Machine Learning Has a Higher Accuracy Than Clinicians in Classifying Hip Fractures

Beate Ehrhardt PhD; Ellen Murphy PhD; Celia L Gregson MBBS; Otto Von Arx MBChB; Michael R Whitehouse PhD; ***Tim Chesser FRCS****; Harinderjit Gill PhD*
University of Bath, Bath, United Kingdom

Purpose: Hip fractures are common injuries in frail patients, are associated with high morbidity and mortality, and carry a considerable burden to any health-care system. Rapid accurate diagnosis, treatment and rehabilitation has shown to improve care for this patient group. There is often controversy between clinicians with radiographic classifications, and any improvement in classification and hence treatment may have significant benefits. The aim of the study was to create a machine learning-based method to automatically classify hip fractures using radiographs, with the aim of standardizing classification.

Methods: The machine learning classification involved 2 stages, first to automatically detect the hip joint(s) in a radiograph, and then classify the fracture given this selected region. A fully convolutional network was trained to automatically locate the hip joint, using 905 manually annotated radiograph images split 60:20:20 into training, validation, and test sets, respectively. Next, transfer learning of a pre-trained convolutional neural net (CNN) was used to automatically classify the type of fracture, utilizing a total data set of 638 non-fractured and 2365 fractured AP radiographs of hips. This was divided as above into training, validation, and test sets. The training data set was increased to a total of 21,856 by flipping, inverting, and random shifts. Each fractured hip was classified by at least 2 experts, with 2 further reviewers deciding the classification when there was disagreement. The classifications were no, trochanteric, and intracapsular fracture.

Results: Locating the hip was 99% accurate. For the fracture classificatio,n human experts agreed on only 59% of cases, while the machine learning algorithm had an overall accuracy of 92%. The accuracy varied between classes—94% for no fracture, 91% for intertrochanteric, and 89% for intracapsular.

Conclusion: The machine learning method was more accurate in classifying hip fractures than human observers. This has potential for providing a more accurate and standardized method for hip fracture classification. This suggests greater use of artificial intelligence would benefit patients.