Predicting Clinical Trial Success for Clostridium difficile Infections Based on Preclinical Data

Fangzhou Li[1,2,3†], Jason Youn[1,2,3†], Christian Millsop[1,2], and Ilias Tagkopoulos[1,2,3*]

[1]Department of Computer Science, University of California, Davis, Davis, CA, USA
[2]Genome Center, University of California, Davis, Davis, CA, USA
[3]USDA/NSF AI Institute for Next Generation Food Systems (AIFS), University of California, Davis, Davis, CA, USA
[†]These authors contributed equally to this work.
[*]Correspondence: itagkopoulos@ucdavis.edu
[*]5313 GBSF, 451 Health Sciences Dr., Davis, CA 95616, USA
[*]530-752-4821

Declarations of Interest: None.

Abstract

Preclinical models are ubiquitous and essential for drug discovery, yet our understanding of how well they translate to clinical outcomes is limited. In this study, we investigate the translational success of treatments for *Clostridium difficile* infection from animal models to human patients. Our analysis shows that only 36% of the preclinical and clinical experiment pairs result in translation success. Univariate analysis shows that the recurrent endpoint is correlated with translation failure (SRC=-0.20, *p*-value=1.53 x 10^{-54}), and explainability analysis of multi-variate random forest models shows that both recurrent endpoint and subject age are negative predictors of translation success. We have developed a recommendation system to help plan the right preclinical study given factors such as drug dosage, bacterial dosage, and preclinical/clinical endpoint. With an accuracy of 0.76 (F1 score of 0.71) and by using only 7 features (out of 68 total), the proposed system boosts translational efficiency by 25%. The method presented can extend to any disease and can serve as a preclinical to clinical translation decision support system to accelerate drug discovery and de-risk clinical outcomes.

Keywords: artificial intelligence, machine learning, drug discovery, clinical trial, recommendation system

Introduction

Clostridium difficile is a spore-forming anaerobic bacteria widely distributed in the intestinal tract of humans and animals and in various environmental contexts[1]. Over the past decade, the frequency and severity of *C. difficile* infection (CDI) have been increasing worldwide to become a leading nosocomial (hospital-acquired) pathogen[2]. It is estimated to affect approximately 3 million individuals worldwide every year[3], underscoring its significant public health impact. Although various treatments, such as metronidazole and oral vancomycin[4–6], have been approved for CDI management, the sustained efficacy, the effectiveness of a treatment after the treatment is no longer administered, is low[7]. This is particularly concerning given the recurrent nature of CDI[3,8], where the sustainability of treatment efficacy (the ability to prevent recurrence post-therapy) is crucial.

A predominant challenge in the development of treatments for *Clostridium difficile*, as with many diseases, lies in the limited rate of translational success from preclinical to clinical stages. For example, the chance of a potential drug candidate identified in the preclinical trials demonstrating efficacy in human studies and ultimately receiving approval is a mere 0.1%[9]. Therefore, the development of a new drug is a time-consuming and costly process that often takes an average of 13 years and costs approximately US$1 billion[10] from the preclinical testing stage to FDA approval. The major causes for such translation failures are the lack of appropriate animal models for predicting the efficacy of the drug in humans[9,11], concerns for the efficacy and safety of the drugs[12], poor study design, ineffective site selection, poor recruitment, patient burden, and poor trial execution[13]. Efforts to enhance translational success have included the use of humanized animals, which exhibit more human-like responses to medical interventions[14], and the application of biomarkers to reduce subjectivity in evaluating drug efficacy and safety[15].

Machine learning-based approaches[16–18] have also been explored, predominantly focusing on attrition rates across different phases of clinical trials. However, these approaches often lack explainability due to the ‘black box’ nature of the models employed. Although machine learning models have shown promising results in other areas of life sciences[19–21], their application in bridging the gap between preclinical and clinical outcomes is hindered by a scarcity of expert-curated datasets. This limitation is particularly pronounced in the context of *C. difficile*, where the complexity of the disease and its treatment modalities necessitates highly specialized and accurate data for effective model training and validation.

