Towards realizing the vision of precision medicine: AI based prediction of clinical drug response

Johann de Jong,[1] Ioana Cutcutache,[2] Matthew Page,[2] Sami Elmoufti,[3] Cynthia Dilley,[4] Holger Fröhlich[1,5,6,†,‡] and Martin Armstrong[7,†]

†These authors contributed equally to this work.

Abstract

Accurate and individualized prediction of response to therapies is central to precision medicine. However, due to the generally complex and multifaceted nature of clinical drug response, realizing this vision is highly challenging, requiring integrating different data types from the same individual into one prediction model. We use the anti-epileptic drug brivaracetam as a case study and combine a hybrid data-/knowledge-driven feature extraction with machine learning to systematically integrate clinical and genetic data from a clinical discovery dataset (n=235 patients). As such, we construct a model that successfully predicts clinical drug response (AUC=0.76) and show that even with limited sample size, integrating high-dimensional genetics data with clinical data can inform drug response prediction. After a further validation on data collected from an independently conducted clinical study (AUC=0.75), we extensively explore our model to gain insights into the determinants of drug response, and identify various clinical and genetic characteristics predisposing to poor response. Finally, we assess the potential impact of our model on clinical trial design and demonstrate that, by enriching for probable responders, significant reductions in clinical study sizes may be achieved. To our knowledge, our model represents the first retrospectively validated machine learning model linking drug mechanism of action and the genetic, clinical and demographic background in epilepsy patients to clinical drug response. Hence, it provides a blueprint for how machine learning-based multimodal data integration can act as a driver in achieving the goals of precision medicine in fields such as neurology.

Author affiliations:

1 Data & Translational Sciences, UCB Biosciences GmbH, 40789, Monheim am Rhein, Germany

2 Data & Translational Sciences, UCB Pharma, Slough, SL1 3WE, UK

3 Late Development Statistics, UCB Biosciences Inc., Raleigh, NC 27617, USA

4 Head of Asset Strategy, UCB Inc., Smyrna, GA 30080, USA

5 Fraunhofer Institute for Scientific Computing and Algorithms (SCAI), Business Area Bioinformatics, 53757 Sankt Augustin, Germany

6 Bonn-Aachen International Center for IT, University of Bonn, 53115 Bonn, Germany

7 Data & Translational Sciences, UCB Pharma, 1420 Braine l'Alleud, Belgium

‡The author performed most of this work while previously working at UCB Biosciences GmbH, Monheim.

Correspondence to: Johann de Jong

UCB Biosciences GmbH, Alfred-Nobel-Straße 10, 40789, Monheim am Rhein, Germany

E-mail: johann.dejong@ucb.com

Correspondence may also be addressed to: Holger Fröhlich

University of Bonn, Bonn-Aachen International Center for Information Technology (B-IT), Friedrich Hirzebruch-Allee 5, 53115 Bonn, Germany

E-mail: holger.froehlich@scai.fraunhofer.de

Martin Armstrong

UCB Pharma, Chemin du Foriest 1, 1420 Braine l'Alleud, Belgium
E-mail: martin.armstrong@ucb.com

Running title: AI Based Prediction of Clinical Drug Response

Downloaded from https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awab108/6178276 by guest on 26 March 2021

ScholarOne, 375 Greenbrier Drive, Charlottesville, VA, 22901 Support (434) 964 4100

Introduction

In recent years, precision medicine has emerged as a new paradigm for improved and more individualized patient care. Its key objective is to provide the right treatment, to the right patient at the right time, by basing medical decisions on individual patient characteristics (including biomarkers), rather than on averages of those characteristics across a whole patient population. Precision medicine has the potential to widely impact both patient care and drug development, via [1,2]

- Early disease diagnosis and prevention using individual patient characteristics.

- Better disease management by more efficient and effective prescribing practices at the level of the individual patient, including reducing the risk of side effects and adverse events.

- More efficient design of clinical trials by enriching for likely responders at baseline.

Precision medicine is closely connected to and partially dependent on pharmacogenetics, a field which has developed since the 1980s resulting from advances in genetics and molecular biology. In pharmacogenetics, statistical associations between genetic variants and drug response are studied within a defined target population. According to the US Food and Drug Administration (FDA) more than 400 pharmacogenetic biomarkers and biomarker signatures have appeared on approximately 160 drug labels [3], almost half of which are within the field of oncology. However, in many cases, and specifically outside oncology, these tests have had a limited impact on prescribing practices. One of the reasons for this is that the nature of drug

response is highly complex, resulting from the interaction of a multitude of factors including environmental, anthropometric and genetic factors, as well as biological subsystems affected by the disease [4]. In this respect, it is unlikely that any single biomarker or other single stratifying factor will fully capture this complexity and, as result, the field of precision medicine is still in its infancy [5].

Recent advances in the ability to generate molecular data, as well as parallel advances in the fields of artificial intelligence (AI), specifically machine learning (ML), and high-performance computing, now allow for integrating different types of multivariate data from the same individual into one multimodal prediction model [2]. Such ML models can be used to make predictions at the individual patient level, and their interpretation can drive the development of a more holistic understanding of drug response. ML models can thus be instrumental in translating precision medicine into practice, and in optimizing clinical trial design by predictive enrichment for drug responders and the testing of therapies in selected populations [6,7]. However, before ML models can be moved into clinical routine, a rigorous validation pipeline is required that consists of (1) validation on an independent test set, (2) further validation on an independent study (owing to the fact that all clinical studies are unavoidably biased by patient selection), (3) a prospective validation via a clinical study to show a benefit compared to standard of care, and (4) following a regulatory pathway to get approval as a diagnostic tool [2]. Due to this complexity, currently there are still only few examples where ML models have reached clinical routine, and all of them are from the oncology field [8,9]. On the other hand, neurology is widely seen as a field, in which there is a high potential for precision medicine, as well as a high demand due to the increasing prevalence of neurological disorders [10,11].

Epilepsy is one such neurological disorder that has a pressing need for the development of more individualized patient care paths in the context of precision medicine. The treatment of epilepsy and the development of new antiepileptic drugs (AEDs) are hampered by high rates of patient non-response [12]. Despite almost 30 AEDs being licensed for use in the US [13], over 30% of patients remain inadequately controlled on their medication [14–16]. The reasons for this remain largely unknown, and iterative prescribing practices remain the norm.

Patients with drug-resistant epilepsy suffer more frequently from comorbid illnesses, psychological dysfunction, social stigmatization and reduced quality of life. Moreover, they are at increased risk of mortality and have an overall decreased life expectancy. As such, these patients represent most of the disease burden in the epilepsy patient population [17].

Additionally, in relation to the need to drive sustainable health care systems, patients with refractory/uncontrolled epilepsy contribute disproportionately to the cost of treatment. Studies of epilepsy sub-populations consistently found substantially higher per-person costs for patients with poor seizure control, ranging from ~2 to ~10 times higher compared to patients with unrefractory/controlled epilepsy [18]. With a high patient and socio-economic burden, there is therefore a pressing need to better understand and predict AED response, in order to support the development of new AEDs and to ensure that patients are optimally treated earlier in their disease course.

While previous work has been performed in an attempt to improve our understanding of AED response [19–31], the focus has mostly been on specific pharmacokinetics related drug metabolizing enzymes and transporters, neglecting other genetic factors that could be important in determining response, such as a drug's mechanism of action (MOA) or underlying disease mechanisms. However, as outlined above, focusing on single stratifying factors or biomarkers has severe limitations in its potential for achieving the goals of precision medicine. As such, it has so far been difficult for the – sometimes even conflicting – results to impact clinical practice [32]. We are aware of only one study that jointly analyzed clinical and genetic data to predict AED resistance in patients with mesial temporal lobe epilepsy [33]. Whilst showing promise, the analysis was, again, limited to pharmacokinetic related genes. Moreover, the authors did not validate their model in an independent patient dataset, so the extent to which their results generalize outside their discovery dataset is unclear. Finally, the study addressed the relatively general question of resistance against any AED. However, in order to achieve the objectives of precision medicine and to support optimized prescription in clinical practice, specific drug response prediction models are needed, directly linking drug mechanism of action to patient-specific response.

In further related work no attempts were made to link drug mechanism of action to patient response [34–38], response was not predicted directly [36,38], only very limited numbers of SNPs were used [35], no genetics data was used at all [34–38], and/or the results were not validated on external data [34,37,38].

