Supplementary Material for MIRACL: A Robust Multi-Label Learning Framework on Noisy Multimodal Electronic Health Records

Anonymous submission

Appendix A: Experimental Setup & Implementation Details

Prediction Tasks

- Phenotyping(PHE): A multi-label classification problem that classifies which of 25 acute care conditions are present in a given patient ICU stay record.
- Diagnosis(DIA): A multi-label classification problem that predicts 14 diagnosis conditions.

Dataset and Preprocessing

Due to the limited number of multimodal multi-label learning datasets, we choose 3 EHR-based datasets to further validate our approach.

- MIMIC-IV (PHE, DIA) (Johnson et al. 2023b,a) Phenotyping, Diagnosis: We use the same preprocessing procedure as (Xu et al. 2024) for MIMIC-IV.
- MIMIC-III (PHE) (Johnson et al. 2016): We use the same preprocessing procedure as (Harutyunyan et al. 2019) for EHR. For clinical notes, we adapt the (Khadanga et al. 2019) to extract clinical notes and use a maximum length of 512 for each clinical note.
- MIMIC-IV (Johnson et al. 2023b,a) Phenotyping, Diagnosis: We use the same preprocessing procedure as (Xu et al. 2024) for MIMIC-IV.
- MIMIC-III (Johnson et al. 2016) Phenotyping: We use the same preprocessing procedure as (Harutyunyan et al. 2019) for EHR. For clinical notes, we adapt the (Khadanga et al. 2019) to extract clinical notes.

Table 1: Statistics of the Multimodal Multi-Label dataset.

Task	Prediction Task	Modality	# Number	C	L_{avg}
MIMIC-IV PHE	Clinical Phenotype	$\{TS, T\}$	59,798	25	4.575
MIMIC-IV DIA	Clinical Diagnosis	$\{TS, T\}$	132,576	14	2.246
MIMIC-III PHE	Clinical Phenotype	$\{TS, T\}$	41,904	25	4.126

Baseline Description

Baselines: We use FlexCare (Xu et al. 2024) as the backbone model for noisy multi-label approaches and compare our approach to several baseline models with same hyperparameter setting:

- **Focal Loss (Focal)** (Lin et al. 2017): Addresses class imbalance by focusing more on hard-to-classify examples.
- **Asymmetric Focal Loss (ASL)** (Ridnik et al. 2021): Modifies Focal Loss to better handle label imbalance in multi-label settings by assigning different weights to relevant and irrelevant labels.
- **Generalised Cross-Entropy (GCE)** (Zhang and Sabuncu 2018): A robust loss function designed for noisy multi-label classification, combining properties of mean absolute error (MAE) and cross-entropy (CE) for better noise tolerance.
- **MLLSC** (Ghiassi, Birke, and Chen 2023): Handles missing and corrupted labels by leveraging loss values for both true-positive and false-positive labels to improve model robustness.
- **MultiT** (Li et al. 2022): Utilises label correlations to estimate a transition matrix for noisy multi-label learning, effectively aligning observed labels with true labels to mitigate label noise.

We also compare our approach against existing multimodal healthcare model:

- **M3Care** (Zhang et al. 2022): Proposes an end-to-end multimodal framework that addresses missing modalities in healthcare data by imputing missing modalities information from similar patients.
- **MedFuse** (Hayat, Geras, and Shamout 2022): A lightweight and flexible multimodal model that projects each modality (e.g., time-series EHR, medical images) into a shared latent space using modality-specific encoders, by a LSTM fusion-based module
- **FlexCare** (Xu et al. 2024): A flexible multimodal multitask framework that decomposes parallel task prediction into asynchronous single-task outputs, uses task-agnostic representation learning with covariance regularization across modalities, and integrates these via a task-guided hierarchical fusion module to support multimodal multi-label EHR prediction.

