Nevertheless, despite any disagreements and puzzlements, I appreciate the contributions of both discussants to the topic of causal inference with surrogate outcomes, and I agree with both that much remains to be done.

References

Cochran, W. G. & Cox, G. M. (1950). *Experimental designs*. Wiley, New York.
Cox, D. R. (1958). *Planning of experiments*. Wiley, New York.
Fisher, R. A. (1926). The arrangement of field experiments. *J. Ministry Agric Great Britain* **33**, 503–513.
Frangakis, C. F., Rubin, D. B. & Zhou, X.-H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. With discussion and rejoinder. *Biostatistics* **3**, 147–177.
Kempthorne, O. (1952). *The design and analysis of experiments*. Wiley, New York.
Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9.*Statist. Sci.* **5**, 465–480 (Translated, 1990).
Rosenbaum, P. R. (1995). *Observational studies*. Springer-Verlag, New York.
Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **7**, 34–58.
Rubin, D. B. (2002). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services Outcomes Res. Methodol.* **2**, 169–188.

Donald B. Rubin
E-mail: rubin@stat.harvard.edu

ELJA ARJAS

After having read the comments of Lauritzen and Aalen I felt first disappointed as they had not really challenged the ideas presented in the paper with Parner (A&P). It seemed that there would be little I could respond to. But then some of the issues that they had brought up, also in their comments to Rubin, made me think more about the connections between the MPP approach on the one hand, and the potential outcome framework and the graphical models on the other. I hope that these comments will help the reader to see what these three approaches have in common, and also how they differ.

To make such comments transparent, I will try to rephrase Case a of Rubin's macaque example, summarized in Display 3 of his paper. Suppose for simplicity that the treatment is chosen at time $t = 0$, the intermediate variable (antibody response) is measured at time $t = 1$, and the vital status Y is determined at time $t = 2$. I will generally adopt the notation of Lauritzen and denote the treatment by T and the intermediate variable by S. (Following A&P, the logical notation for the treatment would be A_0, with the corresponding covariate X_0 being a dummy, while the intermediate variable would be naturally viewed as a covariate X_1 and A_1 would again be a dummy. Considering antibody response as a treatment variable A_1 would not seem natural in the context, as corresponding causal inferences would very likely be severely confounded). Like the other authors, I will drop from the notation the index i referring to an individual macaque as redundant. Finally, in order to correspond to Rubin's example as closely as possible, suppose that each individual macaque can be characterized by a latent variable U, which, if it were known, would tell how S is going to depend on T. Such a variable also appears in Lauritzen's graphical models, and the notation is the same. Here U has three possible values, all *a priori* equally likely so that $P(U = 1) = P(U = 2) = P(U = 3) = 1/3$. Considering then treatment arm $T = 1$, let $P(S = L \mid U = 1, T = 1) = P(S = L \mid U = 2,$

$T = 1) = P(S = L \mid U = 3, T = 1) = 1$. Finally, copying from Rubin's Display 3, let $P(Y = 1 \mid U = 1, T = 1) = 0$, $P(Y = 1 \mid U = 2, T = 1) = 0.4$ and $P(Y = 1 \mid U = 3, T = 1) = 0.8$. For treatment arm $T = 2$ the corresponding probabilities could be similarly picked from Display 3.

Note that these last three probabilities remain the same if they are additionally conditioned on S, because T and U already determine S. In the actual experiment U is unknown, and therefore predictions concerning future values of S and/or Y cannot be based on such hypothetical knowledge. But as treatment assignment was assumed to be random, knowledge about the value of T does not tell one anything about U. Assignment is therefore non-informative also in the sense of A&P.

A simple computation, by averaging over the three possible values of U, shows now that $P(Y = 1 \mid T = 1) = 0.4$. Similarly, if one takes into account possible observation of S that becomes available at time $t = 1$, we have (see also Rubin's Display 3) that $P(Y = 1 \mid T = 1, S = L) = 0.2$ and $P(Y = 1 \mid T = 1, S = H) = 0.8$. In other words, the observation of S updates the 'crude' survival prediction $P(Y = 1 \mid T = 1) = 0.4$ downwards to $P(Y = 1 \mid T = 1, S = L) = 0.2$ if $S = L$ is observed, and upwards to $P(Y = 1 \mid T= 1, S = H) = 0.8$ if $S = H$. One of the *martingale* properties referred to by Aalen now says that the expected change in the predictions at time $t = 1$ is 0. This is because, apart from their opposite signs, the expectations of the upward and downward moves (innovations) are the same: an upward jump of size $0.8 - 0.4 = 0.4$ will happen with probability $P(S = H \mid T = 1) = 1/3$, and a downward jump $0.2 - 0.4 = -0.2$ will happen with probability $2/3$.

