Received: 13 October 2020 | Revised: 12 December 2020 | Accepted: 1 January 2021

DOI: 10.1002/bimj.202000305

Biometrical Journal

DISCUSSION PAPER

On the logic of collapsibility for causal effect measures

Vanessa Didelez[1,2] | Mats Julius Stensrud[3]

[1] Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany

[2] Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany

[3] Department of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Correspondence
Vanessa Didelez, Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany.
Email: didelez@leibniz-bips.de

KEYWORDS
causal inference, collapsibility, confounding, survival analysis

Liu et al. (2020) discuss the relation between efficacy measures within subgroups and efficacy measures on the population level, which can be obtained by merging the subgroups. They come to the conclusion that neither odds ratios (for binary endpoints) nor hazard ratios (for time-to-event endpoints) are suitable measures of efficacy in this context. This insight is not new, and more general settings have been considered previously (Daniel, Zhang, & Farewell, 2020; Greenland & Pearl, 2011; Greenland, Robins, & Pearl, 1999; Huitfeldt, Stensrud, & Suzuki, 2019; Martinussen & Vansteelandt, 2013; Pang, Kaufman, & Platt, 2013; Sjölander, Dahlqwist, & Zetterqvist, 2016). While we largely agree with their conclusion, we do so for different reasons and would like to point out a number of important subtleties that have perhaps not been appreciated by Liu et al. (2020). These should be carefully understood to avoid any further misleading interpretations. In particular, we want to emphasise, like many before, that confounding and non-collapsibility are separate issues (Didelez et al., 2010; Greenland, 1996; Greenland & Pearl, 2011; Greenland et al., 1999; Pand, Kaufman, & Platt, 2013; Pang et al., 2013; Shrier & Pang, 2015); to cite Greenland (2011): ‘confounding may occur with or without non-collapsibility, and non-collapsibility may occur with or without confounding’. Moreover, in view of patients and investigators preferring contrasts in terms of absolute risks (Murray, Caniglia, Swanson, Hernández-Díaz, & Hernán, 2018), we are sceptical about the emphasis on relative median survival time proposed in Liu et al. (2020).

Notation, terminology and set-up

We adapt the notation of Liu et al. (2020). Let Y denote the binary or time-to-event outcome (or endpoint) with distribution $P(Y)$, let $T \in \{Rx, C\}$ be the binary treatment indicator and let $X \in \{g^-, g^+\}$ be the indicator for the two subgroups. We suppress the index i denoting the individual as it is not relevant to the key issues.

We use $A \perp\!\!\!\perp B \mid C$ and B are conditionally independent of each other given C (Dawid, 1979).

Collapsibility of associational measures of dependence

Collapsibility has been investigated from many different angles (see Greenland, 2011 and references therein). Let $\phi = \phi(Y, T)$ be a measure of association between Y and T; that is, ϕ is a functional of the joint $P(Y, T)$ (in particular, ϕ could be a parameter of $P(Y \mid T)$). Let $\phi_x = \phi(Y, T \mid X = x)$ be a measure of conditional association between Y and T given $X = x$; that is, ϕ_x is a functional of the conditional distribution $P(Y, T \mid X = x)$ (in particular, ϕ_x could be a parameter of $P(Y \mid T, X = x)$). The measure ϕ is called *collapsible* over X if ϕ is a weighted average of the $\phi_x, x \in \mathcal{X}$ (Greenland, 2011; Huitfeldt et al., 2019). Note that a collapsible measure is thus 'logic-respecting' as defined in Liu et al. (2020). When $\phi_x = \phi_{x'}$, for all x, x', *strict* collapsibility demands that $\phi = \phi_x$ (Greenland et al., 1999). In the context of odds-ratios 'collapsibility' is often used to mean 'strict collapsibility'.

