Supplementary: Causally-Aware Intraoperative Imputation for Overall Survival Time Prediction

The supplementary document is organized as follows:

- Sec. 1 makes a brief summary of terms and priories.
- Sec. 2 provides a statistical description of data.
- Sec. 3 describes the primary findings in our pilot study.
- Sec. 4 elaborates the network architectures.
- Sec. 5 provides more experimental results.
- Sec. 6 shows more visualization results.
- Sec. 7 provides the theoretical analysis.
- Sec. 8 provides some insightful discussion of the potential of this work.

1. Preliminaries

In this section, we will make a brief summary of the technical terms, and medical priories related to this work.

1.1. Technical Terms

Generally, structured MRIs are typically composed of the following six modalities:

- T_1-WI: T_1-weighted imaging T_1.
- T_2-WI: T_2-weighted imaging T_2.
- DWI: Defusion weighted imaging.
- ADC: Apparent deffusion coefficient.
- T_1-IP/OP: T_1 in phase / out of phase.
- T_1ce: Contrast-enhanced T_1 at the arterial phase (T_1ce-AP), portal phase (T_1ce-PP), and lag phase (T_1ce-LP).

The surgery-related indicators and the abbreviations are as follows:

- AFP: Alpha-fetoprotein.
- TBil: Total bilirubin.
- ALT: Alanine aminotransferase.
- GGT: Gamma-Glutamyltransferase.
- HBs-Ag: Hepatitis B virus surface antigen.
- HBV-DNA: Hepatitis B DNA.
- HCV-Ab: Hepatitis C virus antibody.
- PT: Prothrombin time.
- G-Score: A metric for evaluating the severity of inflammation. According to *Metavir scoring system* [1], it has five levels from G0 to G4.
- S-Score: A metric for evaluating the severity of fibrosis. According to *Metavir scoring system*, it has five levels from S0 to S4.

1.2. Medical Priories

Magnetic Resonance Imaging (MRI). MRI is an imaging method that radiates energy from substances in the body to the surrounding environment through high-frequency magnetic field in vitro [3]. Generally speaking, MRI is the most important reference for the early diagnosis of liver cancer. The appearance of the tumor will show different brightness or contrast against the whole liver on different modality of MRI, which provides the most significant indication for doctors to judge the priority of cancer at the stage of early diagnosis.

Surgery-related indicators. In fact, the overall treatment cycle for primary liver cancer tend to extend from years up to several decades, in which tumor resection surgery is only one of those complex procedures. Taking the first surgery as the dividing line, we summarize all indicators into three categories:

- Preoperative – collected before the surgery.

- Intraoperative – collected during the surgery.

- Postoperative – collected after the surgery.

Concretely, in addition to the patient's basic information *e.g.* age and gender, the preoperative indicators mainly derive from the results of blood tests *e.g.* alpha-fetoprotein, hepatitis B antigen test, and hepatitis C antigen test, *etc,*. The intraoperative indicators include direct records during the surgery *e.g.* tumor number, operation duration, bleeding, *etc,* and pathological analysis results obtained from samples collected during the surgery *e.g.* cell type, differentiation, clinicopathological cirrhosis, *Metavir* Metrics, *etc,*. Postoperative information includes the adjuvant therapy adopted by the doctor and the recurrence of the patient.

Typical diagnostic process of liver cancer. A typical diagnostic process of liver cancer usually begins with the patient's report of symptoms (*e.g.* abdominal pain) and abnormity in blood test indicators (*e.g.* high alpha-fetoprotein). Combined with MRI, doctors are supposed to pass judgement on the severity of the tumor, and thus make a basic estimation on the overall survival (OS) time, *e.g.* longer or shorter than five years. However, on one hand, the information available for early diagnosis is too limited. On the other hand, there are too many uncertain factors in the whole span from early diagnosis to cancer-related death, such as the surgical level, adjuvant therapy, and the lifestyle of the individual. These factors make it difficult for doctors and clinicians to make an accurate and confident estimation on OS time.