In this study, we aim to curate an Animal-to-Human (A2H) dataset for *C. difficile* infections that contains preclinical to clinical translation information and use a machine learning-based model to predict the translational success (**Fig. 1a**). We expand this predictor to a recommendation system that provides an explanation of what parameters are important and why (**Fig. 1b**).

Results

Experimental features correlated to preclinical to clinical translation success. **Figure 2a** depicts the spectral biclustering[22] of the 5,851 preclinical-clinical pair samples, excluding control intervention and inconvertible unit samples from the original 6,918, across the 68 features after performing one-hot encoding to the original 42 variables. From top to bottom, the fourth and sixth row clusters were associated with the lowest and highest average translation success rates of 0.25 and 0.45, respectively. We found that the row cluster with the lowest average translation success rate differentiated from other clusters due to its unique disease model, which challenged first and then treated animals with clindamycin (*p*-value<4.77 x 10^{-122}). Similarly, the cluster with the highest average translation success rate had adopted *C. difficile* strains (e.g., VA11, 2926, VA5, TTU 614) that were significantly different from those used in other clusters (*p*-value<7.2 x 10^{-13}). A t-SNE plot for the A2H dataset can also be found in **Supplementary Figure 2**. The distribution of success metrics in both preclinical and clinical trials, specifically focusing on the survival and recovery rates, respectively, are shown in **Fig. 2b-c**. These rates are skewed towards the right, partly due to the use of existing drugs like vancomycin and metronidazole as controls in case-control studies[23]. Delta (δ), the difference between the recovery rate and survival rate used to assign the target variable (translation success/failure) (see **Methods**), was modeled using a normal distribution as $\delta \sim N(0.09, 0.41)$ (**Fig. 2d**). We labeled the preclinical and clinical trial pairs (see **Methods**) that fell within ±0.5 standard deviation of $\delta = 0.0$ as ‘translation success.’ (3,746 samples) and ‘translation failure’ otherwise (2,105 samples) (**Fig. 2d**). Spearman correlation coefficient of the features with the dependent variable lists 8 preclinical features as the top 10 most correlated features (**Fig. 2e**), amongst which sustained endpoint of the clinical trial was most negatively correlated to translation success (SRC=-0.20, *p*-value=1.53 x 10^{-54}).

Machine learning models accurately predict translation success. We implemented the model selection pipeline on A2H datasets created using different translation thresholds c (0.0625, 0.125, 0.25, 0.5, 1.0, and 2.0) (see **Methods**). In every scenario during the cross-validation process, the random forest model emerged as the top-performing classifier (**Supplementary Table 1**). Notably, we observed an improvement of the F1 score when applying SMOTE, especially for thresholds defined by smaller c (F1 improved 126.3%, 124.9%, 39.3%, and 2.7% for c of 0.0625, 0.125, 0.25, 0.5, respective; **Supplementary Figure 3**). Running the sequential feature selection[24] (SFS) in a parsimonious setting (smallest feature subset that is within one standard error of the best cross-validation F1 score) on the best pipeline with $c = 0.5$ (FS: none, MVI: Simple, OS: SMOTE, CLS: random forest) significantly reduced the required number of features by 76.5% from 68 to 16 with negligible 0.8% performance loss (validation set F1 score decrease from 0.75 to 0.74) as shown in **Fig. 3a**, where the results for other value of c can be found in **Supplementary Figure 4**. Moreover, we were able to achieve validation set F1 score of 0.73 with only 7 features identified using the Kneedle elbow method[25] (**Fig. 3a**). **Table 1** further shows the holdout test set performance for different numbers of features for the best model. The best model pipeline for the benchmark A2H dataset ($c = 0.5$) on the holdout test set achieved a 25% better F1 score than a random baseline (0.69 vs. 0.56, respectively), while AUCPR and AUCROC were 0.68 and 0.82, respectively (**Fig. 3b-d**). For all six different translation thresholds c except when $c = 2.0$, we had better performance than the random baseline (**Fig. 3c** and **Supplementary Table 2**).

