Prospectively validated disease-agnostic predictive medicine with augmented intelligence

Bragi Lovetrue[1] & Idonae Lovetrue[1]

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

Despite numerous remarkable achievements in a variety of complex challenges[1,2], standalone artificial intelligence (AI) does not unlock life science from the long-term bottleneck of linearly extracting new knowledge from exponentially growing new data[3] that severely limits the clinical success rate of drug discovery[4]. Inspired by state-of-the-art AI training methods, we propose a human-centric 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, a real-world validation— the public, prospective prediction of pivotal ongoing clinical trial success outcomes at large scale (PROTOCOLS)—was also developed to benchmark HAI's performance with experts' baseline in the prospective prediction sensitivity of successful clinical programs with only preclinical data as inputs. HAI achieved a 11.4-fold improvement from the baseline with a 99% confidence[7] in the PROTOCOLS validation. The decision-support integration of HAI at the stage of selecting development candidates to enter first-in-human studies could significantly increase the average clinical success rate of investigational new drugs from 7.9%[8] to 90.1% for nearly any disease[9,10]. This result confirms that exponentially extracted knowledge alone may be sufficient to accurately predict clinical success, effecting a complete reversal of Eroom's law[4]. The

trained 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]. This study demonstrates that disruptive breakthroughs necessitate the smallest team size to attain the largest HAI for optimal knowledge extraction from high-dimensional, low-quality data spaces, thus establishing the first prospective proof of the previous discovery that small teams disrupt[12]. The global adoption of training HAI may provide a new path to mass producing scientific and technological breakthroughs via exponential knowledge extraction and better designs of data labels for training better-performing AI[13].

1 INTRODUCTION

Despite recent successes[1,2], AI does not enable life science to overcome the long-term bottleneck of linearly extracting new knowledge from exponentially growing new data[3]. The long-term shortage of knowledge in life science has accelerated the decline of productivity in drug discovery even after the widespread adoption of AI for biomedical data analysis[4]. For example, recent studies have shown that only 20% of biomedical research generated reproducible data[14] and fewer than 10% of animal studies successfully predicted the clinical efficacy of new drugs[15]. It has been proposed that increasing the scale of data collection or the efficiency of knowledge extraction could address this issue[3].

High dimensionality and low quality are two defining characteristics of biomedical data, as biological systems involve large quantities of entities across multiple scales. These attributes make it more difficult to reproduce and structure biomedical data. Deep neural networks (DNNs) are well

suited for extracting knowledge from the high-dimensional, high-quality data spaces[2]. In contrast, biological neural networks (BNNs) are adept in extracting knowledge from low-dimensional, low-quality data spaces. As a result, neither DNNs nor BNNs alone are suitable for improving the efficiency of knowledge extraction from biological data.

The multi-faceted correspondence between DNNs and BNNs[5] suggests the novel possibility of combining their complementary advantages. As such, we propose adapting best practices for DNN training to a BNN, augmenting the resulting network for learning from high-dimensional data. In principle, an augmented BNN (ABNN) is well suited to learning knowledge from the high-dimensional, low-quality life science data and informing better designs of life science data labels (Figure 1a). As an ABNN is necessarily hosted in a human brain and might be translated into high-performing DNNs, it is a form of human-centric augmented intelligence (HAI).

There are several key advantages of training an ABNN over training an DNN (Figure 1b). Firstly, there is no need to re-design network architecture or to tune network hyperparameters for any ABNN, because the human brain, as the host of an ABNN, has been optimized to learn via evolution. Secondly, there is no need to pre-process data for training an ABNN, because all publicly available life science data are human-interpretable and thus ABNN-friendly, even though most of those data are unlabeled, unstructured or uncleaned, thus inscrutable for an DNN without adequate preprocessing. Thirdly and more importantly, training an ABNN may not suffer from overfitting because ABNN has sufficient generalizability to learn from the opposite-category data of what it will be applied to, whereas overfitting is somewhat inevitable for training an DNN

because DNN lacks generalizability and must learn from the same-category data as what it will be applied to (Figure 1b). In the context of life science and drug discovery, an ABNN could be trained with biological data about normal physiological states only, and then tested solely with clinical data about abnormal pathological states. Such a trained ABNN is supposed to directly extrapolate from the normal states of human physiology to the abnormal states of human diseases, even before research data become available for the abnormal states of a new disease like COVID-19.

Here we introduce the first ABNN, developed by training a BNN to learn like an DNN (Figure 2a and 2b) for eight years. Since larger DNNs very often perform better[16,17], we propose that the largest ABNN could only be attained by limiting the research team size to a single brain to maximally increase the efficiency of knowledge extraction (Figure 2d). As such, the best practices of DNN training were adapted to a single ABNN that learns from all-encompassing life science data in an uninterrupted pass without imposing intermediate milestones (Figure 2e). The trained ABNN achieved exponential knowledge extraction from an exhaustive body of publicly accessible life science publications and databases, thereby constructing a foundation model[6] of human physiology and over 130 specific models of human disease (Figure 2c).

The quality of extracted knowledge by the ABNN was assessed by measuring resulting increases in the productivity of drug discovery. Inspired by the CASP challenge[18], which serves as a benchmark for evaluating an DNN's performance against experts' performance in protein structure prediction, we designed a real-world validation—the public prospective prediction of pivotal ongoing clinical trial success outcomes at large scales (PROTOCOLS) challenge—to benchmark

ABNN's performance against experts' performance in clinical success prediction (Figure 3a, 3b and Methods section 'Validation design').

The rationale for the PROTOCOLS challenge is that pivotal clinical trials (phase 2 and phase 3) are statistically powered to assess a drug's clinical success by evaluating its clinical efficacy and safety, which is largely determined by the underlying biological processes that can be distilled into scientific knowledge. Historical clinical success rates for investigational drugs are well-documented[8,19] and provide an accurate baseline for evaluating ABNN's performance in the PROTOCOLS challenge.

Clinical success prediction is a critical go/no-go decision step in drug development because only a subset of developable drug candidates with highest clinical success potential can be selected to transition from non-clinical to clinical development due to the prohibitive cost of human clinical trials[20]. As the best practice in the biopharmaceutical industry[21], a group of committed experts review both publicly available and privately accessible preclinical data to predict which drug candidates will achieve clinical success, and initiate clinical trials only for those drug candidates that are prospectively predicted to be clinically successful (Figure 3a). However, the real-world industrial practice has yielded a prospective prediction sensitivity of 7.9% (that is for every 100 clinical programs that experts prospectively predicted to achieve clinical success, only 7.9 of them achieve actual clinical success)[19], which is only slightly better than chance[22].

The real-world PROTOCOLS validation is designed to mirror the real-world industrial practice so as to minimize the gap of performance from validation to application: The ABNN used

the publicly available subset of experts-accessible preclinical data to predict which drug candidates will achieve clinical success, tracked the readouts of clinical trials especially for those drug candidates that are prospectively predicted to be clinically successful (Figure 3a). In contrast to the experts' baseline performance of 7.9%, the ABNN followed identical procedures with fewer data yet yielded a prospective prediction sensitivity of 90.1% (aka. for ever 100 clinical programs that were prospectively predicted to achieve clinical success, 90.1 of them achieve actual clinical success), representing a 11.4-fold improvement over the baseline in realistic settings.

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 have little gap between the validation performance of ABNN in the PROTOCOLS challenge and the clinical utility of ABNN in real-world development candidate selection[23,24].

2 RESULTS

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

The samples of validation clinical trials and newly initiated clinical trials since 2020 were shown to follow identical distributions across all major therapeutic areas (two-sample Kolmogorov-Smirnov test, statistic = 0.333; P = 0.73; Figure 4). The PROTOCOLS thus provided an unbiased

evaluation of ABNN's disease-agnostic performance.

Minimum sample size (131) and event size (118) requirements were determined for evaluating ABNN's performance with 99% confidence[7] (Figure 3c and Methods section 'Statistical analysis'). Both requirements were met as clinical readouts were available for 157 of 265 predictions by the cutoff date (Methods section 'Clinical readout'). A ground truth was then determined for each readout and was then applied to validation of the corresponding prediction (Methods section 'Prediction Validation), producing a set of confusion matrices for varying subgroups of the validation set (Figure 7).

It should be noted that positive instances (true positives and false negatives) are far more important than negative instances (true negatives and false positives) in real-world scenarios of life science and drug discovery. True positives are also more interesting than true negatives because the historical clinical success rate for investigational drugs entering human studies is only 7.9% across all diseases[8]. As such, false positives have only a marginal clinical impact because human clinical trials can still fail without doing harm to patients. In contrast, false negatives will have a substantial clinical impact because they not only deny clinical benefits to patients, but also eliminate sizable revenue opportunities.

Accordingly, we chose sensitivity and F1 score as appropriate performance metrics for the PROTOCOLS validation[25]. Our choice of sensitivity realistically reflects the best practice in actual drug development: Every investigational new drug greenlit to enter first-in-human studies is then genuinely believed by qualified experts to achieve clinical success, so the clinical success rate from

phase 1 to approval is construed as a statistical measure of sensitivity[19]. We also reported other standard measures (accuracy, specificity, etc.) to provide a balanced view of ABNN's performance (Data Table 2).

Since the ABNN was trained to functionally reconstruct human physiology and diseases (Methods section 'Model training'), its disease-agnostic performance was first assessed for the full validation set (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 over the realistic baseline sensitivity of 7.9% and an almost 6.0-fold improvement over the conservative baseline sensitivity of 15.1% (Figure 5a; Methods section 'Baseline performance'). The 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). These data clearly demonstrate that the ABNN has extracted large-scale, high-quality knowledge, enabling its excellent disease-agnostic performance in real-world settings of the PROTOCOLS validation.

Since the ABNN should not require access to any clinical data before making predictions of investigational new drugs' clinical success (Figure 3a) for development candidate selection, further analysis of subgroups was conducted for the validation set, with varying degrees of clinical data availability.

First-in-class clinical trials for novel drug-disease pairs represent a fairly challenging subset of the PROTOCOLS validation because no prior clinical data are available for those drug-disease

pairs across all therapeutic areas. Since PROTOCOLS included 128 first-in-class clinical trials with clinical readouts, the minimum sample size and event size requirements for a 99% confidence interval (CI) have been met. The 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 exhibited superior disease-agnostic prospective prediction sensitivity for first-in-class clinical trials, compared with all 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 to its performance 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'). These results clearly demonstrate that ABNN's outstanding performance in the PROTOCOLS validation is independent of clinical data availability for arbitrary drug-disease pairs.

Oncology clinical trials represent a different type of challenge for the ABNN, as they conventionally suffer from the lowest historical clinical success rates: The realistic baseline sensitivity for oncology is 5.3%, in contrast to the highest sensitivity of 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. As such, a clinically meaningful assessment of ABNN's performance in oncology should primarily focus on sensitivity, F1 score and accuracy.

In addition, oncology clinical trials often include quite complex trial designs and typically involve active controls and combination therapies as a standard design for pivotal clinical trials. Human cancer patients are further stratified by certain biomarkers[26], 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 set, the minimum sample size and event size requirements for a 90% CI have been met (see Methods section 'Statistical analysis'). The 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 over the realistic baseline sensitivity of 5.3% and an almost 8.2-fold improvement over 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). These data clearly demonstrate that the ABNN has extracted sufficiently complex and intricately nuanced knowledge of the human body and cancer etiology to accurately predict the clinical success of cancer drugs for biomarker-differentiated heterogeneous patient populations.

