Interpreting risk reduction in clinical trials for pulmonary arterial hypertension

Annie C. Lajoie[1,2], Sébastien Bonnet[1,2,3], Yves Lacasse[2,3], Jean-Christophe Lega[4,5] and Steeve Provencher[1,2,3]

Affiliations: [1]Pulmonary Hypertension Research Group, Université Laval, Quebec, Canada. [2]Institut universitaire de cardiologie et de pneumologie de Québec Research Center, Université Laval, Quebec, Canada. [3]Dept of medicine, Université Laval, Quebec, Canada. [4]Dept of internal and vascular medicine, Centre Hospitalier Lyon Sud, Pierre-Bénite cedex, France. [5]UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, CNRS, Lyon, France.

Correspondence: Steeve Provencher, Pulmonary Hypertension Research Group, Institut universitaire de cardiologie et de pneumologie de Québec Research Center, Laval University, Quebec, G1V 4G5, Canada. E-mail: steve.provencher@criucpq.ulaval.ca

ABSTRACT Because of scepticism concerning study results when relying solely on relative effect estimates, the number needed to treat (NNT) has been used extensively to quantify the net clinical benefit of an intervention, and is reported increasingly in randomised trials and observational studies. This method is a simple measure representing the number of patients who would need to be treated to prevent one additional adverse event. However, like relative risk, the NNT is an inherently time-dependent measure. Thus, its calculation may lead to misleading interpretations, especially for studies involving varying follow-up times or recurrent outcomes. In addition to study duration and the efficacy of the therapy and the comparator, multiple other factors directly influence the NNT and should be taken into account in its interpretation as for comparative effectiveness of therapies. Its accurate estimation and interpretation, as well as its limitations, are therefore crucial to avoid erroneous clinical and public health decisions. We discuss the calculation and the interpretation of risk reduction and the NNT in the context of the changing landscape of clinical trials in pulmonary arterial hypertension.

Introduction

Throughout the past 20 years, numerous specific pharmacological agents, including phosphodiesterase-5 inhibitors, endothelin receptor antagonists, prostaglandins, soluble guanylate cyclase stimulators and, more recently, selective prostacyclin receptor agonists have emerged for the treatment of pulmonary arterial hypertension (PAH) [1]. During the same period, the landscape of PAH clinical trials changed dramatically. Early clinical trials were typically of short-term duration, comparing the effects of PAH-targeted therapies *versus* placebo and using exercise tolerance as the primary end-point. A meta-analysis of these trials documented a reduction in short-term mortality of ≈40% with monotherapy [2]. More recently, we have witnessed a progressive shift in PAH study designs with longer event-driven trials comparing the effects of combination therapy on clinical worsening that is perceived as a more

clinically relevant outcome measure [3]. A meta-analysis demonstrated that combination therapy significantly reduced the risk of clinical worsening by ≈35% as compared to monotherapy alone [4].

However, many physicians express scepticism when efficacy is presented only as odds or risk ratios (RR) because of the dangers of misinterpreting the importance of a therapy when relying solely on relative effect estimates. For example, a therapy reducing the risk of an event from 3% to 2% or reducing this risk from 30% to 20% would both represent a ~35% relative risk reduction. In an attempt to provide information on the reduction in absolute risk, it is thus not surprising that the number needed to treat (NNT), a global measure representing the number of patients who would need to be treated to prevent one additional adverse event, has been embraced an alternative to express results. Since its description >25 years ago [5] this method has been used extensively to quantify the net clinical benefit of an intervention, as well as for cost comparison and cost-effectiveness analyses. There is indeed an "implicit belief by physicians that the NNT value adequately captures the overall worth of a treatment" [6]. However, NNT and risk ratios are inherently time-dependent measures that few medical students, and even physicians interpret correctly (figure 1) [7, 8].

Indeed, multiple factors, in addition to the efficacy of the therapy and the comparator, may directly influence the NNT and RR and should be taken into account in their interpretation. As a result, the added value of the NNT over the absolute risk reduction has been questioned [9]. Herein, we expand on some of these issues raised by the interpretation of our recent meta-analyses on combination therapies in PAH [4, 10] and discuss the calculation and the interpretation of risk reduction and the NNT in the context of the changing landscape of clinical trials in pulmonary arterial hypertension.

