Prospectively validated disease-agnostic predictive medicine with hybrid AI

Bragi Lovetrue[1] & Idonae Lovetrue[1]

[1]*Demiurge Technologies AG, 14 6300, Baarerstrasse, Zug, Switzerland*

Despite numerous superhuman achievements in complex challenges[1,2], standalone AI does not free life science from the long-term bottleneck of linearly extracting new knowledge from exponentially growing new data[3], severely limiting the success rate of drug discovery[4]. Inspired by the state-of-the-art AI training methods, we trained a human-centric hybrid augmented intelligence[5] (HAI) to learn a foundation model[6] that extracts all-encompassing knowledge of human physiology and diseases. To evaluate the quality of HAI's extracted knowledge, we designed the public, prospective prediction of pivotal ongoing clinical trial outcomes at large scale (PROTOCOLS) challenge to benchmark HAI's real-world performance of predicting drug clinical efficacy without access to human data. HAI achieved a 11.4-fold improvement from the baseline with 99% confidence[7] in the PROTOCOLS validation, readily increasing the average clinical success rate of investigational new drugs from 7.9%[8] to 90.1% for almost any human diseases[9,10]. The validated HAI confirms that exponentially extracted knowledge alone is sufficient for accurately predicting drug clinical efficacy, effecting a total reversal of Eroom's law[4]. HAI is also the world's first clinically validated model of human aging that could substantially speed up the discovery of preventive medicine for all age-related diseases[11]. Our results demonstrate that disruptive break-

throughs necessitate the smallest team size to attain the largest HAI for optimal knowledge extraction from high-dimensional low-quality data space, thus establishing the first prospective proof of the previous discovery that small teams disrupt[12]. The global adoption of training HAI provides a well-beaten economic path to mass-produce scientific and technological breakthroughs via exponential knowledge extraction and better designs of data labels for training better-performing AI[13].

1 INTRODUCTION

Despite numerous superhuman achievements in complex challenges[1,2], standalone AI does not free life science from the long-term bottleneck of linearly extracting new knowledge from exponentially growing new data[3]. The shortage of knowledge in life science accelerates the decline of productivity in drug discovery[4]. Only 20% of biomedical research generated reproducible data[14]. Less than 10% of animal studies successfully predicted the clinical efficacy of novel drugs[15].

The life science community divides on how to best address the knowledge shortage. Many believe that increasing the scale of data collection is sufficient. In contrast, others believe that increasing the efficiency of knowledge extraction is necessary[3].

High dimensionality and low quality are two defining characteristics of life science data. Biological systems have intractably many entities at multiple scales. Biomedical data are difficult to reproduce and hard to structure. Artificial neural networks (ANNs) are good at learning knowledge from the high-dimensional space of high-quality data[2]. In contrast, biological neural networks

(BNNs) are good at learning knowledge from the low-dimensional space of low-quality data. As a result, neither ANNs nor BNNs alone are suitable for improving the efficiency of knowledge extraction from life science data.

The multi-faceted correspondence between ANNs and BNNs[5] suggests a novel possibility of combining their complementary advantages. We thus hypothesized that if the best practices of training ANNs were adapted to train a BNN, then the resultant BNN would be augmented with ANN's advantage of learning from high-dimensional data. In principle, an augmented BNN (ABNN) would be good at learning knowledge from the high-dimensional low-quality life science data and informing better designs of life science data labels (Figure 1).

Here we introduce the first ABNN in life science by training a BNN to learn like an ANN (Figure 2a and 2b). Inspired by the fact that larger ANNs almost always perform better[16,17], we hypothesized that the largest ABNN could only be attained by limiting team size to one to maximally increase the efficiency of knowledge extraction (Figure 2d). We adapted the best practices of training ANNs to train a single ABNN that learns from all-encompassing life science data in an uninterrupted pass without imposing interim milestones (Figure 2e). After eight years of training, the ABNN achieves exponential knowledge extraction from 2000+ life science publications and databases and builds a foundation model[6] of human physiology and 130+ specific models of human diseases (Figure 2c).

We assessed the quality of extracted knowledge in ABNN by measuring to what extent ABNN could increase the productivity in drug discovery. Inspired by the CASP challenge[18] that

serves to benchmark ANN performance for protein structure prediction, we designed the public prospective prediction of pivotal ongoing clinical trial outcomes at large scale (PROTOCOLS) challenge to benchmark ABNN performance for drug efficacy prediction (Figure 3a and Methods section 'Validation design').

The rationale of the PROTOCOLS challenge is that pivotal clinical trials (phase 2 and phase 3) are statistically powered to evaluate drug clinical efficacy preordained mainly by the underlying biological processes that can be distilled into scientific knowledge. Historical clinical success rates for investigational drugs are well-curated and duly updated[8,19] to provide an accurate baseline for evaluating ABNN performance in the PROTOCOLS challenge.

The PROTOCOLS challenge evaluated the real-world performance of ABNN in actual clinical settings with an emphasis on the generalizability of ABNN to novel drug-disease pairs even with scarce or uninformative prior clinical data (Figure 3b). Therefore, there would be no gap between the validation performance of ABNN in the PROTOCOLS challenge and the clinical utility of ABNN in real-world drug discovery[20,21].

2 RESULT

The PROTOCOLS validation consists of 265 then-ongoing pivotal clinical trials following a predefined set of screening criteria (Figure 3c and Methods section 'Validation data collection').

The sample of validation clinical trials and the sample of newly initiated clinical trials since

2020 are shown to follow identical distribution across all major therapeutic areas (two-sample Kolmogorov-Smirnov test, statistic = 0.333; P = 0.73; Figure 4). The PROTOCOLS validation could thus provide an unbiased evaluation of ABNN's disease-agnostic performance.

The minimum sample size requirement (131) and the minimum event size requirement (118) were determined for evaluating ABNN's performance with 99% confidence[7] (Figure 3c and Methods section 'Statistical analysis'). Both requirements are met as clinical readouts were available for 157 of 265 predictions by the cutoff date (Methods section 'Clinical readout'). Ground truth was determined for each readout and was then applied to validating the corresponding prediction (Methods section 'Prediction Validation), resulting in confusion matrices over different subgroups of the validation dataset (Figure 7).

It should be noted that positive instances (true positive and false negative) are much more important than negative instances (true negative and false positive) in the real-world scenarios of life science and drug discovery. True positives are more interesting than true negatives because the historical clinical success rate for investigational drugs entering human studies is only 7.9% for all diseases[8]. False positives will only have a marginal clinical impact because human clinical trials will still fail, bringing no clinical harm to patients and only generating the status quo level of sunk R&D costs for drug developers. In contrast, false negatives will have a substantial clinical impact because they deny clinical benefits to patients and eliminate sizable revenue to drug developers that dwarfs saved R&D costs.

Correspondingly, we chose sensitivity and F1 score as the most appropriate performance

measures for the PROTOCOLS validation[22]. Our choice of sensitivity realistically reflects the best practice in actual drug development: 100% of investigational new drugs greenlit to enter first-in-human studies are then believed to achieve clinical success, so the clinical success rate from phase 1 to approval is exactly a measure of sensitivity[19]. We also reported other standard measures (accuracy, specificity, etc.) to offer a balanced view of ABNN's performance (Data Table 2).

As ABNN was trained to functionally reconstruct human physiology and almost all major diseases that could possibly occur to human body (see Methods section 'Model training'), we first evaluate ABNN's disease-agnostic performance for the full validation dataset (157 predictions; Supplementary Table 1). ABNN achieved 90.1% prospective prediction sensitivity (99% CI 80.0%, 96.9%; $P < 0.001$; Figure 5a;), a nearly 11.4-fold improvement from the realistic baseline sensitivity of 7.9% and an almost 6.0-fold improvement from the conservative baseline sensitivity of 15.1% (Figure 5a; Methods section 'Baseline performance'). ABNN achieved 0.86 F1 score (99% CI 0.79, 0.92; $P < 0.0001$; Figure 6c; Data Table 1), 82.8% accuracy (99% CI 74.2%, 89.8%; $P < 0.0001$; Figure 6c; Data Table 1) and 72.7% specificity (99% CI 57.5%, 85.3%; $P < 0.0001$; Figure 6b; Data Table 1). Those data clearly demonstrate that ABNN has extracted large-scale high-quality knowledge to enable its superb disease-agnostic performance in the PROTOCOLS validation.

As ABNN does not require prior clinical data to predict investigational drugs' clinical efficacy (Figure 3a), we conduct a couple of further analyses for subgroups of the validation dataset with varying degrees of reduced availability of prior clinical data.

First-in-class clinical trials for novel drug-disease pairs represent the most challenging subset of the PROTOCOLS validation because no prior clinical data are available for those drug-disease pairs across all therapeutic areas. As there are 128 first-in-class clinical trials in the validation dataset, the minimum sample size and event size requirements for 99% confidence have been met. ABNN achieved 92.4% prospective prediction sensitivity (99% CI 80.8%, 99.2%; $P < 0.009$; Figure 5b, 6b; Data Table 1), 0.85 F1 score (99% CI 0.76, 0.92; $P < 0.0001$;Figure 6c; Data Table 1), 82.5% accuracy (99% CI 72.8%, 90.3%; $P < 0.0001$; Figure 6c; Data Table 1), and 71.7% specificity (99% CI 55.6%, 85.0%; $P < 0.0001$; Figure 6b; Data Table 1).

Notably, ABNN's disease-agnostic prospective prediction sensitivity for first-in-class clinical trials is shown to be even superior over that for all validation clinical trials ($\Delta = 2.3\%$; 99% CI -0.92%, 5.52%; $P < 0.002$ for superiority at a 2% margin; Figure 5b; Methods section 'Statistical analysis'). ABNN's prospective prediction accuracy is also non-inferior compared to that for all clinical trials ($\Delta = 0.3\%$; 99% CI -0.87%, 1.47%; $P < 0.0001$ for non-inferiority at a 2% margin; Methods section 'Statistical analysis'). Our results clearly demonstrate that ABNN's superb performance in the PROTOCOLS validation is independent of the availability of prior clinical data for any drug-disease pair.

Oncology clinical trials represent a different type of challenge for ABNN because they have the lowest historical clinical success rates: The realistic baseline sensitivity is 5.3% for oncology and 23.9% for hematology[8] (Methods section 'Baseline performance'). Consequentially, the importance of already preferred positive instances is even higher in oncology than in other therapeutic

areas, to the reasonable extent that negative instances are no longer interesting. Therefore, a clinically meaningful assessment of ABNN's performance in oncology should primarily focus on F1 score, accuracy, and sensitivity.

Moreover, oncology clinical trials feature the most complex trial designs with active controls and combo therapies as the standard design of pivotal clinical trials. Human cancer patients are further stratified by certain biomarkers[23], stages of cancer, and prior treatment history, all of which add to the complexity of the scientific knowledge required to make accurate predictions of drug clinical efficacy.

As there are 48 oncology clinical trials in the validation dataset, the minimum sample size and event size requirements for 90% confidence have been met (see Methods section 'Statistical analysis'). ABNN achieved 88.9% prospective prediction sensitivity (90% CI 75.7%, 94.3%; $P < 0.015$; Figure 5c, 6b; Data Table 1), a nearly 16.8-fold improvement from the realistic baseline sensitivity of 5.3% and an almost 8.2-fold improvement from the conservative baseline sensitivity of 10.8% for oncology clinical trials (Methods section 'Baseline performance'). Remarkably, ABNN achieved 0.85 F1 score (90% CI 0.77, 0.90; $P < 0.0004$; Figure 6c; Data Table 1) and 77.1% prospective prediction accuracy (90% CI 65.1%, 84.9%; $P < 0.0004$; Figure 6c; Data Table 1). Those data clearly demonstrate that ABNN has extracted sufficiently complex and intricately nuanced knowledge of the human body and cancer etiology to accurately predict cancer drug clinical efficacy for biomarker-differentiated heterogeneous patient populations.

COVID-19 represents the ultimate test for the quality of existing knowledge in ABNN be-

cause no knowledge can be extracted (due to zero prior data). The entire disease model for COVID-19 can only be built by generalizing non-COVID-19 knowledge to COVID-19 clinical symptoms. No clinical data were available worldwide when the preprint of ABNN's COVID-19 disease model was published on 20. March 2020[24].

As there are 49 infectious disease clinical trials in the validation dataset, the minimum sample size and event size requirements for 90% confidence have been met (see Methods section 'Statistical analysis'). 94% (46/49) are COVID-19 pivotal clinical trials that come with neither prior clinical data nor prior life science data (Supplementary Table 4).

ABNN achieved 88.9% prospective prediction sensitivity (90% CI 68.3%, 95.4%; $P < 0.05$; Figure 5c; Data Table 1) for infectious disease clinical trials in the validation dataset, a nearly 6.7-fold improvement from the realistic baseline sensitivity of 13.2% and an almost 3.9-fold improvement from the conservative baseline sensitivity of 22.8% for infectious disease clinical trials (Methods section 'Baseline performance'). ABNN achieved 0.78 F1 score (90% CI 0.65, 0.86; $P < 0.0012$; Figure 6c; Data Table 1) and 81.6% prospective prediction accuracy (90% CI 70.1%, 88.4%; $P < 0.002$; Figure 6c; Data Table 1). Those data clearly show that the quality of extracted knowledge in ABNN has reached a critical mass to enable the de novo generation of new knowledge for a new disease like COVID-19 while maintaining the same quality as that of the new knowledge extracted from prior life science data.

For all the other therapeutic areas in the PROTOCOLS validation (60 predictions in total; Supplementary Table 7), ABNN achieved 92.4% 7.8% prospective prediction sensitivity (Figure

5c, 6b; Data Table 2) and 0.91 0.10 F1 score (Figure 6d; Data Table 2). Although the minimum requirements for 90% confidence are not met for any of those therapeutic areas (Methods section 'Statistical analysis'), those data demonstrate that ABNN's performance is stable and consistent across all major therapeutic areas in the PROTOCOLS validation.

We also present Alzheimer's Disease (AD) data in the PROTOCOLS validation to demonstrate the robustness of ABNN's extracted knowledge. AD represents an extreme case because the historical clinical success rate of disease-modifying drugs is next to zero[25]. A single hit (true positive) becomes the most appropriate performance measure of ABNN for AD because both false positives and negative instances are equally uninteresting in real-world scenarios. AD thus represents a tough challenge for ABNN because there is a consensus that AD is too complex, the quantity of existing data is too low, and the quality of existing data is too poor[25]. What's worse, existing data are probably more misleading than informative.

ABNN achieved 100% accuracy and 100% precision with non-zero true positive and zero false positive (Supplementary Table 6) for three prospective predictions of AD pivotal clinical trials in the PROTOCOLS validation. ABNN's hit is masitinib, a first-in-class c-Kit inhibitor that met the efficacy primary endpoint in the pivotal phase 3 trial (NCT01872598). Our results demonstrate that ABNN has extracted large-scale, high-quality knowledge of AD from existing life science data.

Aging is an evolutionarily conserved phenomenon across all mammalian species[26] and is clinically recognized as the main risk factor of numerous diseases in multiple therapeutic areas[27].

To evaluate the quality of extracted knowledge about aging in ABNN, we set out to identify all age-related pivotal clinical trials in the PROTOCOLS validation (Methods section 'Validation data collection') and measured ABNN's performance (Figure 8).