For this purpose, we set out to use ML approaches to integrate a broader set of genetic and clinical data, to predict response to one specific AED. Specifically, we used the AED brivaracetam as a case study. Brivaracetam (trade name: Briviact) is an AED whose mechanism of action is via blockade of the vesicle membrane protein SV2A and is licensed for the treatment of partial-onset (focal) seizures in people 4 years of age and older. During

phase III clinical studies, approximately 30-40% of subjects reached the primary endpoint of >50% reduction in seizure frequency [39]. By integrating clinical and whole-genome sequencing (WGS) data from 235 participants in these phase III clinical studies, we built and compared several ML models predictive of brivaracetam response. Our results show how patient and clinical data, as well as genetic makeup (comprising of epilepsy-specific and drug mechanism-specific genetic features) each uniquely contribute to predicting drug response. In addition, we demonstrate that our best ML model achieves a good prediction performance in a retrospective validation using an additional independent phase III clinical study. Finally, we quantify the impact that our best model could have on clinical trial design by its ability to enrich for responders in future clinical studies. We thus realize an important step towards precision medicine in neurological disorders.

Materials and methods

Processing the clinical data

Efficacy data were pooled from two phase III clinical trials (N01252 [NCT00490035], N01358 [NCT01261325]) for a therapeutic dose of > 50 mg/d enrolling adults (=16 years) with focal seizures. Both initial clinical trials received approval by appropriate institutional review committees and were conducted in accordance with all regulatory, ethical, and good clinical practice requirements. After an 8-week prospective baseline period, patients were randomized to brivaracetam or placebo without titration and entered a 12-week treatment period. Patients randomized to brivaracetam > 50 mg/d (100 and 200 mg/d) or placebo were included in the analysis. Although trial N01252 also included patients treated with 50 mg/d, these were excluded from the analysis in order to maximize correspondence between the validation dataset (N01252) and the training dataset (N01358), which did not include any patients treated with 50 mg/d. Data was used for subjects who consented to give a DNA sample and were treated with either 100mg or 200mg of brivaracetam per day, the approved brivaracetam doses.

Sample collection, WGS and processing

Total genomic DNA >850ng (>10ng/uL minimum concentration) was extracted from blood of 235 patients (NCT01261325) and 47 patients (NCT00490035) treated with 100mg or

200mg of brivaracetam per day. DNA samples were sequenced to 30x mean coverage on Illumina HiSeq X instruments with 150bp, paired-end sequencing. Reads were aligned to the GRCh38/hg38 reference using BWA-mem 7.15.r1140 [40,41]. Picard [42] was used to mark duplicates and GATK4 [43] for base quality score recalibration, which was applied using a 4-bin base quality scheme. Picard was used to aggregate all data for each sample into individual CRAM files using lossless compression. SNPs and indels were called, per sample, using HaplotypeCaller (GATK 3.5) [43] and individual gVCFs consolidated into a single gVCF. Pre-computed CADD scores [44] were used for SNVs and CADD was run separately to generate scores for indels. The scores were converted to GRCh38 coordinates using CrossMap [45]. The gVCF was normalized and split to generate per-sample VCFs using bcftools [46]. snpEff [47] was used to annotate each sample VCF with CADD scores. MANTA v1.3.1 [48] was used on all samples for CNV/SV calling.

Constructing the gene sets

Gene sets (Supplementary Material Table S1) were identified by enrichment of Gene Ontology Molecular Function terms; MetaBase canonical pathways; or phenotype categories, with genes that have an etiological association to epilepsy or that are putatively linked to the mechanism of action of brivaracetam. Epilepsy genes were combined from different molecular evidence sources: rare genetic epilepsies or rare disorders that display phenotype traits ontologically descendant from "seizure" (excluding symptomatic seizures and focal sensory seizures) and "epileptic encephalopathy" in the Human Phenotype Ontology [49]; nearest mapped genes for significant SNPs from epilepsy GWAS studies in GWAS Catalog [50]; direct genetic associations from gene-disease databases DisGeNET [51] and OpenTargets [52]; experimentally validated target genes for miRNAs dysregulated in epilepsy from miRmap [53] and EpimiRBase [54]; and epigenetic mechanisms linked to epilepsy. In turn, drug mechanism of action genes were drawn from: canonical pathways related to synaptic vesicle biology; direct neighbors of SV2A, SV2B and SV2C in a protein-protein interaction network contextualized with brain gene expression data from GTEx [55]; genes encoding drug-metabolizing enzymes for brivaracetam; genes associated with comorbidities that influence drug response; and the targets of co-administered AEDs (carbamazepine, phenytoin, phenobarbital, levetiracetam). Epilepsy genes and drug mechanism of action genes were linked to gene sets from Gene Ontology Molecular Function terms using (1) TopGO [56], (2)

Metabase [57] neurological canonical signaling pathways using a hypergeometric test and (3) knowledge-based epilepsy phenotype categories [58] based on gene membership.

Variant filtering and gene mapping

After pre-processing as outlined above, the WGS data represented ~20 million variants. From these, only variants that satisfied at least one of the following conditions were retained in subsequent gene set-based analyses:

- A significant association to the disease phenotype in Genome Wide Association Studies (GWAS), as listed in the GWAS Catalog [50],

- Strong Linkage Disequilibrium (LD) to any of the above GWAS SNPs ($r^2 > 0.8$).

- Proximity of at most 10kb to SV2A and a CADD [44] score of more than 15.

- eQTL evidence for SV2A and a CADD [44] score of more than 15.

This resulted in approximately 14000 unique single nucleotide variants. These filtered variants were mapped to genes based on proximity in the genome (in ±10kb region), brain cis-eQTLs from GTEx and significant SNPs from epilepsy related GWAS studies (Supplementary Material Table S2). Hg38 genomic intervals for genes were extracted from the SnpEff database [47] using the "dump" argument. Genomic coordinates for each gene were extended ±10kb and converted to BED format. Significant brain cis-eQTLs were downloaded from https://gtexportal.org/home/datasets (v7) and CrossMap [45] applied to convert to coordinates on the GRCh38 assembly before converting to BED format. GWAS Catalog [50] v1.0.1 was used to select all the SNPs significantly associated with specific epilepsy/seizure traits. For each such SNP the position, risk allele and mapped gene information were extracted from the catalog. A BED file containing these positions and their corresponding LD blocks was prepared. Each BED formatted annotation file was used to annotate sample VCF files and a custom Python script used to extract a simplified annotation of each variant which collapses together multiple gene mapping sources keeping: chromosome, position, reference allele, alternate allele, genotype, CADD scores, Entrez ids of mapped genes, snpEff annotation. Gene mapping based on GWAS and eQTL information was done only if the expected risk allele was observed in the sample.

Gene set-wise mutational burden scores

Gene set-wise mutational burden scores were calculated by counting the number of unique mutations in the genes corresponding to each gene set.

Polygenic risk score

We retrieved log2 odds ratios for all SNPs with an epilepsy phenotype from the PheWAS Catalog [59], and for each subject computed the dot product of the SNP log2 odds-ratios with the patient-specific SNP genotypes. More specifically, let N be the total number of epilepsy related SNPs, i be a patient indicator, j be a SNP indicator, x_j be the log2 odds-ratio for SNP j, g_{ij} be the genotype of patient i for SNP j. Then the polygenic risk score y_i for patient i is computed as follows:

$$y_i = \sum_{j=1}^{N} x_j g_{ij}$$

Model training

All models were trained using nested cross-validation, with 10 inner folds for hyperparameter optimization, and 10 outer folds for performance estimation. To assess the dependence of performance estimation on the chosen folds, 20 repeats were done using different folds. For all models except the multimodal neural network, hyperparameter settings were selected using a grid search. Because of the large number of hyperparameters to be optimized for the multimodal neural network, as well as the computationally expensive model fitting procedure, hyperparameter settings were selected using Bayesian optimization instead [60]. Moreover, for the multimodal neural network, each hyperparameter optimization in the inner loop of the nested cross-validation was repeated three times for stability purposes. The models were implemented using the following software packages:

- Sparse multi-block PLS-DA: mixOmics [61].

- Multimodal neural network: Tensorflow 2 [62]. A multimodal architecture was set up by limiting connections in the first layer(s) of the neural network to connections only between features from the same data modality. Given the limited number of samples, this model is highly complex, and needs to be strongly regularized to avoid

overfitting. For this purpose, we implemented (1) a group lasso penalty [63] on the input layers for the different data modalities, (2) overall l1 [64] and l2 [65] penalties on the subsequent layers, (3) dropout [66] and batch normalization [67]. The neural networks were trained using an Adam optimizer [68], with beta1 = 0.9 and beta2 = 0.999. The learning rate, together with the regularization penalties, dropout rate, batch size and architecture, was optimized using Bayesian optimization [69].

- Elastic net: glmnet [70].

- Gradient-boosted decision trees classifier: xgboost [71].