Methods to calculate the NNT

The NNT is calculated by taking the reciprocal of the absolute risk reduction between two treatment options. In practice, it is generally measured by the difference in the cumulative incidence of the outcome over a fixed follow-up time period between two groups of patients as a proxy for treatment efficacy [11]. Using the data from AMBITION study [12] in which 46 of the 253 and 77 of the 247 patients randomised to combination therapy and monotherapy, respectively, met the primary efficacy outcome, it would seem appealing to compute the NNT for preventing one event of clinical failure as:

$$NNT = 1/(\text{proportion of patients with an event in the placebo arm} - \text{proportion of patients with an event in the treated arm})$$

$$NNT = 1/(31.1\% - 18.2\%)$$

$$NNT = 7.7 \approx 8$$

However, this calculation does not account for varying times in individual study arms. Indeed, the NNT has traditionally been derived from single trials measuring dichotomous outcomes with equal follow-up periods for all patients, and its computation must be performed with care in trials involving more complex study designs with varying follow-up periods such as recent long-term event-driven PAH trials. For example, patients in the AMBITION trial were treated on average for 550 days in the combination-therapy group and 484 days in the pooled monotherapy group.

Various methods to calculate the NNT have been proposed for studies with different follow-up times [13]. The incidence rates of the outcome may be used, the NNT being computed as the reciprocal of the between-group difference in incidence density rates (or person-time incidence rates) (table 1). This difference represents the incidence rate of prevented events per person-moment. In the example above, the incidence rates of the primary outcome would be 0.120 *versus* 0.235 events per person-years, yielding an NNT=1/(0.235–0.120)=8.7≈9 per person-years. However, some authors have computed the corresponding NNT with the interpretation that it represents the number of patients who need to be treated to prevent one outcome over a given time period. Interpreting this NNT as the need to treat nine persons for 1 year to prevent an outcome would indeed be inaccurate. Because this NNT is computed from the incidence rate of events, its inverse does not represent persons, but rather person-time. Moreover, this computation assumes that patient-time is intrinsically interchangeable, so that 1 year of follow-up in one patient is equivalent to 6 months of follow-up of two patients, which is frequently inaccurate. Ultimately, doing the same calculation for the other recent long-term trials would then suggest that the addition of selexipag might be associated with the lowest NNT among the evaluated therapies (table 1). Alternatively, the event probability in the active treatment group at a fixed time can be computed from the event probability in the control group at that time and the hazard ratio [2]. While the hazard ratio is an estimate of the instantaneous risk at that specific time rather than a cumulative over an entire study [18], both methods
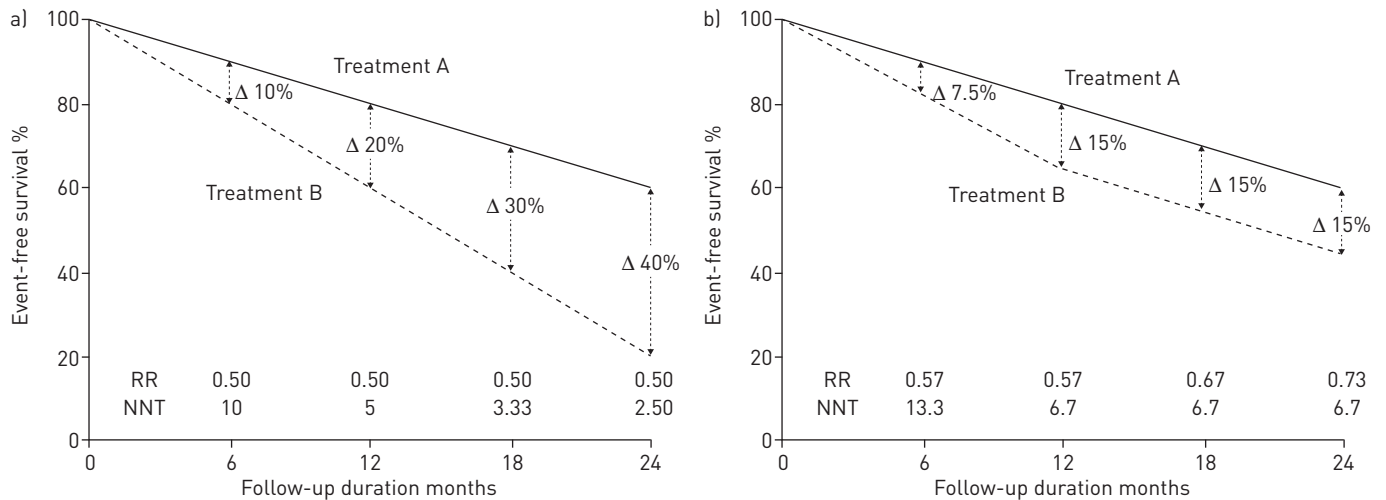
FIGURE 1 Risk ratio (RR) and number needed to treat (NNT) are time-dependent measures. a) When an intervention is associated with constant relative risk reduction over time, the NNT progressively decreases as the number of events increases; b) conversely, when the absolute risk reduction remains constant over time, the relative risk reduction progressively decreases (increased relative risk).

