Radiographic Annotation Accuracy – The "Good Doctor" Performance Competing with AI

Yuan Chai, A. Mounir Boudali, John Farey, William L. Walter

Sydney Musculoskeletal Health, Kolling Institute, Faculty of Medicine and Health,
University of Sydney, St. Leonards, NSW 2064, Australia

Introduction

Personalized surgeries are planned from patient-specific anatomical landmarks. Due to differences in anatomies and radiographic qualities, the regions of landmark annotations cannot be quantified. In contrast, recent advances in machine learning landmarking techniques present its accuracy in heatmap format representing the landmark size at different confidence levels. However, the measurement accuracies can only be analyzed at a parameter dimension regardless of the heatmap information. There is a demand for quantifying landmark sizes under clinical practice to be comparable with AI.

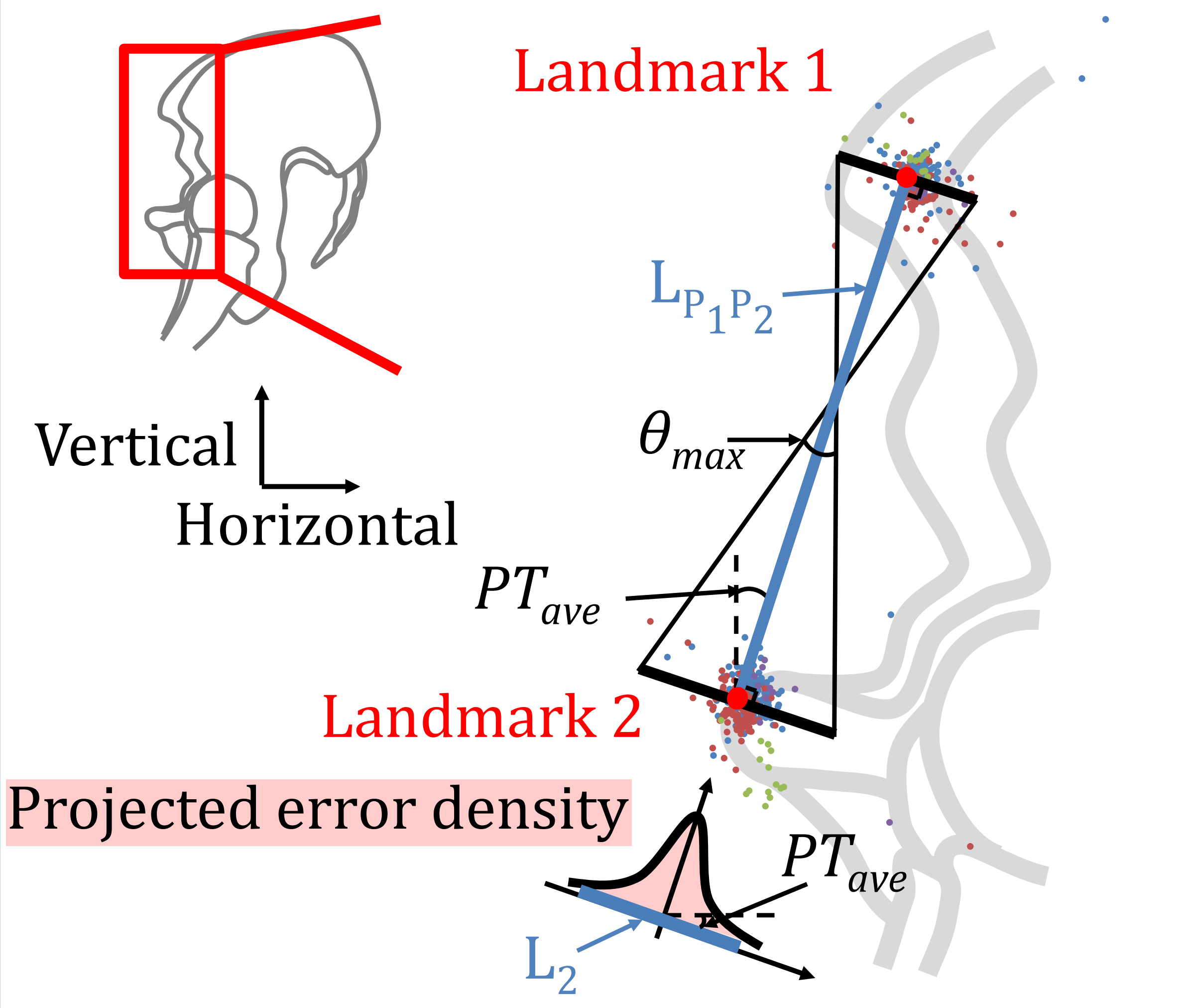

Figure 1. Diagram of calculating the coordinate of each landmark annotation.