2. Data Description

In this section, we will provide a statistical feature description for either numeric (Tab. 1) and categorical variables (Fig. 1) available in our own dataset. As far as we know, there is **NO** open source tumor dataset that contains sufficient intraoperative information. Therefore, we unfortunately cannot test our proposed method on any classical data set for the time being. Note that, the dataset on which we train and test is basically in-house, and will **NOT** be open source in the short run.

There are a total of 11 preoperative and 20 intraoperative variables in our dataset. All the preoperative variables except *Gender* are in numeric type. 11 of the intraoperative variables are in categorical type, while 9 are in numeric type. We performed feature statistics on the 361 samples after primary screening. Tab. 1 shows the statistical feature description of all the preoperative and intraoperative variables of numeric type, in terms of *Min*, *Max*, *Mean*, and *Std*. Fig. 1 shows the distribution of categorical variables, with the number of samples in each category.

Table 1. Statistical Feature Description of Numeric Variables.

Variable	Attribute	Min	Max	Mean±Std
Age	Preoperative	21	85	53.7±11.7
AFP		1.0	60500.0	1726.7±8115.6
Albumin		32.0	50.8	40.8±3.1
TBil		3.8	38.2	12.2±5.0
GGT		11.0	955.8	77.1±88.9
HBs-Ag		0.0	10334.0	4591.6±3108.6
HBV-DNA		$1.6e3$	$1.7e6$	$(1.5\pm2.4)e5$-
HCV-Ab		0.0	28.0	0.2±0.2
PT		10.2	21.0	12.2±1.0
Ascites	Intraoperative	0.0	200.0	2.6±14.6
Cirrhosis Nodules		0.1	0.9	0.4±0.1
Tumor1 Diameter		0.6	19.0	4.1±2.7
Tumor2 Diameter		0.3	5.0	1.7±1.0
Tumor3 Diameter		0.2	2.5	1.0±0.6
Sum of Tumor Diameter		0.6	21.3	4.4±3.0
Boundary		0.8	3.0	1.9±0.9
Vessel Bleeding		0.0	1000.0	151.1±133.5
Portal Occlusion		0.0	15.0	6.2±6.6

3. Pilot Study

In this section, we will describe the findings in our pilot study. Our proposed model (CAWIM) and novel methodology (CaDAG) are basically motivated by these preliminary conclusions.

3.1. Experiments

OS time prediction. We set a baseline model using only MRI and preoperative information for the OS time classification model and trained on a single fold with T_1ce-AP modality as *MRI+Pre*. On the other hand, we use only ground-truth intraoperative indexes for both training and testing, as *(gt)Intra*. Then we directly applied ground-truth intraoperative information to the model *MRI+Pre* by concatenating the indexes together, as *MRI+Pre+(gt)Intra*.

Correlation of intraoperative indexes. We are informed that the intraoperative indexes are related with one another. So, we calculated the correlation coefficient between each pair of intraoperative variables after zero-mean normalization.

3.2. Primary Findings

Significance of intraopertive indexes. Tab. 2 shows the performance of (a) *MRI+pre*, (b) *(gt)Intra*, and (c) *MRI+Pre+(gt)Intra*. We found that intraoperative indicators should play a crucial role in improving OS time prediction, with the support of the facts as follows:

- Using only intraoperative indexes (row (b)) can improve the model performance by 5.25% compared to the baseline model (row (a)) on F_1-score.

- Concatenating all the feature indexes together boosts the optimal performance (row (c)) by a promotion of 9.11% compared to using MRI and preoperative indexes (row (a)), and 3.86% to using only intraoperative indexes (row (b)).