- Stacked gradient-boosted decision trees classifier: xgboost [71]. Individual xgboost classifiers were trained for each data modality. Predictions of these individual classifiers were then stacked [72] using logistic linear regression.

Power calculations

For a given classifier threshold h, the positive predictive value for drug-treated subjects (ppv_{trt}) was determined using the classifier trained on both clinical and genetics data, as described above and in the main text. Since no WGS data was available for placebo-treated patients, we constructed an additional model using clinical data alone, and applied this model (with the same threshold h) to the placebo-treated patients to determine the positive predictive value ppv_{pcb}. These two proportions were interpreted as responder rates in a hypothetical confirmatory trial, and power calculations were performed for a one-sided two-sample test for proportions (using the R stats package; at significance level 0.01 and power 90%), to determine the minimum sample size for which ppv_{trt} is significantly larger than ppv_{pcb}. Note that because WGS data is not available for placebo-treated patients, ppv_{pcb} is necessarily based on clinical data only, whereas ppv_{trt} is based on clinical and genetics data, and hence the calculations are dependent on the assumption that the inclusion of epilepsy- and brivaracetam-specific genetics variants do not inflate placebo response too strongly.

Learning curve

We re-trained our classifier, as described above, but with increasing sample sizes in the inner cross-validation loop (ranging from n = 45 to n = 211), while recording the AUC on a left-out set of fixed size (n = 24) in the outer cross-validation loop. This was repeated 20 times.

Data availability

Clinical studies (NCT01261325, NCT00490035) can be accessed via ClinicalTrials.gov.

Results

Defining discovery and validation datasets

In order to develop a predictive model for drug response, we used two datasets, (1) a discovery dataset and (2) a retrospective validation dataset [2]. We constructed a discovery dataset from phase III clinical trial data (NCT01261325) on drug response in 235 adult epilepsy patients, who consented to give a DNA sample and were treated with either 100mg or 200mg of brivaracetam per day, the approved brivaracetam doses (Table 1) [73]. The overall response rate in this dataset was 37%, with an average age of 40.7 years, 48% female, and focus localization mostly confirmed to be frontal or temporal. This discovery dataset was used for constructing, optimizing and internally validating the model. Additionally, we constructed a validation dataset by collecting data on 47 adult epilepsy patients from an independently conducted clinical trial (NCT00490035). These 47 patients consented to give a DNA sample and were treated with 100mg of brivaracetam per day (Table 1) [74]. The overall response rate in this dataset was 36%, with an average age of 41.5 years, 43% female, and focus localization mostly confirmed to be frontal or temporal. We used this dataset to determine the extent to which the internally validated model generalized to unseen patient data, with potentially different patient characteristics.

In addition to drug response (defined as >50% seizure frequency reduction 12 weeks after study baseline), available data modalities in both datasets included a comprehensive range of demographic and clinical characteristics, as well as WGS data for each patient. More details about the demographic and clinical data available for each patient can be found in [75,76(p15),77] and the Supplementary Material (Figure S1, Table S3). The data was pre-processed as described in section Materials and Methods.

Taking a knowledge-based approach to feature extraction from WGS data

WGS data contains potentially millions of variants. This is orders of magnitude larger than the typical number of patients enrolled in a clinical study, and therefore poses a high risk of overfitting, i.e. bad generalization performance of the ML model [78]. The sparsity, particularly of rare variants, further complicates this problem. Hence, we took a knowledge-based approach by extracting 40 disease-mechanism and drug mechanism of action (MOA) associated features from the original ~20 million variants, to be used as input into our machine learning model (Figure 1A). More specifically, we first defined literature-derived sets of genes that related to (1) epilepsy disease mechanisms and (2) brivaracetam's mechanism of action (Figure 1B and section Materials and Methods). With respect to (1), we only considered single nucleotide variants (SNVs) that showed a significant association to the disease phenotype in genome wide association studies (GWAS), as listed in the GWAS Catalog [50], or showed strong linkage disequilibrium (LD) to any of those SNVs (r^2>0.8). With respect to (2), we only considered SNVs proximal to SV2A (±10kb) and/or with expression quantitative trait locus (eQTL) evidence for SV2A, as documented by the Genotype-Tissue Expression (GTEx) project [55]. These SNVs were then mapped to genes by proximity of at most 10kb upstream or downstream of transcription start site, or by significant influence on gene expression in brain tissues, as documented by GTEx.

To arrive at a single mutational load score per gene set, we counted the number of SNVs mapped to the genes in each gene set. This resulted in 38 features to be used as input into our machine learning models. In addition, we defined a polygenic epilepsy risk score using the PheWAS catalog [59] (section Materials and Methods), and a feature indicating the presence of at least one structural variant overlapping brivaracetam's receptor gene SV2A (±10kb). This resulted in a total of 40 features derived from the genetics data, which by the approach taken above, we treated as three distinct genetic data modalities, (1) the gene set-wise mutational load scores, (2) the polygenic risk score and (3) the SV2A structural variance feature. Together with the clinical data, this resulted in four data modalities. For more details, we refer the reader to the section Materials and Methods.

Machine learning-based data integration successfully predicts AED response

We applied and evaluated several machine learning approaches to systematically integrate our four data modalities for predicting response to brivaracetam:

- Sparse multi-block partial least squares discriminant analysis [61,79], with blocks defined by the data modalities.

- A multimodal neural network. Here, each data modality is represented by its own sub-network at the input of the neural network, with initially no connections allowed between the sub-networks. Only at deeper network levels, the different sub-networks are connected to jointly compute a response probability. For more details regarding the architecture and hyperparameters, refer to section Materials and Methods.

- An elastic net classifier [70,80], trained on all data modalities jointly.

- A gradient-boosted decision trees classifier [71], trained on all data modalities jointly.

- Logistic regression-based stacking [72] of gradient-boosted trees classifiers.

These machine learning approaches spanned the whole range of data integration strategies [81–83], from early (e.g. sparse multi-block partial least squares discriminant analysis) to intermediate (the multimodal neural network) to late (the stacked classifier). Our best model, a single gradient-boosted trees classifier trained jointly on all data modalities, significantly outperformed all other models and achieved an AUC of 0.76 ($p = 3.8E-6$; Figure 2A,B). Interestingly, the two linear methods (elastic net and sparse multi-block partial least squares discriminant analysis) showed substantially poorer performance than the two tree-based classifiers. This suggests that our problem of brivaracetam response prediction is, to an important extent, non-linear in nature. The multimodal neural network also performed relatively poorly. This is likely due to the large number of parameters specified by such a network, relative to the limited number of samples available for training the network.

Our model generalizes to independent clinical trial data

Each clinical study has unavoidable biases due to the selection and sampling process of patients from the overall population. To make sure that the internally cross-validated performance of our model was not driven by certain characteristics unique to the discovery

dataset, we collected data from an additional independently conducted clinical trial (NCT00490035, n = 47; Table 1), and used this data to externally validate our model. Despite significant overall differences in clinical and patient characteristics between the two datasets (Supplementary Material Figure S1), our classifier generalized well: It achieved an AUC of 0.75 (Figure 2C), very close to the cross-validated AUC of 0.76 on the discovery dataset, although due to the limited size of the validation dataset, the asymptotic 95% confidence interval was wide (0.6 – 0.9).

Both clinical and genetic features uniquely impact drug response

A better understanding of the mechanisms driving drug (non-)response is instrumental in precision medicine. Hence, we next analyzed the predictive strength of our model from the level of entire data modalities down to the level of individual clinical and genetic patient features. To this end, we trained gradient-boosted trees classifiers on each of the four individual data modalities alone and compared their performance with our best model. Our best model significantly outperformed all models trained on individual data modalities (p = 1.9e-06, Figure 3A), demonstrating the unique contribution of both clinical and genetic factors to predicting drug response.

To investigate in more detail which individual patient features were most important in determining drug response, we computed feature-wise absolute SHAP values (SHapley Additive exPlanations) [84] (Figure 3B), averaged across samples. SHAP value estimation decomposes each prediction into a sum of feature importance scores and a bias term, where each feature importance score represents the change in the expected prediction when conditioning on that feature [84]. Consistent with the results in Figure 3A, the main determinants of drug response prediction represented a mix of clinical and genetic patient features. Strongly predictive clinical features included prior use of levetiracetam and epileptic focus localization. In the model, prior use of levetiracetam, an AED having the same primary receptor as brivaracetam, negatively impacted the probability of response to brivaracetam (Figure 3C). Likewise, extra-temporal epileptic focus localization negatively impacted the probability of response (Figure 3C). These two observations were also reflected in the patient population, where non-responders more often had used levetiracetam prior to study entry (Figure 3D). This was consistent with a previous study suggesting that brivaracetam non-responders are more likely to suffer from drug-refractory epilepsy, and thus have generally used relatively more AEDs prior to study entry [85]. It should be noted that the brivaracetam

responder rate in the levetiracetam-naïve group is roughly twice as high as the responder rate in the levetiracetam-exposed group (Supplementary Material Figure S2A). This association is highly significant, even when correcting for prior exposure to any of the other AEDs in our data, or when correcting for confounders not restricted to AEDs (Supplementary Material Figure S2B). This indicates that prior levetiracetam use is uniquely informative in predicting brivaracetam response, relative to the other AEDs and all other variables in the model, which is likely due to levetiracetam having the same primary receptor gene as brivaracetam.