assume that the RR reduction with a therapy is relatively constant over time, which is generally inaccurate. More importantly, physicians often make the mistake of trying to intrapolate or even to extrapolate beyond the original study duration to standardise comparisons between interventions.

Ultimately, the NNT can be estimated using the Kaplan–Meier approach using the difference in the cumulative incidence of the outcome between the two groups at a specific time of follow-up [16]. This method generally accounts for the heterogeneous risk reduction over varying follow-up and represents the number of patients who need to be treated to prevent one patient with the outcome over the period of interest. However, the Kaplan–Meier approach is also subject to distorted NNT calculation, especially when outcomes of interest do not occur randomly. In PAH trials, clinical worsening is a composite end-point generally defined as a combination of death, hospitalisation, lung transplant, treatment escalation and symptomatic progression [3]. While death and PAH-related hospitalisation generally occur erratically, symptomatic progression and treatment escalation, which contributed to 40–80% of the first clinical-worsening event in recent combination trials [12, 14, 15, 17], probably cluster around study visits. This tends to results in inhomogeneous decline in event-free survival on Kaplan–Meier curves (figure 2). Comparing the cumulative incidence of the outcome between the two groups at a specific time of follow-up may thus result in aberrant estimation of the NNT, as probably occurred when estimating the NNT at week 16 from the AMBITION study [12] before clinical worsening was confirmed in a required subsequent study visit (table 1).

The impact of study and treatment duration on NNT and RR: are long-term event-driven trials necessary in PAH?

The impact of time dependency on NNT and RRs is even more problematic when trials of various durations are compared inappropriately. Indeed, NNT and RR are inherently time-dependent measures (figure 1): when the RR reduction is constant over time, increasing follow-up duration will progressively decrease the NNT as the absolute event rate increases, whereas RR will progressively increase (thus decreasing the RR reduction) with longer follow-up when absolute risk reduction is constant over time [4]. For example, clinical trials evaluating the addition of phosphodiesterase type 5 inhibitors resulted in a lower RR compared to endothelin receptor antagonists (RR 0.44, 95% CI 0.31–0.63 versus RR 0.76, 95% CI 0.64–0.90), suggesting that the addition of the former might have reduced clinical worsening more effectively. However, clinical trials evaluating phosphodiesterase type 5 inhibitors were of shorter duration (31±27 weeks versus 90±56 weeks). Interestingly, in our recent meta-analysis, the RR of clinical worsening correlated tightly with study duration in recent trials comparing combination therapy versus monotherapy, whereas the NNT did not [10]. This is consistent with long-term event-driven trials in which the treatment effect on clinical worsening was dominantly observed during the first 6–12 months of treatment. In these trials, the NNT progressively decreased until 52 weeks of follow-up, event-free survival curves being essentially parallel thereafter [12, 14, 15, 17], thus implying a relatively constant absolute risk reduction thereafter [10]. Thus, study duration alone may largely influence differences in RR reduction.

