Tables Guide Vision: Learning to See the Heart through Tabular Data Supplementary Material

Marta Hasny[1,2] Maxime Di Folco[1,2] Keno Bressem[4] Julia A. Schnabel[1,2,3]
[1] School of Computation, Information and Technology, Technical University of Munich, Germany
[2] Institute of Machine Learning in Biomedical Imaging, Helmholtz Munich, Germany
[3] School of Biomedical Engineering and Imaging Sciences, King's College London, UK
[4] TUM University Hospital, Technical University of Munich, Germany

1. Detailed Data Description

1.1. Tabular Attributes

Table 1 presents a comprehensive list of tabular attributes from the UK Biobank that were used for tabular similarity calculation during pretraining. These attributes were consistently used across all baseline methods that incorporated tabular data during pretraining. Attributes marked as *extracted* were derived from automated cardiovascular MRI analysis, as described in [1]. Additionally, Table 2 lists attributes that were used for evaluation but were not included in the pretraining process, along with their corresponding field IDs.

1.2. Multilabel CAD Classification Details

One of our evaluation tasks is multilabel CAD classification, encompassing four cardiovascular disorders within ICD-10 categories I20–I25. Each disease is defined by the following codes:

- Angina pectoris: I200, I201, I208, I209;
- Myocardial infarction: I210, I211, I212, I213, I214, I219;
- Other acute ischaemic heart diseases: I240, I248, I249;
- Chronic ischaemic heart disease: I250, I251, I252, I253, I254, I255, I256, I258, I259.

For CAD classification training and evaluation, we consider a subject positive only if they were diagnosed with CAD prior to the date of their cardiac MRI. Due to the nature of cardiovascular disease, it is possible that some subjects labeled as healthy may have undiagnosed CAD. Addressing this limitation in the UK Biobank is left for future work.

Table 1. A list of tabular attributes and their UK Biobank field IDs that were used for the tabular similarity calculation at pretraining.

Tabular Feature	UK Biobank Field ID
Sex	31
Smoking status	20116
Date of birth	33
Date I20 first reported (angina pectoris)	131296
Date of myocardial infarction	42000
Date I24 first reported (other acute ischaemic heart diseases)	131304
Date I25 first reported (chronic ischaemic heart disease)	131306
LVEF	Extracted
LVEDV	Extracted
LVESV	Extracted
LVSV	Extracted
LVEDM	Extracted
LVCO	Extracted
RVEDV	Extracted
RVESV	Extracted
RVSV	Extracted
RVEF	Extracted
RVCO	Extracted
MYOESV	Extracted
MYOEDV	Extracted

Table 2. Tabular feature names and UK field IDs fo attributes that were not used for the pretraining, but included in the evaluations.

Tabular Feature	UK Biobank Field ID
Height	12144
Weights	21002
Body mass index (BMI)	21001

2. Implementation Details

2.1. Baselines

We compare TGV against a mean-guess baseline (used only for cardiac phenotype prediction), a supervised 3D ResNet-50 model [8], four image-based contrastive learning approaches, and one image-tabular contrastive learning method. This section details the implementation of each baseline.