As there are 119 clinical trials of age-related diseases in the validation dataset, the minimum sample size and event size requirements for 99% confidence have been met. ABNN achieved 91.0% prospective prediction sensitivity (99% CI 79.1%, 98.4%; $P < 0.005$; Data Table 1; Figure 9a), 0.85 F1 score (99% CI 0.76, 0.92; $P < 0.0001$; Data Table 1; Figure 6c), 81.5% prospective prediction accuracy (99% CI 71.3%, 89.7%; $P < 0.0001$; Data Table 1; Figure 6c), and 69.2% specificity (99% CI 51.8%, 83.9%; $P < 0.0001$; Data Table 1; Figure 6b).

Remarkably, ABNN showed non-inferior prospective prediction sensitivity for all age-related clinical trials to that for all first-in-class clinical trials ($\Delta = 1.4\%$; 99% CI -1.12%, 3.92%; $P < 0.003$ for superiority at a 2% margin; Methods section 'Statistical analysis'). ABNN's prospective predictive accuracy for age-related clinical trials is also shown to be non-inferior over its performance for first-in-class clinical trials ($\Delta = 1.0\%$; 99% CI -1.14%, 3.14%; $P < 0.0001$ for non-inferiority at a 2% margin; Methods section 'Statistical analysis').

ABNN's prospective predictive sensitivity for all age-related clinical trials is statistically significantly higher than that for the full validation dataset ($\Delta = 0.9\%$; 99% CI -1.13%, 2.93%; $P < 0.0001$ for superiority at a 2% margin; Figure 9a; Methods section 'Statistical analysis'). Our results clearly demonstrate that ABNN has extracted sufficient knowledge of aging that significantly contributed to its consistent performance across all the age-related indications in the PROTOCOLS

validation (Figure 9b).

We also investigated the relationship between the quality of extracted knowledge in ABNN and the 'hardness' of therapeutic areas by computing the correlation between F1 scores and realistic baseline sensitivities (or LOA: historical clinical success rates for investigational new drugs since phase 1. See Methods section 'baseline performance') for all validation subgroups (Figure 6d). F1 scores and LOA are weakly negatively correlated (Exact Pearson r = -0.260), indicating that the quality of ABNN's knowledge about human diseases is orthogonal to the quality of collective knowledge for the corresponding therapeutic areas. ABNN performance is better for the more difficult therapeutic areas (Figure 6d; Data Table 2).

Finally, we explored the relationship between the quality of extracted knowledge in ABNN and drug modality (Figure 10). ABNN achieved 91.5% prospective prediction sensitivity (90% CI 80.8%, 95.7%; $P < 0.015$; Figure 11a; Data Table 1) for biologics (68 biological trials; Methods section 'Drug classification') and 85.7% prospective prediction sensitivity (90% CI 71.9%, 92.2%; $P < 0.009$; Figure 11a; Data Table 1) small molecule drugs (77 NME trials; Methods section 'Drug classification'). ABNN's performance is superb across drug modalities, slightly superior for biologics to for small molecules ($\Delta = 5.8\%$; 90% CI 0.78%, 10.82%; $P < 0.0001$ for superiority at a 2% margin; Figure 11a; Methods section 'Statistical analysis'). We further computed the correlation between F1 scores and realistic baseline sensitivities (or LOA: historical clinical success rates for investigational new drugs since phase 1. See Methods section 'baseline performance') across drug modalities. F1 scores and LOA are strongly positively correlated (Exact Pearson r = 0.975;

Figure 11b), indicating that the quality of extract knowledge in ABNN may be a natural fit for drug modality with higher target specificity.

3 DISCUSSION

We present the brand-new possibility of deep-learning-augmented human intelligence to build a foundation model of human physiology that enables highly accurate data-free predictions of drug clinical efficacy for almost any human disease. This achievement is intractable for data-driven machine intelligence, whose demand for clinical data is so high and the efficiency of knowledge extraction is so low that the gap between promise and proof becomes nearly unbridgeable (Figure 1).

Humans have designed machine learning methods to train artificial neural networks that have successfully learned black-box models outperforming human experts in playing games[1] and predicting protein structures[2]. Our work shows that machine learning methods can be adapted to train augmented biological neural networks (ABNNs) to successfully learn quasi-white-box models[24] to predict drug clinical efficacy (Figure 2).

ABNN could thus safely deliver AI-surpassing clinical utility while averting the numerous challenges in privacy and explainability faced by medical AI systems[21]. ABNN can also inform better design of data labels for training better-performing ANNs with fewer data[13].

Furthermore, in contrast to domain experts who have privileged access to private data in ad-

dition to public data, ABNN was trained with public data only and started from an underprivileged position. Nonetheless, our ABNN has extracted orders-of-magnitude more knowledge from far less data than experts.

We designed and conducted the most rigorous validation of the quality of extracted knowledge in ABNN: the public, prospective prediction of pivotal ongoing clinical trial outcomes at large scale (PROTOCOLS) challenge (Figure 3 and Figure 4). At 99% confidence, ABNN successfully achieved > 90% prospective prediction sensitivity, >0.85 F1 score and > 82% prospective prediction accuracy across all therapeutic areas, regardless of the availability of prior clinical data (Figure 5 and Figure 6; Data Table 2).

As ABNN takes a drug's mechanism of action and a trial's design protocol as the only inputs for evaluating the clinical efficacy of a novel drug (Figure 3a), ABNN's performance in the PROTOCOLS validation confirms that large-scale, high-quality scientific knowledge alone is sufficient for accurately predicting drug clinical efficacy.

Given that both inputs for ABNN are available before first-in-human studies and positive instances matter much more than negative cases in actual drug discovery and development, ABNN's positive predictions of drug clinical efficacy alone could readily increase the clinical success rates of investigational new drugs from 7.9%[8] to more than 90% and bring a permanent reversal of Eroom's law[4].

As ABNN is also the world's first clinically validated model of human aging (Figure 8 and

Figure 9), applying ABNN to cellular rejuvenation programming could substantially speed up the discovery of preventive medicine for age-related diseases[11].

ABNN also leaves room for improvement as 17.2% (27/157) predictions are either false positives (18/157) or false negatives (9/157) in the PROTOCOLS validation. Further work will be planned to investigate why ABNN performs better for harder diseases (Figure 6d). The exponential efficiency of knowledge extraction enables ABNN to translate each false prediction into a substantial improvement of the corresponding disease model.

While our ABNN could readily reshape medicine, its bigger impact comes from the training method of ABNNs, which provides a well-beaten path to drastically increase the efficiency of knowledge extraction in every discipline that is characterized by large-scale, low-quality data space (e.g., synthetic biology, material science, and climate science). Our result provides the first positive evidence that the smallest team is necessary for delivering the most disruptive breakthrough, corroborating the previous discovery that small teams disrupt[12].

The advent of ABNN at the peak of global productivity decline utters in the new era of human-centric AI. The global adoption of training ABNNs will generate a paradigm shift in research culture and organizational structure to mass-produce scientific and technological breakthroughs. ABNNs will give rise to general purpose technologies that could potentially overcome the productivity J-curve and the productivity paradox recurrently observed in the past [28].

4 METHOD

TRAINING DATA COLLECTION

The training dataset comprises all publicly accessible sources of life science data (including peer-reviewed publications and curated databases) that can't be enumerated due to limited space.

MODEL TRAINING

See Figure 2. The ABNN was first trained to learn a foundation model of human physiology that was then adapted to learn specific models of human diseases. ABNN is principally disease-agnostic, and yet it leaves room for improvement in hematological disorders, rare neurodevelopmental disorders, and rare musculoskeletal disorders (collectively as Unfinished Disorders). Unfinished disorders are excluded from validation performance evaluation, yet they are included in validation data collection because the validation/invalidation of prospective predictions for those Unfinished Disorders can provide additional insights to accelerate the model building process for those disorders.

VALIDATION METHOD

See Figure 3a.

VALIDATION DESIGN

Validation is critical for an objective assessment of ABNN's general-purpose exponential knowledge extraction from life science data. The most straightforward way is to test to what extent ABNN could reverse Eroom's law, yet it would be unrealistically time-consuming and non-scalable to directly start new drug programs. Running virtual clinical trials is an effective alternative: Using drug mechanism of action and clinical trial design protocol as the only inputs, ABNN would predict new drug clinical efficacy using extracted knowledge only without access to patient data or third-party confidential information. Highly accurate virtual clinical trials could be run before first-in-human studies and prevent drug developers from starting pivotal real-world trials that are doomed to failure[29].

To meet the stringent criteria of the PROTOCOLS challenge as shown in Figure 3b, the most rigorous validation of ABNN would be to publish prospective predictions of pivotal clinical trial outcomes for all human diseases with tamper-proof timestamps and calculate prediction accuracy with a 99% confidence interval based on the ground truth determined by actual clinical readouts. There is no selection of lead over nonlead indications, although nonlead indications have far lower success rate[19].

There are a few hundreds of new pivotal clinical trials whose protocols are publicly accessible in a centralized database as required by FDA[30] and whose clinical readouts are usually published by sponsors in a standardized format. Public, prospective predictions of pivotal ongoing clinical trial outcomes could serve as a validation benchmark for ABNNs as much as ImageNet served as a benchmark validation in visual object recognition for ANNs[31].

BASELINE PERFORMANCE

Conservative Baseline Sensitivity (CBS) is defined as the probability of FDA approval for drugs in phase 2 development (called Phase2 Likelihood of Approval, or P2LOA in abbreviation). It is calculated as the corresponding P2LOAs in a 2011-2020 survey[8]. This baseline sensitivity is conservative because it does not take into account that ABNN only needs two inputs (drug mechanisms of action and clinical trial design protocols) to generate predictions, and both are readily available in the real world even before first-in-human studies (phase 1). Nevertheless, this baseline sensitivity is more balanced in consideration of both the phase distribution of clinical trials in the validation dataset (157 clinical readouts (3.18% (5/157) phase 1 trials, 43.9% (69/157) phase 2 trials, and 54.1% (85/157) phase 3 trials, and the irrelevance of phase transitions for ABNN to make predictions.

Realistic Baseline Sensitivity (RBS) is defined as the probability of FDA approval for drugs in phase 1 development (called Phase1 Likelihood of Approval, or P1LOA in abbreviation). It is calculated as the corresponding P1LOAs in a 2011-2020 survey[8]. This baseline sensitivity is realistic as it reflects that ABNN would make identical predictions at earlier stages of development with drug mechanism of action and clinical trial design protocol as the only inputs available as early as preclinical stages.

RBA is not directly available from previous survey studies[8,19] for two subgroups of the validation datasets: all first-in-class indications (128 trials) and all age-related indications (119 trials). As diseases in those two subgroups come from therapeutic areas featuring lower-than-average his-

torical clinical success rates (oncology, neurology, cardiovascular, etc.), the real RBAs for both subgroups should be lower than the RBA for all indications (7.9%[8]). We thus assumed that a reasonable estimate of RBA for the subgroup of all first-in-class indications (128 trials) would be 6.32%, equal to 80% of the RBA for all indications (7.9%[8]). Adding that no anti-aging therapeutics have ever been clinically approved, we assumed that a reasonable estimate of RBA for the subgroup of all age-related indications (119 trials) would be 5.69%, equal to 72% of the RBA for all indications (7.9%[8]).

VALIDATION DATA COLLECTION

The validation dataset comprises the registration information of human clinical trials on www.clinicaltrials.gov with the screening criteria as shown in Figure 3c (phase 1/2 is categorized as phase 1, and phase 2/3 is categorized as phase 2). We estimate that there are approximately 200-300 clinical trials per year that would match the screening criteria, which is consistent with 424 clinical trials per year with a much looser set of screening criteria (no limit on primary completion date, no preference of first-in-class trials, etc.) in a comprehensive survey study[19]. Please see the Supplementary Information for the full validation data.

A clinical trial in the validation dataset is designated as "first-in-class" if the drug-indication pair hasn't been approved for marketing.

A clinical trial in the validation dataset is designated as "age-related" if age is a clinically identified risk factor for the indication thereof.

PREDICTION GENERATION

Per the industry-wise best practice[32], We predict SUCCESS for a pivotal clinical trial if at least one of its prespecified efficacy primary endpoint(s) is believed to be met with predefined statistical significance. ABNN predicts FAILURE for a pivotal clinical trial if none of its prespecified efficacy primary endpoint(s) is believed to be met with predefined statistical significance (Figure 3d).

If a pivotal clinical trial has multiple co-primary endpoints, we could predict PARTIAL SUCCESS, PARTIAL FAILURE if at least one of its prespecified efficacy primary endpoint(s) is believed to be met with predefined statistical significance and at least one of its prespecified efficacy primary endpoint(s) is believed to be not met with predefined statistical significance.

PREDICTION PUBLICATION

265 predictions were published as timestamped tweets @DemiugeTech to establish a publicly accessible track record of prospective predictions. The immutability of tweets ensures that each published prediction cannot be modified afterward. We also indexed each tweet to prevent any intentional deletions of false predictions that go unnoticed to ensure a reliable performance evaluation. Nevertheless, index gaps do exist for several legitimate cases that do not compromise the rigor of validation: a) an index was unnoticedly skipped and remained unused from then on; b) an indexed prediction was made but later became disqualified as a prospective prediction, because the corresponding result had been announced after the indexed prediction was made but before

the indexed prediction was published. To distinguish those two cases of index gaps, we used DocuSign to electronically sign every indexed prediction before publishing it as a tweet. Hence, no DocuSign certificates are available for unused indices. 001 and 286 are the indices of the first and the last published prospective predictions, respectively. 21 index gaps have been detected, 5 of which were unused, 5 of which were for disqualified predictions that don't have clinical readouts as of the cutoff date (March 1, 2022), 9 of which were for disqualified predictions that clinical readouts as of the cutoff date with 66.7% accuracy (3 false predictions and 6 true predictions), and 2 of which were for qualified unpublished predictions whose results are still not available. DocuSign certificates for disqualified and qualified unpublished predictions are available upon written request (Figure 3d).

CLINICAL READOUT

It is customary for private companies and mandatory for public companies in the biopharmaceutical industry to forecast the estimated date of clinical readouts and to publish the actual data of clinical readouts for human clinical trials. We track the availability of clinical readouts for every published prospective prediction from multiple online sources, including https://www.globenewswire.com/ for press releases, https://www.biospace.com/ for curated news, and financial reports of public companies. We usually published the link to an available clinical readout as a timestamped tweet by replying to the original timestamped tweet of the corresponding prediction, thus illustrating the clearly established timeline of predictions followed by readouts (Figure 3d).

GROUND-TRUTH DETERMINATION

Under the FDA guidelines[33], a clinical trial with available readout is given the ground-truth label success if and only if at least one efficacy primary endpoint is met according to interim/topline analyses or final data published by the trial sponsors (Figure 3d).

Similarly, a clinical trial with available readout is given the ground-truth label failure if and only if (1) none of the efficacy primary endpoint(s) is met according to topline analyses or final data or deemed likely to be met according to interim analyses, or (2) the trial is terminated for non-recruitment or non-business reasons, (3) the post-trial development is discontinued for non-business or unspecified reasons, according to the official announcements by the trial sponsors.

