Title

Predicting clinical trial success for Clostridium difficile infections based on preclinical data.

Authors

Li, Fangzhou

Youn, Jason

Millsop, Christian

et al.

Publication Date

2024

DOI

10.3389/frai.2024.1487335

Peer reviewed

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 09 October 2024
DOI 10.3389/frai.2024.1487335

OPEN ACCESS

EDITED BY
Tse-Yen Yang,
Asia University, Taiwan

REVIEWED BY
TaChen Chen,
Nihon Pharmaceutical University, Japan
Hsin-Yi Lo,
China Medical University (Taiwan), Taiwan

*CORRESPONDENCE
Ilias Tagkopoulos
✉ itagkopoulos@ucdavis.edu

†These authors have contributed equally to this work

RECEIVED 27 August 2024
ACCEPTED 26 September 2024
PUBLISHED 09 October 2024

CITATION
Li F, Youn J, Millsop C and Tagkopoulos I (2024) Predicting clinical trial success for *Clostridium difficile* infections based on preclinical data.
Front. Artif. Intell. 7:1487335.
doi: 10.3389/frai.2024.1487335

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, United States, [2]Genome Center, University of California, Davis, Davis, CA, United States, [3]USDA/NSF AI Institute for Next Generation Food Systems, University of California, Davis, Davis, CA, United States

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 sustained response endpoint is correlated with translation failure (SRC = -0.20, p-value = 1.53×10^{-54}), and explainability analysis of multi-variate random forest models shows that both sustained response 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

machine learning, translational research, 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 (Smits et al., 2016). 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 (Czepiel et al., 2019). It is estimated to affect approximately 3 million individuals worldwide every year (Cole and Stahl, 2015), underscoring its significant public health impact. Although various treatments, such as metronidazole and oral vancomycin (Zar et al., 2007; Johnson et al., 2014; Teasley et al., 1983), have been approved for CDI management, the sustained efficacy, the effectiveness of treatment after the treatment is no longer administered, is low (Van Giau et al., 2019). This is particularly concerning given the recurrent nature of CDI (Cole and Stahl, 2015; McFarland et al., 1999), 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% (Seyhan, 2019). 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 (Ciociola et al., 2014) 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 (Seyhan, 2019; Paul et al., 2010), concerns for the efficacy and safety of the drugs (Kola and Landis, 2004), poor study design, ineffective site selection, poor recruitment, patient burden, and poor trial execution (Fogel, 2018). Efforts to enhance translational success have included the use of humanized animals, which exhibit more human-like responses to medical interventions (Shultz et al., 2007), and the application of biomarkers to reduce subjectivity in evaluating drug efficacy and safety (Yu, 2011). Machine learning-based approaches (Shah et al., 2019; Toh et al., 2019; Gayvert et al., 2016) 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, interfering with their application in decision-making with high stakes (Lipton, 2017). Although machine learning models have shown promising results in other areas of life sciences (Lysenko et al., 2018; Wang et al., 2020; Eetemadi and Tagkopoulos, 2019), their application in bridging the gap between preclinical and clinical outcomes is hindered by a scarcity of expert-curated and harmonized datasets (Austin, 2021). 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, to address the data scarcity, we manually curate the Animal-to-Human (A2H) translation dataset by extracting and pairing preclinical and clinical data for *C. difficile* infections from the scientific literature and ClinicalTrials.gov, respectively. Using our A2H dataset, we train a machine learning-based classifier to predict translational success (Figure 1a). Next, to address model interpretability, we apply an explainable AI method (Lundberg and Lee, 2017), and then we expand this predictor to a recommendation system (Figure 1b).

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 (Ventz et al., 2018; Clinical and Designs, 2019). For instance, if a trial investigates two distinct treatment dosages, 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 variables. The raw data and variable description can be found in Supplementary Data 1.

Data compendium

Due to the different interests in the endpoints measured in animal and human subjects, the number of preclinical and clinical trials that share the same endpoints is limited. For example, the survival rate is predominantly measured in preclinical trials, while the recovery rate is more often used in clinical trials for CDI. The survival rate indicates the ratio of living subjects, while the recovery rate indicates the ratio of healthy patients in the group at the point of measure. These two rates both reflect drug efficacy (Zhuang et al., 2009), making them more comparable and relevant for evaluating the effectiveness of treatments in both preclinical and clinical trials. In this section, we describe a strategy to derive translation outcome using these survival and recovery rates.

