BMC Medical Research Methodology

RESEARCH ARTICLE **Open Access**

Multiple Score Comparison: a network meta-analysis approach to comparison and external validation of prognostic scores

Sarah R. Haile[1†], Beniamino Guerra[1†], Joan B. Soriano[2], Milo A. Puhan[1,3*] for the 3CIA collaboration

Abstract

Background: Prediction models and prognostic scores have been increasingly popular in both clinical practice and clinical research settings, for example to aid in risk-based decision making or control for confounding. In many medical fields, a large number of prognostic scores are available, but practitioners may find it difficult to choose between them due to lack of external validation as well as lack of comparisons between them.

Methods: Borrowing methodology from network meta-analysis, we describe an approach to Multiple Score Comparison meta-analysis (MSC) which permits concurrent external validation and comparisons of prognostic scores using individual patient data (IPD) arising from a large-scale international collaboration. We describe the challenges in adapting network meta-analysis to the MSC setting, for instance the need to explicitly include correlations between the scores on a cohort level, and how to deal with many multi-score studies. We propose first using IPD to make cohort-level aggregate discrimination or calibration scores, comparing all to a common comparator. Then, standard network meta-analysis techniques can be applied, taking care to consider correlation structures in cohorts with multiple scores. Transitivity, consistency and heterogeneity are also examined.

Results: We provide a clinical application, comparing prognostic scores for 3-year mortality in patients with chronic obstructive pulmonary disease using data from a large-scale collaborative initiative. We focus on the discriminative properties of the prognostic scores. Our results show clear differences in performance, with ADO and eBODE showing higher discrimination with respect to mortality than other considered scores. The assumptions of transitivity and local and global consistency were not violated. Heterogeneity was small.

Conclusions: We applied a network meta-analytic methodology to externally validate and concurrently compare the prognostic properties of clinical scores. Our large-scale external validation indicates that the scores with the best discriminative properties to predict 3 year mortality in patients with COPD are ADO and eBODE.

Keywords: Prognostic scores, External validation, Multiple score comparison, Chronic obstructive pulmonary disease

Background

Prediction models, which combine predictors using regression coefficients, and simpler prognostic scores, which typically assign point values to predictors based on prediction models, have become increasingly popular [1, 2]. They aid in decision making in public health, clinical research and clinical practice [3] by estimating a person's risk of developing a disease or other outcome. In several medical fields, a variety of prediction models have been developed to assess the individual risk of adverse outcomes. A great example for this was a very recent systematic review regarding validated risk factor models for neurodevelopmental outcomes in children born very preterm or with very low birth weight [4]; 78 original studies (including 222 prediction models) were extracted. Most of the models were not intended to be used for clinical practice and only four studies (5%) had performed a validation. Another example regards models

* Correspondence: miloalan.puhan@uzh.ch
†Equal contributors
[1]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland
[3]Epidemiology & Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA
Full list of author information is available at the end of the article

predicting risk of type 2 diabetes mellitus with genetic risk models on the basis of established genome-wide association markers; a systematic review deemed to be eligible 21 articles representing 23 studies [5]. Concerning the risk of developing cardiovascular disease, over the past two decades, numerous prediction models have been developed, to estimate the risk of developing cardiovascular disease [6]. Only 36% of them were externally validated and only 19% by independent investigators. In the case of chronic obstructive pulmonary disease (COPD), several prognostic scores have been developed to predict mortality, starting with the BODE score [7]. But scores also exist to predict exacerbations [8], or the course of health-related quality of life [9, 10]. Prognostic scores suffer from a reluctance of general practitioners to use them [11, 12] as well as from scepticism because they lack internal and external validation which are requirements for generalizability [13, 14]. The external validation studies are often simply poorly designed or reported [15]. The lack of comparisons among available prognostic scores provides an additional hurdle to their widespread applicability, as practitioners may not be able to decide among them based on the information available [16].

Luckily, the collection of "big data" [17] and the growing availability of individual patient data (IPD) data analyses [18–23] provide researchers with new opportunities and challenges [24, 25]. Furthermore, the call of the medical community for data sharing [26] improves the possibilities of checking a model's predictive performance across clinical settings, populations, and subgroups [25]. The COCOMICS study [27] is a rare example of prognostic scores being directly compared with each other and simultaneously externally validated after pooling all the databases in a single cohort [16]. Our approach, multiple score comparison network meta-analysis (MSC), extends the simple pooling approach to pool direct comparisons taken from different studies, as a meta-analysis across studies provides in general higher quality information compared to the analysis of a database, constituted pooling together the single studies [28, 29]. This methodology allows to take into account heterogeneity of the individual studies and obtain more generalizable results [25].

Methods

Various methodological approaches have been proposed for network meta-analysis for comparison of treatments [30–36], which is sometimes referred to as network meta-analysis, multiple (or mixed) treatment comparisons meta-analysis (MTC meta-analysis) or multiple treatments meta-analysis [37, 38]. For diagnostic test performance, the first steps of network meta-analysis were undertaken (e.g. in terms of sensitivity and specificity) [39, 40]. No similar methodology exists to compare the performance of prognostic scores or prediction models. Nevertheless, network meta-analysis may provide an attractive solution to the problem of comparing the performance of prognostic scores.

Changing from comparing effects of treatments to comparing performance of prediction models or prognostic scores, however, reveals a number of differences between the two settings, and care must be taken to ensure that the unique features of multiple score comparison (MSC) meta-analysis (as we will refer to this new method) are considered properly in the analysis.

A number of features distinguish a MSC meta-analysis of prognostic scores from a meta-analysis of treatments. In network meta-analysis of treatments outcomes are summarized separately within each treatment arm of a randomized trial, and combined to obtain estimates of treatment effect (for example, mean difference or log odds ratio); instead, the MSC meta-analysis uses performance measures of each score in a cohort that can be calculated on the same sample of patients. Additionally, the number of prognostic scores assessed in a given cohort is not limited by the practicalities of study design, so that it would be easily possible to have more than, say, four scores within one cohort, while such a large number of treatment arms in an RCT is relatively unlikely due to considerations of power and sample size along with practical aspects of conducting clinical trials. Consideration of multi-score studies properly, including the correlations inherent in such comparisons, in MSC is therefore of great importance.

We developed a comprehensive approach to MSC to assess various prediction models using network meta-analysis with individual patient data, providing external validation and concurrent comparison of the scores, and applied it to risk prediction scores for mortality in COPD [41, 42]. After careful methodological issues (see also online-only material, where we go deeper into the statistical background) the following approach was developed: we calculated aggregated summary statistics for each cohort and score. Then, we examined the network structure by grouping the cohorts according to which scores could be evaluated. We adapted methodology from network meta-analysis [35] to concurrently externally validate and compare prognostic scores from individual patient data across different cohorts, explicitly including correlations [43] between the scores on a cohort level.

We will also re-interpret NMA as a two-stage meta-regression model, as proposed in [35]:

1. Ordinary meta-analysis to gain the direct estimates for corresponding pooled effect estimates (using the inverse-variance weighted means of the

corresponding cohorts). Cohorts at our disposal are classified into "groups" according to which scores it is possible to evaluate by their data.

2. Based on the direct estimates and their variances from the first stage, they obtain to find the optimal estimate of the pooled effect parameters that obeys the fundamental consistency equations. In this stage we merge the group estimates, looking for the weighted least squares solution to the regression problem equation.

The last steps were to confirm that transitivity is a plausible assumption and to check for possible inconsistency and heterogeneity.

