M³Care: Learning with Missing Modalities in Multimodal Healthcare Data

Chaohe Zhang∗
wangyasha@pku.edu.cn
Key Lab of High Confidence Software Technologies, Ministry of Education
School of Computer Science, Peking University
Beijing, China

Xu Chu∗
Department of Computer Science and Technology, Tsinghua University
Key Lab of High Confidence Software Technologies, Ministry of Education
Beijing, China

Yasha Wang†
wangyasha@pku.edu.cn
Key Lab of High Confidence Software Technologies, Ministry of Education
National Engineering Research Center of Software Engineering, Peking University
Beijing, China

Jiangtao Wang
Center for Intelligent Healthcare, Coventry University
Coventry, UK

Liantao Ma
yinhao Zhu
Junfeng Zhao

Yasha Wang
Chaohe Zhang
Liantao Ma
Yinghao Zhu
Jiangtao Wang
Junfeng Zhao

Corresponding Author.

ABSTRACT
Multimodal electronic health record (EHR) data are widely used in clinical applications. Conventional methods usually assume that each sample (patient) is associated with the unified observed modalities, and all modalities are available for each sample. However, missing modality caused by various clinical and social reasons is a common issue in real-world clinical scenarios. Existing methods mostly rely on solving a generative model that learns a mapping from the latent space to the original input space, which is an unstable ill-posed inverse problem. To relieve the underdetermined system, we propose a model solving a direct problem, dubbed learning with Missing Modalities in Multimodal healthcare data (M³Care). M³Care is an end-to-end model compensating the missing information of the patients with missing modalities to perform clinical analysis. Instead of generating raw missing data, M³Care imputes the task-related information of the missing modalities in the latent space by the auxiliary information from each patient’s similar neighbors, measured by a task-guided modality-adaptive similarity metric, and then conducts the clinical tasks. The task-guided modality-adaptive similarity metric utilizes the uncensored modalities of the patient and the other patients who also have the same uncensored modalities to find similar patients. Experiments on real-world datasets show that M³Care outperforms the state-of-the-art baselines. Moreover, the findings discovered by M³Care are consistent with experts and medical knowledge, demonstrating the capability and the potential of providing useful insights and explanations.

CCS CONCEPTS
• Applied computing → Health informatics; • Information systems → Data mining.

KEYWORDS
healthcare informatics, multimodal data, electronic health record

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. 2022. M³Care: Learning with Missing Modalities in Multimodal Healthcare Data. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’22), August 14–18, 2022, Washington, DC, USA. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3534678.3539388

1 INTRODUCTION
Multimodal data can provide complementary information from various modalities that reveal the fundamental characteristics of real-world subjects [6, 8, 21, 52, 53]. Thus, many clinical applications, such as disease diagnosis and mortality prediction [21, 39, 44, 53, 59], require multimodal electronic health record (EHR) data to achieve good diagnostic or prognostic results. Conventional approaches usually assume that each sample is associated with the unified uncensored modalities, and all modalities are available for each sample [40, 58]. However, missing modality is a common issue...
Also, patients may lack some specific modalities due to patient (i.e., combinations of available modalities) make it more complex in the prior spatial (or spatio-temporal) correlations required to modalities refer to the samples, which are permutable. This results matrix are not permutable. Different from that, the rows in missing alization is from the matrix space, such that the row vectors in a as image [22] or geosensory data [54], often assumes a sample re-
in the matrix space. Thus the traditional imputation methods do than thereof in different modalities, exhibiting a coherent behavior features inside the same modality are naturally more correlated the same modalities is missing entirely. On the other hand, the missing modalities manifest themselves by column-wise consec-
tures) [35]. Whereas, as illustrated in the right part of Figure 1, be imputed via modeling correlations between the columns (fea-
 incoherence, confer [14] for a survey), and the missingness can complete the block-wise missingness are not present in the missing modalities. In a word, conceptually, there is a gap between the missing modalities (column-wise consecutive missingness) and the existing imputation-based methods for random or block-wise missingness.

In methodology, some deep generative methods [8, 33, 43, 45] are proposed for missing modalities. Essentially, such methods are usually based on the manifold assumption: the probability mass of real-world objects is supported on low-dimensional manifolds [1]. In terms of EHR data imputation, the manifold assumption can be interpreted as the low-rankness and stability of the feature covariance matrix. In other words, there exist a set of low-dimensional basis vectors and a deterministic mapping \( T \) subject to: (a) the basis vectors span the pre-image of the mapping \( T \), and (b) the observed EHR feature vectors live on the image of the mapping \( T \). Existing EHR generative-based completion methods tackle the problem by solving a generative model which learns a mapping from a latent space (spanned by basis vectors) to the original input space [8, 33, 43, 45]. Solving such mapping is essentially an ill-posed inverse problem (low- to high-dimensional underdetermined system [13]), whose solution is often non-unique and unstable [26]. On the other hand, such complex auxiliary models may introduce extra noise, which has negative impacts [9, 15, 51]. To this end, the problem of completing missing modalities requires a different way.