Mean-guess. The mean-guess baseline is applied to nu-

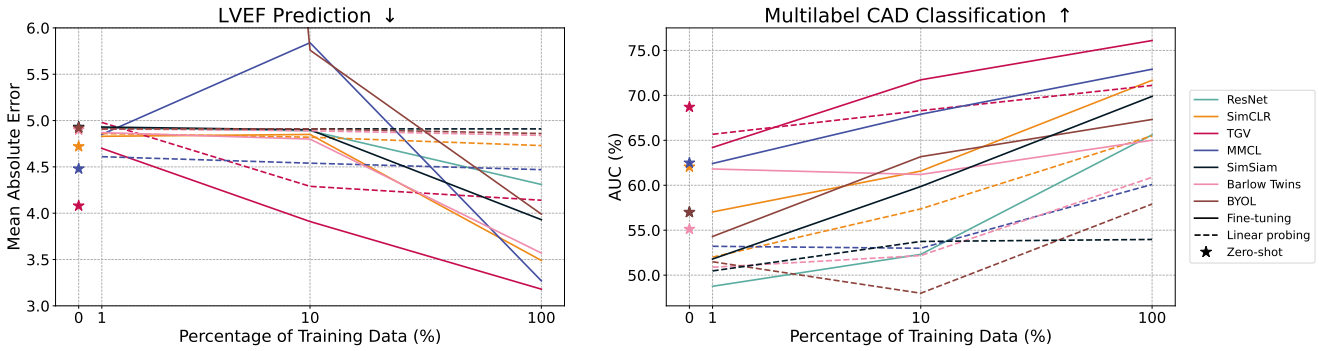


Figure 1. Performance of all evaluated methods on LVEF prediction and CAD classification using zero-shot prediction, and linear probing and fine-tuning under limited data regimes. The LVEF prediction result of BYOL [6] with fine-tuning at 1% is clipped; the value is 26.68. The reported results are obtained using strategies: fine-tuning (solid lines), linear probing (dashed lines), and zero-shot prediction (stars), applied consistently across all included methods.

Table 3. The results of zero-shot predictions on TGV under varying size of representative set P. The size is indicated by N, with the performance of changes relative to $N = 2000$ shown in parenthesis. Red text indicated a decrease of performance.

N	CAD ↑	LVEF ↓	LVEDM ↓	LVEDV ↓	RVEF ↓	RVEDV ↓	MYOESV ↓
2000	68.70±0.83	4.08±0.00	7.64±0.04	13.63±0.07	3.98±0.01	12.43±0.05	8.18±0.06
1000	67.83±1.43 (↓-1%)	4.11±0.00 (↑1%)	7.79±0.07 (↑2%)	13.84±0.11 (↑2%)	4.01±0.01 (↑1%)	12.63±0.03 (↑2%)	8.35±0.10 (↑2%)
500	67.58±1.22 (↓-2%)	4.15±0.02 (↑2%)	8.09±0.13 (↑6%)	14.13±0.31 (↑4%)	4.05±0.03 (↑2%)	12.99±0.12 (↑5%)	8.62±0.13 (↑5%)
100	66.70±1.17 (↓-3%)	4.32±0.04 (↑6%)	8.74±0.45 (↑14%)	15.34±0.61 (↑13%)	4.30±0.06 (↑8%)	14.34±0.55 (↑15%)	9.32±0.43 (↑14%)

merical cardiac phenotype values. The mean value for each phenotype is computed from the training set and used as the prediction on the test set. We report the mean absolute error (MAE) on the test set, calculated as the average difference between the ground truth and the calculated mean.

Supervised. We include one supervised baseline: 3D ResNet-50 [8]. We use the implementation available in the MONAI framework [2], training the model separately for each task using the same hyperparameters as described in Sec. 2.2.

Contrastive. All contrastive learning baselines are built on the 3D ResNet50 architecture [2, 8]. SimCLR [3] is implemented using the NT-Xent loss provided by the lightly framework [12]. The remaining contrastive methods, including image augmentation-based approaches such as BYOL [6], Barlow Twins [13], and SimSiam [4], as well as the image-tabular method MMCL [7], are based on a publicly available repository[1] implemented by Hager et al. [7]. We adapt this codebase to support 3D ResNet-50 and modify it to match our 10-epoch training schedule, ensuring fair comparison. All contrastive baselines use the same image augmentation pipeline, designed to be compatible with 3D+t medical imaging:

- RandomHorizontalFlip(probability=0.5),
- RandomResizedCrop(size=128, scale=(0.6, 1.0)),
- RandomRotation(degrees=45).

All augmentations are implemented using the torchvision library and were selected for their suitability to medical data and volumetric inputs.