TABLE 1 Methods to estimate number needed to treat (NNT) in recent event-driven clinical trials comparing combination therapy *versus* monotherapy in pulmonary arterial hypertension

	SERAPHIN [14]		AMBITION [12]		GRIPHON [15]	
	CombiTx	**MonoTx**	**CombiTx**	**MonoTx**	**CombiTx**	**MonoTx**
Study characteristics						
Duration weeks	104[+]	85[+]	79	69	71[§]	64[§]
Patients at risk	154	154	253	247	462	458
Number of events	50	68	46	77	124	183
Cumulative proportion %	32.5	44.2	18.2	31.2	26.8	42.1
Event rates per patient-year	0.162	0.269	0.120	0.235	0.197	0.344
NNT calculations						
Cumulative incidence	8.6≈9		7.7≈8		6.5≈7	
Incidence rate[#] patient-years	9.4≈10		8.7≈9		6.8≈7[f]	
HR[¶] (estimated at month 24)	7.4≈8		6.1≈7		6.7≈7	
Kaplan–Meier						
At week 16	10.5≈11		33.3≈34		12.2≈13	
At month 6	7.9≈8		12.0≈12		10.4≈11	
At month 12	6.6≈7		7.1≈8		8.0≈8	
At month 24	6.0≈6		6.1≈7		7.5≈8	

Using the Kaplan–Meier method, the NNT appears to be somewhat comparable across studies, especially at 12 months and beyond, where the NNT is relatively constant, implying a relatively constant absolute risk reduction. Note that in the AMBITION trial, the NNT estimation at week 16 is probably distorted because clinical worsening did not occur randomly, clustering around study visits. Per consensus, the NNT is most frequently rounded up to the nearest whole number. Moreover, although only mean NNT are presented, the 95% confidence intervals around the NNT, specifying the "limits" within which we can confidently state the true NNT lies, should be presented in index articles when the NNT is presented [16]. Finally, the results of the COMPASS trial [17] are not computed in the table, since NNT calculation is controversial in the absence of statistically significant results. CombiTx: combination therapy; monoTx: monotherapy; HR: hazard ratio. [#]: the incidence rate of the outcome may be used as a measure of outcome frequency. It is computed as the number of patients with the outcome divided by the total amount of person-time or patient-time, to account for varying follow-up times. This NNT represents the number of patient-years to prevent one clinical worsening event. [¶]: the event probability in the active treatment group at a fixed time can be computed from the event probability in the control group at that time and the HR. If at some specified time, t, the survival probability in the control group is $S_c(t)$ then the survival probability in the active group is $[S_c(t)]^{HR}$ [14]. The NNT is estimated as $1/\{[S_c(t)]^{HR}-S_c(t)\}$. As example, the AMBITION trial provides an HR of 0.50. At 104 weeks, ≈63% of patients did not meet the primary end-point (figure 2). Then, NTT = $1/[0.63^{0.50}-0.63]$ = 6.1≈7. [+]: mean duration for the overall study population, whereas event rates and NNT are estimated for the subgroup of patients on background therapy only (*i.e.* combination therapy *versus* monotherapy). [§]: median duration of follow-up. [f]: assuming that median follow-up duration approaches mean follow-up duration.

This concept is important in the PAH field, which has witnessed a shift in study design and duration within the past two decades.

More pragmatically, this constant absolute risk reduction after 6–12 months of follow-up questions the requirement for long-term event-driven trials in PAH, especially that morbidity events occurring at 3, 6 and 12 months were recently shown to reliably predict subsequent deaths [19]. Indeed, most recent event-driven studies lasted 4–6 years, patients being exposed to the study drugs on average for ≈2 years [12, 14, 15, 17]. In the context of an orphan disease with limited recruitment and the rapidly changing treatment paradigm in PAH, the optimal duration of future trials should be revisited, balancing study power (longer follow-up will probably be associated with increased numbers of events) with the possibility for patients to contribute to subsequent trials and benefit from newer PAH-target therapies and treatment algorithms [1].

Baseline risk of events, comparison between interventions and external validity of the NNT and RRs