Calculation of aggregated summary statistics

First, the performance measure for comparison of the various prognostic scores was defined as the area under the curve (AUC) of the corresponding receiver operating characteristic curve (ROC). This is a graphical plot that illustrates the ability of a binary classifier system (diagnostic or prognostic) as its discrimination threshold is varied (in particular plotting true vs false positive rate). Differences in AUC, denoted ΔAUC, provided an estimate of the relative discrimination ability when comparing scores. For this purpose, we use of a common comparator (CC) model (in our case the GOLD classification, since it is a variable supposed to be present in each COPD cohort); it constitutes a reference value for the performance of other scores, the value from which to subtract the possibly common biases [44].

Variance and covariance estimates for the ΔAUC values were estimated numerically using bootstrapping. We also confirmed consistency of bootstrapped variance estimates to those of the analytical formula for variances of paired differences in AUC [45] (results not shown).

Aggregated data on the cohort level for a cohort with k scores consist therefore of *k – 1* ΔAUC estimates and a corresponding *(k – 1) x (k – 1)* variance-covariance matrix.

To further clarify the methodology, we show the main steps with a small fictional example. Suppose we had 2 cohorts where score A and B could be evaluated (group 1: AB; cohorts P, R), 2 cohorts where A and C could be evaluated (group 2: AC; cohort S, T), 2 cohorts where A, B and D could be evaluated (group 3: ABD; cohorts U, V), and a final 2 cohorts where A, B, C, and D could be evaluated (group 4: ABCD, cohorts X and Y). Let us focus on group 3, constituted by the cohorts X and Y in which the scores A, B, and D can be used. We would obtain performance difference of the scores B and D in comparison to the score A for each of the cohorts, as reported in Table 1.

Analogously, in group 3, we would obtain variance-covariance matrices, like the ones reported in Table 2.

Table 1 Point estimate of the difference of AUC of the scores B and D with the score A in the group 3 of the fictional example

Cohort	ΔAUC – AB	ΔAUC – AC	ΔAUC – AD
X	0.05	–	0.07
Y	0.06	–	0.18

Examination of network structure

Once the aggregated summary statistics were computed, we explored the structure of the network. In a first step, we divided the cohorts into groups based on which sets of scores could be evaluated.

Each group is represented by a polygon, that passes by all the scores (i.e. the vertices) which can be evaluated in the cohorts constituting that group. The thickness of the polygon is directly proportional to the number of deaths in the group (Fig. 1).

Head-to-head comparisons within a group can be performed between any two scores connected in the same polygon.

For example, in group 4, A and D can be compared because they are both connected by the same polygon, even though there is no line directly connecting the two scores in that group.

According to Table 3, group 1, group 2, group 3 and group 4 have a cumulative number of 4000, 1000, 3000 and 2000 patients, respectively.

Multiple score comparison

The method of Lu et al. [35] was used to perform the multiple score comparison meta-analysis with Der Simonian-Laird random effects [46–49]. This method, which reinterprets frequentist NMA as a two-stage meta-regression model (using inverse variance weighted least squares estimation), was chosen as, compared to most of the network meta-analytic techniques, it can easily handle multi-score cohorts, and does not make unnecessary simplifications with respect to the correlations inherent in such trials, as discussed above. In the first stage, cohorts in which the same set of scores have been assessed are grouped together and meta-analysed separately.

Table 2 Variance-covariance matrices of the difference of AUC of the scores B and D with the score A in the group 3 of the fictional example

Cohort X		
	0.0012	0.0005
	0.0005	0.0009
Cohort Y		
	0.0068	0.0051
	0.0051	0.0109

An estimation T^2 of the between - cohort variance (τ^2) (i.e., the variance of the true performance difference

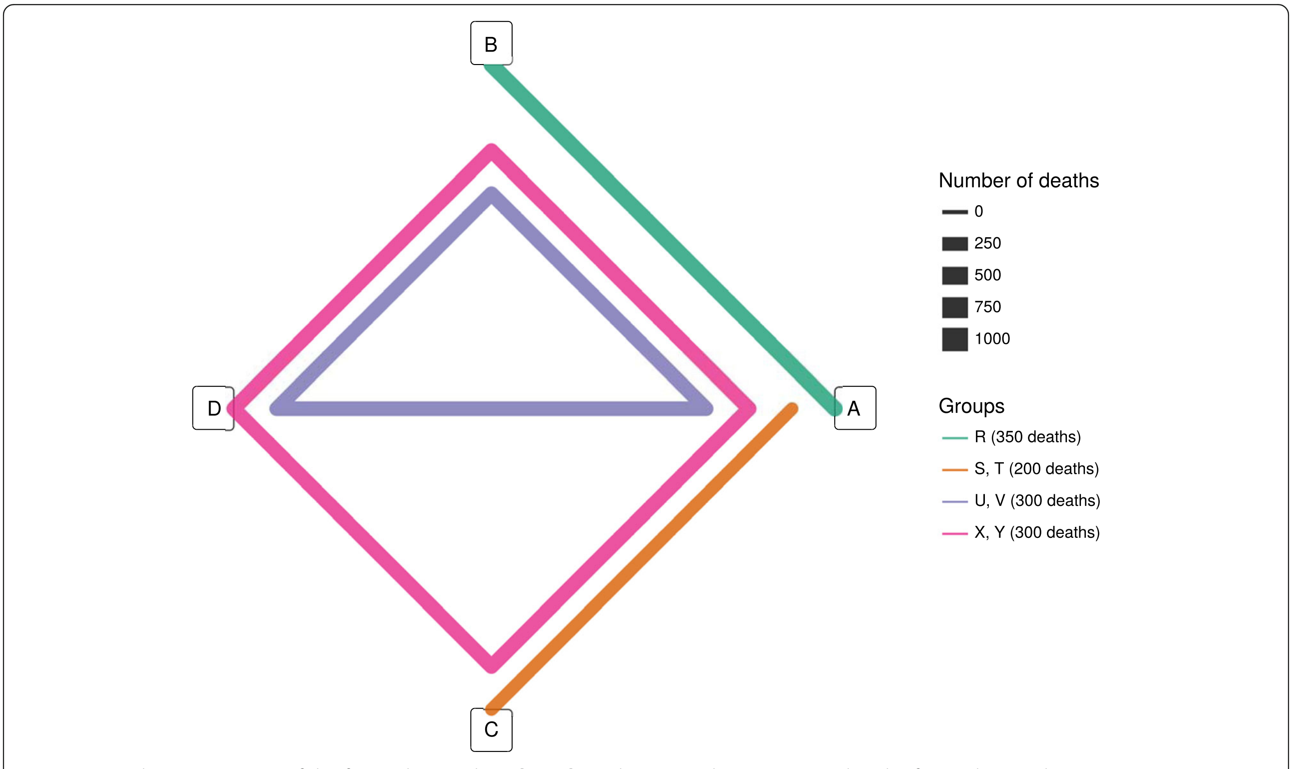

Fig. 1 Network representation of the fictional example. **a**, **b**, **c**, **d** are the scores that are assessed in the fictional example

across all studies) is the Der Simonian-Laird method [47] adapted to the network meta-analysis case [35]. Indeed, the *Q* statistic (adapted to network meta-analysis) is referred to a χ^2 distribution with degrees of freedom $df_g = (M_g - 1)(N_g - 1)$, where M_g is the number of scores compared in the group g and N_g is the number of cohorts belonging to the group *g*. Thus, the degrees of freedom are $df_1 = 1*0 = 0$, $df_2 = 1*1 = 1$, $df_3 = 2*1 = 2$, $df_4 = 3*1 = 3$. Table 3 allows us to calculate pooled τ^2 (according to the methods of moments) [46] with which we evaluate the weights used to obtain the weighted average of the performance estimate for the whole network (reported in the first 4 rows in Table 4).