2.2. Fine-tuning & Linear Probing

We use the same training hyperparameters for both fine-tuning and linear probing. All downstream tasks are trained for 35 epochs. Cardiac phenotype prediction and CAD classification use a learning rate of 1e-3 and 1e-4 respectively. Cardiac phenotype prediction is trained without augmentations under the full data regime, and with an augmentation rate of 0.6 in the 10% and 1% data settings. CAD classification uses a fixed augmentation rate of 0.4 across all data regimes. We employ L1 loss for cardiac phenotype prediction and binary cross-entropy with logits for CAD classification.

2.3. Zero-shot Prediction

Our zero-shot prediction approach is based on using a representative sample of images to compare the query image against. These representative subsamples are drawn from the disease-balanced training set used for fine-tuning the CAD classification model. The size of the disease balanced data (6,424 samples) allows to construct three non-overlapping subsamples, each containing 2,000 images.

[1] https://github.com/paulhager/MMCL-Tabular-Imaging

Table 4. Mean and standard deviation of the zero-shot prediction over three non-overlapping representative sets P. The best results is **bold**, while the second best is underlined.

Model	CAD ↑	LVEF ↓	LVEDM ↓	LVEDV ↓	RVEF ↓	RVEDV ↓	MYOESV ↓	BMI ↓	Height ↓	Weight ↓	Sex ↑
Image Augmentation											
SimCLR [3]	62.05±1.36	4.72±0.00	13.01±0.07	23.26±0.11	4.71±0.02	21.89±0.16	13.48±0.41	2.54±0.03	6.24±0.03	8.43±0.08	75.04±0.66
BYOL [6]	56.99±0.79	4.92±0.01	16.43±0.01	27.81±0.19	4.95±0.03	26.29±0.19	16.19±0.37	3.33±0.44	7.02±0.01	10.44±0.13	69.72±0.36
SimSiam [5]	57.01±1.58	4.93±0.01	16.04±0.12	27.46±0.09	4.94±0.04	25.62±0.22	16.14±0.13	2.86±0.02	6.93±0.03	8.80±1.30	68.19±0.31
Barlow Twins [13]	55.12±1.59	4.90±0.04	16.54±0.21	27.17±0.04	4.97±0.08	25.98±0.14	16.52±0.17	3.16±0.03	7.02±0.06	10.88±0.09	72.90±0.35
Tabular Supervision											
MMCL [7]	62.49±0.50	4.48±0.06	8.74±0.11	15.12±0.07	4.55±0.05	13.70±0.12	9.54±0.13	2.96±0.07	5.13±0.03	8.68±0.24	92.60±0.15
Tabular Guidance											
TGV (Ours)	**68.70±0.83**	**4.08±0.00**	**7.64±0.04**	**13.63±0.07**	**3.98±0.01**	**12.43±0.05**	**8.18±0.06**	**2.39±0.08**	**4.88±0.02**	**7.44±0.07**	**98.16±0.13**