Given that the reliability of ground-truth determination is capped by the 85.1% theoretical maximal prediction accuracy of drug clinical efficacy by human clinical trials[34], we should refrain from directly labeling missed predictions as false without further consideration of subsequent developments. In particular, we should pay attention to missed failure predictions for phase 2b trials that are less statistically powered than phase 3 trials. Missed failure predictions should be more likely to be true rather than false if the trial sponsor decides not to conduct a phase 3 trial for non-business or unspecified reasons.

PREDICTION VALIDATION

Compared to the corresponding ground-truth label, a prospective prediction of success is validated as TRUE if its corresponding ground-truth label is also success or invalidated as FALSE if its corresponding ground-truth label is failure. On the other hand, a prospective prediction of

failure is validated as TRUE if its corresponding ground-truth label is also failure or invalidated as FALSE if its corresponding ground-truth label is success. The confusion matrices of validation performance across on a disease-agnostic basis and on a per-therapeutic-area basis are shown in Figure 7.

DRUG CLASSIFICATION

Drugs in the PROTOCOLS validation were generally classified by the FDA classification codes for new drug applications: new molecular entity (NME), biologic, and non-NME that compromises vaccines and other modalities.

In particular, NME drugs are novel small molecule drugs or novel combinations of multiple old or new small molecule drugs. Biologics comprise a broad range of drug types such as antibodies, peptides, RNAi therapies, cell therapies, gene therapies, microbiome therapies. non-NMEs comprise vaccines, cytokines, enzymes, insulins and other reformulation of approved drugs.

STATISTICAL ANALYSIS

To evaluate the standalone performance of ABNN, we designed the public, prospective prediction of pivotal ongoing clinical trial outcomes at large scale (PROTOCOLS) challenge as the most rigorous external validation of a clinical prediction model with a binary outcome[7].

We make some reasonable assumptions to calculate minimal sample size and minimal event size to detect prediction accuracy, sensitivity, specificity, and F1 score at 99% confidence, using

the previously reported method[7]:

The anticipated clinical readout event proportion (φ) is set as 0.9 because as of the cutoff date an average of 75% of clinical trials on clinicaltrial.gov report clinical trial results in compliance with the FDAAA 801 and the Final Rule tracked by https://fdaaa.trialstracker.net/, and more than 90% of big-pharma-sponsored clinical trial report results[35].

The ratio of total observed clinical readout events and total expected clinical readout events (O/E) is set as 1 with a width of 0.15 to account for about 12% of clinical trials that are prematurely terminated without clinical trial results[36].

Accordingly, the minimal sample size requirement is 131 predictions, and the minimal event size requirement is 118 readouts for 99% confidence. From 11. Feb 2020 to 22. Dec 2020, 265 prospective predictions have been published to meet the minimal sample size requirement. By the study cutoff date (March 1, 2022), 157 clinical readouts (5 phase 1 trials, 68 phase 2 trials, and 85 phase 3 trials) were available (Supplementary Table 1), so the minimal event size requirement has been met.

All other factors held constant, for subgroups of the validation datasets with smaller event sizes, performance measures are detected at 90% confidence with the minimal sample size requirement of 52 predictions the minimal event size requirement of 47 readouts. For subgroups of the validation datasets with even smaller even sizes, only simple means and standard deviations are calculated over the corresponding proportions in those subgroups.

To evaluate the direct translatability between prospective validation and clinical application of ABNN, a two-sample Kolmogorov-Smirnov test was used as the standard method to detect if the sample of validation clinical trials and the sample of initiated clinical trials since 2020 follow identical distribution (see Figure 4). Oncology in the PROTOCOLS validation has the same proportion as in benchmark[19].

To ensure sufficient statistical significance in real-world scenarios of actual clinical settings, confidence intervals on proportions are Agresti-Coull Intervals[37]. Confidence intervals on the difference are Wald Intervals[37]. Two-sided p-values[38] are calculated for every proportion and difference. A 2% absolute margin is chosen for non-inferiority and superiority comparisons. Our statistical significance threshold is 0.01.

We used the functions of python SciPy and NumPy libraries.

5 DATA AVAILABILITY

The full dataset of the PROTOCOLS validation is publicly available at https://twitter.com/demiurgetech.

6 ACKNOWLEDGEMENT

TBA.

7 AUTHOR CONTRIBUTIONS

B.L. and I.L. contributed to the study conception. B.L. and I.L. contributed to the study design. B.L. and I.L. contributed to data collection and interpretation. B.L. performed the statistical analysis. B.L. and I.L. contributed to the interpretation of the results and writing of the manuscript. All authors read and approved the final manuscript.

8 COMPETING INTERESTS

This study was funded by Demiurge Technologies AG ('Demiurge'). B.L. and I.L. are employees, board members, and shareholders of Demiurge.

9 SUPPLEMENTAL INFORMATION

The full information of the PROTOCOLS validation is available in Excel spreadsheets.

References

References

1. Silver, D. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).

2. Tunyasuvunakool, K. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

3. Nurse, P. Biology must generate ideas as well as data. *Nature* **597**, 305–305 (2021).

4. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov* **11**, 191–200 (2012).

5. Richards, B. A. A deep learning framework for neuroscience. *Nat. Neurosci* **22**, 1761–1770 (2019).

6. Bommasani, R. On the opportunities and risks of foundation models. Preprint at https://arxiv.org/abs/2108.07258 (2021).

7. Riley, R. D. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med* **40**, 4230–4251 (2021).

8. BIO, Informa, QLS. Clinical Development Success Rates and Contributing Factors 2011–2020. URL: https://www.bio.org/clinical-development-success-rates-and-contributing-factors-2011-2020. (2021) (Accessed: 28th February 2022)

9. High-Accuracy MoA-based Clinical Efficacy Prediction Validation. Demiurge Technologies AG (2022). URL: https://www.demiurge.technology/validation. (Accessed: 25th February 2022)

10. Lovetrue, B. The AI-discovered aetiology of COVID-19 and rationale of the irinotecan+ etoposide combination therapy for critically ill COVID-19 patients. *Med. Hypotheses* **144**, 110180–110180 (2020).

11. Browder, K. C. In vivo partial reprogramming alters age-associated molecular changes during physiological aging in mice. *Nat. Aging* (2022).

12. Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378–382 (2019).

13. Jiang, L. *et al.* Learning data-driven curriculum for very deep neural networks on corrupted labels. *International Conference on Machine Learning* 2304–2313 (2018).

14. Mullard, A. Cancer reproducibility project yields first results. *Nat. Rev. Drug Discov* **16**, 77–77 (2017).

15. Perrin, S. Preclinical research: Make mouse studies work. *Nature* **507**, 423–425 (2014).

16. Brown, T. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst* **33**, 1877–1901 (2020).

17. Kaplan, J. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).

18. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins Struct. Funct. Bioinforma* **89**, 1607–1617 (2021).

19. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol* **32**, 40–51 (2014).

20. Wu, E. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med* **27**, 582–584 (2021).

21. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med* **28**, 31–38 (2022).

22. Chicco, D., Tötsch, N. & Jurman, G. The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min* **14**, 1–22 (2021).

23. Liu, R. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* **592**, 629–633 (2021).

24. Lovetrue, B. The AI-Discovered Aetiology of COVID-19 and Rationale of the Irinotecan+Etoposide Combination Therapy for Critically Ill COVID-19 Patients. Preprint at https://www.preprints.org/manuscript/202003.0341/v1 (2020).

25. E, T, G., T, Douglas, D. S. & G. Alzheimer's disease: The right drug, the right time. *Science* **362**, 1250–1251 (2018).

26. Fei, Z., Raj, K., Horvath, S. & Lu, A. *Universal DNA Methylation Age Across Mammalian Tissues. Innov. Aging* **5**, 412–412 (2021).

27. Niccoli, T. & Partridge, L. Ageing as a Risk Factor for Disease. *Curr. Biol* **22**, 741–752 (2012).

28. Brynjolfsson, E., Rock, D. & Syverson, C. The productivity J-curve: How intangibles complement general purpose technologies. *Am. Econ. J. Macroecon* **13**, 333–372 (2021).

29. Woo, M. An AI boost for clinical trials. *Nature* **573**, 100–100 (2019).

30. Medicine, N. L., Of, Clinicaltrials & Gov. *National Library of Medicine (US) Available* 27–27 (2022).

31. Russakovsky, O. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis* **115**, 211–252 (2015).

32. Mcleod, C. Choosing primary endpoints for clinical trials of health care interventions. *Contemp. Clin. Trials Commun* **16**, 100486–100486 (2019).

33. U.S.Food and Drug Administration, (2005). URL http//www.fda.gov/RegulatoryInformation/Guidances/ucm127069.htm.

34. Burt, T., Button, K. S., Thom, H. H. Z., Noveck, R. J. & Munafò, M. R. The Burden of the "False-Negatives" in Clinical Development: Analyses of Current and Alternative Scenarios and Corrective Measures. *Clin. Transl. Sci* **10**, 470–479(2017).

35. Devito, N. J., Bacon, S. & Goldacre, B. Compliance with legal requirement to report clinical

trial results on ClinicalTrials. gov: a cohort study. *Lancet* **395**, 361–369 (2020).

36. Williams, R. J., Tse, T., Dipiazza, K. & Zarin, D. A. Terminated trials in the ClinicalTri- als. gov results database: evaluation of availability of primary outcome data and reasons for termination. *PLoS One* **10**, 127242–127242 (2015).

37. Fagerland, M. W., Lydersen, S. & Laake, P. Recommended tests and confidence intervals for paired binomial proportions. *Stat. Med* **33**, 2850–2875 (2014).

38. Altman, D. G. & Bland, J. M. How to obtain the P value from a confidence interval. *BMJ* **343**, 2304–2304 (2011).

Figure 1. **The challenge and solution of extracting knowledge from life science data**

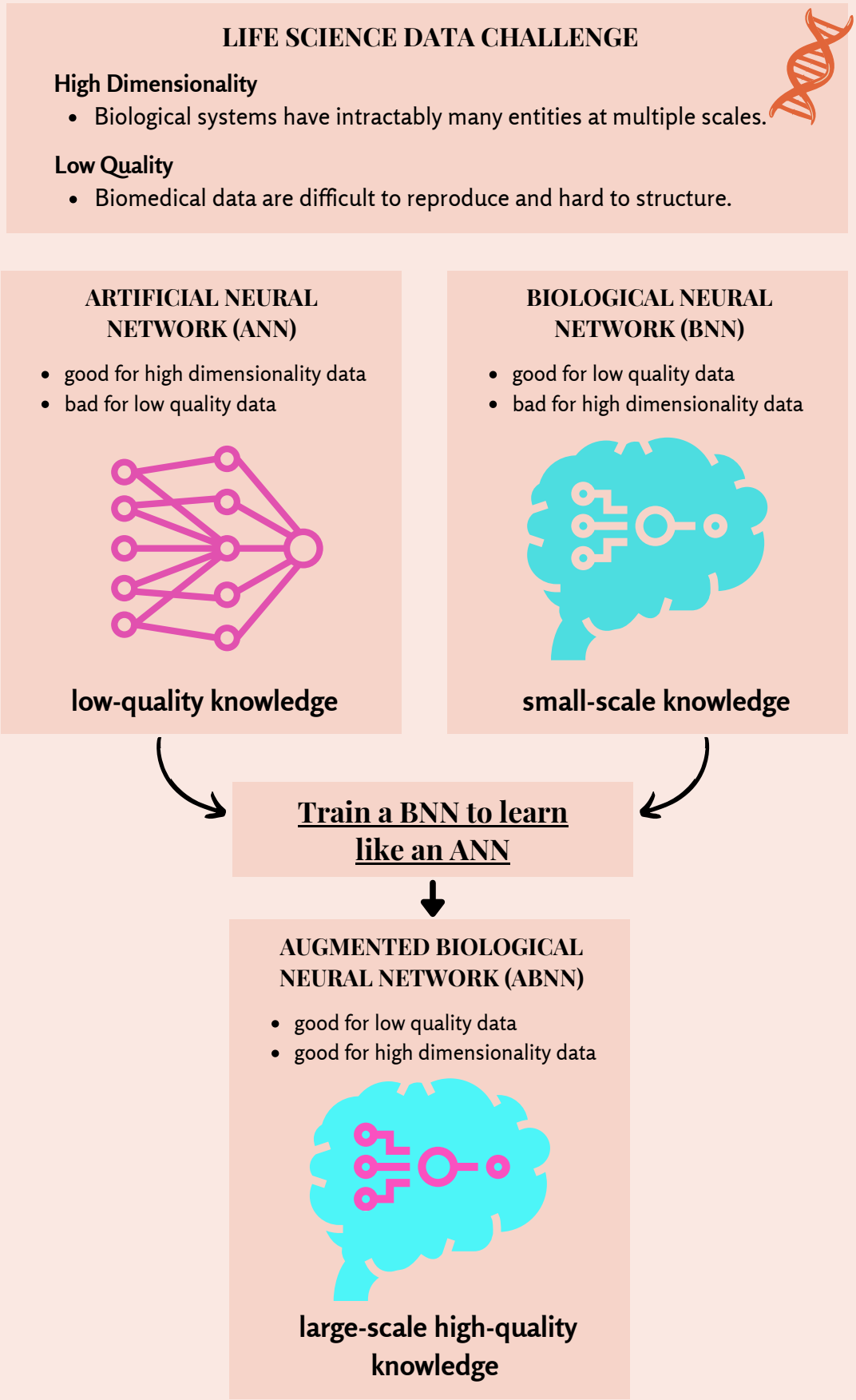


Figure 1. The challenge and solution of extracting knowledge from life science data. High dimensionality and low quality are two defining characteristics of life science data. Artificial neural networks (ANNs) are good at learning knowledge from the high-dimensional space of high-quality data, whereas biological neural networks (BNNs) are good at learning knowledge from the low-dimensional space of low-quality data. An augmented biological neural network (ABNN) combines the complementary advantages of ANNs and BNNs would become good at learning knowledge from the high-dimensional low-quality life science data.

