Online supplementary file 3: risk of bias assessment tables

Risk of bias assessment for systematic reviews

For the assessment of the risk of bias of the systematic review, included in our literature search, we used the revised tool of A MeaSurement Tool to Assess systematic Reviews (AMSTAR 2). Table 1 shows the item by item results for the paper of Burmester et al. Each item can be responded to as Yes, Partial yes or No. The overall rating of confidence for this systematic review is low.

Table 1.: A MeaSurement Tool to Assess systematic Reviews (AMSTAR 2) (used for Burmester et al. 2017)

AMSTAR-2		YES	Partial YES	NO
Item nr.	QUESTION			
1.	Were components of PICO included in research question and inclusion criteria?			X
2.	Was an 'a priori' design provided?	X		
3.	Was selection of included study designs explained?	X		
4.	Was a comprehensive literature search performed?			X
5.	Was there duplicate study selection?			X
6.	Was there duplicate data extraction?			X
7.	Was a list of excluded studies provided inclusive justification of the exclusion?			X
8.	Was included studies described in adequate detail?		X	
9.	Was satisfactory technique used for assessing risk of bias in individual studies?			X
10.	Was source of funding for included studies reported?		X	
11.	Were appropriate methods used for statistical combination of results?		X	
12.	Was the potential impact of risk of bias in individual studies on the results of the meta-analysis assessed?			X
13.	Was risk of bias in individual studies accounted for when interpreting results?			X
14.	Was a satisfactory explanation for, and discussion of, any heterogeneity observed being provided?			X
15.	Was adequate investigation of publication bias performed and likely impact on results discussed?			X
16.	Were potential sources of conflict of interest (including funding received) reported?	X		

Risk of bias assessment of randomised controlled trials

For the assessment of the risk of bias of RCTs, we used the revised Cochrane Risk of Bias 2 tool (RoB 2 tool). Risk of bias judgement per domain (randomisation process, intended interventions, missing outcome data, measurement of outcome and selection of reported results) are shown in table 2 and are rated as low, some concerns or high. In the last column of this table, the overall risk of bias is shown and this represents the worst domain score.

Table 2.: revised Cochrane Risk of bias 2 tool for randomized controlled trials (RoB 2 tool)

study	Bias arising from the randomisation process	Bias due to deviations from intended interventions	Bias due to missing outcome data	Bias in measurement of outcome	Bias in selection of the reported result	Overall bias
Goss et al. 2018	low	low	low	low	low	**low**
Krzysiek et al. 2009	low	some concerns	high	some concerns	low	**high**
Takeuchi et al. 2009	low	low	low	low	low	**low**
Ducourau et al. 2020	low	some concerns	low	high	low	**high**
Abdallah et al. 2017	low	low	low	low	low	**low**
Burmester et al. 2014	low	low	low	low	low	**low**
Burmester et al. 2015	low	low	low	low	low	**low**
Iwahashi et al. 2014	some concerns	low	low	low	low	**some concerns**
Ogata et al. 2014	low	low	low	low	low	**low**
St. Clair et al. 2002	low	some concerns	low	some concerns	low	**some concerns**
Wells et al. 2019	low	some concerns	low	some concerns	low	**some concerns**
Westhovens et al. 2015	low	low	low	low	low	**Low**
Zhuang et al. 2012	low	some concerns	low	some concerns	low	**some concerns**
l'Ami et al. 2018	low	some concerns	low	some concerns	low	**some concerns**
Breedveld et al. 2018	low	low	low	low	low	**Low**
Inman et al. 2018	low	low	low	low	low	**Low**
Braun et al. 2008	low	low	low	some concerns	some concerns	**some concerns**
Abe et al. 2006	low	high	low	some concerns	high	**high**
Syversen et al. 2020 (EULAR abstract)	low	high	no information	some concerns	low	**high**

l'Ami et al. 2020 (EULAR abstract)	some concerns	high	high	some concerns	low	**high**

Quality assessment of observational studies

Observational studies that were included in the body of evidence for the evaluation of prognostic factors, were assessed using the Quality in Prognosis Studies tool (QUIPS), see table 3. All other cohort or case-control studies were assessed using the Newcastle-Ottowa scale for observational (cohort or case-control) studies (NOS), see table 4a and 4b.

The QUIPS evaluates six areas for potential bias: participation, attrition, prognostic factor measurement, outcome measurement, study confounding and statistical analysis and reporting. Each area contains several questions to be rated as low, moderate or high. Table 3 summarises the worst score of the questions in each area. A question mark was scored when one or more of the questions in a particular area could not be answered (because of lack of information in the paper under assessment) and none of the other questions in that area were scored high.

With the NOS, a study can be awarded a maximum of one star for each item within the categories of selection (max. 4*) and outcome/exposure (max. 3*). Category comparability includes one item that can be rewarded a maximum of two stars.

