *et al.* Prediction of Mortality Using Different COPD Risk Assessments - A 12-Year Follow-Up. *Int J Chron Obstruct Pulmon Dis* **16**, 665-675 (2021). https://doi.org:10.2147/COPD.S282694

18 Heikes, K. E., Eddy, D. M., Arondekar, B. & Schlessinger, L. Diabetes Risk Calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes. *Diabetes Care* **31**, 1040-1045 (2008). https://doi.org:10.2337/dc07-1150

19 Bang, H. *et al.* Development and validation of a patient self-assessment score for diabetes risk. *Ann Intern Med* **151**, 775-783 (2009). https://doi.org:10.7326/0003-4819-151-11-200912010-00005

20 Herman, W. H., Smith, P. J., Thompson, T. J., Engelgau, M. M. & Aubert, R. E. A new and simple questionnaire to identify people at increased risk for undiagnosed diabetes. *Diabetes Care* **18**, 382-387 (1995). https://doi.org:10.2337/diacare.18.3.382

21 Freedman, A. N. *et al.* Colorectal cancer risk prediction tool for white men and women without known susceptibility. *J Clin Oncol* **27**, 686-693 (2009). https://doi.org:10.1200/JCO.2008.17.4797

22 Park, Y. *et al.* Validation of a colorectal cancer risk prediction model among white patients age 50 years and older. *J Clin Oncol* **27**, 694-698 (2009). https://doi.org:10.1200/JCO.2008.17.4813

23 Gail, M. H. *et al.* Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* **81**, 1879-1886 (1989). https://doi.org:10.1093/jnci/81.24.1879

24 Gail, M. H. *et al.* Projecting individualized absolute invasive breast cancer risk in African American women. *J Natl Cancer Inst* **99**, 1782-1792 (2007). https://doi.org:10.1093/jnci/djm223

25 Rockhill, B., Spiegelman, D., Byrne, C., Hunter, D. J. & Colditz, G. A. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *J Natl Cancer Inst* **93**, 358-366 (2001). https://doi.org:10.1093/jnci/93.5.358

26 Costantino, J. P. *et al.* Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst* **91**, 1541-1548 (1999). https://doi.org:10.1093/jnci/91.18.1541

27 Matsuno, R. K. *et al.* Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. *J Natl Cancer Inst* **103**, 951-961 (2011). https://doi.org:10.1093/jnci/djr154

28 Banegas, M. P. *et al.* Projecting Individualized Absolute Invasive Breast Cancer Risk in US Hispanic Women. *J Natl Cancer Inst* **109** (2017). https://doi.org:10.1093/jnci/djw215

29 Bzdok, D., Engemann, D. & Thirion, B. Inference and Prediction Diverge in Biomedicine. *Patterns (N Y)* **1**, 100119 (2020). https://doi.org:10.1016/j.patter.2020.100119

30 Breiman, L. Statistical Modeling: The Two Cultures. *Statistical Science* **16**, 199-231 (2001).

31 Thompson B. & G.M., B. The importance of structure coefficients in regression research. *Educ. Psychol. MeasEduc. Psychol. Meas* **45**, 203-209 (1985).

32 Taylor, J. & Tibshirani, R. J. Statistical learning and selective inference. *Proc Natl Acad Sci U S A* **112**, 7629-7634 (2015). https://doi.org:10.1073/pnas.1507583112

33 Lundberg, S. & Lee, S. A Unified Approach to Interpreting Model Predictions. *arXiv* **1705.07874 [cs.AI]** (2017).

34 Kessler, R. C. *et al.* Age of onset of mental disorders: a review of recent literature. *Curr Opin Psychiatry* **20**, 359-364 (2007). https://doi.org:10.1097/YCO.0b013e32816ebc8c

35 Kessler, R. C. *et al.* Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry* **6**, 168-176 (2007).

36 Kessler, R. C. *et al.* Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry* **62**, 593-602 (2005). https://doi.org:10.1001/archpsyc.62.6.593

37 Roehrig, C. Mental Disorders Top The List Of The Most Costly Conditions In The United States: $201 Billion. *Health Aff (Millwood)* **35**, 1130-1135 (2016). https://doi.org:10.1377/hlthaff.2015.1659

38 Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R. & Walters, E. E. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry* **62**, 617-627 (2005). https://doi.org:10.1001/archpsyc.62.6.617

39 National Institute of Mental Health. *Strategic Plan (online: accessed 4/24/2024)*,

40 Garcia-Argibay, M. *et al.* Predicting childhood and adolescent attention-deficit/hyperactivity disorder onset: a nationwide deep learning approach. *Mol Psychiatry* (2022). https://doi.org:10.1038/s41380-022-01918-8

41 Maniruzzaman, M., Shin, J. & Al Mehedi Hasan, M. Predicting Children with ADHD Using Behavioral Activity: A Machine Learning Analysis *Appl. Sci.* **12**, 2737 (2022). https://doi.org: https://doi.org/10.3390/app12052737

42 Ter-Minassian, L. *et al.* Assessing machine learning for fair prediction of ADHD in school pupils using a retrospective cohort study of linked education and healthcare data. *BMJ Open* **12**, e058058 (2022). https://doi.org:10.1136/bmjopen-2021-058058