Overall Training Procedure

Warm-Up Phase: As demonstrated by (Arazo et al. 2019), cross-entropy loss distribution naturally fits a mixture model with theoretical justification. Similarly, in the multi-label setting, the binary cross-entropy (BCE) loss $\mathcal{L}_{\text{bce}} = -\Big(\tilde{Y} \cdot \log(\hat{Y}) + (1 - \tilde{Y}) \cdot \log(1 - \hat{Y})\Big)$. We treat $\mathcal{L}_{\text{cons}}$ as a regularisation term and incorporate it into the final objective:

$$\mathcal{L}_{\text{warmup}} = \mathcal{L}_{\text{bce}} + \lambda_{\text{cons}} \cdot \mathcal{L}_{\text{cons}}, \tag{1}$$

where λ_{cons} is a weighting coefficient controlling the strength of the contrastive regularization.

Correction Phase : We fit the computed selection scores to Gaussian Mixture Models from the last epoch over the entire dataset. We then perform Label correction based on the sample selection mechanism derived from the GMMs. We train using the corrected BCE loss $\mathcal{L}_{\text{corr}} = \mathcal{L}_{\text{bce}}(\tilde{Y}_{\text{corr}}, \hat{Y})$ for the remainder of the training period: At the beginning of training, the model relies more on the corrected loss to mitigate the influence of label noise. As training progresses and the model learns more robust representations, the weights gradually shifts towards the standard BCE loss, balancing the contributions of both components dynamically.

Mathematically, the weighted loss $\mathcal{L}_{\text{weighted}}$ is defined as:

$$\mathcal{L}_{\text{weighted}} = \beta_t \, \mathcal{L}_{\text{corr}} + (1 - \beta_t) \, \mathcal{L}_{\text{bce}} + \lambda_{\text{cons}} \cdot \mathcal{L}_{\text{cons}}, \tag{2}$$

where T represents max epoch, β_t increases linearly with epoch t that transitions smoothly from an initial value $\beta_0 = 1$ to a final value $\beta_f = 0.5$ for stabilising the final stages of training in label correction framework (Arazo et al. 2019). The detailed algorithm is shown below in Algorithm 1.

$$\mathbb{E}[Y^{lc} \mid c] = U^{(lc)} \cdot \hat{Y}^{lc} + (1 - U^{(lc)}) \cdot \tilde{Y}^{lc}. \tag{3}$$

$$\tilde{Y}^{l0}_{\text{corr}} = \begin{cases} \tilde{Y}^{l0}, & \tilde{Z}^{l0} \in S_{\text{clean}} \wedge (X, \tilde{Y}^{l0}) \in Z^- \\ \mathbb{E}[Y^{lc} \mid c = 0], & \tilde{Z}^{l0} \in S_{\text{unce}} \wedge (X, Y^{l0}) \in Z^- \\ \hat{Y}^{l0}, & \tilde{Z}^{l0} \in S_{\text{noisy}} \wedge (X, Y^{l0}) \in Z^- \end{cases}, \tag{4}$$

where $c = 0$ indicates the observed negative class; $\hat{Y}^{l0}$ represents model prediction for label l and observed class 0.

$$\tilde{Y}^{l1}_{\text{corr}} = \begin{cases} \tilde{Y}^{l1}, & \tilde{Z}^{l1} \in S_{\text{clean}} \\ \mathbb{E}[Y^{lc} \mid c = 1], & \tilde{Z}^{l1} \in S_{\text{unce}} \\ \hat{Y}^{l1}, & \tilde{Z}^{l1} \in S_{\text{noisy}} \end{cases}, \tag{5}$$

where $c = 1$ indicates the observed positive class; $\hat{Y}^{l1}$ represents model prediction for label l and observed class 1; $\mathbb{E}[Y^{lc} \mid c = 1]$ denotes the expected soft label refined by the uncertainty-aware correction strategy (see Equation 3).

Model Implementation

All experiments are performed on the High-Performance Computing infrastructure using PyTorch 1.11.0 and an NVIDIA A100 GPU. To maintain fairness in comparisons, we apply consistent hyperparameter settings and neural network architecture across all experiments. Early stopping is not used, as we assume the unavailability of a clean validation set, reflecting real-world conditions.

