

DCCTNET: KIDNEY TUMORS SEGMENTATION BASED ON DUAL-LEVEL COMBINATION OF CNN AND TRANSFORMER

Bingzhen Hou¹, Guimei Zhang^{1*}, Huiqun Liu², Yipeng Qin³, Ying Chen¹

Nanchang Hangkong University¹, Jiangxi Cancer Hospital², Cardiff University³

*Guimei Zhang is the corresponding author

ABSTRACT

The hybrid model of CNN(Convoluation Neural Networks) and Transformer is a popular method in segmenting kidney images, but most existing hybrid models directly fused local features from CNN with global features from Transformer, ignoring the issue of semantic gaps between distinct features. Furthermore, feature fusion is typically performed solely at the feature level, without considering alignment at the mask (prediction map) level. To address these limitations, we propose a novel segmentation method called Dual-level Combination of CNN and Transformers Network (DCCTNet). Specifically, we select similar features from both CNN and Transformer to reduce semantic gaps at the feature level. Additionally, we further utilize the global information of the Transformers by reducing the difference between the prediction maps in the coding stage at the mask level. We evaluate DCCTNet on the KiTS19 dataset, achieving 97.3% dice score for kidneys segmentation and 81.2% dice score for kidney tumors segmentation, respectively. <https://github.com/hou-bz/DCCTNet>.

Index Terms— CNN, Transformer, Kidney tumors, Medical image segmentation

1. INTRODUCTION

Kidney cancer is a prevalent global disease that seriously threatens human health [1]. Improving the precise segmentation of kidney tumors can greatly assist physicians in devising treatment plans and diagnosing the disease [2]. With the development of deep learning and medical imaging, CNN models have gained significant attention in medical image processing [3, 4].

Kidneys and kidney tumors segmentation play a crucial role in medical image segmentation, which has attracted many researchers [5, 6, 7]. For example, Guo et al. [8] proposed a model called RAU-Net for kidney tumors segmentation, although RAU-Net can improve boundary segmentation accuracy, the generalization ability of this method is weak. Zhang et al [9] designed a simple network, which used a dynamic segmentation header that allowed segmentation on partially labeled datasets based on different organs and tumors, and achieved good results in kidney

tumors segmentation. Zhu et al. [10] proposed FSNet, which adapted transfer learning method. By conducting pre-training on large-scale dataset and fine-tuning it on kidney tumors dataset. However, the local bias characteristic in CNN hampers its ability to capture long-range dependencies, thereby hindering further improvement of segmentation performance [11, 12, 13].

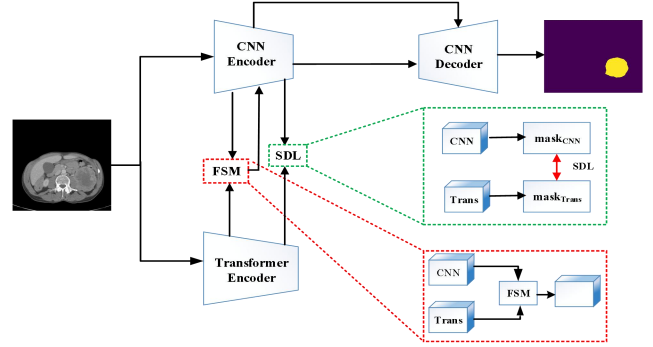


Fig. 1. Flow chart of DCCTNet. $mask_{CNN}$ denotes the mask generated by the features output in the CNN encoding stage, whereas $mask_{Trans}$ denotes the mask generated by the features output in the Transformer encoding stage.

Inspired by the effectiveness of Transformer in capturing global information [14], many researchers have incorporated Transformer into CNN to address the challenge of missing global information. TransUNet [15] is an excellence medical image segmentation model that stacks convolutions and Transformer in the encoder stage in a cascade manner. This powerful encoder achieves better results in medical image segmentation. Huang et al. [16] redesigned the feed-forward in Transformer blocks to supplement local information, modeling global and local information simultaneously. Lin et al. [17] designed a double-branch Transformer structure, capturing global information of the images at different resolutions in the encoder, achieving good segmentation performance on polyp images. Heidari et al. [18] proposed a hybrid encoder by combining CNNs and Transformers to enhance the fusion of global and local information more efficiently, incorporating a Double-Level Fusion (DLF) module into the skip connection.