Even with similar relative efficacy of drug therapies, the NNT varies inversely with baseline risk of experiencing an adverse event [20]. Contemporary trial participants generally have lower baseline risk than those enrolled in earlier trials. Novel therapies may thus be associated with larger NNTs, even if they are similarly effective. Consequently, the calculated NNT is entirely specific to a single comparison in a particular study population and NNT cannot be generalised to a particular subgroup of patient, therapy,
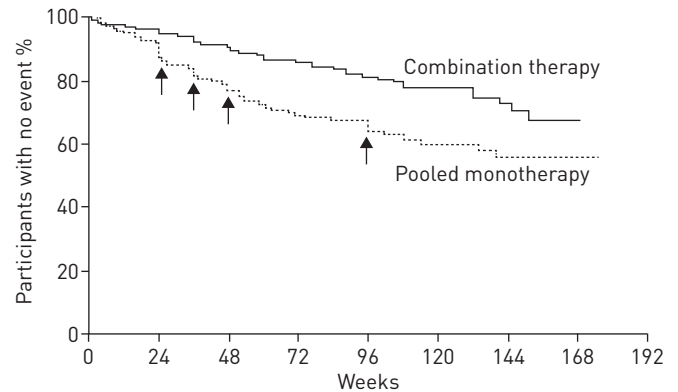
FIGURE 2 Kaplan–Meier curves for the probability of a first adjudicated primary end-point event in the AMBITION trial, suggesting that the primary outcome events frequently clustered around study visits. Compared to death, pulmonary arterial hypertension-related hospitalisation and transplantation which generally occur erratically, symptomatic progression and treatment escalation, which contributed to 40–80% of the first clinical-worsening events in recent event-driven trials, are defined by the investigator, clustering around study visits. This may contribute to the inhomogeneous decline in event-free survival on Kaplan–Meier curves (arrows) and the distorted number needed to treat calculations when estimated at a specific time of follow-up. Reproduced and modified from [12] with permission from the publisher.

drug class or treatment strategy. Therefore, comparisons for different therapies based on NNT is most commonly misleading, unless therapies were tested in study populations with the same stage of disease, comorbidities and background therapies, for their effects on the same outcomes, against the same comparator and over the same time frame. Even in these circumstances, confidence intervals for the NNT are commonly not reported, precluding valid comparison between various randomised controlled trials [9, 16]. Similarly, generalising an NNT from a particular trial to routine care may lead to erroneous conclusions due to selection bias, enhanced comorbidities, lower intensity of monitoring by physicians, lower levels of adherence and higher use of co-interventions.

Outcomes overlooked by the NNT and RRs: treatment impact or utility?

The NNT is generally expressed for binary primary outcomes (*e.g.* clinical worsening). However, most therapies impact more than one predefined outcome. Yet, secondary or continuous outcome measures (*e.g.* exercise capacity) may still be relevant to PAH patients. Of note, a method for estimating the proportion of patients who benefit from a treatment and its NNT when the outcome is a continuous variable has been suggested [21]. Similarly, the NNT calculation using time-to-first event analyses overlooks recurrent events that can occur more than once during the patient follow-up, such as hospitalisation. Thus, the calculation of the NNT to prevent one clinical worsening event in the AMBITION trial [12] does not mean that only one out of between six and nine patients benefited upfront combination therapy *versus* monotherapy.

Conversely, the NNT to "prevent" clinical worsening may be misleading, as the NNT best applies to acute conditions without long-term sequelae. In chronic conditions such as PAH, adverse outcomes are not permanently avoided, but are merely postponed [22]. Thus, it has been suggested that it may be more accurate to describe the potential impact of these chronic therapies in terms of average duration of life without clinical worsening gained rather than focusing on differential event-free survival at a single time point [23]. For these reasons, the NNT is an expression of the frequency of a specific outcome in a specific study population treated with a specific therapy and comparator, rather than a surrogate for the intervention's utility. Patients, physicians and eventually regulatory authorities vary their treatment decisions depending on cost, side-effect profile, the outcome it prevents and personal/societal values. Thus, NNT alone is insufficient to fully capture the global benefit of a therapy and conclude whether a therapy should be used, and other considerations need to be incorporated into treatment decision making [24]. This is of particular importance in PAH where drugs are generally expensive, constraining (*e.g.* parenteral prostaglandins) or associated with potential side-effects.

Conclusions

The NNT is a simple, appealing and valuable measure of the treatment impact. However, its computation and interpretation may be misleading, especially in trials with varying follow-up times. The NNT and RR reduction are time-dependent measures, both decreasing with longer follow-up when relative and absolute

risk reductions are constant over time, respectively. This concept is important in the PAH field, which witnessed a progressive shift in study design and duration. However, even with appropriate computation, the comparison of the NNT and RR between therapies is generally misleading unless therapies were tested in similar study populations with the same disease severity, the same outcomes, against the same comparator and over the same time frame. Even when this fundamental premise is respected, these comparisons are only indirect and subject to artefacts and should therefore be interpreted with extreme caution in the absence of head-to-head clinical trials.