For preclinical studies, 480 arms with 27 variables were gathered, including specifics of animal (e.g., species, strain, sex, age, weight), disease model (e.g., administration sequence, disease strain), and drug (e.g., dosing, duration). For clinical studies, 272 arms with 15 variables 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 variables from the raw datasets (Supplementary Information Section 1.1; Supplementary Data 2). To analytically assign the binary dependent variable, we first calculated the difference between recovery and survival rates, denoted as δ (Equation 1), for each sample in the paired dataset as follows:

$$-1.0 \le \delta = r_r - r_s \le 1.0, \tag{1}$$

where $0.0 \le r_s \le 1.0$ is the survival rate for animal subjects in the preclinical study, and $0.0 \le r_r \le 1.0$ is the recovery rate for human subjects in the clinical study. We then fit a normal distribution (Equation 2) to these deltas as

$$\delta \sim N(\mu, \sigma), \tag{2}$$

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

$$\begin{aligned} 1\,(\textit{translation success}) &: |\delta| < c * \sigma, \\ 0\,(\textit{translation failure}) &: |\delta| \ge c * \sigma, \end{aligned} \tag{3}$$

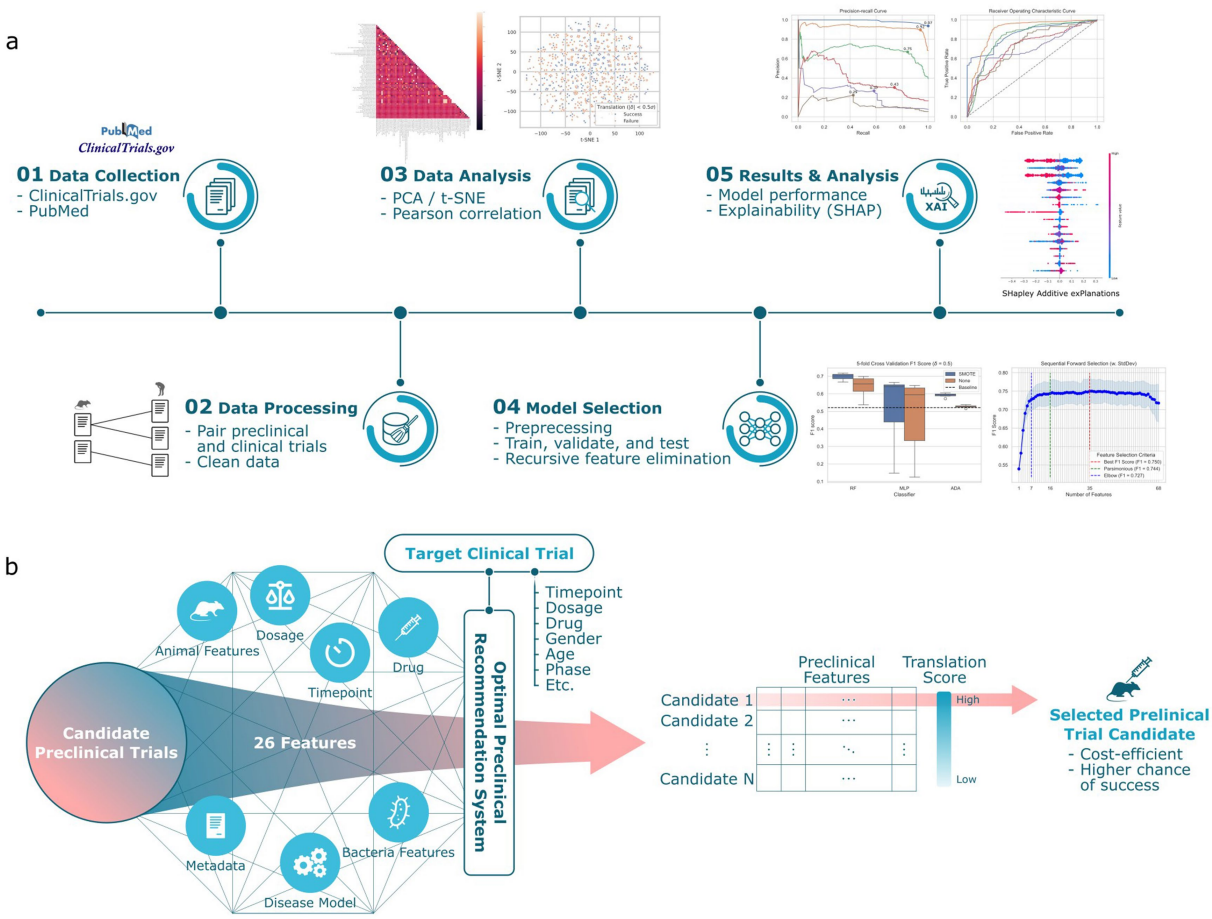

FIGURE 1
Overview of the preclinical recommendation system. (a) We collect data from publicly available preclinical and clinical trial information about *Clostridium difficile* infection. This dataset, designated as A2H, is constructed by pairing the preclinical trial with the clinical trial that shares 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.

where c is a coefficient that controls the strictness of the translation success (Equation 3). 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 categorical variables were transformed into input features applicable to machine learning using one-hot encoding. The pipeline includes, in the order specified, 2 feature scaling [Standard (Pedregosa et al., 2011) and MinMax (Pedregosa et al., 2011)], 1 missing value imputation (MVI) method [Simple (Pedregosa et al., 2011)], 1 oversampling (OS) [SMOTE (Chawla et al., 2002)], and 3 classifiers (CLS) [random forests (Breiman, 2001), AdaBoost (Freund and Schapire, 1997), MLP (Hinton, 1990)]. 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 (Supplementary Information Section 1.2). 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. We have provided more in detailed information in Supplementary Information Section 1.3.

Model interpretability

To increase the interpretability of the model, we applied the Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) 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.

Results

Experimental features correlated to preclinical to clinical translation success

Figure 2a depicts the spectral biclustering (Kluger et al., 2003) 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<4.77×10^{-122}). Similarly, the cluster with the highest average translation success rate had adopted *C. difficile* strains (e.g., VA11, 2,926, VA5, TTU 614) that were significantly different from those used in other clusters (p-value<7.2×10^{-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 Figures 2b,c. These rates are