In this paper, considering the complementary roles of CNN and Transformer, we propose DCCTNet, a CNN-Transformer hybrid network for kidneys and kidney tumors segmentation. In Fig.1, the encoder composes of a CNN and a Transformer, with the CNN serving as the primary component and the Transformer as the secondary one. Additionally, we incorporate the Feature Selection Module (FSM) and Semantic Distillation Loss (SDL) to effectively capture global information from the Transformer. Moreover, a multi-scale feature pyramid structure is employed in the decoder to enhance context information extraction.

Building upon previous works, our research has made the following main contributions.

(1) This paper proposes a novel dual-level combination hybrid model that efficiently utilizes long-range dependencies of Transformer and local features of CNN.

(2) We propose a Feature Selection Module (FSM) to selectively enhance features with high semantic similarity and weaken those with large semantic differences, thereby alleviating the semantic gaps during features fusion. Furthermore, we design a Semantic Distillation Loss (SDL) that enables CNN to learn global information from Transformer network at the mask level.

(3) We perform kidneys and kidney tumors segmentation on the KiTS19 dataset. Compare to other hybrid models or kidney tumor-specific methods, our proposed method achieves superior segmentation results.

2. PROPOSED METHOD

2.1. Overview of DCCTNet

The DCCTNet model consists of four modules: hybrid encoder(hybrid encoder structure that combines both ResNet34 as a backbone and Swin Transformer as an auxiliary encoder), decoder, feature selection module, and semantic distillation loss. Fig.2 illustrates the architecture of DCCTNet.

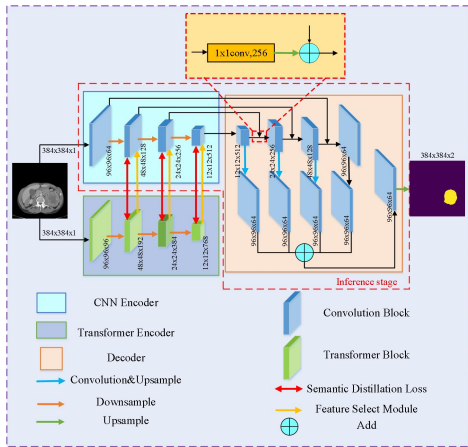


Fig.2. Detailed network architecture diagram of DCCTNet

2.2. Feature Selection Module (FSM)

Due to the existence of semantic gaps between the features extracted by Transformer and CNN, directly fusing them without FSM would inevitably lead to some negative impact. To address this issue, we propose a novel FSM, as show in Fig.3. The FSM automatically calculates semantic similarity between the features extracted by Transformer and CNN networks, and generates a similarity probability map. Based on this map, high weights are assigned to features with high similarity while low weights to those with low similarity. Subsequently, the weighted features obtained from both networks are multiplied with the similarity map. Therefore, features with high similarity are selectively enhanced. Furthermore, channel attention is applied to process the original features in order to obtain a weight map for each channel which is then assigned to the enhanced feature representation. Finally, these two types of features are effectively fused together, mitigating the existing semantic gaps present within their respective network extractions.

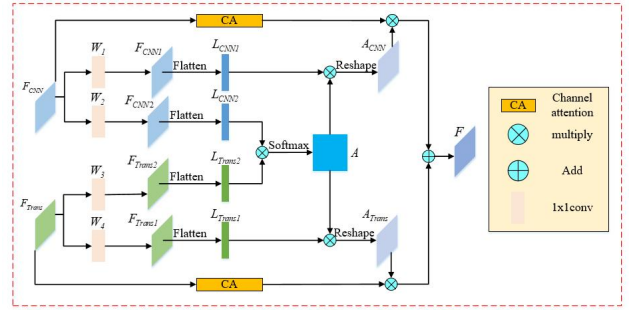


Fig. 3. Feature Selection Module (FSM)