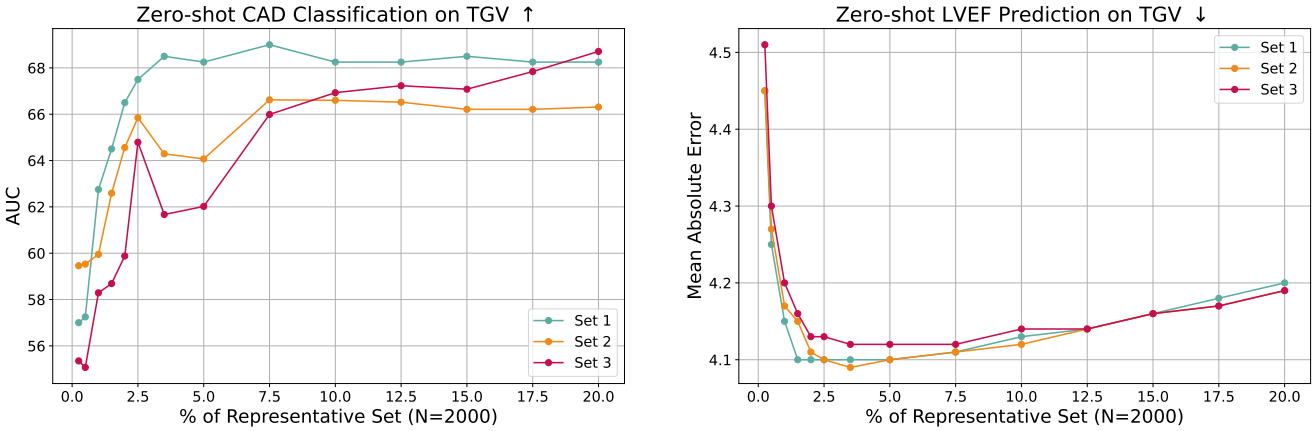


Figure 2. Zero-shot CAD classification and LVEF prediction performance as a function of the percentage of samples used for the mean aggregation out of all the samples in the representative set, i.e., number of nearest neighbors used for making the prediction (MP Eq. 9). The size of the representative sets is N=2,000. Results are reported on the validation set; the best-performing configuration is selected for final evaluation on the test set.

The final result is reported by averaging the predictions across these three subsamples.

3. Additional Cardiac Experiments

3.1. Performance under Low-Data Regimes (Complete)

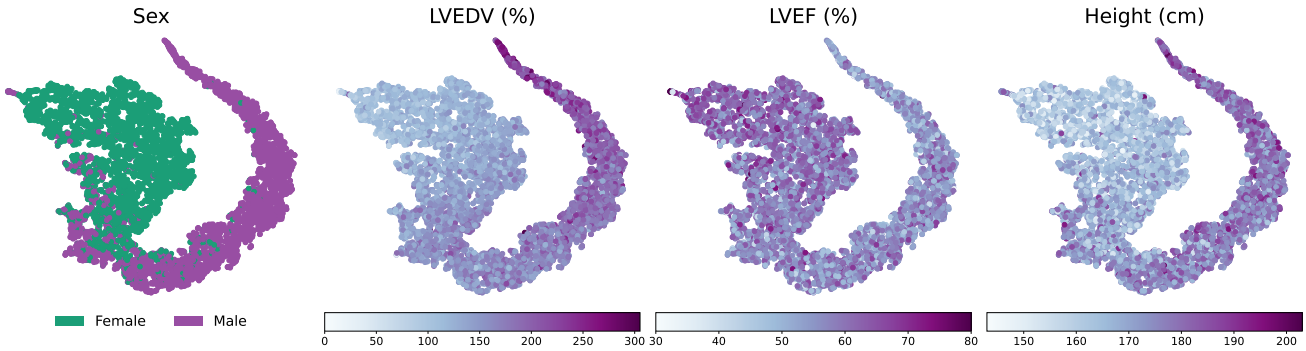
Fig. 1 presents the results on CAD classification and LVEF prediction under low-data regimes for all the baselines, which were omitted for clarity in the main body of the paper. TGV outperforms the other methods on nearly all the data regimes and all tasks, with some exceptions. MMCL [7] and SimCLR [3] are typically the second best approach, while BYOL [6], Barlow Twins [13], and SimSiam [5] report the worst overall performance.