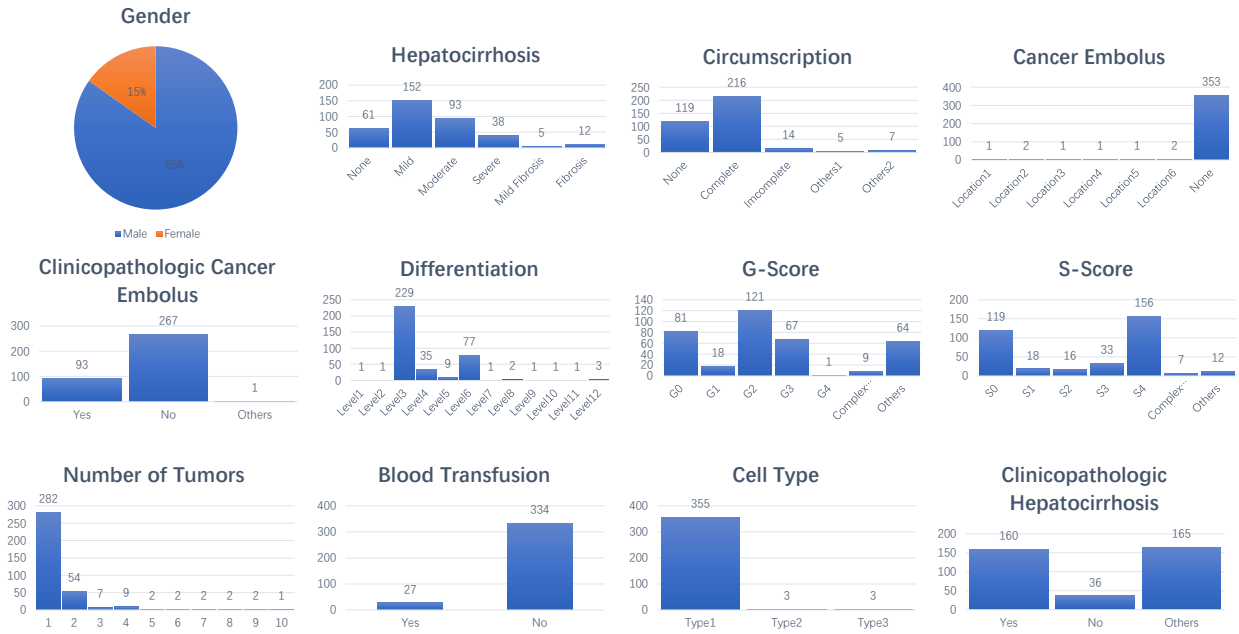

Figure 1. Statistical Feature Description of Categorical Variables.

Table 2. Model Performace on a **SINGLE** Fold with Ground Truth Intraoperative Indexes on T_1-ce AP modality.

	Model	Precision	Recall	F_1-Score
(a)	MRI+Pre	37.86	34.32	35.02
(b)	(gt)Intra	45.42	44.78	40.27
(c)	MRI+Pre+(gt)Intra	**54.58**	**46.06**	**44.13**

Correlation of intraopertive indexes. We found that there are adequate but intricate correlations between certain pairs of intraoperative variables. Fig. 2 shows the correlation heatmap of all the intraoperative indexes available to us. We observed the following facts:

- Among all the variables, the most notable positive correlation is observed between *Sum of Tumor diameter* and *Tumor1 diameter* ($r = 0.83$), as well as *G-score* and *S* ($r = 0.83$).
- There are relatively obvious correlation between *Hepatocirrhosis* and *Cirrhosis nodes* ($r = 0.54$), *Number of tumors* and *Tumor2 diameter* ($r = 0.58$), *Number of tumors* and (*Tumor3 diameter*) ($r = 0.63$), *Tumor2 diameter* and *Tumor3 diameter* ($r = 0.49$).
- Among all the negatively correlated variable pairs, the correlation coefficients between *Blood transfusion* and *G-score* ($r = -0.41$) as well as *S-score* ($r = -0.36$) are the most prominent.

3.3. Interpretation and Analysis

The validity of intraoperative information lies in that it is collected during the operation and often better describes tumor related characteristics than those at the earlier stage *e.g.* MRI images and preoperative information. Therefore, it can better indicate OS time when directly applied to the classification model. However, the intraoperative indexed cannot be obtained during at the early stage, thus cannot be leveraged in the inference phase.

On the other hand, most of the correlation we observe are in line with common sense, or can be explained by medical prior knowledge. *i.e.* the correlation between the diameter of each tumor and the sum of them all is in line with common sense. While most of the patients (64.8%) only have one tumor, *Tumor1 diameter* should contribute most to *Sum of Tumor diameter*; *G-score* and *S-score* are naturally highly correlated, according to the definition by *Metavir scoring system* representing respectively for the level inflammation and fibrosis. Nevertheless, it is hard to explain the negative correlation between *Blood transfusion* and either *Metavir* metric (*G-score* and *S-score*), to our best known.