The parameters ϕ or ϕ_x are measures of association, which may or may not have a causal interpretation. It is well known that the regression coefficient in a linear regression of Y on T alone, versus on (T, X) jointly is not collapsible when T and X are correlated, that is, when X is unbalanced in the treatment arms. As a further example consider the associational risk-difference $\phi^{RD} = P(Y = 1 \mid T = Rx) - P(Y = 1 \mid T = C)$. Note that

$$\phi^{RD} = \sum_x [P(Y = 1 \mid T = Rx, X = x)P(X = x \mid T = Rx)) \tag{1}$$

$$(-P(Y = 1 \mid T = C, X = x)P(X = x \mid T = C)]\,. \tag{1}$$

Hence, it can easily be seen that the marginal associational risk difference ϕ^{RD} equals a weighted average of $\phi_x^{RD} = P(Y = 1 \mid T = Rx, X = x) - P(Y = 1 \mid T = C, X = x)$ with weights $P(X = x \mid T = Rx)$ under the conditions that $Y \perp\!\!\!\perp X \mid T = C$ and $P(X = x \mid T = C)$ can be completely different distributions and ϕ^{RD} is not a weighted average of ϕ_x^{RD}, as illustrated by the example of tables 2 and 3 in Liu et al. (2020). Note that what the authors calculate as the 'true' marginal efficacies in these tables are really the crude (marginal) associations and not the causal efficacy (see below).

In both cases considered above (linear regression and risk differences), $T \perp\!\!\!\perp X$ is balanced in the treatment arms), and this independence is implied when T is randomised and causal effects of T are identified. However, we do not generally expect collapsibility otherwise. These points hint at a relation between collapsibility and causal inference. Causally interpretable measures are also what Liu et al. (2020) are interested in, as indicated by the title of their paper. Thus, we will discuss the issue of collapsibility from a causal point of view and argue that a formal consideration of causal concepts is crucial.

1 | A CAUSAL POINT OF VIEW

Causal contrasts, comparing a treatment versus control, are commonly expressed in terms of potential outcomes or, more generally, their distributions. Thus, $Y(Rx)$ is the outcome if a person is assigned to treatment and $Y(C)$ is the outcome if the same person is assigned to control. Note that this notation is oblivious as to whether the data are obtained form an observational study or a randomised controlled trial (RCT), and we use it to be formal and explicit about estimands and assumptions. In other scientific communities this is also expressed as interventional distributions such that $P(Y = y; \mathrm{do}(T = Rx))$ stands for the distribution of the endpoint under an intervention that sets treatment to Rx, and similar for controls (Pearl, 2009). For our purposes we can regard $P(Y(t) = y)$ as equal to $P(Y = y; \mathrm{do}(T = t))$.

1.1 | Causal collapsibility

Liu et al. (2020) state that their interest is 'true effects'. While they do not provide a formal definition of 'true effects', we assume that they refer to some causal notion and its population as opposed to estimated value. To *define* causal effects it is immaterial whether X is balanced between treatment arms, though of course, estimation may be affected by such an imbalance. Let $f(P)$ be a summary measure of a distribution, for example, its mean or the odds. We denote this summary for the subgroups g^+, g^- in each treatment arm as

$$\mu_x^t = f(P(Y(t) \mid X = x)), \quad t \in \{Rx, C\},\ x \in \{g^+, g^-\}.$$

The above can be interpreted as applying Rx (or C) to each subgroup and then summarising the distribution using f separately by subgroups. Similarly, over the entire patient population (or marginally) we have

$$\mu^t = f(P(Y(t)), \quad t \in \{Rx, C\}.$$

This can be interpreted as applying Rx (or C) to the entire patient population and then summarising the distribution using f (as in Liu et al. (2020), start of section 3). Further, let μ_x denote a contrast of μ_x^{Rx} versus μ_x^C, for example, the difference or ratio; and let $\mu_{\{g^+,g^-\}}$ denote the corresponding contrast of μ^{Rx} versus μ^C. In other words, μ_x is the causal counterpart to ϕ_x and $\mu_{\{g^+,g^-\}}$ to ϕ, respectively. In this context, we define *collapsibility of a causal measure* over X if it holds that $\mu_{\{g^+,g^-\}}$ is a weighted average of μ_{g^+} and μ_{g^-} (Huitfeldt et al., 2019).

For instance, $\mu_{\{g^+,g^-\}} = P(Y(Rx) = 1) - P(Y(C) = 1)$ is the causal population risk difference while $\mu_{g^+} = P(Y(Rx) = 1 \mid X = g^+) - P(Y(C) = 1 \mid X = g^+)$ is the causal risk difference in subgroup g^+. Note that by properties of conditional probabilities

$$\mu_{\{g^+,g^-\}} = \sum_{x \in \{g^+,g^-\}} [P(Y(Rx) = 1 \mid X = x) - P(Y(C) = 1 \mid X = x)]P(X = x). \tag{2}$$

When comparing (1) and (2), it becomes clear why we need to distinguish associational measures from causal effect measures: an intervention that fixes treatment to either Rx or C 'eliminates' the dependence between treatment and baseline covariates such as X. Hence, causal collapsibility is not the same as associational collapsibility.

