Supplemental Appendix 1 provides methodological details about identifying eligible orthodontists, collecting data, and assessing intra-rater reliability.

Process for Identifying Eligible Orthodontists

In brief, orthodontists selected from the membership list were excluded from further data collection if they could not be found on the AAO's online directory. For orthodontists in the online directory, a Google search was performed to find all office addresses listed in the member's online directory entry. For office addresses that could not be found using this approach, a more in-depth search was conducted using the word "orthodontist" and the orthodontist's name, in an attempt to find any offices. If none of the offices under the searched orthodontist's name could be found online, the orthodontist was excluded from further investigation for having no online presence. Office addresses that were identified using Google, were further investigated (e.g., visit the office website, the Google maps office description, the Yelp address description) to determine whether the office was limited to orthodontics. An orthodontist was excluded from further investigation if all of that orthodontist's offices were: closed, located outside of Canada and/or the USA, previously included under another orthodontist already in the study, or if the practice was not in an orthodontics-only office setting (e.g., general dentistry / pediatric dentistry / periodontics / university / hospital-based offices).

Data Collection

Data were collected about each orthodontist and the offices listed in that orthodontist's membership directory entry. The office addresses and graduation year of each orthodontist were recorded. If the orthodontist had offices that were limited to orthodontics, then further information was recorded for each of those offices separately for both Google and Yelp reviews. For Google and Yelp, the collected data were: presence of reviews (yes or no), presence of negative reviews (yes or no), date of website visit, website link, overall rating, number of 1-star, 2-star, 3-star, 4-star, 5-star reviews, number of reviews without textual content, presence of an online response to a negative review (yes or no), negative review updates (yes or no), negative review updated to a positive review (yes or no). For Google, the number of reviews without textual content was recorded. For Yelp, whether the office had been claimed (yes or no) was also recorded. All negative (i.e., 1-star and 2-star) Google and Yelp textual reviews, as well as office responses to negative reviews, were saved verbatim for analysis.

Negative Review Content Analysis

All negative Google and Yelp reviews, and any corresponding office responses, were scored for content using a scoring template. The scoring template consisted of a series of codes that describe the type of problem being reported, who was described as being responsible for the problem, and when during the different stages of orthodontic care the problem occurred. Also, if an office posted a response to a negative review, that response was scored for content. If an identical negative review about an office was posted on both Yelp and Google websites, that response was scored only once for content.

The initial draft of the scoring template was created based on common themes reported in previous research on orthodontic patient satisfaction. Then, a pilot study was conducted to evaluate and refine that scoring template. Three examiners (AMS, RK, DSR) independently scored batches of 10-15 negative online reviews using the draft template. The raters would then meet to discuss each review and any coding disagreements. The scoring template was revised after each meeting to better reflect the content of the reviews and improve the reliability of the raters. This process continued until the template was no longer being refined, and the three examiners were scoring similarly. Sixty randomly selected reviews were evaluated during this preliminary phase.

During the study, two examiners independently scored each of the 956 negative online reviews. One examiner (AMS) scored all reviews, while the two other examiners (DSR, RK) each scored half (478) of the reviews. After all reviews were scored, disagreements in coding were identified and two examiners discussed and resolved the disagreement to arrive at a final code.

Intra-Rater Reliability

The reliability of the scoring procedure was assessed by having two examiners (AMS, DSR) independently re-score the same subset of 50 randomly selected reviews from the 956 negative reviews. The two reviewers then met to discuss and resolve disagreements to arrive at a final code. The reliability of scoring was calculated by comparing the codes assigned to those 50 re-scored reviews with the codes assigned several months previously using the identical scoring procedure. Reliability was calculated as the average of percentage agreement of codes regarding what the complaint was about, as well as the average of percentage agreement regarding who was considered the cause of the complaint. Re-scoring what the complaints were about in a negative review yielded 95.7% agreement, while re-scoring who was responsible for the problems was less reliable with 88.8% agreement. These high levels of reliability demonstrate consistency in how the scoring of reviews was conducted.