Purpose: Accuracy of respiratory surrogate models of tumor displacement is influenced by selection of appropriate training data, the simultaneous tumor localizations and respiratory surrogate measurements used to create the model. The object of this study was to explore ways in which (1) varying timing and (2) selectively discarding samples of training data can improve surrogate model accuracy.

Methods: Motion from 125 lung, 10 liver, and 47 pancreas SBRT treatment fractions from 92 patients was analyzed. Each treatment fraction dataset included radiographically measured tumor positions, positions of three surrogate markers affixed to the torso, and tumor positions predicted by a commercial tumor tracking system from marker measurements. Partial-least-squares regression models of tumor position were trained on samples from 26 Hz position data. Low-frequency (0.05-0.10 Hz) and high-frequency (26 Hz) training datasets 5-150 seconds in duration were created. The effect of discarding samples with the highest leverage (as measured by Hotelling's T² statistic) was evaluated. Tumor localization errors were measured over 20 minutes.

Results: The baseline tumor localization error (mean +/- standard deviation) in training data was 1.9 +/- 1.7 mm. Mean errors for high-frequency training data were 4.7-4.9 mm and did not differ significantly with number (1-2), timing, or duration (5-30 sec) of acquisitions. Mean +/- standard deviation errors for low-frequency training data varied from 4.4 +/- 2.7 mm (0.10 Hz, 2.5-min acquisition) to 6.6 +/- 16.1 mm (0.05 Hz, 45-sec acquisition). Removal of high-leverage samples reduced tumor localization errors for high-frequency-derived models by 4-7% but did not reduce errors for low-frequency data.

Conclusions: Lengthening training data acquisitions improves models derived from low-frequency but not high-frequency training data. Conversely, removing high-leverage training samples benefits high-frequency data only. Training sample selection involves interplay between sampling rate, acquisition duration, and number of training samples.

Funding Support, Disclosures, and Conflict of Interest:

This work was supported in part by grant # CA124766 from the NIH/NCI and by the Achievement Rewards for College Scientists (ARCS) scholarship.