It can easily be seen that the causal risk difference and causal risk ratio are collapsible effect measures while the causal odds ratio is not. Huitfeldt et al. (2019) show that the marginal causal risk ratio (or relative response in Liu et al. (2020)) is a weighted average of subgroup causal risk ratios with weights $P(X = x \mid Y(C) = 1)$ (see also Miettinen, 1972).

1.1.1 | Prognostic and predictive biomarkers

Liu et al. (2020) distinguish prognostic and predictive properties of a biomarker X. For our considerations here and for the causal claims made by the authors, we want to stress that the biomarker needs to be *pre-treatment* (or baseline), that is, known to not be affected by treatment. The causal assumptions detailed below do not generally hold when conditioning on any post-treatment variables which may be on some causal pathway from T to Y.

Under this premise, we interpret the definition in Liu et al. (2020) of a biomarker X to be 'treatment-effect prognostic' as $Y(Rx) \not\perp\!\!\!\perp X$, deviate from the prevalence of the biomarker $\gamma^x = P(X = x)$, when it is prognostic for the controls.

In contrast, a biomarker is predictive if $\mu_{g^+} \neq \mu_{g^-}$, that is, it is an effect modifier on the chosen scale. It is worth mentioning that effect modification depends on the scale, for example, if X is not predictive on the additive scale (risk difference) it will be predictive on the multiplicative scale (risk ratio). We recommend VanderWeele and Robins (2007) and VanderWeele (2015) for further insights into the causal interpretation of statistical interactions and their distinction form causal effect modification.

1.2 | Structural assumptions

Inference on causal effects from any data always relies on certain structural assumptions (in addition to, say, parametric or other modelling assumptions). These structural assumptions refer to the relation between potential outcomes and observables and are the basis for using the observed data for inference on causal effects. Some of these assumptions can easily be derived from causal diagrams representing prior assumptions on the causal structure (Greenland & Pearl, 2011; Pearl, 2009). As we will explain, the key structural assumptions are often easier to justify when the data result from an RCT.

The first assumption is that of *positivity* demanding that $0 < P(T = Rx \mid X = x) < 1$; this is always satisfied in an RCT at least for the population eligible for the trial. Further, note that the potential outcomes $Y(Rx)$ and $Y(C)$ can never be observed jointly (one of them will be 'counterfactual'); however under the assumption of *causal consistency*, we can at least observe one of the potential outcomes. This states that when $T = t$, then we observe $Y = Y(t)$. Hence, under consistency we have for the observable outcome Y:

$$Y = Y(Rx)I\{T = Rx\} + Y(C)I\{T = C\},$$

where $I\{\cdot\}$ is the indicator function. There is no need for an upper index on Y as used in Liu et al. (2020). Also note that in the standard RCT context, consistency will typically hold. It might be violated if the administration of treatment in the trial (under special medical supervision, say) is extremely different, in a way that substantially affects the outcome, from how it would ever be administered in real life.

1.3 | Collapsibility and confounding

Besides positivity and consistency, a key assumption is that of ‘ignorability’ (aka ‘exchangeability’ or ‘no unmeasured confounding’).

1.3.1 | Ignorability

We have *ignorability* if $Y(t) \perp\!\!\!\perp T$. This can be interpreted as T being independent of any baseline characteristics that predict the potential outcomes (this ensures covariate balance in expectation). It is easily seen that ignorability (with consistency) implies

$$P(Y(t) = y) = P(Y(t) = y \mid T = t) = P(Y = y \mid T = t).$$

The above equality proves non-parametric identifiability of any marginal (i.e. population) causal measure of efficacy in an RCT. In particular it implies that, in an RCT, the distribution of Y in the treatment arm can be seen as a sample from the distribution of $Y(Rx)$, and the control arm as a sample from the distribution of $Y(C)$. Note that like always with finite samples, especially when small, these can happen to be ‘bad’ samples due to sampling variability, but they will not systematically be so. An accidental imbalance of 1/3 versus 2/3 as considered in table 3 of Liu et al. (2020) is unlikely in sufficiently large RCTs.