COVID-19 represents the ultimate test for the quality of existing knowledge in the ABNN because no knowledge of COVID-19 can be directly extracted (due to a lack of prior data shortly

after the outbreak). The entire disease model for COVID-19 can only be constructed by generalizing other knowledge to COVID-19 clinical symptoms. In addition, no clinical data were available worldwide when the preprint of ABNN's COVID-19 disease model was published on March 20, 2020[27].

As there are 49 infectious disease clinical trials in the validation set, the minimum sample size and event size requirements for a 90% CI have been met (see Methods section 'Statistical analysis'). In total, 94% (46/49) are COVID-19 pivotal clinical trials (Supplementary Table 4).

The ABNN achieved 88.9% prospective prediction sensitivity (90% CI 68.3%, 95.4%; $P < 0.05$; Figure 5c; Data Table 1) for infectious diseases in the validation set. This represents 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'). The ABNN also 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). These data clearly indicate the quality of extracted knowledge in ABNN has reached a critical mass, enabling 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.

Across all the other therapeutic areas in the PROTOCOLS validation (60 predictions in total; Supplementary Table 7), the 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 a 90% CI were not met for these data (Methods section 'Statistical analysis'), the results demonstrate that ABNN's performance is stable and consistent.

Alzheimer's Disease (AD) was also included in the PROTOCOLS validation and served to further demonstrate the robustness of ABNN's extracted knowledge. AD represents an extreme case because the historical clinical success rate of AD-modifying drugs is nearly zero[28]. A single true positive then becomes an appropriate performance metric for AD because both false positives and negative instances are equally uninteresting in real-world settings. As such, AD represents a tough challenge for ABNN because of a consensus that it is overly complex, the quantity of existing data is too low, and the quality of existing data is too poor[28].

Despite this, the ABNN achieved 100% accuracy and 100% precision with non-zero true positives and zero false positives (Supplementary Table 6) for three prospective predictions of AD pivotal clinical trials in the PROTOCOLS validation. The first true positive was Masitinib, a first-in-class c-Kit inhibitor that met the efficacy primary endpoint in a pivotal phase 3 trial (NCT01872598). These results demonstrate the potential of the ABNN for extracting large-scale, high-quality knowledge of AD from existing life science data.

Aging is an evolutionarily conserved phenomenon across all mammalian species[29] and is clinically recognized as the main risk factor of numerous diseases in multiple therapeutic areas[30]. 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 set, the minimum sample size and event size requirements for a 99% CI have been met. The 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's prospective prediction sensitivity for all age-related clinical trials is non-inferior to its performance for all first-in-class clinical trials (Δ = 1.4%; 99% CI -1.12%, 3.92%; $P < 0.003$ for superiority at a 2% margin; Methods section 'Statistical analysis'). The ABNN's prospective predictive accuracy for age-related clinical trials is also shown to be non-inferior to its performance for first-in-class clinical trials (Δ = 1.0%; 99% CI -1.14%, 3.14%; $P < 0.0001$ for non-inferiority at a 2% margin; Methods section 'Statistical analysis').

The ABNN's prospective predictive sensitivity for all age-related clinical trials is statistically significantly higher than its performance for all clinical trials in the validation set (Δ = 0.9%; 99% CI -1.13%, 2.93%; $P < 0.0001$ for superiority at a 2% margin; Figure 9a; Methods section 'Statistical analysis'). These results clearly demonstrate that the 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 the ABNN and the difficulty of specific therapeutic areas by calculating correlations 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 were negatively and weakly correlated (Exact Pearson r = -0.260), indicating that the quality of ABNN's knowledge concerning human diseases is orthogonal to the quality of collective knowledge for the corresponding therapeutic areas. Surprisingly, ABNN's performance was higher for increasingly difficult therapeutic areas (Figure 6d; Data Table 2).

Finally, we explored the relationship between the quality of extracted knowledge in the ABNN and drug modality (Figure 10). The 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) for small molecule drugs (77 NME trials; Methods section 'Drug classification'). The ABNN's performance is consistently high 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 calculated correlations 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 were positively and strongly 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

In this study, the possibility of a deep-learning-augmented human intelligence (HAI) has been presented for constructing a foundation model of human physiology, potentially enabling highly accurate data-free predictions of clinical success for investigational new drugs for almost any disease. This achievement is intractable for data-driven machine intelligence, on which the demands for clinical data are unrealistically high and the efficiency of knowledge extraction is sufficiently low that gaps between promise and proof have become nearly unbridgeable (Figure 1).

Humans have designed machine learning methods to train deep 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 the same approach could be adapted to training augmented biological neural networks (ABNNs) in learning quasi-white-box models[27] to predict the clinical success of investigational new drugs (Figure 1b).

The resulting ABNN could safely deliver AI-surpassing clinical utility while averting the privacy and explainability challenges faced by medical AI systems[24]. The ABNN could also inform the better design of labels for training better-performing DNNs with fewer data[13].

Furthermore, in contrast to domain experts who have privileged access to private preclinical data in addition to public preclinical data, the presented ABNN was trained solely using public preclinical data yet has extracted orders-of-magnitude more knowledge from far less data than experts (Figure 3a).

PROTOCOLS represents the most rigorous validation design to date for evaluating the real-world utility of extracted knowledge from large-scale, low-quality data spaces (Figure 3 and Figure 4). With a 99% CI, the trained ABNN achieved a prospective prediction sensitivity of more than 90%, an F1 score of more than 0.85, and a prospective prediction accuracy of more than 82% across all therapeutic areas, independent of the availability of prior clinical data (Figure 5 and Figure 6; Data Table 2).

In addition, the drug mechanism of action and clinical trial design protocol were the only inputs used by the ABNN to evaluate the clinical success of novel drugs (Figure 3a). The ABNN's performance in PROTOCOLS confirmed that large-scale, high-quality scientific knowledge alone may be sufficient to accurately predict drug clinical success.

Given that both ABNN inputs are available prior to 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 serve as a decision-support intelligence for the selection of drug candidates for clinical development, significantly increasing the clinical success rates for investigational new drugs from 7.9%[8] to 90.1% and bringing a permanent reversal of Eroom's law[4] (Figure 3a).

The trained ABNN is also the world's first clinically validated model of human aging (Figure 8 and Figure 9). Its application to the development of cellular rejuvenation programming could substantially speed up the discovery of preventive medicines for age-related diseases[11].

The ABNN also leaves room for improvement as 17.2% (27/157) of predictions were either false positives (18/157) or false negatives (9/157) in the PROTOCOLS validation. As such, further work is planned to investigate why the ABNN performs better for harder diseases (Figure 6d). The exponential efficiency of knowledge extraction enabled the ABNN to translate each false prediction into substantial improvements in the corresponding disease model.

While this work could significantly reshape medicine, its broader impact lies in the presented training methodology, which provides a new path for drastically increasing the efficiency of knowledge extraction in every discipline that is characterized by large-scale, low-quality data spaces (e.g., synthetic biology, material science, and climate science). These results also provide the first prospective proof that the smallest research team is necessary for delivering the most disruptive breakthroughs[12].

The advent of ABNN ushers in the new era of human-centric augmented intelligence. The global adoption of ABNN training could produce a paradigm shift in research culture and organizational structure, in an effort 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[31].

4 METHODS

TRAINING DATA COLLECTION

The included training dataset comprises all publicly accessible sources of life science data including peer-reviewed publications and curated databases, which cannot be enumerated here due to space limitations.

MODEL TRAINING

See Figure 2. The ABNN was first trained to learn a foundation model of human physiology, which was then adapted to learn specific models of human diseases. The ABNN is principally disease-agnostic yet leaves room for improvement in identifying hematological disorders, rare neurodevelopmental disorders, and rare musculoskeletal disorders. These unfinished disorders were excluded from the performance validation yet were included in validation data collection because prospective predictions can provide additional insights to accelerate model building for those disorders.

VALIDATION METHODOLOGY

See Figure 3a and Figure 3b.

VALIDATION DESIGN

Validation is critical for an objective assessment of ABNN's capabilities in exponentially extracting general-purpose knowledge from life science data. The simplest approach is to determine the extent to which ABNN could reverse Eroom's law, yet it would be unrealistically time-consuming and non-scalable to directly begin new drug development. As such, running virtual

clinical trials is an effective alternative. Drug mechanism of action and clinical trial design protocol would serve as the only inputs and the ABNN would predict new drug clinical efficacy using extracted knowledge only without access to patient data or confidential third-party information. Highly accurate virtual clinical trials could be run before first-in-human studies and prevent drug developers from initiating pivotal real-world trials that are doomed to fail[32].

A rigorous validation of the ABNN, including the stringent criteria of the PROTOCOLS challenge as shown in Figure 3b, would involve publishing prospective predictions for pivotal clinical trial outcomes for all human diseases. Tamper-proof timestamps could be included and prediction sensitivity could be determined with a 99% CI, based on a ground truth determined by actual clinical readouts. No selection of lead over non-lead indications would be included, although non-lead indications have a far lower success rate[19].

There are a few hundred new pivotal clinical trials each year whose protocols are publicly accessible in a centralized database (as required by the FDA[33]) and whose clinical readouts are typically published by sponsors in a standardized format. Public, prospective predictions of pivotal ongoing clinical trial outcomes could serve as a validation benchmark for ABNNs, similar to the way in which ImageNet provided a benchmark validation for visual object recognition using DNNs[34].

BASELINE PERFORMANCE

Conservative Baseline Sensitivity (CBS) is defined as the probability of FDA approval for

drugs in phase 2 development (Phase2 Likelihood of Approval, or P2LOA in abbreviation). It is calculated as the corresponding P2LOA in a 2011-2020 survey[8]. This baseline sensitivity is conservative because it does not consider that an ABNN only requires two inputs to generate predictions (drug mechanisms of action and clinical trial design protocols), and both are readily available in the real world even before first-in-human studies (phase 1). Regardless, this baseline sensitivity is more balanced in consideration of both the phase distribution of clinical trials in the validation set (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 (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 acknowledges that an ABNN would make identical predictions at earlier stages of development with drug mechanisms of action and clinical trial design protocols serving as the only inputs, both of which are available as early as preclinical stages.

RBA is not directly available from previous survey studies[8,19] for two subgroups of the validation sets: 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 historical 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 set comprised the registration information for 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 each year that would meet these the screening criteria, which is consistent with the 424 clinical trials each year exhibiting 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 set is designated as "first-in-class" if the drug-indication pair hasn't been approved for marketing worldwide.