4. Network Architectures

We removed the final fully-connect layers of the official Resnet-34 network as our image encoder. In causally-aware intraoperative reasoning, we used FC-BN-RELU (Fully-connect layer, batch norm and RELU activation) to down-

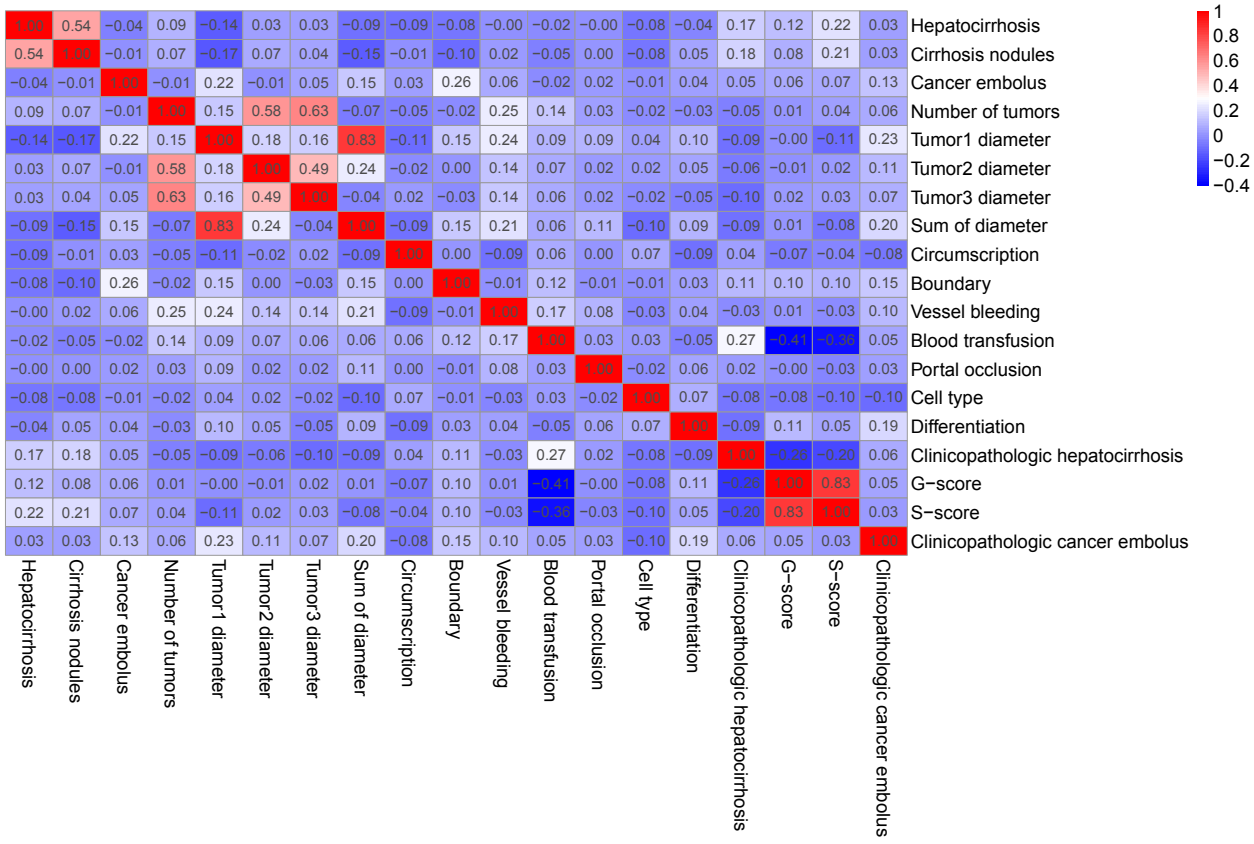

Figure 2. Correlation Heatmap of all Intraoperative Indexes.

sample the image feature to 85 dimension and then concatenated it with 11 preoperative features to predict the intraoperative variables, where we utilized a single fully-connect layer as our predict netwotk. Besides, we also used similar structure to predict the OS time. We firstly employed FC-BN-RELU to downsample the image feature to 85 dimension and then concatenated it with 11 preoperative features and other prediceted intraoperative features to predict the OS time.