1.3.2 | Conditional ignorability

A different assumption, which Liu et al. (2020) do not clearly distinguish from ignorability, is that of *conditional ignorability* given X: This demands that $Y(t) \perp\!\!\!\perp T \mid X$. Note that this is not the same as (unconditional) ignorability. Conditional ignorability is often assumed in the context of observational studies where $X \not\!\perp\!\!\!\perp T$ does not balance baseline covariates), where X is a set of covariates (not only a biomarker) that capture all confounding between T and Y.

1.3.3 | Randomisation

Under *randomisation* of T we have the even stronger property that $(Y(t), X) \perp\!\!\!\perp T$, and this implies both ignorability and conditional ignorability. Under randomisation of T (or under conditional ignorability) we have that

$$P(Y(t) = y \mid X = x) = P(Y(t) = y \mid T = t, X = x) = P(Y = y \mid T = t, X = x).$$

This proves that under randomsiation of T any conditional (i.e. subgroup) causal measure of efficacy is non-parametrically identified. In the case where (Y, T, X) are binary, no further assumptions are required. Causally interpretable marginal or conditional risk differences, risk ratios or odds ratios can consistently be estimated; also note that under independent censoring, marginal and conditional survival curves can consistently be estimated, for example, by the Kaplan Meier estimator.

1.3.4 | Confounding

The amount of confounding, say in an observational study, is sometimes measured by comparing an estimate of the simple associational measure ϕ with an estimate of the causal marginal effect μ, where μ is obtained by suitable adjustment such as standardisation or inverse probability of treatment weighting (IPTW) (Greenland et al., 1999). Essentially, this boils down to comparing whether (a summary of) $P(Y = y \mid T = t)$ is different from (the same summary of) $P(Y(t) = y)$, though one would need to exclude other sources of structural bias as well.

The assumption of ignorability thus implies that there is no confounding of the effect of T on Y; the assumption of conditional ignorability given X implies that there is no confounding, other than possibly by X, of the effect of T on Y.

Under either assumption, and in particular under randomisation of T any of the above quantities μ_x and $\mu_{\{g^+,g^-\}}$ can consistently be estimated from an RCT as they are known functions of $P(Y(t) = y)$ or $P(Y(t) = y \mid X = x)$, for $t = Rx, C$, and these can be obtained by the observable $P(Y = y \mid T = t)$ or $P(Y = y \mid T = t, X = x)$, respectively.

Moreover, the relation between any marginal and conditional measures is mathematically determined: We have

$$P(Y(t) = y) = \sum_x P(Y(t) = y \mid X = x)P(X = x), \tag{3}$$

which follows the subgroup mixable principle as stressed by Liu et al. (2020). The marginal causal odds ratio given by

$$MCOR = \frac{P(Y(Rx) = 1)P(Y(C) = 0)}{P(Y(Rx) = 0)P(Y(C) = 1)}$$

can be re-expressed in terms of the probabilities in the subgroups via (3); all of these quantities can consistently be estimated with data from an RCT or with data on (Y, T, X) from an observational study if conditional ignorability holds (see, e.g. Zhang, 2008). The same holds for the subgroup causal odds-ratios. The fact that the marginal causal odds ratio is not a weighted average of the subgroup causal odds ratios, that is, its non-collapsibility, pertains to its mathematical properties and has nothing to do with absence or presence of confounding (Greenland, 1996, 2011; Greenland et al., 1999; Hernán et al., 2011).

1.3.5 | Adjustment

In section 3.1 Liu et al. (2020) address adjustment for imbalances. Their motivation for this is somewhat unclear as they otherwise consider RCTs. Indeed, we do not consider it of much importance that, in RCTs, small sample sizes can result in random imbalances – ‘confounding’ is a source of *structural* bias which does not go away with increasing sample size, but such finite-sample imbalances under randomisation are not sources of bias, they are the random variation we expect with small samples.

In the context of subgroups, one considers conditional effects. As emphasised by Daniel et al. (2020), ‘adjusting’ and ‘conditioning’ should not be confused. Adjusting typically means that the analysis takes observed confounders into account. This can be achieved by a number of different methods, such as regression adjustment, stratification, matching or IPTW (see, e.g. Goetghebeur, le Cessie, De Stavola, Moodie, & Waernbaum, 2020, for a review). Of these, regression and stratification use the principle of conditioning. However, when the aim is to estimate a marginal causal effect, then additional standardising, on the correct scale, with respect to the confounder distribution is required. Liu et al. (2020) are instead interested in subgroup-specific measures of efficacy, that is, they are conditional on the subgroup indicator. Their reason for the conditioning, in an RCT, is thus a different one than adjustment.