Strongly predictive genetic features included the presence of structural variants overlapping brivaracetam's receptor gene SV2A, and the mutational load in the GO:0051011 gene set, a GO term representing microtubule minus end binding (Figure 3C). Note that epilepsy can be caused by disorders of neurodevelopmental processes, and microtubule associated proteins are known to play a role in the regulation of such processes [86,87]. In the model, having more mutations in the GO:0051011 gene set positively impacted the probability of response to brivaracetam (Figure 3D). On the other hand, having structural variants overlapping brivaracetam's receptor gene SV2A (±10kb) negatively impacted the probability of response (Figure 3D). These two observations were also reflected in the patient population. Patients with a structural variant overlapping SV2A (±10kb) were more likely not to respond to brivaracetam (Figure 3D). Moreover, responders had more mutations in the GO:0051011 gene set (Figure 3D).

SHAP dependence plots and brivaracetam response associations for all features can be found in Supplementary Material Figure S3 and S4.

Placebo response is not related to model determinants

Some of the features identified in Figure 3B could potentially also be associated with placebo response. Therefore, we wanted to assess to what extent our model was also partly capturing placebo response. As WGS data was not available for placebo-treated patients, we focused on the model trained on only clinical data from brivaracetam-treated patients to address this question. We applied this model to 235 additional, placebo treated, patients collected from the same clinical study as our discovery dataset (Table 1, clinical study NCT01261325) [73]. Placebo response could not be predicted significantly better than random (AUC = 0.56±0.088), indicating that our models indeed did not capture placebo response to any significant degree (Supplementary Material Figure S5A). In addition, we trained a new

classifier on all clinical data available for the placebo-treated patients, for predicting placebo response directly. Supporting the analysis above, the model did not perform better than random (AUC = 0.5±0.019; Supplementary Material Figure S5C).

Finally, to determine whether placebo response was related to any of the individual main clinical determinants of the integrated model (Figure 2A,B), we univariately tested the association of placebo response with the main clinical determinants of that model (Figure 3B). No single significant feature could be identified at a false discovery rate of 10% (Supplementary Material Figure S6).

Reducing sample size of confirmatory studies by enriching for drug responders

Clinical trials in epilepsy, but also in other disease areas, generally suffer from high non-responder rates [12,88]. A higher non-responder rate means that more patients will need to be enrolled to establish a statistically significant difference in responder rates between drug- and placebo-treated subjects. However, models trained to predict drug response on data from an initial exploratory study, could be used to enrich for responders in subsequent confirmatory studies (so-called enrichment trial according to FDA terminology[1]). As such, minimum required sample sizes for these confirmatory studies could be reduced.

More specifically, note that the output of our machine learning model is a probability for each individual patient to respond positively to brivaracetam after 12 weeks. In screening patients for a confirmatory study, a specific threshold could be applied to this probability to decide which patients would be accepted and which would not. Choosing to accept only patients with a very high probability of response results in a high specificity (many accepted patients are responders), but low sensitivity (many responders are not accepted). Conversely, accepting patients even at low predicted response probabilities results in a higher sensitivity (many responders are accepted) but lower specificity (many accepted patients are non-responders). This trade-off between sensitivity and specificity is visualized by the ROC in Figure 2B.

As a concrete example, setting the classifier threshold at 0 would result in a specificity of 0% and a sensitivity of 100% (Figure 2B), which would imply accepting all patients at screening,

---

[1] https://www.fda.gov/media/121320/download

and lead to the original responder rate in our discovery dataset (38%; Figure 4A). On the other hand, setting the classifier threshold at 0.45 would result in a specificity of 66% and a sensitivity of 74% (Figure 2B), which could result in a 1.5 times higher responder rate (~57%; Figure 4A). This higher responder rate could subsequently reduce the minimum sample size required for a confirmatory study (at 90% statistical power and 0.01 significance threshold) from ~230 subjects in each study treatment arm, to ~50 subjects (classifier thresholds at 0 and 0.45 respectively; Figure 4B), a reduction of ~65%. This reduction would come at the expense of stricter confirmatory study inclusion and exclusion criteria, as determined by the classifier threshold: In the current example, setting the classifier threshold at 0.45 would imply that ~50% of patients would not meet the criteria for study entry (Figure 4A). A classifier with even better performance than the current model should be able to increase this acceptance rate, by showing a relative decrease in false negative calls at the same responder rate.

Note that the response rates for drug-treated subjects were estimated using both clinical and genetics data, whereas the response rates for placebo-treated subjects were necessarily estimated using clinical data alone, as no WGS data was available for placebo-treated subjects (for more details, please refer to section Materials and Methods).

Note also that, even when estimating both placebo and treatment response rates using only clinical data, substantial reductions in required sample size can still be observed (Supplementary Material Figure S7).

We then wanted to determine whether, given the current modeling approach, better models could likely be constructed. As outlined above, such models would be able to achieve higher study acceptance rates at similar responder rates, or higher responder rates at similar acceptance rates. For this purpose, we re-trained our classifier with increasing sample sizes in the inner cross-validation loop (ranging from n = 45 to n = 211), while recording the AUC on a left-out set of fixed size (n = 24) in the outer cross-validation loop. The roughly linearly increasing AUC showed no signs of reaching a plateau for larger n, which indeed suggested that with more patients, even better performing models can be built (Figure 4C).
To get an indication of how many patients would be needed to achieve even better performance, we fitted a robust linear regression model to the data in Figure 4C. Extrapolating the linear trend using this regression model suggested that it may be reasonable to expect that at least 350 patients would be needed to achieve a performance of AUC ~= 0.9 (Figure 4D).

Discussion

In this work, we presented a retrospectively validated ML model for predicting clinical drug response to the AED brivaracetam. We demonstrated that integrating WGS data with clinical data significantly improves predictive performance compared to a model based on clinical data or genetic data alone, i.e. that both clinical and genetic patient features uniquely contribute to model performance. Importantly, we also showed that it is unlikely that our model captured placebo response to any significant degree. This analysis was restricted to clinical determinants since, due to the cost considerations of generating WGS data and our focus on the determinants of clinical drug (non-)response, only genomes of brivaracetam-treated patients were sequenced.

In principle, WGS allows for a hypothesis-free way of analyzing the influence of genomic variation on an outcome of interest. However, the dimensionality of WGS data is huge relative to the small number of patients available in our study, posing serious risks of overfitting. Hence, it is clear that prior knowledge needs to be used to maximally reduce the WGS feature space in the most informative way possible. For this reason, we initially analyzed all variants in the WGS data at the level of sets of genes known to be related to epilepsy and/or brivaracetam's MOA. Genes were organized into functional categories based on prior biological knowledge in the form of biological pathways, molecular functions, and human phenotypes. Such an approach allows us to simultaneously reduce the feature space based on those genes and linked genetic variants most likely to exert an effect on drug response, accommodate sparse rare variants, and generate features with a functional interpretation. In the end, gene set-wise mutational load scores were computed by simply counting the number of mutations mapped to the genes in a gene set. This can be thought of as a linear model that maps input variants, each with a coefficient equal to 1, to gene set-wise scores. One could think of more involved approaches, for example to estimate the coefficients of such a (possibly regularized) linear model using the class labels, as an integral part of the nested-cross validation. However, in our case this led to substantial overfitting.

Our results confirm that the mutational burden in our epilepsy-and MOA-related gene sets are important factors in explaining and predicting response. Two of the most informative genetic features in the model were the presence of structural variants overlapping the SV2A gene, and the mutational load in a gene set representing microtubule minus end binding. Brivaracetam is a high affinity SV2A ligand whose purported mechanism of action involves

blocking the SV2A membrane glycoprotein and inhibiting neurotransmitter release [89]. We showed that brivaracetam non-responders were significantly more likely to carry SV2A structural variants than responders (p=0.0055, Fig 3d). However, previous published work failed to find any impact of single SV2A variants on response to levetiracetam, an AED in the same drug class as brivaracetam, but with a less specific SV2A binding profile [90]. This again highlights the necessity for summarizing genetic variation into polygenic risk scores, gene-level and even gene set-level scores, to effectively address the challenges that are posed by the analysis of individual variants, which are often rare events.