Analogously, extending the definitions from meta-analysis [46] to network meta-analysis [35] we calculate the variables C, Q and τ (τ represents the heterogeneity and deserves further discussion in the text later).

Table 3 Group characteristics of a fictional network (*g* identifies the group. *n* is the total number of subjects and *d* is the total number of deaths in each group. Additional characteristics are also listed: the *Q* statistic describing heterogeneity has *df* degrees of freedom, and τ gives the square root of the τ^2 statistics for between - cohort heterogeneity)

g	Scores	Cohorts	*n*	*d*	*Q*	*df*	τ
1	A, B	R	4000	350	28	0	0.019
2	A, C	S, T	1000	200	15.5	1	0.014
3	A, B, D	U, V	3000	300	13.4	2	0.004
4	A, B, C, D	X, Y	2000	300	9.6	3	0.036

In Stage II the inverse variance weighted least square solution across all groups is found, thus we obtain the performance vector related to each score, best fitting the results of Stage I (for more details see Additional file 1). In the last row of Table 4, the final results of the MSC meta-analysis for the fictional example of are reported.

Transitivity, heterogeneity and inconsistency

The main assumptions to be met for performing a network meta-analysis are transitivity (a key assumption related to consistency), heterogeneity (differences in estimates of the same treatment or score contrasts coming from different studies) and inconsistency (comparing direct and indirect estimates, sometimes referred to as incoherence) [44, 50]. A key assumption of consistency is transitivity (sometimes referred to as similarity [51])

Table 4 Stage I and Stage II results of the MSC meta-analysis in the fictional example of Table 1 (comparison with the A score)

Stage	G	B	C	D
I	1	0.09 (0.07, 0.12)		
I	2		0.18 (0.07, 0.29)	
I	3	0.10 (0.08, 0.12)		0.22 (−0.05, 0.50)
I	4	0.08 (0.04, 0..12)	0.15 (0.01, 0.29)	0.18 (0.05, 0.31)
II		0.09 (0.06, 0.13)	0.17 (0.10, 0.25)	0.21 (0.07, 0.35)

among the treatment effects [34, 44, 51–55], that is, that indirect comparisons are valid estimates of (unobserved) direct comparisons. Therefore one statistical approach to check for transitivity in our case is to explore the distribution variables giving case-mix across groups [56, 57], which we have adopted here using meta-regression. In practice, we used the definition of transitivity from a review paper on the topic [44] better matching our methodology, namely that the different sets cohorts do not differ with respect to the distribution of variables that could generate case mix variation. Thus, we evaluated by meta-regression analysis [58] the distribution of the variables that could generate case mix variation (like median and variability of age [25], range and variance of obstruction severity (i.e., FEV1% pred.), exercise capacity, size, mortality rate).

In case of variables directly affecting the performance, we used analysis of variance (ANOVA) to see whether the distribution of the identified variables was imbalanced in the groups and could consequently generate imbalance in the performance group by group. In case of homogenous groups, we cannot reject the null hypothesis of transitivity. With this method we assess as well, the eventuality that within-cohort heterogeneity could affect the analysis when "case-mix" is present (i.e. heterogeneity in the variables representing heterogeneity in the cohorts, like FEV%predicted range, that could affect the discriminative properties in the specific cohorts).

Heterogeneity could be described using a multivariate version of the usual τ^2 statistic, which in the Lu et al. [35] approach is considered on a group level at stage 1. They suggest that a pooled τ^2 may be a natural solution to situations where there are singleton groups (i.e. constituting only a cohort).

Inconsistency was primarily assessed visually by comparing direct and indirect comparisons from node-splitting side by side [59]. As a further check of inconsistency, we further examined the Q statistic (that is, the residual sum of squares) which can be used to reject the hypothesis of inconsistency between direct and indirect estimates if Q is greater than the χ^2 statistic with N − K + 1 degrees at freedom at the 100(1 − α)% level [35], where N is the sum of the number of contrasts in each group, and K is total number of scores. Furthermore local consistency could be assessed at the group level by examining residuals and leverage statistics. Furthermore, we considered ways to calculate direct and indirect evidence within the network. Direct comparisons were computed by including only cohorts where both scores under consideration were present, and then performing the usual random effects meta-analysis [46]. However, defining loops of any order for indirect comparisons proved to be difficult in our setting, where the network is highly connected, and most cohorts have between 4 and 9 scores being assessed. Due to the various difficulties presented by studies with multiple scores, we chose to examine inconsistency in the network using "node-splitting" [59]. This approach avoids the need to define loops of any order, and includes all possible indirect evidence.

Consideration of missing data

The main analysis was performed without any imputation technique. A sensitivity analysis, using multiple imputation was also performed and it is shown in the online-only material. The results were not significantly different in the two cases.

COPD data description

Following the recommendation for large prospective studies [41], we based our analysis on a large-scale database (provided by the COPD Cohorts Collaborative International Assessment (3CIA) consortium [42]) from a diverse set of 24 cohort studies and 15,762 patients with COPD (1871 deaths and 42,203 person-years of follow-up). The cohorts were heterogeneous concerning geographic location, sample size, number of events and correspond to a broad spectrum of patients with COPD from primary, secondary and tertiary care settings. Mean FEV1 ranged from 30 to 70% of the predicted values, mean modified Medical Research Council (mMRC) dyspnea scores from 1.0 to 2.8 (the scale goes from 0 to 4, with 4 being the worst), mean number of exacerbations in the previous year (where available) from 0.2 to 1.7 and mean 6-min walk distance (where available) from 218 to 487 m. The follow-up period varied from cohort to cohort, thus we decided to use a minimum common time frame of 3 years. The mean age varies between 58 and 72 years. The outcome of interest was 3-year all-cause mortality. A table summarizing the clinical characteristics of the cohorts is reported in Additional file 1.

Results

To illustrate an MSC meta-analysis, we compared the prognostic ability of various scores to predict mortality in patients with COPD. The COPD Cohorts Collaborative International Assessment (3CIA) [42] initiative contains individual data for around 16,000 patients (approx. 70,000 person years) with COPD from 26 cohorts in seven countries. Patients were considered to have COPD if the ratio of forced expiratory volume in 1 s (FEV1) to forced vital capacity (FVC) was less than 70%, regardless of the Global Initiative for Chronic Obstructive Lung Disease (GOLD) (2007) stage (I–IV) [60]. The minimum required set of variables for each cohort included vital status (up to death, loss to follow-up, or last data collection in June 2013), age, sex, pre-bronchodilator and post-bronchodilator FEV_1 and dyspnoea MRC grade

[42]. Most cohorts included many more variables allowing for the calculation of a total of 10 prognostic scores: GOLD (2007), GOLD (2011), updated ADO, BODE, updated BODE, eBODE, BODEx, DOSE, SAFE and optimised B-AE-D [7, 10, 61–65].

Examination of network structure

We apply the MSC network meta-analysis of prognostic scores for 3-year mortality from the 3CIA data.

Based on the availability of the 10 scores in each cohort, the cohorts could be divided into 6 groups. The network structure is shown in Table 5 and in Fig. 2.