Figure 2. Train a BNN to learn like an ANN

a Learn from Multiple Levels of Abstraction

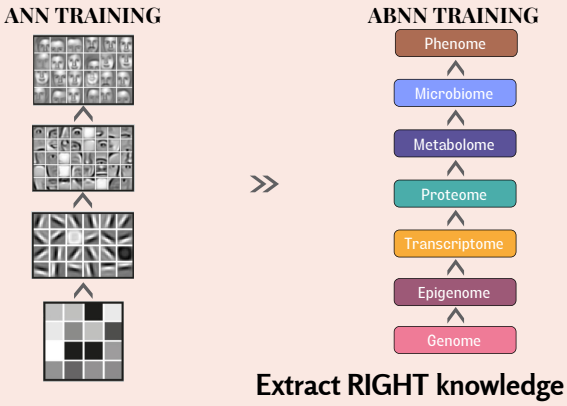


Figure 2a left. ANNs learn better representations of data with more levels of abstraction than with fewer levels and BNNs share the hierarchical architecture of ANNs. **Figure 2a right.** ABNNs extract right knowledge by integrating more multiomics data (e.g. genome, proteome, transcriptome, epigenome, metabolome and microbiome).

b Learn via the Backpropagation Algorithm

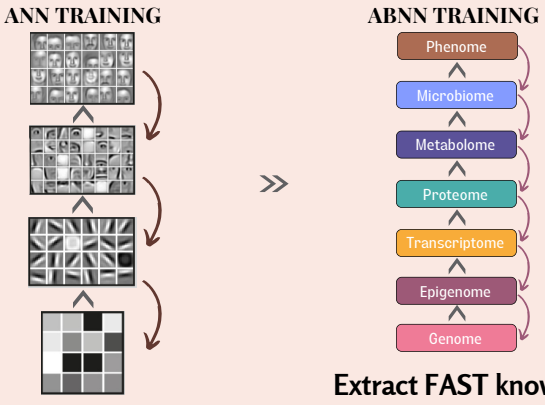


Figure 2b left. ANNs use backpropagation to improve learned representations of data by sequentially updating its internal parameters from the highest to the lowest level of abstraction.[2] The ubiquitous feedback connections in BNNs may implement backpropagation-like learning rules. **Figure 2b right.** ABNNs extract fast knowledge by iteratively updating internal representations in the strict order from the most macroscopic level (phenotype) through the intermediate level (endophenotype) to the most microscopic level (genotype).

c Learn for a Foundation Model

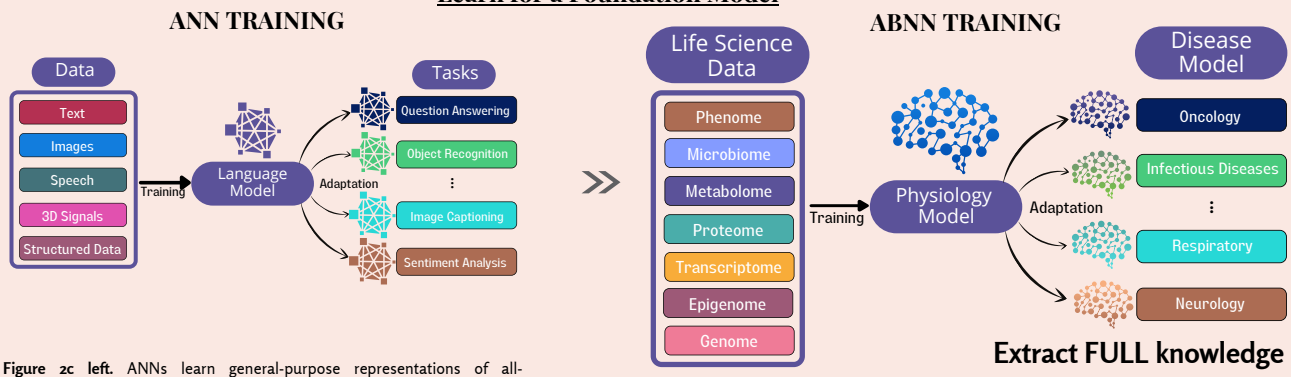


Figure 2c left. ANNs learn general-purpose representations of all-encompassing data to deliver maximal performance in a wide range of special-purpose tasks. Both BNNs and ANNs could learn to command general linguistic abilities for numerous tasks.

Figure 2c right. ABNNs extract full knowledge by first learning a general-purpose model of human physiology from the all-encompassing data, before starting to learn numerous human diseases models from disease-specific data.

d Learn in the Biggest Network

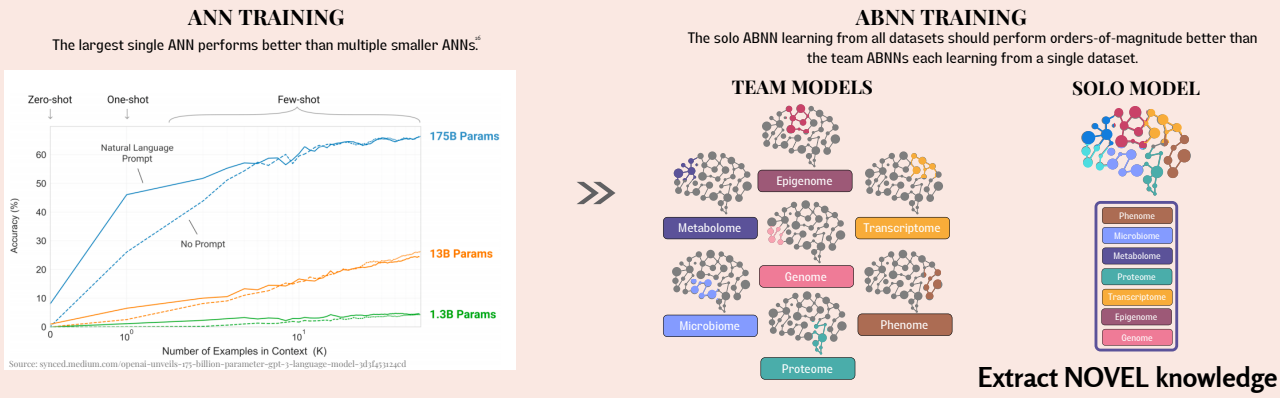


Figure 2d left. ANNs maximize the robustness of learned representations of data by increasing the sheer size of network parameters.

Figure 2d right. ABNNs extract novel knowledge if a single ABNN of the largest size learns from the entirety of the all-encompassing data, rather than multiple ABNNs of smaller sizes learn from siloed sub-datasets.

e Learn with an Uninterrupted Pass

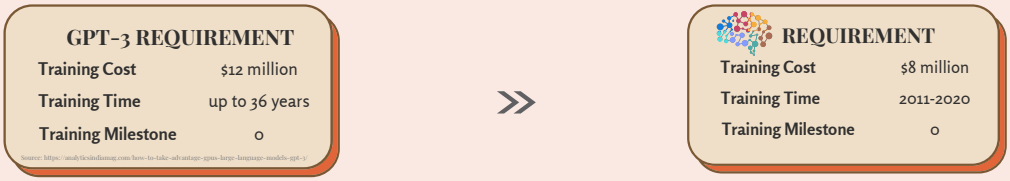


Figure 2e left. ANNs learn best representations of data given sufficient time and ample resources to allow for an uninterrupted pass of the full training data before which no meaningful performance milestones can be defined.

Figure 2e right. ABNNs extract testable knowledge if a milestone-free supply of time and fund is guaranteed to ensure an uninterrupted pass of learning so that ABNNs won't be evaluated at all before the completion of learning both human physiology and human diseases.

Figure 3. A large-scale, public prospective external validation of the trained ABNN

a Validation Method

DRUG CLINICAL EFFICACY PREDICTION

- Predict new drug clinical efficacy using extracted knowledge only.
- Access to patient data or confidential information is not required.

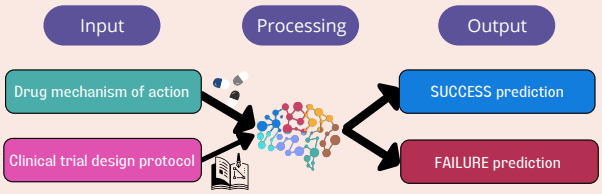


Figure 3a. Validation is critical for an objective assessment of ABNN's potential for improving the efficiency of global drug research and development. As the foundation model is supposed to be able to functionally reconstruct almost all major diseases that could occur to a single human patient, drug clinical efficacy prediction is the most effective validation method for ABNN to evaluate the clinical efficacy of novel drugs using mechanism of action data only and to prevent drug developers from starting pivotal real-world trials that are doomed to failure.[21]

b Validation Design

Prospective, not retrospective

- good for evaluating real-world performance in actual clinical settings.

Past (Retrospective) — Present — Future (Prospective) ✓

No gap between research and application

External, not internal

- good for evaluating the model's generalizability to novel clinical settings.

One dataset (Internal): Training | Validation VS. Separate Datasets (External): Training | Validation ✓

No gap between mature and novel applications

Large-scale, not small-scale

- good for reliably estimating the precise performance.

99% Confidence Interval ✓

Public, not private

- good for unbiased rigorous validation.

100% Twitter-based Validation ✓

Figure 3b. Prospective validation is much more rigorous than retrospective validation for examining the real-world performance of AI-based systems in actual clinical practices.[21] Large-scale validation is much more rigorous than small-scale validation for enabling an unbiased estimation of the actual predictive power of AI-based systems. External validation using separate datasets independent of the training datasets is much more rigorous than internal validation using held-out portions of the training datasets to evaluate the performance generalizability of AI-based systems to targeted clinical applications. Public validation is much more rigorous than private validation for ruling out all the possibilities of contaminating the performance estimates.

c Validation Dataset

ONGOING PIVOTAL CLINICAL TRIALS

TRIAL SCREENING CRITERIA
https://www.clinicaltrials.gov/

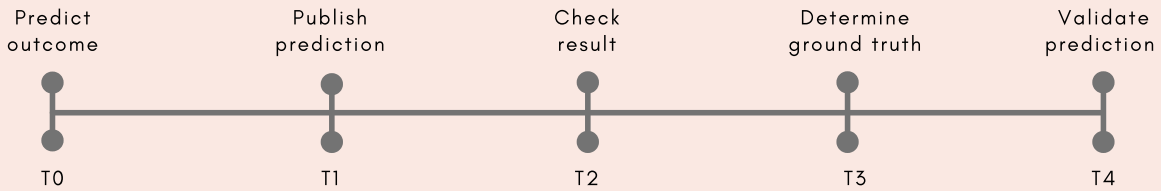
Phase:	Phase 2 and Phase 3
Study Type:	Interventional
Study Results:	No results yet
Study Disease:	Any
Study Status:	Recruiting / Active, not recruiting
Study Country:	United States, China, Japan, Germany, UK, France
Sponsor Type:	Industry
Primary Completion Date:	1.January 2020 - 31. December 2022
Additional Criteria:	first-in-class preferred (lead or non-lead)

»

- **Total Clinical Trials: 265**
- **Total Therapeutic Areas: 18**
- **Minimal Trial Sample Size: 131**
- **Minimal Readout Event Size: 118**

Figure 3c. The list of screening criteria used to identify pivotal clinical trials for ABNN validation. To evaluate the stand-alone performance of the foundation model, we designed public prospective predictions of pivotal ongoing clinical trial outcomes at large scale (PROTOCOLS) as the most rigorous external validation of a clinical prediction model with a binary outcome. To detect prediction accuracy, sensitivity, and specificity at 99% confidence, 265 prospective predictions were published to meet the minimal sample size requirement of 131 and 158 readouts were collected to meet the minimal event size requirement of 118 (see Methods section 'Statistical analysis').

d Validation Workflow



Success or Failure
- whether the primary endpoint of a clinical trial is predicted to be met (success) or not met (failure).

Public Tweet
- Every prospective prediction is published as an indexed tweet that is both timestamped and immutable @DemiurgeTech.

Clinical Readout
- Clinical readouts are usually published as press releases, financial reports, research papers, etc. by trial sponsors.

Ground Truth
- The ground truth of whether a clinical trial is a success or a failure can be determined directly from clinical readouts.

Validated Prediction
- The prediction is validated (invalidated) as true (false) if it is identical (opposite) to the ground-truth label of the trial.

Figure 3d. The workflow of every validation step from start to end (see Methods section for details). To meet the stringent criteria of the PROTOCOLS validation, we used Twitter as the best venue to conduct the PROTOCOLS validation of ABNN with immutable timestamped tweets to build a publicly verifiable track record on the publicly accessible account of @DemiurgeTech. ABNN made prospective predictions of pivotal clinical trial outcomes for all human diseases (except for neurodevelopmental and skeletomuscular rare disorders and hematological disorders). There is no selection of lead over nonlead indications, though nonlead indications have far lower success rate.[19] There are a few hundreds of new pivotal clinical trials whose protocols are publicly accessible in a centralized database as required by FDA[33] and whose clinical readouts are usually published by sponsors in a standardized format. We calculated prediction accuracy with 99% confidence interval based on the ground truth determined by actual clinical readouts.

Figure 4. Proportions of Clinical Trials Per Key Therapeutic Area

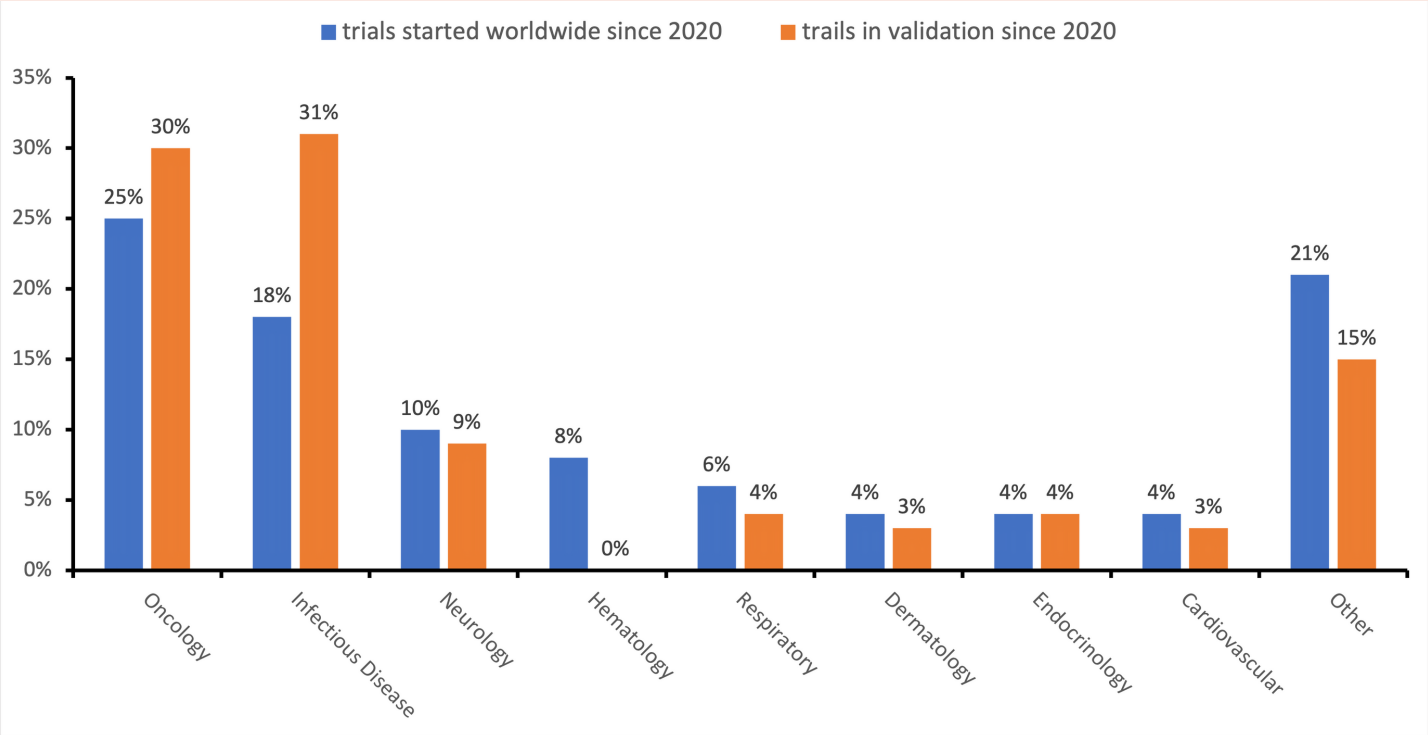


Figure 4. To evaluate the direct translatability between prospective validation and clinical application, a two-sample Kolmogorov-Smirnov test was used as the standard method to show that the sample of validation clinical trials and the sample of initiated clinical trials since 2020 follow identical distribution (statistic = 0.333 p-value = 0.73).