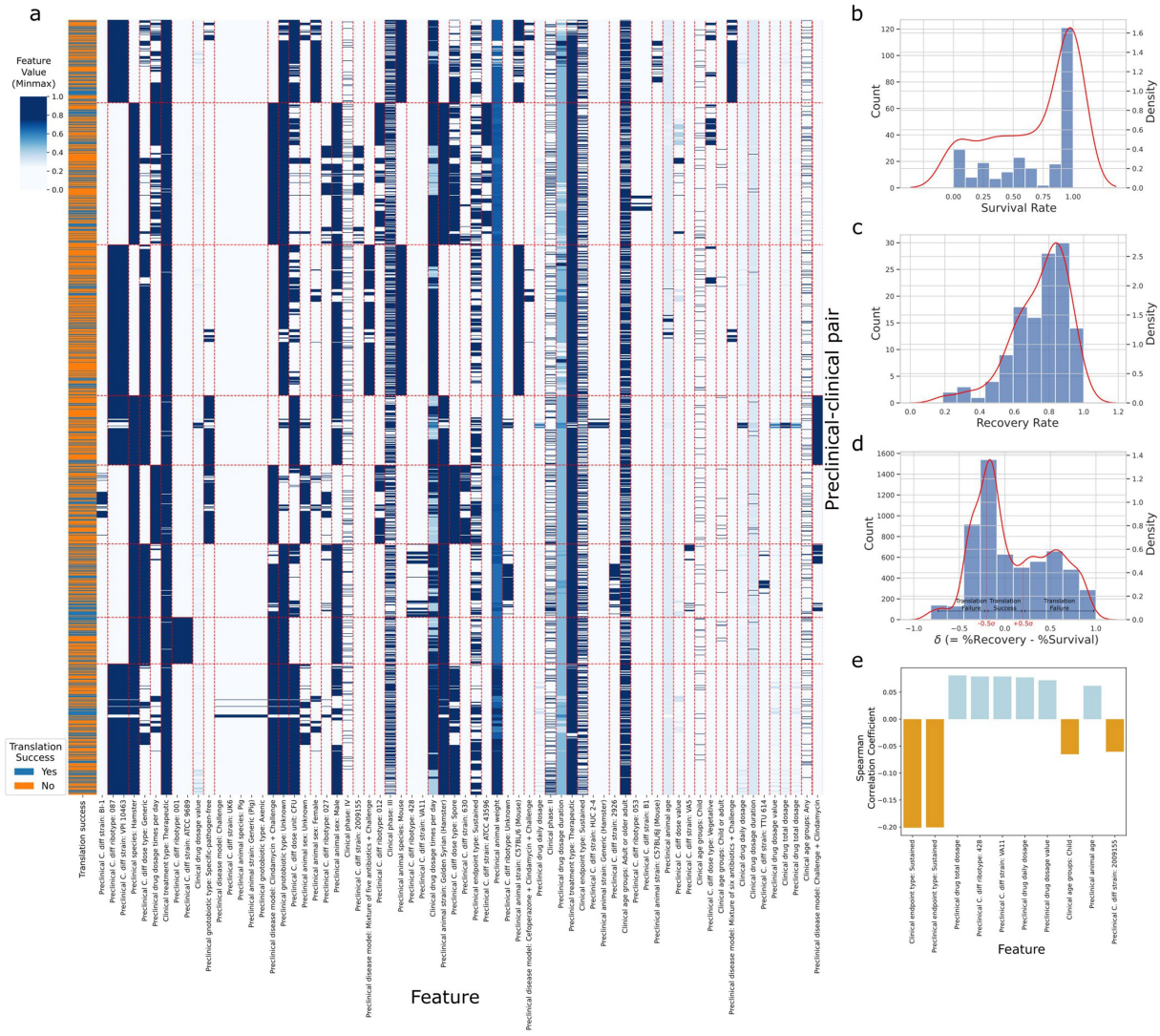

FIGURE 2
Statistics of the A2H dataset. (a) The spectral biclustering (K_{row} = 8 and K_{col} = 59) plot of the A2H dataset with $|\delta|<0.5\,\sigma$. The vertical and horizontal red dashed lines separate column and row clusters, respectively. The features on the x-axis with colons in their names 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.

skewed toward the right, partly due to the use of existing drugs like vancomycin and metronidazole as controls in case–control studies (Kaye et al., 2005). 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 $\delta \sim N(0.09, 0.41)$ (Figure 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) (Figure 2d). Spearman correlation coefficient of the features with the dependent variable lists 8 preclinical features as the top 10 most correlated features (Figure 2e), among which sustained response endpoint (i.e., outcome measured 14 days after the treatment) of the clinical trial was most negatively correlated to translation success (SRC = -0.20, p-value = 1.53×10^{-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 (Raschka, 2018) (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 Figure 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 (Satopaa et al., 2011; Figure 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 (Figures 3b–d). For all six different translation thresholds c except when $c = 2.0$, we had better performance than the random baseline (Figure 3c; Supplementary Table 2).

Sustained response 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 (Sun et al., 2022), 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 (Best et al., 2012), 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 (Manor and Lipsitz, 2013).

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 endpoints, such as percent weight change, patient-reported, and quality-of-life assessments, to provide a more comprehensive evaluation of trial success as well as the clinical utility. This, however, would result in a more complicated definition for translation success, which, in the future, would require us to provide