Specifically, the 1x1 convolution is employed to map features F_{Trans} and F_{CNN} , and ensure their dimensions are consistent, resulting in the generation of F_{C1} , F_{C2} from F_{CNN} , and F_{T1} , F_{T2} from F_{Trans} . To facilitate semantic information interaction among these features, we flatten F_{C1} , F_{C2} , F_{T1} and F_{T2} to generate sequences L_{C1} , L_{C2} , L_{T1} and L_{T2} , respectively. These sequences are then multiplied element wise to obtain a similarity map A_i that can be computed using Eq. (1).

$$A_i = \frac{\exp(L_{C1_i} L_{T1_i})}{\sum_{i=1}^N \exp(L_{C1_i} L_{T1_i})} \quad (1)$$

where, N is the length of the flattened vector.

To enhance the features with semantics similarity, the generated similarity map A_i is used as weights and multiplied with L_{C1} and L_{T1} , respectively. Subsequently, we can obtain an enhanced feature with global information that can be computed using Eq. (2) and (3).

$$A_{CNN} = \text{Reshape}(\text{Conv}_C(\sum_{i=1}^N (A_i L_{CI}))) \quad (2)$$

$$A_{Trans} = \text{Reshape}(\text{Conv}_T(\sum_{i=1}^N (A_i L_{TI}))) \quad (3)$$

where, Conv_C and Conv_T are convolution operations on the features extracted from CNN and Transformer networks, respectively. Reshape is the process of converting flattened vectors into feature maps.

To further enhance the semantic information of A_{CNN} and A_{Trans} , we utilize the channel attention to amplify the most important features through re-weighting. Finally, these re-weighting features are summed up, which can be computed using Eq. (4).

$$F = CA(F_{CNN}) \times A_{CNN} + CA(F_{Trans}) \times A_{Trans} \quad (4)$$

where, $CA(\cdot)$ represents channel attention manipulation.

2.3. Semantic Distillation Loss (SDL)

We propose a Feature Selection Module (FSM) to further mitigate semantic gaps and compensate for the insufficiency of global information in CNN at the feature level. However, if there are significant semantic gaps between features, it becomes challenging for the FSM to effectively learn global information at the feature level. To address this issue, inspired by knowledge distillation models [19], this paper proposes a Semantic Distillation Loss (SDL) that learns global information from Transformers at the mask level.

Specifically, we up-sample features from the second to fourth layers of the hybrid encoders to match the dimensions of the label mask. To enhance alignment between the generated masks from both networks, we introduce a KL (Kullback-Leibler) loss that guides the main encoder in capturing global information at the mask level from the Transformer. Following previous works on knowledge distillation, the mask is classified pixel-wise. The loss can be computed by Eq. (5).

$$L_{KL} = \frac{1}{N} \sum_{i \in N} KL(P_i^c \parallel P_i^t) \quad (5)$$

where, N presents the number of pixels in the mask, i is the index of each pixel, P_i^c and P_i^t are the masks generated by CNN and Transformer models, respectively.

To minimize the difference between the mask and label, we utilize a label-based constraint on the mask using KL divergence. Additionally, we introduce a consistency constraint through calculating L2 regularization calculation to improve model robustness. The regularization loss can be computed using Eq. (6).

$$L_{RKL} = \frac{1}{N} \sum_{i \in N} \|KL(P_i^c \parallel P_i^t) - KL(P_i^t \parallel P_i^c)\|_2 \quad (6)$$

where P_i^t is the probability map of a label, and i represents the index of each pixel, P_i^c and P_i^t are the masks generated by CNN and Transformer, respectively.

The Semantic Distillation Loss can be computed by Eq. (7):

$$L_{SD} = L_{KL} + L_{RKL} \quad (7)$$

2.4. Loss Function

The total loss function of our method is defined as Eq.(8), which integrates Dice Loss, Cross-Entropy (CE) Loss, and Semantic Distillation Loss.

$$L_{total} = L_{Dice} + L_{ce} + \lambda L_{SD} \quad (8)$$

where λ is a hyper-parameter, L_{SD} is Semantic Distillation Loss, L_{Dice} is the Dice Loss, L_{ce} is Cross-Entropy Loss, and L_{total} is the total loss.

