Chronological Contrastive Learning: Few-Shot Progression Assessment in Irreversible Diseases

Clemens Watzenböck [*1,2] CLEMENS.WATZENBOECK@MEDUNIWIEN.AC.AT
Daniel Aletaha[3], **Michaël Deman**[4], **Thomas Deimel**[1,3], **Jana Eder**[3], **Ivana Janickova**[1,2], **Robert Janiczek**[4], **Peter Mandl**[3], **Philipp Seeböck**[1,2], **Gabriela Supp**[3], **Paul Weiser**[1,2,5], **Georg Langs**[*1,2] GEORG.LANGS@MEDUNIWIEN.AC.AT
[1] *Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab, Medical University of Vienna, Vienna, Austria,* [2] *Comprehensive Center for Artificial Intelligence in Medicine, Medical University of Vienna, Vienna, Austria,* [3] *Department of Medicine III, Division of Rheumatology, Medical University of Vienna, Wien, Austria,* [4] *Janssen R&D, Data Science, Analytics and Insight,* [5] *Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, Massachusetts, USA*

Editors: Under Review for MIDL 2026

Abstract

Quantitative disease severity scoring in medical imaging is costly, time-consuming, and subject to inter-reader variability. At the same time, clinical archives contain far more longitudinal imaging data than expert-annotated severity scores. Existing self-supervised methods typically ignore this chronological structure. We introduce ChronoCon, a contrastive learning approach that replaces label-based ranking losses with rankings derived solely from the visitation order of a patient's longitudinal scans. Under the clinically plausible assumption of monotonic progression in irreversible diseases, the method learns disease-relevant representations without using any expert labels. This generalizes the idea of Rank-N-Contrast from label distances to temporal ordering. Evaluated on rheumatoid arthritis radiographs for severity assessment, the learned representations substantially improve label efficiency. In low-label settings, ChronoCon significantly outperforms a fully supervised baseline initialized from ImageNet weights. In a few-shot learning experiment, fine-tuning ChronoCon on expert scores from only five patients yields an intraclass correlation coefficient of 86% for severity score prediction. These results demonstrate the potential of chronological contrastive learning to exploit routinely available imaging metadata to reduce annotation requirements in the irreversible disease domain.

Keywords: Unsupervised Learning, Contrastive Learning, Few-Shot Learning, Representation Learning, Longitudinal Medical Imaging, Disease Progression, Rheumatoid Arthritis

1. Introduction

Time is of the essence in clinical settings. Time series – repeated scans of the same patient over multiple visits – capture essential information about disease evolution and treatment response. Although this information is routinely available in clinical archives, it is rarely used for representation learning. Most deep-learning approaches rely on large annotated datasets, yet expert scoring is expensive, time-consuming, and subject to inter-reader variability. In addition, discrete ordinal scores introduced to make expert assessment feasible and comparable capture only a coarse approximation of continuous disease severity. Often, they introduce quantization errors.

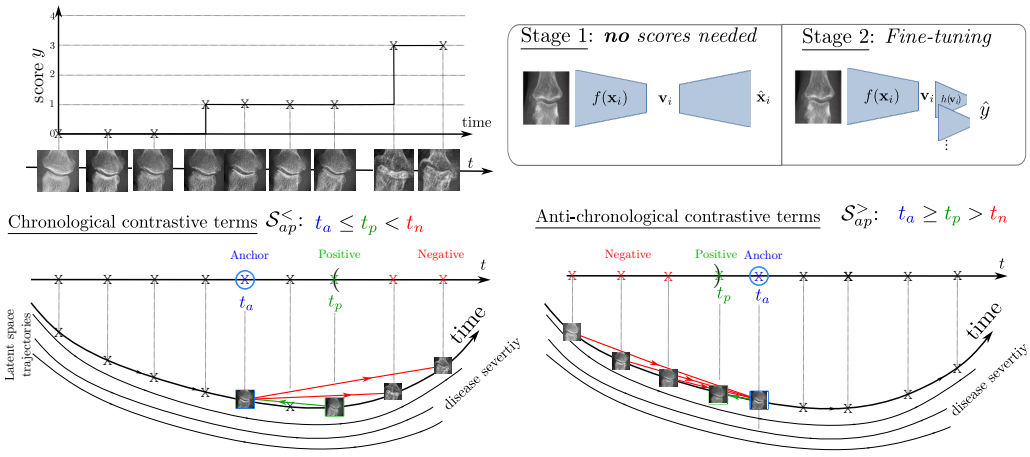

Figure 1: Chronological contrastive learning objective illustrated using a case of monotonically worsening JSN in a patient's interphalangeal (IP) joint. *Bottom:* Anti-/chronological contrastive terms. The loss aligns disease trajectories in latent space, capturing severity automatically. *Top right:* Training stages. In stage 1, no labels beyond timestamps and patient+ROI IDs are required. In stage 2, the model is fine-tuned for score prediction.

We introduce *ChronoCon*, a chronological contrastive learning objective function that uses temporal examination order to train a model for mapping imaging data to quantitative severity scores. The idea is motivated by a simple example: consider a patient with an irreversible disease who is imaged at times $t_1 < t_2 < t_3$. In the latent-disease representation, the second scan should be at least as similar to the first scan as the third is to the first. Formally, for encoded features $\mathbf{v}_i$, we expect: $\text{sim}(\mathbf{v}_1, \mathbf{v}_2) \leq \text{sim}(v_1, v_3)$. A corresponding relation holds when comparing later visits to earlier ones. $\text{sim}(\mathbf{v}_2, \mathbf{v}_3) \leq \text{sim}(\mathbf{v}_1, \mathbf{v}_3)$. These ordering constraints, illustrated in Figure 1, allow the model to learn a progression-aware feature space without using any severity labels.

Related work Recent work on image series and latent-space alignment incorporates temporal or pairwise information by jointly processing image pairs in both supervised (Kamran et al., 2025) and unsupervised settings (Bannur et al., 2023). While effective at capturing pairwise differences, these approaches do not leverage the full temporal trajectory available in longitudinal patient data.

In supervised contrastive learning, several methods define positive and negative pairs based on label ordering (Gong et al., 2022; Zha et al., 2023). (Janíčková et al., 2025) used a triplet loss with time-dependent margins as hyperparameters, which makes the approach difficult to apply to nonlinear progressions in irregularly sampled time series. Conversely, while (Couronné et al., 2021) handles irregular sampling, the soft-rank loss does not enforce discriminability across more distant visits: it preserves ordering without explicitly pushing farther-apart time points away in latent space, similar to label-distribution smoothing or feature-distribution smoothing (Yang et al., 2021).