In fact, for missing modalities, solving the generative model is not necessary. By assuming the low-rankness and the stability of the covariance matrix of EHR features, the locally similar row vectors (patients with similar features from some uncensored modalities) in a sub-matrix imply globally similar row vectors (those patients should have similar missing features) in the data matrix. Moreover, if a local row vector \( X \) falls in a convex hull spanned by a set of local row vectors, then the global row vector corresponding to \( X \) is likely to fall into the convex hull spanned by the particular set of global row vectors. Thus, instead of solving the inverse problem of the low- to high-dimensional mapping, we can solve a less underdetermined problem: comparing similarities of local row vectors and impute the missing entries in a row by referring to the uncensored entries of locally similar rows. The similarity comparison can be conducted in the original input space with a sophisticated metric on data manifolds, or in a learned latent space with a more straightforward
metric. More importantly, modeling the similarity relationship in the low-dimensional latent space is a direct (or forward, namely) problem of solving a mapping from a high-dimensional space to a low-dimensional space, which is less complicated than the inverse problem of solving a generative model.

On the other hand, this intuition is also in agreement with the real-world clinical practice, i.e., how doctors use the relationships between patients to assist the clinical analysis. If two patients are similar in one clinical modality, they are more likely to be similar in another one. Thus, as shown in Figure 2, for a patient with missing modalities, we can utilize the other complete modalities of the patient (local row vector) and other patients who also have these modalities (local uncensored row vectors), to find similar patients. Although seeming straightforward, applying this intuition to clinical tasks will face the following challenges:

**Challenge 1.** Which space can be used to perform imputation? There are at least two options: the original input space or a learned latent space. Existing methods usually estimate missing data in the original input space. However, the probability distribution in the original input space contains task-relevant and task-irrelevant information. Imputation in the original input space treats task-irrelevant information equally, weakening the task-specific information conveyed in EHR data. This results in an indiscriminate loss of task-relevant and task-irrelevant information, leading to inferior performance.

**Challenge 2.** What metric(s) should be used to model the similarity relationship? T-LSTM uses autoencoders to generate patient representations, based on which the similarity is obtained. SMIL uses multivariate Gaussian to assign similarity weights. However, they did not associate the connotation of similarity with clinical tasks. In different clinical tasks (e.g., mortality prediction and disease diagnosis), the patient characteristics that need attention are different, so two patients considered similar in one clinical task may be considered not so similar in another. More importantly, the similarity metric might vary for different modalities. The manifolds in different modalities in the original input spaces are naturally equipped with different metrics. A learned deterministic mapping from the original input spaces to the learned latent space is not guaranteed to result in a unified metric for different mapped modalities in the latent space.

**Challenge 3.** How to infer the local-to-global similarities of patients? Metrics computing the similarities between local row vectors is not sufficient to describe the similarities of global row vectors. Thence modeling the intra- and inter-correlations of features from various modalities is challenging but indispensable to aggregate the information from different modalities.

By jointly considering the above issues, we propose a model learning with Missing Modalities in Multimodal healthcare data (M³Care), an end-to-end inductive learning model to compensate for the missing modalities and perform clinical tasks. In summary, our main contributions are summarized as follows:

- We propose M³Care to compensate for the modality-missing patient in the latent space and perform clinical tasks with EHR data.

Since the latent representations are highly compressed and task-supervised, this results in less loss of task-relevant information and is thus more beneficial for subsequent tasks in an end-to-end learning schema (Response to Challenge 1).

- Methodologically, (a) M³Care uses task-guided deep kernels in the latent space of each modality as the metric to compute patient similarities (Response to Challenge 2). (b) M³Care captures intra-correlations within each modality and inter-correlations between modalities by a self-attentive multi-modal interaction module so that the local metrics are aggregated to calculate the similarities of global row vectors (Response to Challenge 3).

- Extensive experiments show that M³Care outperforms all state-of-the-art models under multiple levels of incompleteness in different evaluation metrics. Besides, the findings discovered by M³Care are in accord with experts and medical knowledge, which shows it can provide useful insights and explanations.

## 2 RELATED WORK

**Multimodal learning for healthcare.** With the advancement of medical technology, comprehensive healthcare is burgeoning to meet the demands of patients. This has allowed for multiple medical modalities (e.g., medical image, clinical notes, etc.) to be analyzed to offer patients with feedback, as well as physicians with insights on clinical applications.

To this end, multimodal learning for healthcare has attracted the interest of researchers. For example, RAIM is proposed for jointly analyzing continuous monitoring data (e.g., ECG, heart rate) and discrete clinical events (e.g., intervention, lab test) to predict patient decompensation. Gao et al. utilize a multimodal inference model to jointly encode trial criteria text and patient EHR tabular data for patient-trial inference. Huang et al. develop and compare different multimodal fusion architectures to classify Pulmonary Embolism (PE) cases. Ma et al. and Hoang et al. develop distillation frameworks to leverage the multimodal EHR data to enhance the prognosis. Although the methods above work well, one common drawback is that they can only handle samples with complete modalities. Limitations exist while modeling multimodal interactions with the presence of missing modalities.

**Methods for missing modalities.** Currently, there have been research interests in handling missing modalities, which are mainly divided into two types: deleting incomplete samples or imputing missing modalities. For the first type, FitRec performs profile forecasting based on multimodal user data, which discards samples with missing modalities. Wang et al. propose a knowledge distillation framework on samples with complete modalities, while distilling the supplementary information from the incomplete ones. However, such methods cannot handle the samples in need with missing modalities and have limitations to be applied in rigid demand domains like healthcare. Besides, such methods dramatically reduce training data and result in over-fitting of deep learning models, especially when there are many modalities and many different missing combination patterns.