Even if it would make sense to use absolute performance, we used relative performance of each score in comparison to a Common Comparator score, in order to get rid of possible common biases. We chose as Common Comparator score the GOLD classification. One cohort (COCOMICS Requena I) was excluded from the analysis because it only had sufficient variables to evaluate a single score (GOLD) and it would not contribute to the analysis. We had to further exclude the cohort A1ATD because there were no cases in the follow-up considered for our analysis (3 years) and the lack of events does not allow calculating an AUC. Of the remaining 24 cohorts, 4 had 2 scores (GOLD (2007), updated ADO), and the other 20 had between 3 and 9 scores assessed. We note that in no cohort all the 10 scores could be evaluated.

As GOLD (2007) is commonly used to classify the grade of severity of COPD patients, it could be assessed in all cohorts [60]. We note that direct evidence was available for 41 of 45 score comparisons, indirect evidence was available for other 16 comparisons (among which the four cases in which the direct comparison was missing).

Multiple score comparison meta-analysis (MSC)

Transitivity, heterogeneity and inconsistency

To examine whether transitivity was fulfilled, we analysed the distribution of a number of possible variables potentially generating case-mix, following epidemiological reasoning and literature (age median and variability [25], FEV1 percent predicted range and variance, mortality percentage, exercise capacity range, number of events) across the groups using meta-regression. For the variables generating case-mix (whose meta-regression analysis were significant), the ANOVA analysis showed that the variables were balanced in the groups. Thus, we cannot reject the null hypothesis of transitivity.

Stage I group level results are presented in the top of Table 6, while the bottom rows show the stage II overall results from the network meta-analysis. GOLD (2007) scores ranged from 0.481 to 0.731, with a median of 0.614, and interquartile range (0.587, 0.641). Of the scores, the one that predicted mortality best was updated ADO with an average AUC 0.083 higher than that of GOLD (2007) (95% confidence interval: 0.069, 0.097), followed by the updated BODE which was associated with a 0.072 better AUC than GOLD (95% confidence interval: 0.051, 0.093) and eBODE (+0.069, 95% confidence interval: 0.044, 0.093). DOSE (+0.027, 95% confidence interval: 0.010, 0.045), optimised B-AE-D (+0.016, 95% confidence interval: −0.007, 0.038) and GOLD (2011) (+0.014, 95% confidence interval: 0.001, 0.028) and showed the worst performance in predicting mortality, only slightly better than GOLD (2007). The other scores, BODE, SAFE and BODEx showed moderate performance, between +0.045 and +0.064 improvement in AUC over GOLD.

Table 5 Group characteristics of the network

g	Scores	Cohorts[a]	*n*	*d*	*Q*	*df*	*τ*
1	GOLD – ADO	Copenhagen*, HUNT, Japan, SEPOC*	4323	378	2.8	3	0
2	GOLD – ADO – BODE – BODEupd	Barmelweid*, Basque*, Galdakao†, Pamplona†, Zaragoza I†	1208	215	15.5	12	0.014
3	GOLD – GOLD (2011) – ADO – BODE – BODEupd	Mar de Plata Argentina, PACECOPD*, Son Espases Mallorca	556	61	10.9	8	0.025
4	GOLD – GOLD (2011) – ADO – BODE – BODEupd – SAFE	COPDgene	4484	337	7.46E-29	0	NA
5	GOLD – GOLD (2011) – ADO – BODEx – DOSE – BAED	Genkols, ICE COLD ERIC, Initiatives BPCO, Sevilla†, Terrassa I†, Terrassa III†, Zaragoza II†	4346	722	48.1	30	0.011
6	GOLD – GOLD (2011) – ADO – BODE – BODEupd – eBODE – BODEx – DOSE – BAED	La Princesa Madrid, Requena II†, Tenerife†, Terrassa II†	845	125	34.5	24	0.014

Abbreviations: *g* group, *n* number of subjects, *d* number of deaths, *Q* likelihood ratio statistic, *df* degrees of freedom, *τ* heterogeneity within the group, *GOLD* Global initiative for chronic Obstructive Lung Disease, *BODE* Body mass index, airflow Obstruction, Dyspnoea and severe Exacerbations, *BODE upd.* BODE updated, *ADO* Age, Dyspnoea, airflow Obstruction (we use in the our analysis the updated version of the ADO score), *e-BODE* severe acute exacerbation of COPD plus BODE, *BODEx* Body mass index, airflow Obstruction, Dyspnoea, severe acute Exacerbation of COPD, *DOSE* Dyspnoea, Obstruction, Smoking and Exacerbation frequency, *SAFE* Saint George's Respiratory Questionnaire (SGRQ) score, Air-Flow limitation and Exercise capacity, *B-AE-D* Body-mass index, Acute Exacerbations, Dyspnoea

[a] Cohorts belonging to the ADO or COCOMICS groups are marked with * or † respectively. We notice that for group 4 the value of heterogeneity τau is not available (NA); indeed, that is a singleton group, where we cannot evaluate heterogeneity

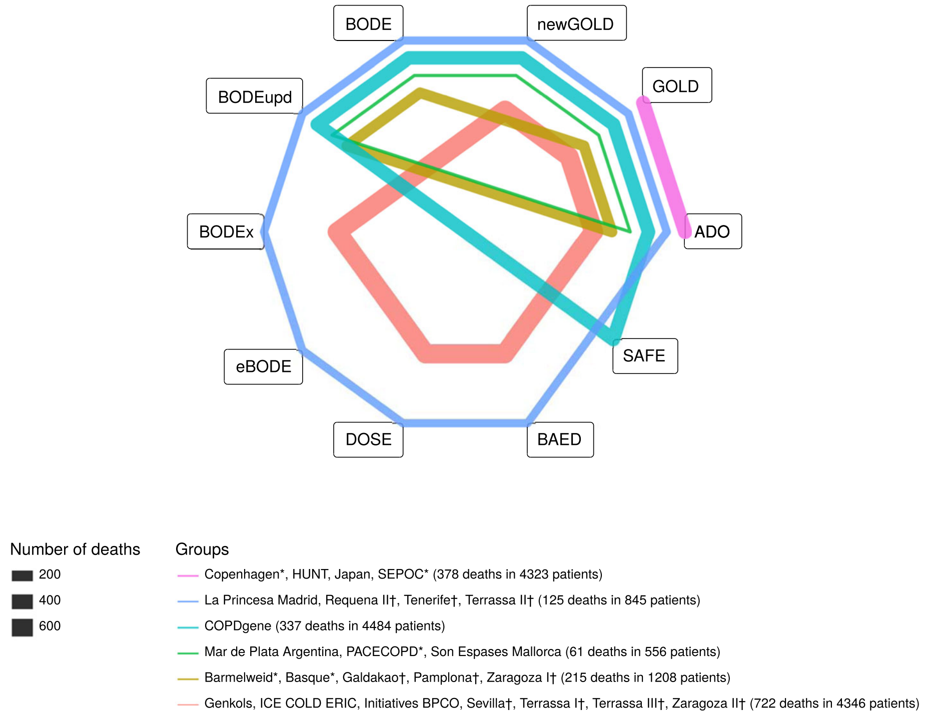
BODE
newGOLD
BODEupd
GOLD
BODEx
ADO
eBODE
SAFE
DOSE
BAED
Number of deaths
200
400
600
Groups
Copenhagen*, HUNT, Japan, SEPOC* (378 deaths in 4323 patients)
La Princesa Madrid, Requena II†, Tenerife†, Terrassa II† (125 deaths in 845 patients)
COPDgene (337 deaths in 4484 patients)
Mar de Plata Argentina, PACECOPD*, Son Espases Mallorca (61 deaths in 556 patients)
Barmelweid*, Basque*, Galdakao†, Pamplona†, Zaragoza I† (215 deaths in 1208 patients)
Genkols, ICE COLD ERIC, Initiatives BPCO, Sevilla†, Terrassa I†, Terrassa III†, Zaragoza II† (722 deaths in 4346 patients)