Conflict of interest: None declared.

References

1   Galiè N, Humbert M, Vachiery JL, et al. 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: the Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS). Eur Respir J 2015; 46: 903–975.

2   Galiè N, Manes A, Negro L, et al. A meta-analysis of randomized controlled trials in pulmonary arterial hypertension. Eur Heart J 2009; 30: 394–403.

3   McLaughlin VV, Badesch DB, Delcroix M, et al. End points and clinical trial design in pulmonary arterial hypertension. J Am Coll Cardiol 2009; 54: Suppl. 1, S97–S107.

4   Lajoie AC, Lauziere G, Lega JC, et al. Combination therapy versus monotherapy for pulmonary arterial hypertension: a meta-analysis. Lancet Respir Med 2016; 4: 291–305.

5   Kimura M, Taniguchi H, Kondoh Y, et al. Pulmonary hypertension as a prognostic indicator at the initial evaluation in idiopathic pulmonary fibrosis. Respiration 2013; 85: 456–463.

6   Nathan SD, Shlobin OA, Ahmad S, et al. Serial development of pulmonary hypertension in patients with idiopathic pulmonary fibrosis. Respiration 2008; 76: 288–294.

7   Handa T, Nagai S, Miki S, et al. Incidence of pulmonary hypertension and its clinical relevance in patients with sarcoidosis. Chest 2006; 129: 1246–1252.

8   Bright-Thomas RJ, Ray SG, Webb AK. Pulmonary artery pressure in cystic fibrosis adults: characteristics, clinical correlates and long-term follow-up. J Cyst Fibros 2012; 11: 532–538.

9   Hutton JL. Number needed to treat and number needed to harm are not the best way to report and assess the results of randomised clinical trials. Br J Haematol 2009; 146: 27–30.

10   Lajoie AC, Guay CA, Lega JC, et al. Trial duration and risk reduction in combination therapy trials for pulmonary arterial hypertension: a systematic review. Chest 2018; 153: 1142–1152.

11   Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988; 318: 1728–1733.

12   Galiè N, Barberà JA, Frost AE, et al. Initial use of ambrisentan plus tadalafil in pulmonary arterial hypertension. N Engl J Med 2015; 373: 834–844.

13   Suissa D, Brassard P, Smiechowski B, et al. Number needed to treat is incorrect without proper time-related considerations. J Clin Epidemiol 2012; 65: 42–46.

14   Pulido T, Adzerikho I, Channick RN, et al. Macitentan and morbidity and mortality in pulmonary arterial hypertension. N Engl J Med 2013; 369: 809–818.

15   Sitbon O, Channick R, Chin KM, et al. Selexipag for the treatment of pulmonary arterial hypertension. N Engl J Med 2015; 373: 2522–2533.

16   Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. BMJ 1999; 319: 1492–1495.

17   McLaughlin V, Channick RN, Ghofrani HA, et al. Bosentan added to sildenafil therapy in patients with pulmonary arterial hypertension. Eur Respir J 2015; 46: 405–413.

18   Spruance SL, Reid JE, Grace M, et al. Hazard ratio in clinical trials. Antimicrob Agents Chemother 2004; 48: 2787–2792.

19   McLaughlin VV, Hoeper MM, Channick RN, et al. Pulmonary arterial hypertension-related morbidity is prognostic for mortality. J Am Coll Cardiol 2018; 71: 752–763.

20   Alter DA, Manuel DG, Gunraj N, et al. Age, risk-benefit trade-offs, and the projected effects of evidence-based therapies. Am J Med 2004; 116: 540–545.

21   Guyatt GH, Juniper EF, Walter SD, et al. Interpreting treatment effects in randomised trials. BMJ 1998; 316: 690–693.

22   Tan LB, Murphy R. Shifts in mortality curves: saving or extending lives? Lancet 1999; 354: 1378–1381.

23   Lubsen J, Hoes A, Grobbee D. Implications of trial results: the potentially misleading notions of number needed to treat and average duration of life gained. Lancet 2000; 356: 1757–1759.

24   Andersen KH, Iversen M, Kjaergaard J, et al. Prevalence, predictors, and survival in pulmonary hypertension related to end-stage chronic obstructive pulmonary disease. J Heart Lung Transplant 2012; 31: 373–380.