**Supplementary Table 2. EPHPP Scores**

| | **Selection bias** | **Study design** | **Confounders** | **Blinding** | **Data collection method** | **Withdrawals/ dropouts** | **Average score** |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| Elbers (1999) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Gooren (2004) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Mueller (2011) | 2 | 2 | 2 | 1 | 3 | 3 | 2.2 |
| Wierckx (2014) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Gava (2016) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Auer (2016) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Auer (2018) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Jarin (2017) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Defrayne (2018) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Vita (2018) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Klaver (2018) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Olson-Kennedy (2018) | 2 | 2 | 2 | 1 | 3 | 1 | 1.8 |
| Tack (2018) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Tack (2017) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Scharff (2019) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Wiik (2019) | 2 | 2 | 2 | 1 | 3 | 3 | 2.2 |
| Van Caenegem (2014) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Haraldsen (2007) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| SoRelle (2019) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Greene (2019) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Roberts (2014) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Lapauw (2008) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Jain (2019) | 2 | 2 | 2 | 1 | 3 | cc | 2 |
| Sharula (2012) | 2 | 2 | 2 | 1 | 3 | cc | 2 |

**Notes on EPHPP scores:** The studies recruited transgender participants from gender clinics and were assessed as moderate for selection bias. All studies were assessed as moderate for study design as they used either retrospective or prospective cohort studies, in which measurements were conducted before and after hormone transition to assess possible changes. The studies were assessed as moderate for controlling confounding factors. Since none of the studies were blinded, all were considered weak on this variable of assessment, while all studies used valid and reliable medical records and as such were considered strong in terms of quality of data collection. Withdrawal and drop-outs were not applicable in the case-control studies but were appropriately described in two cohort studies[31,57] with low dropout levels and these were assessed as strong. A third cohort study[66] had a dropout level of greater than 40% and was assessed as weak. Based on the mean scores, all studies were categorized as moderate in quality (average scores between 1.8 and 2.2).