The second type is generating the missing modalities at first. However, the incompleteness of modalities leads to column-wise consecutive missingness of features, which makes traditional

---

*We also substantiate this intuition by mining the real-world clinical datasets, please refer to the Intuition discovery experiments in Appendix.*

*e.g., five modalities can result in $2^5 - 1 = 31$ missing patterns.*
methods like matrix completion can not be used [51]. Some advanced generative methods such as autoencoders [33, 43] and generative adversarial networks (GAN) [8, 45] have been proposed. These solutions, however, may introduce unwanted extra noise [9]. Especially when the size of samples with complete modalities is small, yet the number of modalities is large, the modalities imputed by such methods may have a negative effect [15, 51]. Moreover, while facing data with many modalities or missing patterns, the number of generators required is also large, which is difficult to train. Chen and Zhang [9] propose a method to enable multimodal fusion of incomplete data and get good performance. However, the method is transductive and needs pre-training, indicating difficulty in applying it when a new sample comes. Therefore, in this paper, we propose a new inductive framework to tackle the above limitations in an end-to-end schema.

3 PROBLEM FORMULATION

In this section, we define the input data and the modeling problem in this paper. Besides, the necessary notations used in the paper are listed in Table 1 for ease of understanding.

Definition 1 (Patient multimodal EHR data). In multimodal EHR data, each patient can be represented as a collection of observations from multiple modalities (data sources), e.g., medical images, clinical notes, lab tests, etc. Suppose $M$ is the number of modalities, $N$ is the number of patients (samples), and let $n$ be the subscript referring to a specific patient, the patient multimodal EHR dataset can be denoted as: $X = \{X_n\}_{n=1}^{N} = \{(x_{n1}, x_{n2}, ..., x_{nM})\}_{n=1}^{N}$. The observed data of a patient are represented as: $X_n = (x_{n1}, x_{n2}, ..., x_{nM})$, where $0 < M' < M$. It should be noted that, we used the most relaxed setting, i.e., the modality missingness is irregular across patients. In all the training and test (validation) sets, each modality is potentially missing, but at least one modality is present for each patient.

Problem 1 (Disease diagnosis). Given a patient’s multimodal EHR data $X_n$ with some missing modalities, we formulate the disease diagnosis task as a binary or multi-label classification problem, which is diagnosing the disease $y \in \{0, 1\}^{|C|}$ of the patient, where $|C|$ is the number of the unique modality based on disease categories.

Table 1 shows the notations used in the paper.

4 METHODOLOGY

4.1 Overview

Figure 3 and Figure 4 show the architecture of M$^3$Care. It consists of two main sub-models. The first one is used to compensate for the missing information in the latent space (Corresponding to Figure 3). The other utilizes the processed representations to perform the clinical tasks (Corresponding to Figure 4). Specifically, M$^3$Care includes the following detailed components:

- The Unimodal Representation Extraction module maps the original input features of a patient in each modality to the latent space by encoders with various backbones. The backbones are different due to different input modalities (Left part in Figure 3).

- The Similar Patients Discovery and Information Aggregation module computes patient similarities of each modality with learned task-guided deep kernels. The similarities induce patient graphs. With graph information propagation, the information from similar patients is aggregated (Middle and right parts in Figure 3).

- The Adaptive Modality Imputation module imputes the missing modality in the latent space with the aggregated information, and fuses the existing modality and the auxiliary information to enhance the representation learning (Right part in Figure 3).

- The Multimodal Interaction Capture module takes the intra- and inter-modality dynamics into consideration to perform the final clinical tasks (Figure 4).

4.2 Unimodal Representation Extraction

For a specific patient $n$, it is hard to model the interactions among the raw data, since his/her data $X_n$ is high-dimensional and inconsistent with respect to different data structures in different modalities [9]. Therefore the unimodal representation extraction models are in need to extract the task-relevant feature latent representation in the latent space of each modality. Here, suppose $f_m(·; \Theta_m)$ be the modality $m$’s unimodal representation extraction model with learnable parameter $\Theta_m$. For modality $m$’s raw input, the corresponding latent representation can be obtained via:

$$h_{n}^{m} = f_{m}(x_{n}^{m}; \Theta_{m}) ,$$

where $h_{n}^{m} \in \mathbb{R}^{N_b}$ and $N_b$ is the dimension of the representation for modality $m$. We use the lowercase letter $x_{n}^{m}$ to denote a modality for a single patient $n$.

Three types of $f_m(·; \Theta_m)$ are taken into consideration in this paper: 1) ResNet [20] for embedding image modalities; 2) Transformer Encoder [50] for embedding sequential modalities, such as time-series data like patient laboratory test, medication, and free texts like clinical notes; and 3) multi-layer perceptron (MLP) for embedding vector-based modalities, such as demographic information. As shown in the left part in Figure 3, the black boxes denote the missing information.
4.3 Similar Patients Discovery and Information Aggregation

The above section describes how to deal with a single sample, and now we focus on a group of samples. Given a batch of patients, the representations of modality $m$ for them are denoted as:

$$H^m = [h^m_1, h^m_2, \ldots, h^m_B]^T \in \mathbb{R}^{B \times N_m},$$

where $B$ is the batch size. For sequential modalities, we select the representation of the last timestamp as the one.