The closest work to ours is *Rank-N-Contrast* (RnC) (Zha et al., 2023). RnC defines the conditional probability that the positive (p) is the correct match for the anchor (a) among

its negatives $n \in \mathcal{S}^{\bullet}_{ap}$ as

$$P(v_p \mid v_a, \mathcal{S}^{\bullet}_{ap}) = \frac{\exp\big[\mathrm{sim}(\mathbf{v}_a, \mathbf{v}_p)\big]}{\exp\big[\mathrm{sim}(\mathbf{v}_a, \mathbf{v}_p)\big] + \sum_{n \in \mathcal{S}^{\bullet}_{ap}} \exp\big[\mathrm{sim}(\mathbf{v}_a, \mathbf{v}_n)\big]}. \tag{1}$$

The corresponding per-pair loss is $\ell^{\bullet}_{ap} = -\log P(v_p \mid v_a, \mathcal{S}^{\bullet}_{ap})$. Negatives are selected based on distances in label space $\mathcal{S}^{\mathrm{RnC}}_{ab} := \big\{k \;\big|\; k \neq i,\ |y_a - y_n| \geq |(y_a - y_p|\big\}$, a strategy well suited for fully supervised regression problems. RnC has since been applied to segmentation (Weng et al., 2025), visual-concept explanation (Obadic et al., 2024), and extended to survival prediction (Saeed et al., 2024). However, neither RnC nor these generalizations[1] can be used *without labels* and thus cannot be applied directly to timestamps.

For a patient with visits at $t_1 = 0$, $t_2 = 1$, and $t_3 = 3$ years, RnC would imply that features between first and second visit are more similar than those between the second and third. In irreversible diseases, however, progression is nonlinear: long periods of stability may be followed by abrupt worsening. Consequently, absolute time intervals are not meaningful distances.

2. Methods

For each image $\mathbf{x}$, the corresponding relative time point t within a patient's examination series is available. For some images, an additional ordinal expert-annotated score y is provided. We use this information to encourage representations that capture disease progression. Let $\mathcal{D} = \{(\mathbf{x}_i, t_i, y_i, \mathtt{id}_i)\}_{i=1\ldots N}$ denote a dataset of N examples with imaging, time, and scoring information, where $\mathtt{id}_i$ is the group identifier determining which samples may be contrasted against each other. We propose a two-stage learning procedure. In the *first stage*, only imaging data and metadata are required. We learn a mapping $f : \mathbb{R}^{l \times w} \to \mathbb{R}^d$, $\mathbf{x} \mapsto \mathbf{v}$ from the image space to a latent representation space using ChronoCon. The goal of the *second stage* is to learn a scoring function, mapping latent representations to an estimate of the ordinal score, $h : \mathbb{R}^d \to \mathbb{R}$, $\mathbf{v} \mapsto \hat{y}$, trained using only an MSE loss.

Chronological contrastive learning. To apply contrastive learning to time-stamps, we introduce the sets of chronological negatives $\mathcal{S}^{<}_{ap}$ and anti-chronological negatives $\mathcal{S}^{>}_{ap}$ as

$$\begin{aligned} \mathcal{S}^{<}_{ap} &= \{n \;\mid\; n \neq a,\ \ \mathtt{id}_a = \mathtt{id}_p = \mathtt{id}_n,\ \ (t_a \leq t_p < t_n)\ \}, \\ \mathcal{S}^{>}_{ap} &= \{n \;\mid\; n \neq a,\ \ \mathtt{id}_a = \mathtt{id}_p = \mathtt{id}_n,\ \ (t_a \geq t_p > t_n)\ \}. \end{aligned} \tag{2}$$

Trivial pairs without valid negatives are excluded from normalization, and we define *ChronoCon* loss as the balanced[2] sum of the forward and backward chronological contributions:

$$\begin{aligned} L^{\mathrm{Chrono}}_{\mathrm{Con}} &= \tfrac{1}{|\mathcal{P}^{<}_{+}|} \textstyle\sum_{(a,p)\in\mathcal{P}^{<}_{+}} \ell^{<}_{ap} \;+\; \tfrac{1}{|\mathcal{P}^{>}_{+}|} \textstyle\sum_{(a,p)\in\mathcal{P}^{>}_{+}} \ell^{>}_{ap}, \\ \mathcal{P}^{<}_{+} &= \{(a,p) \;\mid\; \mathtt{id}_a = \mathtt{id}_p,\ (t_a \leq t_p),\ |\mathcal{S}^{<}_{ap}| > 0\}, \\ \mathcal{P}^{>}_{+} &= \{(a,p) \;\mid\; \mathtt{id}_a = \mathtt{id}_p,\ (t_a \geq t_p),\ |\mathcal{S}^{>}_{ap}| > 0\}. \end{aligned} \tag{3}$$

1. In (Saeed et al., 2024), time-to-event was also the prediction target, and repeated imaging for the same patient was not considered.
2. One might also attempt to use only forward chronological contributions, but then early images would be under-represented as positives and late images over-represented as negatives.

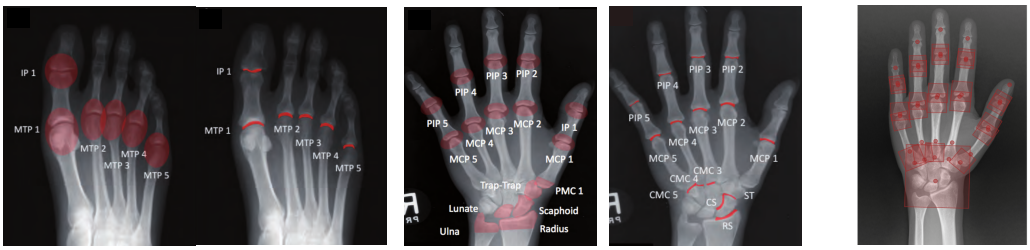

Figure 2: *Left:* Joint contributions to the total SvHS, highlighting erosive areas and joint spaces. Image reproduced from (Peterfy et al., 2011). *Right:* Regions of interest extracted during fully automatic preprocessing of hand radiographs.

Our loss provides a natural way of enforcing order via ranking with respect to t for all samples sharing the same group identifier (`id`). We explicitly avoid imposing a metric on t. Intuitively, such ordering should also support improved prediction of the target y in downstream tasks, provided y and t exhibit a monotone relationship.

Ordinal contrastive learning. – This property also makes the loss well suited for ordinal regression. While not our primary focus, we evaluate the loss on ordinal disease-severity labels by adjusting the group identifiers accordingly. To distinguish this setting from our main objective—the unsupervised application to time-stamps—we denote the loss used on labels y as $L_{\text{Con}}^{\text{Ordinal:Y}}$, emphasizing the *ordinality* in the pair selection process.[3]

Data augmentation. Without augmentation, only image series of length *three or more* would contribute to the loss, as at least one anchor, one positive, and one negative are required. To enable training on series with only two visits, we apply double-crop augmentation following (Zha et al., 2023).

Application of ChronoCon in rheumatoid arthritis (RA) radiographs We evaluate this approach on radiographs of 778 patients with RA to demonstrate that chronological information in routine imaging can yield clinically meaningful representations even without expert annotations. Disease severity in RA is commonly quantified using the Sharp–van der Heijde (SvH) score, which aggregates erosion and joint-space narrowing (JSN) subscores for multiple joints, resulting in a total score ranging from 0 to 448. These subscores are discrete and costly to obtain, making this domain a representative and challenging test case for label-efficient learning (van der Heijde, 2000).