It is worth noting that, except for prior use of levetiracetam, the clinical data alone contains no information that is specific to the action of, or response to brivaracetam. In this respect, a model built on clinical data alone would likely have some utility in predicting response to other drugs, i.e. would mostly have prognostic value. Including genetic features that are specific to the MOA and disposition of brivaracetam means that the model is likely to have diagnostic value, i.e. be specific to predicting response to brivaracetam. This is supported by our observation that genetic features contributed significantly to the performance of the integrated model (Figure 3A) and suggests that genetic variation is at least to some extent, complementary to clinical information in our setting. In other words, both genetic variation and clinical information each describe different aspects of brivaracetam response. Future work will entail pursuing this hypothesis by testing the model in equivalent data sets derived from other AEDs.

The above discussion demonstrates that our results can provide guidance as to the type of genetic features that are important in determining drug response and which should be considered when translating these approaches into practice, with a focus on genes related to the disease understudy and the drugs' mechanism of action, and pharmacokinetics.

Among the clinical features in the present study, we found that prior levetiracetam use was the most important predictor of brivaracetam response. Previous work has indeed shown that levetiracetam-naïve patients exhibited stronger response to brivaracetam [76,85]. This demonstrates that using ML, we are able to objectively identify variables for which clinical relevance has previously been established in the literature. In the context of this finding, it is important to note that despite showing a weaker response to brivaracetam, levetiracetam-exposed patients showed a better-than-placebo response [76,85]. Moreover, levetiracetam-exposed patients generally showed an attenuated response to a wide range of AEDs, leading to the suggestion that the observed brivaracetam non-response was due to a more general

AED-refractoriness [85]. In our data however, we observed that prior use of levetiracetam was uniquely informative for predicting brivaracetam response relative to all other AEDs in the model, likely due to levetiracetam having the same receptor gene as brivaracetam.

In applying our model to responder enrichment for clinical trial design, we observed that it is likely that better models could be built when using more patients than available in our current discovery dataset. However, below n $\sim$= 100 patients, performance was little better than random, suggesting limited utility in responder enrichment for numbers lower than 100. Thus, there is a balance that needs to be considered when applying these approaches to inform clinical study design: Exploratory studies would need to be large enough to support the building of models that would allow for sufficiently strong predictive enrichment in, and reduction in size of, confirmatory studies.

The above considerations represent the first step in the translational pipeline towards clinical practice, supporting the development of companion diagnostics. The actual translation of our present method into routine clinical practice would require several additional steps. First, further validation of our model is required: After internal and retrospective validation as performed here, the next step would be a prospective clinical validation study, in which the benefit compared to standard routes for patient selection has to be demonstrated [2]. After this prospective clinical validation, the test itself, genotyping and algorithm, would need to be made available in a format, and on a platform that could be employed in a routine clinical care environment, such as a hospital. In the current format our model uses 106 clinical and 40 diverse genetic features, comprising a total of 4695 genetic variants, hence being technically challenging and relatively labor and cost intensive. In this context, it should also be noted that clinical features captured in clinical trials are not necessarily in line with what is measured in routine clinical practice. Future work should thus focus on streamlining the model towards translation into clinical practice, including development of a cost-effective assay for the most informative genetic variants, routes required to make algorithms available and accessible and inclusion of clinical features that might be more routinely collected. A systematic way of assessing feature importance in ML models, such as the SHAP values used in our work, can support this translation by helping to deconvolute the model and identify the features that are most informative.

In case of success, the regulatory implications for an impact on the drug label and companion diagnostic co-approval need to be considered. Regulatory authorities are interested in the ability to target treatments to patients who would most benefit, needing to ensure that the

label accurately reflects the enrichment strategies used to select the patients, and that there is an approved method available to identify these patients once any drug is approved. In general, the potential effect of enrichment strategies on labeling and the route of approval of any test as a companion diagnostic should be part of an ongoing dialogue with regulatory bodies during drug development[2].

Our work represents the first retrospectively validated machine learning model linking drug mechanism of action and the genetic, clinical and demographic background of epilepsy patients to clinical drug response. As such, it provides a blueprint for applying ML approaches to achieve the key ambition of precision medicine across a variety of disease areas, including neurological disorders. In particular, extracting informative features from very high dimensional and sparse genomic data is challenging but relevant to many disease areas. Our work demonstrates that even with limited sample size, integrating genetics data with clinical data for informing drug response prediction is possible, when using advanced ML methods combined with literature knowledge to manage the huge dimensionality of the genetics data. Furthermore, such models have substantial potential utility for clinical practice and in future clinical trial design by their ability to enrich for responders. Integration of additional data modalities, such as EEG or imaging, may further improve model performance. Additionally, this ability to identify and characterize subpopulations more likely to respond, shows that within the domain of neurological disorders and specifically epilepsy, we can begin to think more systematically about rational and personalized clinical decision making and reducing the disease burden for patients on an individual level.

Acknowledgements

Funding

---

[2] https://www.fda.gov/media/121320/download

employees, were responsible for data interpretation, revising the manuscript for intellectual content, and approving of the manuscript for submission.

Competing interests

Author contributions

Designing and implementing the computational methods and experiments: J.d.J.; preparing the data: I.C., S.E., M.P.; analyzing the results: all authors; writing the manuscript: J.d.J.; co-writing the manuscript: H.F., M.A., defining the project: C.D., H.F., M.A.; supervising the project: H.F., M.A.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Vogenberg FR, Isaacson Barash C, Pursel M. Personalized medicine: part 1: evolution and development into theranostics. *P & T : a peer-reviewed journal for formulary management*. 2010;35(10):560-576.

2. Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. *BMC Medicine*. 2018;16(1):150. doi:10.1186/s12916-018-1122-7

3. FDA. Table of Pharmacogenomic Biomarkers in Drug Labeling. Published 2020. Accessed June 1, 2020. https://www.fda.gov/drugs/science-and-research-drugs/table-pharmacogenomic-biomarkers-drug-labeling

4. Armstrong M. The genetics of adverse drug reactions: Promises and problems. *Methods in Pharmacology and Toxicology*. Published online 2008:121-147. doi:10.1007/978-1-59745-439-1-7

5. Peck RW. Precision Medicine Is Not Just Genomics: The Right Dose for Every Patient. *Annual Review of Pharmacology and Toxicology*. 2018;58(1):105-122. doi:10.1146/annurev-pharmtox-010617-052446

6. Harrer S, Shah P, Antony B, Hu J. Artificial Intelligence for Clinical Trial Design. *Trends in Pharmacological Sciences*. 2019;40(8):577-591. doi:https://doi.org/10.1016/j.tips.2019.05.005

7. De Martini D. Empowering phase II clinical trials to reduce phase III failures. *Pharmaceutical Statistics*. 2020;19(3):178-186. doi:10.1002/pst.1980

8. Paik S, Shak S, Tang G, et al. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *New England Journal of Medicine*. 2004;351(27):2817-2826. doi:10.1056/NEJMoa041588

9. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(6871):530-536. doi:10.1038/415530a

10. Strafella C, Caputo V, Galota MR, et al. Application of Precision Medicine in Neurodegenerative Diseases. *Front Neurol*. 2018;9:701. doi:10.3389/fneur.2018.00701

11. Tan L, Jiang T, Tan L, Yu J-T. Toward precision medicine in neurological diseases. *Ann Transl Med*. 2016;4(6):104. doi:10.21037/atm.2016.03.26

12. Daniel F, French J. Designing better trials for epilepsy medications: the challenge of heterogeneity. *Clinical Investigation*. 2013;3:927-934. doi:10.4155/cli.13.77

13. CenterWatch. *FDA Approved Drugs: Listings in Epilepsy*.; 2020.

14. Kwan P, Brodie MJ. Early Identification of Refractory Epilepsy. *New England Journal of Medicine*. 2000;342(5):314-319. doi:10.1056/NEJM200002033420503

15. Chen Z, Brodie MJ, Liew D, Kwan P. Treatment Outcomes in Patients With Newly Diagnosed Epilepsy Treated With Established and New Antiepileptic Drugs: A 30-Year

Longitudinal Cohort Study. *JAMA Neurology*. 2018;75(3):279-286. doi:10.1001/jamaneurol.2017.3949

16. Perucca E, Brodie MJ, Kwan P, Tomson T. 30 years of second-generation antiseizure medications: impact and future perspectives. *The Lancet Neurology*. 2020;19(6):544-556. doi:https://doi.org/10.1016/S1474-4422(20)30035-1