For a specific patient with missing modalities, as shown in Figure 1, the representations from the missing modalities cannot be obtained, which results in a lack of modality information. As discussed in Section 1, we can compare similarities of local patient data and impute the missing modalities of the patient by referring to the uncensored modalities of locally similar patients. There is thus a problem with similar patient discovery: What metric(s) should be used to model the similarity relation? In practical applications, one often adopted strategy is to try different kinds of similarity measures on the learned representations in each modality space, such as Cosine, Euclidean distance, and so on, and then select the best similarity measure [28]. However, this approach is time-consuming, and even if one tries different similarity metrics, it is found that those traditional similarity metrics often fail to consider the local environment of data points, and may learn incomplete and inaccurate relationships. In this case, it is unlikely that the traditional similarity function will be adequate to capture the local manifold structure precisely [27]. Moreover, complex relationships such as higher-order statistics are failed to capture in that way [27].

To this end, we extend this idea to kernel spaces and select the RBF kernel. Given two samples $h^m_i$ and $h^m_j$, the similarity calculated from the RBF kernel is defined as:

$$k(h^m_i, h^m_j) = \exp\left(-\frac{\|h^m_i - h^m_j\|^2}{2\sigma^2}\right), (2)$$

where $\sigma$ is the bandwidth to control the extent to which similarity of small distances is emphasized over large distances. Following [12], we set $\sigma$ as a fraction of the mean distance between examples. Expanding the exponential via Taylor series, we can see that the RBF kernel implies an infinite dimension mapping, capturing the higher-order statistics.

Furthermore, as mentioned in Challenge 2, since the data are in multiple modalities, M$^3$Care is required to calculate similarities in each modality space and associate the connotation of similarity with clinical tasks. Thus, a task-guided modality-semantic-adaptive similarity metric is needed. We extend the standard RBF kernel to deep kernel [36] to build an adaptive kernel with a learnable network to fit the representations in a modality data-driven way. Specifically, the kernel is denoted as:

$$k_{\text{sim}}(h^m_i, h^m_j) = \left(1 - \delta_m\right)k(\phi_{\text{sim}}(h^m_i), \phi_{\text{sim}}(h^m_j)) + \delta_m q(h^m_i, h^m_j), (3)$$

where $\phi_{\text{sim}}$ is a network with parameters $\omega_m$ for modality $m$. $k$ and $q$ are different RBF kernels with different $\sigma$. The $\delta_m \in (0, 1)$ is a learnable safeguard to preventing the learned kernel from going extremely far-away from the right direction.

Now, back to the batch of patient representations, the pairwise similarities with respect to each modality are calculated as (i.e., the
4.4 Adaptive Modality Imputation

Now we obtain two different representations for the batch of patients in each modality: $H_m^m$ and $H_m^m$. The former focuses on the patients themselves for modality $m$, while the latter refers to the information aggregated from similar patients. For a specific patient $i$, if the data of modality $m$ are missing, we can directly impute the representation with the auxiliary information aggregated from similar patients. While for a patient whose modality $m$ is complete, such auxiliary information can also be fused into the original representation, making the representation smoother to reduce noise, thus enhancing the representation learning. Here, we use an attention fusion to adaptively extract the proper amount of information from them (i.e., $H_m^m$ and $H_m^m$). Specifically, two weights $\alpha^m, \beta^m \in \mathbb{R}^{B \times 1}$ are introduced to determine the importance of the above two representations, which are obtained by fully connected layers:

$$a^m = \text{Sigmoid}(H_m^m W_0), \quad \beta^m = \text{Sigmoid}(H_m^m W_1),$$  

(9)

where $W_0, W_1 \in \mathbb{R}^{N_h \times 1}$ are the weight matrices. $\alpha$ and $\beta$ indicate the importances of self-information and information of similar patients. We add a constraint $\alpha + \beta = 1$ by calculating $\alpha = \frac{\beta}{\alpha + \beta}$, $\beta = 1 - \alpha$. The final imputed and enhanced representations can be obtained as (i.e., the No.3 arrow in Figure 3):

$$h_i^m = \begin{cases} \hat{h}_i^m & \text{if modality $m$ of sample $i$ is missing} \\ e_i^m \cdot h_i^m + \beta_i^m \cdot \hat{h}_i^m & \text{otherwise} \end{cases}$$  

(10)

where $h_i^m$ and $\hat{h}_i^m$ are the $i$-th sample of $H_m^m$ and $H_m^m$, respectively.

4.5 Multimodal Interaction Capture

Back to a specific patient, so far, the representations of the missing modalities have been imputed through the above sections. These representations are used to perform the clinical tasks. Thus, we need to consider complex correlations among multimodal EHR, including intra-correlations within each modality and inter-correlations between modalities. Inspired by Akbari et al. [2], Kim et al. [30], a context-aware multimodal interaction capture is built. Specifically, for sequential modalities, the internal positional encoding is added: $H_{\text{seq}} = [h_{1seq}, h_{2seq}, ..., h_{Nseq}] + PE_{\text{seq}}$, where $PE_{\text{seq}}$ is the positional encoding for sequential modality $\text{seq}$ and $N_{\text{seq}}$ is the length of the sequence. Next, the representations are added with the corresponding modality type embeddings and concatenated to form the input:

$$z_0^i = [H^1 + TE^1; H^2 + TE^2; ...; H^M + TE^M].$$  

(11)

where $TE^m$ is the corresponding type embedding to identify each modality. And the multimodal interactions are captured through:

$$z_i^l = \text{LayerNorm}(z_i^{l-1} + \text{MHSA}(z_i^{l-1})),$$  

$$z_i^l = \text{LayerNorm}(z_i^{l-1} + \text{FFN}(z_i^{l-1})), $$  

(12)