43 Menon, S. S. & Krishnamurthy, K. Multimodal Ensemble Deep Learning to Predict Disruptive Behavior Disorders in Children. *Front Neuroinform* **15**, 742807 (2021). https://doi.org:10.3389/fninf.2021.742807

44 Toenders, Y. J. *et al.* Predicting Depression Onset in Young People Based on Clinical, Cognitive, Environmental, and Neurobiological Data. *Biol Psychiatry Cogn Neurosci Neuroimaging* **7**, 376-384 (2022). https://doi.org:10.1016/j.bpsc.2021.03.005

45 Xiang, Q. *et al.* Prediction of the trajectories of depressive symptoms among children in the adolescent brain cognitive development (ABCD) study using machine learning approach. *J Affect Disord* **310**, 162-171 (2022). https://doi.org:10.1016/j.jad.2022.05.020

46 Foland-Ross, L. C. *et al.* Cortical thickness predicts the first onset of major depression in adolescence. *Int J Dev Neurosci* **46**, 125-131 (2015). https://doi.org:10.1016/j.ijdevneu.2015.07.007

47 Rocha, T. B. *et al.* Identifying Adolescents at Risk for Depression: A Prediction Score Performance in Cohorts Based in 3 Different Continents. *J Am Acad Child Adolesc Psychiatry* **60**, 262-273 (2021). https://doi.org:10.1016/j.jaac.2019.12.004

48 Cohen, J. R., Thakur, H., Young, J. F. & Hankin, B. L. The development and validation of an algorithm to predict future depression onset in unselected youth. *Psychol Med* **50**, 2548-2556 (2020). https://doi.org:10.1017/S0033291719002691

49 SCORE2 working group and ESC Cardiovascular Risk Collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J* **42**, 2439-2454 (2021). https://doi.org:10.1093/eurheartj/ehab309

50 Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc Natl Acad Sci U S A* **116**, 15849-15854 (2019). https://doi.org:10.1073/pnas.1903070116

51 Guo, A., Beheshti, R., Khan, Y. M., Langabeer, J. R., 2nd & Foraker, R. E. Predicting cardiovascular health trajectories in time-series electronic health records with LSTM models. *BMC Med Inform Decis Mak* **21**, 5 (2021). https://doi.org:10.1186/s12911-020-01345-1

52 Men, L., Ilk, N., Tang, X. & Liu, Y. Multi-disease prediction using LSTM recurrent neural networks. *Expert Systems with Applications* **177**, 114905 (2021).

53 D'Ascoli, S., ., Refinetti, M., Biroli, G. & Krzakala, F. Double Trouble in Double Descent: Bias and Variance(s) in the Lazy Regime. *Proceedings of the 37th International Conference on Machine Learning, PMLR* **19**, 2280-2290 (2020).

54 Jernigan, T. L., Brown, S. A. & Dowling, G. J. The Adolescent Brain Cognitive Development Study. *J Res Adolesc* **28**, 154-156 (2018). https://doi.org:10.1111/jora.12374

55 Garavan, H. *et al.* Recruiting the ABCD sample: Design considerations and procedures. *Dev Cogn Neurosci* **32**, 16-22 (2018). https://doi.org:10.1016/j.dcn.2018.04.004

56 Volkow, N. D. *et al.* The conception of the ABCD study: From substance use to a broad NIH collaboration. *Dev Cogn Neurosci* **32**, 4-7 (2018). https://doi.org:10.1016/j.dcn.2017.10.002

57 Barch, D. M. *et al.* Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: Rationale and description. *Dev Cogn Neurosci* **32**, 55-66 (2018). https://doi.org:10.1016/j.dcn.2017.10.010

58 Lisdahl, K. M. *et al.* Adolescent brain cognitive development (ABCD) study: Overview of substance use assessment methods. *Dev Cogn Neurosci* **32**, 80-96 (2018). https://doi.org:10.1016/j.dcn.2018.02.007

59 Jager, S., Allhorn, A. & Biessmann, F. A Benchmark for Data Imputation Methods. *Front Big Data* **4**, 693674 (2021). https://doi.org:10.3389/fdata.2021.693674

60 Dhillon, I. S. & Sra, S. Generalized Nonnegative Matrix Approximations with Bregman Divergences. *Advances in Neural Information Processing Systems* **18** (2006).

61 Tandon R. & Sra, S. Sparse nonnegative matrix approximation: new formulations and algorithms. *Max Planck Institute for Biological Cybernetics Technical Report* **193** (2010).

62 Xu, J. *et al.* NMF-Based Approach for Missing Values Imputation of Mass Spectrometry Metabolomics Data. *Molecules* **26** (2021). https://doi.org:10.3390/molecules26195787

63 McConaughy, S. H. in *Educational Psychology, Handbook of Psychoeducational Assessment* (eds J.J.W. Andrews, D.H. Saklofske, & H.L. Janzen) Ch. 10, 289-324 (Academic Press, 2001).

64 Fried, L. P. *et al.* The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* **1**, 263-276 (1991). https://doi.org:10.1016/1047-2797(91)90005-w

65 Kursa, M. B., Jankowski, A. & Witold, R. R. Boruta – A System for Feature Selection. Fundamenta Informaticae 101, 271-285 (2010). *Fundamenta Informaticae* **101**, 271-285 (2010).