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

PREDICTION GENERATION

Per the industry-wise best practices[35], the ABNN predicted SUCCESS for a pivotal clinical trial if at least one of its prespecified efficacy primary endpoint(s) was believed to be met with predefined statistical significance. In contrast, the ABNN predicted FAILURE for a pivotal clinical trial if none of its prespecified efficacy primary endpoint(s) were 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

A total of 265 predictions were published as timestamped tweets @DemiugeTech to establish a publicly accessible track record for prospective predictions (Supplementary Figure 1). The immutability of tweets ensures that each published prediction cannot be post hoc modified. We also indexed each tweet to prevent any intentional deletions of false predictions that go unnoticed to ensure a reliable performance evaluation. However, index gaps do exist for several legitimate cases, which does not compromise the rigor of validation for the following reasons: 1) An index was unintentionally skipped and remained unused from that point forward. 2) 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. We used DocuSign to distinguish these two cases of index gaps by electronically

signing every indexed prediction before publishing it as a tweet. As such, no DocuSign certificates are available for unused indices. Th indices 001 and 286 are the indices of the first and the last published prospective predictions, respectively. A total of 21 index gaps were detected, 5 of which were unused and 5 of which were for disqualified predictions and did not have clinical readouts as of the cutoff date (March 1, 2022). Nine predictions were made for disqualified predictions with clinical readouts as of the cutoff date (3 false predictions and 6 true predictions). Two predictions 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 tracked the availability of clinical readouts for every published prospective prediction from multiple online sources, including https://www.globenewswire.com/ for press releases and 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[36], a clinical trial with available readouts 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) are met according to topline analyses or final data or are 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 an 85.1% theoretical maximum prediction accuracy for drug clinical success by human clinical trials[37], we should refrain from directly labeling missed predictions as false without further consideration of subsequent developments. Specifically, we consider missed failure predictions for phase 2b trials that are less statistically powered than phase 3 trials. Missed failure predictions should thus 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

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. Similarly, 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. Confusion matrices are provided to assess ABNN's validation performance 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

The ABNN's performance in clinical success prediction was evaluated using the public, prospective prediction of pivotal ongoing clinical trial success outcomes at large scale (PROTOCOLS) challenge, a rigorous external validation of a clinical prediction model with a binary outcome[7].

A number of reasonable assumptions wee included to calculate minimal sample and event sizes, in order to determine to prospective prediction sensitivity, F1 score, accuracy, and specificity

at a 99% CI using the previously reported method[7].

The anticipated clinical readout event proportion (φ) was set to 0.9 because as of the cutoff date an average of 75% of clinical trials on clinicaltrial.gov have reported clinical trial results in compliance with the FDAAA 801 and the Final Rule tracked by https://fdaaa.trialstracker.net/. More than 90% of big-pharma-sponsored clinical trials report results[38].

The ratio of total observed clinical readout events and total expected clinical readout events (O/E) was set to 1 with a width of 0.15 to account for about 12% of clinical trials that are prematurely terminated without clinical trial results[39].

As such, the minimum sample size requirement was 131 predictions, and the minimal event size requirement was 118 readouts for a 99% CI. From February 11, 2020 to December 22, 2020, 265 prospective predictions were published to meet this minimal sample size requirement. By the tiem of 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). Thus the minimum event size requirement has been met.

All other factors held constant, performance measures were detected at a 90% CI for validation subgroups with smaller event sizes, with a minimum requirement of 52 predictions and a required event size of 47 readouts. Only simple means and standard deviations were calculated for validation subgroups with even smaller event sizes over the corresponding proportions.

The direct translatability of prospective validations to clinical ABNN applications was evaluated using a two-sample Kolmogorov-Smirnov test as the standard method to compare validation clinical trials and initiated clinical trials since 2020. The results followed an identical distribution (see Figure 4) and the oncology data in the validation had the same proportion as in the benchmark[19].

The statistical significance in real-world scenarios was ensured for actual clinical settings, using Agresti-Coull Intervals[40] for the proportions and Wald Intervals[40] for the differences. Two-sided p-values[41] were calculated for every proportion and difference and a 2% absolute margin was selected for non-inferiority and superiority comparisons, using a statistical significance threshold of 0.01. The SciPy and NumPy libraries in python were used for all function commands.

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 detailed summary 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. Pritchard, J. F. Making Better Drugs: Decision Gates in Non-Clinical Drug Development. *Nat. Rev. Drug Discov* **2**, 542–553 (2003).

21. Seyhan, A. A. Lost in translation: the valley of death across preclinical and clinical divide - identification of problems and overcoming obstacles. *Transl. Med. Commun* **4**, 18–18 (2019).

22. Hingorani, A. D. Improving the odds of drug development success through human genomics: modelling study. *Sci. Rep* **9**, 18911–18911 (2019).

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

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

25. 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).

26. Liu, R. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature*

592, 629–633 (2021).

27. 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).

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

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

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

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

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

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

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

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

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

37. 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).

38. 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).

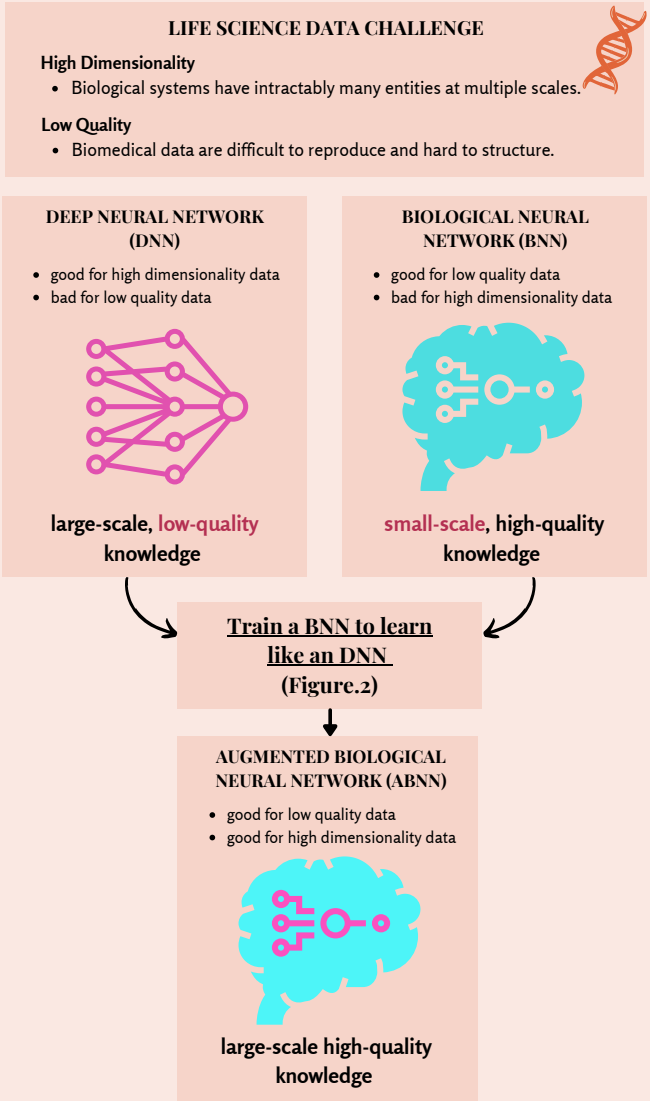
39. Williams, R. J., Tse, T., Dipiazza, K. & Zarin, D. A. Terminated trials in the ClinicalTrials. gov results database: evaluation of availability of primary outcome data and reasons for termination. *PLoS One* **10**, 127242–127242 (2015).

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

41. 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

a Combining The Complementary Advantages of DNN and BNN



b The Training Advantages of ABNN over DNN

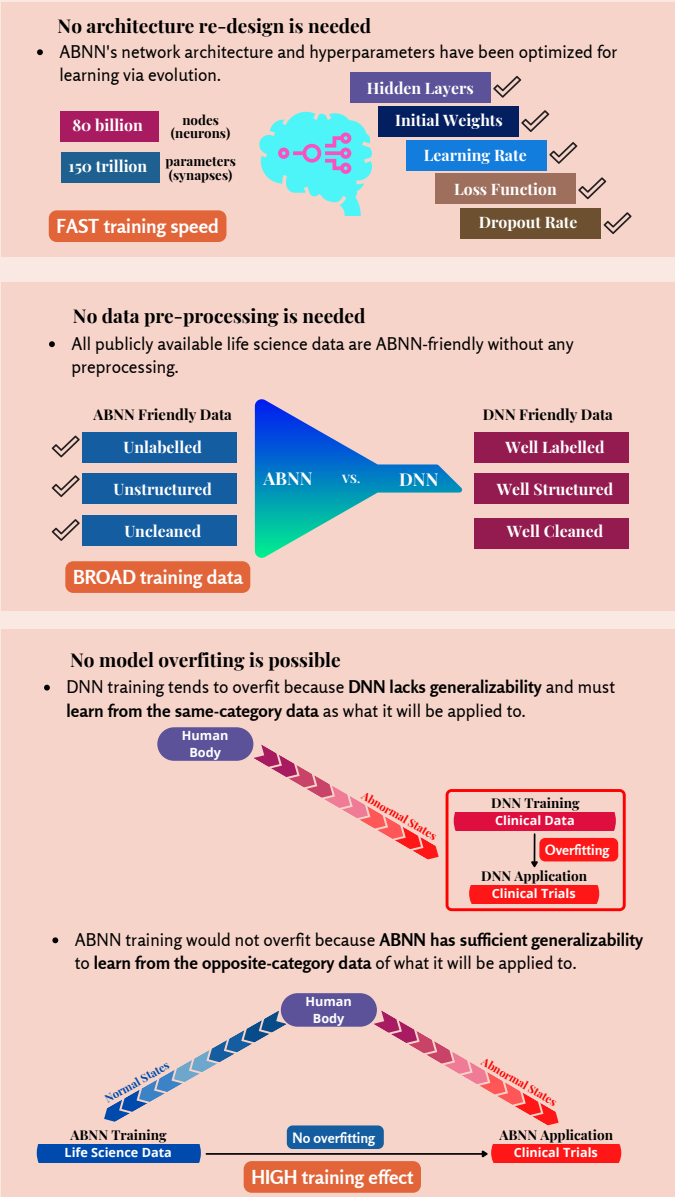


Figure 1a. The challenge and solution of extracting knowledge from life science data. High dimensionality and low quality are two defining characteristics of life science data. Deep neural networks (DNNs) are good at learning knowledge from the high-dimensional space of high-quality data, whereas biological 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 DNNs and BNNs would become good at learning knowledge from the high-dimensional low-quality life science data.

Figure 1b. The training advantages of ABNN over DNN. The training of ABNN is much more efficient than that of DNN because the former spends no time and cost on designing network architecture, tuning network hyperparameters or cleaning training data. Furthermore, the application of ABNN to clinical trials is much more effective than that of DNN because the former has sufficient generalizability to learn from large-scale life science data whereas the latter has limited generalizability to learn from small-scale clinical data at the risk of overfitting.