Methods

Measuring pelvic tilt (PT) as an example, this study recruited 115 sagittal pelvic radiographs for the measurement of two PT definitions. We proposed a method to unify the scale of images that allows horizontal comparisons of landmarks and calculated the maximum possible error using a density vector (Fig. 1). Traditional descriptive statistics were also applied

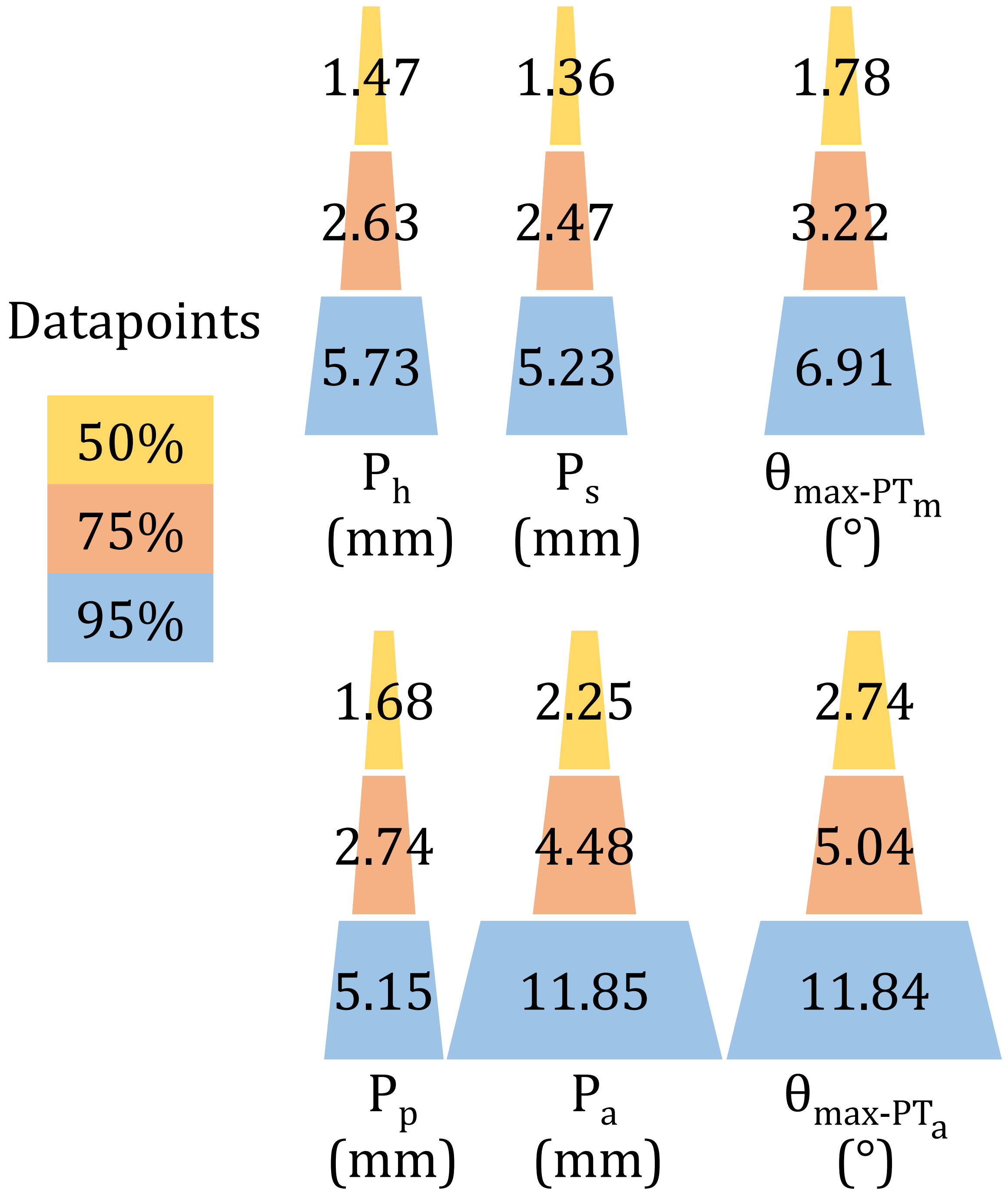

Figure 2. The cloud diameter of each landmark (calculation method excluding wrong landmarks) and its parameter-wise maximum impact at 50%, 75%, and 95% data points.