Recurrent endpoint and subject age as predictors of translation success. We analyzed the feature importance of the best model for each c using five ranking methods: sequential feature selection, linear discriminant analysis (LDA), Pearson correlation coefficient (PCC), impurity-based feature importance of random forest (RF), and SHAP as shown in **Figure 4a** for c = 0.5. Of the 16 features selected by SFS, only three were from clinical features. The five ranking methods consensually identified whether clinical and preclinical endpoints were sustained or acute as most influential to the translation prediction (mean rank = 1 and 2.2). We found that RF and SHAP could highlight the importance of dosage-relevant features, while linear methods like LDA and PCC could not. A further investigation of SHAP values provided more detailed insights into the relationship between feature values and their impact on predictions. Specifically, the model considered sustained preclinical and clinical endpoints would decrease the translation success probability (mean SHAP value = -0.14 for both). This observation can be explained by the significantly lower translation success for samples with sustained preclinical and clinical endpoints compared to those with at least one acute endpoint (*p*-value = 3.3×10^{-10}). The model also considered younger subjects for both animals and humans would be more likely to result in translation failure, with the animal age being highly correlated with the SHAP value (p-value = 3.9×10^{-298}), with a smaller animal age value resulting in a more negative impact on translation success probability. Also, for the human subjects, the SHAP value of the child age group was significantly smaller than the more-aged group (mean SHAP value = 0.01 vs. -0.18; *p*-value = 8.7×10^{-164}). The SHAP performance for other c can be found in **Supplementary Figure 5**.

Discussion

Our research underscores the importance of refined preclinical strategies in drug development, a principle that holds true across various medical fields. The necessity for improved preclinical approaches, as indicated by the frequent phase III failures due to a lack of responder hypothesis-based trials[26], aligns with our findings, where a machine learning model driven by a selective feature set significantly enhanced the predictability of translational success.

Our choice to focus on *C. difficile* in this study stems from several key considerations. Firstly, the existence of well-established rodent models for *C. difficile* infection closely mimics the human disease and provides a robust basis for preclinical studies, therefore allowing for more accurate predictions of clinical outcomes. Additionally, the pressing need for improved treatment strategies for *C. difficile* infections, given their increasing prevalence and public health impact, underscores the practical significance of our research. Furthermore, the localized nature of *C. difficile* infections in the gut[27], as opposed to systemic diseases, presents a unique opportunity. It allows for more controlled study parameters and a clearer understanding of treatment effects, which are critical for the successful application of machine learning techniques in predicting translational outcomes. This aspect is particularly vital in lightening the complexity that often accompanies the study of systemic diseases, where multiple organ systems and a myriad of physiological factors can confound results[28].

There are a few areas of improvement. First, we assumed a direct and linear relationship between preclinical survival rates and clinical recovery rates. Yet, it is important to acknowledge that these metrics, while informative, may not fully capture the multifaceted nature of trial outcomes. Future studies could benefit from incorporating additional parameters, such as percent weight change for preclinical and quality-of-life assessments for clinical trials, to provide a more comprehensive evaluation of trial success. Second, rather than employing delta δ, which represents the difference between survival and recovery rates in current work, an alternative approach could involve using a ratio of these rates. This change would be significant because a 10% difference in lower rates has different implications compared to a 10% difference in higher rates. Third, the primary challenge of the data curation process is its dependency on expert-guided manual data curation. Implementing an automated data extraction pipeline, leveraging transformer-based large language models (LLMs)[29–31], would significantly enhance the efficiency of extracting data from existing literature. This enhancement would be beneficial not only for *C. difficile* infection but also for a broader range of bacterial diseases, such as streptococcal infections, tuberculosis, and salmonellosis. By creating a dataset enriched with multi-omics information for these diverse diseases, we can develop a more generalizable ML-based predictor that demonstrates higher performance. Additionally, this enriched dataset would facilitate intra-clinical predictions, such as forecasting the outcomes of clinical trial phase 2 based on phase 1 data.

