A validation and calibration process for self-reported tobacco use with participants' cotinine levels: an example from the Building Blocks trial

Chao Huang PhD[1], Zoe Roberts PhD[2], Rebecca Cannings-John PhD[3], Julia Sanders PhD[4], Kate Pickett PhD[5], Alan Montgomery PhD[6], Michael Robling PhD[3]

[1]Hull York Medical School, University of Hull, UK; [2]School of Medicine, Cardiff University, UK; [3]Centre for Trials Research, Cardiff University, UK; [4]School of Healthcare Sciences, Cardiff University, UK; [5]Department of Health Science, University of York, UK; [6]School of Medicine, University of Nottingham, UK

Corresponding author: Chao Huang, PhD, Hull York Medical School, University of Hull, 3rd Floor, Allam Medical Building, Cottingham Road, Hull, HU6 7RX, UK; chao.huang@hyms.ac.uk; Tel: + 44(0)1482463281

ABSTRACT

Introduction: Reducing smoking in pregnancy was a primary outcome in our Building Blocks trial of the Family Nurse Partnership[1]. We calibrated maternal reports of smoking using cotinine values derived from urine samples to assess tobacco use [2]. This involves identifying the extent to which an individual accurately reports smoking and requires complete and synchronized data collection over time. However, some urine samples may be missed or collected at a different time from self-report (non-synchronised).

Methods: We used statistical validation processes to address both non-synchronized and incomplete data. First, we examined consistency in reporting behaviours at baseline and follow-up for participants grouped by extent of non-synchronized time of collection. Second, we used data from complete cases to infer values for mothers with missing urine samples at follow-up. We then used Markov chain transition rate matrix constructed to assess the robustness of such inferences.

Results: Maternal under- and over-reporting of smoking were consistent across the 870 participants grouped by different levels of non-contemporary data collection (Breslow-Day test: p=0.24; Chi-squared test: p=0.69). Using participants' baseline reporting behaviours to infer their follow-ups provided comparable smoking outcomes (4.5 cigarettes per day with SD of 5.5) to the simulated counterparts (4.5 cigarettes per day with SD of 6.0).

Conclusion: We have demonstrated consistent reporting behaviour over time and minimal impact due to non-aligned follow-up urine sample collection. For studies collecting smoking data this proposed method provided a pragmatic solution to facilitate the calibration process of self-reported tobacco use and retain adequate power without introducing undue bias.

IMPLICATIONS

Synchronized and completed data collection is essential but very often hard to achieve in smoking related studies. When violated, proper statistical validation process should be followed to minimize the potential bias and loss of power in trial analyses. For this purpose, we provided the Building Block trial as an example to demonstrate how to deal with the non-synchronization and incompleteness issues in data collection.

INTRODUCTION

The Family Nurse Partnership (FNP) is a licensed intensive home-visiting intervention developed in the USA that involves up to 64 structured home visits by specially recruited and trained family nurses, from early pregnancy until the child's second birthday. The Building Blocks trial [1] aimed to assess the effectiveness of FNP in England, for adolescent first-time mothers. This pragmatic, non-blinded, randomised controlled, parallel-group trial in community midwifery settings was carried out within 18 partnerships between local authorities and primary and secondary care organisations in England.

In the Building Blocks trial [1], one of the primary outcomes was to investigate the effectiveness of FNP in reducing smoking behaviour during pregnancy. We collected self-reported data on number of cigarettes smoked per day, at trial baseline and late in pregnancy. Self-report is an effective method of data collection in terms of time, efficiency and feasibility, and accuracy can be reasonable within research studies, especially with less sensitive health behaviours. However, self-reported smoking can be inaccurate and some participants are likely to report smoking fewer (or more) cigarettes than they actually do [2,3]. Recall bias, the difficulty in remembering information over time, could randomly cause both over-reporting and under-reporting [4, 5]. Other factors tend to lead to bias by underreporting only, for example, younger people and pregnant women were more likely than others to answer in ways that are congruent with social norms (social desirability bias) [6, 7, 8, 9].

The prevalence of misreporting of self-reported smoking is not trivial. In one trial with adolescent participants, 30% were either under-reported, over-reported or both [4]. In another study, almost 26% of pregnant women reporting themselves as non-smokers were subsequently classified as smokers after validation by means of serum cotinine measurement [8].

The accuracy of self-reported health behaviours had been discussed widely and the use of validation techniques and objective measures were strongly recommended to minimize bias [10]. For self-reported smoking data, several biochemical measures were available as validation techniques [11 - 17]. In the Building Blocks trial, we used urinary cotinine levels to supplement the information gained from participants' self-reported behaviours [1]. Together we used the analytical approach from Dukic et al. [2] to combine both self-report and biochemical information to increase the precision of the smoking measurements. This calibration approach requires both complete and well-synchronized collection of self-reports and urine samples, i.e., the self-report of number of cigarettes and urine samples should be collected on the same day. This is especially important among pregnant women, as it is known that their smoking patterns fluctuate over short periods of time [18,19].

In the Building Blocks trial we achieved high levels of synchronization in data collection for self-report and cotinine measures at baseline, as both were obtained during a single face-to-face interview. However, we faced greater challenges in synchronized collection at follow-up in late pregnancy, when separate collection approaches were necessary (women self-reported their smoking during telephone interviews and were asked to return urine samples by post). For most participants follow-up self-reports and urine samples collection occurred on different dates. In addition, there was more missing data at follow-up. Some participants provided only self-reports, while some had neither self-reports nor urine samples. These issues created the risk of losing participants for the analyses (loss of power) and potential bias.