Figure 5a. Prospective Prediction Sensitivity of Clinical Trials for All Therapeutic Areas

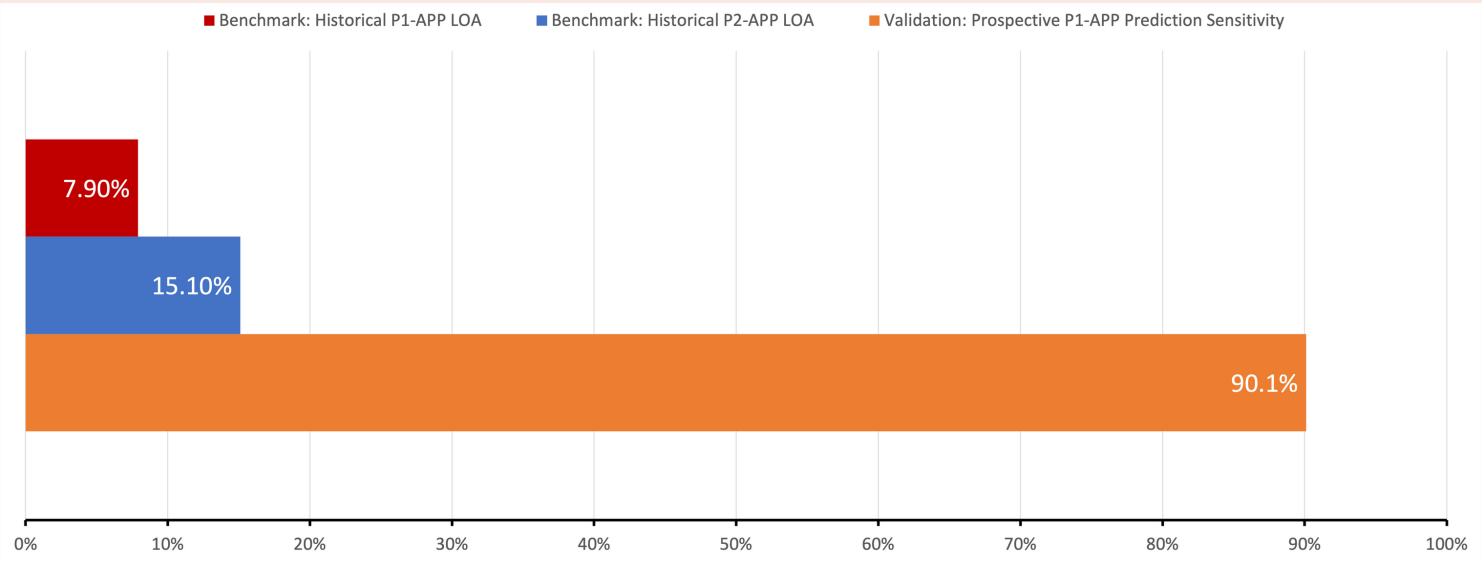


Figure 5a. The PROTOCOLS validation uses a realistic baseline sensitivity of 7.90% (Historical P1-APP LOA: historical likelihood of success from phase 1 to approval) and a conservative baseline sensitivity of 15.1% (Historical P2-APP LOA: historical likelihood of success from phase 2 to approval) to benchmark ABNN's performance (see Methods section 'Baseline performance' for details). ABNN's overall prospective prediction sensitivity of pivotal clinical trials in the PROTOCOLS validation is 90.1% (99% CI 80.0%, 96.9%; $P < 0.001$) .

Figure 5b. Prospective Prediction Sensitivity of Clinical Trials for All Trials and First-in-class Trials

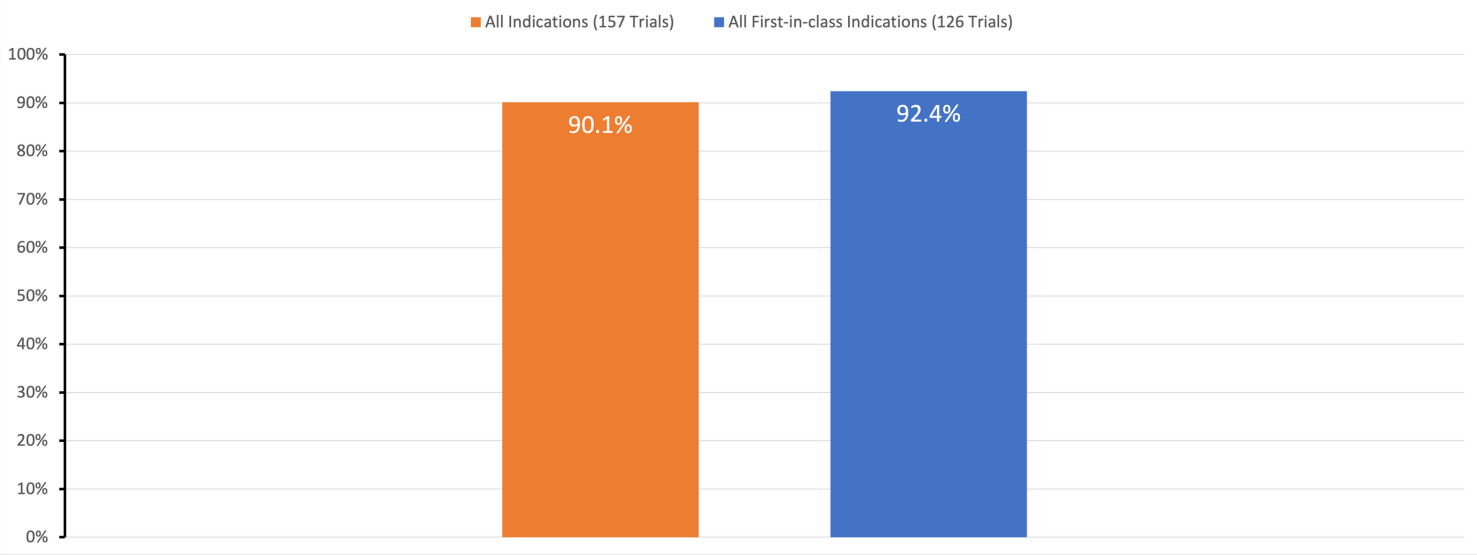


Figure 5b. ABNN's overall prospective prediction accuracy of first-in-class pivotal clinical trials is superior compared with that of all pivotal clinical trials (Δ = 2.3%; 99% CI -0.92%, 5.52%; P < 0.002; 2% margin for non-inferiority) for all therapeutic areas excluding neurodevelopmental, skeletalmusclar rare disorders, and hematological disorders (see Methods section 'Model training' for details).

Figure 5c. Prospective Prediction Sensitivity of Clinical Trials Per Therapeutic Area

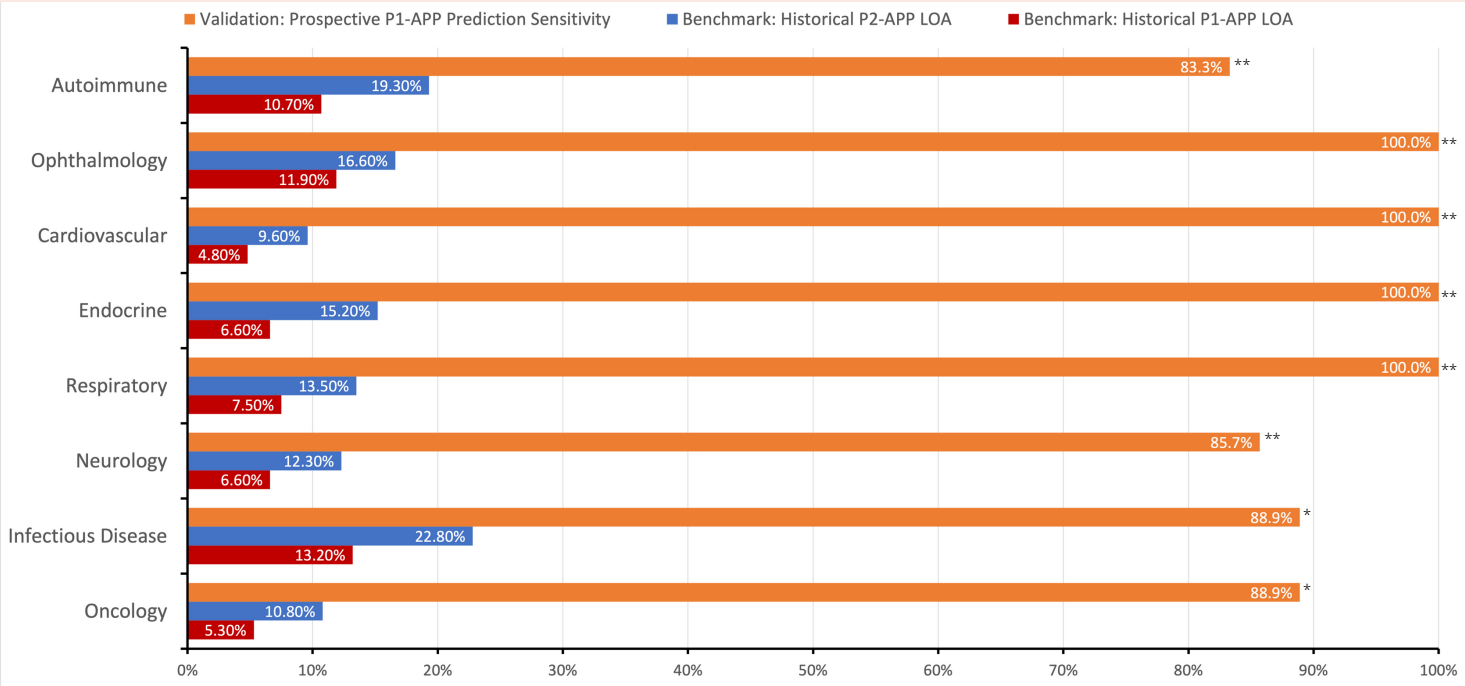


Figure 5c. The PROTOCOLS validation uses a realistic baseline sensitivity (Historical P1-APP LOA: historical likelihood of success from phase 1 to approval) and a conservative baseline sensitivity (Historical P2-APP LOA: historical likelihood of success from phase 2 to approval) to benchmark ABNN's performance (see Methods section 'Baseline performance' for details). Oncology: 77.1% (90% CI 65.1%, 84.9%; P < 0.0004); Infectious Disease: 81.6% (90% CI 70.1%, 88.4%; P < 0.002); *The minimum sample size and event size requirements for 90% confidence have been met (see Methods section 'Statistical analysis'). **The minimum sample size and event size requirements for 90% confidence have not been met (see Methods section 'Statistical analysis').

Figure 6a. Performance of ABNN: Precision and Recall

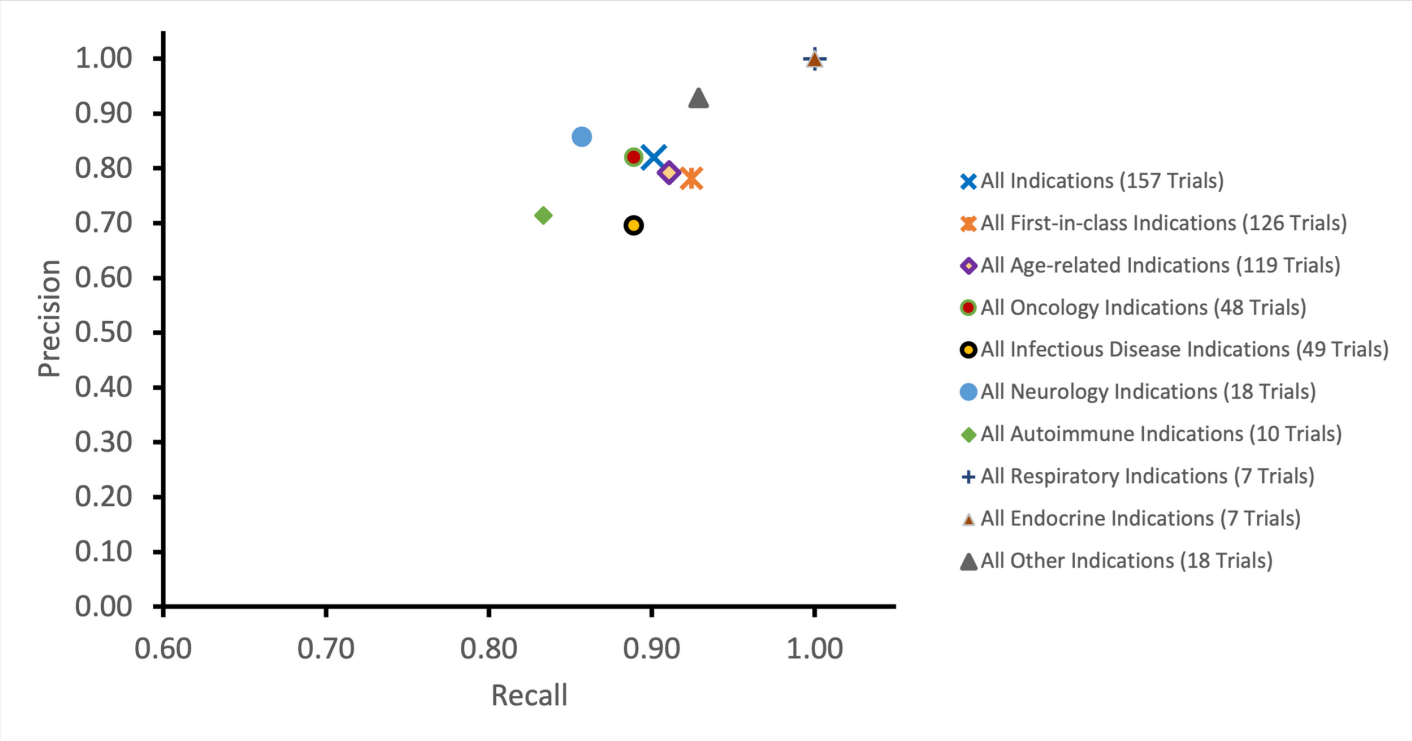


Figure 6a. ABNN's performance in terms of precision and recall (see Data Table 1 for more details).