First, we localize the joints with an automatic landmark-detection method (Payer et al., 2019; Jonkers et al., 2025) and extract an image patch for each detected joint (illustrated in Figure 2). For the first stage of training, contrastive pairs are constructed only from patches belonging to the *same patient, side, and joint type.* $\text{sim}(\mathbf{v}_i, \mathbf{v}_j) = -L_2(\mathbf{v}_i, \mathbf{v}_j)/\tau$ with temperature $\tau = 1$ was used as similarity metric throughout . To stabilize training, we add a standard denoising autoencoder (DAE) reconstruction loss L_2^{DAE}.

In the second stage, a multi-headed regressor replaces the decoder, with one head for each score type (59 in total). During fine-tuning, the encoder parameters are also updated, but with a reduced learning rate. The only loss used in this stage is mean-squared error (MSE) on the ERO/JSN scores.

3. One might in this respect refer to $L_{\text{Con}}^{\text{Chrono}}$ as $L_{\text{Con}}^{\text{Ordinal:t}}$, but we refrain from this to avoid confusion.

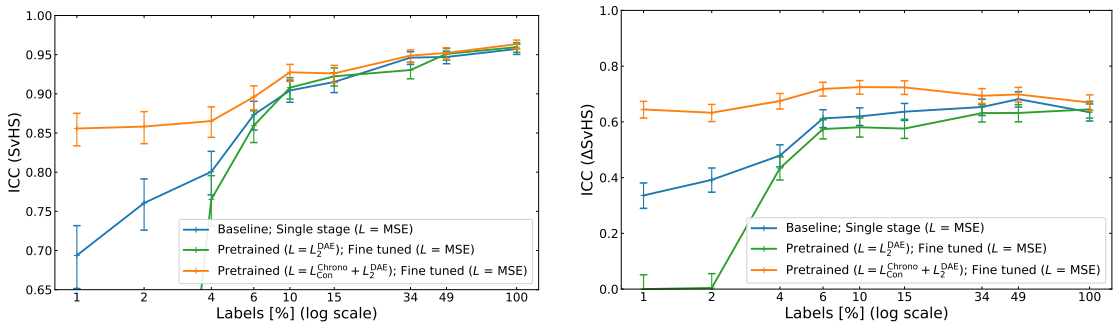

Figure 3: ICC of standard of reference and estimated SvHS as a function of training set size: (*left*) $SvHS$, and (*right*) change $\Delta SvHS$; blue: only supervised training, green: pre-trained with reconstruction loss; orange: pre-trained with ChronoCon and reconstruction loss. Error bars indicate 95% CI.

Quality measures. All models are trained solely to predict *cross-sectional* SvH scores. Individual JSN and ERO scores are then summed to obtain the total SvH score, and differences between visits (ΔSvHS) are computed afterward. Performance is evaluated for both single–time point predictions and the derived progression ΔSvHS. Agreement with ground truth is quantified using the intraclass correlation coefficient (ICC), root-mean squared error (RMSE), and Pearson's correlation coefficient ρ. Full metric definitions are in Appendix C.

3. Experiments and Results

Dataset The dataset consists of hand and foot radiographs from 778 patients with RA. It comprises 13 742 radiographic images and a total of 407 045 individual scores across 59 score types. As detailed in Table A in the appendix, the score distribution is highly imbalanced: fewer than 1% of erosion scores fall into the highest category, and fewer than 4% of JSN scores do. The dataset also exhibits only short longitudinal series, with a median of 4 visits per patient (IQR [3,5]).

We use a patient-level split to avoid leakage of longitudinal information. Training, validation, and test sets contain 466/155/157 patients with 8 157/2 753/2 832 images and 241 701/81 501/83 843 scores, respectively.

Model and training details We used ResNet18 as the encoder for all models (He et al., 2015). For the DAE task, the decoder mirrored the encoder architecture, using transposed convolutions for upsampling.

A hierarchically grouped dataloader was used to improve temporal consistency: patches from the same ROI and patient were typically placed in the same batch and oversampled based on intra-patient median ERO/JSN scores to mitigate score imbalance. Early stopping on validation mean absolute error (MAE) with a 10-epoch patience was applied for fine-tuning and for the supervised baseline, restoring the best-performing model.

All preprocessing steps and model and training configurations will be made available upon acceptance of this manuscript at https://github.com/cirmuw. See also Appendix C.

Dataset size				Cross sect. (SvHS)		Longit. (ΔSvHS)	
scores [%]	scores [N]	images	patients	RMSE ↓	ICC ↑	RMSE ↓	ICC ↑
1	2 442	82	5	19.9 $^{\mathbf{-7.0}}_{\pm 2.5}$	86 $^{\mathbf{+17}}_{\pm 2}$	9.5 $^{\mathbf{-1.9}}_{\pm 0.7}$	64 $^{\mathbf{+30}}_{\pm 3}$
2	4 475	152	10	19.3 $^{\mathbf{-4.5}}_{\pm 2.6}$	86 $^{\mathbf{+10}}_{\pm 2}$	9.0 $^{\mathbf{-2.1}}_{\pm 0.6}$	63 $^{\mathbf{+24}}_{\pm 3}$
4	9 413	319	20	18.7 $^{\mathbf{-4.7}}_{\pm 2.7}$	87 $^{\mathbf{+7}}_{\pm 2}$	8.1 $^{\mathbf{-2.3}}_{\pm 0.5}$	67 $^{\mathbf{+19}}_{\pm 3}$
6	14 719	499	31	17.2 $^{\mathbf{-2.1}}_{\pm 2.7}$	90 $^{\mathbf{+3}}_{\pm 2}$	7.7 $^{\mathbf{-1.4}}_{\pm 0.6}$	72 $^{\mathbf{+11}}_{\pm 2}$
10	23 546	808	46	14.5 $^{\mathbf{-2.9}}_{\pm 2.1}$	93 $^{\mathbf{+3}}_{\pm 1}$	7.8 $^{\mathbf{-1.1}}_{\pm 0.5}$	72 $^{\mathbf{+10}}_{\pm 2}$
15	36 243	1 238	71	14.5 $^{\mathbf{-1.6}}_{\pm 2.1}$	93 $^{\mathbf{+1}}_{\pm 1}$	7.7 $^{\mathbf{-1.0}}_{\pm 0.5}$	72 $^{\mathbf{+8}}_{\pm 2}$
34	79 799	2 745	156	12.5 $^{\mathbf{-0.5}}_{\pm 1.6}$	95 $^{\mathbf{+0}}_{\pm 1}$	8.2 $^{\mathbf{-0.5}}_{\pm 0.7}$	69 $^{\mathbf{+4}}_{\pm 3}$
49	116 448	3 989	226	12.1 $^{\mathbf{-1.2}}_{\pm 1.6}$	95 $^{\mathbf{+0}}_{\pm 1}$	8.0 $^{\mathbf{-0.2}}_{\pm 0.6}$	70 $^{\mathbf{+2}}_{\pm 3}$
100	237 733	8 157	466	10.8 $^{\mathbf{-1.0}}_{\pm 1.3}$	96 $^{\mathbf{+0}}_{\pm 1}$	8.4 $^{\mathbf{-0.6}}_{\pm 0.7}$	67 $^{\mathbf{+4}}_{\pm 3}$

Table 1: Test-set performance of our model after pretraining with ChronoCon loss on 466 patients and fine-tuning on a fraction of labeled data. Superscripts in green/red show improvement over the supervised baseline; subscripts give half the 95% CI.