3. EXPERIMENTS

This study conducts experiment analysis on the KiTS19 dataset [20], which comprises 210 annotated abdominal CT images that depict kidneys and kidney tumors. We randomly divided 168 cases as training data and 42 cases as testing data. To improve training efficiency, each case was sliced into images of size 384x384. Considering the low contrast in the kidney CT images and the small kidney region in the CT images, we applied normalization and pixel truncation techniques to reduce background noise. Specifically, pixel values were truncated to a range between -200 and 300.

All experiments are implemented using PyTorch 1.8 framework on two RTX 2080Ti GPUs (11GB) with a batch size of 8 and the number of training epochs to 60. To improve the generalization ability of DCCTNet, we apply various data augmentation techniques, including scaling, random rotation, and flipping. This paper adopts the same evaluation metrics as the KiTS19 challenge, which include the Dice Score, Volumetric Overlap Error (VOE) and Relative Volume Difference (RVD). Hyper-parameters are empirically set $\lambda=0.4$.

3.1. Experimental Results

3.1.1. Ablation Study

The objective of this experiment is to analyze the respective contributions of different modules to the network. From Table 1, it is noted integrating a Feature Select Module (FSM) into this hybrid structure results in an impressive improvement of 2.9%, employing Semantic Distillation Loss (SDL) achieves a notable increase of 2.1% in the Dice score. These results demonstrate the effectiveness of our method.

Table 1. Ablation studies of different components.

Method	Kidney Tumor		
	Dice	VOE	RVD
Baseline	0.766	0.205	0.031
Baseline + FSM	0.795	0.129	-0.025
Baseline + SDL	0.787	0.176	-0.006
DCCTNet	0.812	0.122	-0.003

3.1.2. Comparative Experiments with Transformers

To validate the superiority of DCCTNet, we conduct comparative segmentation experiments on kidney and tumor segmentation. We compared DCCTNet with several mainstream Transformer-based segmentation methods. The experimental results are presented in Table 2.

Table 2. Comparative experiments of mainstream Transformer-based segmentation methods.

Method	Kidney Tumor			Kidney		
	Dice	VOE	RVD	Dice	VOE	RVD
TransUNet[15]	0.765	0.023	0.096	0.972	0.032	0.012
Swin-UNet[14]	0.637	0.223	-0.031	0.970	0.031	0.005
DStransUNet[17]	0.686	0.306	0.165	0.943	0.031	-0.044
MISSFormer[16]	0.501	0.332	-0.053	0.965	0.035	0.003
HiFormer[18]	0.709	0.183	-0.093	0.965	0.032	-0.002
DCCTNet	0.812	0.122	-0.003	0.973	0.025	-0.002

As represented in Table 2, our approach outperforms other hybrid models in both kidneys and tumors segmentation tasks. For tumor segmentation, we achieve a Dice score of 0.812, which significantly surpasses other models and achieves the smallest errors in VOE and RVD. For kidneys segmentation, we achieves a Dice score of 0.973 and the superior results in RVD and VOE error metrics, thereby demonstrating the effectiveness of our method.

