Article Type: Letter to the Editor

Annotations & Reflections

Pregnancy and paracetamol: Methodological considerations on the study of associations between in utero exposure to drugs and childhood neurodevelopment.

Dear Editor,

We have observed an increasing interest in the study of childhood neurodevelopment following in utero exposure to drugs (1-7). We appreciate the efforts to address an important, difficult and poorly studied subject, but we believe there are some major methodological pitfalls to avoid. With specific reference to recent papers, we wish to discuss some general and specific, methodological issues that we believe complicate the interpretation of such data. Of special interest in this perspective is the publication by Brandlistuen *et al.* (1) that has stirred some controversy and attracted significant public attention (8-9).

General points to consider: Quality and validity of the outcome data

In studies associating in utero drug exposure to poor childhood neurodevelopment, this is perhaps the most pivotal area of interest. While pregnancy outcomes, such as malformations, miscarriages, small-for-gestational age, etc., are relatively easily quantified, such is not the case for childhood neurodevelopment. Prior to conducting an epidemiological study, one should therefore very carefully assess the data at hand: are these data a valid surrogate for the outcome of interest? While several clinical scoring systems have been developed, the gold standard being the Baylor III score (10), parent-assessed questionnaires are most commonly used. These are substantially more easy to apply in large-scale settings and inherently cost-effective. Quite a few of such questionnaires exist - the most commonly used being the Age and Stages Questionnaires (ASQ), the Motor and Social Development (MSD) scale and the Strength and Difficulties Questionnaire (SDQ) (11-13). Two of the most prominent recent population studies on paracetamol exposure during pregnancy and childhood neurodevelopment from Norwegian and Danish cohorts (1-2, 14-15) primarily used questionnaire-based outcomes, ASQ and SDQ, respectively. Importantly, these and other related questionnaires were originally developed as a cost-effective screening tool to identify children at risk of delayed neurodevelopment for further referral to diagnostic testing and observation (11-13, 16). They were not developed or validated as a means of analysing subtle differences in neurodevelopment within a healthy paediatric population. We thus have serious doubts as to what the scores calculated from such questionnaires (1, 17)

a) are actually measuring in terms of neurodevelopment, and
b) to what extent this translates into quantifiable parameters that can be meaningfully handled and interpreted.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/bcpt.12322

Despite the use of what papers typically quote as *"validated scores",* it is unclear to what extent these itemized questionnaires actually represent measurements of neurodevelopment. To what extent are these questionnaires meaningful in terms of a measurable and reproducible scientific outcome, and, not least, to the mothers observing their children? The content validity of e.g. the Norwegian version of a questionnaire is well documented (18). Unfortunately, the more important construct validity (e.g. to what extent does this proxy actually represent the outcome of interest), of the questionnaire is not sufficiently documented. Thus, clinically meaningful conclusions inferred from slopes derived from a regression analysis cannot be made. Two studies have made a direct assessment of construct validity of the questionnaire-based approach, comparing ASQ and MSD questionnaire-based scoring to the reference standard of clinically assessed Baylor III scores (19, 20). While p-values for comparison document a strong *correlation*, r-squared values are thoroughly unimpressive (0.24 to 0.31). SDQ has not been validated against the Baylor III score. A recent additional study on the validity of the ASQ specifically studied the validity compared to other tools of intelligence quotient (IQ) assessment in preterm children at the age of 5 years (21). ASQ was a reasonable tool for detection of severe developmental delay but not suitable for lesser degrees of delayed development in IQ. Alas, while a statistically significant correlation between the outcome measurement applied and the reference may be demonstrated, the degree to which this proxy actually explains the neurodevelopment in children is poor. This severely compromises the clinical inferences that can be allowed from such data.

Case study of the paper by Brandlistuen *et al.*

In the case of Brandlistuen *et al.* (1), the handling of data illustrates some pitfalls that, in our opinion, may lead to potential wrongful interpretations.