Figure 2. Train a BNN to learn like an DNN

a Learn from Multiple Levels of Abstraction

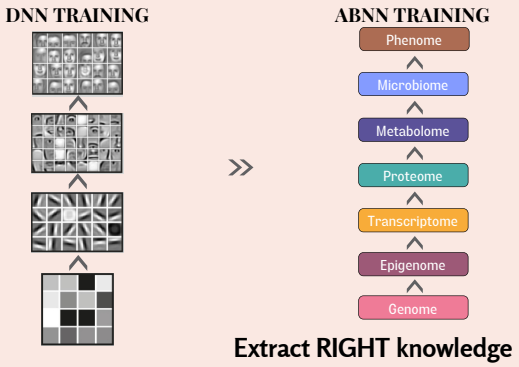


Figure 2a left. DNNs learn better representations of data with more levels of abstraction than with fewer levels and BNNs share the hierarchical architecture of DNNs. **Figure 2a right.** ABNNs extract right knowledge by integrating more multiomics data (e.g. genome, proteome, transcriptome, epigenome,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

b Learn via the Backpropagation Algorithm

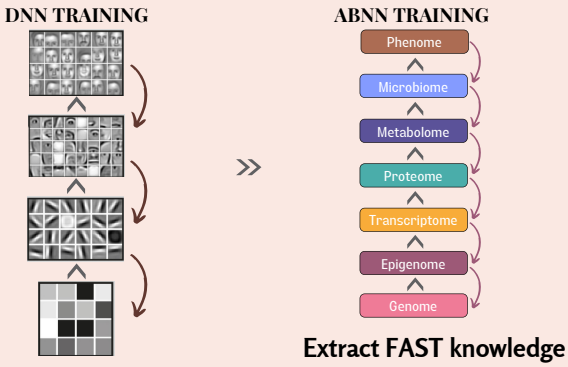


Figure 2b left. DNNs use backpropagation to improve learned representations of data by sequentially updating its internal parameters from the highest to the lowest level of abstraction. The ubiquitous feedback connections in BNNs may implement backpropagation-like learning rules. **Figure 2b right.** 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

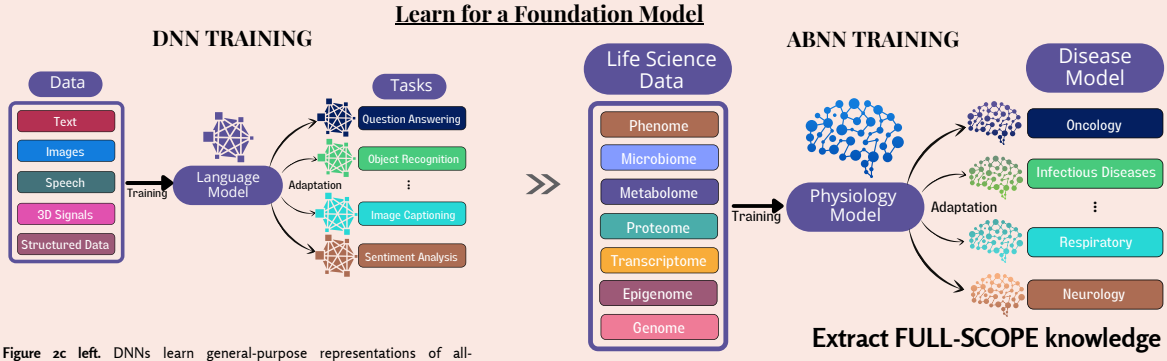


Figure 2c left. DNNs learn general-purpose representations of all-encompassing data to deliver maximal performance in a wide range of special-purpose tasks. Both BNNs and DNNs could learn to command general linguistic abilities for numerous tasks.

Figure 2c right. ABNNs extract full-scope 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

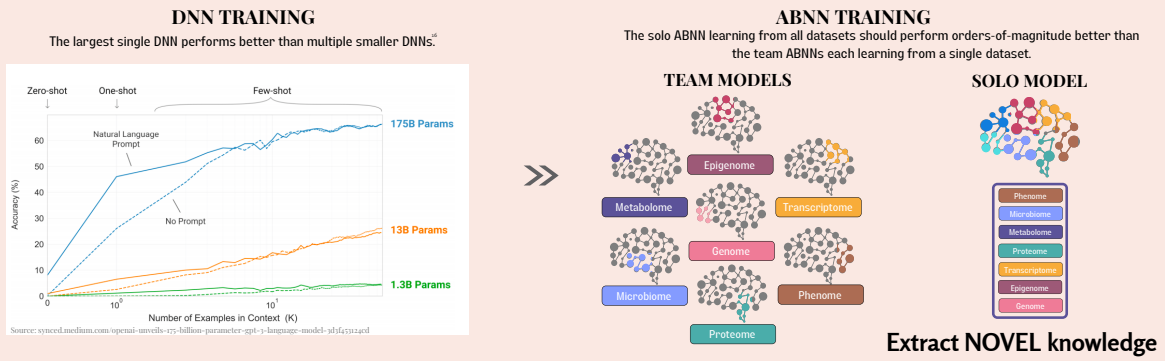


Figure 2d left. DNNs maximize the robustness of learned representations of data by increasing the sheer size of network parameters.

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. DNNs 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 real-world, large-scale, public, prospective, external validation of the trained ABNN

a Real-world ABNN validation and Real-world Industrial Practice in Prospective Predictions of Clinically Successful Drug Candidates Share Identical Process and Measure

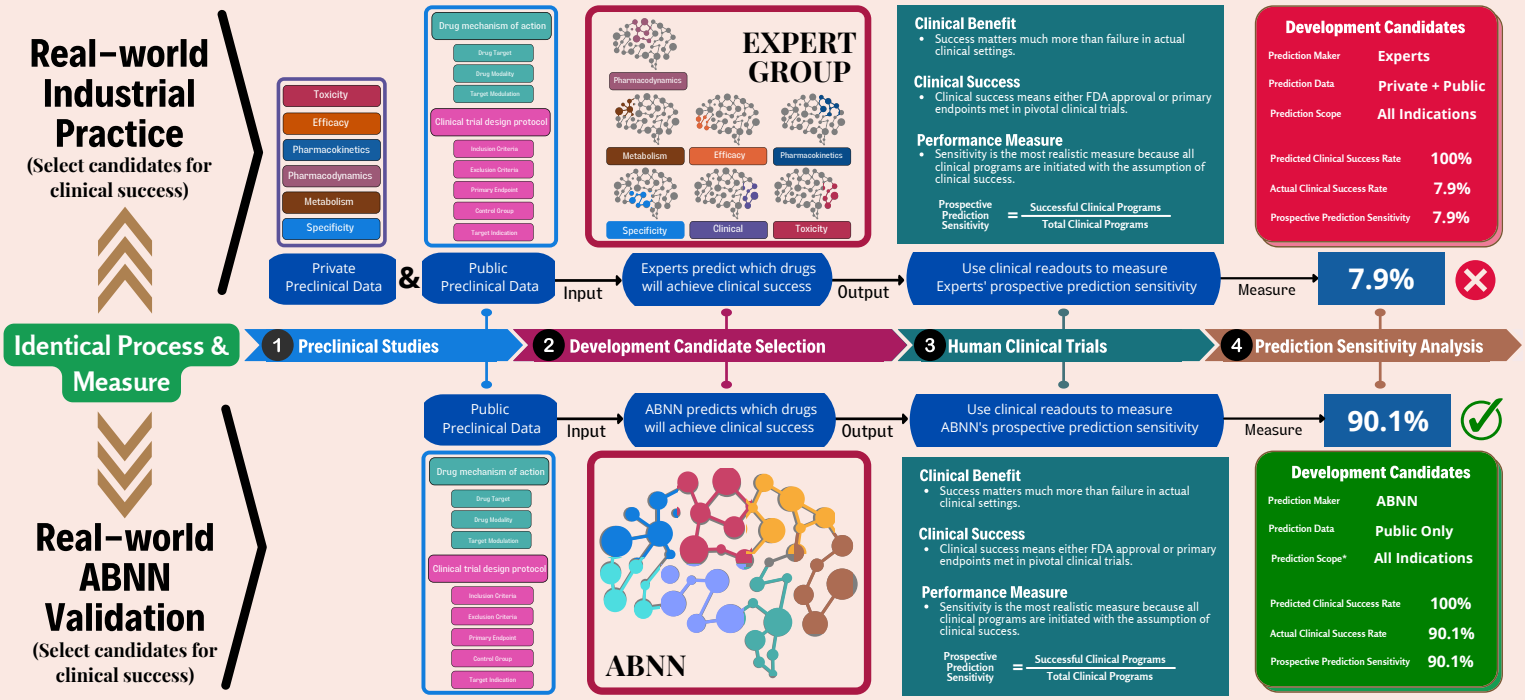


Figure 3a. Real-world validation is critical for an objective assessment of ABNN's potential for improving the efficiency of global drug research and development. As the foundation model is supposed to be able to functionally reconstruct almost all major diseases that could occur to a single human patient, 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. There are several key go/no-go decision steps in drug development as illustrated by Step 1-4 in the figure. In Step 1, preclinical studies aims to decide which drug candidates are developable and generate sufficient data that are best indicative of efficiency and safety in humans for each developable candidate. Biopharmaceutical companies have a considerable data advantage over ABNN in this step because they have exclusive access to internally generated private data in addition to public data, whereas ABNN only has access to public data of drug mechanism of action (drug target, drug modality, and target modulation) and clinical trial design protocol (inclusion criteria, exclusion criteria, primary endpoint, control group, and target indication).

Step 2 is a critical go/no-go decision step in drug development because only a subset of developable drug candidates with highest clinical success potential can be selected to transition from non-clinical to clinical development due to the prohibitive cost of human clinical trials. As the best practice in the biopharmaceutical industry, a group of committed experts use private + public preclinical data to predict which drug candidates will achieve clinical success (Step 2) , initiate clinical trials only for those drug candidates that are prospectively predicted to be clinically successful (Step 3), yet deliver only 7.9% prospective prediction sensitivity (aka. for every 100 clinical programs that experts prospectively predicted to achieve clinical success, only 7.9 of them achieve actual clinical success, Step 4). The real-world PROTOCOLS validation is designed to mirror the real-world industrial practice (Step 2 to Step 4) so as to ensure the transferability of ABNN's performance from validation to application: ABNN used public preclinical data only to predict which drug candidates will achieve clinical success (Step 2) , tracked the readouts of clinical trials only for those drug candidates that are prospectively predicted to be clinically successful (Step 3), and delivered 90.1% prospective prediction sensitivity (aka. for ever 100 clinical programs that were prospectively predicted to achieve clinical success, 90.1 of them achieve actual clinical success, Step 4). *All indications cover all diseases except for hematological disorders, rare neurodevelopmental disorders, and rare musculoskeletal disorders.

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

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. Training | Separate Datasets (External): 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. 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 success 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

Predict outcome	Publish prediction	Check result	Determine ground truth	Validate prediction
T0	T1	T2	T3	T4
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[8] 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

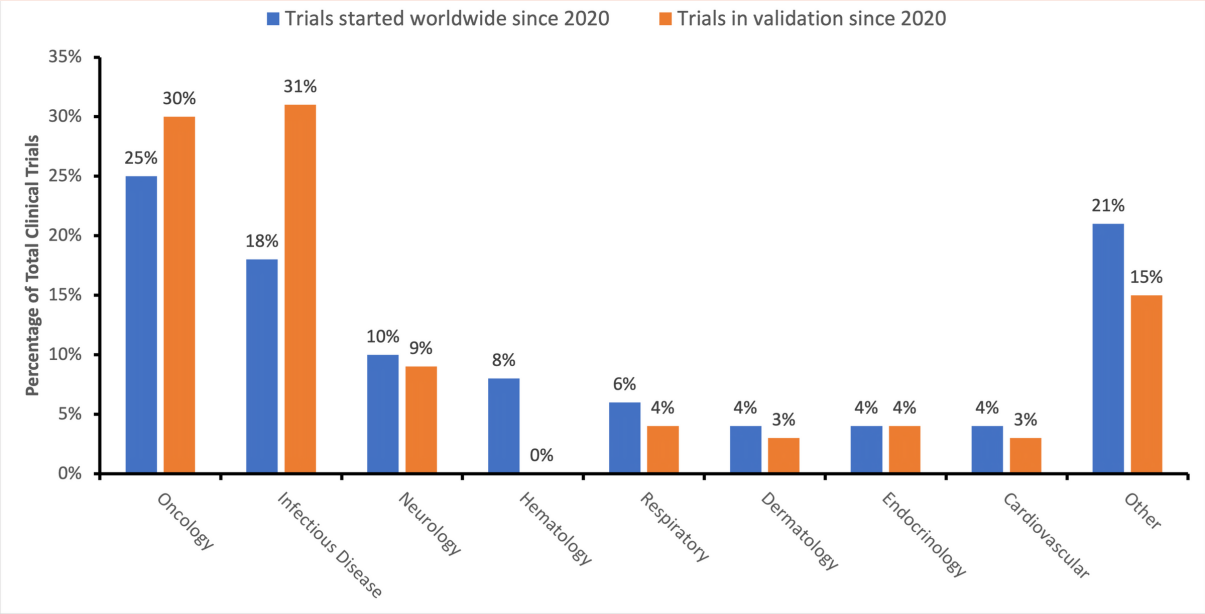


Figure 4. To evaluate the direct translatability between prospective validation and clinical application, a two-sample Kolmogorov-Smirnov test was used as the standard method to show that the sample of validation clinical trials and the sample of initiated clinical trials since 2020 follow identical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