where $l = 1, ..., L$ refers to the number of such stacked layers. MHSA refers to the Multi-Head Self-Attention [50], FFN refers to a feed-forward network and LayerNorm is the layer normalization [4]. The predictor is built via:

$$\hat{y}_i = \text{Sigmoid}(x_i^L W_{\text{final}}),$$  

(13)
where $W_{\text{final}} \in \mathbb{R}^{N_h \times 1}$ is the weight matrix. The Sigmoid function is used to turn the output into the probability. In this case, the cross-entropy loss is used as the prediction loss function:

$$L_{\text{pre}} = -\frac{1}{B} \sum_{i=1}^{B} (y_i^T \log(\hat{y}_i) + (1 - y_i)^T \log(1 - \hat{y}_i)),$$

where $B$ is the batch size, $\hat{y}_i \in \{0, 1\}^{C_i}$ is the predicted probability, and $y_i \in \{0, 1\}^{C_i}$ is the ground truth. The overall loss function is:

$$L = L_{\text{pre}} + \lambda L_{\text{stab}},$$

where $\lambda$ is the hyperparameter to control the loss. For ease of understanding, we summarize $\text{M}^3\text{Care}$ in Algorithm 1 in Appendix.

5 EXPERIMENT

We evaluate $\text{M}^3\text{Care}$ on the following datasets: Ocular Disease Intelligent Recognition (ODIR) Dataset and Ophthalmic Vitrectomy (OV) Dataset. The model code is provided in $^3$.

5.1 Data Description and Task Formulation

We use the following datasets and tasks to evaluate our model.

- **Ocular Disease Intelligent Recognition (ODIR) Dataset**
  comes from an ophthalmic database, which is meant to represent real-life set of patients collected from hospitals [34]. 3,500 patients are extracted to construct this dataset to diagnose ocular diseases. This dataset contains the following modalities (incomplete modalities exist): demographic information, clinical text for both eyes, and fundus images for both eyes. The detailed statistics are presented in Table 7 in Appendix.

- **Ophthalmic Vitrectomy (OV) Dataset**
  comes from an ophthalmic hospital in clinical practice, after vitrectomy, the intraocular pressure (IOP) may increase abnormally. This symptom cannot be predicted by doctors. We collect 832 patients to predict whether the IOP will increase abnormally. This dataset contains six modalities (incomplete modalities exist): demographic, clinical notes, medications, admission records, discharge records, and surgical consumables. The detailed statistics are presented in Table 7.

- **Task**. The task on the dataset is binary classification tasks. Following existing works [21, 57], we assess performance using the area under the receiver operating characteristic curve (AUPRC), the area under the ROC curve (AUROC), and accuracy (ACC). AUPRC is the most informative and primary evaluation metric, especially while dealing with skewed real-world data [11, 14, 57]. Due to the size of the dataset, we employ 10-fold cross-validation to assure the consistency of the performance and report the average performance with standard deviations.

5.2 Experimental Setup and Baselines

To conduct the experiment, we use the Adam optimization algorithm in Pytorch 1.5.1. More details are in the Appendix. We include these state-of-the-art models as our baseline models:

- **MEN** [56] captures view-specific and cross-view interactions, and summarizes them with a multi-view gated memory module.
- **MVIT** [49] utilizes directional pairwise cross-modal transformers to attend to interactions between multimodal data.
- **ViLT** [30] commissions the transformer module to extract and process all the multimodal features simultaneously.
- **CM-AEs** [43]: The cross-modal autoencoders, which generate missing modalities first and make predictions.
- **SMIL** [40] approximates the missing modality using a weighted sum of manually defined modality priors learned from the dataset.

The following ablation studies are also conducted:

- $\text{M}^3\text{Care}_{1-}$ does not use the task-guided deep kernels of each modality. It directly calculates similarity via cosine similarity.
- $\text{M}^3\text{Care}_{2-}$ does not consist of the Information Aggregation and the Adaptive Modality Imputation module. It directly computes the mean similarity from each modality and approximates the missing-modality representations via the similar patients.

Specifically, on the ODIR Dataset, the number in ( ) denotes the standard deviation of bootstrapping for 1,000 times. The results show that, compared with the best baseline method, $\text{M}^3\text{Care}$ achieves relative improvements of 4.9% in micro-AUC. On the OV Dataset, the number in ( ) denotes the standard deviation of 10-fold cross-validation. We can see that, compared with the best baseline method, $\text{M}^3\text{Care}$ achieves relative improvements of 6.1% in AUPRC and 6.0% in AUROC. Among these baseline methods, some ones

### Table 2: Results on the ODIR Dataset

| Methods | ODIR Dataset (Multi-label Classification) |
|---------|------------------------------------------|
|         | micro-AUC $\uparrow$ | macro-AUC $\uparrow$ | test loss $\downarrow$ |
| MFN [56] | 0.7877 (0.030) | 0.7766 (0.029) | 0.1772 (0.020) |
| ViLT [49] | 0.7944 (0.028) | 0.8032 (0.026) | 0.2399 (0.019) |
| CM-AEs [43] | 0.8028 (0.030) | 0.7624 (0.029) | 0.1731 (0.016) |
| SMIL [40] | 0.8092 (0.032) | 0.7978 (0.025) | 0.2278 (0.032) |
| HGMF [9] | 0.8080 (0.030) | 0.8103 (0.031) | 0.1810 (0.022) |
| $\text{M}^3\text{Care}_{1-}$ | 0.8130 (0.031) | 0.8059 (0.032) | 0.1781 (0.020) |
| $\text{M}^3\text{Care}_{2-}$ | 0.8030 (0.031) | 0.8138 (0.029) | 0.1631 (0.018) |
| $\text{M}^3\text{Care}$ | **0.8490** (0.025) | **0.8245** (0.026) | **0.1543** (0.018) |

$^*$: $p < 0.01$, **: $p < 0.05$
We conduct several further experiments. Due to the limitation of pages, some of the experiments are in the Appendix.