3.1. Label efficiency

A key advantage of our loss is that it enables learning meaningful feature representations without access to scores y. To evaluate label efficiency, we created progressively smaller training subsets by reducing the number of patients with labeled data. The full training set comprises images from 466 patients. All splits were performed at the patient level, allowing us to simulate how performance changes when labels are available for only a subset of patients. The validation and test sets remained fixed across all experiments (155 and 157 patients, respectively). Details on the splits are provided in Table 3.1.

Setup. As a *baseline*, the encoder (initialized with ImageNet weights) and the multi-headed regressor were trained in a single stage using MSE loss.

Training with the *ChronoCon* loss was conducted in two stages. In the first stage, the loss was $L = L_{\text{Con}}^{\text{Chrono}}(\{\mathbf{v}_i, t_i, \texttt{id}_i\}_{i=1}^{2b}) + 10^3 \cdot L_2^{\text{DAE}}(\mathbf{x}_i, \hat{\mathbf{x}}_i)$, with a double-crop augmentation in the contrastive term, effectively doubling the batch size b. In the second stage, the decoder was removed and the multi-headed regressor added for fine-tuning on (a subset of) scores; the encoder learning rate was reduced by a factor of 10. Performance for this model is shown as the orange line in Figure 3, with numerical values in Table 3.1; differences relative to the baseline without pretraining appear as red/green superscripts.

As an additional baseline, the same two-stage procedure was repeated using DAE pretraining only (green line in Figure 3); this variant is not shown in the table.

Observations and Interpretation. Our method using $L_{\text{Con}}^{\text{Chrono}} + L_2^{\text{DAE}}$ outperforms the baseline across all label sub-splits. The improvement is driven by $L_{\text{Con}}^{\text{Chrono}}$, not by L_2^{DAE}; in fact, DAE pretraining alone performs worse.

Performance gains are most pronounced in the low-label setting. Even in a *few-shot* scenario with labels from only 5 patients, our method achieves an ICC of 0.86 and RMSE of

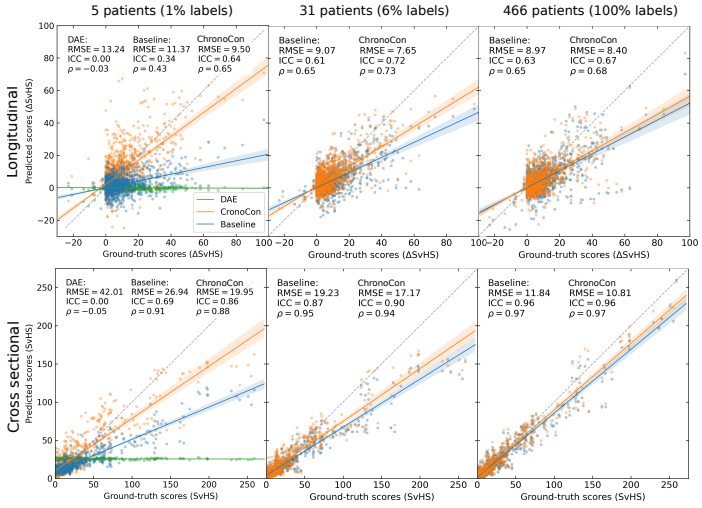

Figure 4: Scatter plots comparing ground truth and model predictions for the baseline model and our method trained on labels from 5, 31, and 466 patients. *Top:* Longitudinal evaluation in terms of score differences between visits. *Bottom:* Cross-sectional prediction performance for total SvH scores.

19.9. For context, the SvHS has a standard deviation of 46 in the full dataset. Remarkably, ChronoCon trained on just 5 patients performs on par with a recently published model (RMSE = 23.6) trained on 367 patients [4]

Longitudinal evaluation (ΔSvHS) shows even larger gains from ChronoCon in the low-label regime. Notably, its performance remains almost constant over a wide range of training-set sizes, unlike the baseline. This stems from explicitly learning patient-specific progression in an unsupervised manner. With only 4% of labels, results match those obtained using 100% of labels. Interestingly, longitudinal performance peaks at 6–15% of labels rather than at full supervision, suggesting that strong (and noisy) labels may override the features learned during pretraining.

Scoring consistency—The progression error are substantially better than expected from subtracting two noisy estimates. If $\hat{y}_i = y_i + \epsilon_i$ with noise variance $\sigma^2 = \mathbb{E}[\epsilon_i^2]$, then $\mathrm{MSE}(\Delta\mathrm{SvHS}) = 2\sigma^2(1-c)$, where $\mathbb{E}[\epsilon_i\epsilon_j]/\sigma^2 = c$ is the error correlation ($i \neq j$). For uncorrelated errors, RMSE(ΔSvHS) should be $\sqrt{2}$ times RMSE(SvHS). However, our errors are highly correlated ($c = 0.91$ for 4% labels; $c = 0.70$ for 100%), indicating strong error cancellation—i.e., *scoring consistency*—at least partly due to ChronoCon pretraining.

3.2. Learned feature space

To further investigate the feature space learned in the first stage with ChronoCon, the embedding is visualized in Figure 5 *a* and *b*. The 512-dimensional features were reduced to 2 dimensions using principal component analysis (PCA). The training process is completely invariant to global time shifts for each patient because only repeatedly acquired patches of

4. and substantially outperforms (Moradmand and Ren, 2025a) (trained on 428 patients/visits) who reported RMSE = 44.28 on scores 0–270.

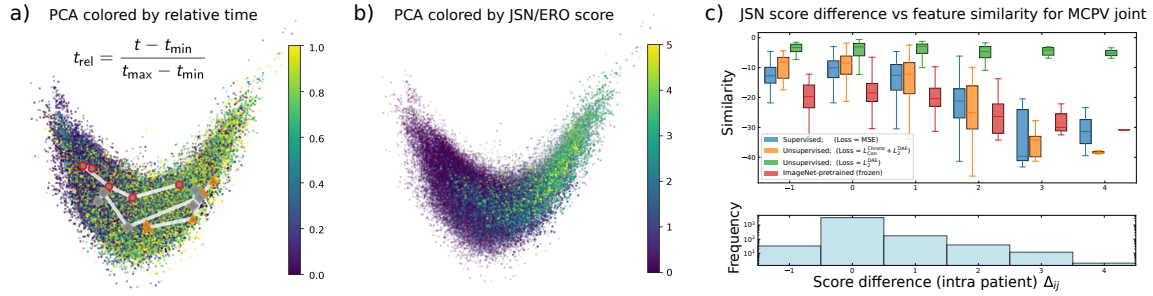

Figure 5: Feature space (PCA) of the unsupervised model pretrained with ChronoCon. *Left (a):* Colored by relative time t_{rel} (0 = first visit, 1 = last); white lines show example patient–joint trajectories. *Middle (b):* Same embedding colored by ground-truth scores (no score information was used during training). *Right (c):* Feature similarity between chronologically ordered visits compared to the corresponding joint-space–narrowing (JSN) score differences for the MCPV joint.

the same patient, side, and joint type are contrasted against each other. In Figure 5 *a*, the embedding is colored by relative time, with three example trajectories shown as white lines. In Figure 5 *b* the same embedding is colored by JSN/ERO labels. Importantly, no labels were used during training. Points are displayed in order of increasing score to highlight the transition from low to high severity. All ERO and JSN patches are shown in the same plot, even though their respective maximum scores differ (5 and 4).