2 | TIME-TO-EVENT ENDPOINTS

While it has been known (at least) since the 1970s that the odds ratio is not collapsible (Whittemore, 1978), it has taken a little longer to appreciate the corresponding problem with hazard or rate ratios (Aalen, Cook, & Røysland, 2015; Greenland, 1996; Sjölander et al., 2016). The issue with rate (hazard) contrasts is more subtle than for odds ratios, and we want to address two particular aspects here.

2.1 | Non-collapsibility of rate/hazard differences/ratios

It is curious that while (causal) risk differences and risk ratios are collapsible, rate (hazard) differences and rate (hazard) ratios are not. The key here is that rates (hazards) are based on conditional probabilities, namely, conditional on prior survival. A rate (hazard) can be converted into a risk, the probability of an event before a given time, through a non-linear transformation which ‘destroys’ the collapsibility of risk differences and risk ratios, as described in Daniel et al. (2020). As with odds ratios, these phenomena occur under ignorability or randomisation, and are therefore not linked to, or indicative of, any confounding.

However, additive hazard differences (in continuous time) satisfy the special condition of *strict collapsibility* (Daniel et al., 2020); in particular, when an additive Aalen model with no interaction terms fits the data-generating mechanism during the entire follow-up (which is a strong parametric restriction), then the marginal and conditional hazard differences are equal (Daniel et al., 2020; Sjölander et al., 2016). This strict collapsibility does not imply that the hazard difference is a collapsible effect measure in general (see also Section 2.4).

2.2 | 'The hazards of hazard ratios'

The title of this subsection is a quote from Hernán (2010), who explains why hazard ratios are problematic as measures of causal contrasts. The issues pertain to the differential effects in (possibly latent) subgroups. Indeed, the causal interpretation of hazard ratios as well as other hazard-based contrasts is ambiguous, as we explain next (Martinussen, Vansteelandt, & Andersen, 2020; Stensrud, Aalen, Aalen, & Valberg, 2019a; Stensrud & Hernán, 2020).

As mentioned in Section 2.1, rates and hazards are based on conditional probabilities, conditioning on prior survival. Using potential outcomes notation, this means that we consider

$$P(t \leq Y(j) < t + h \mid Y(j) \geq t), \quad j \in \{Rx, C\},$$

where $Y(j)$ is the time-to-event when assigned to treatment arm $j \in \{Rx, C\}$. The hazard ratio is therefore

$$\frac{P(t \leq Y(Rx) < t + h \mid Y(Rx) \geq t)}{P(t \leq Y(C) < t + h \mid Y(C) \geq t)}.$$

This hazard ratio, which is the target of inference in most randomised trials with time-to-event outcomes, is problematic because a causal contrast should compare the *same population* under the scenarios of treatment versus control. However, the numerator of the hazard ratio is a probability for the 'population' characterised by $\{Y(Rx) \geq t\}$ while the denominator gives a probability for 'population' $\{Y(C) \geq t\}$ – these can be very different groups.

To explain the distinction between conditioning on $\{Y(Rx) \geq t\}$ and $\{Y(C) \geq t\}$, assume Z is a latent frailty which interacts with treatment. While Z is balanced at baseline due to randomisation of T, the treated individuals who survive a given time t may tend to be more 'frail' (if treatment is beneficial and lowers mortality) than the controls who survive the given time t (Hernán, 2010). Thus, as t increases, a contrast of hazards compares possibly increasingly differing groups of 'survivors'.

2.3 | Testing the null hypothesis

A special case deserves attention, especially in the context of clinical trials where testing the null hypothesis of 'no effect' is often described as the prime interest. When $Y \perp\!\!\!\perp T \mid X$, then the causal odds ratio and hazard ratio *are* both strictly collapsible. In other words, we do not have to worry about the validity of statistical tests as the significance level is preserved; in an RCT, the marginal and the subgroup odds/hazard ratios all equal 1 under the null hypothesis. The issues discussed by Liu et al. (2020) are therefore relevant when the central aim is to quantify efficacy, which is more meaningful than hypothesis testing in many (if not most) practical settings. However, when null hypotheses are of interest, these should be formulated in terms of survival probabilities instead of hazard ratios (Stensrud et al., 2019a; Stensrud, Røysland, & Ryalen, 2019b).