Conclusion

This study aims to help translate preclinical findings to clinical outcomes for *Clostridium difficile* infections, leveraging machine learning to enhance predictive accuracy and interpretability. Our model identifies key factors influencing translational success, streamlining drug development for CDI and potentially other diseases. This approach not only promises more effective treatments but also exemplifies the transformative impact of integrating computational methods in modern medicine, paving the way for advancements in personalized healthcare.

Acknowledgments

We would like to thank the members of the Tagkopoulos lab for the helpful discussions and comments. This work was supported by the United States Department of Agriculture-National Institute of Food and Agriculture AI Institute for Next Generation Food Systems (AIFS), USDA-NIFA award number 2020-67021-32855.

Author Contributions

C.M. collected the data and performed initial analysis. F.L. and J.Y. performed all computation analyses and created the figures. J.Y., F.L., and I.T. contributed to the critical analysis and wrote the paper. I.T. conceived and supervised all aspects of the project.

Conflict of Interest Statement

The authors declare no competing interests.

Materials and methods

Raw data acquisition. Clinical trial data about *C. difficile* infection (CDI) were collected from *ClinicalTrials.gov*, a comprehensive database of privately and publicly funded clinical studies. This study focused exclusively on completed interventional clinical trials that have published results to ensure the reliability and validity of the data. Parallel to clinical trial data collection, a thorough search was conducted on PubMed to identify publications that tested the same intervention (i.e., drug candidate) in an animal model as one of the clinical trials in our collection. Note that within the scope of a single trial, multiple experimental arms may be present, each contributing to the collective dataset. Here, an ‘arm’ is delineated as a cohort or subset of subjects receiving a particular therapeutic regimen[32,33]. For instance, if a trial investigates two distinct dosages of treatment, each dosage arm is a cohort that can have multiple individuals (or samples; animals for preclinical and humans for clinical studies, respectively). This resulted in a preclinical dataset of 480 arms from 43 preclinical trials, collectively consisting of 3 animal species, 60 interventions (drug candidates), and 29 variables. Similarly, the clinical dataset has 158 arms from 52 clinical trials, collectively consisting of 53 interventions (drug candidates) and 21 total features.

Data compendium. While there were other measures of success, such as changes in body weight and the number of *C. difficile* spores, we focused primarily on the survival rate and recovery rate of the preclinical and clinical trials, respectively. These two rates share similarities in reflecting patient outcomes, making them more directly comparable and relevant for evaluating the effectiveness of treatments in both preclinical and clinical trials.

For preclinical studies, 480 arms with 27 features were gathered, including animal strain, sex, age, weight, specifics of the disease model, dosing, and duration information. For clinical studies, 272 arms with 15 features were collected, encompassing aspects like dosage details, intervention class, therapeutic approach, and participant demographics. We then paired the preclinical and clinical trial arms that tested the same intervention (drug candidate) to construct an A2H dataset, which consists of 6,918 samples and 42 variables (27 from preclinical trials and 15 from clinical trials) after data cleaning and dropping 8 features from the raw datasets (**Supplementary Information 1.1**). To analytically assign the binary dependent variable, we first calculated the difference between recovery and survival rates, denoted as δ, for each sample in the paired dataset as follows:

$$-1.0 \leq \delta = r_r - r_s \leq 1.0, \qquad \text{Equation 1}$$

where $0.0 \leq r_s \leq 1.0$ is the survival rate for animal subjects in the preclinical study, and $0.0 \leq r_r \leq 1.0$ is the recovery rate for human subjects in the clinical study. We then fit a normal distribution to these deltas as

$$\delta \sim N(\mu, \sigma), \qquad \text{Equation 2}$$

where mean (μ) and standard deviation (σ) estimate the standard distribution of δ. We assigned the binary label as follows:

$$\begin{array}{c} 1\ (translation\ success)\text{: } |\delta| < c * \sigma, \\ 0\ (translation\ failure)\text{: } |\delta| \geq c * \sigma, \end{array} \qquad \text{Equation 3}$$

where c is a coefficient to control the strictness of the translation success. We visualized our performance statistics using the A2H dataset with labels assigned with $c = 0.5$. However, different choices of c were also analyzed and reported (**Supplementary Figure 1**).