Figure 5 *c* shows intra-patient feature similarities between pairs of visits for the MCPV joint. Equivalent plots for all other joints are provided in the appendix (Figure 6). The visit pairs are chronologically ordered, though not necessarily consecutive for patients with more than two visits. The lower panel displays a histogram of JSN score differences between the two visits. Scores typically increase over time, although decreases of up to –1 also occur, likely due to quantization effects in the discrete scoring system.

The DAE baseline is shown in green, our main model—which combines the reconstruction loss with ChronoCon—is shown in orange, and the model using only the first 17 layers of ResNet18 (pretrained on ImageNet) as encoder is shown in red. None of these three models had access to ERO/JSN scores. The only model which used scores during training is the baseline shown in blue.

Observations and interpretation. Among the unsupervised baselines, the DAE shows the weakest behavior: its feature similarities exhibit little correspondence with progression severity. In contrast, the frozen ImageNet encoder can already distinguish progression to some extent—the red box-plots roughly follow the trend of the supervised baseline. This aligns with the observation that the supervised baseline, when trained on labels from only 5 patients, still achieves an ICC of 0.34 for ΔSvHS.

Our full model provides the clearest separation of progression patterns. The embeddings are effectively ordered only along each *individual* trajectory and disease-related features are learned automatically. Remarkably, coloring the embedding by visit time appears more disordered than when coloring by severity. This suggests that, in the second stage, the regression heads only need to learn which regions of feature space correspond to which

model	Cross sectional (SvHS)		Longitudinal (ΔSvHS)	
	RMSE ↓	ICC ↑	RMSE ↓	ICC ↑
single stage baseline	11.8 ± 1.4	95.7 ± 0.6	9.0 ± 0.8	63.5 ± 3.1
L_2^{DAE}	12.9 ± 1.1	96.0 ± 0.6	9.1 ± 0.6	64.5 ± 3.0
$L_{\text{Con}}^{\text{Chrono}}$	11.0 ± 1.3	96.4 ± 0.5	<u>8.4</u> ± 0.7	66.8 ± 2.8
$L_{\text{Con}}^{\text{Chrono}} + L_2^{\text{DAE}}$	<u>10.8</u> ± 1.3	<u>96.3</u> ± 0.6	<u>8.4</u> ± 0.7	<u>67.0</u> ± 2.8
L^{RnC}	10.4 ± 1.1	96.6 ± 0.5	8.4 ± 0.7	67.6 ± 2.8
$L^{\text{RnC}} + L_{\text{Con}}^{\text{Chrono}}$	10.9 ± 1.2	96.3 ± 0.6	8.7 ± 0.7	65.1 ± 3.0
$L^{\text{RnC}} + L_{\text{Con}}^{\text{Chrono}} + L_2^{\text{DAE}}$	10.8 ± 1.2	96.3 ± 0.6	8.6 ± 0.7	66.7 ± 2.8
$L_{\text{Con}}^{\text{Ordinal:Y}}$	**9.8** ± 0.9	**97.1** ± 0.4	**8.2** ± 0.6	**70.6** ± 2.6
$L_{\text{Con}}^{\text{Ordinal:Y}} + L_{\text{Con}}^{\text{Chrono}}$	10.7 ± 1.3	96.4 ± 0.6	8.3 ± 0.7	68.4 ± 2.7
$L_{\text{Con}}^{\text{Ordinal:Y}} + L_{\text{Con}}^{\text{Chrono}} + L_2^{\text{DAE}}$	10.4 ± 1.1	96.6 ± 0.5	8.5 ± 0.7	67.1 ± 2.8

Table 2: Different pre-training strategies on the full training set of 466 patients. Values are shown with 95% CI. Underlined metrics indicate the best methods that do not require labels during pretraining; bold values indicate the best overall performance. Except for the baseline, all models were trained in two stages.

scores—explaining the strong performance even with labels from just 5 patients. Consistently, the feature similarities of our model (orange) closely follow those of the supervised baseline (blue), which was trained on patch–score pairs from all 466 patients.

3.3. Combination with other pre-training methods and ablation study

Different pre-training strategies are summarized in Table 3.3. All models were trained on the full set of 466 patients and their corresponding scores. The table is organized into four groups: (i) the supervised baseline (top), (ii) unsupervised methods where no labels y are used in the first stage, (iii) supervised first-stage training with the original RnC loss applied to labels, and (iv) supervised first-stage training with our loss applied to *labels* ($L_{\text{Con}}^{\text{Ordinal:Y}}$).

Whenever a contrastive loss is applied to labels, stratification via `id` uses the score type (e.g., `IP_JSN`, `PIPII_ERO`, ...), so any–vs–any patient pairs are allowed as long as they correspond to the same ROI and score type.

Observations and interpretation. Neither as a standalone pretraining task nor in combination with other losses did the reconstruction/denoising objective L_2^{DAE} yield meaningful improvements. In contrast, ChronoCon pretraining improved performance—particularly ICC(ΔSvH)—over the other unsupervised methods. However, when all labels were also used during pretraining (via $L_{\text{Con}}^{\text{Ordinal:Y}}$ or L^{RnC}), these methods performed best. Overall, the strongest results were obtained with our $L_{\text{Con}}^{\text{Ordinal:Y}}$ loss applied to *labels*, outperforming even the original RnC pretraining.

We speculate that this advantage arises because JSN/ ERO are ordinal scores, and our formulation respects this, whereas the RnC loss does not. $0 \to 1$ differs from $1 \to 2$.

4. Discussion

Summary We introduced ChronoCon, a chronological contrastive loss that exploits visitation order in longitudinal imaging to learn disease-relevant representations without expert labels. In RA radiographs, the learned feature space captured both cross-sectional severity and longitudinal progression, and clearly improved performance in low-label scenarios compared with purely supervised or reconstruction-based pretraining. These findings highlight that chronological information routinely present in clinical archives can serve as a powerful and inexpensive supervisory signal.

Compared with typical representation-learning approaches that operate on individual images or unordered pairs, our method leverages temporal ordering as an additional inductive bias. While several self-supervised objectives have been explored in medical imaging, few explicitly account for longitudinal structure, and we are not aware of prior work using visitation order as a supervisory signal for predominantly irreversible diseases. The observed improvements in low-label settings suggest that temporal ordering can provide complementary information in an irreversible disease setting.