Fig. 2 Depiction of network structure with lines weighted by the total number of deaths in the group. Abbreviations: GOLD, Global initiative for chronic Obstructive Lung Disease; BODE, Body mass index, airflow Obstruction, Dyspnoea and severe Exacerbations; BODE upd., BODE updated; ADO, Age, Dyspnoea, airflow Obstruction (we use the updated version of the ADO score in our analysis); e-BODE, severe acute exacerbation of COPD plus BODE; BODEx, Body mass index, airflow Obstruction, Dyspnoea, severe acute Exacerbation of COPD; DOSE, Dyspnoea, Obstruction, Smoking and Exacerbation frequency; SAFE, Saint George's Respiratory Questionnaire (SGRQ) score, Air-Flow limitation and Exercise capacity; B-AE-D, Body-mass index, Acute Exacerbations, Dyspnoea. [a]Cohorts belonging to the ADO or COCOMICS groups are marked with * or † respectively. [b]The thickness of the lines is proportional to the number of deaths of the respective cohort. [c]Below we report the composition of the group; each of them is identified by a specific color: 🟪 Copenhagen*, HUNT, Japan, SEPOC* (378 deaths in 4323 patients), 🟨 Barmelweid*, Basque*, Galdakao†, Pamplona†, Zaragoza I† (215 deaths in 1208 patients), 🟩 Mar de Plata Argentina, PACECOPD*, Son Espases Mallorca (61 deaths in 556 patients), 🟦 COPDgene (337 deaths in 4484 patients), 🟥 Genkols, ICE COLD ERIC, Initiatives BPCO, Sevilla†, Terrassa I†, Terrassa III†, Zaragoza II† (722 deaths in 4346 patients), 🟦 La Princesa Madrid, Requena II†, Tenerife†, Terrassa II† (125 deaths in 845 patients)

Concerning heterogeneity, due to the group containing only a single cohort (group 6), we primarily considered a random effects analysis calculated using a pooled estimate of τ^2 for all groups, which was 0.00015, indicating a relatively low heterogeneity. The results of the MSC network meta-analysis were not substantially different when using the group-specific τ^2 estimates.

Possible inconsistency between direct and indirect comparisons was assessed using the Q statistic as described above. Overall, Q for the random effects analysis was 22.1 with 16 degrees of freedom. Keeping in mind that in this case (as in classical network meta-analysis) the inconsistency test has low power, since Q was smaller than the corresponding χ^2 statistic of 26.3, we do not reject the hypothesis of consistency ($P = 0.14$).

Both direct and indirect estimates of the score comparisons were calculated using node-splitting [59], and compared visually. The results are similar to each other and to the estimates provided by the network meta-analysis (see Additional file 1 for further discussion).

Consideration of missing data

As a secondary analysis, the entire meta-analysis was repeated in a multiple imputation framework, as described above. The results were similar to the main analysis without imputation (see Additional file 1) [1, 66, 67].

Discussion

To the best of our knowledge, the MSC meta-analysis proposed in this paper represents the first methodology to evaluate the comparative prognostic properties of prediction models that synthesizes all available (direct and indirect) evidence. The application of the MSC meta-analysis could provide different clinical fields with a clear indication of which is the best-performing prediction model, paving the way for a standardized clinical application. While there are a number of issues when adapting usual NMA methodology to MSC, they can be addressed in a straightforward manner. Multi-score studies are considered in our approach by explicitly using covariance estimates for the various prognostic

Table 6 Stage I and Stage II results of the MSC meta-analysis (comparison with the GOLD classification)

Stage	g	ADO	BODEupd	eBODE	BODE	SAFE	BODEx	DOSE	BAED	newGOLD
I	1	0.097 (0.07, 0.123)								
I	2	0.098 (0.057, 0.139)	0.124 (0.078, 0.17)		0.098 (0.059, 0.137)					
I	3	0.044 (−0.03, 0.117)	0.023 (−0.054, 0.099)		0.019 (−0.054, 0.091)					−0.011 (−0.053, 0.03)
I	4	0.042 (0.01, 0.074)	0.043 (0.008, 0.078)		0.049 (0.017, 0.081)	0.037 (0.005, 0.069)				−0.008 (−0.038, 0.022)
I	5	0.099 (0.076, 0.123)					0.056 (0.035, 0.076)	0.036 (0.015, 0.057)	0.032 (0.005, 0.058)	0.028 (0.008, 0.047)
I	6	0.076 (0.027, 0.126)	0.043 (−0.006, 0.092)	0.048 (0.004, 0.093)	0.043 (0.001, 0.085)		0.030 (−0.015, 0.074)	0.021 (−0.023, 0.065)	−0.017 (−0.079, 0.045)	0.008 (−0.031, 0.047)
II		0.083 (0.069, 0.097)	0.072 (0.051, 0.093)	0.069 (0.044, 0.093)	0.064 (0.045, 0.082)	0.052 (0.022, 0.082)	0.045 (0.029, 0.061)	0.027 (0.010, 0.045)	0.016 (−0.007, 0.038)	0.014 (0.001, 0.028)

Abbreviations: *MSC* Multiple Score Comparison, *GOLD* Global initiative for chronic Obstructive Lung Disease, *BODE* Body mass index, airflow Obstruction, Dyspnoea and severe Exacerbations, *BODE upd.* BODE updated, *ADO* Age, Dyspnoea, airflow Obstruction (we use in our analysis the updated version of the ADO score), *e-BODE* severe acute exacerbation of COPD plus BODE, *BODEx* Body mass index, airflow Obstruction, Dyspnoea, severe acute Exacerbation of COPD, *DOSE* Dyspnoea, Obstruction, Smoking and Exacerbation frequency, *SAFE* Saint George's Respiratory Questionnaire (SGRQ) score, Air-Flow limitation and Exercise capacity, *B-AE-D* Body-mass index, Acute Exacerbations, Dyspnoea

The first six rows show the Stage I results (group by group). The last row shows the Stage II results (namely the final results of the multiple score comparison meta-analysis)

The scores are ordered by performance of the prognostic scores in Stage II scores. Calculation of such estimates using bootstrapping may be computationally intensive but is not difficult to implement. The approach presented here can be used to compute prognostic score comparisons for the entire network of evidence, as well as both direct and indirect comparisons between scores.

Despite these adaptations, the results of the MSC meta-analysis are clear, and may be interpreted in a fashion similar to standard network meta-analysis results. The only difference is that the performance measure is not mean difference between treatments, or log odds ratio, but difference in performance measure such as AUC. Measures of heterogeneity and inconsistency can however be calculated and interpreted in the usual fashion [44]. For instance, a definition similar to the one used for the heterogeneity for meta-analysis of direct comparisons, can be used for the heterogeneity of network meta-analysis, adapting a definition used for multi-arm trials to multiple score comparison. Since we have singleton groups in our MSC data (group 6 in our database), it is recommended in our case to use pooled estimate of the τ^2 (τ^2_{pooled}) [35], i.e. a multivariate version of the pooled estimate for the heterogeneity variance (more technical details are provided in Additional file 1).

We used one of the scores as a common comparator, which would not generally be necessary, but may be easily possible in this MSC setting.