5. More Results

In this section, we provide more experimental results based on our proposed CAWIM.

Different modalities. We apply our CAWIM on each modality of MRI, *i.e.* (a) T_1-IP, (b) T_1-WI, (c) T_2-WI, and (d) T_1ce-AP. The results are shown in Tab. 3. We found that the *T_1ce-AP* performs best on Precision (45.58%) and F_1-score (42.21%), while *T_1-WI* best on Recall (45.61%). Therefore, we decide on T_1ce-AP to conduct subsequent experiments.

Theoretical upper-bound. Intuitively, training the model with ground-truth intra-operative information can reach a theoretical upper-bound to our approach. The results are listed in Tab. 4. Our model performs slightly below this theoretical upper limit. It is intuitive and reasonable, as the gap is attributed to the error in the intermediate process of predicting intra-operative information. This further verifies the efficacy of our method.

Other baselines. Tab. 5 shows some other baselines from previous work. It can be seen that our CAWIM performs much better than these methods.

Table 3. The Average Model Performance on **FIVE** Folds on Different Modalities of MRI.

	Modality	Precision	Recall	F1-score
(a)	T_1-IP	33.74	37.81	32.90
(b)	T_1-WI	40.19	**45.61**	41.32
(c)	T_2-WI	32.65	34.50	31.31
(d)	T_1ce-AP	**45.58**	43.70	**42.21**

Table 4. Theoretical upper-bound of our proposed method.

Model	Precision	Recall	F_1-Score
MRI + Pre. + Intra. (GT)	50.11±3.56	42.31±2.50	42.23±1.37
Ours	45.58±5.86	43.70±5.89	42.21±4.92

Table 5. Other baselines.

Model	Precision	Recall	F_1-Score
MCAN [9]	32.45	33.16	32.78
MBT [6]	35.65	37.23	36.40
GCN [4]	25.96	23.74	22.47
Bilinear Pooling [5]	34.75	35.06	34.19
Ours	**45.58**	**43.70**	**42.21**

6. More Visualization

In this section, we provide more visualization results to demonstrated the effectiveness of our model. Fig. 3 shows the high-response area with and *w/o* our CaDAG. Fig. 4 shows the developing process of high-response area in the training phase, in which the high-response area gradually focuses on the liver and tumor related areas with the progress of the training process. It can be drawn that our CaDAG managed to guide the model to locate on liver-related regions.

7. Theoretical Analysis

For completeness, we first introduce basic assumptions and the algorithm for identifying $\mathbf{M}_i$ in the i-th random splitting, in our main context.

Assumption 1 (Causal Graph)**.** *We assume the causal graph over* $\mathbf{C}$ *is a directed acyclic graph (DAG) and denote the corresponding SCM as* $\mathcal{M} := \langle G := (\mathbf{C}, \mathbf{E}), \mathcal{F}, P(\varepsilon) \rangle$.

Assumption 2 (Markovian and Faithfulness)**.** *For triplets of disjoint sets* $\mathbf{V}_i, \mathbf{V}_j, \mathbf{V}_k$*, it holds that* $\mathbf{V}_i \perp_d \mathbf{V}_j | \mathbf{V}_k \leftrightarrow \mathbf{V}_i \perp \mathbf{V}_j | \mathbf{V}_k$*, where* $\perp_d$ *and* $\perp$ *respectively mean d-separation and probability independence.*

Assumption 3 (Distributional Faithfulness)**.** *If* $X_i \to X_j$ *and at least* $E \to V_i$ *or* $E \to V_j$ *holds,* $\{P^e(V_i|V_j, \mathbf{Z})\}$ *is dependent to* $\{P^e(V_j|\mathbf{Z})\}$*, where* $\mathbf{Z}$ *denotes the minimal deconfounding set* [1].