Limitations and ethical concerns. The usefulness of ChronoCon depends on the presence of a valid ordering variable t within subgroups of shared `id`. When t denotes visit time, the loss assumes a predominantly monotonic progression. This is a reasonable approximation for erosive changes in RA (van der Heijde, 2000), but may not hold in diseases with non-monotonic or treatment-reversible patterns.

Beyond this conceptual limitation, several practical aspects should be noted. First, our experiments are based on a single-center dataset, and broader multi-center validation will be necessary. Second, the method relies on sufficient longitudinal coverage; in datasets dominated by single visitations, the benefit of chronological contrastive learning is limited.

Furthermore, the choice of subgroup identifiers also warrants careful consideration. In our RA study, `id` was defined at the joint level to avoid semantically implausible comparisons. More broadly, subgrouping can encode clinically meaningful structure, but our approach could be used with demographic or biologically sensitive categories which raises ethical concerns. Depending on the application, subgroup definitions may either improve representation quality or inadvertently entrench biases, making transparent justification essential.

Conclusions. Chronological contrastive learning provides a simple and effective way to leverage unlabeled longitudinal imaging data for representation learning. By using only visitation order, it generalizes label-based contrastive ranking to a setting where expert scores are not required, enabling strong performance even when labels are scarce. Our experiments on RA radiographs demonstrate that chronological signals embedded in routine clinical workflows contain exploitable structure for learning progression-aware feature spaces. The approach has potential relevance for other predominantly irreversible diseases and may help reduce annotation burden in domains where expert scoring is costly or inconsistent.

Acknowledgments

This project is supported by the Innovative Health Initiative Joint Undertaking and its members, and other contributing partners, under grant agreement No. 101194766. C.W. thanks Marlene Steiner and Simon Schürer-Waldheim for many insightful discussions.

References

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing, March 2023. URL http://arxiv.org/abs/2301.04558. arXiv:2301.04558 [cs].

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

Zhiyan Bo, Laura C. Coates, and Bartlomiej W. Papiez. Interpretable Rheumatoid Arthritis Scoring via Anatomy-aware Multiple Instance Learning, August 2025. URL http://arxiv.org/abs/2508.06218. arXiv:2508.06218 [cs] version: 1.

Raphaël Couronné, Paul Vernhet, and Stanley Durrleman. Longitudinal self-supervision to disentangle inter-patient variability from disease progression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 231–241. Springer, 2021.

Thomas Deimel, Paul J. Weiser, Martin Urschler, Christian Payer, Georg Langs, and Daniel Aletaha. autoscora: Deep learning to automate sharp/van der heijde scoring of radiographic damage in rheumatoid arthritis. Manuscript under review (revision), 2025.

Yu Gong, Greg Mori, and Frederick Tung. Ranksim: Ranking similarity regularization for deep imbalanced regression, 2022. URL https://arxiv.org/abs/2205.15236.

Darshana Govind, Zijun Gao, Chaitanya Parmar, Kenneth Broos, Nicholas Fountoulakis, Lenore Noonan, Shinobu Yamamoto, Natalia Zemlianskaia, Craig S. Meyer, Emily Scherer, Michael Deman, Pablo Damasceno, Philip S. Murphy, Terence Rooney, Elizabeth Hsia, Anna Beutler, Robert Janiczek, Stephen S. F. Yip, and Kristopher Standish. Vision Transformer Model for Automated End-to-End Radiographic Assessment of Joint Damage in Psoriatic Arthritis. In Xuanang Xu, Zhiming Cui, Islem Rekik, Xi Ouyang, and Kaicong Sun, editors, *Machine Learning in Medical Imaging*, pages 94–103, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73284-3.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Ivana Janíčková, Yen Y. Tan, Thomas H. Helbich, Konstantin Miloserdov, Zsuzsanna Bago-Horvath, Ulrike Heber, and Georg Langs. Temporal Representation Learning of Phenotype Trajectories for pCR Prediction in Breast Cancer . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15974. Springer Nature Switzerland, September 2025.

Jef Jonkers, Luc Duchateau, Glenn Van Wallendael, and Sofie Van Hoecke. landmarker: A toolkit for anatomical landmark localization in 2d/3d images. *SoftwareX*, 30:102165, 2025. ISSN 2352-7110. doi: https://doi.org/10.1016/j.softx.2025.102165. URL https://www.sciencedirect.com/science/article/pii/S2352711025001323.

Mohammed Kamran, Maria Bernathova, Raoul Varga, Christian Singer, Zsuzsanna Bago-Horvath, Thomas Helbich, Georg Langs, and Philipp Seeböck. *LesiOnTime - Joint Temporal and Clinical Modeling for Small Breast Lesion Segmentation in Longitudinal DCE-MRI*, pages 329–339. Springer, Cham, 09 2025. ISBN 978-3-032-05558-3. doi: 10.1007/978-3-032-05559-0_33.

Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, June 2016. ISSN 1556-3707. doi: 10.1016/j.jcm.2016.02.012. URL https://doi.org/10.1016/j.jcm.2016.02.012.

Deyu Ling, Wenxin Yu, Zhiqiang Zhang, and Jinmei Zou. An attention network with self-supervised learning for rheumatoid arthritis scoring. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2024. doi: 10.1109/ISCAS58744.2024.10558434.

Krzysztof Maziarz, Anna Krason, and Zbigniew Wojna. Deep learning for rheumatoid arthritis: Joint detection and damage scoring in x-rays, 2022. URL https://arxiv.org/abs/2104.13915.

Hajar Moradmand and Lei Ren. Multistage deep learning methods for automating radiographic sharp score prediction in rheumatoid arthritis. *Scientific Reports*, 15(1), January 2025a. ISSN 2045-2322. doi: 10.1038/s41598-025-86073-0. URL http://dx.doi.org/10.1038/s41598-025-86073-0.

Hajar Moradmand and Lei Ren. Multistage deep learning methods for automating radiographic sharp score prediction in rheumatoid arthritis. *Scientific Reports*, 15(1): 3391, January 2025b. ISSN 2045-2322. doi: 10.1038/s41598-025-86073-0. URL https://www.nature.com/articles/s41598-025-86073-0. Publisher: Nature Publishing Group.

Ivica Obadic, Alex Levering, Lars Pennig, Dario Oliveira, Diego Marcos, and Xiaoxiang Zhu. Contrastive pretraining for visual concept explanations of socioeconomic outcomes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 575–584, 2024. doi: 10.1109/CVPRW63382.2024.00062.

Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Medical Image*

Analysis, 54:207–219, 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.03.007. URL https://www.sciencedirect.com/science/article/pii/S1361841518305784.