Performance of the considered prognostic scores can be computed from the individual patient data (IPD) directly; this is how we approach the problems having at our disposal a large-scale IPD database. The group results (Stage I) arise from averaging the cohort results, that, in turn, are calculated using the IPD of each cohort. If no IPD are available, instead, there are two possibilities: use published results, or send cohorts code to extract the aggregated performance measures individually. Use of published results requires that comparisons have been reported for more than one score, which in practice may almost never be the case. Sending code to obtain aggregated measures may be an optimal approach in cases where no large-scale collaboration exists, and published results are not detailed enough.

We used all-cause mortality as outcome. Apart from being clinically relevant, mortality is the easiest outcome that we can expect to evaluate in a cohort, with a hard definition. This makes it easier to reduce the problems related to miss-classification or missingness of the outcome [68, 69].

Given the patterns of missing data (in general, the variables are completely or almost completely missing or not missing at all) a sensitivity analysis performed after multiple imputation is providing similar results to the analysis without imputation (a comparison is provided in Additional file 1). Analogously, a sensitivity analysis using heterogeneity group by group gives similar results than using a pooled heterogeneity (here recommended because of the network structure; more details are available in Additional file 1).

There are a few limitations to this approach to MSC. Although the analysis can be implemented as outlined by Lu et al. [35] (Additional file 1), creating an input dataset in a spreadsheet may be less than straightforward. We have therefore provided example R code to convert a dataset of prognostic scores to a MSC meta-analysis, without first making a table of cohort-level summary statistics, as is often performed. We note however that such a dataset including a column for each cell of the variance-covariance matrices could be analysed using mvmeta in Stata. Creating that kind of summary dataset might be useful to go along with the network meta set of commands in Stata [70]. We focused on implementing this approach starting from the raw prognostic scores from individual patients, which had been calculated using raw data from the international collaboration [42].

Conclusions

In summary, we have adapted methodology from network meta-analysis to compare prognostic scores from individual patient data across different cohorts. This approach permits concurrent external validation of the scores in a consistent analysis explicitly including correlations between the scores on a cohort level. Estimates of differences in performance can be estimated for the entire network, as well as for both direct and indirect comparisons of scores. Results of the MSC meta-analysis can be interpreted in a manner similar to that of the usual network meta-analysis, regardless of the performance measure used. Our application to prognostic scores showed that the ADO and updated BODE scores have the best discriminative performance to predict mortality for patients with COPD. The meta-analysis could also be repeated for a number of different performance measures in order to describe multiple facets of the prognostic scores (e.g. discrimination and calibration [1]) or using reclassification methods (like the net reclassification index, NRI [71]) or to aid in the interpretation of the results. Development of clearer data input formats as well as more automated would provide opportunities for further methodological research in MSC meta-analysis.

Additional file

Additional file 1: Multiple Score Comparison: A network meta-analysis approach to comparison and external validation of prognostic scores. Additional statistical background is provided, as well as further details on heterogeneity, transitivity, inconsistency, indirect evidence, and multiple imputation, as well as a table summarizing the scores compared. (DOCX 750 kb)

Abbreviations

COPD: Chronic obstructive pulmonary disease; IPD: Individual patient data; MSC: Multiple score comparison

Acknowledgements

We would like to thank Hélène Aschmann, Henock Yebyo and Rafael Sauter (all University of Zurich, Zurich, Switzerland) for their comments.
3CIA Collaborators: Prof. Bernd Lamprecht, Dr. Ana S. Ramírez, Dr. Pablo Martinez-Camblor, Dr.Bernhard Kaiser, Dr. Inmaculada Alfageme, Dr. Pere Almagro, Dr. Ciro Casanova, Dr. Cristóbal Esteban-González, Dr. Juan J. Soler-Cataluña, Dr. Juan P. de Torres, Prof. Marc Miravitlles, Prof. Bartolome R. Celli,Dr. Jose M. Marin, Dr. Gerben ter Riet, Dr. Patricia Sobradillo-Ecenarro, Prof. Peter Lange, Prof. JudithGarcia-Aymerich, Prof. Josep M. Antó-Boqué, Dr. Alice M. Turner, Dr. Meilan K. Han, Dr. ArnulfLanghammer, Dr. Linda Leivseth, Dr. Per Bakke, Dr. Ane Johannessen, Dr. Oga Toru, Dr. Borja Cosío, Dr.Julio Ancochea-Bermúdez, Dr. Andres Echazarreta, Prof. Nicolas Roche, Prof. Pierre-Régis Burgel, Prof.Don D. Sin, Prof. Joan B. Soriano, Prof. Milo A. Puhan.

Funding

Financial support specific to the current paper (salary support for PhD candidate BG) came from the University of Zurich (Zurich, Switzerland), who had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. No additional funding was available.

Availability of data and materials

The dataset used during the current study are available from the corresponding author on reasonable request.

Authors' contributions

SRH drafted the manuscript. BG and SRH performed the statistical analysis. All the authors conceived and designed the study, and performed a critical revision of the manuscript. JBS and MAP supervised the study. All authors have read and approved the final version of the manuscript.

Ethics approval and consent to participate

All cohort studies were approved by their respective ethics committee and patients provided written informed consent. Details are included in the paper related to the 3CIA collaboration and in the individual studies:

- de Torres JP, Casanova C, Marin JM, et al. Prognostic evalation of COPD patients: GOLD 2011 versus BODE and the COPD comorbidity index COTE. Thorax 2014; 69: 799–804.
- Puhan MA, Garcia-Aymerich J, Frey M, et al. Expansion of the prognostic assessment of patients with chronic obstructive pulmonary disease: the updated BODE index and the ADO index. Lancet 2009; 374: 704–11.
- Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. COPD 2010; 7: 32–43.
- Vestbo J, Prescott E, Lange P. Association of chronic mucus hypersecretion with FEV1 decline and chronic obstructive pulmonary disease morbidity. Copenhagen City Heart Study Group. Am J Respir Crit Care Med 1996; 153: 1530–35.
- Puhan MA, Hansel NN, Sobradillo P, et al. Large-scale international validation of the ADO index in subjects with COPD: an individual subject data analysis of 10 cohorts. BMJ Open 2012; 2: e002152.
- Esteban C, Quintana JM, Aburto M, et al. The health, activity, dyspnea, obstruction, age, and hospitalization: prognostic score for stable COPD patients. Respir Med 2011; 105: 1662–70.
- Johannessen A, Nilsen RM, Storebo M, Gulsvik A, Eagan T, Bakke P. Comparison of 2011 and 2007 Global Initiative for Chronic Obstructive Lung Disease guidelines for predicting mortality and hospitalization. Am J Respir Crit Care Med 2013; 188: 51–59.
- Leivseth L, Brumpton BM, Nilsen TI, Mai XM, Johnsen R, Langhammer A. GOLD classifications and mortality in chronic obstructive pulmonary disease: the HUNT Study, Norway. Thorax 2013; 68: 914–21
- Frei A, Muggensturm P, Putcha N, et al. Five comorbidities reflected the health status in patients with chronic obstructive pulmonary disease: the newly developed COMCOLD index. J Clin Epidemiol 2014; 67: 90–111.
- Roche N, Deslee G, Caillaud D, et al. Impact of gender on COPD expression in a real-life cohort. Respir Res 2014; 15: 20.
- Garcia-Aymerich J, Gomez FP, Benet M, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. Thorax 2011; 66: 430–37.
- Soler JJ, Sanchez L, Roman P, Martinez MA, Perpina M. Risk factors of emergency care and admissions in COPD patients with high consumption of health resources. Respir Med 2004; 98: 318–29.
- Soler-Cataluña JJ, Martinez-Garcia MA, Roman Sanchez P, Salcedo E, Navarro M, Ochando R. Severe acute exacerbations and mortality in patients with chronic obstructive pulmonary disease. Thorax 2005; 60: 925–31.
- Soler-Cataluña JJ, Martinez-Garcia MA, Sanchez LS, Tordera MP, Sanchez PR. Severe exacerbations and BODE index: two independent risk factors for death in male COPD patients. Respir Med 2009; 103: 692–99.
- Alfageme I, Reyes N, Merino M, et al. The effect of airflow limitation on the cause of death in patients with COPD. Chron Respir Dis 2010; 7: 135–45.
- Casanova C, Cote C, de Torres JP, et al. Inspiratory-to-total lung capacity ratio predicts mortality in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2005; 171: 591–97.
- Almagro P, Calbo E, Ochoa de Echaguen A, et al. Mortality after hospitalization for COPD. Chest 2002; 121: 1441–48.
- Sanjaume M, Almagro P, Rodriguez-Carballeira M, Barreiro B, Heredia JL, Garau J. Post-hospital mortality in patients re-admitted due to COPD. Utility of BODE index. Rev. Clin Esp 2009; 209: 364–70 (in Spanish).
- Almagro P, Salvado M, Garcia-Vidal C, et al. Recent improvement in long-term survival after a COPD hospitalisation. Thorax 2010; 65: 298–302.
- Marin JM, Soriano JB, Carrizo SJ, Boldova A, Celli BR. Outcomes in patients with chronic obstructive pulmonary disease and obstructive sleep apnea: the overlap syndrome. Am J Respir Crit Care Med 2010; 182: 325–31.
- Pillai AP, Turner AM, Stockley RA. Global Initiative for Chronic Obstructive Lung Disease 2011 symptom/risk assessment in alpha1- antitrypsin deficiency. Chest 2013; 144: 1152–62.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland. [2]Servicio de Neumología, Instituto de Investigación del Hospital Universitario de la Princesa (IISP), Universidad Autónoma de Madrid, Madrid, Spain. [3]Epidemiology & Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA.