Algorithm 1 Identify causal directions among $\mathbf{M}$.

INPUT: E; skeleton and M via Alg. 1 in main context.
OUTPUT: Directed graph among $\mathbf{M}$.

1: For each adjacent (A_i, A_j) such that one of $V_i \in \mathbf{M}$,
2: Detect deconfounding set $\mathbf{Z}$.
3: Calculate $\hat{\Delta}_{A_i \to A_j | \mathbf{Z}, E} > \alpha$ and $\hat{\Delta}_{A_j \to A_i | \mathbf{Z}, E} > \alpha$.
4: Determine $A_i \to A_j$ or $A_j \to A_i$.

Lemma 1. *Under assumps 1-3, for each random splitting* i*, we have that* $\mathbf{M}_i, \mathrm{De}(\mathbf{M}_i), \mathrm{PA}(\mathbf{M}_i), \mathrm{PA}(\mathrm{De}(\mathbf{M}_i))$ *are identifiable.*

[1] $\mathbf{Z}$ is a deconfounding set between V_i and V_j if $V_i \perp V_j | \mathbf{Z}$ and $\mathbf{Z} \cap (\mathrm{De}(V_i) \cup \mathrm{De}(V_j)) = \emptyset$.

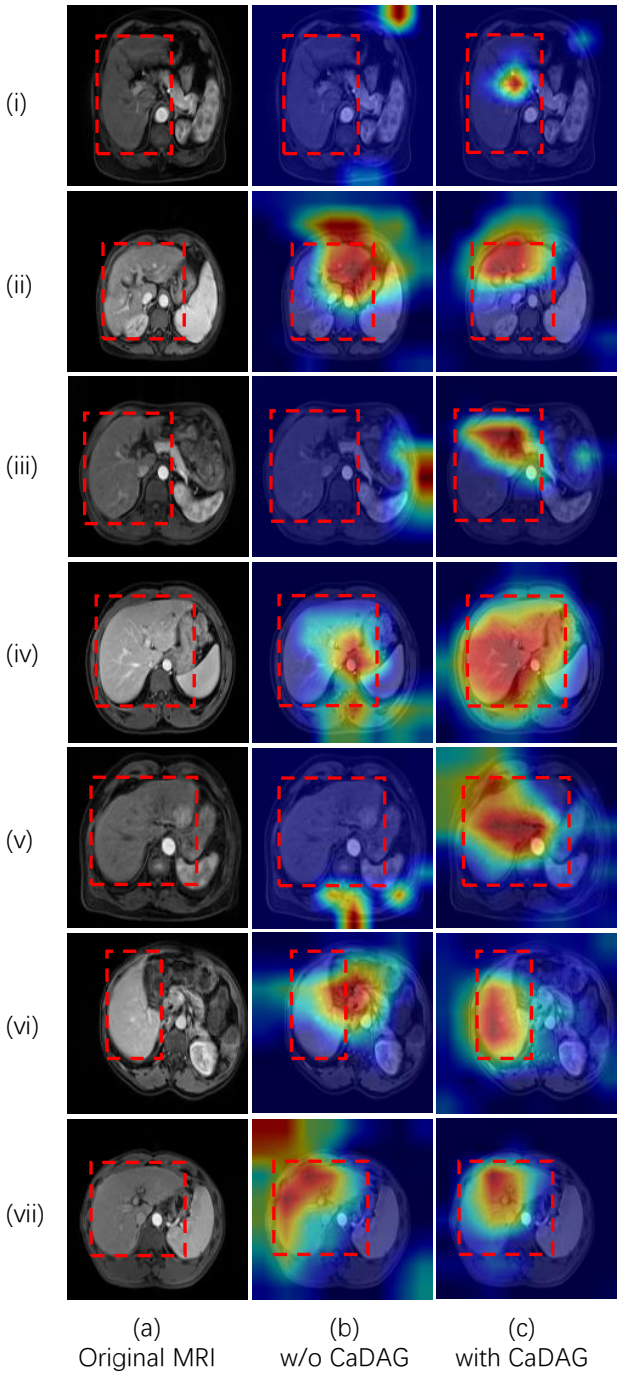

Figure 3. More Results Visualization of Heat Maps [8] with and *w/o* CaDAG.