Charles G Peterfy, Chun Wu, Ying Lu, and Mark Genovese. Comparison of the genant-modified sharp and van der heijde-modified sharp scoring methods for radiographic assessment in rheumatoid arthritis. *International Journal of Clinical Rheumatology*, 6(1): 15, 2011.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Numan Saeed, Muhammad Ridzuan, Fadillah Adamsyah Maani, Hussain Alasmawi, Karthik Nandakumar, and Mohammad Yaqub. SurvRNC: Learning Ordered Representations for Survival Prediction using Rank-N-Contrast . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland, October 2024.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Dongmei Sun, Thanh M. Nguyen, Robert J. Allaway, Jelai Wang, Verena Chung, Thomas V. Yu, Michael Mason, Isaac Dimitrovsky, Lars Ericson, Hongyang Li, Yuanfang Guan, Ariel Israel, Alex Olar, Balint Armin Pataki, Gustavo Stolovitzky, Justin Guinney, Percio S. Gulko, Mason B. Frazier, Jake Y. Chen, James C. Costello, S. Louis Bridges, and RA2-DREAM Challenge Community. A Crowdsourcing Approach to Develop Machine Learning Models to Quantify Radiographic Joint Damage in Rheumatoid Arthritis. *JAMA network open*, 5(8):e2227423, August 2022. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2022.27423.

DREAM RA2 Challenge Team. Ra2-dream challenge: Winning solution. https://www.synapse.org/Synapse:syn21680642/wiki/604549, 2020. Accessed: 2025-12-05.

D. van der Heijde. How to read radiographs according to the sharp/van der heijde method. *The Journal of Rheumatology*, 27(1):261–263, 2000. Corrected and republished.

Yuzhe Weng, Haotian Wang, Tian Gao, Kewei Li, Shutong Niu, and Jun Du. Enhancing multimodal sentiment analysis for missing modality through self-distillation and unified modality cross-attention. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889029.

William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2025. URL https://CRAN.R-project.org/package=psych. R package version 2.5.6.

Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International conference on machine learning*, pages 11842–11851. PMLR, 2021.

	Score distribution (counts)						
Type	0	1	2	3	4	5	NS
ERO hands	160 000	11 838	3 657	1 603	395	412	2 613
ERO feet	66 581	8 733	2 472	980	221	424	2 177
JSN hands	63 165	24 048	7 043	5 248	3 093	–	1 548
JSN feet	24 279	9 221	2 338	2 225	1 640	–	1 091

Table 3: Erosion and joint-space narrowing scores on joint-(part) level. NS = Not scoreable (e.g. surgical spacers, fused joints, missing fingers, ...).

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression, 2023. URL https://arxiv.org/abs/2210.01189.

Appendix A. Vienna RA dataset details

Data were collected between 1997 and February 2018 at the Division of Rheumatology, Department of Internal Medicine III, Medical University of Vienna. The study protocol outlining the retrospective data analysis was approved by the local ethical committee of the Medical University of Vienna (vote number : 1206/2018). Table A shows the score distribution for erosion and joint-space narrowing speperated by hands and feet joints.

Appendix B. Related work for RA and SvH score estimation

Most published work on SvHS prediction in RA has relied on fully supervised learning without unsupervised pretraining beyond ImageNet initialization (Sun et al., 2022). (Maziarz et al., 2022) a combined objective of ROI segmentation together and smoothed label classification highlighting to address the quantification error in JSN and ERO scores. (Bo et al., 2025) proposed an attention-based multiple-instance learning model to obtain an interpretable SvHS predictor. (Moradmand and Ren, 2025b) used a vision transformer to aggregate per-joint predictions into a total SvH score, achieving strong performance in common, less severe cases. The winning RA2-DREAM Challenge approach (Team, 2020) used a pipeline that combined joint localization through segmentation with a subsequent model that integrates all joint scores per patient. More recent work explored self-supervision (Ling et al., 2024) and unsupervised pretraining of a vision transformer in a cohort of patients with psoriatic arthritis (Govind et al., 2025). To our knowledge, no existing model explicitly leverages the time-series structure of longitudinal RA imaging.

Appendix C. Training-/ evaluation details and additional results

Our full preprocessing and training code, including all parameter settings, will be made publicly available after acceptance of this manuscript at https://github.com/cirmuw.

Preprocessing All right-hand and right-foot radiographs were horizontally mirrored for consistency. Images originally encoded in the DICOM `MONOCHROME2` format (black foreground on white background) were converted to `MONOCHROME1`. Radiographs containing both hands or both feet were split at the midline.

Joint detection Joint localization was performed using the Spatial Configuration Network (SCN) (Payer et al., 2019), implemented as described in (Jonkers et al., 2025). After training on 480 radiographs, landmark detection achieved a mean median point-to-point error of $<$1.0 mm (mean over ROIs; median over samples) on a test set of 40 radiographs. Square patches of size 156×156 pixels were extracted around each region of interest (ROI).

Data augmentation Data augmentation included random rotations (up to 10°), translations (up to 17 pixels), followed by a center crop to 128×128 pixels to avoid padded boundaries. Photometric augmentations consisted of random intensity scaling, intensity shifting, contrast adjustment, and histogram shifting. Additional robustness augmentations included Gaussian smoothing and light noise perturbations. All image patches were finally normalized to the intensity range $[0, 1]$.

Encoder All experiments used a ResNet18 encoder. Prior to training, the encoder was initialized with ImageNet-pretrained weights.

Decoder For reconstruction or denoising tasks, a decoder mirroring the ResNet18 architecture was employed, with deconvolution (transposed convolution) layers replacing convolutional downsampling layers. When reconstruction was used, Gaussian noise of magnitude 10^{-5} (with clip) was added to the input to implement a denoising autoencoder. Whenever the decoder was included, memory requirements doubled and the batch size was therefore halved.

Regression heads Score prediction used a multi-headed regression module comprising 59 independent heads (one per score subtype). Each head was a multilayer perceptron with two hidden layers of dimension 128. Whenever supervised regression was used, the MSE loss was applied.

Score summation and extrapolation For erosion, the proximal and distal parts of the affected joints were scored separately. In the computation of the total SvH score, these two parts were summed, and the model predictions were combined in the same way. For foot joints, the total erosion score per joint must not exceed 10 (5 for each joint part). Accordingly, the outputs of the regression heads ($y \in \mathbb{R}$) were clipped to the range $[0, 5]$ for each foot joint part. For the PIP and MCP joints of the hand, the sum of proximal and distal parts was clipped to the range $[0, 5]$.

Metrics and analysis Erosion and joint-space-narrowing scores were summed per visit to obtain the total SvH score. When subscores were missing (e.g., not scoreable due to surgery), the total score was estimated via linear interpolation. Visits with more than 25% missing subscores were excluded. To assess the model's ability to capture progression, we evaluated the change in total SvH score between visits, $\Delta\text{SvHS} = \text{SvHS}(t_2) - \text{SvHS}(t_1)$.

Inter-class correlation coefficients were computed using a two-way mixed-effects model with single measures and absolute agreement (ICC3-1 in the terminology of (Shrout and

Fleiss, 1979)). All ICC values and confidence intervals were calculated with the *R* package `psych` (R Core Team, 2021; William Revelle, 2025).

In scatter plots of true vs. predicted SvH or ΔSvH scores, Pearson's correlation coefficient ρ is reported. Error bars represent 95% confidence intervals (CI). For both the root-mean-squared error (RMSE) and ρ, confidence intervals were obtained via bootstrapping. In tables, reported ± values correspond to half the width of the 95% CI.