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

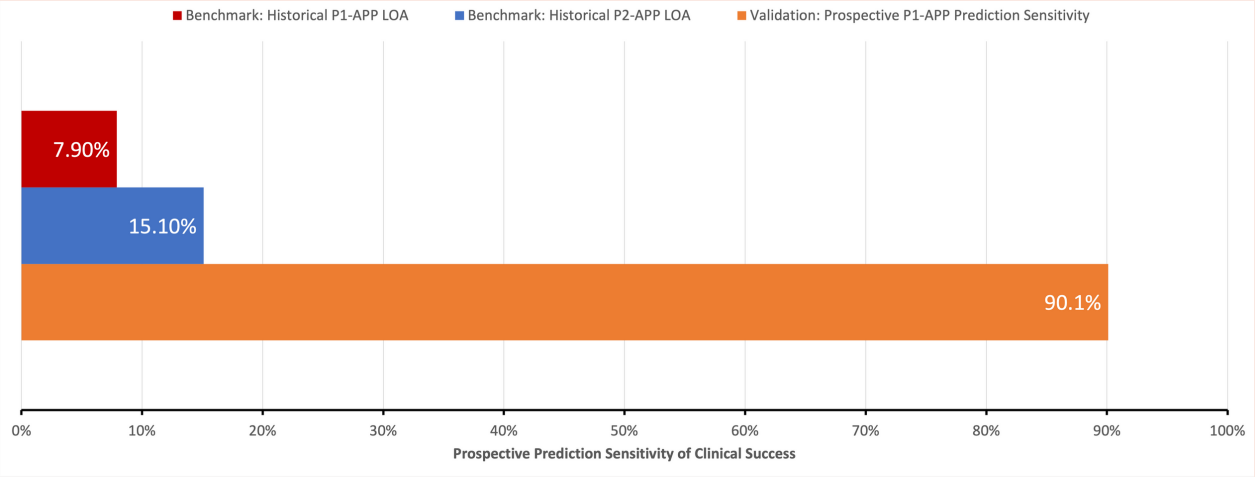


Figure 5a. The PROTOCOLS validation uses a realistic baseline sensitivity of 7.90% (Historical P1-APP LOA: historical likelihood of success from phase 1 to approval) and a conservative baseline sensitivity of 15.1% (Historical P2-APP LOA: historical likelihood of success from phase 2 to approval) to 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

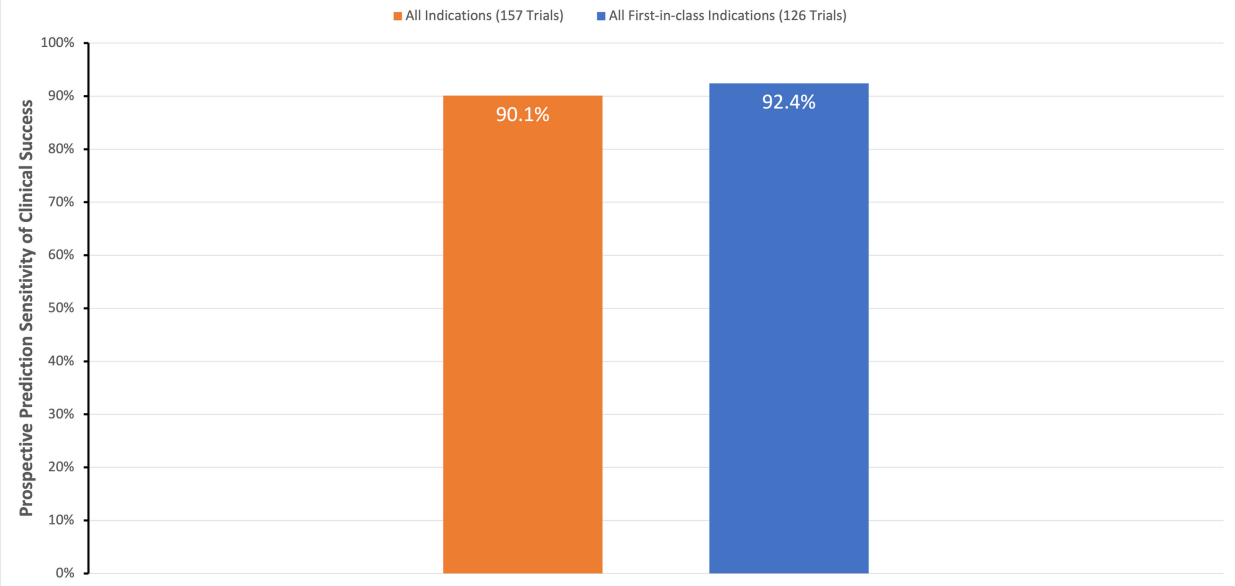


Figure 5b. ABNN's overall prospective prediction accuracy of first-in-class pivotal clinical trials is superior compared with that of all pivotal clinical trials (Δ = 2.3%; 99% CI -0.92%, 5.52%; P < 0.002; 2% margin for non-inferiority) for all therapeutic areas excluding neurodevelopmental, skeletalmusclar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

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

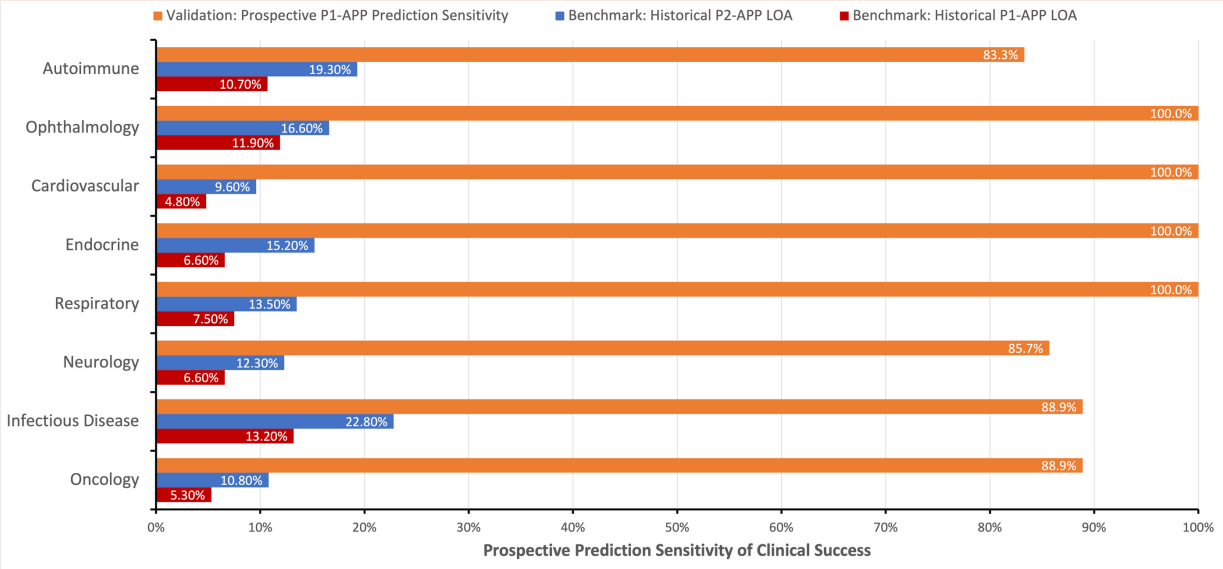


Figure 5c. The PROTOCOLS validation uses a realistic baseline sensitivity (Historical P1-APP LOA: historical likelihood of success from phase 1 to approval) and a conservative baseline sensitivity (Historical P2-APP LOA: historical likelihood of success from phase 2 to approval) to benchmark ABNN's performance (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

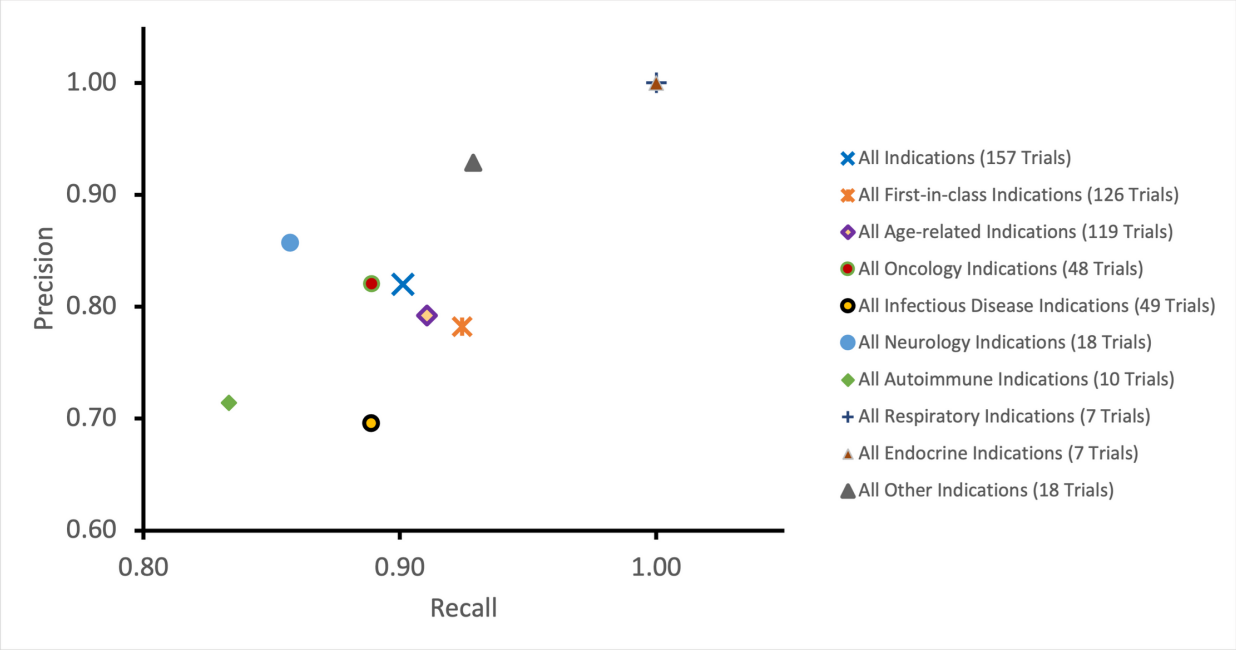


Figure 6a. ABNN's performance in terms of precision and recall (see Data Table 1 for more details).