It is a simple matter to verify that the martingale property remains true at $t = 1$ also when $T = 2$. (In fact, the martingale property at $t = 1$ is a simple consequence of the identity $E(E(Y \mid T, S) \mid T) = E(Y \mid T)$, saying that the expected value of conditional predictions based on S is the same as the 'crude' prediction. Using a similar argument one easily verifies that the martingale property remains true for the MPP models considered in A&P and with respect to probabilities P_A, where A is a given assignment rule).

In epidemiological studies, the practical value of intermediate variables like S is in that they can often be used as surrogate endpoints if Y has not been measured on some individuals, typically because of right censoring. For this to be useful in practice, S should naturally be a good predictor for Y. For example, in leukaemia studies relapse could serve as a surrogate outcome for survival if the patient is still alive. Once such estimation has been carried out, one can follow the recipe given in A&P and predict, for each treatment T, the future response (of a generic patient) according to P_{A}. The prediction can be done either jointly for the pair (S, Y), or marginally for S or Y. In this context it is important to recall that one should never condition such predictions on intermediate variables (see e.g. Keiding, 1999). This is obvious, and the stronger S is as a predictor of Y the less important it is to know also T in predicting Y. One obvious possibility to consider the 'added value' of S in such predictions would be to compute the expectation of the conditional variances, $E(\mathrm{Var}(Y \mid T, S) \mid T)$, and compare it with the 'crude' variance $\mathrm{Var}(Y \mid T)$. In particular, the expression $1 - E(\mathrm{Var}(Y \mid T, S) \mid T)/\mathrm{Var}(Y \mid T) = \mathrm{Var}(E(Y \mid T, S) \mid T)/\mathrm{Var}(Y \mid T)$ has all its values between 0 and 1; it is equal to 0 if knowledge of S is redundant in predicting Y, and equal to 1 if it is perfect.

Attempting to connect these considerations to graphical models, the situation here would correspond to the graph displayed on the left-hand side of Lauritzen's Figure 2, where the missing arrow from U to T guarantees that U does not confound the effect of T on S and Y. In the terminology of A&P, this corresponds to the requirement that the choice of T is non-informative. However, apart from a short passage in Example 1, A&P does not contain anything on the issue that appears quite central in the discussion between Rubin and Lauritzen: whether the treatment effect is 'completely mediated by the intermediate variable' (as would be the case in the left-hand graph of Lauritzen's Figure 1, where the arrow leading

directly from T to Y is missing). As was also noted by Lauritzen, the possible truth of the conditional independence statement $T \perp Y \mid (S, U)$ depends on what variable U would be considered. In practice the choice of U is rarely made explicit, and even if one would have a particular U in mind it would be impossible to verify the conditional independence from data because, by definition, U is not observed. The situation is not much different in the Rubin framework, in which the conditioning is on both $S(1)$ and $S(2)$, and these cannot be observed jointly. I find the analysis by Lauritzen, drawing a connection between these two ways of thinking, illuminating.

Conditioning on an unobserved U, or on a pair of potential surrogate outcomes, is an attractive mental exercise that may lead to interesting interpretations in real life examples. At the level of 'principal strata' this seems to be possible, but again the 'truth' of precise probabilistic statements based on such conditioning cannot be verified empirically. In Rubin's macaque example, the somewhat contradictory looking results between predictions conditioned only on past observations and those conditioned on the potential surrogate outcomes is one more example of 'Simpson's paradox' (which Rubin himself makes note of in a different context). One can always introduce into the models latent variables that will match with observed data in the sense of giving correct marginal distributions, but which may have even seemingly 'incorrect' orderings between them. In the macaque example, as noted also by Rubin, treatment $T = 2$ is superior to $T = 1$ already by a comparison of the 'crude' predictions $P(Y = 1 \mid T = 1) = 0.4$ and $P(Y = 1 \mid T = 2) = 0.6$. That the adjusted predictions, which can be made after observing S, do not distinguish between the two treatments does not contradict this. The reason is clear from Rubin's Display 3: when $T = 2$, the higher survival prediction is issued more often than when $T = 1$ as in the former case the 'good news' $S = H$ is observed more often (in fact, with twice the probability).