Data handling

As discussed above, we do not agree on the handling of the presented Likert-scale questionnaire data as continuous variables in this specific context. However, we do acknowledge that the handling of such data from an ordinate scale by calculation of means and mean differences is subject to disagreement and controversy (22, 23). One may choose to accept arguments put forward by Norman (23) and others: The statistical handling of the "numbers" (i.e. assigned and calculated mean scores) themselves are insensitive to the underlying constructs. However, just therein *lies the rub*: as long as these values do not reflect the issue at hand, clinically meaningful conclusions cannot readily be made from any inferential statistics applied to the data. In order to justify inferential conclusions made from handling the available data as a continuous variable, the authors would have to convince us that a meaningful and reproducible, objective estimate of neurodevelopment status can be applied to, say, an average difference of .23 score on 4-6 three-point Likert items assessed by the observing mother.

Sibling design issues

While we fully appreciate the brilliant idea of using sibling controls to partially adjust for genetic and environmental factors, we believe that important elements of the authors' data handling and

analysis have been improperly or insufficiently addressed. The authors have included the entire cohort of unexposed (table 4) and the concordant siblings (table 3) in the regression models. We believe that the main analysis should have been confined to discordant siblings and, additionally, the analysis should take the dependence within these sibling pairs into account, e.g. by using a conditional regression model. By including other sibling pairs or the entire cohort of unexposed in the regression models, some of the sibling adjustment effect is lost, and the study becomes vulnerable to exactly the same confounders that the authors intend to avoid. Additionally, all concordant pairs are used as reference in table 3. This reference population includes a substantial proportion where both siblings are exposed, thus making the coefficient estimates virtually uninterpretable. Lastly, the importance of sibling order as source of significant bias is handled improperly. The authors conclude that there is no such effect, as the confidence intervals of the respective point estimates ”overlap”. Such inference simply cannot be made (24). The sibling order should have been entered into the regression model.

Time of exposure

Another highly plausible biological rationale is that the timing of exposure during pregnancy could be of significance. The authors have tested for this by applying qualitatively and quantitatively unspecified ”numerous tests” which failed to reach statistical significance. The authors have applied the Bonferroni correction to address the issue of multiple testing. The approach taken by the authors appears counter-intuitive. The concern here is not the risk of a type 1 error but the opposite: you want to be very sure that this variable is not excluded on the basis of a false negative inference. Hence, if anything, the level of significance should have been increased rather than adjusted downwards, as it is preferable to accept some noise in the model as opposed to excluding an important parameter. Again, the more obvious and simpler approach is to enter the time variable in the regression model.

Clinical extrapolation

The attempt by the authors to quantify the extent of adverse neurodevelopment (using the term “substantially adverse”) and translate their findings into a relative risk in the general population is unjustified. As we argue above, based on insufficient construct validity of the principal outcome measurement and inadequate statistical handling of the data, there is simply no way to infer a clinically meaningful interpretation of statistically significant β values from the regression analyses performed.

Conclusion

The validity of outcome parameters and the translation of questionnaire-based scores of neurodevelopment into a single continuous scale variable should be undertaken with great caution. We believe that a temptation to overstretch the information value that can reasonably be extracted from such questionnaires is imminent. Clinically meaningful interpretation of such data in terms of regression analyses are in our opinion unjustified or, at the very least, subject to underlying assumptions than cannot be verified. Within the highly sensitive field of pregnancy and drugs,

diligence is paramount to avoid dissemination of information that otherwise may serve to confuse and worry physicians, pregnant women and the public alike.

Related note on Letter to the Editor and scientific discussions of published papers

We have had the experience of having a Letter to the Editor on the paper by Brandlistuen *et al.* (1) rejected for purely generic *("..we receive many more papers than we can publish…")* reasons (25). This was disturbing as we believe that an open scientific discussion of published data is fundamental to scientific development and essential to dissemination of knowledge. Therefore, we take this opportunity to emphasize our opinion on this subject: Editors should by and large give way for scientific discussions of published papers in their respective journals; in fact, we suggest that this be a prioritized obligation.