Figure 6b. Performance of ABNN: Sensitivity and 1 - Specificity

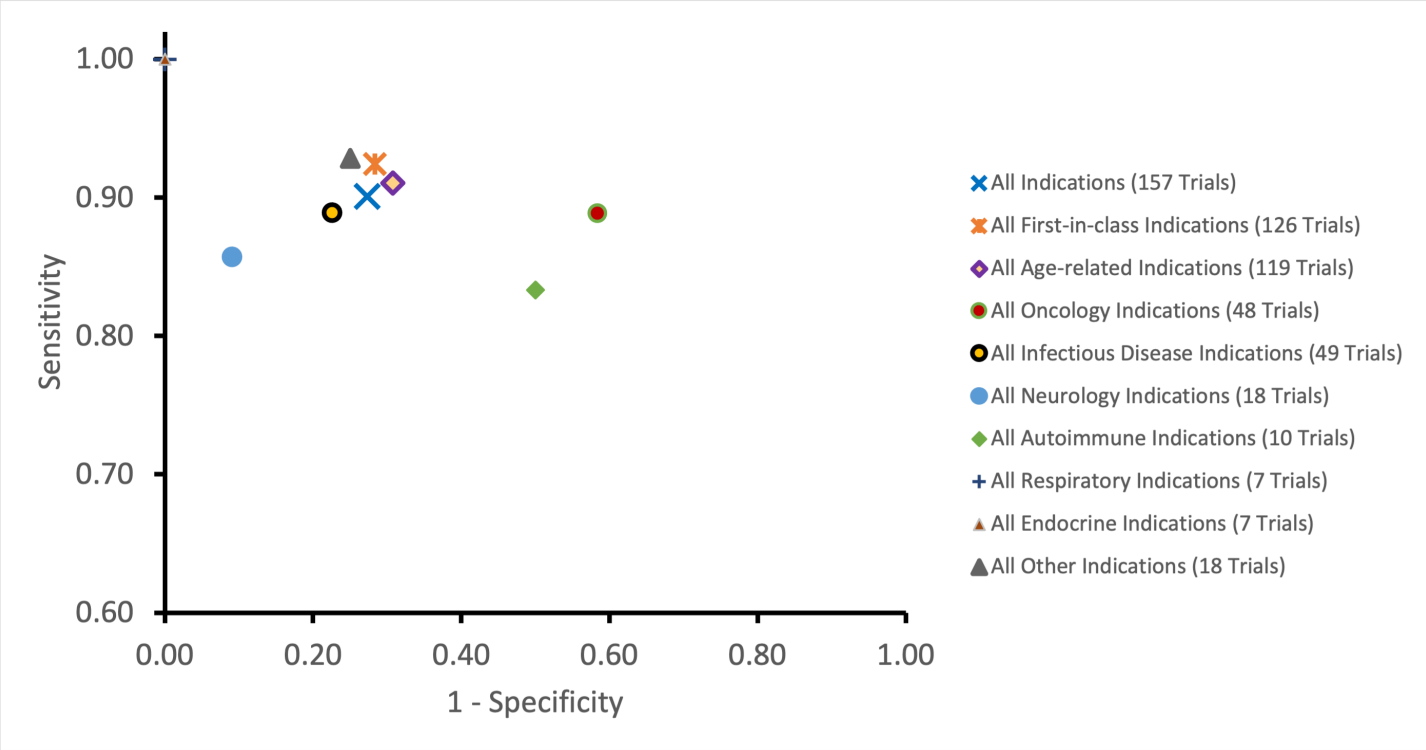


Figure 6b. ABNN's performance in terms of sensitivity and 1 - specificity (see Data Table 1 for more details).

Figure 6c. Performance of ABNN: Accuracy and F1 Score

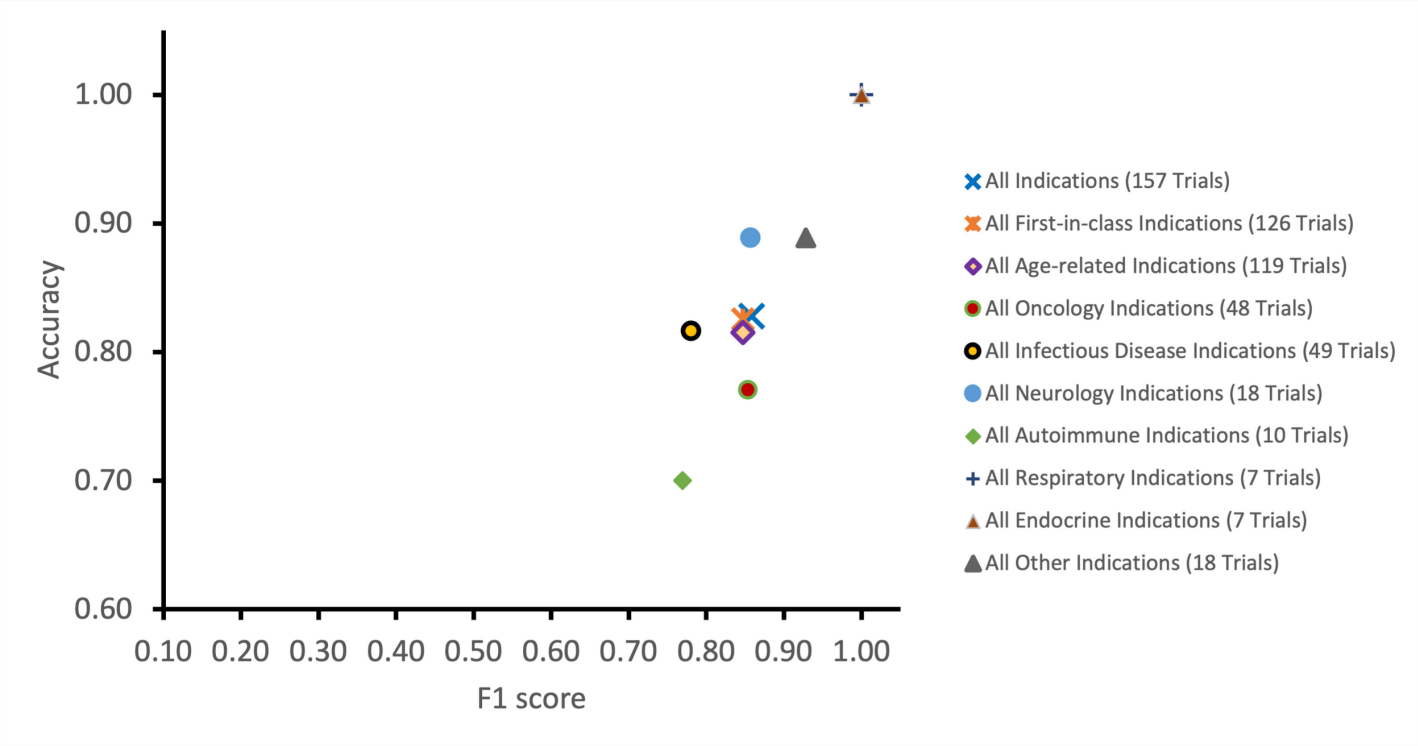


Figure 6c. ABNN's performance in terms of accuracy and F1 score (see Data Table 1 for more details).

Figure 6d. Performance of ABNN: LOA and F1 Score

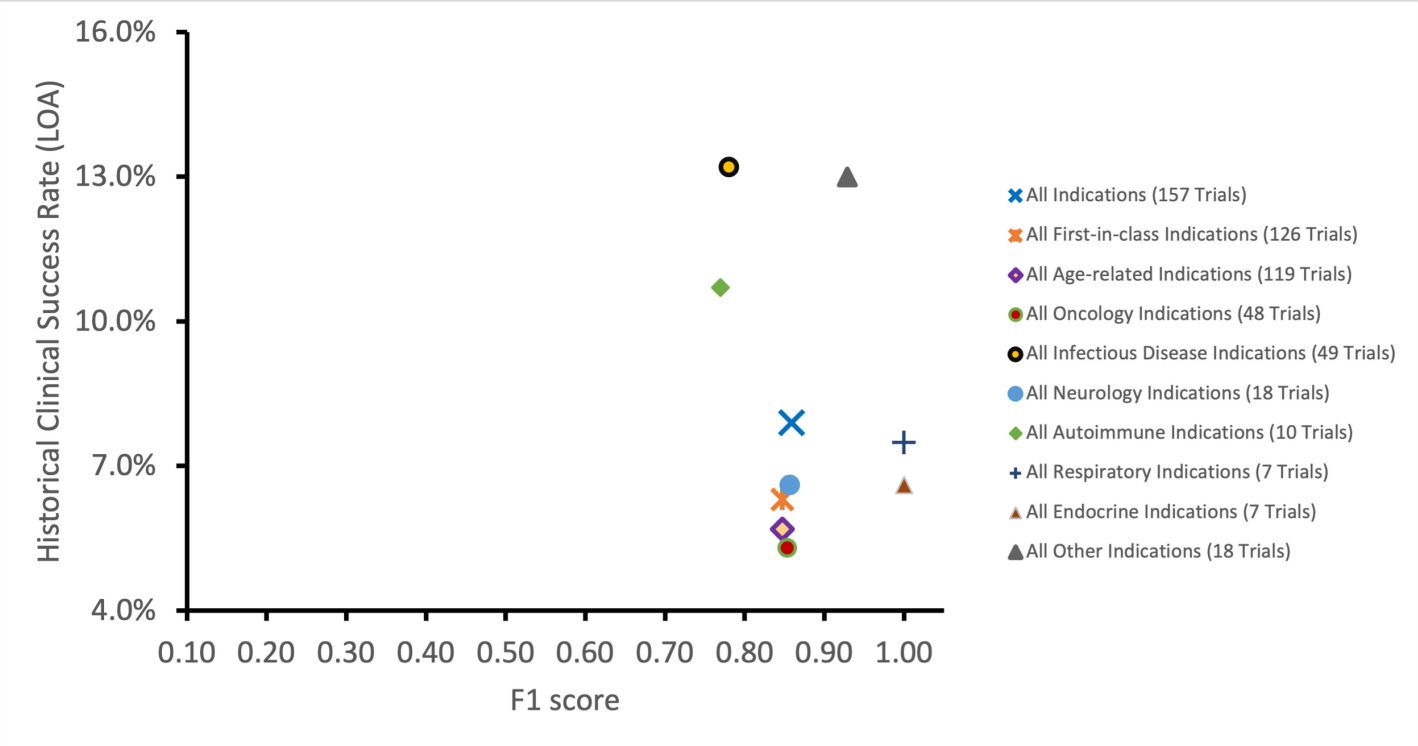


Figure 6d. ABNN's performance in terms of historical clinical success rates for investigation new drugs (LOA) and F1 score (see Data Table 1 for details). LOA is identical to the realistic baseline accuracy (see Methods section 'Baseline performance'). F1 scores and LOAs across overall validation subgroups are negatively weakly correlated (exact Pearson coefficient = -0.260)

Figure 7a. Performance of ABNN: Confusion Matrices for All Therapeutic Areas

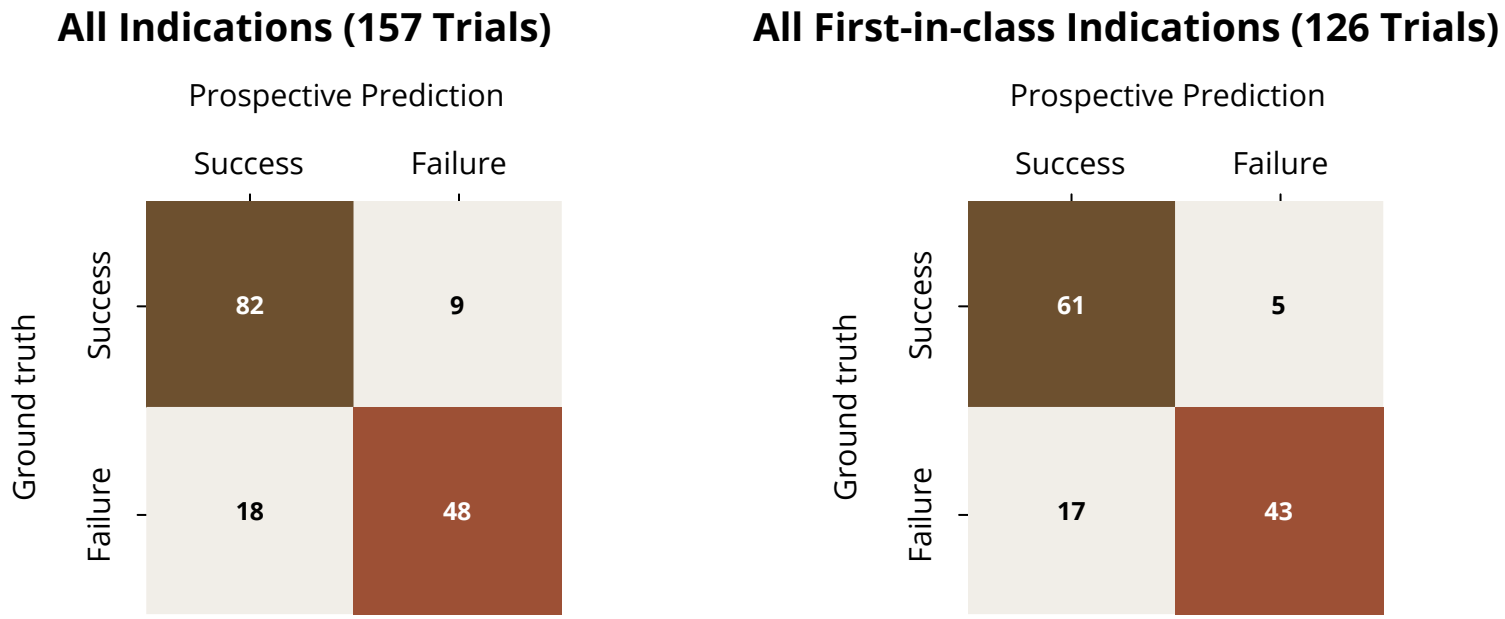


Figure 7a. ABNN's performance in terms of confusion matrices for all therapeutic areas.