Proof. Identification of $\mathbf{M}_i$ has been shown in Alg. 1. **We first show the identification of** $\mathrm{De}(\mathbf{M}_i)$.

- Line 5 to 10 are based on (i) the structure $E \to \cdots \to$

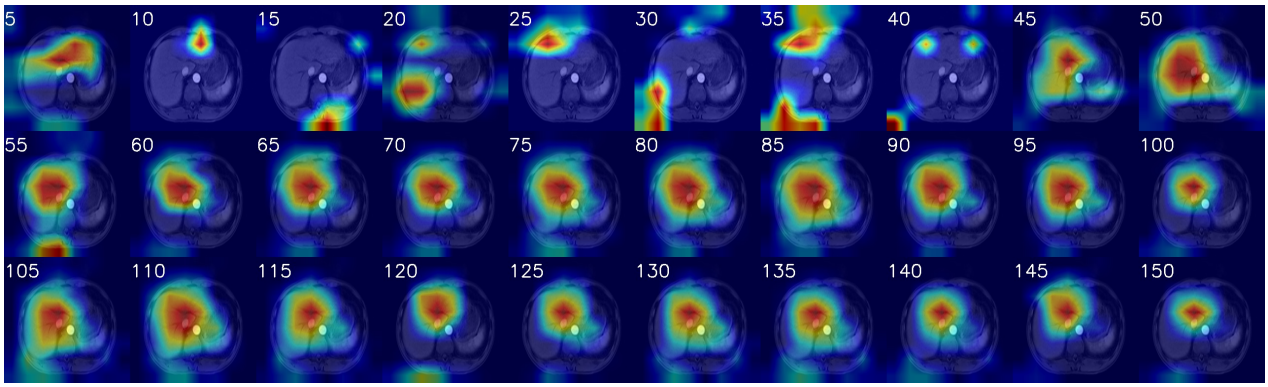

Figure 4. Visualization of the Development of High-response Area with CaDAG during Training.

Algorithm 2 Detection of $\mathrm{De}(\mathbf{M}_i) \cup \mathbf{M}_i$

Start with $\mathbf{C} = \mathbf{D} = \mathbf{M}_i$ and $\mathrm{visited}(V_i)$ =FALSE
while $\mathbf{D} \neq \emptyset$ **do**
  for $A_j \in \mathbf{D}$ **do**
    for $A_i \in \mathrm{Adj}(A_j)$ **do**
      if $A_i \notin \mathbf{M}_i$ and $A_i \perp E | \mathbf{C}_{e,x_i} \cup \{A_j\} \setminus \mathbf{D}_{a_i,e}$ **then**
        $\mathbf{C} = \mathbf{C} \cup \{A_i\}$
        if $\mathrm{visited}(A_i)$ =FALSE **then**
          $\mathbf{D} = \mathbf{D} \cup \{A_i\}$
        end if
      end if
      if $A_i \in \mathbf{M}_i$ and $\widehat{\Delta}_{A_j \to A_i} < \widehat{\Delta}_{A_i \to A_j}$ **then**
        $\mathbf{C} = \mathbf{C} \cup \{A_i\}$
        if $\mathrm{visited}(A_i)$ =FALSE **then**
          $\mathbf{D} = \mathbf{D} \cup \{A_i\}$
        end if
      end if
    end for
    Let $\mathbf{D} = \mathbf{D} \setminus \{A_j\}$
  end for
end while

$A_j - A_i$ and (ii) A_i and E are not adjacent.

- Line 11 to 19: In this case, we identify the direction between A_i and A_j by the "Independent Causal Mechanism (ICM) Principle" following [2], where $\widehat{\Delta}_{A_j \to A_i}$ and $\widehat{\Delta}_{A_i \to A_j}$ are the estimated HSIC (see Eq. 17 in [2] for the detailed formulation of $\widehat{\Delta}$).