5.4 Further Analysis

We conduct several further experiments. Due to the limitation of pages, we report two cases in the test set here. As shown in Figure 5, the intraocular pressure (IOP) of the two patients increases abnormally and M^Caresuccessfully predicts the outcome. The rows and columns show the Query and Key multimodal records, which are the abbreviations of each modality, i.e., demographic information, clinical notes, medications, admission records, discharge records, and surgical consumables, respectively.

We notice that M^Caregives strong focus on Med0 and Med2 (i.e., the first and third medications) of patient a, and Med0 of patient b. In all these medications, the patients received the two drugs: Tropicamide Phenylephrine Eye Drops (Mydrin-P) and Prednisolone Acetate Ophthalmic Suspension (Pred Forte). These drugs are used for ophthalmologic examinations, prior to ocular surgery [42] or treat eye swelling caused by allergy, infection, injury, or other conditions [48]. Our model discovers that these two drugs may have a strong relationship with the abnormally increased intraocular pressure. This is highly consistent with medical literature [3, 29, 41, 46] and clinician experience, which confirm that the two drugs can lead to adverse reactions like elevation of IOP and should be used with caution for specific patients in clinical practice.

6 CONCLUSIONS

In this paper, we propose M^Care, an end-to-end model to compensate for the missing information of the patients with missing modalities and perform clinical prediction as well as analysis. For a patient with missing modalities, M^Care finds similar patients with a task-guided modality-adaptive similarity metric. Instead of generating raw missing data, M^Care imputes the hidden representations of the missing modalities in the latent space by the auxiliary information from these similar ones, and conducts the clinical tasks. Experiments show that M^Care outperforms all baseline models. Besides, the findings are in accord with experts and medical knowledge, which shows it can provide useful insights.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.62172011). L. Ma is supported by the China Postdoctoral Science Foundation (2021TQ0011). J. Wang is supported by EPSRC New Investigator Award under Grant No.EP/V043544/1.

REFERENCES

[1] Arvind Agarwal, Samuel Gerber, and Hal Daume. 2010. Learning multiple tasks using manifold regularization. Advances in neural information processing systems 23 (2010).
[2] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. arXiv preprint arXiv:2104.11178 (2021).
[3] Eray Atalay, Nivshar Tamçelik, Ceyhun Arci, Ahmet Ozkizil, and Metin Dastan. 2015. The change in intraocular pressure after pupillary dilation in eyes with pseudoxefoliation glaucoma, primary open angle glaucoma, and eyes of normal subjects. International ophthalmology 35, 2 (2015), 215–219.
[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. arXiv preprint arXiv:1607.06450 (2016).
[5] Tian Bai and Slobodan Vucetic. 2019. Improving medical code prediction from clinical text via incorporating online knowledge sources. In The World Wide Web Conference. 72–82.
[6] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. The World Wide Web Conference.
[7] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware LSTM networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 423–443.
[8] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 1158–1166.

Table 3: Results on the OV Dataset

| Methods      | AUPRC ↑ | AUROC ↑ | ACC ↑ |
|--------------|---------|---------|-------|
| MFN [56]    | 0.6456 (0.038) | 0.6789 (0.032) | 0.6627 (0.032) |
| MultiT [49] | 0.6814 (0.047) | 0.6891 (0.045) | 0.6988 (0.031) |
| ViLT [30]   | 0.6987 (0.051) | 0.7249 (0.048) | 0.6627 (0.033) |
| CM-AEs [43] | 0.6891 (0.051) | 0.6927 (0.040) | 0.6747 (0.029) |
| SMIL [40]   | 0.7109 (0.045) | 0.7041 (0.033) | 0.6867 (0.032) |
| HGMF [9]    | 0.7037 (0.050) | 0.7544 (0.027) | 0.7100 (0.032) |
| M^Care1−   | 0.6849 (0.054) | 0.7472 (0.052) | 0.7080 (0.064) |
| M^Care2−   | 0.7110 (0.044) | 0.7562 (0.057) | 0.7234 (0.072) |
| M^Care     | 0.7549 (0.065) | 0.7998 (0.049) | 0.7438 (0.058) |

Figure 5: Attention weight visualization for two patients with abnormally increased intraocular pressure in the test set of the OV dataset. Best viewed in color.

like CM-AEs [43], SMIL [40], HGMF [9] use various mechanisms to handle the missing modalities, and thus they achieve relative higher performance. However, the performance boost demonstrates the effectiveness of M^Care. Besides, it is worth mentioning that the OV Dataset only contains the multimodal EHR data of 832 patients yet has six modalities, demonstrating that M^Care performs well on a small dataset while the number of modalities is large, which is suitable for the real-world scenario, where data has a large number of modalities or missing patterns.