Table 3.: QUality In Prognosis Studies tool (QUIPS) (summary rating)

study	**Risk of Bias**					
	study participation	study attrition	prognostic factor measurement	Outcome measurement	Study confounding	Statistical analysis and reporting
Arad 2019	low	high	high	low	moderate	moderate
Arends 2010	moderate	high	high	low	high	moderate
Arstikyte 2015	high	high	high	low	high	moderate
Bartelds 2011	low	moderate	low	low	moderate	low
Bartelds 2007	moderate	moderate	high	low	high	low
Bastida 2018	moderate	high	low	low	moderate	low
Bouman 2017	moderate	moderate	moderate	low	high	low
Chen 2015 (1)	high	?	moderate	low	high	moderate
Chen 2015 (2)	high	high	?	low	high	low
Daien 2012	high	low	low	low	high	low
Dervieux 2012	moderate	?	?	low	high	low
Dong 2019	moderate	high	moderate	low	high	moderate
Ducourau 2011	low	moderate	high	low	high	low
Eng 2015	low	?	?	low	high	moderate
Eng 2016	low	high	moderate	low	high	moderate
Gehin 2019	low	high	low	low	high	low

Jamnitski 2012	high	high	low	low	low	low
Jani 2015	moderate	moderate	moderate	moderate	moderate	low
Jani 2017	low	moderate	low	low	Low	low
Kneepkens 2015	moderate	moderate	low	low	low	low
Kneepkens 2014	moderate	moderate	moderate	moderate	moderate	low
Kneepkens 2015 (2)	moderate	moderate	moderate	moderate	moderate	low
Marsman 2016	moderate	high	moderate	moderate	high	low
Martinez-Feito 2019	low	high	low	low	high	low
Moots 2017	moderate	high	?	moderate	high	low
Mulleman 2011	moderate	high	moderate	low	high	low
Plasencia 2015	high	moderate	moderate	low	moderate	moderate
Pouw 2015	moderate	high	low	low	high	low
Rosas 2017	low	high	moderate	low	high	moderate
Sigaux 2017	moderate	high	high	moderate	high	moderate
Siljehult 2018	moderate	high	moderate	moderate	high	moderate
Takeuchi 2017	high	high	high	low	high	moderate
Van Kuijk 2010	high	high	moderate	low	high	low
Vogelzang 2014	high	high	Low	low	moderate	low
Vogelzang 2015	high	high	Low	low	high	low
Wolbink 2006	moderate	high	Low	low	high	low
Wolbink 2005	low	high	moderate	low	moderate	low
Jani 2020	moderate	high	high	low	high	low
Paul 2020	moderate	high	high	high	high	low
Ding 2020	moderate	high	?	low	high	low
D'Agostino 2017	low	high	?	low	high	low
Van den Bemt 2013	low	high	?	low	high	low
Ulijn 2020	moderate	high	moderate	high	high	low
Vincent 2016	moderate	moderate	high	moderate	high	moderate
Plasencia 2013	moderate	moderate	moderate	low	moderate	moderate
Jamnitski 2011	low	low	moderate	low	low	moderate
Bartelds 2010	low	low	moderate	low	low	low
Bastida 2020	high	?	moderate	high	moderate	low
Lamers-Karnebeek 2019	low	low	high	moderate	high	low
Chen 2016	high	high	low	low	moderate	high
Redondo 2018	high	high	low	low	moderate	moderate

L'Ami 2019	low	moderate	moderate	low	high	low
Bingham 2009	low	low	low	low	moderate	moderate
van der Bijl 2008	moderate	low	high	low	high	moderate
Jurado 2017	moderate	moderate	moderate	low	low	low
Mazilu 2014	high	moderate	moderate	moderate	high	low
Thurlings 2010	high	high	high	high	high	moderate
Ancuta 2018 (EULAR abstract)	high	high	high	moderate	high	high
Martinez-Feito 2019 (EULAR abstract)	high	high	high	low	moderate	moderate
Hooijberg 2019 (EULAR abstract)	high	high	?	low	high	moderate
Stamp 2019 (ACR abstact)	high	high	?	low	?	moderate

Table 4a.: Newcastle-Ottowa Scale (NOS) for observational (cohort) studies

	Risk of Bias		
study	Selection	Comparability	Outcome
Meric 2011	**	-	**
Mulleman 2009	**	-	*
Zänker 2018	*	-	*
Jani 2018 (EULAR abstract)	**	*	*
Van der Maas 2012	**	-	**
Martinez-Feito 2018	***	**	**
Mulleman 2010	**	-	*
Rosas 2014	**	-	*
Senabre 2019	**	**	*
Ducourau 2014	**	-	**
Kneepkens 2017	***	-	***
Sanmarti 2015	***	-	**
Ternant 2015	**	-	**
Van Herwaarden 2015	***	-	**
Benucci 2016	***	-	***
Boumans 2013	**	-	*
Chimenti 2016	**	-	**
De Vries 2009	***	-	***
De Vries 2007	***	*	***
Finckh 2010	***	*	*
Inciarte-Mundo 2016	*	-	*
Paramarta 2014	***	-	**
Pascual-Salcedo 2011	***	-	**