17. Laxer KD, Trinka E, Hirsch LJ, et al. The consequences of refractory epilepsy and its treatment. *Epilepsy and Behavior*. 2014;37:59-70. doi:https://doi.org/10.1016/j.yebeh.2014.05.031

18. Begley CE, Durgin TL. The direct cost of epilepsy in the United States: A systematic review of estimates. *Epilepsia*. 2015;56(9):1376-1387. doi:10.1111/epi.13084

19. Aronica E, Gorter JA, Ramkema M, et al. Expression and Cellular Distribution of Multidrug Resistance–related Proteins in the Hippocampus of Patients with Mesial Temporal Lobe Epilepsy. *Epilepsia*. 2004;45(5):441-451. doi:10.1111/j.0013-9580.2004.57703.x

20. Feldmann M, Asselin M-C, Liu J, et al. P-glycoprotein expression and function in patients with temporal lobe epilepsy: a case-control study. *The Lancet Neurology*. 2013;12(8):777-785. doi:10.1016/S1474-4422(13)70109-1

21. Klotz U. The Role of Pharmacogenetics in the Metabolism of Antiepileptic Drugs. *Clinical Pharmacokinetics*. 2007;46(4):271-279. doi:10.2165/00003088-200746040-00001

22. Ghosh C, Marchi N, Desai NK, et al. Cellular localization and functional significance of CYP3A4 in the human epileptic brain. *Epilepsia*. 2011;52(3):562-571. doi:10.1111/j.1528-1167.2010.02956.x

23. Emich-Widera E, Likus W, Kazek B, et al. CYP3A5*3 and C3435T MDR1 Polymorphisms in Prognostication of Drug-Resistant Epilepsy in Children and Adolescents. *BioMed Research International*. 2013;2013.

24. Boussadia B, Ghosh C, Plaud C, et al. Effect of status epilepticus and antiepileptic drugs on CYP2E1 brain expression. *Neuroscience*. 2014;281:124-134. doi:https://doi.org/10.1016/j.neuroscience.2014.09.055

25. Dauchy S, Dutheil F, Weaver RJ, et al. ABC transporters, cytochromes P450 and their main transcription factors: expression at the human blood–brain barrier. *Journal of Neurochemistry*. 2008;107(6):1518-1528. doi:10.1111/j.1471-4159.2008.05720.x

26. Nurmohamed L, Garcia-Bournissen F, Buono RJ, Shannon MW, Finkelstein Y. Predisposition to epilepsy—Does the ABCB1 gene play a role? *Epilepsia*. 2010;51(9):1882-1885. doi:10.1111/j.1528-1167.2010.02588.x

27. Grover S, Kukreti R. A systematic review and meta-analysis of the role of ABCC2 variants on drug response in patients with epilepsy. *Epilepsia*. 2013;54(5):936-945. doi:10.1111/epi.12132

28. Bournissen FG, Moretti ME, Juurlink DN, Koren G, Walker M, Finkelstein Y. Polymorphism of the MDR1/ABCB1 C3435T drug-transporter and resistance to anticonvulsant drugs: A meta-analysis. *Epilepsia*. 2009;50(4):898-903. doi:10.1111/j.1528-1167.2008.01858.x

29. Balan S, Bharathan SP, Vellichiramal NN, et al. Genetic Association Analysis of ATP Binding Cassette Protein Family Reveals a Novel Association of ABCB1 Genetic Variants with Epilepsy Risk, but Not with Drug-Resistance. *PLOS ONE*. 2014;9(2):1-10. doi:10.1371/journal.pone.0089253

30. van der Weide J, Steijns LSW, van Weelden MJM, de Haan K. The effect of genetic polymorphism of cytochrome P450 CYP2C9 on phenytoin dose requirement. *Pharmacogenetics and Genomics*. 2001;11(4). https://journals.lww.com/jpharmacogenetics/Fulltext/2001/06000/The_effect_of_genetic_polymorphism_of_cytochrome.2.aspx

31. Dombrowski SM, Desai SY, Marroni M, et al. Overexpression of Multiple Drug Resistance Genes in Endothelial Cells from Patients with Refractory Epilepsy. *Epilepsia*. 2001;42(12):1501-1506. doi:10.1046/j.1528-1157.2001.12301.x

32. Cavalleri GL, McCormack M, Alhusaini S, Chaila E, Delanty N. Pharmacogenomics and epilepsy: the road ahead. *Pharmacogenomics*. 2011;12(10):1429-1447. doi:10.2217/pgs.11.85

33. Silva-Alves MS, Secolin R, Carvalho BS, et al. A Prediction Algorithm for Drug Response in Patients with Mesial Temporal Lobe Epilepsy Based on Clinical and Genetic Information. *PLOS ONE*. 2017;12. doi:10.1371/journal.pone.0169214

34. Petrovski S, Szoeke CE, Sheffield LJ, D'souza W, Huggins RM, O'brien TJ. Multi-SNP pharmacogenomic classifier is superior to single-SNP models for predicting drug outcome in complex diseases. *Pharmacogenet Genomics*. 2009;19(2):147-152. doi:10.1097/FPC.0b013e32831d1dfa

35. Shazadi K, Petrovski S, Roten A, et al. Validation of a multigenic model to predict seizure control in newly treated epilepsy. *Epilepsy Research*. 2014;108(10):1797-1805. doi:10.1016/j.eplepsyres.2014.08.022

36. Devinsky O, Dilley C, Ozery-Flato M, et al. Changing the approach to treatment choice in epilepsy using big data. *Epilepsy Behav*. 2016;56:32-37. doi:10.1016/j.yebeh.2015.12.039

37. Zhang J-H, Han X, Zhao H-W, et al. Personalized prediction model for seizure-free epilepsy with levetiracetam therapy: a retrospective data analysis using support vector machine. *Br J Clin Pharmacol*. 2018;84(11):2615-2624. doi:10.1111/bcp.13720

38. Yao L, Cai M, Chen Y, Shen C, Shi L, Guo Y. Prediction of antiepileptic drug treatment outcomes of patients with newly diagnosed epilepsy by machine learning. *Epilepsy Behav*. 2019;96:92-97. doi:10.1016/j.yebeh.2019.04.006

39. Steinhoff BJ, Staack AM. Levetiracetam and brivaracetam: a review of evidence from clinical trials and clinical experience. *Therapeutic Advances in Neurological Disorders*. 2019;12:1756286419873518. doi:10.1177/1756286419873518

40. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

41. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2010;26(5):589-595. doi:10.1093/bioinformatics/btp698

42. *Picard Toolkit*. Broad Institute; 2019. http://broadinstitute.github.io/picard/

43. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110

44. Rentzsch P, Witten D, Cooper G, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*. 2019;47:D886-D894.

45. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*. 2013;30(7):1006-1007. doi:10.1093/bioinformatics/btt730

46. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993. doi:10.1093/bioinformatics/btr509

47. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92. doi:10.4161/fly.19695

48. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220-1222. doi:10.1093/bioinformatics/btv710

49. Köhler S, Carmody L, Vasilevsky N, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res*. 2019;47(D1):D1018-D1027. doi:10.1093/nar/gky1105

50. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*. 2018;47(D1):D1005-D1012. doi:10.1093/nar/gky1120

51. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2020;48(D1):D845-D855. doi:10.1093/nar/gkz1021

52. Carvalho-Silva D, Pierleoni A, Pignatelli M, et al. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research*. 2018;47(D1):D1056-D1065. doi:10.1093/nar/gky1133

53. Vejnar CE, Blum M, Zdobnov EM. miRmap web: Comprehensive microRNA target prediction online. *Nucleic Acids Res*. 2013;41(Web Server issue):W165-W168. doi:10.1093/nar/gkt430

54. Mooney C, Becker BA, Raoof R, Henshall DC. EpimiRBase: a comprehensive database of microRNA-epilepsy associations. *Bioinformatics*. 2016;32(9):1436-1438. doi:10.1093/bioinformatics/btw008

55. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*. 2013;45(6):580-585. doi:10.1038/ng.2653

56. Alexa A, Rahnenfuhrer J. *TopGO: Enrichment Analysis for Gene Ontology*. http://www.bioconductor.org/packages/release/bioc/html/topGO.html

57. Nikolsky Y, Kirillov E, Zuev R, Rakhmatulin E, Nikolskaya T. Functional analysis of OMICs data and small molecule compounds in an integrated "knowledge-based" platform. *Methods Mol Biol*. 2009;563:177-196. doi:10.1007/978-1-60761-175-2_10

58. Wang J, Lin Z-J, Liu L, et al. Epilepsy-associated genes. *Seizure - European Journal of Epilepsy*. 2017;44:11-20. doi:10.1016/j.seizure.2016.11.030

59. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*. 2013;31(12):1102-1111. doi:10.1038/nbt.2749

60. Bischl B, Richter J, Bossek J, Horn D, Thomas J, Lang M. *MlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions*.; 2017. http://arxiv.org/abs/1703.03373

61. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*. 2017;13(11):1-19. doi:10.1371/journal.pcbi.1005752

62. Martín Abadi, Ashish Agarwal, Paul Barham, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*.; 2015. https://www.tensorflow.org/

63. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*. 2006;68:49-67.

64. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58(1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

65. Tikhonov AN, Arsenin VY. *Solution of Ill-Posed Problems*. John Wiley & Sons; 1977.

66. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*. 2012;abs/1207.0580. http://arxiv.org/abs/1207.0580

67. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. JMLR.org; 2015:448-456.

68. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: Bengio Y, LeCun Y, eds. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. ; 2015. http://arxiv.org/abs/1412.6980

69. Mockus JB, Mockus LJ. Bayesian Approach to Global Optimization and Application to Multiobjective and Constrained Problems. *J Optim Theory Appl*. 1991;70(1):157-172.

70. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, Articles*. 2010;33(1):1-22. doi:10.18637/jss.v033.i01

71. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: KDD '16. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785

72. Wolpert DH. Stacked generalization. *Neural Networks*. 1992;5(2):241-259. doi:https://doi.org/10.1016/S0893-6080(05)80023-1

73. ClinicalTrials.gov. *Double-Blind, Randomized Study Evaluating the Efficacy and Safety of Brivaracetam in Adults With Partial Onset Seizures*.; 2016.

74. ClinicalTrials.gov. *Brivaracetam Efficacy and Safety Study in Subjects With Partial Onset Seizures*.; 2016.

75. Biton V, Berkovic SF, Abou-Khalil B, Sperling MR, Johnson ME, Lu S. Brivaracetam as adjunctive treatment for uncontrolled partial epilepsy in adults: a phase III randomized, double-blind, placebo-controlled trial. *Epilepsia*. 2014;55(1):57-66. doi:10.1111/epi.12433

76. Klein P, Schiemann J, Sperling MR, et al. A randomized, double-blind, placebo-controlled, multicenter, parallel-group study to evaluate the efficacy and safety of adjunctive brivaracetam in adult patients with uncontrolled partial-onset seizures. *Epilepsia*. 2015;56(12):1890-1898. doi:10.1111/epi.13212

77. Rheims S, Ryvlin P. Pharmacotherapy for tonic–clonic seizures. *Expert Opinion on Pharmacotherapy*. 2014;15(10):1417-1426. doi:10.1517/14656566.2014.915029

78. Johnstone IM, Titterington DM. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2009;367(1906):4237-4253. doi:10.1098/rsta.2009.0159

79. Singh A, Shannon CP, Gautier B, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*. 2019;35(17):3055-3062. doi:10.1093/bioinformatics/bty1054

80. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005;67(2):301-320. doi:10.1111/j.1467-9868.2005.00503.x

81. Ahmad A, Fröhlich H. Integrating Heterogeneous omics Data via Statistical Inference and Learning Techniques. *Genomics and Computational Biology*. 2016;2(1):e32. doi:10.18547/gcb.2016.vol2.iss1.e32

82. Pavlidis P, Weston J, Cai J, Grundy WN. Gene Functional Classification from Heterogeneous Data. In: *Proceedings of the Fifth Annual International Conference on Computational Biology*. RECOMB '01. Association for Computing Machinery; 2001:249-255. doi:10.1145/369133.369228

83. Maragos P, Gros P, Katsamanis A, Papandreou G. Cross-Modal integration for Performance Improving in multimedia: A review. In: Maragos P, Potamianos A, Gros P, eds. *Multimedia Processing and Interaction, Audio, Video, Text*. Vol 33. Multimedia Systems and Applications. Springer; 2008:3-45.

84. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.; 2017:4765-4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

85. Asadi-Pooya AA, Sperling MR, Chung S, et al. Efficacy and tolerability of adjunctive brivaracetam in patients with prior antiepileptic drug exposure: A post-hoc study. *Epilepsy Research*. 2017;131:70-75. doi:https://doi.org/10.1016/j.eplepsyres.2017.02.007

86. Bonini SA, Mastinu A, Ferrari-Toninelli G, Memo M. Potential Role of Microtubule Stabilizing Agents in Neurodevelopmental Disorders. *International journal of molecular sciences*. 2017;18(8):1627-.

87. Lasser M, Tiber J, Lowery LA. The Role of the Microtubule Cytoskeleton in Neurodevelopmental Disorders. *Frontiers in Cellular Neuroscience*. 2018;12:165. doi:10.3389/fncel.2018.00165

88. Hwang TJ, Carpenter DP, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA internal medicine*. 2016;176 12:1826-1833.

89.   Klitgaard H, Matagne A, Nicolas J-M, et al. Brivaracetam: Rationale for discovery and preclinical profile of a selective SV2A ligand for epilepsy treatment. *Epilepsia*. 2016;57(4):538-548. doi:10.1111/epi.13340

90.   Lynch JM, Tate SK, Kinirons P, et al. No major role of common SV2A variation for predisposition or levetiracetam response in epilepsy. *Epilepsy Research*. 2009;83(1):44-51. doi:https://doi.org/10.1016/j.eplepsyres.2008.09.003

Figure legends

Figure 1. Analysis pipeline and gene set definition. (A) Combining hybrid data-/knowledge-driven feature extraction with advanced machine learning to systematically integrate clinical and genetic data for predicting brivaracetam response. **(B)** Defining literature-derived gene sets that relate to (1) epilepsy disease etiology and (2) brivaracetam's mechanism of action.

Figure 2. Model performance and external validation. (A) Performance of several machine learning approaches to systematically integrate the genetic and clinical data modalities for predicting response to brivaracetam, as estimated using repeated cross-validation with 20 repeats (each dot represents a repeat). The performance is shown in terms of area under ROC (AUC). An AUC of 0.5 corresponds to chance level, and an AUC of 1 is the best achievable performance. Statistical significance was assessed by a paired Wilcoxon-test between the best and the second-best model. **(B)** Receiver operating characteristic (ROC) of our best model, a gradient-boosted decision trees classifier trained jointly on all data modalities. In light gray, the individual ROCs for the 20 repeats, in red the average ROC across the 20 repeats and the associated empirical 95% confidence interval based on the 20 repeats. The ROC depicts the trade-off between specificity and sensitivity of a classifier, while varying the diagnostic cutoff for the probability of response: Choosing to treat only patients with high predicted probability of response results in a high specificity (most treated patients will benefit from the treatment), but a low sensitivity (many patients that could potentially have benefited from treatment are not treated). Conversely, treating even patients with low predicted response probability results in a high sensitivity (most patients who could

potentially benefit are treated) but low specificity (many treated patients do not benefit from the treatment). The area under ROC (AUC) is a measure of overall prediction performance. An AUC of 0.5 corresponds to chance level, and an AUC of 1 is the best achievable performance. **(C)** Receiver operating characteristic (ROC) of validating of our best model on the independent validation dataset.

Figure 3. Determinants of drug response probability. (A) Performance of gradient-boosted trees classifiers on the individual data modalities, compared with the integrated model. Statistical significance was assessed by a paired Wilcoxon-test between the clinical-only model and the integrated model. **(B)** All patient features with non-zero average absolute SHAP values. Error bars represent the empirical 95% confidence intervals based on 20 cross-validation repeats. **(C)** SHAP dependence plots for four selected variables; top-to-bottom, left-to-right: Prior use of levetiracetam, extra-temporal focus localization, mutational load in gene set GO:0051011 and structural variants overlapping SV2A. Each circle represents a single patient (note that circles are often super-imposed). **(D)** Univariate associations of selected variables with brivaracetam response; top-to-bottom, left-to-right: Prior use of levetiracetam, extra-temporal focus localization, mutational load in gene set GO:0051011 and structural variants overlapping SV2A. Statistical significance was assessed by a Wilcoxon test (for GO:0051011) or a Fisher's exact test (for the other three features).

