Preventing Youth Soccer, NCT 04266925 - Study Protocol and Statistical Analysis Plan

Last updated to add a cover page: 5/8/20

Last update for clarity of wording: 5/1/20

Last update for substantive material: 1/31/20

Study Methods

We conducted a randomized cross-over trial with boys' and girls' youth soccer games randomized to have one vs. three referees, and pairs of teams serving as the unit of analysis. Player-to-player collisions, fouls and player injuries were assessed through behavioral coding of recorded games to evaluate whether additional referees reduced the risk of child injury. The research protocol was approved by the Institutional Review Board at the University of Alabama at Birmingham. Referees provided informed consent to participate. Since we were videotaping public events in public locations, the consent of players, coaches, and others captured on the videotapes was exempted. All videotapes were stored as highly confidential research data. We monitored for study-related adverse events and noted none.

Randomization of games took place in advance so that either one or three referees could be scheduled to work. Each game was pseudo-randomly assigned by a referee scheduling group to have one or three referees, and then matched pairs were identified to determine sequence approximating a 1:1 allocation ratio (half the pairs had 1 referee first and half had 3 referees first). Sequence was therefore concealed prior to assignment and the referee scheduling group, who were responsible for assigning referees to each game and worked independently from the coaches, players and researchers, applied the randomized number of referees. The referee scheduling group also did not serve as referees for any of the matches. Masking of condition to any party was not possible given the nature of the research design and goals.

Youth Soccer League and Game Settings. Participating teams were part of a consortium of elite youth soccer clubs in the greater Birmingham, Alabama metropolitan area and extending to nearby cities like Huntsville and Tuscaloosa. The clubs served a racially, geographically, and socioeconomically diverse group of communities and families and were based in rural, suburban, and urban areas. Both boys and girls teams participated, and we focused our research on the U10 (under age 10) and U11 (under age 11) leagues, which were the oldest groups of teams previously playing matches officiated by a single referee. Those teams were populated with players almost entirely ages 9-10, with a few 7- and 8-year-olds, generally highly-skilled players, "playing up" with teams serving primarily older children. All teams serving those age groups were eligible for participation; no cultural, language, health or other exclusion criteria applied.

Games were played throughout the metropolitan area, with teams traveling as far as about 100 miles (160 km) for matches against clubs in neighboring cities. In many cases, teams played a few matches on the same day or weekend, as part of a round-robin tournament. The crossover design study was originally planned using PASS (NCSS Statistical Software, Kaysville, UT) to detect differences in outcomes of at least 0.27 SD assuming 150 pairs of games (300 total games) with 90% power, and assuming a two-sided alpha level of 0.05. The methodology applied accounts for the paired nature of the study design, which addresses within-participant variability. Because changes to the league occurred between the planning stage of the study and the implementation, and due to logistical challenges in data collection, we achieved a smaller sample size than planned. In total, we included 49 pairs of games in the study (98 total games), which under the same assumptions offered power to detect an effect size of at least 0.47 SD.

The recorded 98 games involved 13 clubs and included 38 games for U10 boys, 28 for U10 girls, 18 for U11 boys, and 14 for U11 girls. All games were paired, such that the same clubs played each other, once with 1 referee on the field and the other time with 3 referees on the field. All games occurred during the same Fall 2017 season, with the first games occurring on September 9, 2017 and the last on November 11, 2017. Games were scheduled to last 50 minutes ($M = 53.1$, $SD = 6.5$), with a few games shortened by weather conditions or lengthened to

account for injury-related stoppage. Climate during the games ranged widely, with hot and humid weather typical at the start of the season and crisp, cooler weather at the end (M temperature = 80.9º F, SD = 5.7º). Light rain was present at two games.

We recorded all games from a location in the center of the field sideline using a digital video camera with a wide-angle lens placed on a tripod. Undergraduate students, graduate students, and paid research staff were trained to follow the action of the ball and record both sound and video. Elevated angles were used when logistically feasible. Appropriate gear was present to handle recording in the rain. All games scheduled to be recorded were recorded successfully, but technical issues (e.g., dead battery, full memory card) led to small portions of games not recorded on rare occasions.

Referees. Referees were hired by the league. All had appropriate training and certification to serve as youth soccer referees. The full panel of 72 referees used by the league included 61 men (85%) and 11 women (15%) with a mean age of 27.2 years (SD = 15.9; range = 14.1-71.2 years). They were 65% Caucasian, 18% African-American, 13% Hispanic, and 3% Asian American; one referee declined to provide their race/ethnicity. Most of these referees were present at one or more of the games included in this research, and all provided informed consent via a very short online questionnaire that included consenting processes and a brief demographic survey. Researchers were available to answer questions about the research and consenting process by email or telephone, or in person during recorded soccer matches. All referees serving the league were eligible for inclusion in the study; no exclusion criteria applied.