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

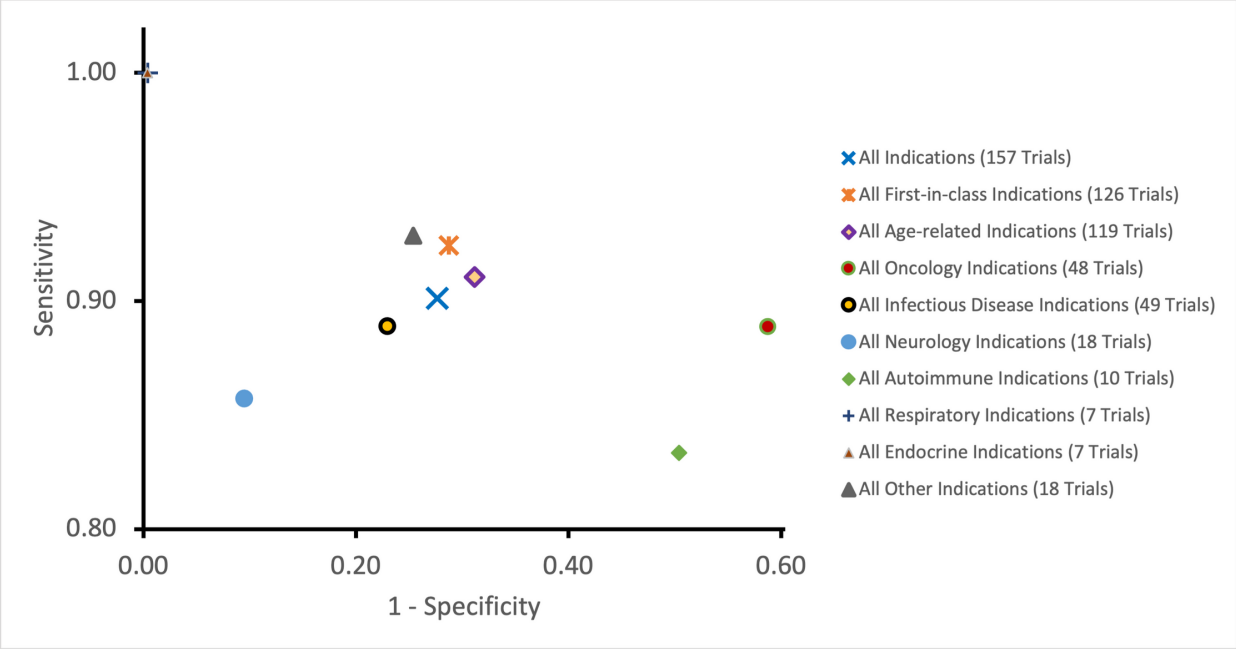


Figure 6b. ABNN's performance in terms of sensitivity and 1 - specificity (see Data Table 1 for more details).

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

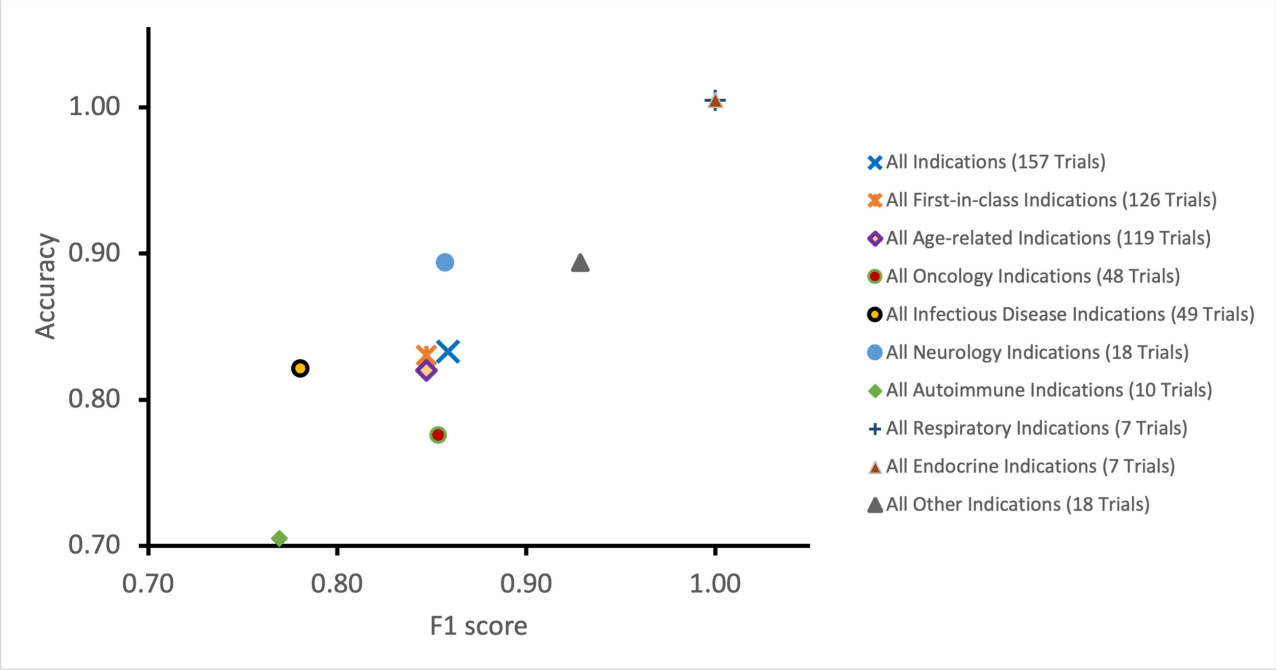


Figure 6c. ABNN's performance in terms of accuracy and F1 score (see Data Table 1 for more details).

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

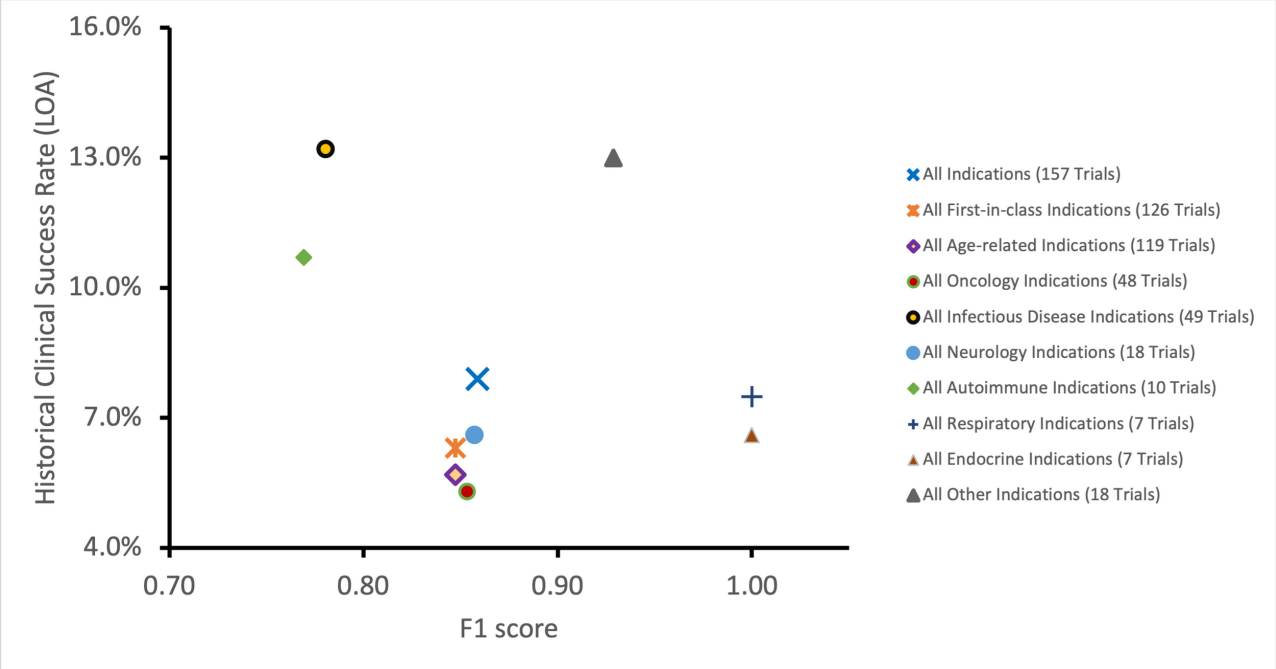


Figure 6d. ABNN's performance in terms of historical clinical success rates for investigation new drugs (LOA) and F1 score (see Data Table 1 for details). LOA is identical to the realistic baseline accuracy (see Methods section 'Baseline performance'). F1 scores and LOAs across overall validation subgroups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

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

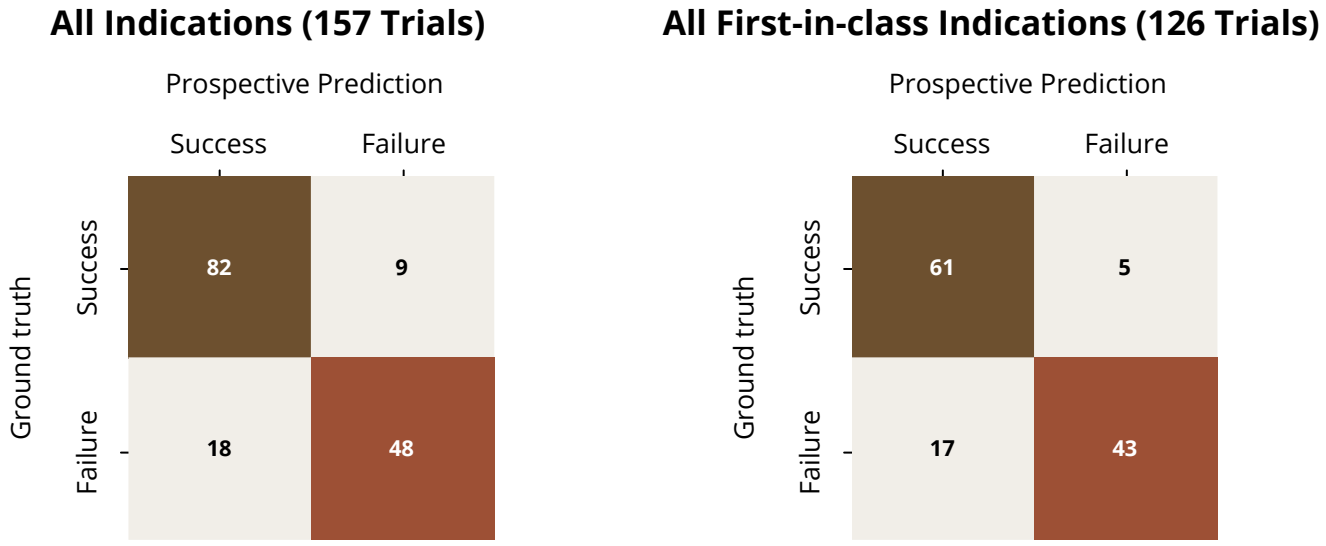


Figure 7a. ABNN's performance in terms of confusion matrices for all therapeutic areas.