Model selection. To find the most predictive machine learning model for our preclinical-to-clinical translation, we implemented a model selection pipeline that chooses the best data preprocessing combination and classifier. The pipeline includes, in the order specified, 2 feature scaling (Standard[34] and MinMax[34]), 1 missing value imputation (MVI) method (Simple[34]), 1 oversampling (OS) (SMOTE[35]), and 3 classifiers (CLS) (random forests[36], AdaBoost[37], MLP[38]). We rigorously tested each possible permutation of these preprocessing steps combined with a classifier using a 5-fold cross-validation approach to ensure robust evaluation, where each split was stratified, and samples from the same preclinical and clinical pair were grouped while splitting. Moreover, a grid search was performed on the classifiers to find the optimal hyperparameters using the validation set. Ultimately, the model candidate with the highest F1 score was selected as the best model.

Model interpretability. To increase the interpretability of the model, we applied the Shapley Additive Explanations (SHAP)[39] algorithm. The greater the magnitude of the SHAP value of a feature, the more influence that feature has on the model output. SHAP can provide the local explanation for each sample and the global explanation for an entire class by summarizing the overall importance of features across all data points. In this study, we used SHAP to analyze features that are influential in general to determine translation success.

References

1 Smits WK, Lyras D, Lacy DB, *et al.* Clostridium difficile infection. *Nat Rev Dis Primer*. 2016;2:1–20.

2 Czepiel J, Dróżdż M, Pituch H, *et al.* Clostridium difficile infection. *Eur J Clin Microbiol Infect Dis*. 2019;38:1211–21.

3 Cole SA, Stahl TJ. Persistent and recurrent Clostridium difficile colitis. *Clin Colon Rectal Surg*. 2015;28:65–9.

4 Zar FA, Bakkanagari SR, Moorthi K, *et al.* A comparison of vancomycin and metronidazole for the treatment of Clostridium difficile--associated diarrhea, stratified by disease severity. *Clin Infect Dis*. 2007;45:302–7.

5 Johnson S, Louie TJ, Gerding DN, *et al.* Vancomycin, metronidazole, or tolevamer for Clostridium difficile infection: results from two multinational, randomized, controlled trials. *Clin Infect Dis*. 2014;59:345–54.

6 Teasley D, Olson M, Gebhard R, *et al.* Prospective randomised trial of metronidazole versus vancomycin for Clostridium-difficile-associated diarrhoea and colitis. *The Lancet*. 1983;322:1043–6.

7 Van Giau V, Lee H, An SSA, *et al.* Recent advances in the treatment of C. difficile using biotherapeutic agents. *Infect Drug Resist*. 2019;12:1597.

8 McFarland LV, Surawicz CM, Rubin M, *et al.* Recurrent Clostridium Difficile Disease: Epidemiology and Clinical Characteristics. *Infect Control Hosp Epidemiol*. 1999;20:43–50.

9 Seyhan AA. Lost in translation: the valley of death across preclinical and clinical divide--identification of problems and overcoming obstacles. *Transl Med Commun*. 2019;4:1–19.

10 Ciociola AA, Cohen LB, Kulkarni P, *et al.* How drugs are developed and approved by the FDA: current process and future directions. *Off J Am Coll Gastroenterol ACG*. 2014;109:620–3.

11 Paul SM, Mytelka DS, Dunwiddie CT, *et al.* How to improve R&D productivity: the pharmaceutical industry’s grand challenge. *Nat Rev Drug Discov*. 2010;9:203–14.

12 Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*. 2004;3:711–6.

13 Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp Clin Trials Commun*. 2018;11:156–64.

14 Shultz LD, Ishikawa F, Greiner DL. Humanized mice in translational biomedical research. *Nat Rev Immunol*. 2007;7:118–30.

