Purpose: To develop a quality assurance method for deformable image registration (DIR) of thoracic CT images and to evaluate how this algorithm generalizes to diverse datasets and registration algorithms.

Methods: We previously developed a three-dimensional neural network to infer registration accuracy based on registered images labelled by expert identified point landmarks and evaluated performance across two different datasets, the publicly available Dirlab dataset (Dirlab) and an in-house database of longitudinal 4DCT scans (Long4DCT) and two DIR algorithms, elastix b-spline (B-spline) and DRAMMS. In this study, we evaluated and improved algorithm robustness for different dataset and DIR combinations. Different combinations of training data were evaluated on held-out testing datasets: 1) Dirlab registered by B-spline (Dirlab / B-spline); 2) Dirlab registered by B-spline and DRAMMS (Dirlab / B-spline / DRAMMS; 3) Dirlab and Long4DCT registered by B-spline (Dirlab / Long4DCT / B-spline).

Results: 1) The Area Under Curve (AUC) for training Dirlab / B-spline tested on held out Dirlab / B-spline was 0.99, on Long4DCT / B-spline was 0.94, on Dirlab / DRAMMS was 0.96, and on Long4DCT / DRAMMS was 0.94. 2) Training on multiple registration algorithms Dirlab / B-spline / DRAMMS gave AUC 0.99 for Dirlab / B-spline, 0.94 for Long4DCT / B-spline, 0.99 for Dirlab / DRAMMS, and 0.93 for Long4DCT / DRAMMS. 3) By training on Dirlab / Long4DCT / B-spline, the AUC for testing on Dirlab / B-spline was 0.99, for Long4DCT / B-spline was 0.97, for Dirlab / DRAMMS was 0.97, and for Long4DCT / DRAMMS was 0.96.

Conclusion: Our registration QA algorithm showed reliable inference ability across various datasets and DIR, however, training with diverse datasets provided slightly better performance than by adding DIR algorithms. Enriching training data by different dataset combinations obtained some improvement, demonstrating that our model's robustness has the potential for improvement.