3.2. Evaluating Robustness of the Zero-shot Predictions

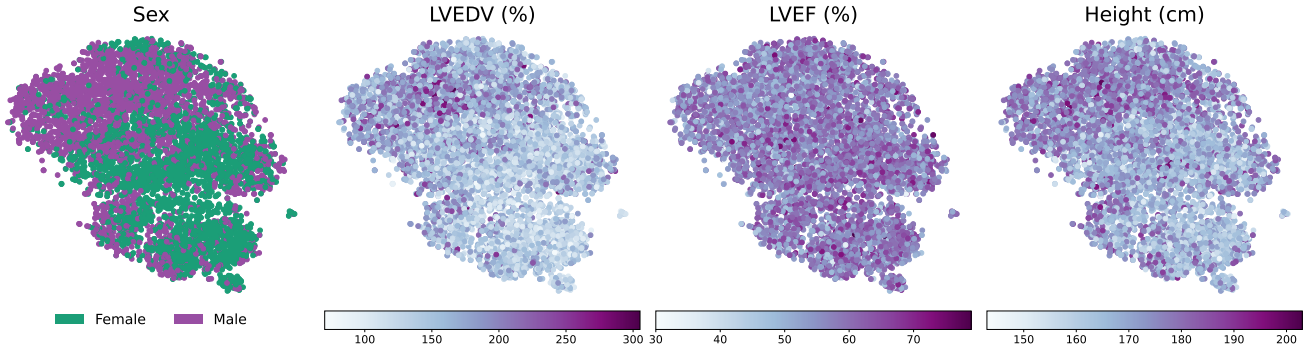
We evaluate the robustness of our zero-shot approach in terms of two conditions: (1) how changing the size of the representative set impacts performance, and (2) how changing the samples included in the representative set affects the predictions.

1) Robustness to representative set size. We evaluate the robustness of the zero-shot predictions under different sizes of the representative set P. The experiment is performed using the image encoder pretrained with TGV and the results are reported in Table 3. We consider the N=2000 as the baseline and report the changes in the performance against it. Reducing the size of the representative set leads to a worsening in the performance of the zero-shot prediction, as some target values might end up being underrepresented. However, the performance even at N=100, still outperforms the results of mean-guess for cardiac phenotype prediction and is comparable with supervised ResNet for both CAD classification and cardiac phenotype prediction (Main Paper (MP) Table 1), showing the robustness of our zero-shot prediction approach.

2) Robustness across different representative sets. Table 4 reports the mean and standard deviation of zero-shot prediction performance across three different representative sets P. CAD prediction shows the highest standard deviation, which is reflective of the small number of CAD



(a) MMCL



(b) SimCLR

Figure 3. t-SNE visualization of sex, LVEDV, LVEF, and height for a) MMCL [7] and b) SimCLR [3]. Sex, LVEDV, and LVEF were included as attributes in the pretraining of MMCL, height was not. SimCLR is an image augmentation based method. Thus, no tabular attributes were used during its training.

positive cases in the UK Biobank. Generally, methods with lower overall performance also exhibit higher standard deviation, suggesting that stronger representations yield more robust zero-shot predictions that are less sensitive to changes in the representative sets.

3.3. Nearest Neighbor Number Selection for Zero-Shot Prediction

The final zero-shot prediction is computed as the mean of the K nearest neighbors of the query image (MP Eq. 9). The number K is chosen empirically based on validation set performance. Fig. 2 illustrates the results across different values of K, expressed as a percentage of the representative set P, i.e., the amount of samples from P used for mean aggregation, for both CAD classification and LVEF prediction. The optimal K varies depending on the task and representative set. Choosing too few neighbors may yield unrepresentative predictions, while too many can bias results toward simply outputting the mean value of the representative set, especially for cardiac phenotype prediction. Therefore, selecting an appropriate K is crucial to maximize performance for the given task and data.

3.4. Representation Evaluation using t-SNE

We evaluate the representations generated by MMCL [14] and SimCLR [3] using t-SNE. Fig. 3 presents visualizations of the representations with respect to sex, LVEDV, LVEF, and height. Among these four attributes, only height was not included in the pretraining of MMCL. SimCLR, being an image augmentation-based method, did not use any tabular data during pretraining. MMCL demonstrates superior clustering ability compared to SimCLR across all attributes. While the comparison between SimCLR and MMCL already underscores the importance of tabular data-informed pretraining, TGV produces even more coherent and clinically meaningful clusters (MP Fig. 4), further solidifying the value of incorporating tabular data into contrastive learning frameworks.