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

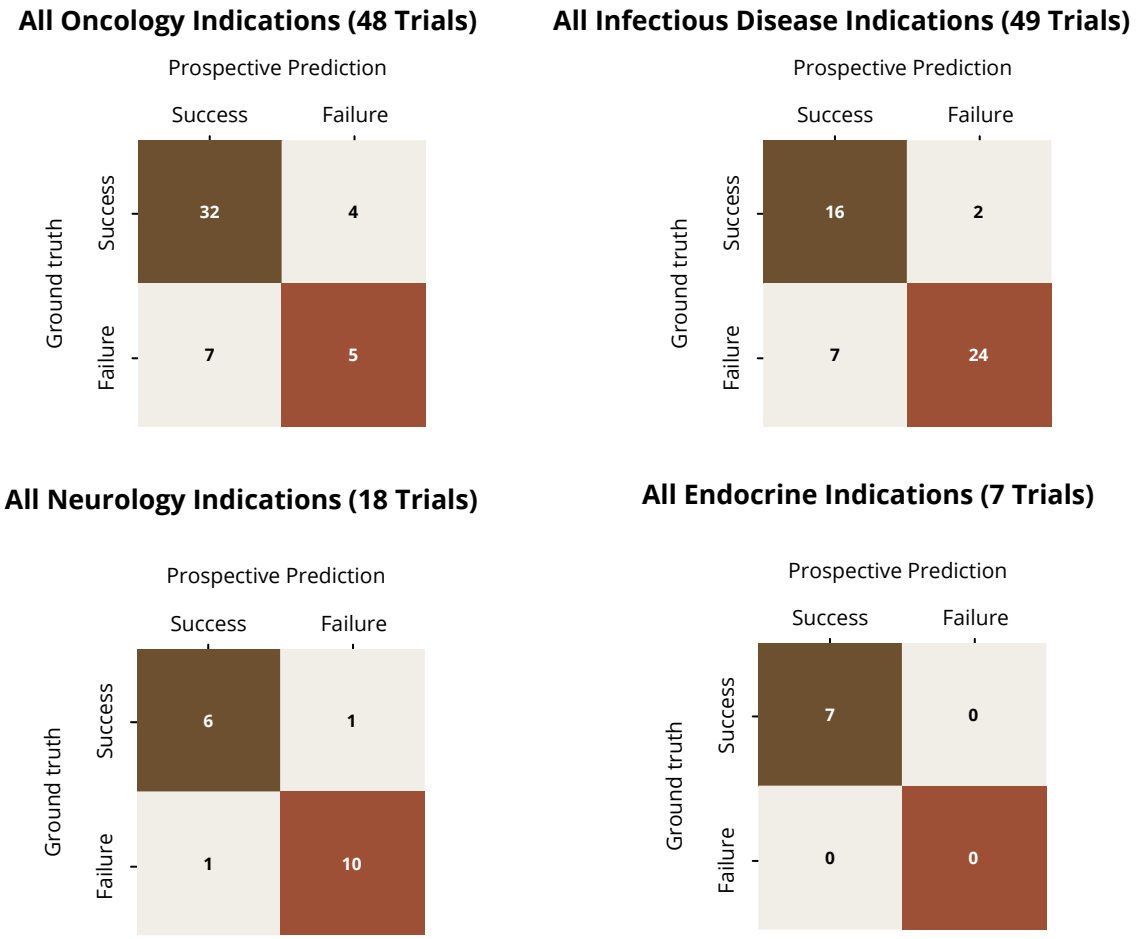


Figure 7b. ABNN's performance in terms of confusion matrices per therapeutic area.

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

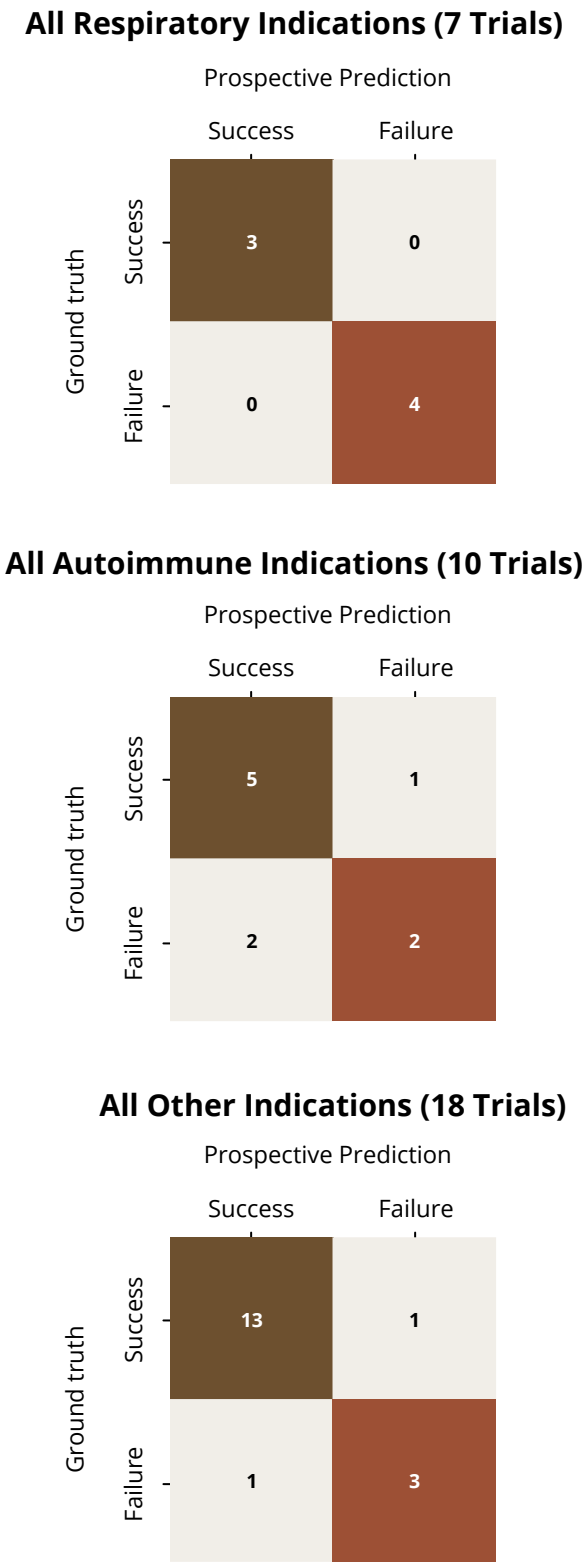


Figure 7c. ABNN's performance in terms of confusion matrices per therapeutic area.

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

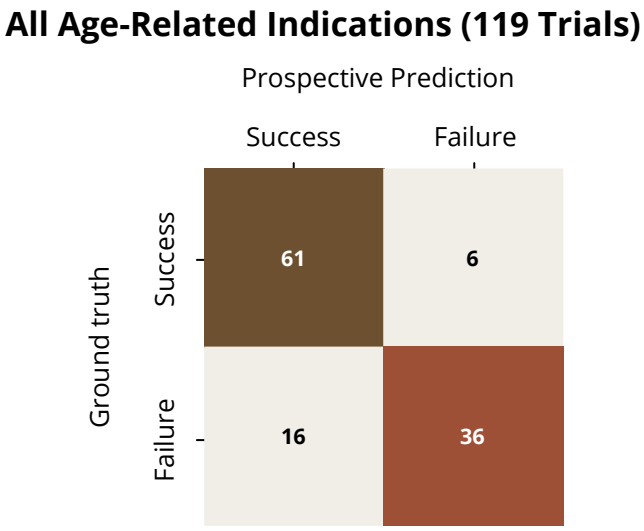


Figure 8. ABNN's performance in terms of confusion matrices for all age-related diseases.

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

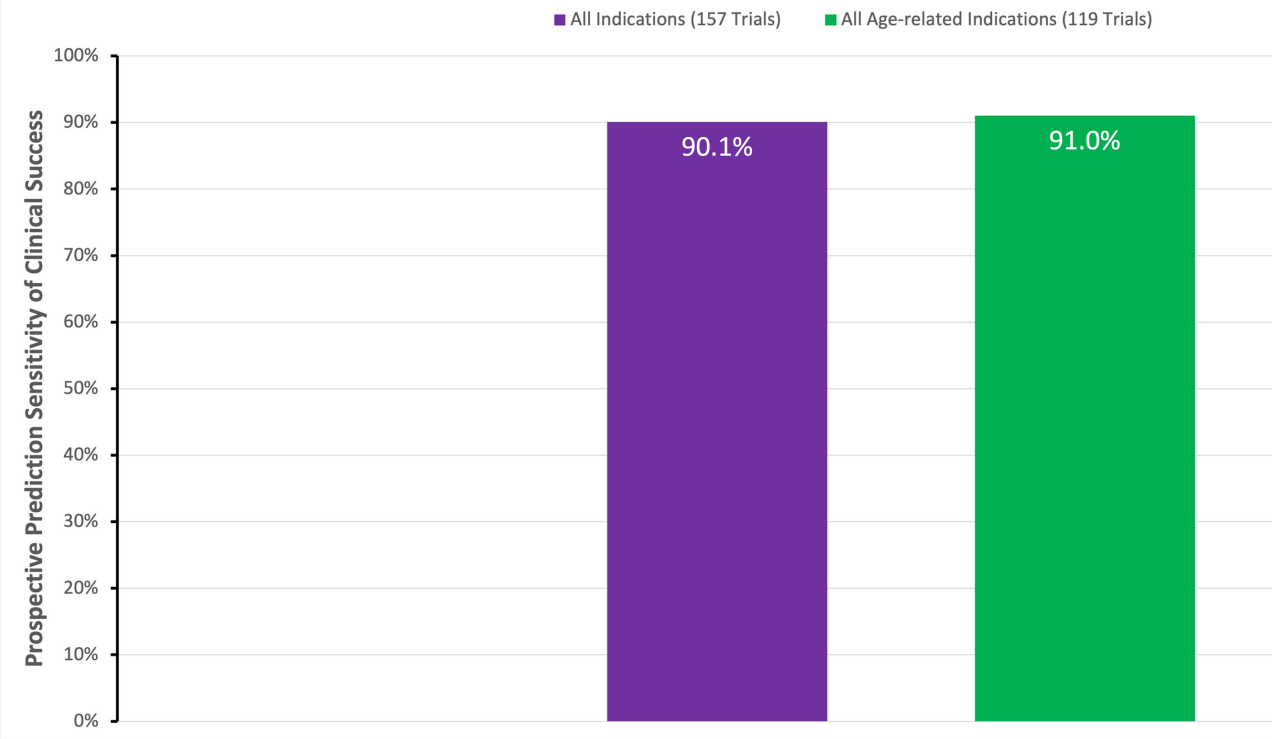


Figure 9a. ABNN's overall prospective prediction sensitivity of all age-related pivotal clinical trials is 91.0% (99% CI 79.1%, 98.4%; P < 0.005) in the PROTOCOLS validation (Methods section 'Statistical analysis').

