

**Online Appendix A: Detailed results of quality assessment of quantitative studies**

Author, year, reference	Quality Rating Components																				Global Score						
	Selection Bias (A)			Sampling (B)				Confounders (C)			Blinding (D)			Data collection methods (E)			Withdrawals/dropouts (F)			Analysis (G)	A	B	C	D	E	F	GR
	1	2	R	1	2	3	R	1	2	R	1	2	R	1	2	R	1	2	R	1							
Alajbegovic et al, 2002 <sup>27</sup>	2	5	2	2	-	-	3	3	4	3	3	3	2	1	1	1	4	5	2	1	2	3	3	2	1	2	3
Barnes et al, 2004 <sup>28</sup>	2	1	2	2	-	-	3	3	4	3	3	3	2	3	3	3	4	5	2	1	2	3	3	2	3	2	3
Blight et al, 2008 <sup>29</sup>	1	2	2	1	1	1	1	3	4	3	3	3	2	1	1	1	4	2	2	1	2	1	3	2	1	2	2
Chaaya et al, 2006 <sup>31</sup>	1	1	1	1	1	1	1	3	4	3	3	3	2	1	1	1	4	1	1	1	1	1	3	2	1	1	2
Creson et al, 1996 <sup>32</sup>	2	5	2	2	-	-	3	3	4	3	3	3	2	3	3	3	4	5	2	1	2	3	3	2	3	2	3
Delic-Ovcina, 2011 <sup>33</sup>	4	5	3	2	-	-	3	3	4	3	3	3	2	1	1	1	4	5	2	1	3	3	3	2	1	2	3
Eytan et al, 2011 <sup>34</sup>	1	5	2	1	1	1	1	3	4	3	3	3	2	1	1	1	4	5	2	1	2	1	3	2	1	2	2
Farhood et al, 2010 <sup>35</sup>	2	1	2	2	-	-	3	2	4	3	3	3	2	1	1	1	4	1	1	1	2	3	3	2	1	1	3
Giuliani et al, 2010 <sup>36</sup>	1	1	1	1	1	1	1	3	4	3	3	3	2	1	1	1	4	1	1	1	1	1	3	2	1	1	2
Giuliani et al, 2012 <sup>37</sup>	1	1	1	1	1	1	1	3	4	3	3	3	2	1	1	1	4	1	1	1	1	1	3	2	1	1	2
Harel-Ftsch et al, 2010 <sup>39</sup>	1	5	2	1	1	1	1	3	4	3	3	3	2	1	1	1	4	5	2	1	2	1	3	2	1	2	2
Harris et al, 2012 <sup>41</sup>	1	5	2	2	-	-	3	3	4	3	3	3	2	3	3	3	4	5	2	1	2	3	3	2	3	2	3
Harris et al, 2011 <sup>40</sup>	4	5	3	2	-	-	3	3	4	3	3	3	2	3	3	3	4	5	2	1	3	3	3	2	3	2	3
Helweg-Larsen et al, 2008 <sup>42</sup>	4	5	3	2	-	-	3	3	4	3	3	3	2	3	3	3	4	5	2	1	3	3	3	2	3	2	3
Israelowitz et al, 2006 <sup>43</sup>	2	5	2	2	-	-	3	3	4	3	3	3	2	3	3	3	4	5	2	1	2	3	3	2	3	2	3
Keinan-Boker et al, 2011 <sup>44</sup>	1	3	3	2	-	-	3	3	4	3	3	3	2	3	3	3	4	3	3	1	3	3	3	2	3	3	3
Khader et al, 2009 <sup>45</sup>	1	1	1	1	1	1	1	3	4	3	3	3	2	1	1	1	4	1	1	1	1	1	3	2	1	1	2
Koupil et al, 2007 <sup>47</sup>	1	2	2	1	2	2	2	3	4	3	3	3	2	1	1	1	4	5	2	1	2	2	3	2	1	2	2