Sincerely,

Per Damkier, MD, PhD

Anton Pottegård, MScPharm, PhD

René dePont Christensen, Statistician, PhD

Jesper Hallas, MD, DMSc

References

1. Brandlistuen RE, Ystrom E, Nulman I, Koren G, Nordeng H. Prenatal paracetamol exposure and child neurodevelopment: a sibling-controlled cohort study. Int J Epidemiol 2013;42:1702-1713.
2. Liew Z, Ritz B, Rebordosa C, Lee P-C, Olsen J. Acetaminophen Use During Pregnancy, Behavioral Problems, and Hyperkinetic Disorders. JAMA Pediatr. 2014;168:313-320.
3. Batton B, Batton E, Weigler K, R.N. Aylwardn G, Batton D. In Utero Antidepressant Exposure and Neurodevelopment in Preterm Infants. Am J Perinatol 2013;30:297–302.
4. Harrington RA, Lee Li-Ching, Crum RM, Zimmerman AW, Herz-Picciotto I. Prenatal SSRI use and offspring with autism spectrum disorder or developmental delay. Pediatrics 2014. Doi:10.1542/peds.2013-3406.
5. Malm H, Artama M, Brown AS, Gyllenberg D, Hinkka-Yli-Salomaki S, McKeague I et al. Infant and childhood neurodevelopmental outcomes following prenatal exposure to selective serotonin reuptake inhibitors: overview and design of a Finnish register-based study (FinESSI). BMC Psychiatry 2012;12:217-225.
6. Nulman I, Koren G, Rovet J, Barrera M, Pulver A, Streiner D et al. Neurodevelopment of children following prenatal exposure to venlafaxine, selective serotonin reuptake inhibitors, or untreated maternal depression. Am J Psychiatry 2012;169:1165-1174.
7. Austin M-P, Karatas JC, Mishra P, Christl B, Kennedy D, Oei J. Infant neurodevelopment following in utero exposure to antidepressant medication. Acta Pædiatrica 2013;102:1054-1059.

8. NBC News. Too much Tylenol in pregnancy could affect child's development, study finds. November 22, 2013. Available from www.nbcnews.com accessed February 4, 2014.
9. American Medical Network. Too much Tylenol in pregnancy could affect neurodevelopment. November 23, 2013. Available from www.health.am accessed February 4, 2014.
10. Bayley N. Bayley scales of infant and toddler development. 3rd edn. San Antonio TX: Hardcourt Assessment, 2006.
11. Squires J, Bricker D, potter L. Revision of a parent-completed development screening tool: Ages and Stages Questionnaires. J Pediatr. Psychol 1997;22:313-318
12. Poe S Poe GS. Design and procedures for the 1981 Child Health Supplement to the National Health Interview Survey. Working paper series. Hyattsville, Maryland: National Center for Health Statistics 1986.
13. Goodman R. The Strengths and Difficulties Questionnaire: A Research Note. Journal of Child Psychology and Psychiatry. 1997;38:581-586.
14. Magnus P, Irgens LM, Haug K, Nystad W, Skjaerven R, Stoltenberg C; MoBa Study Group. Cohort Profile: The Norwegian mother and child cohort study (MoBa). Int J Epidemiol 2006;35:1146-1150.
15. Andersen AM, Olsen J. The Danish National Birth Cohort: Selected scientific contributions within perinatal epidemiology and future perspectives. Scan J Public health 2011;39 (Suppl 7):115-120
16. Glascoe FP. Are overreferrals on developmental screening tests really a problem? Arch Pediatr Adolesc Med. 2001;155:54-9.
17. Information for researchers and professionals about the Strengths & Difficulties Questionnaires,. Available from http://www.sdqinfo.com Accessed July 25, 2014.
18. Richter J et al. A validation of the Norwegian version of the ages and stages questionnaire. Acta Peadiatrica 2007;96:748-752.
19. Belfort MB et al. Using parent questionnaires to assess neurodevelopment in former preterm infants: a validation study. Paediatr Perinat Epidemiol 2013;27:199-207.
20. Schonhaut L et al. Validity of the ages and stages questionnaires in term and preterm infants. Pediatrics 2013;131.e1468-e1474 .
21. Halbwachs M, Muller JB, Nguyen TTS, de La Rochebrochard E, Gascoin G, Branger B et al. Usefulness of parent-completed ASQ for neurodevelopmental screening of preterm children at five years of age. PLoS One. 2013;8:e71925. doi: 10.1371/journal.pone.0071925. eCollection 2013
22. Jamieson, Susan (2004). "Likert Scales: How to (Ab)use Them," Medical Education, 2004;38:1217-1218.
23. Norman G. Likert scales, levels of measurement and the "laws" of statistics". Advances in Health Science Education. 2010;15:625-632
24. Sedgwick P. Confidence intervals and statistical significance. BMJ 2012;344:e2238.
25. Personal Communication. Letter from the Editor, Int J Epidemiol, March 2014.