15 Yu D. Translational research: current status, challenges and future strategies. *Am J Transl Res*. 2011;3:422.

16 Shah P, Kendall F, Khozin S, *et al.* Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ Digit Med*. 2019;2:1–5.

17 Toh TS, Dondelinger F, Wang D. Looking beyond the hype: applied AI and machine learning in translational medicine. *EBioMedicine*. 2019;47:607–15.

18 Gayvert KM, Madhukar NS, Elemento O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem Biol*. 2016;23:1294–301.

19 Lysenko A, Sharma A, Boroevich KA, *et al.* An integrative machine learning approach for prediction of toxicity-related drug safety. *Life Sci Alliance*. 2018;1.

20 Wang W, Kiik M, Peek N, *et al.* A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PloS One*. 2020;15:e0234722.

21 Eetemadi A, Tagkopoulos I. Genetic Neural Networks: an artificial neural network architecture for capturing gene expression relationships. *Bioinformatics*. 2019;35:2226–34.

22 Kluger Y, Basri R, Chang JT, *et al.* Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res*. 2003;13:703–16.

23 Kaye KS, Harris AD, Samore M, *et al.* The Case-Case-Control Study Design: Addressing the Limitations of Risk Factor Studies for Antimicrobial Resistance. *Infect Control Hosp Epidemiol*. 2005;26:346–51.

24 Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J Open Source Softw*. 2018;3:638.

25 Satopaa V, Albrecht J, Irwin D, *et al.* Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. *2011 31st International Conference on Distributed Computing Systems Workshops*. 2011:166–71. https://doi.org/10.1109/ICDCSW.2011.20

26 Sun D, Gao W, Hu H, *et al.* Why 90% of clinical drug development fails and how to improve it? *Acta Pharm Sin B*. 2022;12:3049–62.

27 Best EL, Freeman J, Wilcox MH. Models for the study of Clostridium difficile infection. *Gut Microbes*. 2012;3:145–67.

28 Manor B, Lipsitz LA. Physiologic complexity and aging: Implications for physical function and rehabilitation. *Prog Neuropsychopharmacol Biol Psychiatry*. 2013;45:287–93.

29 Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2019;btz682.

30 Vaswani A, Shazeer N, Parmar N, *et al.* Attention is All you Need. *Advances in Neural Information Processing Systems*. Curran Associates, Inc. 2017. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed 12 December 2023)

31 Liu X, Zheng Y, Du Z, *et al.* GPT understands, too. *AI Open*. Published Online First: 26 August 2023. doi: 10.1016/j.aiopen.2023.08.012

32 Ventz S, Cellamare M, Parmigiani G, *et al.* Adding experimental arms to platform clinical trials: randomization procedures and interim analyses. *Biostat Oxf Engl*. 2018;19:199–215.

33 Nair B. Clinical Trial Designs. *Indian Dermatol Online J*. 2019;10:193–201.

34 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–30.

35 Chawla NV, Bowyer KW, Hall LO, *et al.* SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002;16:321–57.

36 Breiman L. Random Forests. *Mach Learn*. 2001;45:5–32.

37 Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J Comput Syst Sci*. 1997;55:119–39.

38 Hinton GE. 20 - CONNECTIONIST LEARNING PROCEDURES11This chapter appeared in Volume 40 of Artificial Intelligence in 1989, reprinted with permission of North-Holland Publishing. It is a revised version of Technical Report CMU-CS-87-115, which has the same title and was prepared in June 1987 while the author was at Carnegie

Mellon University. The research was supported by contract N00014-86-K-00167 from the Office of Naval Research and by grant IST-8520359 from the National Science Foundation. In: Kodratoff Y, Michalski RS, eds. *Machine Learning*. San Francisco (CA): Morgan Kaufmann 1990:555–610. https://doi.org/10.1016/B978-0-08-051055-2.50029-8

39 Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. 2017. https://doi.org/10.48550/arXiv.1705.07874

Tables

Table 1. The holdout test confusion matrix for the best translation model with $|\delta| < 0.5\sigma$, where RF denotes random forest classifier and different feature selection criteria (best = 35 features, parsimonious = 16 features, elbow = 7 features) selected from the full set of 68 features.