10.3389/frai.2024.1487335

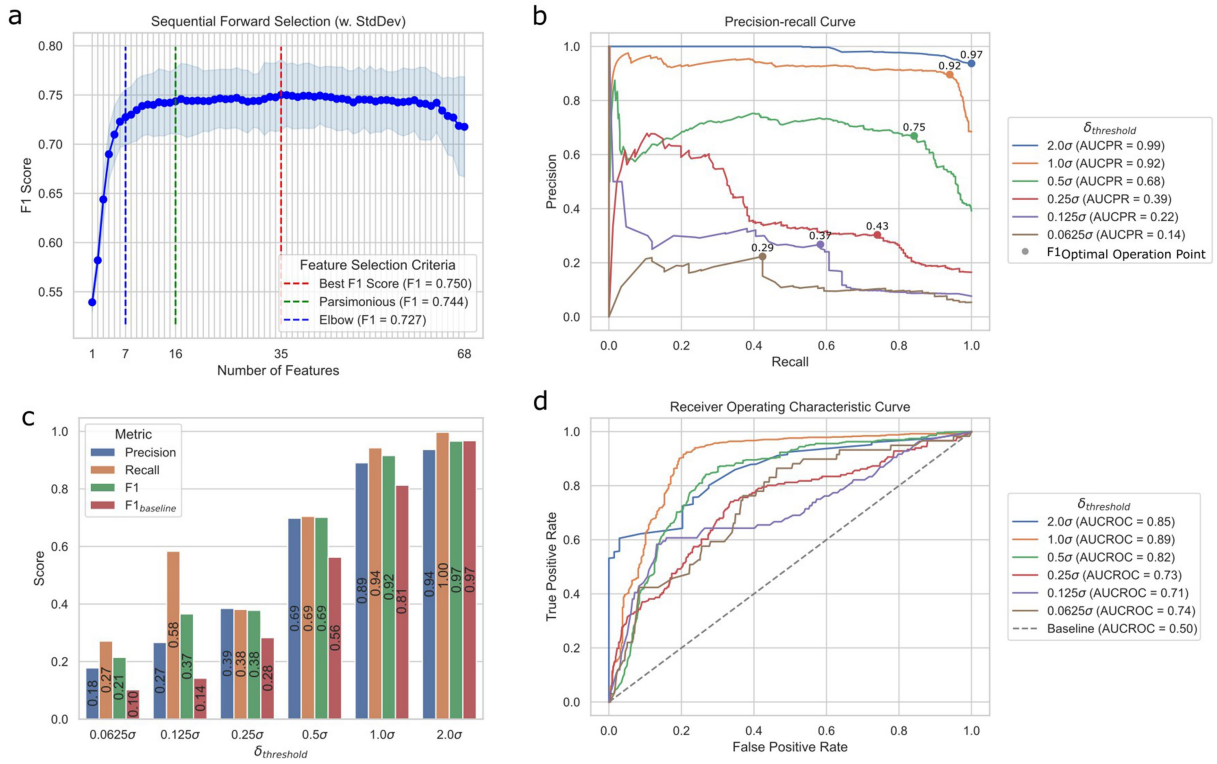

FIGURE 3
Prediction performance of the ML translation predictor. (a) Sequential feature selection results for different modes of selection criteria using the validation set. *Best* is the smallest feature subset when the F1 score was the best, *parsimonious* is the smallest feature subset that is within one standard error of the cross-validation performance, and *elbow* is the smallest feature subset based on the Kneedle 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 best F1 score. The values above the dots indicate the F1 scores of OOPs. AUCPR stands for the area under the PR curve. (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. AUCROC stands for the area under the ROC curve.

TABLE 1 The holdout test confusion matrix for the best translation model with $|\delta| < 0.5\,\sigma$.

Model	TP	FN	FP	TN	Precision	Recall	F1	Accuracy
Baseline	**430**	**0**	668	0	0.39	**1**	0.56	0.39
Random forest								
- Best ($K = 35$)	298	132	**121**	**547**	**0.71**	0.69	0.70	**0.77**
- Parsimonious ($K = 16$)	303	127	131	537	0.70	0.70	0.70	**0.77**
- Elbow ($K = 7$)	335	95	173	495	0.66	0.78	**0.71**	0.76

The baseline model would always predict positive (translation success). For the random forest model, we showed three scenarios with different feature selection criteria. *K* denotes the number of features selected from the total 68 features. While deploying the model in real life, fewer features would reduce data collection costs. For each metric, the best performance was highlighted in bold, and the second best was highlighted with an underscore.