3.1.3. Comparative Experiments with CNNs

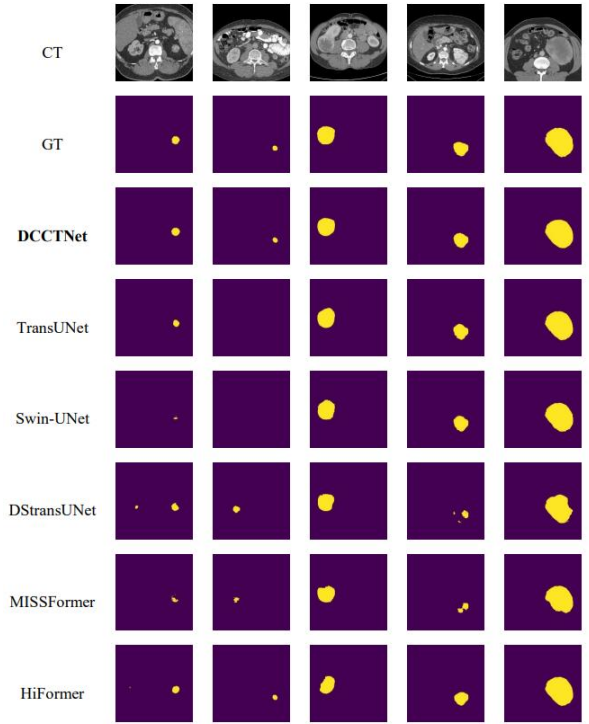
To validate the advancement of our method, we compare it with the latest methods for segmenting kidneys and tumors based on Dice criterion. The experiment results are illustrated in Table 3.

Table 3. Comparisons with other methods for segmenting kidneys and kidney tumors.

Method	Kidney Dice	Kidney Tumor Dice	Composite Dice
BiSC-UNet [6]	0.954	0.741	0.847
MSS UNet[7]	0.969	0.805	0.887
DoDNet[9]	0.965	0.775	0.870
RAU-Net[8]	0.962	0.768	0.865
RBCA-Net[5]	0.941	0.715	0.842
FS-Net[10]	0.966	0.768	0.867
DCCTNet	0.973	0.812	0.892

The results from Table 3 demonstrates the superior performance of DCCTNet in segmenting both kidneys and kidney tumors compared to other methods. Our approach combines CNN with Transformer to effectively learn global information at the feature and mask levels. In contrast, the latest kidneys and kidney tumors segmentation models. Despite these method achieve notable achievements, they were still limited by the small and fixed receptive field of CNN, our method achieves the best segmentation performance against others by incorporating Transformer networks, Feature Selection Module (FSM) and Semantic Distillation Loss (SDL).

3.1.4. Visualization Experiment

**Fig. 4.** Examples of the kidney tumors segmentation results

To intuitively justify the superiority of our method, we conducting visualization experiments in the kidney tumors segmentation. The results of kidney tumors segmentation are shown in Fig. 4. the CT, GT and DCCTNet denote the input kidney slices, Ground Truths(GT) and segmentation results of DCCTNet, respectively. We can see from the CT images of kidney tumors, the boundaries of the tumors are blurred and their shapes vary greatly, some tumors are too small to be clearly visible. Our method appears to be more similar to Ground Truth. Those processed by Swin-UNet.

DStransUNet TransUNet, HiFormer and MISSFormer have large deviations from the Ground Truth. This finding validates that hybrid models are more effective at capturing local information and global information. Consequently, segmentation performance is significantly superior to the comparative models.

4. CONCLUSIONS

In this study, we propose a dual-level hybrid model, called DCCTNet, for segmenting kidneys and kidney tumors. DCCTNet integrates the specific characteristics of CNN, multi-scale fusion and Transformer into a unified framework. FSM and SDL modules are employed in DCCTNet to leverage global information from the Transformer. DCCTNet enhances the modeling capacity of both local spatial information and global contextual information. It outperforms traditional CNNs, Transformers and hybrid model methods, achieving a Dice score of 0.973 in kidney segmentation and 0.812 in kidney tumor segmentation. These two experiments proved that the proposed method surpasses most state-of-the-art networks in kidneys and kidney tumors image segmentation.

5. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (Grant 62361043).