Results

Our result shows that all the measurements had excellent reliabilities (intraclass correlation coefficients > 0.9), yet 84 landmarks (6.09%) were identified as wrong from the secondary review. The landmark point clouds present landmark sizes of different annotation strategies at different probability levels (Fig. 2 and 3), which are comparable to the machine learning outcomes. The outcome also presents the maximum impact on corresponding parameters, the landmarks' regional shapes, and observer preferences. With 95% data points, the clinical reference of measuring pelvic tilt can reach a maximum disagreement of 6.9° - 11.8° (Fig. 2).

Discussion and Conclusion

The landmarks with excellent reliability still have a chance (at least 6.09% in our case) of making wrong landmark decisions. Identifying skeletal contours is at least 24.64% more accurate than estimating landmark locations (Fig. 3, $\widehat{\text{hat}}$ distributions are estimated). The landmark at a clear skeletal contour is more likely to generate systematic errors. Due to landmark ambiguity, a very careful surgeon measuring PT could make a maximum 11.8° random difference in 95% of cases, serving as a "good doctor benchmark" to qualify good landmarking techniques.

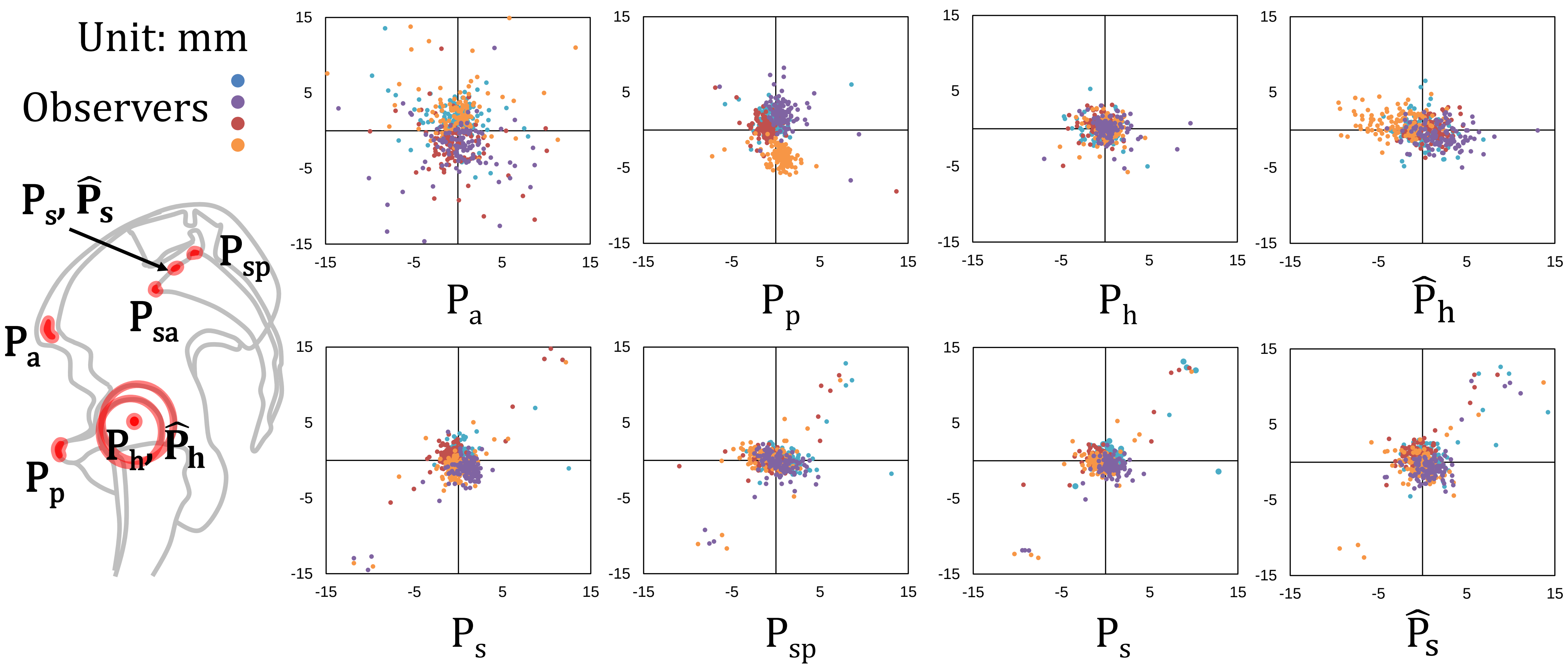

Figure 3. Scaled data point distribution of each landmark.

Contact

Yuan Chai PhD.
Sydney Musculoskeletal Health, Kolling Institute
Faculty of Medicine and Health, University of Sydney
St. Leonards, NSW 2064, Australia
E-mail: yuan.chai@sydney.edu.au