better explanations for our recommendations such that they would be trustworthy and actionable for drug development researchers. Second, rather than employing the delta δ, which represents the difference between survival and recovery rates in current work, we 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, due to the complicated nature of (pre) clinical experimental designs, comparing results across different studies may have confounding biases resulting from unobserved variables. In this work, for example, we directly modeled the translation between preclinical and clinical trials, while in reality, trials usually consist of comparator groups to account for experimental biases. As a future improvement, we plan to adapt a recent work that used pairwise meta-learning to allow the model to learn across different experiments efficiently (Feng et al., 2024). Fourth, we would like to dive deeper into analyzing the impact of dosage regimens on model predictions. During our SHAP analysis, we found that high dosage amounts in preclinical trials had a positive impact on the model predictions on translation success rate, while they had a negative impact in clinical trials. While interesting, SHAP might produce misleading explanations for highly correlated features, e.g., daily dosage amount vs. dosage times per day (Aas et al., 2021). Dosage information is essential in drug development, and we aim to conduct a focused, rigorous analysis of dosage regimens for our next

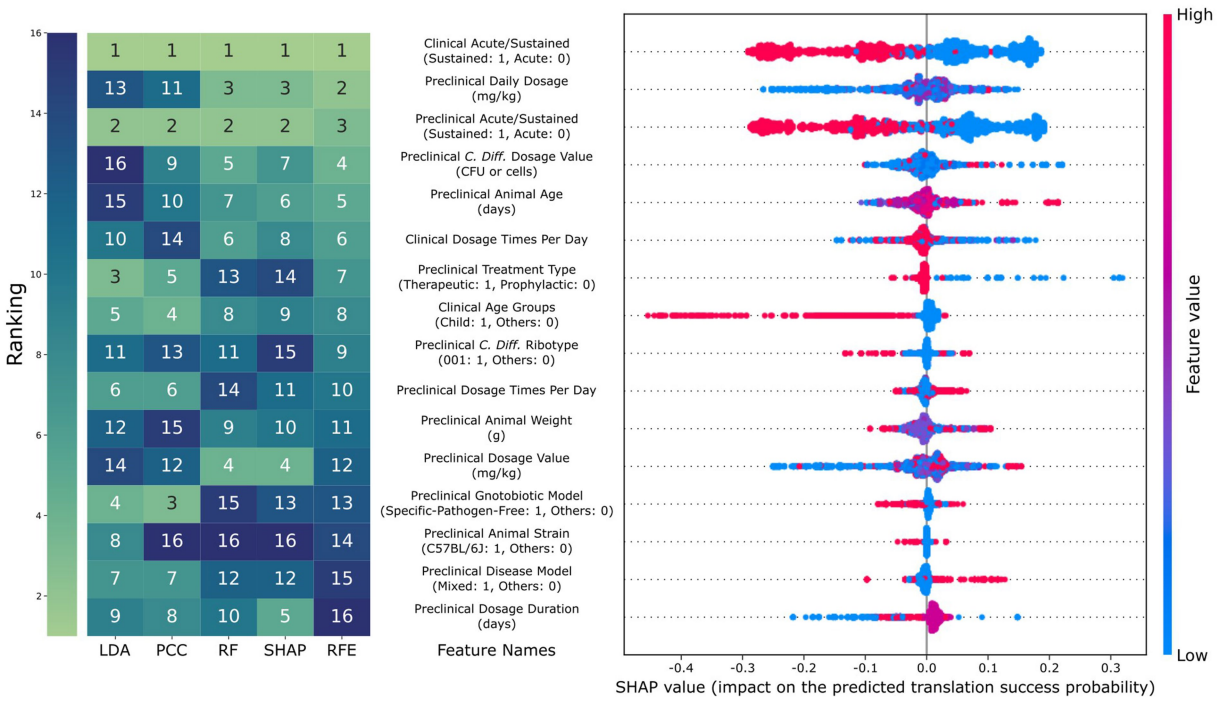

FIGURE 4
Comparison of feature importance and SHAP results. The rank of the 16 features, as selected using parsimonious selection criteria of the sequential feature selection, using different analysis methods (LDA: rank based on the absolute value of linear discriminant analysis, PCC: rank based on the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

study phase. Fifth, the primary challenge of the data curation process is its dependency on expert-guided manual data curation. The current random forest model trained on a small dataset outperformed the state-of-the-art deep learning models, such as neural networks. The advantage of deep neural networks is their capability to generalize the representation to transfer to similar domain datasets (LeCun et al., 2015). Implementing an automated data extraction pipeline, leveraging transformer-based large language models (LLMs)(Lee et al., 2019; Vaswani et al., 2017; Liu et al., 2023), 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. We also consider including more clinical variables, such as patient demographics and health conditions, to further enhance the capability of our predictor. 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. The source code for our A2H recommendations system can be found at https://github.com/IBPA/A2H.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Author contributions