Algorithm 1: MIRACL: Multi-modal Instance Relabelling And Correction for multi-Label noise

```
Input: Multi-label Dataset $\mathcal{D}$, learning rate $\eta$, max epochs $T$, warmup time $t_{warmup}$
Output: Trained Model $M$
Initialise model $M$
for epoch $t = 1$ to $T$ do
    if $t \leq t_{warmup}$ then
        Update $\mathcal{L}_{\text{warmup}}$
    else
        Calibrate $\tilde{Y}$ using Equation (4), (5)
        Update $\mathcal{L}_{\text{weighted}}$ with $\tilde{Y}_{\text{corr}}$ by Equation (2)
    end if
    if $t \geq t_{warmup}$ then
        for $l = 1$ to $C$ do
            for $c = 0$ to $1$ do
                Fit GMM$^{lc}$, select $\mathcal{S}_{\text{clean}}$, $\mathcal{S}_{\text{unce}}$, $\mathcal{S}_{\text{noisy}}$
            end for
        end for
    end if
end for
return Trained Model $M$
```

Implementation Detail Each baseline model is trained independently with the same hyperparameter settings. Each model is trained for 30 epochs using the Adam optimizer, with an initial learning rate of 10^{-3} scheduled via cosine annealing (T$_{\text{max}}$ = 10, η_{min} = 0), batch size of 128, a warm-up period of 5, a correlation threshold of $\tau = 0.02$, a regularisation strength coefficient $\lambda_{cons} = 0.1$, and a selection metric coefficient of $\alpha = 0.5$. For each noise type, experiments are repeated three times with three different random seeds = $\{30, 40, 100\}$. To prevent overfitting to corrected labels, we apply a weight decay parameter of $1e^{-5}$ when initiating label correction. The noise ratios are defined as follows: $\rho_-, \rho_+ \in \{0.2, 0.4\}$ for Symmetric; Asymmetric Flip Noise $\rho_-, \rho_+ \in \{0, 0.2, 0.4\}$; Balanced Noise $\rho_+ = \rho \in \{0.2, 0.4\}, \rho_- = \{0.0448, 0.0896\}$ respectively for MIMIC-IV phenotyping, $\rho_- = \{0.0395, 0.0790\}$ for MIMIC-III phenotyping and $\rho_- = \{0.0382, 0.0764\}$ for diagnosis.

Baseline Setting We use FlexCare (Xu et al. 2024) as the backbone model for noisy multi-label approaches and compare our approach to several baseline models with same hyperparameter setting:

- **FlexCare** (Xu et al. 2024): layers=4, expert_k=2, expert_total=10, hidden_dim = 128, ehr_dim = 76, max-length = 512
- **Focal Loss (Focal)** (Lin et al. 2017): Focusing parameter $\gamma = 2.0$, Alpha-balancing weight $\alpha = 0.25$
- **Asymmetric Focal Loss (ASL)** (Ridnik et al. 2021): Negative focusing parameter $\gamma_- = 4.0$, Positive focusing parameter $\gamma_+ = 1.0$, Probability margin (clipping) $m = 0.05$,
- **Generalised Cross-Entropy (GCE)** (Zhang and Sabuncu 2018): Default parameters, designed to be

Table 2: Comparison of average Last mAP with standard deviation (in bracket) across 3 runs for all models under different noise conditions on the MIMIC-IV diagnosis test dataset. The best average results are highlighted in **bold**.