2.4 | Collapsibility of measures or models?

Collapsibility can be defined 'non-parametrically' as a property of a measure of efficacy (Huitfeldt et al., 2019). Non-parametric estimation is also feasible when we consider a restricted number of discrete variables. However, when extending the considerations to a continuous biomarker X (which is not dichotomised), say, then models are used to impose some smoothness including smoothness over time for time-to-event endpoints. For example, the logistic link ensures that the probability remains inside (0,1), and the Cox model ensures the hazard remains positive (which the Aalen additive hazards model does not ensure). A measure of efficacy that is non-collapsible in general, can be collapsible under certain (restrictive) parametric model assumptions: for instance, the strict collapsibility of the hazard difference implies that the Aalen model without interaction terms is collapsible. However, this is a special case: for instance, if a simple Cox model without interaction terms fits the data-generating mechanism, then the hazard difference is no longer collapsible.

3 | CONCLUSIONS

To summarise we would like to emphasise the following points.

For general clarity and to avoid misunderstandings, associational concepts of dependence should clearly and formally be distinguished from causal contrasts (or causal measures of efficacy). Formal frameworks and notation to do so have been available for over 40 years and keep being refined (Pearl, 2009; Robins, 1986; Rubin, 1974).

As explained above, and as has been pointed out many times before, confounding and non-collapsibility are separate issues and should be kept apart. It is not correct that the odds ratio or the hazard ratio somehow re-introduce confounding into an RCT as claimed by Liu et al. (2020). An RCT *does* guarantee 'no confounding' and this is not a 'belief' but a mathematical fact, which does not need amending. This has nothing to do with the choice of efficacy measure. Of course, RCTs can suffer from other problems, for instance, due to non-adherence or other intercurrent events.

We agree with Liu et al. (2020) that (causal) odds ratios and hazard ratios are problematic as causal contrasts. The non-collapsibility of these parameters is a mathematical property which makes their interpretation awkward, and this is amplified for hazards by their conditioning on survival. Thus, they are also unsuitable measures for transportability between different populations (Martinussen & Vansteelandt, 2013). It is particularly concerning that meta-analyses pool odds ratios or hazard ratios from different studies each possibly using different variables for adjustment where the issue of non-collapsibility is typically ignored. Careful causal considerations are in general required for transportability (see, e.g. Bareinboim & Pearl, 2013; Dahabreh, Robins, Haneuse, & Hernán, 2019).

However, a measure being collapsible does not automatically make it meaningful. Hazard differences derived from an additive Aalen model, for instance, are not easy to interpret and not really useful for decision making. In fact, there is empirical evidence that patients and investigators prefer contrasts in terms of absolute risk (Murray et al., 2018). This makes us sceptical whether the proposed ratio of median survival times will take hold; moreover, its estimation will often need to rely on parametric assumptions, even in a perfectly executed RCT, for instance, when less than 50% of the individuals experience the outcome during follow-up in either the treatment or control arm. Hence, the causal inference community is moving towards using contrasts of risk in time-to-event context, such as differences between suitably adjusted and standardised survival curves (Hernán & Robins, 2020; Robins, 1986). This has also increasingly been recommended in clinical contexts (Stensrud et al., 2019a; Stensrud & Hernán, 2020; Uno et al., 2014). In an RCT, differences between survival probabilities (marginally or in subgroups) can non-parametrically be estimated without proportional hazards assumption at pre-specified times, such as 1-year, 5-year and 10-year survival probabilities, according to the context. As probabilities (i.e. risks), these parameters have the further advantage of complying with the subgroup mixable effects principle.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Vanessa Didelez https://orcid.org/0000-0001-8587-7706

REFERENCES

Aalen, O., Cook, R., & Røysland, K. (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis*, *21*(4), 579–593.

Bareinboim, E., & Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. arXiv preprint arXiv:1312.7485.

Dahabreh, I. J., Robins, J. M., Haneuse, S. J., & Hernán, M. A. (2019). Generalizing causal inferences from randomized trials: Counterfactual and graphical identification. arXiv preprint arXiv:1906.10792.

Daniel, R., Zhang, J., & Farewell, D. (2020). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, https://doi.org/10.1002/bimj.201900297.