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

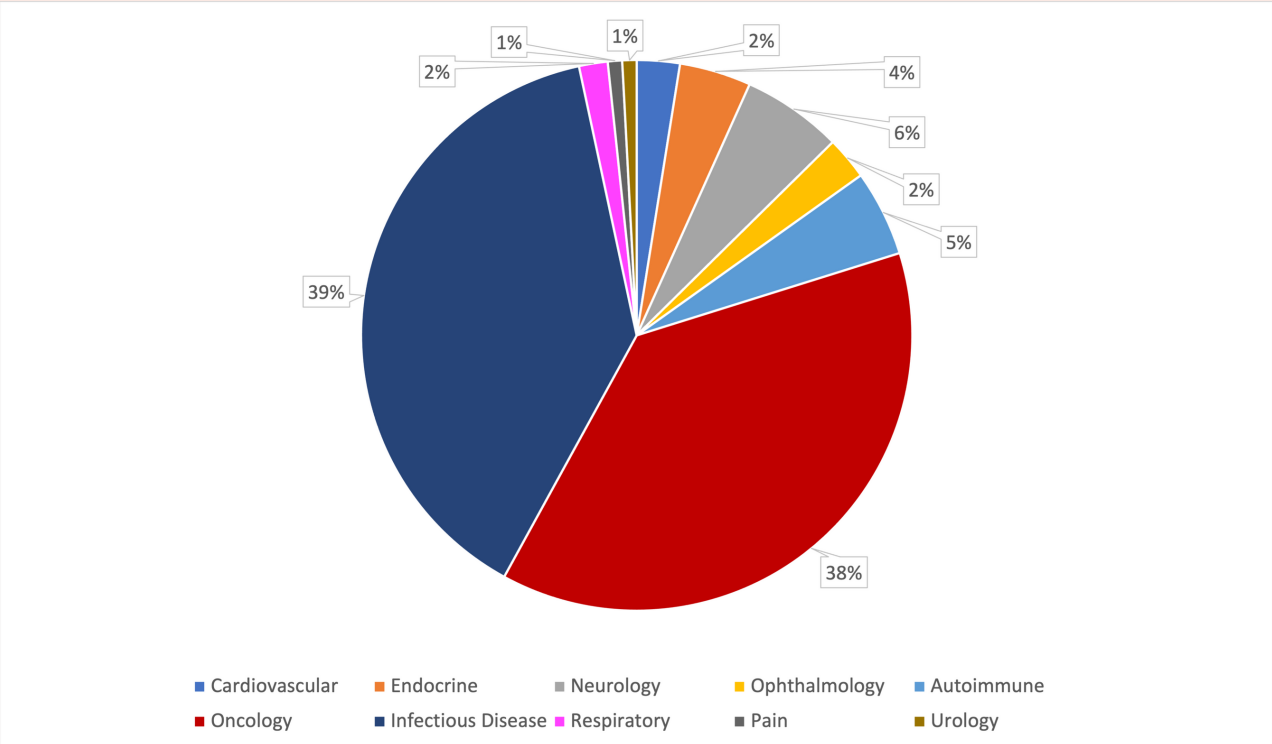


Figure 9b. The distribution of all age-related pivotal clinical trials per therapeutic area in the PROTOCOLS validation (Methods section 'Validation data collection').

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

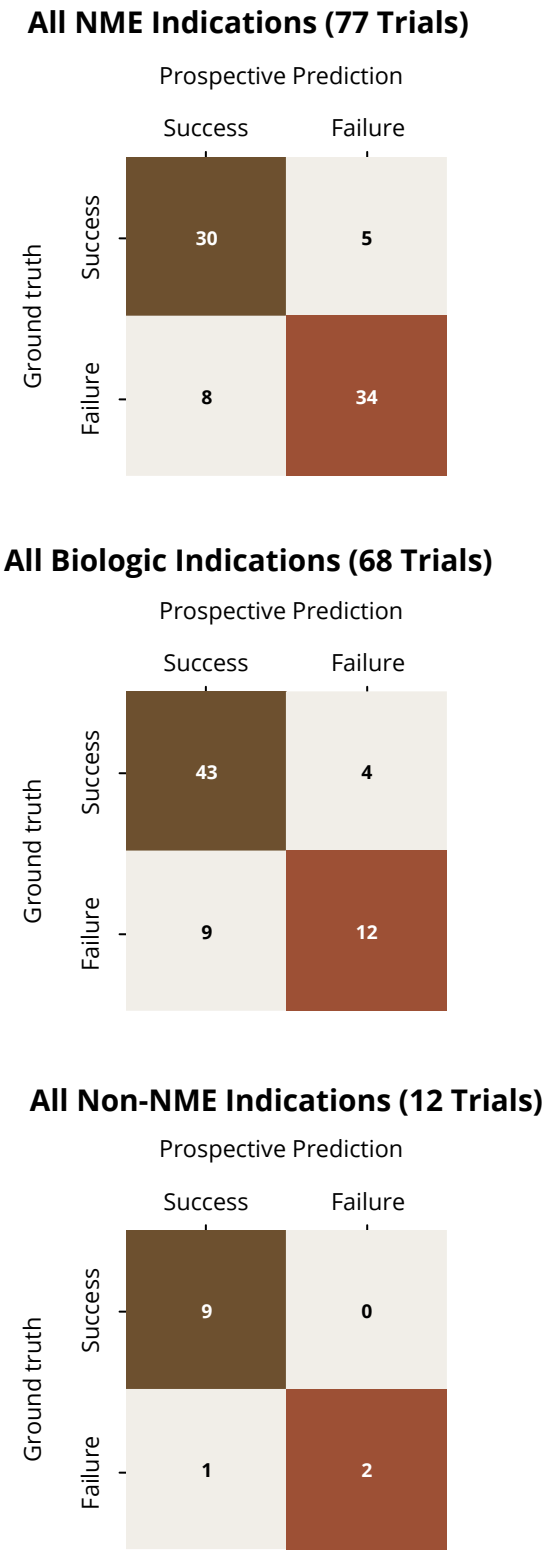


Figure 7d. ABNN's performance in terms of confusion matrices per drug modality.

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

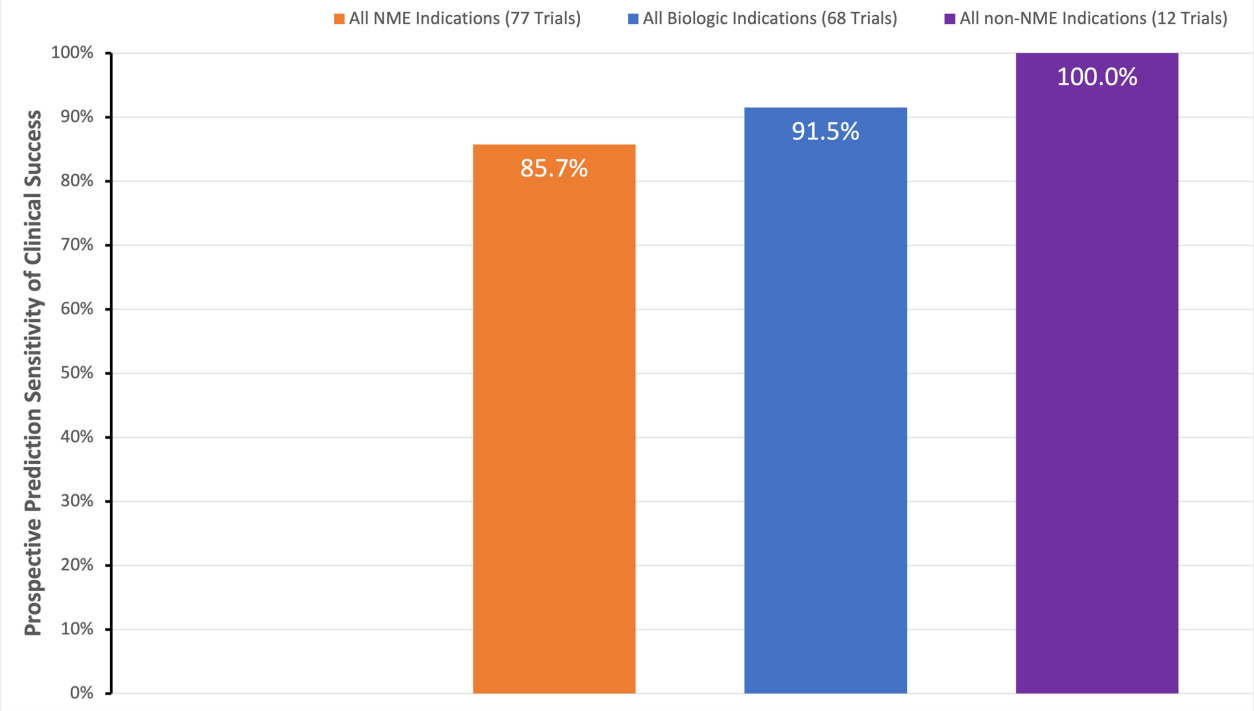


Figure 11a. ABNN's performance in terms of prospective prediction sensitivity across drug modalities (see Data Table 1 for more details). ABNN's achieved superior performance for biologics than for small molecules ((Δ = 5.8%; 90% CI 0.78%, 10.82%; P < 0.0001 for superiority at a 2% margin; Methods section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

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

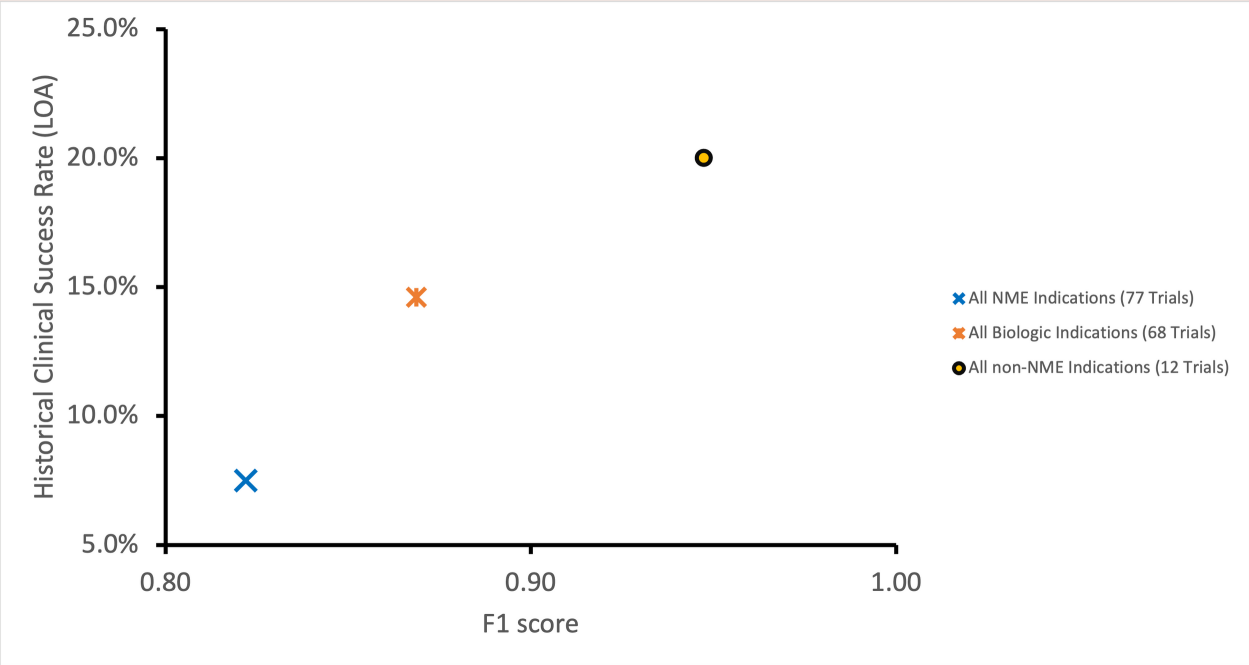


Figure 11b. ABNN's performance in terms of historical clinical success rates for investigation new drugs (LOA) and F1 score (see Data Table 1 for details). LOA is identical to the realistic baseline accuracy (see Methods section 'Baseline performance'). F1 scores and LOAs across overall validation subgroups 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	SEN	F1	ACC	SPE	NPV	PPV	FOR	FDR	FPR	FNR	PLR	NLR	DOR	MCC
All Indications	157	**90.1%**	0.86	82.8%	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	**92.4%**	0.85	82.5%	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	**91.0%**	0.85	81.5%	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	**88.9%**	0.78	81.6%	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	**88.9%**	0.85	77.1%	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	**92.9%**	0.93	88.9%	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	**85.7%**	0.86	88.9%	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	**83.3%**	0.77	70.0%	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	**100.0%**	1.00	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	-	0.00	-	1.00
Endocrine Indications	7	**100.0%**	1.00	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	-	0.00	-	1.00
All NME Indications	77	**85.7%**	0.82	83.1%	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	**91.5%**	0.87	80.9%	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. SEN: Sensitivity; F1: F1-score; ACC: Accuracy; 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);