(ρ_+, ρ_-)	**Symmetric Flip Noise (%)**		**Asymmetric Flip Noise (%)**						**Balanced Noise (%)**	
	(20,20)	**(40,40)**	**(0,20)**	**(0,40)**	**(20,0)**	**(20,40)**	**(40,0)**	**(40,20)**	**(20,3.82)**	**(40,7.64)**
ASL	0.207(0.006)	0.197(0.004)	0.221(0.003)	0.211(0.004)	0.198(0.010)	0.198(0.004)	0.197(0.013)	0.198(0.005)	0.202(0.010)	0.196(0.005)
Focal	0.18(0.000)	0.167(0.011)	0.18(0.000)	0.174(0.011)	0.173(0.012)	0.18(0.000)	0.167(0.011)	0.174(0.011)	0.167(0.011)	0.18(0.000)
GCE	0.193(0.006)	0.193(0.003)	0.208(0.004)	0.195(0.007)	0.193(0.004)	0.194(0.004)	0.189(0.003)	0.192(0.005)	0.19(0.001)	0.193(0.002)
MLLSC	0.157(0.007)	0.154(0.002)	0.153(0.002)	0.157(0.006)	0.159(0.006)	0.157(0.006)	0.159(0.006)	0.157(0.008)	0.157(0.006)	0.159(0.006)
MultiT	0.2(0.004)	0.193(0.003)	0.218(0.006)	0.195(0.002)	0.23(0.001)	0.197(0.006)	0.22(0.016)	0.198(0.007)	0.214(0.021)	0.196(0.009)
M3Care	**0.219(0.000)**	0.206(0.000)	0.222(0.000)	**0.22(0.000)**	0.224(0.000)	**0.215(0.000)**	0.223(0.000)	**0.215(0.000)**	0.224(0.000)	**0.22(0.000)**
MedFuse	0.208(0.001)	0.195(0.001)	0.214(0.002)	0.212(0.001)	0.218(0.004)	0.202(0.001)	0.217(0.001)	0.203(0.001)	0.216(0.003)	0.209(0.001)
FlexCare	0.194(0.007)	0.194(0.004)	0.214(0.009)	0.212(0.010)	**0.231(0.001)**	0.202(0.002)	0.219(0.017)	0.193(0.002)	0.209(0.013)	0.198(0.008)
MIRACL	**0.219(0.000)**	**0.207(0.002)**	**0.223(0.001)**	**0.22(0.001)**	0.228(0.001)	0.214(0.001)	**0.225(0.001)**	0.214(0.001)	**0.225(0.002)**	0.219(0.001)

robust to noise.

- **MLLSC** (Ghiassi, Birke, and Chen 2023): Positive threshold $\tau_{pos} = 0.55$, Negative threshold $\tau_{neg} = 0.6$, Margin $m = 1.0$, Gamma $\gamma = 2.0$
- **MultiT** (Li et al. 2022): Default parameters, designed to perform loss correction based on estimated transition matrix $\hat{T}$.
- **M3Care** (Zhang et al. 2022): hidden_dim = 128, ehr_dim = 76, dropout = 0.1
- **MedFuse** (Hayat, Geras, and Shamout 2022): hidden_dim = 128, ehr_dim = 76, dropout = 0.1

Hyperparameter Tuning

Computational Analysis

Model Name	**Computation Time (h)**
MedFuse	3.108
MultiT	3.979
M3Care	3.118
FlexCare	3.884
MIRACL (Ours)	4.104

Table 3: Computation time comparison for different models under Sym (20,20) condition.

Despite incorporating $L \times 2$ Gaussian Mixture Model (GMM) for dynamic sample selection, MIRACL does not introduce significant computational since it does not rely on re-training strategy against corrected labels. As shown in Table **??**, its total training time remains comparable to other advanced baselines. This efficiency stems from our design choice to fit the GMM once ***per epoch rather than per batch***, and only after the warm-up phase, which amortizes the cost and avoids redundant computation. This demonstrates that MIRACL achieves robust noise correction with gradual increase in training time. **Notably, most of the time complexity stems from the underlying FlexCare backbone shared by MIRACL, rather than the noise modeling module itself.**

Code Availability

We have released part of our code inside: https://github.com/anon-coder-def/CSCNMM. Upon acceptance, the complete code will be open-sourced and made publicly available on GitHub. More importantly, we will release the complete code for the data pre-processing part to ensure reproducibility.

Reproduce Checklist

Reproduce Checklist is attached at the end of main paper.