The superior performance of M^Care than the M^Care1−, (i.e., calculating similarity via cosine similarity) verifies the efficacy of the task-guided deep kernels. Moreover, M^Care outperforms M^Care2−, which demonstrates the superiority of the Information Aggregation and the Adaptive Modality Imputation module.

5.4.1 Clinical implications. To intuitively show the implication of M^Care, we visualize the attention weights of the prediction process. Due to the limitation of pages, we report two cases in the test set here. As shown in Figure 5, the intraocular pressure (IOP) of the two patients increases abnormally and M^Care successfully predicts the outcome. The rows and columns show the Query and Key multimodal records, which are the abbreviations of each modality, i.e.,...
[59] Xi Zhang, Jingyuan Chou, and Fei Wang. 2018. Integrative analysis of patient health records and neuroimages via memory-based graph convolutional network. In 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 767–776.

[60] Tao Zhou, Huazhu Fu, Geng Chen, Jianbing Shen, and Ling Shao. 2020. Hi-net: hybrid-fusion network for multi-modal MR image synthesis. *IEEE transactions on medical imaging* 39, 9 (2020), 2772–2781.
A INTUITION DISCOVERY EXPERIMENT

In each dataset, for each modality, we first take the samples containing the complete modality data and divide them into training, validation, and test sets. Then we use a unimodal classifier (e.g., multilayer perceptron, transformer encoder) to classify the data for each modality. We take the best model from the validation set, apply it to the test set, and collect the representations in the latent space of each modality for each sample. Next, we compute pair-wise similarity matrix $\Pi_m$ between samples in each modality, where $m$ is the corresponding modality.

We want to justify the intuition in our datasets: if two patients are similar in one modality, they are more likely to be similar in another modality with regard to the clinical task. We call intuition the cross-modal transfer of sample similarity: Given modalities $a$ and $b$, the intuition holds if their difference of the pair-wise similarity matrix $\|\Pi_a - \Pi_b\|_{\text{norm}}$ is small enough, where norm is a type of matrix norm. In this experiment, we try different similarity metrics such as normalized Euclidean distance, cosine similarity and RBF kernel. And we also try different norms as metrics of difference, such as Frobenius norm, 2-norm and mean value of all entries in $\|\Pi_a - \Pi_b\|_\text{norm}$.

For comparison, we add noise to the representations in the latent space in one modality (we select modality $b$ here), and calculate the difference $\|\Pi_a - \Pi_b\|_{\text{norm}}$. If the intuition holds, the difference should be bigger than the above original $\|\Pi_a - \Pi_b\|_{\text{norm}}$. We perform this experiment 1,000 times and calculate the average difference to avoid chance.

Furthermore, we also shuffle the representations in the latent space in one modality (we select modality $b$ here), and calculate the difference $\|\Pi_a - \Pi_b\|_{\text{norm}}$. If the intuition holds, the difference should also be bigger than the first original $\|\Pi_a - \Pi_b\|_{\text{norm}}$. In the same way, we perform this experiment 1,000 times and calculate the average difference to avoid chance. The results are shown in Table 4 and 5.

As shown in Table 4, the original difference of the pair-wise similarity matrices in the two modalities is smaller than both the Noise and Shuffle ones with regard to different similarity metrics and different norms. This justifies that if two patients are similar in one modality, they are more likely to be similar in another modality in different view. To this end, we come up with our intuition, i.e., the cross-modal transfer of sample similarity. In another pair, as shown in Table 5, the same conclusion can be drew.

Table 4: Intuition observation results for two modalities: admission records and clinical notes, on the OV Dataset

| Metric          | Norm  | Original | Noise | Shuffle |
|-----------------|-------|----------|-------|---------|
| Normalized      | Frobenius | 183.35   | 403.23| 207.27  |
| Euclidean       | 2     | 153.97   | 385.21| 173.73  |
| Distance        | Mean  | 0.2334   | 0.4871| 0.2724  |
| Cosine similarity | Frobenius | 402.96   | 462.39| 452.72  |
|                 | 2     | 336.31   | 365.02| 376.69  |
|                 | Mean  | 0.5081   | 0.5821| 0.5908  |
| RBF kernel      | Frobenius | 176.35   | 261.05| 199.60  |
|                 | 2     | 147.51   | 196.10| 166.46  |
|                 | Mean  | 0.2241   | 0.3024| 0.2623  |

B ALGORITHM

Algorithm 1 shows the algorithm of $\mathcal{M}^3$Care.

Table 5: Intuition observation results for two modalities: medications and surgical consumables information, on the OV Dataset

| Metric          | Norm  | Original | Noise | Shuffle |
|-----------------|-------|----------|-------|---------|
| Normalized      | Frobenius | 152.96   | 159.02| 370.39  |
| Euclidean       | 2     | 131.37   | 145.77| 360.48  |
| Distance        | Mean  | 0.2021   | 0.2289| 0.5588  |
| Cosine similarity | Frobenius | 384.57   | 406.01| 404.89  |
|                 | 2     | 326.61   | 353.72| 367.75  |
|                 | Mean  | 0.4984   | 0.5328| 0.5671  |
| RBF kernel      | Frobenius | 168.85   | 173.56| 203.48  |
|                 | 2     | 144.43   | 157.96| 188.48  |
|                 | Mean  | 0.2081   | 0.2260| 0.2825  |
Table 6: Micro-AUC on ODIR Dataset under different additional synthesizing multimodal missing rates.