Leavey et al, 2004 <sup>48</sup>	2	2	2	2	-	-	3	3	4	3	3	3	2	1	1	1	4	2	2	1	2	3	3	2	1	2	3
Luitel et al, 2013 <sup>49</sup>	1	1	1	2	-	-	3	3	4	3	3	3	2	1	1	1	4	1	1	1	1	3	3	2	1	1	3
Maksimovic et al, 2011 <sup>51</sup>	2	1	2	2	-	-	3	3	4	3	3	3	2	1	1	1	4	1	1	1	2	3	3	2	1	1	3
McLeod et al, 2005 <sup>52</sup>	2	5	2	2	-	-	3	3	4	3	3	3	2	1	1	1	4	5	2	1	2	3	3	2	1	2	3
Morikawa, 2001 <sup>53</sup>	3	1	2	2	-	-	3	3	4	3	3	3	2	3	3	3	4	5	2	1	2	3	3	2	3	2	3
Mousa et al, 2010 <sup>54</sup>	2	5	2	2	-	-	3	3	4	3	3	3	2	1	1	1	4	5	2	1	2	3	3	2	1	2	3
Nguyen-Van-Tam et al, 1995 <sup>55</sup>	2	1	2	2	-	-	3	3	4	3	3	3	2	3	3	3	4	5	2	1	2	3	3	2	3	2	3
Okello et al, 2014 <sup>56</sup>	2	1	1	2	-	-	3	3	4	3	3	3	2	1	1	1	4	5	2	1	2	3	3	2	1	2	3
Roberts et al, 2013 <sup>57</sup>	1	2	2	1	1	1	1	3	1	3	3	3	2	1	1	1	4	2	2	1	2	3	3	2	1	2	2
Rudatsikira et al 2012 <sup>58</sup>	1	1	1	1	1	1	1	3	4	3	3	3	2	1	1	1	4	5	2	1	2	3	3	2	1	2	2
Sharon et al, 2009 <sup>59</sup>	2	2	2	2	-	-	3	3	4	3	3	3	2	1	1	1	4	2	2	1	2	3	3	2	1	2	3
Siziya et al, 2007 <sup>60</sup>	1	2	1	1	1	1	1	3	4	3	3	3	2	1	1	1	4	5	2	1	2	3	3	2	1	2	2
Siziya et al, 2008 <sup>61</sup>	1	1	1	1	1	1	1	3	4	3	3	3	2	1	1	1	4	5	2	1	2	3	3	2	1	2	2
Sokolova-Djokic et al, 2008 <sup>62</sup>	1	5	2	1	1	1	1	3	4	3	3	3	2	3	3	3	4	5	2	1	2	3	3	2	3	2	3
Stoll, 2008 <sup>63</sup>	2	5	2	2	-	-	3	3	4	3	3	3	2	1	1	1	4	5	2	1	2	3	3	2	1	2	3
Turek et al, 2001 <sup>64</sup>	1	5	2	1	1	1	1	3	4	3	3	3	2	1	1	1	4	5	2	1	2	3	3	2	1	2	2
Weaver et al, 2008 <sup>65</sup>	2	5	2	2	-	-	3	3	4	3	3	3	2	1	1	1	4	5	2	1	2	3	3	2	1	2	3

R = Rating; GR = Global Rating; - = Not applicable

Quality rating based on EPHPP tool<sup>25</sup>. A1=representation; A2=response rate; B1randomisation used; B2=randomization method described; B3=randomization method appropriateness; C1=group differences; C2=controlling of confounders; D1=assessor blinding; D2=participant blinding; E1=data collection tool validity; E2= data collection tool reliability; F1=withdrawal/drop outnumbers/reasons; F2=completion rate; G1=appropriate statistical methods.

A component rating (R) is produced for each component (except G): 1 = 'strong'; 2 = 'moderate'; 3 = 'weak'. A Global Rating (GR) is then produced from component ratings for A to F: 3 = 'weak' – having two or more weak component ratings; 2 = 'moderate' – having only one weak component rating, and 1 = 'strong' – having no weak component rating. As none of the selected studies were intervention studies, adaptations were made for elements relating to intervention studies.