  The ICM principle means that "the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.". That is, the changes of $P(A_i|\mathrm{PA}(A_i))$ does not influence the other mechanisms $P(A_j|\mathrm{PA}(A_j))$ for $j \neq i$. The ICM principle is implied in the definition of "structural causal model" in [7], where each structural equation represents an autonomous physical mechanism.

Next, we identify $\mathrm{PA}(\mathbf{M}_i), \mathrm{PA}(\mathrm{De}(\mathbf{M}_i))$.

Algorithm 3 $\mathrm{PA}(A_i)$ for $A_i \in \mathbf{M}_i \cup \mathrm{De}(\mathbf{M}_i)$.

for $A_j \in \mathbf{M}_i \cup \mathrm{De}(\mathbf{M}_i)$ **do**
  for $A_i \in \mathrm{Adj}(A_j)$ **do**
    if $A_i \notin \mathbf{M}_i$ **then**
      $A_i \in \mathrm{PA}(A_j)$ if $A_i \not\perp E | \{\mathbf{C}_{e,A_i} \setminus \mathbf{D}_{e,A_i} \cup \{A_j\}\}$
    else if $A_j \in \mathbf{M}_i$ and $A_i \in \boldsymbol{X_m}$ **then**
      $A_i \in \mathrm{PA}(A_j)$ when $\widehat{\Delta}_{A_j \to A_i} < \widehat{\Delta}_{A_i \to A_j}$.
    else if $A_j \in \mathbf{M}_i$ and $A_i \notin \boldsymbol{X_m}$ **then**
      $A_i \in \mathrm{PA}(A_j)$ when $E \not\perp A_i | \mathbf{C}_{A_i,e} \cup \{A_j\}$
    end if
  end for
end for

- Line 4: this rule is based on the structure $E \to \cdots \to A_j - A_i$ and $\{A_i, E\}$ are not adjacent.
- Line 6: this rule is based on the HSIC criterion in [2].
- Line 8:this rule is based on the structure $E \to A_i - A_j$ and $\{E, A_j\}$ are not adjacent.

□

Theorem 2. *Under assumptions 1, 2, 3, the learned graph via our CaDAG is a directed acyclic graph.*

Proof. For each random splitting i, we can detect edges that are adjacent to $\mathbf{M}_i$. Then as long as m is large enough such that $\cup_i \mathbf{M}_i = \mathbf{A}$, we can identify all edges among $\mathbf{A}$. This can be achieved by randomness of random splittings. In other words, for each variable $A \in \mathbf{A}$, there exists a random splitting such that the domain variable is dependent to A. It is left to show that the learned graph is a DAG. If the learned DAG has cycles $A_{i_1} \to A_{i_2} \to \ldots \to A_{i_1}$, we orient $(A_{i_k}, A_{i_{k+1}})$ with the minimum frequency among $\{f_{i_k,i_{k+1}}, f_{i_N,i_1}\}$. If this orientation induces another cycle, then we repeat our orientations, until there is no cycle in our learned graph. □

8. Potential of Transfer Learning

In this work, we proposed a general framework for OS time prediction that leveraged the causality of the intraoperative information. Although we had conducted adequate research on medically relevant preliminaries, our proposed CaDAG methodology is essentially data-driven. In other words, the causal relationship among the intraoperative variables are actually mined from the original data with limited artificial interference. On the contrary, the causal structure inferred from the CaDAG can further be considered as prior knowledge to guide the prediction of deep learning models. Therefore, we believe that the methodology of CAWIM can be transferred to various tasks in other fields with little or simple fine-tuning.

References

[1] Pierre Bedossa and Thierry Poynard. An algorithm for the grading of activity in chronic hepatitis c. *Hepatology*, 24(2):289–293, 1996. 1

[2] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(89):1–53, 2020. 6

[3] Scott A Huettel, Allen W Song, Gregory McCarthy, et al. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004. 1

[4] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 5

[5] Tsung-Yu Lin and Subhransu Maji. Improved bilinear pooling with cnns. *arXiv preprint arXiv:1707.06772*, 2017. 5

[6] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 5

[7] Judea Pearl. *Causality*. Cambridge university press, 2009. 6

[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 5

[9] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 5