Dawid, A. P. (1979). Conditional independence in statistical theory (with Discussion). *Journal of the Royal Statistical Society, Series B*, *41*, 1–31.

Didelez, V., Kreiner, S., & Keiding, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statistical Science*, *25*(3), 368–387.

Goetghebeur, E., le Cessie, S., De Stavola, B., Moodie, E.E., & Waernbaum, I. (2020). Formulating causal questions and principled statistical answers. *Statistics in Medicine*, *39*, 4922–4948. https://doi.org/10.1002/sim.8741.

Greenland, S. (1996). Absence of cofounfing does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*, *7*(5), 498–501.

Greenland, S. (2011). Collapsibility. M. Lovric *International Encyclopedia of Statistical Science* (pp. 267–270). Berlin, Heidelberg: Springer.

Greenland, S., & Pearl, J. (2011). Adjustments and their consequences—Collapsibility analysis using graphical models. *International Statistical Review*, *79*(3), 401–426.

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*(1), 29–46.

Hernán, M. (2010). The hazards of hazard ratios. *Epidemiology*, *21*(1), 13–15.

Hernán, M., Clayton, D., & Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, *40*(3), 780–5.

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Boca Raton, FL: Chapman & Hill/CRC.

Huitfeldt, A., Stensrud, M. J., & Suzuki, E. (2019). On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging Themes in Epidemiology*, *16*(1), 1.

Liu, Y. I., Wang, B., Yang, M., Hui, J., Xu, H., Kil, S., & Hsu, J. (2020). Correct and logical causal inference for binary and time-to-event outcomes in randomized controlled trials. *Biometrical Journal*, 1–30.

Martinussen, T., & Vansteelandt, S. (2013). On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Analysis*, *19*(3), 279–296.

Martinussen, T., Vansteelandt, S., & Andersen, P. (2020). Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis*, *26*, 833–855.

Miettinen, O. S. (1972). Standardization of risk ratios. *American Journal of Epidemiology*, *96*(6), 383–388.

Murray, E. J., Caniglia, E. C., Swanson, S. A., Hernández-Díaz, S., & Hernán, M. A. (2018). Patients and investigators prefer measures of absolute risk in subgroups for pragmatic randomized trials. *Journal of Clinical Epidemiology*, *103*, 10–21.

Pand, M., Kaufman, J. S., & Platt, R. W. (2013). Mixing of confounding and non-collapsibility: A notable deficiency of the odds ratio. *The American Journal of Cardiology*, *111*(2), 302–303.

Pang, M., Kaufman, J. S., & Platt, R. W. (2013). Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Statistical Methods in Medical Research*, *25*(5), 1925–1937.

Pearl, J. (2009). *Causality* (2nd edn). Cambridge, MA: Cambridge University Press.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—Application to control for the healthy worker survivor effect. *Mathematical Modelling*, *7*, 1393–1512.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Shrier, I., & Pang, M. (2015). Confounding, effect modification and the odds ratio: Common misinterpretations. *Journal of Clinical Epidemiology*, *68*(5), 470–474.

Sjölander, A., Dahlqwist, E., & Zetterqvist, J. (2016). A note on the noncollapsibility of rate differences and rate ratios. *Epidemiology*, *27*(3), 356–359.

Stensrud, M. J., Aalen, J. M., Aalen, O. O., & Valberg, M. (2019a). Limitations of hazard ratios in clinical trials. *European Heart Journal*, *40*(17), 1378–1383.

Stensrud, M. J., & Hernán, M. A. (2020). Why test for proportional hazards? *Journal of the American Medical Association*, *323*(14), 1401–1402.

Stensrud, M. J., Røysland, K., & Ryalen, P. C. (2019b). On null hypotheses in survival analysis. *Biometrics*, *75*(4), 1276–1287.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., … Wei, L.-J. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, *32*(22), 2380–2385.

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford, UK: Oxford University Press.

VanderWeele, T. J., & Robins, J. M. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, *18*(5), 561–568.

Whittemore, A. S. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statististical Society, Series B*, *40*, 328–340.

Zhang, Z. (2008). Estimating a marginal causal odds ratio subject to confounding. *Communications in Statistics - Theory and Methods*, *38*(3), 309–321.

How to cite this article: Didelez V, Stensrud MJ. On the logic of collapsibility for causal effect measures. *Biometrical Journal*. 2022;*64*:235–242. https://doi.org/10.1002/bimj.202000305