Received: 24 July 2017 Accepted: 20 November 2017
Published online: 21 December 2017

References

1. Steyerberg EW. Clinical prediction models. In: Gail M, Krickeberg K, Sarnet J, Tsiatis A, Wong W, editors. Statistics for biology and health. New York: Springer; 2010. ISBN: 978-1-4419-2648-7.
2. Alonzo TA. Clinical prediction models: a practical approach to development, validation, and updating (book review). Am J Epidemiol. 2009;170:90033.
3. Khalili D, Hadaegh F, Soori H, Steyerberg EW, Bozorgmanesh M, Azizi F. Clinical usefulness of the Framingham cardiovascular risk profile beyond its statistical performance the Tehran lipid and glucose study. Am J Epidemiol. 2012;176:177–86.
4. Linsell L, Malouf R, Morris J, Kurinczuk JJ, Marlow N. Risk factor models for Neurodevelopmental outcomes in children born very preterm or with very

low birth weight: a systematic review of methodology and reporting. Am J Epidemiol. 2017;185:601–12.

5. Bao W, FB H, Rong S, Rong Y, Bowers K, Schisterman EF. Predicting risk of type 2 diabetes mellitus with genetic risk models on the basis of established genome-wide association markers: a systematic review. Am J Epidemiol. 2013;178:1197–207.
6. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, Lassale CM, Siontis GCM, Chiocchia V, Roberts C, Schlüssel MM, Gerry S, Black JA, Heus P, van der Schouw YT, Peelen LM, Moons KGM. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.
7. Celli BR, Cote CG, Marin JM, Casanova C, Montes de Oca M, Mendez RA, Pinto Plata V, Cabral HJ. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. N Engl J Med. 2004;350:1005–12.
8. Guerra B, Gaveikaite V, Bianchi C, Puhan MA. Prediction models for exacerbations in patients with COPD. Eur Respir Rev. 2017;26:1–13.
9. Siebeling L, Musoro JZ, Geskus RB, Zoller M, Muggensturm P, Frei A, Puhan MA, ter Riet G. Prediction of COPD-specific health-related quality of life in primary care COPD patients: a prospective cohort study. NPJ Prim Care Respir Med. 2014;24:14060. Nature Publishing Group
10. Jones RC, Donaldson GC, Chavannes NH, Kida K, Dickson-Spillmann M, Harding S, Wedzicha JA, Price D, Hyland ME. Derivation and validation of a composite index of severity in chronic obstructive pulmonary disease: the DOSE index. Am J Respir Crit Care Med. 2009;180:1189–95.
11. Wyatt JC, Altman DG. Commentary: prognostic models: clinically useful or quickly forgotten? BMJ. 1995;311:1539–41.
12. Puhan MA, Zoller M, Ter Riet G. COPD: more than respiratory (comment). Lancet. 2008;371:26–7.
13. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130:515–24.
14. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. Am J Epidemiol. 2006;163:783–9.
15. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, L-M Y, Moons KGM, Altman DG. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.
16. Collins GS, Moons KGM. Comparing risk prediction models. Br Med J. 2012; 344:e3186.
17. Cook JA, Collins GS. The rise of big clinical databases. Br J Surg. 2015;102:93–101.
18. Broeze KA, Opmeer BC, Bachmann LM, Broekmans FJ, Bossuyt PM, Coppus SF, Johnson NP, Khan KS, ter Riet G, van der Veen F, van Wely M, Mol BW. Individual patient data meta-analysis of diagnostic and prognostic studies in obstetrics, gynaecology and reproductive medicine. BMC Med Res Methodol. 2009;9:22.
19. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32:3158–80.
20. Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM. Assessing risk prediction models using individual participant data from multiple studies. Am J Epidemiol. 2014;179:621–32.
21. Debray TPA, Riley RD, Rovers MM, Reitsma JB, Moons KGM. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Med. 2015;12:1–12.
22. Vale CL, Rydzewska LHM, Rovers MM, Emberson JR, Gueyffier F, Stewart LA. Uptake of systematic reviews and meta-analyses based on individual participant data in clinical practice guidelines: descriptive study. Br Med J. 2015;350:1–9.
23. Ahmed I, Debray TPA, Moons KGM, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol. 2014;14:3.
24. Ahmed I, Sutton AJ, Riley RD. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. BMJ. 2012;344:1–10.
25. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140.
26. Krumholz HM. Why data sharing should be the expected norm. BMJ. 2015; 350:h599.
27. Marin JM, Alfageme I, Almagro P, Casanova C, Esteban C, Soler-Cataluña JJ, De Torres JP, Martinez-Camblor P, Miravitlles M, Celli BR, Soriano JB. Multicomponent indices to predict survival in COPD: the COCOMICS study. Eur Respir J. 2013;42:323–32. Respiratory Dept, Hospital Universitario Miguel Servet, Zaragoza, Spain Respiratory Dept, Valme University Hospital, Seville, Spain Internal Medicine Unit, Hospital Universitari Mütua de Terrassa, Barcelona, Spain Respiratory Dept, Hospital Nuestra Senora
28. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int J Epidemiol. 1999;28:1–9.
29. Bravata DM, Olkin I. Simple pooling versus combining in meta-analysis. Eval Health Prof. 2001;24:218–30.
30. Higgins J, Whitehead A. Borrowing strength from external trials in a meta-analysis. Stat Med. 1996;15:2733–49.
31. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med. 2002;21:2313–24.
32. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med. 2004;23:3105–24.
33. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc. 2006;101:447–59.
34. Salanti G, Higgins J, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. Stat Methods Med Res. 2008;17:279–301.
35. Lu G, Welton NJ, Higgins J, White IR, Ades AE. Linear inference for mixed treatment comparison meta-analysis: a two-stage approach. Res Synth Methods. 2011;2:43–60.
36. Zarin W, Veroniki AA, Nincic V, Vafaei A, Reynen E, Motiwala SS, Antony J, Sullivan SM, Rios P, Daly C, Ewusie J, Petropoulou M, Nikolakopoulou A, Chaimani A, Salanti G, Straus SE, Tricco AC. Erratum to: characteristics and knowledge synthesis approach for 456 network meta-analyses: a scoping review. BMC Med. 2017;15:61.
37. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, Ioannidis JPA, Straus S, Thorlund K, Jansen JP, Mulrow C, Catala-Lopez F, Gotzsche PC, Dickersin K, Boutron I, Altman DG, Moher D. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med. 2015;162:777–84.
38. Panagiotou OA. Network meta-analysis: evidence synthesis with mixed treatment comparison (book review). Am J Epidemiol. 2015;181:288–9.
39. Takwoingi Y. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med. 2013;158:544-54.
40. Takwoingi Y, Guo B, Riley RD, Deeks JJ. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Stat Methods Med Res. 2015; doi:10.1177/0962280215592269.
41. Manolio TA, Weis BK, Cowie CC, Hoover RN, Hudson K, Kramer BS, Berg C, Collins R, Ewart W, Gaziano JM, Hirschfeld S, Marcus PM, Masys D, Mccarty CA, Mclaughlin J, Patel AV, Peakman T, Pedersen NL, Schaefer C, Scott JA, Sprosen T, Walport M, Collins FS. New models for large prospective studies: is there a better way? Am J Epidemiol. 2012;175:859–66.
42. Soriano JB, Lamprecht B, Ramírez AS, Martinez-Camblor P, Kaiser B, Alfageme I, Almagro P, Casanova C, Esteban C, Soler-Cataluña JJ, De-Torres JP, Miravitlles M, Celli BR, Marin JM, Puhan MA, Sobradillo P, Lange P, Sternberg AL, Garcia-Aymerich J, Turner AM, Han MK, Langhammer A, Leivseth L, Bakke P, Johannessen A, Roche N, Sin DD. Mortality prediction in chronic obstructive pulmonary disease comparing the GOLD 2007 and 2011 staging systems: a pooled analysis of individual patient data. Lancet Respir Med. 2015;3:443–50.
43. Franchini AJ, Dias S, Ades AE, Jansen JP, Welton NJ. Accounting for correlation in network meta-analysis with multi-arm trials. Res Synth Methods. 2012;3:142–60.
44. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Synth Methods. 2012;3:80–97.
45. Hanley JA, McNeil BJA. Method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983;148:839–43.
46. Borenstein M, Hedges LV, Higgins J, Rothstein HR. Introduction to meta-analysis. Chichester: Wiley; 2011.
47. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177–88.
48. Nikolakopoulou A, Mavridis D, Salanti G. How to interpret meta-analysis models: fixed effect and random effects meta-analyses. Evid Based Ment Health. 2014;17:64.
49. Nikolakopoulou A, Mavridis D, Salanti G. Demystifying fixed and random effects meta-analysis. Evid Based Ment Health. 2014;17:53–7.