Appendix B: Additional Analysis and Visualizations

Selection Metric Analysis

Figure 1 presents a comparative analysis of BCE loss, ranking, and memorization difficulty across training epochs for both clean and noisy instance-label pairs under symmetric (40%, 40%) label noise. We observe that memorization-based metrics (Figure 1c) exhibit the greatest discriminative power during the early training phase (e.g., epochs 0–40), where the clean and noisy curves for both positive and negative pairs are clearly separable. This aligns with the memorization effect observed in noisy label learning literature, where deep networks tend to memorize clean samples earlier.

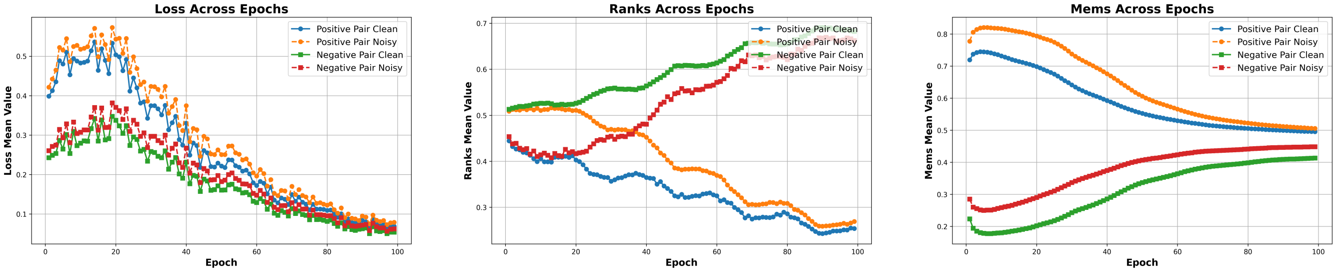
In contrast, ranking-based indicators (Figure 1b) become more stable and reliable in the later stages of training (after epoch 50), where clean positive labels consistently rank higher than noisy ones, and the gap between clean and noisy curves remains steady. This suggests that rank-based selection is more robust after the model has sufficiently learned high-confidence predictions.

Together, these trends validate our design choice: we prioritize memorization difficulty during early epochs to identify clean pairs based on learning dynamics, and shift toward rank-based metrics in later epochs when model predictions become more reliable.

Appendix C: Full Quantitative Results & Checklist

MIMIC-IV Diagnosis Experiment:

While MIRACL demonstrates state-of-the-art performance on the phenotyping task, our results from Table 2 show that it does not consistently outperform M3Care in the diagnosis setting. This discrepancy is attributable to the distinct architectural priorities of the two models in the face of extreme modality missingness. The diagnosis dataset suffers from severe data sparsity, with 76.3% of time-series and 32.6%



(a) Loss Across Epochs

(b) Ranks Across Epochs

(c) Memorization Difficulty Across Epochs

Figure 1: Comparison of metrics such as BCE Loss, Ranks, and Memorization/Forgetting Difficulty across 100 training epochs using the Vanilla FlexCare model under 5% of MIMIC-Phenotyping Dataset with Symmetric (20%, 20%) flip noise.

Table 4: Statistical comparison of MIRACL vs. second-best baseline under each noise setting (based on test mAP over 3 runs) on MIMIC-IV phenotyping.

Noise Type (ρ_+, ρ_-)	Second-best	p-value	Significance
(20,20)	GCE	0.00082	***
(40,40)	GCE	0.00288	**
(0,20)	MultiT	0.00807	**
(0,40)	FlexCare	0.01233	*
(20,0)	MultiT	0.05908	
(20,40)	MultiT	0.00084	***
(40,0)	FlexCare	0.03163	*
(40,20)	FlexCare	0.01548	*
(20,4.48)	FlexCare	0.00002	***
(40,8.96)	MultiT	0.00367	**

of clinical notes absent. M3Care is explicitly designed to handle this challenge through robust modality-specific pathways and dropout mechanisms. In contrast, MIRACL's core strength lies in leveraging cross-modal signals for label noise correction. When one or both modalities are frequently absent, MIRACL's ability to cross-reference evidence is fundamentally limited, reducing its advantage. Nevertheless, MIRACL consistently ranks as the second-best model across most noise configurations, indicating strong generalization despite missing data. This analysis underscores that robustness to label noise and robustness to missing modalities are distinct challenges, and MIRACL is highly specialized for the former. Future work could explore hybrid architectures that combine MIRACL's sophisticated label correction with M3Care's proven robustness to missing data.

