Supplement 3: Certainty of Estimates using GRADE Criteria

Challenges of Applying GRADE to Water Immersion

When using the GRADE Criteria for water immersion, the certainty of the evidence for all outcomes begins at the level of "low" because most water immersion research is conducted as prospective observational studies. GRADE scores observational studies as less certain than randomized controlled trials. Unfortunately, randomized controlled trials of water immersion do not automatically reduce bias because of the nature of the intervention. Blinding of the care provider and participants is not possible and there is no control that can act as a placebo. This increases the risks for performance bias, detection bias, and reporting bias. Uneven attrition is expected as women randomized to water have many legitimate reasons for exiting the water, such as to use the bathroom or to facilitate fetal monitoring. In contrast, women randomized to standard care are unlikely to be asked to enter the water. This attrition bias causes challenges with intention to treat analyses, especially for outcomes that are only relevant if the birth occurs in water. A further challenge occurs in recruiting a sample willing to be randomized. Women who desire water immersion are less willing to be randomized. This selection bias produces a sample that does not represent the population that chooses water immersion for pain control. Given these limitations, randomized controlled trials reduce as much bias as a well-controlled prospective study.

The GRADE criteria assume a study is assessing the superiority of one intervention over another. However, most water immersion studies are interested in equivalency of outcomes. GRADE criteria allow upgrading for large magnitude of effect, but this is not possible when the purpose of a study is to demonstrate no increased risk of poor outcomes. GRADE criteria also allow upgrading for demonstration of a dose-effect. However there is no dose of water immersion; instead women enter and leave the pool at will and the length of immersion is determined by the length of labor. This leaves only one category of upgrading available to studies of water immersion – plausible confounding.

Understanding the limitations of applying the GRADE criteria to water immersion, we recommend readers interpret the results of the GRADE assessment with caution. A GRADE of "low" certainty for water immersion does not necessarily indicate a need for more research. We point to the example of postpartum hemorrhage. Thirteen studies reporting on 63,891 participants have been synthesized to demonstrate there is no increased risk of postpartum hemorrhage with water immersion. Grade assessment indicates the level of certainty is low, but fail-safe analysis indicated an additional 198 studies are needed to change the results to no difference. Fail-safe N is only calculated when the result favors water immersion or the standard care, so these comparisons are not available for outcomes reporting no difference.

Description of Assessment Criteria

Risk of Bias in individual studies are provided in the forest plots for each outcome. Grade criteria reduce certainty of an estimate if an outcome had serious limitations likely to result in a biased estimate, including accounting for the weight of each study to the final estimate.

Inconsistency of estimates between studies was expected as part of this review, as the purpose was to identify reasons for heterogeneity. Because the eligibility criteria for this study reflect intentionally seeking papers in different settings, inconsistency is not a criteria to assess the certainty of the estimate.

Indirectness of the evidence reduces certainty when the population studied is not the population for the intended review. The study of water immersion is limited to women at low risk of birth complications, so this criterion does not affect the certainty of the evidence.

Imprecision of the estimate for a systematic review is generally measuring the ability of the evidence to find a statistically significant result, however one purpose of studies of water immersion is to demonstrate no increased risk of harm. For the purposes of GRADE assessment, certainty was downgraded for imprecision when the sample available for meta-analysis had less than 2000 participants.

Publication bias reduces certainty because it assumes studies with negative results are left unpublished. Prior studies have found publication bias that favors standard care over water immersion. This means the outcome is likely more favorable of water immersion than the estimate suggests and we can be more certain that water immersion is safe. To accommodate the standard Grade format, certainty of a result will be downgraded when the trim and fill test indicate the potential publication bias is enough to change the results.

Certainty of evidence is upgraded when the magnitude of effect is large, using standard risk ratios to define large and very large. For rare outcomes, such as those reported with water immersion, the OR becomes equivalent to the risk ratio, allowing this study to use the standard Grade Criteria for large effect (RR >2 or <0.5) and very large (RR >5 or <0.2) for most outcomes.