50. Jansen JP, Cope S. Meta-regression models to address heterogeneity and inconsistency in network meta-analysis of survival outcomes. BMC Med Res Methodol. 2012;12:152.
51. Donegan S, Williamson P, D'Alessandro U, Tudur Smith C. Assessing key assumptions of network meta-analysis: a review of methods. Res Synth Methods. 2013;4:291–323.
52. Cooper NJ, Sutton AJ, Ades AE, Welton NJ. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. Stat Med. 2009;28:1982–98.
53. Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. BMC Med. 2013;11:159.
54. Cipriani A, Higgins JPT, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. Ann Intern Med. 2013;159:130–7.
55. Efthimiou O, Debray TPA, van Valkenhoef G, Trelle S, Panayidou K, KGM M, Reitsma JB, Shang A, Salanti G. GetReal in network meta-analysis: a review of the methodology. Res Synth Methods. 2016;236-63.
56. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins J. Evaluating the quality of evidence from a network meta-analysis. PLoS One. 2014;9:e99682.
57. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172:971–80.
58. Thompson SG, Higgins J. How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002;21:1559–73.
59. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med. 2010;29:932–44.
60. Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, Anzueto A, Barnes PJ, Fabbri LM, Martinez FJ, Nishimura M, Stockley RA, Sin DD, Rodriguez-Roisin R. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease GOLD executive summary. Am J Respir Crit Care Med. 2013;187:347–65.
61. Azarisman MS, Fauzi MA, Faizal MP, Azami Z, Roslina AM, Roslan H. The SAFE (SGRQ score, air-flow limitation and exercise tolerance) index: a new composite score for the stratification of severity in chronic obstructive pulmonary disease. Postgrad Med J. 2007;83:492–7. Department of Medicine, International Islamic University Malaysia, Jalan Hospital Campus, Kuantan, Pahang, Malaysia. risman1973@hotmail.com
62. Boeck L, Soriano JB, Brusse-Keizer M, Blasi F, Kostikas K, Boersma W, Milenkovic B, Louis R, Lacoma A, Djamin R, Aerts J, Torres A, Rohde G, Welte T, Martinez-Camblor P, Rakic J, Scherr A, Koller M, Van Der Palen J, Marin JM, Alfageme I, Almagro P, Casanova C, Esteban C, Soler-Cataluña JJ, De-Torres JP, Miravitlles M, Celli BR, Tamm M, Stolz D. Prognostic assessment in COPD without lung function: the B-AE-D indices. Eur Respir J. 2016;47:1635–44.
63. From the Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2017. Available from: http://goldcopd.org.
64. Puhan MA, Garcia-Aymerich J, Frey M, ter Riet G, Antó JM, Agusti A, Gómez FP, Rodríguez-Roisín R, Moons KGM, Kessels AG, Held U. Expansion of the prognostic assessment of patients with chronic obstructive pulmonary disease: the updated BODE index and the ADO index_Puhan_2009_210. Lancet Elsevier Ltd. 2009;374:704–11.
65. Soler-Cataluña JJ, Martinez-Garcia MA, Sanchez LS, Tordera MP, Sanchez PR. Severe exacerbations and BODE index: two independent risk factors for death in male COPD patients. Respir Med. 2009;103:692–9. Hospital General de Requena, Unidad de Neumologia, Servicio de Medicina Interna, Paraje Casablanca s/n., 46340 Requena, Valencia, Spain. jjsoler@telefonica.net
66. Rubin DB. Multiple imputation for nonresponse in surveys, Wiley series in probability and mathematical statistics. New York: Harvard Univ; 1987.
67. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol. 2010;10:112.
68. Edwards JK, Cole SR, Troester MA, Richardson DB. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. Am J Epidemiol. 2013;177:904–12.
69. Groenwold RHH, Donders ART, Roes KCB, Harrell FE, Moons KGM. Dealing with missing outcome data in randomized trials and observational studies (practice of epidemiology). Am J Epidemiol. 2012;175:210–7.
70. White IR. Network meta-analysis. Stata J. 2015;15:951–85.
71. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. Am J Epidemiol. 2011;173:1327–35.