| Methods     | 30%       | 40%       | 50%       | 60%       |
|-------------|-----------|-----------|-----------|-----------|
| MFN [56]    | 0.7625 (.03) | 0.7334 (.03) | 0.7260 (.03) | 0.7166 (.03) |
| MulT [49]   | 0.7768 (.02) | 0.7637 (.03) | 0.7480 (.03) | 0.7348 (.03) |
| VILT [30]   | 0.7601 (.03) | 0.7583 (.03) | 0.7492 (.03) | 0.7355 (.02) |
| CM-AEs [43] | 0.7864 (.03) | 0.7787 (.00) | 0.7648 (.03) | 0.7477 (.03) |
| SMIL [40]   | 0.7702 (.02) | 0.7595 (.02) | 0.7485 (.03) | 0.7396 (.03) |
| HGMF [9]    | 0.7831 (.03) | 0.7711 (.03) | 0.7585 (.03) | 0.7427 (.02) |

\textsuperscript{3}Care | 0.8119 \textsuperscript{+} (.02) | 0.7927 \textsuperscript{+} (.03) | 0.7795 \textsuperscript{+} (.03) | 0.7715 \textsuperscript{+} (.02) |

![Figure 6: Attention weight visualization for two patients in the test set of OV dataset.](image)

C FURTHER ANALYSIS

C.1 Multiple levels multimodal incompleteness

To consider more realistic various settings and verify the generalizability of \(^3\text{Care}\), experiments on multimodal data with multiple levels of multimodal incompleteness are conducted. We evaluate the influences of missing modalities by attaching additional synthesizing multimodal incompleteness rates on ODIR Dataset from 30% to 60% with an intermittent 10%. The experiments are repeated test with bootstrapping for 1,000 times, and the results (micro-AUC) are in Table 6.

We can see that all the models’ micro-AUCs decrease as the missing rate increases, and \(^3\text{Care}\) still outperforms all baselines. When the missing rate is the biggest of all settings (60%), \(^3\text{Care}\) also demonstrated significantly better performance than the best baselines CM-AEs [43] and HGMF [9]. Specifically, \(^3\text{Care}\) achieves a micro-AUC of 0.7715, while the baseline models CM-AEs [43] and HGMF [9] achieve 0.7477 and 0.7427, showing 3.2% and 3.8% relative improvement, respectively.

C.2 Clinical implications

To intuitively show the implication of \(^3\text{Care}\), similar to Section 5.4.1, we further visualize the attention weights of the prediction process for another two cases. As shown in Figure 6, the first case has a positive label, and the second one is negative. For the first one, \(^3\text{Care}\) gives a strong focus on Med0 and Med1. In both these medications, the patients received not only the drug mentioned above in Section 5.4.1: Tropicamide Phenylephrine Eye Drops (Mydrin-P), which have been proved by medical literature [3, 29] and clinician experience that they can lead to adverse reactions like elevation of IOP. The patients also received Tobramycin Dexamethasone Eye Drops (Tobradex), which is highly consistent with medical literature [10] and clinician. The drugs can lead to adverse reactions and have the tendency to increase intraocular pressure [10]. For the second one, a healthy patient with a negative label, \(^3\text{Care}\) gives relatively even attention to each data of each modality of the patient, which indicates that \(^3\text{Care}\) does not discover significant signs of elevation of IOP and finally makes a right prediction.

D DETAILS OF EXPERIMENTAL SETTINGS

D.1 Statistics of the Datasets

- **Ocular Disease Intelligent Recognition (ODIR) Dataset**: contains the following modalities (incomplete modalities exist): demographic information, clinical text for both eyes, and fundus images for both eyes. The detailed statistics are presented in Table 7.

- **Ophthalmic Vitrectomy (OV) Dataset**: contains six modalities (incomplete modalities exist): demographic, clinical notes, medications, admission records, discharge records, and surgical consumables. The detailed statistics are presented in Table 7.

| Dataset             | Statistic       | Value               |
|---------------------|-----------------|---------------------|
| Ocular Disease Intelligent Recognition (ODIR) Dataset | # patients | 3,500 |
|                     | # modalities    | 3                   |
|                     | % missing per modality | [0\%, 48.34\%, 0\%] |
|                     | % positive labels | [0.061, 0.060, 0.0046] |
|                     | % female        | .461                |
| Ophthalmic Vitrectomy (OV) Dataset | # patients | 832 |
|                     | # modalities    | 6                   |
|                     | % missing per modality | [0\%, 0.60\%, 15.38\%, 10.33\%, 10.33\%, 4.08\%] |
|                     | % positive labels | 41.7\%             |
|                     | % female        | .456                |

D.2 Model Implementation

The experiment environment is a machine equipped with CPU: Intel Xeon E5-2630, 256GB RAM, and GPU: Nvidia RTX8000. The code is implemented based on Pytorch 1.5.1. The hyper-parameter setting of the proposed \(^3\text{Care}\) is as follows: We set the embedding dimension and hidden dimension as 128/256 for Ocular Disease Intelligent Recognition (ODIR) / Ophthalmic Vitrectomy (OV) dataset, respectively. Since the clinical notes modality in OV dataset is too long per patient, and the number of samples is small (832 patients), we set the batch size as 32 while conducting experiments on OV dataset. For ODIR dataset, we set the batch size as 512.

\textsuperscript{3}The order of the missing rates are corresponding to the above data description.