FL: Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. JY: Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. CM: Data curation, Formal analysis, Writing – review & editing. IT: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. 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.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2024.1487335/full#supplementary-material

References

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. *Artif. Intell.* 298:103502. doi: 10.1016/j.artint.2021.103502

Austin, C. P. (2021). Opportunities and challenges in translational science. *Clin. Transl. Sci.* 14, 1629–1647. doi: 10.1111/cts.13055

Best, E. L., Freeman, J., and Wilcox, M. H. (2012). Models for the study of *Clostridium difficile* infection. *Gut Microbes* 3, 145–167. doi: 10.4161/gmic.19526

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Ciociola, A. A., Cohen, L. B., Kulkarni, P., Kefalas, C., Buchman, A., Burke, C., et al. (2014). How drugs are developed and approved by the FDA: current process and future directions. *Am. J. Gastroenterol.* 109, 620–623. doi: 10.1038/ajg.2013.407

Clinical, N. B., and Designs, T. (2019). Indian. *Dermatol. Online J.* 10, 193–201. doi: 10.4103/idoj.IDOJ_475_18

Cole, S. A., and Stahl, T. J. (2015). Persistent and recurrent *Clostridium difficile* colitis. *Clin. Colon Rectal Surg.* 28, 065–069. doi: 10.1055/s-0035-1547333

Czepiel, J., Dróżdż, M., Pituch, H., Kuijper, E. J., Perucki, W., Mielimonka, A., et al. (2019). *Clostridium difficile* infection. *Eur. J. Clin. Microbiol. Infect. Dis.* 38, 1211–1221. doi: 10.1007/s10096-019-03539-6

Eetemadi, A., and Tagkopoulos, I. (2019). Genetic neural networks: an artificial neural network architecture for capturing gene expression relationships. *Bioinformatics* 35, 2226–2234. doi: 10.1093/bioinformatics/bty945

Feng, B., Liu, Z., Huang, N., Xiao, Z., Zhang, H., Mirzoyan, S., et al. (2024). A bioactivity foundation model using pairwise meta-learning. *Nat. Mach. Intell.* 6, 962–974. doi: 10.1038/s42256-024-00876-w

Fogel, D. B. (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp. Clin. Trials Commun.* 11, 156–164. doi: 10.1016/j.conctc.2018.08.001

Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55, 119–139. doi: 10.1006/jcss.1997.1504

Gayvert, K. M., Madhukar, N. S., and Elemento, O. (2016). A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem. Biol.* 23, 1294–1301. doi: 10.1016/j.chembiol.2016.07.023

Hinton, G. E. (1990). 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. *Mach. Learn.* 1, 555–610.

Johnson, S., Louie, T. J., Gerding, D. N., Cornely, O. A., Chasan-Taber, S., Fitts, D., et al. (2014). Vancomycin, metronidazole, or tolevamer for *Clostridium difficile* infection: results from two multinational, randomized, controlled trials. *Clin. Infect. Dis.* 59, 345–354. doi: 10.1093/cid/ciu313

Kaye, K. S., Harris, A. D., Samore, M., and Carmeli, Y. (2005). The case-case-control study design: addressing the limitations of risk factor studies for antimicrobial resistance. *Infect. Control Hosp. Epidemiol.* 26, 346–351. doi: 10.1086/502550

Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.* 13, 703–716. doi: 10.1101/gr.648603

Kola, I., and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–716. doi: 10.1038/nrd1470

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2019). Bio BERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240. doi: 10.1093/bioinformatics/btz682

Lipton, Z. C. (2017). The mythos of model interpretability. *Comput. Sci.* 2017:3490. doi: 10.48550/arXiv.1606.03490

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., et al. (2023). *GPT understands, too*. AI Open. Available at: https://www.sciencedirect.com/science/article/pii/S2666651023000141.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. Red hook, NY, USA: Curran associates Inc., pp. 4768–4777.

