Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Qualitative publications passing screening, excluded because quantitative data available on the same outcome


Layer EH, Beckham SW, Mgeni, L, Shembilu C, Momburi RB, Kennedy CE. "After my husband's circumcision, I know that I am safe from diseases": women's attitudes and risk perceptions towards male circumcision in Iringa, Tanzania. PLoS ONE 2013; 8(8), e74391.


**Detailed Methods Supplement:**

*Determining consistency:* Consistency of evidence for each outcome was determined on the following basis, using the data points included in Figure 2. Abstracted outcomes were highly consistent if they met one of three criteria: if data points from two or more RCTs were present, all agreed on the significance of effect, and the direction if significant. Otherwise, if there were data points from ≥4 publications, then ≥75% of point estimates agreed on direction of effect, including 2/3 of any RCTs; or all point estimates agree on direction of effect if data points from 3 publications were present. Among outcomes for which evidence was not highly consistent, those with data points from ≥3 publications could be considered for medium consistency. In order to have medium consistency evidence, outcomes must have had at least 67% of point estimates agree on the direction of effect, including the 2 with the highest quality scores (as a percentage of total potential points), the one with the largest sample size, and any RCTs (or 67% of RCTs if there are ≥3). Outcomes for which there were three or more data points but they did not fulfill these criteria were considered to have low-consistency evidence. Those for which there were fewer than three data points were considered to have indeterminate-consistency evidence.

*Designation of study design type:* In most cases this was the study design reported in the publication. However, studies involving secondary analyses of participant characteristics from RCTs which randomized participants on another variable than circumcision were considered cohort or cross-sectional studies with respect to circumcision, rather than RCTs. Similarly, follow-up publications to RCTs of circumcision which collected pre-post data in the circumcision group rather than making comparisons to the control group were considered pre-post studies.

*Identification of duplicate data points:* The goal of the supplementary tables was to make available results on all groups and subgroups for which they were available. For that reason, data points which were otherwise duplicates were not considered duplicated if they contained reports on different subgroups, or a main group and its subgroups, and therefore different results in the tables should not be considered to all necessarily reflect independent data.

Outcome data points were considered duplicates of each other if they were the same outcome or suboutcome, reported in the same group or subgroup; and the publications in which they were found derived data was from the same geographic location, and had inclusion criteria such that the same participants could reasonably have been included in both. In most cases this resulted from the publication of multiple publications from the same study or series of studies.

Once identified, duplicates were then potentially excluded from the outcome-specific tabs in Appendix 1, based on the below criteria. Where nonrandomized longitudinal studies reported on associations of both baseline and incident outcomes with male circumcision, the point estimates for incident outcomes were included in the figure.

*Hierarchy of duplicate data points:* If duplicate data points were found, the following rules were used to determine which was retained in the main supplementary table; the others were moved to the supplementary table of excluded outcomes (supplementary table S3). If the data points reported different data collection time periods, then if those time periods were nested (e.g. 2000-2010 and 2005-2008), the source that captured the widest time range was retained. Peer-reviewed manuscripts were retained over non-peer-reviewed results (abstract, dissertation, white paper). Where both of a set of duplicate data points were peer-reviewed, the one with the later publication date was retained.

*Figure 2:* The goal of the figure was to intuitively display the distribution of the evidence for each outcome. For that purpose it was necessary that each set of data appear only once. For each outcome, only the data point for
the main participant group (not subgroups) from each publication was chosen. When a publication provided no main group estimate, we picked the group most resembling the general population (e.g. HIV- women) rather than inflating the apparent weight of that publication by showing multiple subgroups. This occurred in four cases, all reporting on the HIV outcome: Wawer et al.\(^\text{19}\) where the subgroup with delayed post-MC resumptions of sex was chosen over the subgroup with early resumption; Babalola et al.\(^\text{50}\) where countries with high prevalence of male circumcision rather than those with medium prevalence were chosen for comparison to those with low prevalence; Cuadros et al.\(^\text{58}\) where the more recent dataset was chosen over the older one for comparison of areas with low MC coverage to those with high MC coverage; and Chemtob et al.\(^\text{60}\) where the subgroup with the most recent data was chosen, comparing Israel to the Netherlands rather than to France.

If no main group result was provided and no subgroup was more representative of the general population, but subgroup results were consistent in direction, the less extreme result is displayed. If in this situation no subgroup was more representative of the general population, and the direction of results was consistent between subgroups, the least extreme result was displayed. Conversely, if subgroup results were inconsistent in direction (N=2), the publication was excluded from the graph as being impossible to capture in one data point. In situations where a publication reported results over two or more non-overlapping time periods, we chose to represent the most recent time period in the figure to favor reflecting the effects of VMMC programs rather than traditional circumcision practices. Where a publication reported both a prevalence ratio and unadjusted incidences for the main group (N=1), we reported the incidence rather than prevalence ratio. If direction of effect was not reported (N=2), the result could not be shown.

Confidence intervals not provided in publications were calculated where possible. Because adjusted estimates were preferentially abstracted, confidence intervals were not always calculable in these cases even where numbers of events and participants were provided.

Quality grading: For cohort studies, the level of loss to follow-up deemed ‘acceptable’ was 20% or lower. All domains were considered “key” in assigning final quality levels to RCTs using the GRADE criteria.

Other: For the one included abstract from which the data was subsequently published in a manuscript with refined analyses\(^\text{70}\) (Pintye et al, *Trichomonas vaginalis*), numerical data in the table and figure represent the updated calculations in the manuscript. (Pintye J, Drake AL, Unger JA, et al. Male Partner Circumcision associated with lower *Trichomonas vaginalis* incidence among pregnant and postpartum Kenyan women: a prospective cohort study. Sex Transm Infect 2017;93:137–143.)