Logistics of Soccer Refereeing. Traditionally, professional soccer matches have 3 referees on the field, with the "center" referee responsible for calling all contact fouls while two assistant referees monitor sidelines to regulate out-of-bounds calls and offsides. In the mid-2010s, FIFA, the international soccer monitoring body, recognized that assistant referees on the sidelines may sometimes have a better angle to view fouls than the center referee and therefore at the 2014 World Cup, side referees were granted the right and responsibility to call contact penalties also. This strategy for refereeing elite soccer matches has begun to be adopted more broadly, and is the strategy we implemented for this research when three referees were present. The three referees positioned themselves in "traditional" locations (center referee covering a diagonal pathway across the field and side referees monitoring opposite sidelines), but all three were granted the right and responsibility to call all fouls. They were trained and familiar with this strategy already, as it was used for older children in this and other local community and interscholastic soccer leagues. For games with just one referee, that referee was responsible for monitoring the full field and calling all fouls and sideline violations.

Videotape Coding. Following recording of all games, coding of the videotapes proceeded in three steps. First, as detailed below and following recommendations to code behavioral pediatric psychology data by Chorney and colleagues (2015), objective written criteria were developed, largely through refinement of existing criteria used by our laboratory in previous research (Schwebel, McDaniel, et al., 2006). Joint review of videotapes by the coding team, lab manager and principal investigator was incorporated into the process of developing those criteria. Second, two coders independently reviewed a randomly-selected 15% of games. Inter-rater reliability was established through that review on each variable, with coding agreement of 95% or higher for every categorical variable, $\kappa > .70$ on categorical variables, and intraclass correlation $r \geq .80$ on continuous variables. Intraclass correlations between coders for the four primary outcome measures of collisions, aggressive fouls called by the referees, aggressive fouls detected by the coders, and injuries, were .80, 89, .95, and .90, respectively, all above the

recommended correlation of .75 or above for acceptable agreement (Cicchetti, 1994; Hallgren, 2012). Following establishment of inter-rater reliability, disagreements between coders were resolved through joint tape review and consensus agreement. At that point, the third step of coding proceeded, review of the remaining tapes by a single researcher. Coding was conducted by a mix of doctoral students and paid research assistants, all of whom held a bachelor's degree in psychology, public health, or a related field and many of whom held master's degrees.

Measures. The following four outcome measures were retrieved through videotape review:

(a) Collision, defined as contact between two or more players that resulted in one or more players showing visible signs of pain, falling down as a result of the contact, or experiencing an injury that required adult attention. A collision also was recorded when two or more players had contact as a result of a foul, as called by the referee or judged by the coder. Finally, a collision occurred when there was forceful contact between players because one player was trying to gain a better position or get to the ball.

(b) Aggressive foul (referee), defined as a foul whistled by a referee on the field, live, during the game, that was aggressive in that it involved player-to-player contact. Examples include pushing, tripping, and other aggressive acts. Handballs, offside calls, and illegal throw-ins were not included.

(c) Aggressive foul (coder), defined as a foul identified by our researcher who was coding videotaped games. As in the aggressive referee fouls, only fouls that involved player-to-player contact were included. Thus, the two outcomes of aggressive fouls called by the referee and aggressive fouls coded by the research team were designed to count the same behaviors. The research team benefited, of course, from the opportunity to use replay, slow motion and zooming on videotapes, so their counts were generally higher than the referee-called fouls.

(d) Injury, defined as any sort of pain or tissue damage that was serious enough that an adult attended to the player on the field and/or the player left the game. We further coded injuries into three categories: (a) those occurring as a result of a collision or contact with another player, (b) those occurring as a result of contact with the ball, such as when the ball is kicked and hits a player hard in the stomach, and (c) those occurring as a result of no contact, such as a strained muscle or a cramp. Given the goals of this study, only injuries occurring in the first group – those resulting from a collision or contact with another player – were considered in data analyses. Injuries in this category comprised 80.4% of all injuries that were coded.

Statistical Analysis Plan. All outcome measures were summarized overall, by number of referees, and by gender/age group. To test for differences in each outcome measure count between games with 3 referees and games with 1 referee, separate Poisson mixed models were fitted, allowing for the incorporation of the crossover design. Specifically, correlations within game pairs were addressed with a pair-specific intercept. Offsets were utilized to account for individual recorded game length. This methodology yields rate ratios (RR) to compare the ratio of event rates between 3-referee games and 1-referee games. Given the commonality of competing in youth soccer matches among our sample, the carryover effect from one game to the next should be minimal. Carryover was assessed statistically with multiplicative interaction between group sequence and treatment (Brown, 1980; Willan & Pater, 1986). The level of

significance for all analyses was set at 0.05. Statistical analyses were conducted in SAS 9.4 (SAS Institute, Cary, NC).