Lysenko, A., Sharma, A., Boroevich, K. A., and Tsunoda, T. (2018). An integrative machine learning approach for prediction of toxicity-related drug safety. *Life Sci. Alliance* 1:e201800098. doi: 10.26508/lsa.201800098

Manor, B., and Lipsitz, L. A. (2013). Physiologic complexity and aging: implications for physical function and rehabilitation. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* 45, 287–293. doi: 10.1016/j.pnpbp.2012.08.020

McFarland, L. V., Surawicz, C. M., Rubin, M., Fekety, R., Elmer, G. W., and Greenberg, R. N. (1999). Recurrent *Clostridium difficile* disease: epidemiology and Clinical characteristics. *Infect. Control Hosp. Epidemiol.* 20, 43–50. doi: 10.1086/501553

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., et al. (2010). How to improve R & D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* 9, 203–214. doi: 10.1038/nrd3078

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.48550/arXiv.1201.0490

Raschka, S. (2018). MLxtend: providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J. Open Source Softw.* 3:638. doi: 10.21105/joss.00638

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). *Finding a "Kneedle" in a haystack: detecting knee points in system behavior*. In: 2011 31st international conference on distributed computing systems workshops, pp. 166–171.

Seyhan, A. A. (2019). Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Transl. Med. Commun.* 4, 1–19. doi: 10.1186/s41231-019-0050-7

Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., et al. (2019). Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ Digit. Med.* 2, 1–5. doi: 10.1038/s41746-019-0148-3

Shultz, L. D., Ishikawa, F., and Greiner, D. L. (2007). Humanized mice in translational biomedical research. *Nat. Rev. Immunol.* 7, 118–130. doi: 10.1038/nri2017

Smits, W. K., Lyras, D., Lacy, D. B., Wilcox, M. H., and Kuijper, E. J. (2016). *Clostridium difficile* infection. *Nat. Rev. Dis. Prim.* 2, 1–20. doi: 10.1038/nrdp.2016.20

Sun, D., Gao, W., Hu, H., and Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sin. B* 12, 3049–3062. doi: 10.1016/j.apsb.2022.02.002

Teasley, D., Olson, M., Gebhard, R., Gerding, D., Peterson, L., Schwartz, M., et al. (1983). Prospective randomised trial of metronidazole versus vancomycin for Clostridium-difficile-associated diarrhoea and colitis. *Lancet* 322, 1043–1046. doi: 10.1016/S0140-6736(83)91036-X

Toh, T. S., Dondelinger, F., and Wang, D. (2019). Looking beyond the hype: applied AI and machine learning in translational medicine. *EBioMedicine* 47, 607–615. doi: 10.1016/j.ebiom.2019.08.027

Van Giau, V., Lee, H., An, S. S. A., and Hulme, J. (2019). Recent advances in the treatment of *C. difficile* using biotherapeutic agents. *Infect. Drug Resist.* 12, 1597–1615. doi: 10.2147/IDR.S207572

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). *Attention is all you need*. In: Advances in neural information processing systems. Curran associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Ventz, S., Cellamare, M., Parmigiani, G., and Trippa, L. (2018). Adding experimental arms to platform clinical trials: randomization procedures and interim analyses. *Biostatistics* 19, 199–215. doi: 10.1093/biostatistics/kxx030

Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I. J., Rudd, A. G., et al. (2020). A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS One* 15:e0234722. doi: 10.1371/journal.pone.0234722

Yu, D. (2011). Translational research: current status, challenges and future strategies. *Am. J. Transl. Res.* 3, 422–433

Zar, F. A., Bakkanagari, S. R., Moorthi, K., and Davis, M. B. (2007). A comparison of vancomycin and metronidazole for the treatment of *Clostridium difficile*-associated diarrhea, stratified by disease severity. *Clin. Infect. Dis.* 45, 302–307. doi: 10.1086/519265

Zhuang, S. H., Xiu, L., and Elsayed, Y. A. (2009). Overall survival: a gold standard in search of a surrogate: the value of progression-free survival and time to progression as end points of drug efficacy. *Cancer J.* 15, 395–400. doi: 10.1097/PPO.0b013e3181be231d