6. REFERENCES

- [1] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, et al, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge", *Medical image analysis*, vol. 67, pp. 101821, 2021.
- [2] Robert J Gillies, Paul E Kinahan, Hedvig Hricak, "Radiomics: images are more than pictures, they are data", *Radiology*, vol. 278, no. 2, pp. 563–77, 2016.
- [3] Taehun Kim, Hyemin Lee, Daijin Kim, "Ucanet: Uncertainty augmented context attention for polyp segmentation", *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2167–2175, 2021.
- [4] Gu Zaiwang, Cheng Jun, Fu Huazhu, Zhou Kang, Hao Huaying, Zhao, Yitian, et al, "Ce-net: Context encoder network for 2d medical image segmentation", *IEEE transactions on medical imaging*, vol. 38, no. 10, 2019.
- [5] Hwang Gyeongyeon, Yoon Hakyoun, Ji Yewon, Lee Sang Jun, "RBCA-Net: Reverse Boundary Channel Attention Network for Kidney Tumor Segmentation in CT images", *International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 2114–2117, 2022.
- [6] Chuanxia Wang, Yuting He, Xiaoming Qi, Zhao Ziteng, Guanyu Yang, Xiaomei Zhu, et al, BiSC-UNet: A fine segmentation framework for kidney and renal tumor, *Kidney Tumor Segmentation Challenge: KiTS19, Miccai*, 2019.
- [7] Wenshuai Zhao, Dihong Jiang, Jorge Peña Queraltó, Tomi Westerlund, MSS U-Net: 3D segmentation of kidneys and tumors from CT images with a multi-scale supervised U-Net, *Informatics in Medicine Unlocked*, vol. 19, pp. 100357, 2020.
- [8] Guo Jingna, Zeng Wei, Yu Sen, Xiao Junqiu, RAU-Net: U-Net model based on residual and attention for kidney and kidney tumor segmentation, *IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pp.353–356, 2021.
- [9] Zhang Jianpeng, Xie Yutong, Xia Yong, Shen Chunhua, Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1195–1204, 2021.
- [10] Zhu Xi-Liang, Shen Hongbin, Sun Haitao, Duan, Li Xia, Xu, Yingying, Improving segmentation and classification of renal tumors in small sample 3D CT images using transfer learning with convolutional neural networks, *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no.7, pp. 1303–1311, 2022.
- [11] Shen Zhiqiang, Yang, Hua, Zhang Zhen, Zheng Shaohua, Automated kidney tumor segmentation with convolution and Transformer network, *International Challenge on Kidney and Kidney Tumor Segmentation*, pp. 1–12, 2021.
- [12] Jiangyun Li, Wenxuan Wang, Chen Chen, Tianxiang Zhang, Sen Zha, Hong Yu, et al, Transbtsv2: Wider instead of deeper Transformer for medical image segmentation, *arXiv preprint arXiv:2201.12785*, 2022.
- [13] Yang Jingyi, Jiao Licheng, Shang Ronghua, Liu Xu, Li Ruiyang, Xu Longchang, EPT-Net: Edge Perception Transformer for 3D Medical Image Segmentation, *IEEE Transactions on Medical Imaging*, vol. 42, no. 11, pp. 3229–3243, 2023.
- [14] Cao Hu, Wang Yueyue, Chen Joy, Jiang Dongsheng, Zhang Xiaopeng, Tian Qi, et al, Swin-unet: Unet-like pure Transformer for medical image segmentation, *ECCV 2022 Workshops*, pp. 205–218, 2023.
- [15] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, et al, Transunet: Transformers make strong encoders for medical image segmentation, *arXiv preprint arXiv:2102.04306*, 2021.
- [16] Huang Xiaohong, Deng Zhifang, Li Dandan, Yuan Xueguang, Fu Ying, MISSFormer: An Effective Transformer for 2D Medical Image Segmentation, *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1484–1494, 2022.
- [17] Lin Ailiang, Chen Bingzhi, Xu Jiayu, Zhang Zheng, Lu Guangming, Zhang David, Ds-transunet: Dual swin Transformer u-net for medical image segmentation, *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.
- [18] Moein Heidari, Amirhossein Kazerouni, Milad Soltany Kadarvish, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, et al, HiFormer: Hierarchical multi-scale representations using Transformers for medical image segmentation, *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6202–6212, 2023.
- [19] Geoffrey E. Hinton, Oriol Vinyals, Jeffrey Dean, "Distilling the knowledge in a neural network", *arXiv preprint arXiv:1503.02531*, 2015.
- [20] N. Heller, Niranjana Jude Sathianathan, Arveen A. Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, et al, The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes, *arXiv preprint arXiv:1904.00445*, 2019.