Certainty of evidence is upgraded when the evidence suggests a dose-effect. Water immersion does not have defined doses, instead women enter and exit the tub at will. In general, the length of immersion is determined by the length of labor.

Certainty of evidence is upgraded when controlling for potential sources of confounding are likely to result in a more favorable outcome for water immersion. For this table, studies are upgraded if the result from meta-regression was more favorable than the main analysis.

Supplement 4 Table 1: GRADE Criteria for interventions and outcomes with water immersion for labor and delivery.

			Reduce Grade					Increase Grade					
Outcome	**Studies**	**Sample Size**	**Risk of Bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Publication Bias**	**Magnitude**	**Dose-Effect**	**Plausible Confounding**	**Final Grade**	**Importance**	**Fail-safe N**
Induction	3	2,008	-	n.d.	-	-	-	-	n.d.	-	Low	Limited	-
Amniotomy	5	1,627	-	n.d.	-	↓	-	-	n.d.	-	Low	Limited	-
Augmentation	3	1,420	-	n.d.	-	↓	-	↑	n.d.	-	Low	Important	-
Fetal Monitoring	0	0	-	n.d.	-	-	-	-	n.d.	-	NONE	Limited	-
Opioid	8	27,391	-	n.d.	-	-	-	↑	n.d.	-	Moderate	Important	972
Epidural	7	10,993	-	n.d.	-	-	-	↑	n.d.	-	Moderate	Important	100
Pain	8	1,200	-	n.d.	-	↓	-	↑	n.d.	-	Low	Important	279
Cesarean	8	1,575	-	n.d.	-	↓	-	-	n.d.	-	Very Low	Critical	-
Shoulder Dystocia	7	53,367	-	n.d.	-	-	-	-	n.d.	-	Low	Critical	-
Intact Perineum	14	59,070	-	n.d.	-	-	-	-	n.d.	↑	Moderate	Limited	358
OASI	14	93,690	-	n.d.	-	-	-	-	n.d.	-	Low	Important	-
Episiotomy	13	36,498	-	n.d.	-	-	-	↑↑	n.d.	↑	Very High	Important	1525
Third Stage Management	0	0	-	n.d.	-	-	-	-	n.d.	-	NONE	Limited	-
Postpartum Hemorrhage	13	63,891	-	n.d.	-	-	-	-	n.d.	-	Low	Critical	198
Manual Removal of Placenta	5	2,893	-	n.d.	-	-	-	-	n.d.	-	Low	Critical	-
Maternal Infection	3	32,653	-	n.d.	-	-	-	-	n.d.	-	Low	Important	-
Satisfaction	6	4,144	-	n.d.	-	-	-	-	n.d.	-	Low	Important	133
APGAR	16	100,881	-	n.d.	-	-.	-	-	n.d.	↑	Moderate	Important	-
Neonatal Resuscitation	5	51,028	-	n.d.	-	-	-	-	n.d.	-	Low	Critical	-
Transient Tachypnea	2	1,473	-	n.d.	-	↓	-	-	n.d.	-	Very Low	Limited	-

Outcome	Studies	Sample Size	Reduce Grade					Increase Grade			Final Grade	Importance	Fail-safe N
			Risk of Bias	Inconsistency	Indirectness	Imprecision	Publication Bias	Magnitude	Dose-Effect	Plausible Confounding			
Respiratory Distress	3	32,707	-	n.d.	-	-	-	-	n.d.	-	Low	Critical	-
Neonatal Intensive Unit Admission	0	0	-	n.d.	-	-	-	-	n.d.	-	NONE	Critical	-
Neonatal Death	3	66,544	-	n.d.	-	-	-	-	n.d.	-	Low	Critical	-
Infection in Newborn Period	0	0	-	n.d.	-	-	-	-	n.d.	-	NONE	Important	-
Cord Avulsion	5	50,791	-	n.d.	-	-	-	-	n.d.	-	Low	Limited	5
Breastfeeding Initiation	2	692	-	n.d	-	↓	-	-	n.d.	-	Very Low	Important	-