
Matching clinicians with clinical trials using AI

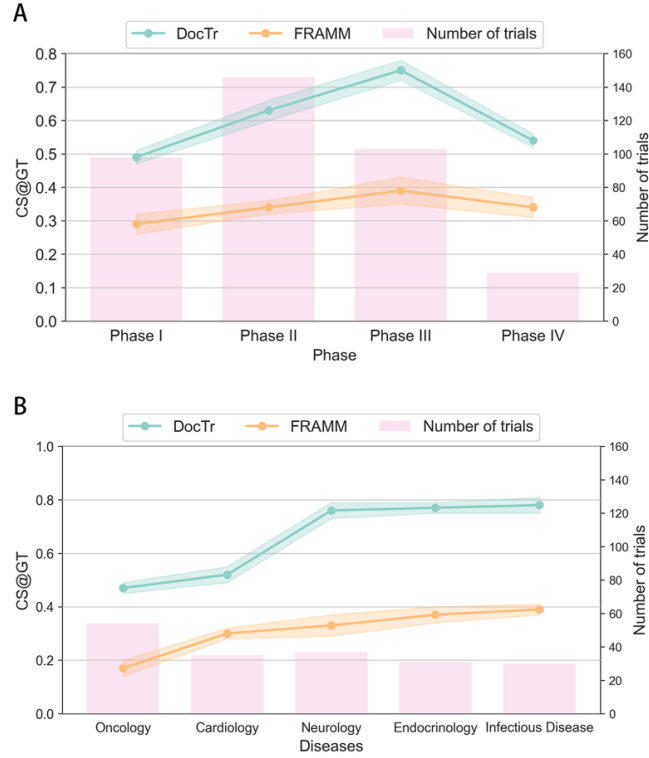
In the format provided by the
authors and unedited

Supplementary Materials

A Recommendation performance for different trial phases and disease categories

In this section, we explore the DocTr’s efficacy across varying trial phases and disease categories, highlighting its adaptability and specialized performance in diverse clinical settings. All evaluations are conducted under the temporal split setting. **Supplementary Supplementary Figure 1-A** shows the performances of DocTr and the best baseline model FRAMM across different phases of clinical trials (Phase I-IV). Here, the pink bars indicate the number of trials in each phase. DocTr consistently outperforms FRAMM across all trial phases, with both models achieving their highest composite similarity scores in Phase III trials. Phase I and Phase IV trials pose greater challenges. The low recommendation performance on Phase I trials may be because there are fewer clinicians with the necessary experience and willingness to participate in these high-risk, early-stage trials[1]. Phase IV trials show reduced performance likely due to the smaller number of trials in these phases, leading to insufficient data for effective training and potential data bias.

Supplementary Supplementary Figure 1-B shows the prediction performance (CS@GT) of DocTr versus the best baseline model FRAMM across different disease categories on the testing set. We select five major disease categories based on ICD codes in our trial dataset: Oncology (C00-D49), Cardiology (I00-I99), Neurology (G00-G99), Endocrinology (E00-E89), and Infectious diseases (A00-B99). The pink bars show the number of trials in different categories. Consistent with the findings across trial phases, DocTr surpasses FRAMM in all disease categories. Notably, both models find it particularly challenging to identify suitable clinicians for oncology and cardiology. A potential contributing factor may be the inherent data sparsity associated with high specialization in these fields. Highly specialized clinicians, such as an oncologist focusing on a single cancer type, tend to have less diverse patient encounter vectors (i.e., fewer unique ICD-10 codes). Our analysis confirms this: the average L0 norm of clinician encounter vectors in oncology and cardiology was lower than clinicians in other domains. This data sparsity can make it more challenging for the model to differentiate between otherwise similar top-tier specialists, thereby impacting recommendation precision.



Supplementary Figure 1 Prediction performance (CS@GT) of DocTr and the best baseline model FRAMM for phase I-IV trials and different disease categories on the testing set. Shadowed area indicates standard deviations. The sample size $n = 5$ independent model evaluations was performed using different random seeds.