Statistical Testing

We perform one-sided Student's t-tests (across 3 runs) comparing MIRACL to the second-best baseline under each noise condition on MIMIC-III (Table 4) Phenotyping and MIMIC-IV phenotyping (Table 5). Significance is marked in Table using *, **, and ***, indicating $p<0.05$, $p<0.01$, and $p<0.001$, respectively. All tests compare test mAP scores under the same seeds. For noise configurations where MIRACL is not the best-performing model (e.g., (40,40)), no significance test is performed.

Table 5: Statistical comparison of MIRACL vs. second-best baseline under each noise setting (based on test mAP over 3 runs) on MIMIC-III phenotyping.

Noise Type (ρ_+, ρ_-)	Second-best	p-value	Significance
(20,20)	ASL	0.00610	**
(40,40)	ASL	0.98795	
(0,20)	ASL	0.00057	***
(0,40)	ASL	0.00068	***
(20,0)	MultiT	0.00328	**
(20,40)	ASL	0.96122	
(40,0)	FlexCare	0.01053	*
(40,20)	ASL	0.90015	
(20,3.95)	MultiT	0.00771	**
(40,7.90)	MultiT	0.01376	*

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2019. Unsupervised Label Noise Modeling and Loss Correction. In Chaudhuri, K.; and Salakhutdinov, R., eds., *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 312–321. PMLR.

Ghiassi, A.; Birke, R.; and Chen, L. 2023. Multi Label Loss Correction against Missing and Corrupted Labels. In Khan, E.; and Gonen, M., eds., *Proceedings of The 14th Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, 359–374. PMLR.

Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; Ver Steeg, G.; and Galstyan, A. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1): 96.

Hayat, N.; Geras, K. J.; and Shamout, F. E. 2022. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In Lipton, Z.; Ranganath, R.; Sendak, M.; Sjoding, M.; and Yeung, S., eds., *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, 479–503. PMLR.

Johnson, A.; Pollard, T.; Horng, S.; Celi, L. A.; and Mark, R. 2023a. MIMIC-IV-Note: Deidentified free-text clinical notes. *MIMIC-IV*.

Johnson, A. E.; Jethani, N.; Shen, L.; Phillips, P.; Lu, Z.; Pollard, T. J.; Moody, B.; Feng, M.; Celi, L. A.; and Mark, R. G. 2023b. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1–8.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3: 160035.

Khadanga, S.; Aggarwal, K.; Joty, S.; and Srivastava, J. 2019. Using Clinical Notes with Time Series Data for ICU Management. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6432–6437. Hong Kong, China: Association for Computational Linguistics.

Li, S.; Xia, X.; Zhang, H.; Zhan, Y.; Ge, S.; and Liu, T. 2022. Estimating Noise Transition Matrix with Label Correlations for Noisy Multi-Label Learning. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24184–24198. Curran Associates, Inc.

Lin, T.-Y.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.

Ridnik, T.; Baruch, E. B.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric Loss For Multi-Label Classification. In *ICCV*, 82–91. IEEE. ISBN 978-1-6654-2812-5.

Xu, M.; Zhu, Z.; Li, Y.; Zheng, S.; Zhao, Y.; He, K.; and Zhao, Y. 2024. FlexCare: Leveraging Cross-Task Synergy for Flexible Multimodal Healthcare Prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3610–3620.

Zhang, C.; Chu, X.; Ma, L.; Zhu, Y.; Wang, Y.; Wang, J.; and Zhao, J. 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, 2418–2428. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393850.

Zhang, Z.; and Sabuncu, M. R. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems*, 31: 8778–8788.