Figure 7b. Performance of ABNN: Confusion Matrices Per Therapeutic Area

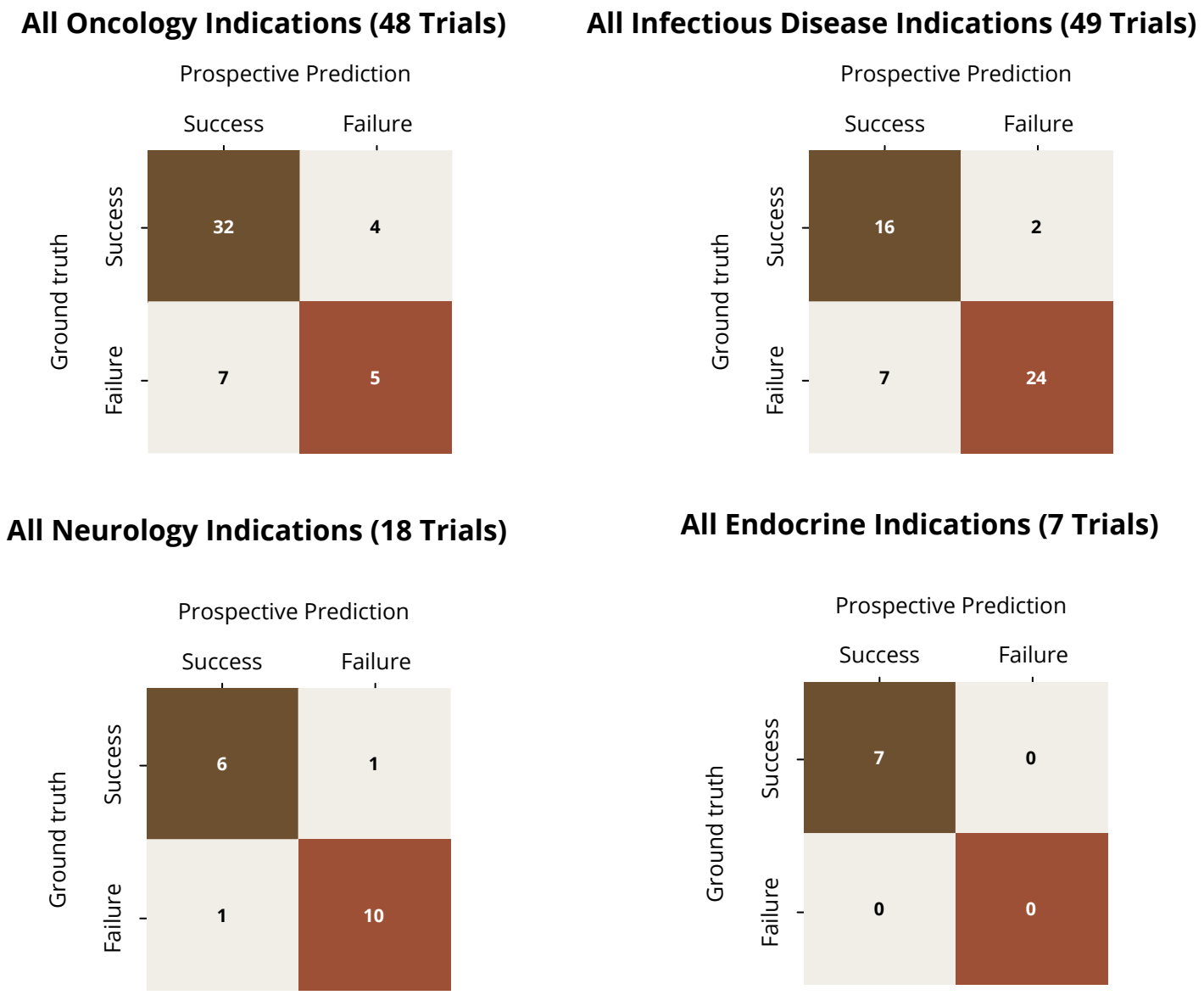


Figure 7b. ABNN's performance in terms of confusion matrices per therapeutic area.

Figure 7c. Performance of ABNN: Confusion Matrices Per Therapeutic Area

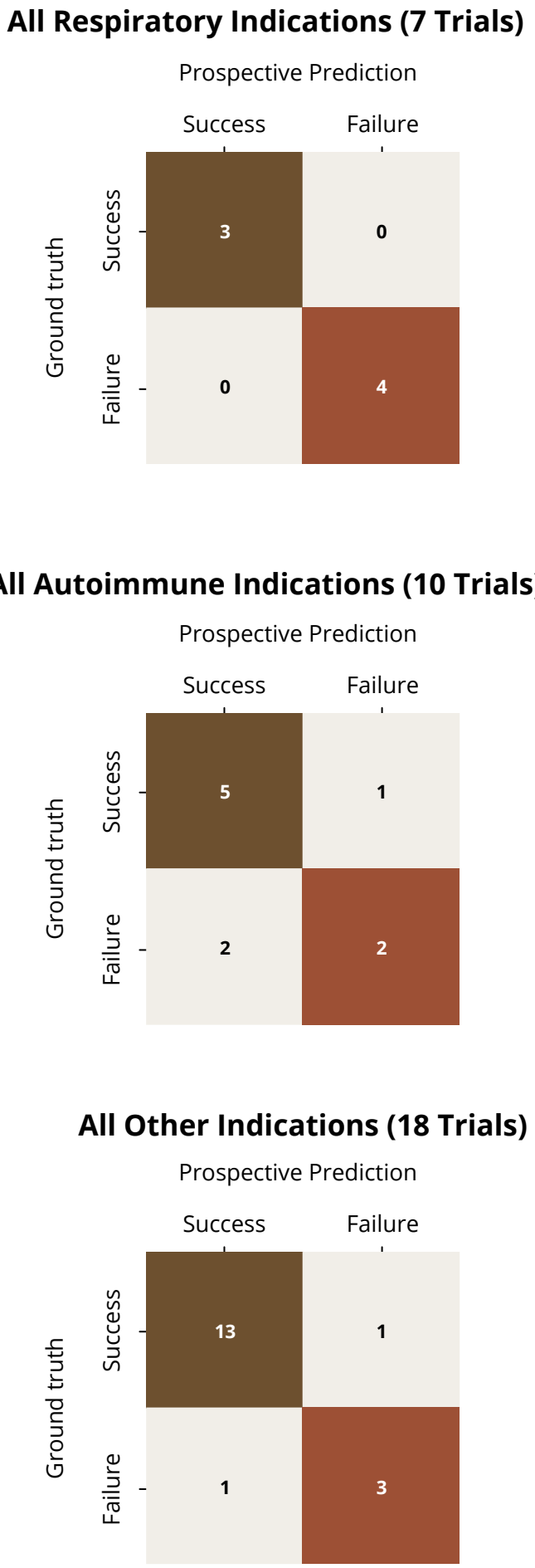


Figure 7c. ABNN's performance in terms of confusion matrices per therapeutic area.

Figure 8. Performance of ABNN: Confusion Matrices for Age-Related Diseases

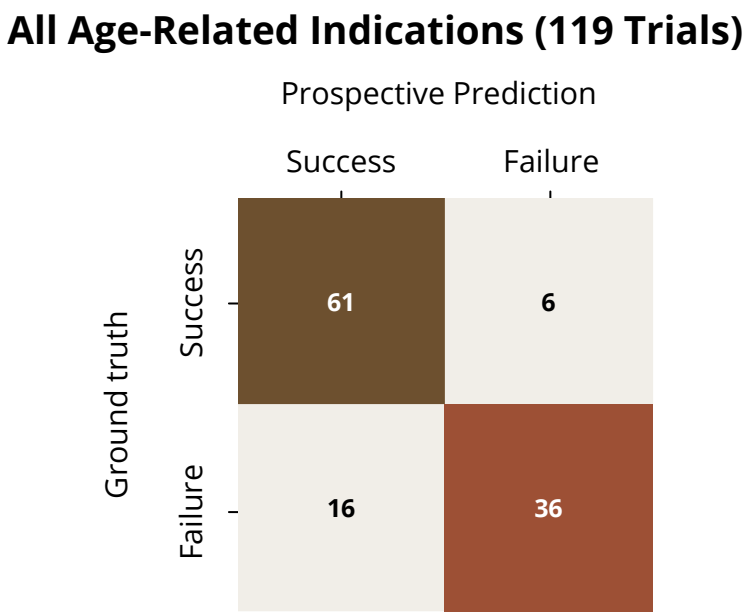


Figure 8. ABNN's performance in terms of confusion matrices for all age-related diseases.

Figure 9a. Prospective Prediction Sensitivity for All Age-related Indications

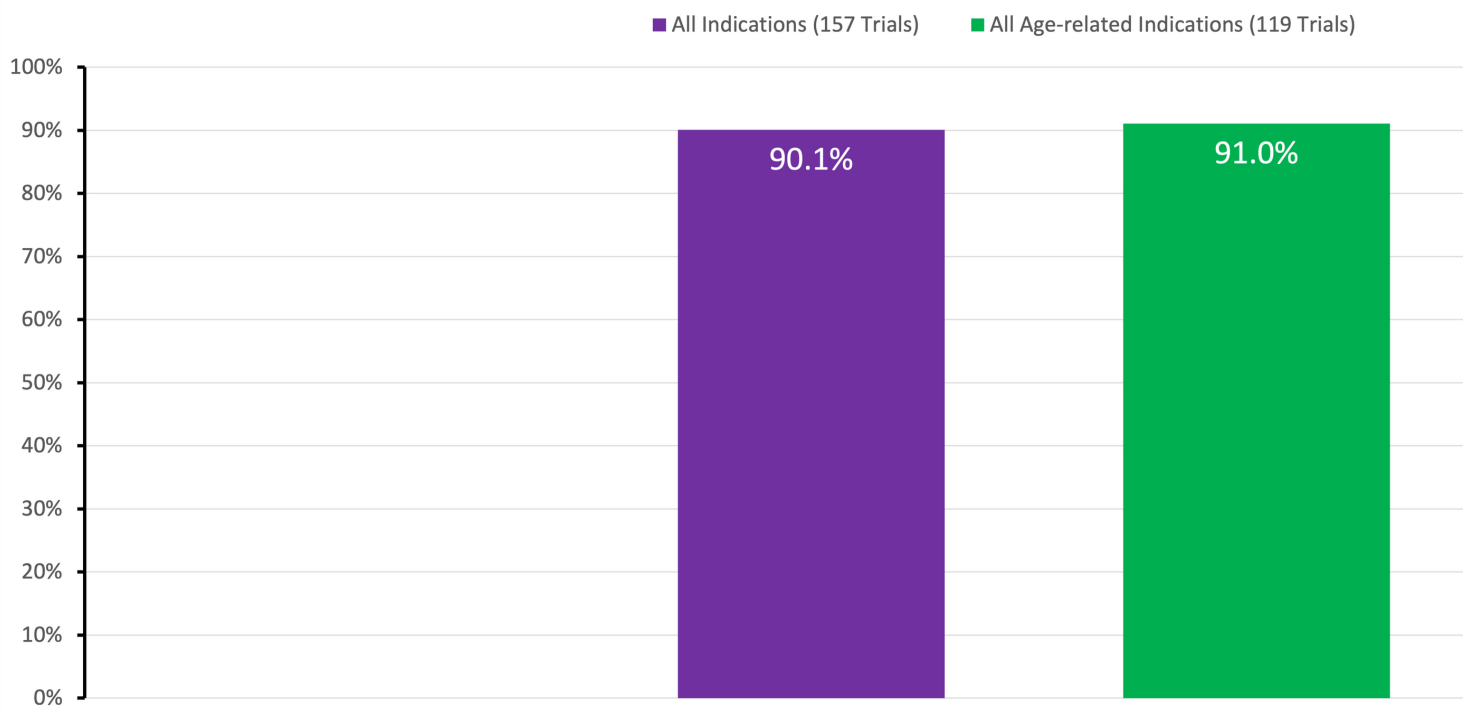


Figure 9a. ABNN's overall prospective prediction sensitivity of all age-related pivotal clinical trials is 91.0% (99% CI 79.1%, 98.4%; P < 0.005) in the PROTOCOLS validation (Methods section 'Statistical analysis').

Figure 9b. Distribution of Age-related Indications per Therapeutic Area

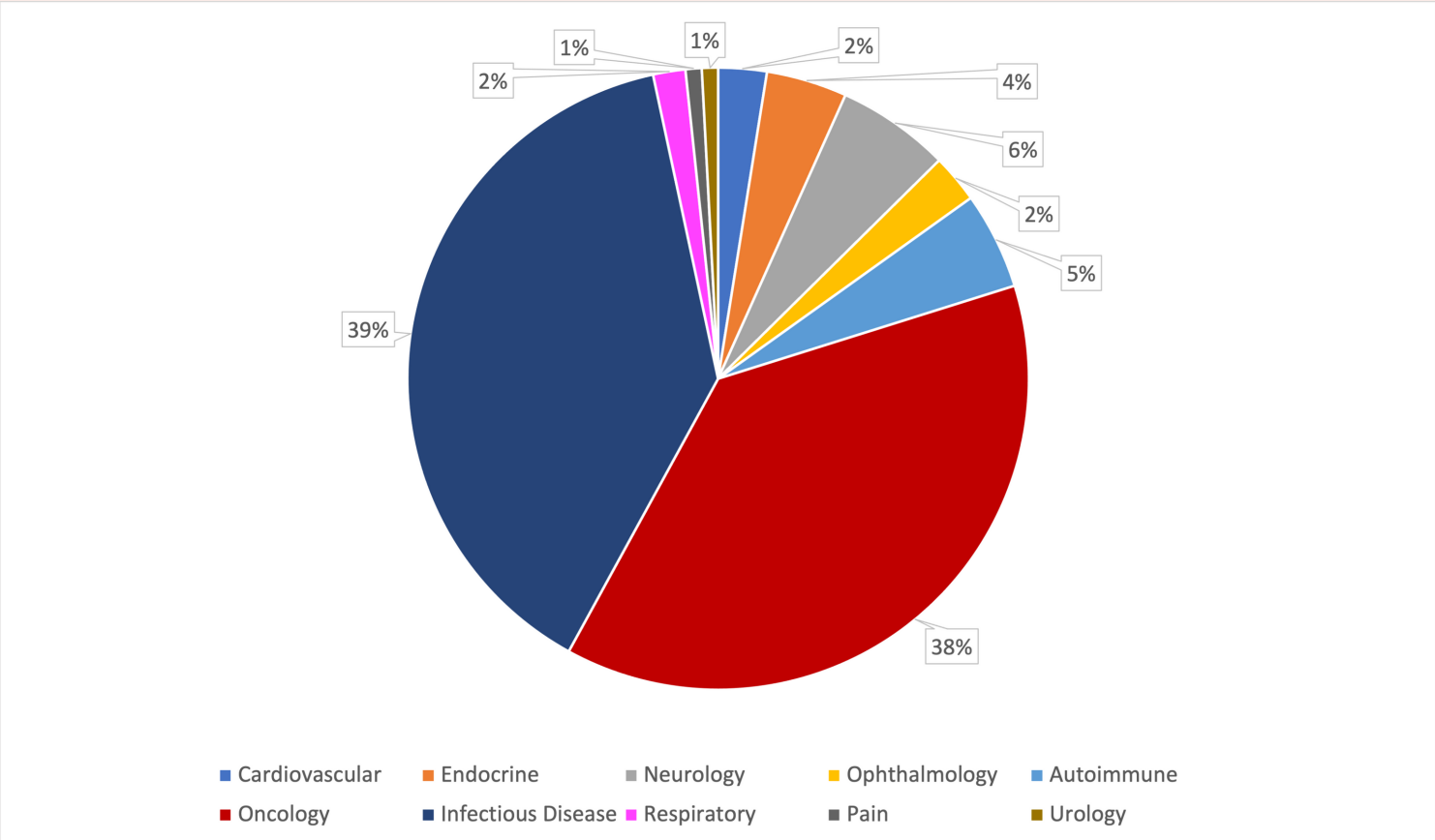


Figure 9b. The distribution of all age-related pivotal clinical trials per therapeutic area in the PROTOCOLS validation (Methods section 'Validation data collection').