B Clinician recommendation performance under regular settings

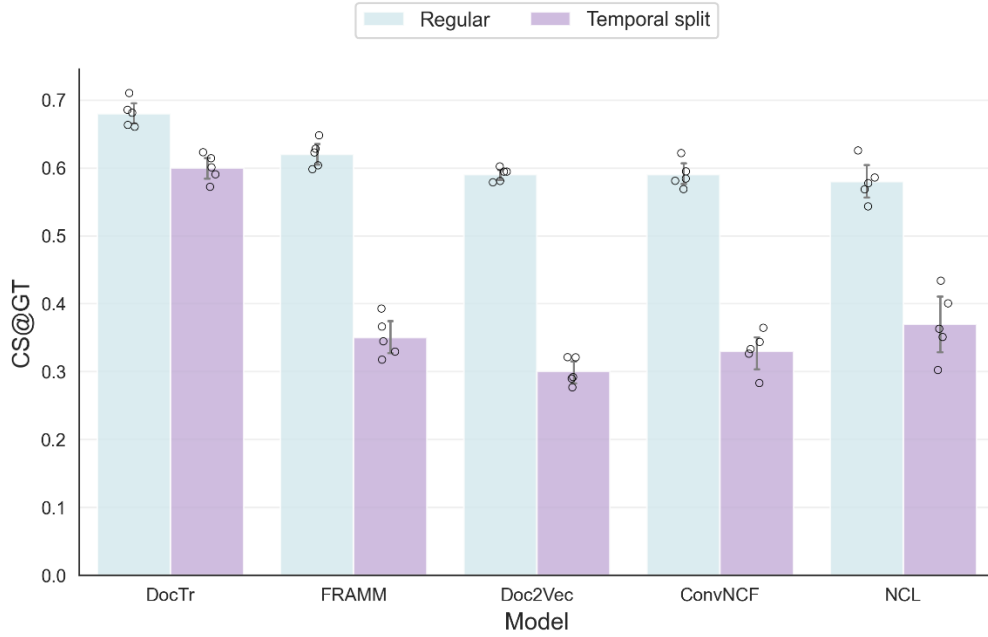
This setup simulates identifying additional suitable clinicians when some potential candidates are already known. We first assess DocTr against baseline models under regular settings, involving a typical random 8:1:1 data split for training, validation, and testing. While the same clinicians and trials might appear across sets, each unique clinician-trial enrollment relationship is exclusive to one set. This regular setting evaluation is designed to simulate a common real-world task where trial managers seek to augment a list of potential sites by recommending additional clinicians from a pool of known investigators. In this scenario, leveraging historical enrollment data (structural information) is highly valuable.

Results (**Supplementary Supplementary Table 1**) demonstrate DocTr significantly surpasses all baselines. Compared to the best-performing baseline, FRAMM[2], DocTr achieves up to 19% higher composite similarity (CS). This score also exceeds the ground truth CS, which measures the similarity among clinicians actually enrolled in the trials. In ablation studies, DocTr-structural (relying solely on structural information) showed high predictive accuracy, whereas DocTr-semantic (using only semantic information) performed significantly lower. This indicates structural information dominates prediction performance in this setting, as recurring clinicians and trials across datasets allow existing connections to effectively identify potential candidates.

Supplementary Table 1 Model performance under regular settings. Performance metrics are reported as mean (standard deviation). An asterisk (*) indicates the performance difference is statistically significant

($p < 0.01$). We evaluate the performance of top K recommended clinicians, shown as $CS@K$. $CS@GT$ sets K to the actual number of clinicians who enrolled in each trial, thereby varying K across different trials. Bold numbers indicate best performance.

Model	$CS@GT$ ($\uparrow$)	$CS@5$ ($\uparrow$)	$CS@10$ ($\uparrow$)	$CS@20$ ($\uparrow$)
Ground Truth	0.30 (-)	0.55 (-)	0.46 (-)	0.38 (-)
Random	0.05 (0.20)	0.10 (0.21)	0.09 (0.25)	0.06 (0.23)
Popular	0.28 (-)	0.38 (-)	0.34 (-)	0.29 (-)
LightGCN	0.56 (0.02)	0.55 (0.02)	0.57 (0.01)	0.57 (0.01)
ConvNCF	0.59 (0.02)	0.61 (0.02)	0.60 (0.03)	0.59 (0.03)
NCL	0.58 (0.03)	0.62 (0.02)	0.62 (0.02)	0.60 (0.02)
Doc2Vec	0.59 (0.01)	0.62 (0.03)	0.60 (0.01)	0.60 (0.02)
FRAMM	0.62 (0.02)	0.63 (0.01)	0.63 (0.02)	0.62 (0.02)
DocTr-semantic	0.35 (0.03)	0.35 (0.02)	0.33 (0.02)	0.36 (0.03)
DocTr-structural	0.65 (0.02)	0.64 (0.02)	0.65 (0.01)	0.65 (0.02)
DocTr	0.68 (0.02)*	0.75(0.02)*	0.73 (0.03)*	0.71 (0.02)*



Supplementary Figure 2 Performance differences ($CS@GT$) between the regular setting and the temporal split setting. We plot the models which have the top 5 highest composite similarities. DocTr has the lowest performance gap. Bars represent the mean and error bars indicate $\pm$ SD. The sample size $n = 5$ independent model evaluations was performed using different random seeds.

C Baseline Models

The detailed introduction for baseline models are as follows:

Naïve recommendation models:

- **Random:** This model randomly selects K clinicians for each trial, serving as a baseline to measure the bare minimum expectation of any recommendation engine.
- **Popular:** This model selects the top K clinicians based on the number of trials they have

previously enrolled in, using popularity as a simple heuristic.