Figure 4. Application to clinical trial design. (A) The trade-off between positive rate (number of patients included in a trial) and positive predictive value (fraction of responders in a trial), as a function of the classifier threshold. 95% confidence intervals were determined using the 20 cross-validation repeats. **(B)** Using the classifier to enrich for responders in confirmatory studies: Minimum required sample size in a confirmatory trial (at 90% statistical power and 0.05 significance level) as a function of classifier threshold. Confidence intervals were determined by computing sample sizes for the confidence intervals in (A). **(C)** Training the model on increasing sample sizes in the inner cross-validation loop, while recording the performance on a set of left-out samples of fixed size ($n = 24$) in the outer cross-validation loop. Each dot represents the performance of a model trained on a different randomly subsampled number of patients (N). 95% confidence intervals were determined using the 20 cross-validation repeats. **(D)** Extrapolation of performance for larger N, using robust linear regression in blue, with a 95% prediction interval.
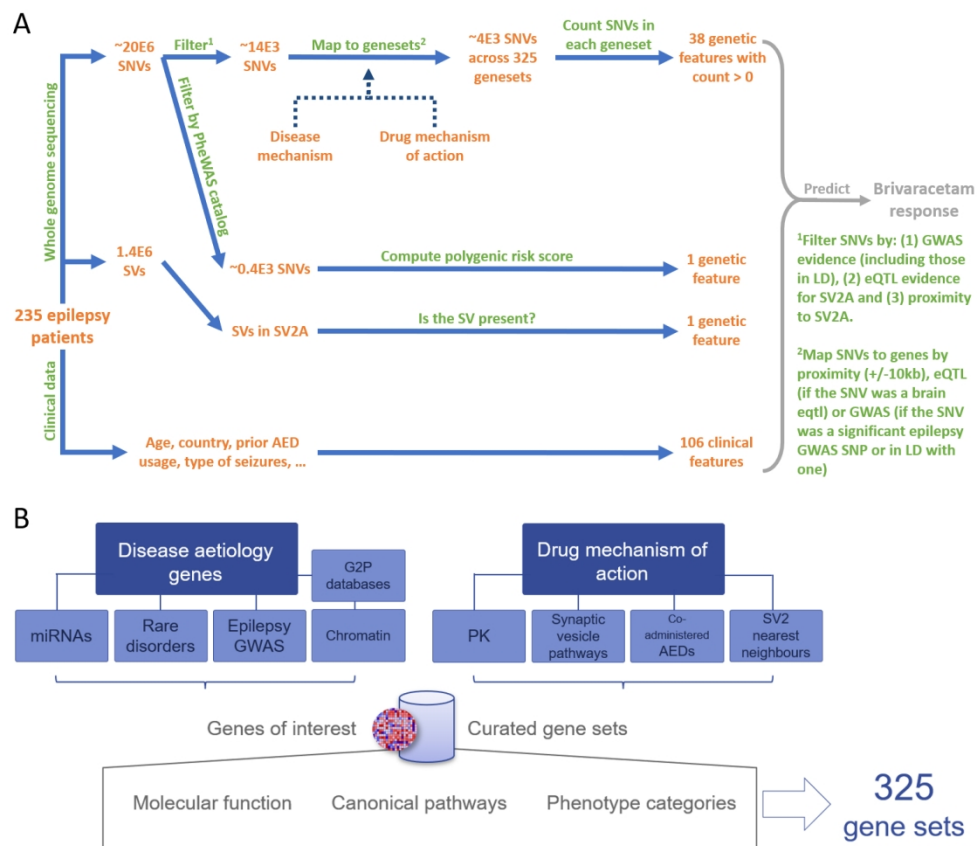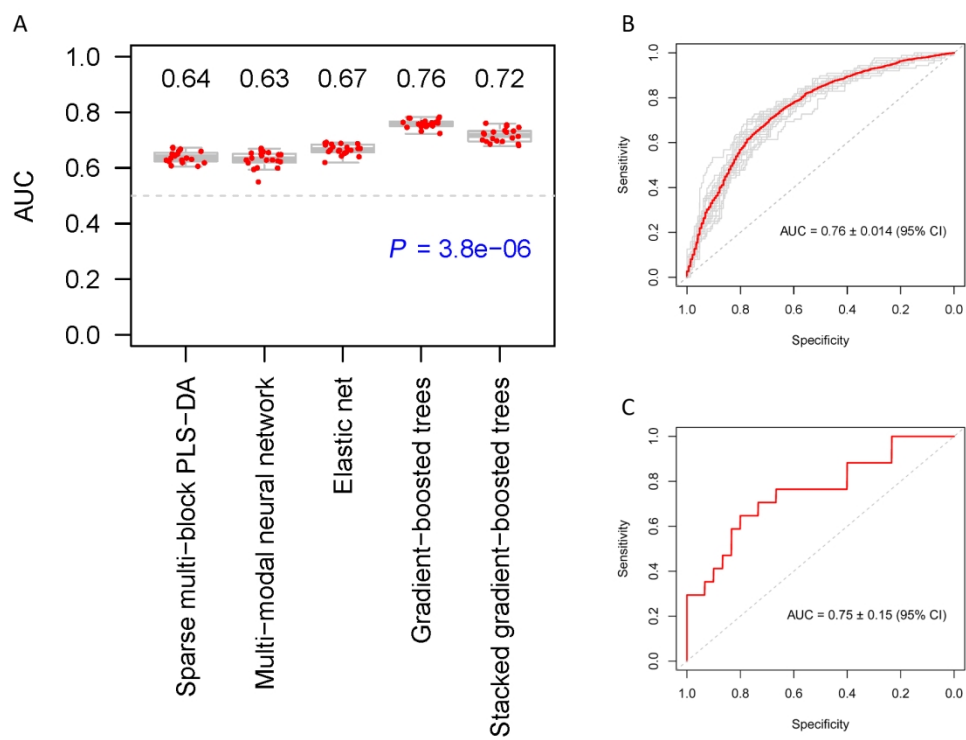
Figure 1

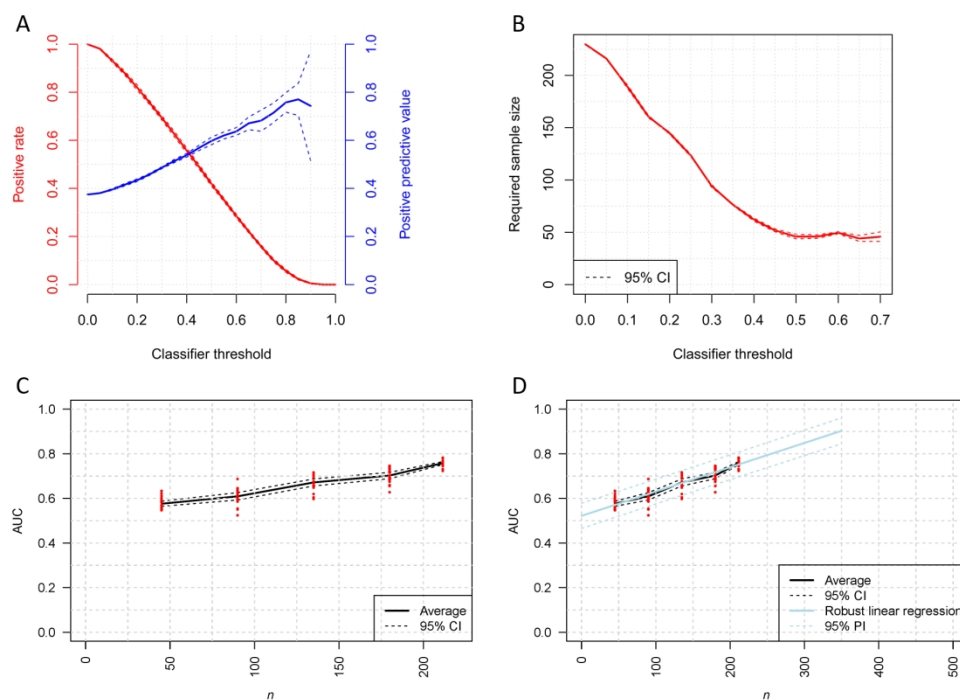266x230mm (150 x 150 DPI)

Figure 2

331x254mm (150 x 150 DPI)

Figure 3

330x262mm (150 x 150 DPI)

Figure 4

332x240mm (150 x 150 DPI)

Table 1 Discovery, validation and placebo datasets

	Discovery dataset	Validation dataset	Placebo dataset
Number of patients	235	47	235
Brivaracetam dose	100mg - 200mg per day	100mg per day	0mg per day
Clinical trial	NCT01261325	NCT00490035	NCT01261325
Whole genome sequencing?	Yes	Yes	No
Responder rate	37%	36%	22%
Age: mean (SD)	40.7 (13.1)	41.5 (13.8)	39.6 (12.9)
Gender	48% female	43% female	50% female
Focus localization			
Frontal	32%	17%	38%
Occipital	5%	6%	6%
Parietal	9%	4%	9%
Temporal	64%	60%	59%