Training parameters Models were trained on NVIDIA A100 GPUs (40 GB VRAM) using the AdamW optimizer. Batch sizes were 512 without a decoder and 256 with a decoder, with learning rates scaled proportionally for smaller batches. A `ReduceLROnPlateau` scheduler was used to reduce the learning rate when the validation loss plateaued.

Hyperparameter search Learning rates and the contrastive temperature τ were optimized using `optuna`'s TPE sampler (Bergstra et al., 2011) on the PIPIII and MCPIII joints (six score types) from the 466 training patients. Search ranges were $\tau \in [0.1, 5]$, encoder LR $\in [10^{-7}, 10^{-2}]$, head LR $\in [10^{-5}, 10^{-2}]$, and weight decay $\in [10^{-8}, 10^{-1}]$. The search yielded $\tau = 1$ as temperature (prefactor to L_2 feature similarity), an encoder LR of $4 \cdot 10^{-4}$, a head LR of $4 \cdot 10^{-5}$, and a weight decay of 10^{-6} for a batch size of 512, with proportional LR scaling for smaller batches.

Abbreviation	**Meaning**
RA	Rheumatoid Arthritis
SvH / SvHS	Sharp–van der Heijde Score
ΔSvHS	Change in SvH score between visits (in chronological order)
ERO	Erosion
JSN	Joint Space Narrowing
ROI	Region of Interest
NS	Not Scoreable
ICC	Intraclass Correlation Coefficient
RMSE	Root-Mean-Squared Error
MAE	Mean Absolute Error
DAE	Denoising Autoencoder
RNC	Rank-N-Contrast (loss)
ChronoCon	Chronological Contrastive Loss (this work)
SCN	Spatial Configuration Network
TPE	Tree-Structured Parzen Estimator (Optuna)

Table 4: List of abbreviations beyond used throughout the manuscript and appendix.

Abbreviation	Description
Hand joints	
PIPII–PIPV	Proximal interphalangeal joints II–V (JSN)
PIPIIED/EP ... VED/EP	PIP joints II–V, distal/proximal erosion
MCPI–MCPV	Metacarpophalangeal joints I–V (JSN)
MCPIED/EP ... VED/EP	MCP joints I–V, distal/proximal erosion
IPIED / IPIEP	Thumb IP joint distal/proximal erosion
Rad_Carp	Radiocarpal joint (JSN)
RadiusE, UlnaE	Radius / Ulna erosion
LunatE, ScaphE, TrapE	Carpal bone erosions (lunate, scaphoid, trapezium)
Sca_Cap, Tra_Sca	Carpal articulation JSN
Base_MCIE	Base of metacarpal I (erosion)
Foot joints	
MTPI–MTPV	Metatarsophalangeal joints I–V (JSN)
MTPIED/EP ... VED/EP	MTP joints I–V, distal/proximal erosion
IP	Hallux interphalangeal joint (JSN)
IPED / IPEP	Hallux interphalangeal distal/proximal erosion
`<joint>_ED`	Distal joint part (erosion)
`<joint>_EP`	Proximal joint part (erosion)

Table 5: Grouped joint and score abbreviations used in this work contributing to Sharp-van der Heijde score . Erosion applies separately to proximal (EP) and distal (ED) joint parts.

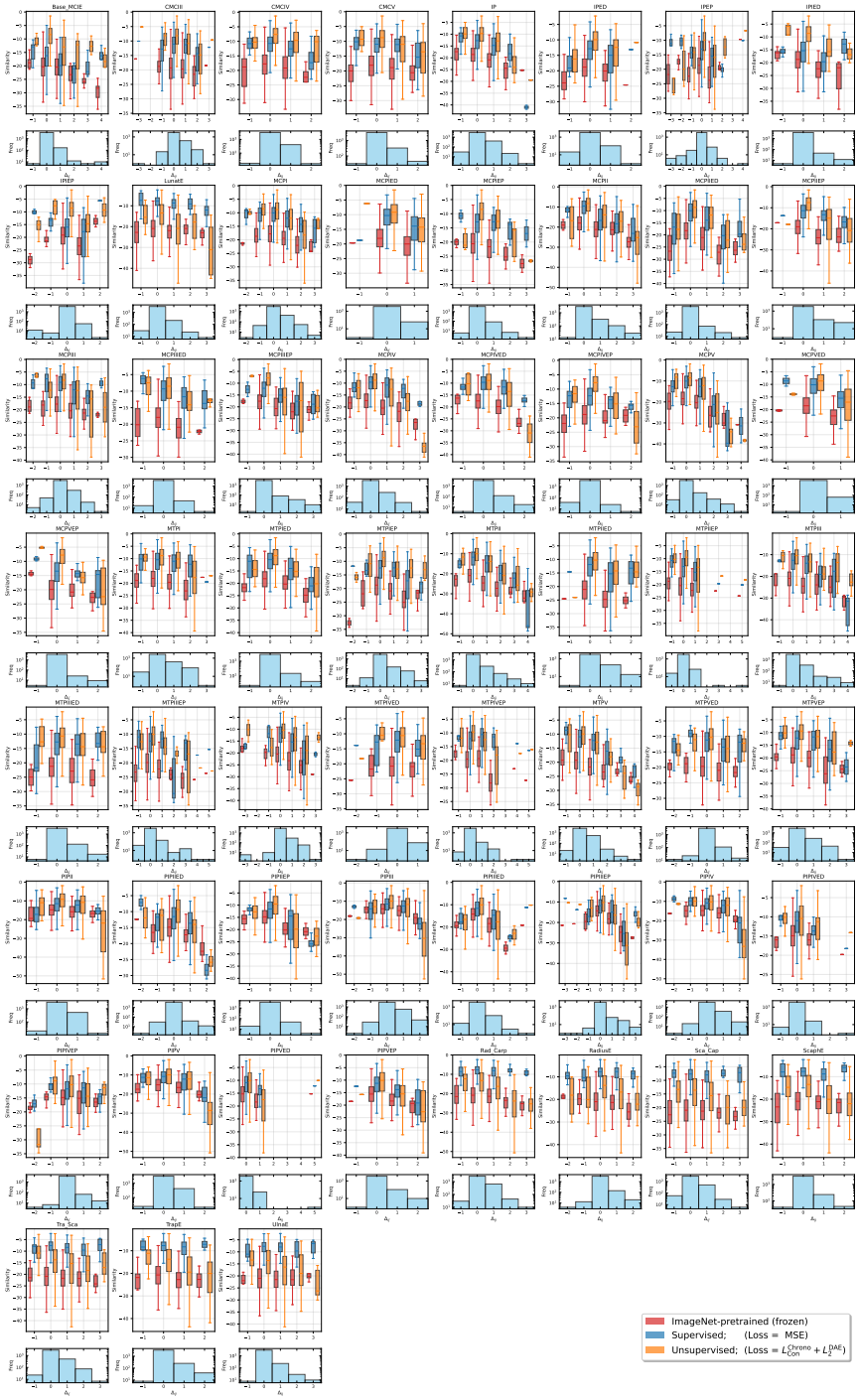

Figure 6: Feature similarity ($-L_2(\mathbf{v}_i, \mathbf{v}_j)$) between different chronologically ordered visits compared to the difference in score (ground truths).