Model	TP	FN	FP	TN	Precision	Recall	F1	Accuracy
Baseline	430	0	668	0	0.39	1	0.56	0.39
$\mathrm{RF}_{\mathrm{best}}$	298	132	121	547	0.71	0.69	0.70	0.77
$\mathrm{RF}_{\mathrm{parsimonious}}$	303	127	131	537	0.70	0.70	0.70	0.77
$\mathrm{RF}_{\mathrm{elbow}}$	335	95	173	495	0.66	0.78	0.71	0.76

Figure Captions

Figure 1. Overview of the preclinical recommendation system. a We collect data from publicly available preclinical and clinical trial information about *Clostridium difficile* infection. This dataset, designated as A2H, is constructed by pairing the preclinical trial with the clinical trial that shares the same drug. A binary classification label is applied to each pair, where a translation is successful (label 1) if the preclinical survival rates and clinical recovery rates are within a threshold δ. Then, a machine learning pipeline chooses the best combination of feature selection, missing value imputation, outlier detection, and classifier. We report the model performance and feature interpretation and predictions. **b** For any specified clinical trial of interest, our system computes a translation score for each candidate preclinical trial. This score quantitatively assesses the potential for successful translation. The preclinical trial that emerges with the highest translation score is then preferentially chosen to inform the design of the ensuing preclinical study.

Figure 2. Statistics of the A2H dataset. a The spectral biclustering ($K_{row} = 8$ and $K_{col} = 59$) plot of the A2H dataset with $|\delta| < 0.5\sigma$. The vertical and horizontal red dashed lines separate column and row clusters, respectively. The features on the x-axis with colons in their names represent categorical features after one-hot encoding, and the string after the colon corresponds to the original category when the encoded feature is 1. The features without colons in their names represent numerical features, and the Min-Max scaling is performed on each numerical feature independently. **b** The distribution of the survival rate from the preclinical trial. **c** The distribution of the recovery rate from the clinical trial. **d** The distribution of delta $\delta = r_r - r_s$, the difference between the clinical trial recovery rate and the preclinical trial survival rate. After fitting the normal distribution $\delta \sim N(\mu, \sigma)$ to the delta, we label the preclinical/clinical trial pairs translation success (label 1) if δ lies between $\pm 0.5\sigma$ around, and translation failure (label 0) otherwise. **e** Top 10 features with the highest absolute Spearman correlation coefficients for thresholds $|\delta| < 0.5 * \sigma$, where δ is the different between clinical recovery and preclinical survival rates, and σ is the standard deviation of δ. All the features have adjusted *p*-value < 0.001.

Figure 3. Prediction performance of the ML translation predictor. a Sequential feature selection results for different modes of selection criteria using the validation set. *Best* is the smallest feature subset when the F1 score was the best, *parsimonious* is the smallest feature subset that is within one standard error of the cross-validation performance, and *elbow* is the smallest feature subset based on the elbow method. **b** Precision-recall (PR) curves of different δ cutoff thresholds on the test set. F1$_{OOP}$ denotes the optimal operating point chosen based on the j-index of the F1 score. **c** Performance metrics of different datasets created with different δ cutoff thresholds on the test set. **d** Receiver operating characteristic (ROC) curves of different δ cutoff thresholds based on the test set.

Figure 4. Comparison of feature importance and SHAP results. The rank of the 16 features, as selected using parsimonious selection criteria of the sequential feature selection, using different analysis methods (LDA: rank based on the absolute value of linear discriminant analysis, PCC: rank based on the absolute value of Pearson

correlation coefficient, RF: rank based on random forest, and SHAP: rank based on the absolute value of SHAP). For the beeswarm plot on the right, each dot represents a data point, where red and blue correspond to high and low feature values, respectively. SHAP value indicates the amount of increase in translation success probability the feature causes for that data point. For example, a red dot with a positive SHAP value means that having a high feature value has a positive impact on predicting translation success for that data point.