4. Assessing TGV's Generalizability

4.1. Dataset

To assess whether TGV can generalize to other domains and datasets, we use the Data Visual Marketing (DVM) car dataset [10]. The dataset contains 1,451,784 images and their corresponding attributes of cars at varying degree angles. Model performance is evaluated on two tasks,

Table 5. The results on car model prediction (accuracy ↑) and price prediction (MAE ↓) on the DVM car dataset. ZS stands for zero-shot, LP for linear probing, and FT for fine-tuning.

	Model Classification ↑			Price Regression ↓		
	ZS	LP	FT	ZS	LP	FT
Mean-Guess	-	-	-	8752.2	8752.2	8752.2
ResNet50 [9]	-	-	92.95	-	-	3411.4
SimCLR [3]	13.52	56.01	89.53	6609.5	5701.3	3674.7
MMCL [7]	81.06	91.64	94.06	3621.1	4440.6	2821.6
TGV (Ours)	**83.37**	**92.52**	**94.23**	**3529.6**	**3759.6**	**2650.6**

car model classification (286 classes) and price prediction (regression). We use the same preprocessing technique as MMCL [7], which yields 70,565 image-tabular data training pairs, 17,642 validation pairs, and 88,207 test pairs. The tabular data attributes include information regarding width, length, height, wheelbase, price, advertisement year, miles driven, number of seats, number of doors, original price, engine size, body type, gearbox type, fuel type, color, and car model. MMCL achieved the best results while using all of the listed attributes, while TGV performed best without including the advertisement year, miles driven, and color. Similarly to our observations on the cardiac dataset, the choice of attributes for TGV is critical; including unrelated attributes leads to noisy and less meaningful similarity scores. For both methods, we report results using their respective best sets of tabular attributes.

4.2. Implementation Details

All the models use a 2D ResNet-50 as a backbone and are pretrained for 500 epochs under a batch size of 512. The best model is selected based on the best validation performance. For the downstream tasks, the models are further trained for 500 epochs for both linear probing and fine-tuning. This follows the setting of previous works [7]. For the baseline models we use the hyperparameters as described by the authors. TGV is pretrained using a learning rate of 1e-3. Model and price predictions are both trained using a learning rate of 3e-4 and a batch size of 512. We use the best performing baselines on UK Biobank as our baselines for DVM, namely SimCLR [3] and MMCL [7]. All the training data is used for the downstream task tuning of the methods, while the representative set is assembled using 10% of the training data.

4.3. Does TGV work on natural images?

We evaluate the generalizability of TGV to natural images by employing the DVM car dataset and evaluation the trained representations on two tasks, car model prediction and price regression. The car model prediction is evaluated using accuracy and price prediction using mean absolute error (MAE). TGV achieves the best performance on both car model classification and price regression using every tuning scheme. This shows that our approach is capable of generalizing to 2D data and natural images without incorporating any changes to the method. Both UK Biobank [11] and DVM datasets are relatively homogeneous, each representing a single organ (heart) or item (car). While the images are often visually similar, tabular data–based similarity provides guidance signals that can capture even subtle differences and enable stronger representations for downstream tasks.

References

[1] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of cardiovascular magnetic resonance*, 20(1):65, 2018. 1

[2] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3, 4, 5

[4] Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive learning. *arXiv preprint arXiv:2106.03719*, 2021. 2

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 3

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 3

[7] Paul Hager, Martin J Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23924–23935, 2023. 2, 3, 4, 5

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 1, 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[10] Jingmin Huang, Bowei Chen, Lan Luo, Shigang Yue, and Iadh Ounis. Dvm-car: A large-scale automotive dataset for visual marketing research and applications, 2023. 4

[11] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3): e1001779, 2015. 5

[12] Aleksandar Susmelj et al. Lightly. https://github.com/lightly-ai/lightly, 2020. 2

[13] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021. 2, 3

[14] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR, 2022. 4