Purpose:
To quantitatively evaluate automatic anatomy segmentation for radiotherapy using a statistical model that can account for the observer variations and to provide a graphical visualization of the segmentation performance.

Methods:
We proposed to use the binary statistical behavior of voxel overlap to evaluate the performance of segmentation. A statistical model parameterized on a Beta distribution was built upon the observations of the volume overlap between the automatic segmentation and the manual reference contours. A statistical performance profile (SPP) was estimated from the model using the generalized maximum likelihood approach. The SPP defined the probability density function characterizing the distribution of performance values. This method was applied to evaluating auto-contour propagation for 10 head-and-neck cancer patients with a total of 122 daily CT scans. One physician contoured the parotid glands on each daily CT from scratch and modified from deformed contours automatically propagated from the planning CT using a deformable image registration method.

Results:
The SPP curves characterizing the overlap between deformed contours and manual contours for both parotids showed a mean performance of 0.71 with standard deviation of 0.07. They also showed that the probability of performance value above 0.90 is almost 0. This demonstrated the high intra-observer variations in contouring head-and-neck patients. The SPP curves for the modified contours showed a high performance for both parotids, with mean value around 0.98 and standard deviation of 0.05. The probability of performance value above 0.99 was around 80%. This indicated that the auto-propagated contours matched well with the physician's judgment. On the other hand, it confirmed that the intra-observer variations in the manual reference contours had a great impact on the evaluation.

Conclusions:
We developed a statistical modeling method for evaluating automatic segmentation. The method can quantify the impact of observer variations and graphically visualize the segmentation performance.