Recommendation models for general tasks:

- **LightGCN**[3]: Initially proposed for collaborative filtering tasks, this model utilizes structural information of the network to enhance recommendation accuracy.
- **ConvNCF**[4]: A convolutional neural network framework that exploits the semantic information embedded in user and item embeddings to improve recommendation performance.
- **NCL**[5]: This model employs graph neural networks to extract structural relationships between items and users, while also clustering them to obtain enriched semantic embeddings.

Clinical trial models:

- **Doc2Vec**[6]: Adapts a memory network to construct doctor representations based on static features and patient visits, which are then used to match with specific trial representations to predict enrollment counts.
- **FRAMM**[2]: As the current state-of-the-art model in clinical trial site recommendation, FRAMM utilizes masked cross attention to construct trial and site embeddings, employing reinforcement learning to generate a ranked list of trial site recommendations.

All models are trained using PyTorch 2.0.1[7] with one NVIDIA V100 GPU. We use the RecBole package to run all recommendation experiments[8]. Hyperparameters for each model are tuned using the grid search strategy on the validation dataset. The DocTr model is trained using the Adam optimizer with a learning rate of 0.001, and the projection dimensions are set at 64. The genetic algorithm’s population size is fixed at 50, with a mutation rate of 0.01. To ensure reliability in our results, all models are run five times with different random seeds, and standard deviations of model performances are calculated to assess consistency and robustness.

D Dataset Statistics

Supplementary Table 2 Trials and clinicians data statistics

Feature	Value
# of clinicians	24,984
# of ICD codes	1,718
# of clinical trials	5,210
Phase I	936
Phase II	1,224
Phase III	1,298
Phase IV	239
Unknown	875
# of clinicians recruited in trials	
Mean (std)	13.4 (25.4)
Median	4.0
Min - Max	1 - 629
# of trials that clinicians enrolled in	
Mean (std)	2.7 (1.0)
Median	3.5
Min - Max	1 - 91

E Algorithms

Supplementary Table 3 The pipeline of the genetic optimization algorithm for fair clinician recommendations.

Input:

Clinicians (with their M attributes and C competing trials), Number of clinicians to recommend K , Population size, Number of generations, Mutation rate γ

Output:

List of K clinicians (best solution)

Training:

Initialize population with random solutions of size K .

for each generation **do**

 Calculate fitness scores for each solution in the population using Eq. 12.

 Select parent solutions based on their fitness scores f .

 Perform crossover by exchanging clinician subsets between selected parents to create new child solutions.

 Mutate child solutions by replacing clinicians with a given mutation rate γ .

 Replace the population with the new child solutions.

end for

Select the best solution from the final population based on fitness score.

Supplementary Table 4 The inference pipeline of the proposed DocTr model.

Input:

K_c candidate clinicians (with their patient encounter vector $\mathbf{c}$, attributes and competing trial information), Clinical trial summary $\mathbf{s}$, eligibility criteria $\mathbf{r}$ and condition code $\mathbf{d}$, clinician-trial enrollment history, target number of clinicians K_t

Output:

List of K_t clinicians

Inferencing:

Use the trained model to generate semantic embeddings $\mathbf{e}_c, \mathbf{e}_s, \mathbf{e}_r$ for clinicians and trials using Eq. 1-2.

Use the trained model to generate enrollment history embeddings $\mathbf{h}_r, \mathbf{h}_s$ using Eq. 3.

Use the trained model to generate structural embeddings $\mathbf{g}_t, \mathbf{g}_c$ for clinicians and trials using Eq. 4-7.

Calculate the matching scores $\hat{\mathbf{y}}$ as initial recommendations using Eq. 8-10.

Run the genetic algorithm to optimize fairness and competing trials of the initial recommendations.

Select top K_t clinicians from the list ranked by fitness score f as final recommendations.

References

1. Karavasilis V, Digue L, Arkenau T, et al. Identification of factors limiting patient recruitment into phase I trials: a study from the Royal Marsden Hospital. *European Journal of Cancer* 2008;**44**(7):978-82.
2. Theodorou B, Glass L, Xiao C, Sun J. FRAMM: Fair ranking with missing modalities for clinical trial site selection. *Patterns* 2024.

-
3. Lightgcn: Simplifying and powering graph convolution network for recommendation. Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval; 2020.
 4. He X, Du X, Wang X, Tian F, Tang J, Chua T-S. Outer product-based neural collaborative filtering. arXiv preprint arXiv:1808.03912 2018.
 5. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. Proceedings of the ACM web conference 2022; 2022.
 6. Doctor2vec: Dynamic doctor representation learning for clinical trial recruitment. Proceedings of the AAAI Conference on Artificial Intelligence; 2020.
 7. Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 2019;**32**.
 8. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. proceedings of the 30th acm international conference on information & knowledge management; 2021.