There is one way, however to get closer to the idea of joint conditioning on the potential surrogate outcomes. This could be called 'play back conditioning' and was earlier considered, in the context of so-called *token causality,* in Arjas (1999). Suppose that we have chosen treatment $T = 1$ and have monitored the life of an individual macaque, observing $S = H$. Suppose, however, that we would be interested in predicting Y in the 'counterfactual' situation in which $T = 2$ had been chosen. This task is different from the original 'crude' prediction $P(Y = 1 \mid S = 2)$, because we now have acquired additional knowledge about the individual in question: $S = H$ combined with the 'factual' treatment $T = 1$ reveals that this macaque must belong to the 'principal stratum' specified by the third row of Rubin's Display 3 (or, using our earlier definitions, we have $U = 3$). Otherwise $S = H$ would have been impossible under $T = 1$. But then Display 3 tells that, if indeed $T = 2$ had been chosen, both $S = H$ and $Y = 1$ would happen with probability 1. Using Rubin's notation, having first obtained a particular value of $S(1)$ we would also be able to tell $S(2)$. But this is a very special case, as can be seen directly by trying out the situation in which the treatment would have remained the same ($T = 1$) but the intermediate observation had been $S = L$.

I conclude with some brief comments on statistical inference and causality. Even when causality is considered in explicit probabilistic terms, it is rarely explained where those probabilities came from. When such reference to statistical inference and data is missing, the obvious suggestion is that the probabilities have been obtained by frequency estimation from samples of 'nearly infinite' size. But this distances such considerations from the problems which a practising statistician faces: that there is a very complex thought experiment, involving uncertainties on several different aspects of the considered causal problem, before one is ready to formalize it even in a tentative way in the chosen framework, be that the graphical model, potential/counterfactual outcome, marked point process, or any other formalism one is accustomed to think in terms of. Moreover, the empirical evidence

available is typically not sufficient for a direct comparison of competing ideas and thoughts. In designed experiments there are usually only limited amounts of data, and in observational studies the possibility of arriving at confounded inferences is a real threat to the validity of the conclusions that can be drawn. Therefore it is a definite bonus if the statistical model, and the inferential method one will use, are made an explicit part of this thought process, and then methods based on probabilistic modelling and likelihood inference can be said to enjoy a canonical status. If this is done it may be less relevant, as Aalen suggests, whether one prefers Bayesian or ML estimation. In particular, individual random effects will be handled in terms of distributions, so for them there is no difference. However, even if the statistical data analysis were done by frequentist methods, I would hesitate to call the results 'objective'.

To Aalen's wish to see 'an example analysed in detail, with actual data and subjective prior distributions', I would respond by referring to older papers such as Arjas & Liu (1995) and Andreev & Arjas (1998). A final point: while A&P naturally tries to argue in favour of the approach that it proposes, its 'superiority' was certainly not 'asserted', nor is it claimed here.

References

Andreev, A. & Arjas, E. (1998). Acute middle ear infection in small children: a Bayesian analysis using multiple time scales. *Lifetime Data Anal.* **4**, 121–137.

Arjas, E. (1999). Probabilistic token causation: a Bayesian perspective. In *Applied probability and stochastic processes* (eds J. G. Shanthikumar & U. Sumita), 63–72. Kluwer, Dordrecht.

Arjas, E. & Liu, L. (1995). Assessing the losses caused by an industrial intervention: a hierarchical Bayesian approach. *J. Roy. Statist. Soc. Ser. C* **44**, 357–368.

Keiding, N. (1999). Event history analysis and inference from observational epidemiology. *Statist. Med.* **18**, 2353–2363.

Elja Arjas
E-mail: elja.arjas@rni.helsinki.fi