Figure 10. Performance of ABNN: Confusion Matrices Per Drug Modality

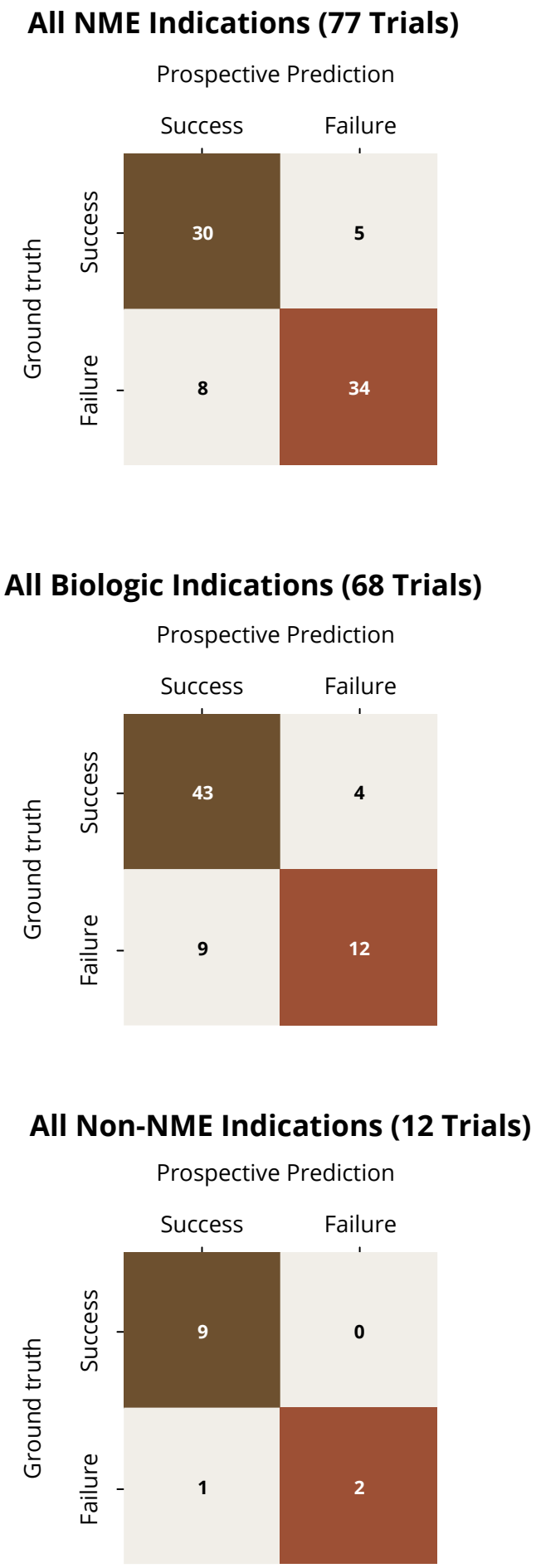


Figure 7d. ABNN's performance in terms of confusion matrices per drug modality.

Figure 11a. Performance of ABNN Per Drug Modality: Prospective Prediction Sensitivity

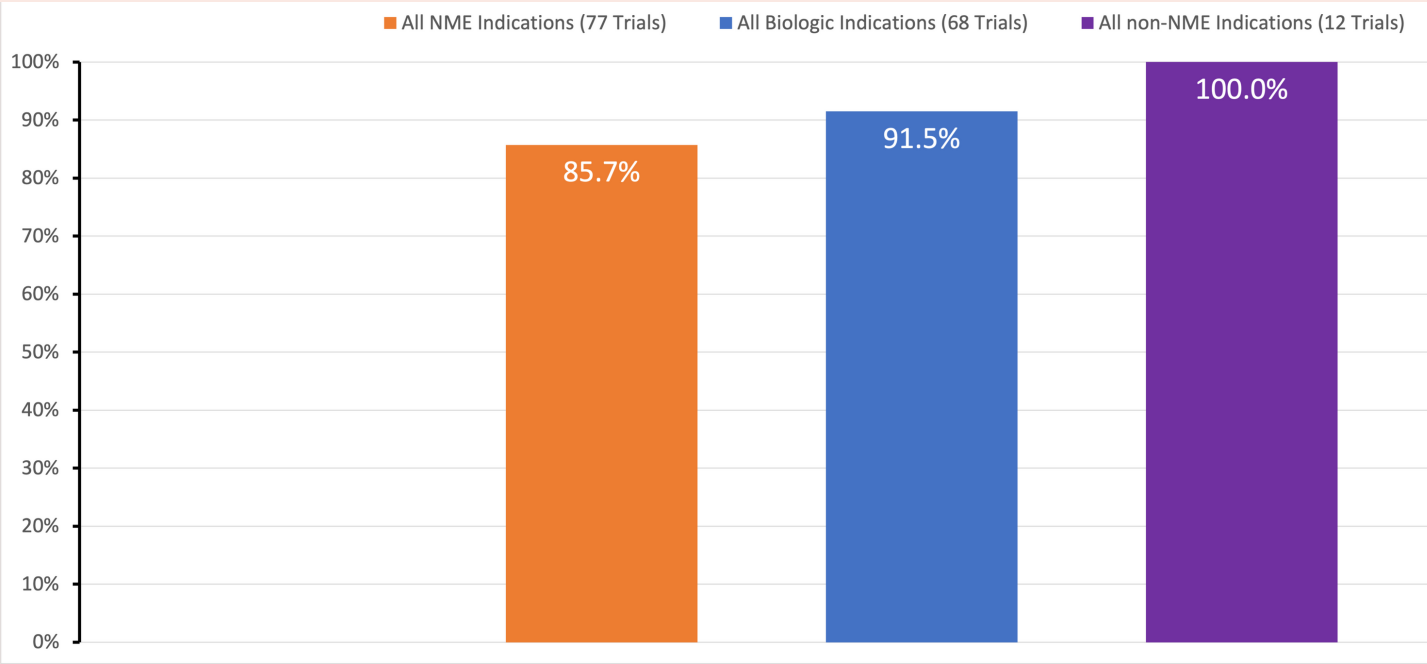


Figure 11a. ABNN's performance in terms of prospective prediction sensitivity across drug modalities (see Data Table 1 for more details). ABNN's achieved superior performance for biologics than for small molecules ((Δ = 5.8%; 90% CI 0.78%, 10.82%; P < 0.0001 for superiority at a 2% margin; Methods section 'Statistical analysis').

Figure 11b. Performance of ABNN Per Drug Modality: LOA and F1 Score

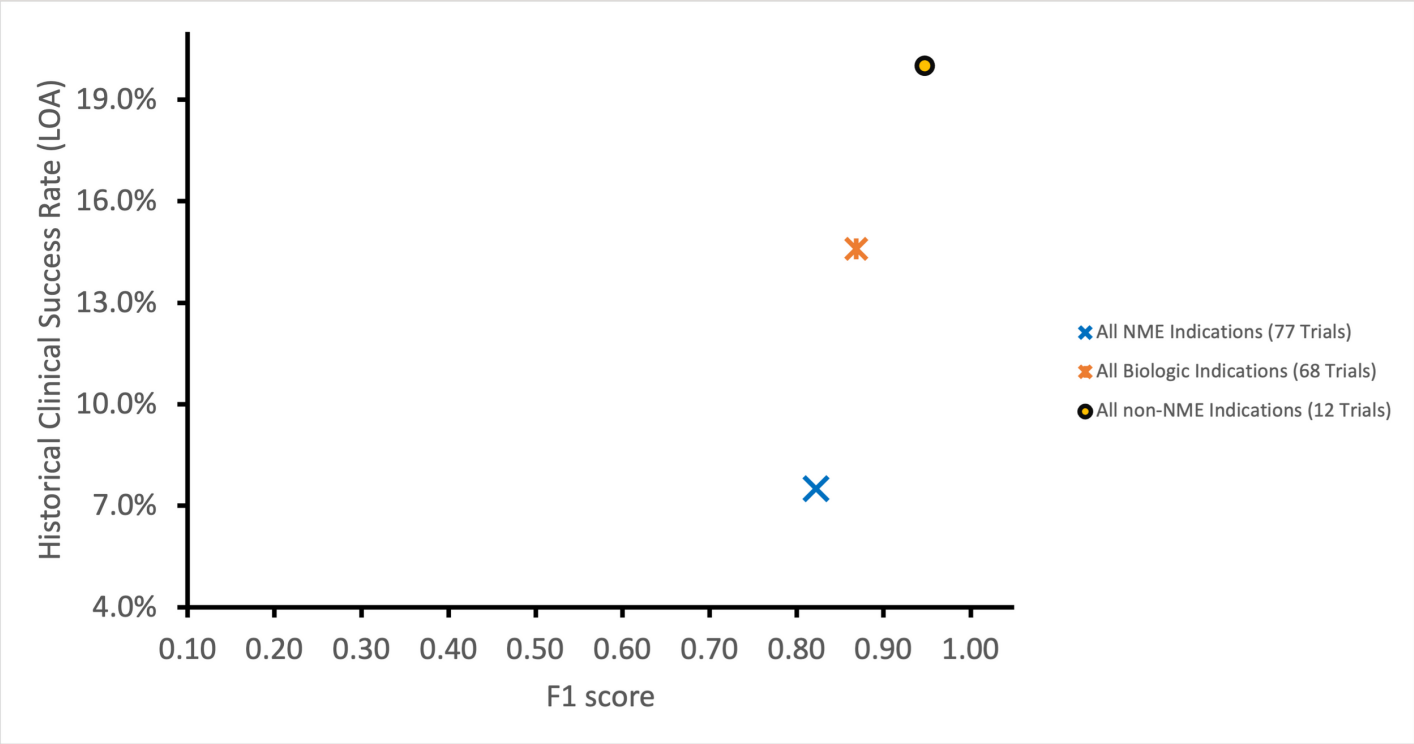


Figure 11b. ABNN's performance in terms of historical clinical success rates for investigation new drugs (LOA) and F1 score (see Data Table 1 for details). LOA is identical to the realistic baseline accuracy (see Methods section 'Baseline performance'). F1 scores and LOAs across overall validation subgroups are strongly positively correlated (exact Pearson coefficient = 0.975)

Data Table 1. ABNN's Performance Measures with Statistical Significance

Validation Subgroups	N	Metric	Performance	Significance	CI (%)	p-value
All Indications	157	Sensitivity	90.1%	99%	(80.0, 96.9)	0.001
All First-in-class Indications	126	Sensitivity	92.4%	99%	(80.8, 99.2)	0.009
All Age-related Indications	119	Sensitivity	91.0%	99%	(79.1, 98.4)	0.005
All Infectious Disease Indications	49	Sensitivity	88.9%	90%	(68.3, 95.4)	0.050
All Oncology Indications	48	Sensitivity	88.9%	90%	(75.7, 94.3)	0.015
All NME Indications	77	Sensitivity	85.7%	90%	(71.9, 92.2)	0.009
All Biologics Indications	68	Sensitivity	91.5%	90%	(80.8, 95.7)	0.015

Validation Subgroups	N	Metric	Performance	Significance	CI (%)	p-value
All Indications	157	F1-Score	0.86	99%	(0.79, 0.92)	0.0001
All First-in-class Indications	126	F1-Score	0.85	99%	(0.76, 0.92)	0.0001
All Age-related Indications	119	F1-Score	0.85	99%	(0.76, 0.92)	0.0001
All Infectious Disease Indications	49	F1-Score	0.78	90%	(0.65, 0.86)	0.0012
All Oncology Indications	48	F1-Score	0.85	90%	(0.77, 0.90)	0.0004
All NME Indications	77	F1-Score	0.82	90%	(0.73, 0.88)	0.0002
All Biologics Indications	68	F1-Score	0.87	90%	(0.80, 0.91)	0.0002

Validation Subgroups	N	Metric	Performance	Significance	CI (%)	p-value
All Indications	157	Accuracy	82.8%	99%	(74.2, 89.8)	0.0001
All First-in-class Indications	126	Accuracy	82.5%	99%	(72.8, 90.3)	0.0001
All Age-related Indications	119	Accuracy	81.5%	99%	(71.3, 89.7)	0.0001
All Infectious Disease Indications	49	Accuracy	81.6%	90%	(70.1, 88.4)	0.0020
All Oncology Indications	48	Accuracy	77.1%	90%	(65.1, 84.9)	0.0004
All NME Indications	77	Accuracy	83.1%	90%	(74.4, 88.6)	0.0002
All Biologics Indications	68	Accuracy	80.9%	90%	(71.3, 87.0)	0.0002

Validation Subgroups	N	Metric	Performance	Significance	CI (%)	p-value
All Indications	157	Specificity	72.7%	99%	(57.5, 85.3)	0.0001
All First-in-class Indications	126	Specificity	71.7%	99%	(55.6, 85.0)	0.0001
All Age-related Indications	119	Specificity	69.2%	99%	(51.8, 83.9)	0.0001
All Infectious Disease Indications	49	Specificity	77.4%	90%	(62.1, 86.4)	0.0040
All Oncology Indications	48	Specificity	41.7%	90%	(23.4, 64.2)	0.0080
All NME Indications	77	Specificity	81.0%	90%	(68.3, 88.3)	0.0018
All Biologics Indications	68	Specificity	57.1%	90%	(39.7, 72.3)	0.0016

Data Table 1. ABNN's primary performance measures.

Data Table 2. ABNN's All Performance Measures

Validation Subgroups	N	F1	ACC	SEN	SPE	NPV	PPV	FOR	FDR	FPR	FNR	PLR	NLR	DOR	MCC
All Indications	157	**0.86**	82.8%	90.1%	72.7%	84.2%	82.0%	18.0%	17.8%	27.3%	9.9%	3.30	0.14	24.29	0.65
All First-in-class Indications	126	**0.85**	82.5%	92.4%	71.7%	89.6%	78.2%	10.4%	21.8%	28.3%	7.6%	3.26	0.11	30.86	0.66
All Age-related Indications	119	**0.85**	81.5%	91.0%	69.2%	85.7%	79.2%	14.3%	20.8%	30.8%	9.0%	2.96	0.13	22.88	0.63
Infectious Disease Indications	49	**0.78**	81.6%	88.9%	77.4%	92.3%	69.6%	7.7%	30.4%	22.6%	11.1%	3.94	0.14	27.43	0.64
Oncology Indications	48	**0.85**	77.1%	88.9%	41.7%	55.6%	82.1%	44.4%	17.9%	58.3%	11.1%	1.52	0.27	5.71	0.34
Other Indications	18	**0.93**	88.9%	92.9%	75.0%	75.0%	92.9%	25.0%	7.1%	25.0%	7.1%	3.71	0.10	39.00	0.68
Neurology Indications	18	**0.86**	88.9%	85.7%	90.9%	90.9%	85.7%	9.1%	14.3%	9.1%	14.3%	9.43	0.16	60.00	0.77
Autoimmune Indications	10	**0.77**	70.0%	83.3%	50.0%	66.7%	71.4%	33.3%	28.6%	50.0%	16.7%	1.67	0.33	5.00	0.36
Respiratory Indications	7	**1.00**	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	-	0.00	-	1.00
Endocrine Indications	7	**1.00**	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	-	0.00	-	1.00
All NME Indications	77	**0.82**	83.1%	85.7%	81.0%	87.2%	78.9%	12.8%	21.1%	19.0%	0.14	4.50	0.18	25.50	0.66
All Biologic Indications	68	**0.87**	80.9%	91.5%	57.1%	75.0%	82.7%	25.0%	17.3%	42.9%	0.09	2.13	0.15	14.33	0.53

Data Table 2. F_1: F_1-score; ACC: Accuracy; SEN: Sensitivity; SPE: Specificity; NPV: Negative Predictive Value. PPV: Positive Predictive Value; FOR: False Omission Rate;FDR: False Discovery Rate; FPR: False Positive Rate; FNR: False Negative Rate; PLR: Positive Likelihood Ratio; NLR: Negative Likelihood Ratio; DOR: Diagnostic Odds Ratio; MCC: Matthews Correlation Coefficient; NME: New Molecular Entity (small molecule drugs);