Plasencia 2012	***	-	**
Radstake 2009	***	-	**
Gehin 2020 (EULAR abstract)	**	**	**
Senabre-Gallego 2020 (EULAR abstract)	**	**	**
Gavan 2018 (EULAR abstract)	*	-	*
Perry 2018 (EULAR abstract)	*	-	**

Table 4b.: Newcastle-Ottowa Scale (NOS) for observational (case-control) studies

	Risk of Bias		
study	**Selection**	**Comparability**	**Exposure**
Bejan-Angoulvant 2017	*	**	*
Bendtzen 2006	**	-	*
Krintel 2013	**	-	*
Bender 2007	**	-	*
Hoxha 2016	**	-	*

Quality assessment of diagnostic studies

For diagnostic studies, comparing two or more types of assay for biopharmaceutical blood concentration or anti-drug antibodies measurement, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) risk of bias tool was used, see table 5. With QUADAS-2, four key domains (patient selection, index test(s), reference standard and flow and timing) are rated in terms of risk of bias and domains 1-3 are additionally rated for concern regarding applicability to the research question. Domains can be scored low, high or unclear. Each domain contains a set of questions to help reach the judgements.

Table 5.: Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) (summary rating)

	Domain 1: patient selection		**Domain 2: index test(s)**		**Domain 3: reference standard**		**Domain 4: flow and timing**
Study	Risk of bias	Concerns regarding applicability	Risk of bias	Concerns regarding applicability	Risk of bias	Concerns regarding applicability	Risk of bias
Alfonso et al. 2016	low	low	low	low	low	low	low
Bader et al. 2017	low	low	low	low	low	low	low
Bodini et al. 2015	low	low	low	low	low	low	low
Bodio et al. 2020	low	low	high	unclear	low	low	low
Clarke et al. 2019	low	low	low	low	low	low	unclear
Corstjens et al. 2013	low	low	low	low	low	low	low
Hock et al. 2019	low	low	low	low	low	low	low

Laserna-Mendieta et al. 2019	low	low	low	low	low	low	low
Martín et al. 2015	low	low	low	low	high	unclear	low
Nasser et al. 2018	low	low	unclear	unclear	low	low	low
Novakovic et al. 2019	low	low	low	low	low	low	low
Steenholdt et al 2013	low	low	unclear	unclear	low	low	low
Steenholdt et al 2014	low	low	low	low	low	low	unclear
Steenholdt et al 2015	low	low	low	low	low	low	unclear
Teixeira et al. 2018	low	low	high	unclear	low	low	low
Van Bezooijen et al. 2016	low	low	high	unclear	high	unclear	low
Van den Bossche et al. 2016	low	low	low	low	low	low	unclear
Verstockt et al. 2018	low	low	low	low	unclear	unclear	low
Willrich et al. 2015	low	low	low	low	high	unclear	low
Yang et al. 2018	high	unclear	low	low	high	unclear	low
Real-Fernández et al. 2019	low	low	low	low	low	low	low
Ogrič et al. 2019	unclear	low	low	low	low	low	low

Quality assessment economic evaluations

Methodological quality of the economic evaluations included in the body of evidence was assessed using the Consensus on Health Economic Criteria (CHEC) list, see table 6. This tool consists of nineteen items comprising specific criteria that have reached consensus as a generic core set of items for the quality assessment of economic evaluations. When information was clearly described in the paper (including supplementary materials) the item was scored 'yes', in case of insufficiently available information, the item is scored 'no'.

Table 6.: Consensus on Health Economic Criteria (CHEC)

		Included studies					
		Krieckaert et al. 2015		Jani et al. 2016		Laine et al. 2016	
Item nr.	**QUESTION**	YES	NO	YES	NO	YES	NO
1.	Is the study population clearly described?		X		N/A		X
2.	Are competing alternatives clearly described?	X			N/A		X

3.	Is a well-defined research question posed in answerable form?	X		X		X	
4.	Is the economic study design appropriate to the stated objective?	X		X		X	
5.	Is the chosen time horizon appropriate in order to include relevant costs and consequences?	X			N/A	X	
6.	Is the actual perspective chosen appropriate?	X		X			X
7.	Are all important and relevant costs for each alternative identified?		X	X		X	
8.	Are all costs measured appropriately in physical units?	X		X			X
9.	Are costs valued appropriately?		X		X	X	
10.	Are all important and relevant outcomes for each alternative identified?	X		X			X
11.	Are all outcomes measured appropriately?	X			N/A		X
12.	Are outcomes valued appropriately?	X			N/A		X
13.	Is an incremental analysis of costs and outcomes of alternatives performed?	X			X		X
14.	Are all future costs and outcomes discounted appropriately?	X			X		X
15.	Are all important variables, whose values are uncertain, appropriately subjected to sensitivity analysis?		X	X			X
16.	Do the conclusions follow from the data reported?	X		X		X	
17.	Does the study discuss the generalizability of the results to other settings and patient/client groups?	X			X		X
18.	Does the article indicate that there is no potential conflict of interest of study researcher(s) and funder(s)?	X		X			X
19.	Are ethical and distributional issues discussed appropriately?		X		X		X