This research aimed to provide a pragmatic solution to facilitate the calibration process of self-reported tobacco use and retain adequate power without introducing undue bias. The integrated

validation and calibration process was proposed to tackle the data incompleteness and non-synchronization issues.

METHODS

Calibration approach

The calibration approach we adopted from Dukic et al [2] could be summarized into five steps. The first two steps were to calculate the cotinine weighted number of cigarettes ($N_{\cot}$) based on the participant's cotinine level and the self-reported weighted number of cigarettes (N_{self}) based on the participant's self-report of number of cigarettes smoked on each of the 3 days prior to interview. The third step was to classify the participants into four reporting groups: over-reporter, accurate reporter, under-reporter and extreme under-reporter, by comparing their $N_{\cot}$ and N_{self} values. Over-reporters were participants whose cotinine level was at least 30% less than that expected according to their self-reported average number of cigarettes smoked. Accurate-reporters had urine cotinine levels within 30% differences of their self-reported average. Under-reporters had cotinine levels 30% - 80% greater than expected for their self-reported average while extreme under-reporters had cotinine levels >80% greater than their self-reported average. At step four, for each reporting group, we calculated the averaged difference between self-reported number of cigarettes and number based on cotinine samples. Finally, for each participant, these averaged differences were used to calibrate the self-reported number of cigarettes (Details on this approach were provided in supplementary materials S1).

Study design and participants

The Building Blocks trial was originally registered with ISRCTN (number ISRCTN23019866). Completeness and synchronization in data collection at baseline were well achieved (95.7% were completely collected and 96.1% of them were collected on the same day). We focused on the non-synchronization at follow-up. We first grouped the trial participants according to their completeness of data collection at follow-ups. Participants who had both self-report and cotinine sample collected at follow-up were categorized as full data cases, for whom only non-synchronization issue existed. Participants with only self-reports at follow-up formed partial data cases (incompleteness in data collection). Participants who neither have self-report nor urine sample at follow-up were categorized as insufficient data cases (not valid for analyses).

Statistical analysis

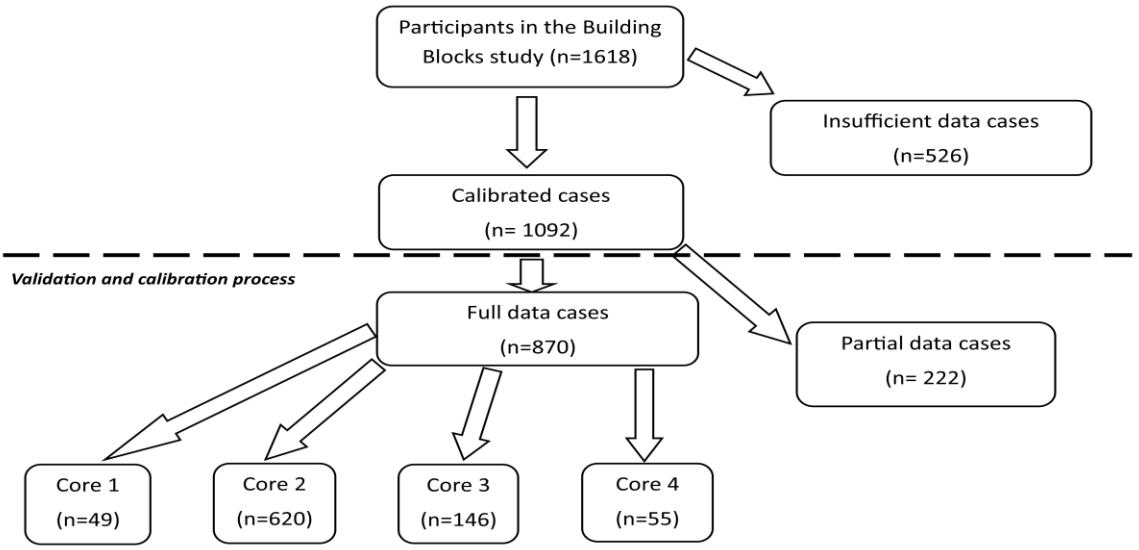
A participant flow chart for the validation and calibration process was presented in Figure 1. Full data cases were further divided into four core groups (Core 1 – Core 4) according to the extent of gaps between self-report and urine sample collection dates. Participants of each core group were further cross-tabulated and the homogeneity in reporting behaviours was assessed by Breslow-Day test, while the baseline/follow-up association was evaluated by Mantel-Haenszel test. A contingency table on reporting behaviour shifting patterns was also employed for Chi-squared test on homogeneity. For partial data cases, the robustness of inferring participants' reporting behaviours at follow-up by their baseline counterparts was examined by a robustness testing process. We first directly imputed the follow-up reporting behaviours by their baseline counterparts and calculate their calibrated tobacco use. We then conducted the Monte Carlo simulation to impute the follow-up reporting behaviours by using the Markov chain transition rate matrix constructed by the full data cases. With multiple imputation, ten simulated data sets were repeatedly generated and the pooled results of calibrated tobacco use were calculated and compared with the results from direct

imputing, which lead to the robustness justification. R language version 3.4 and SPSS 20 were utilized for the analysis.

RESULTS

From 1618 participants in the Building Blocks study, 526 insufficient data cases provided too little information and hence were excluded for analysis. Of 1092 participants for this research, 870 participants were full data cases and 222 participants were partial data cases (Figure 1).

Figure 1. Participants flow chart for the validation and calibration process



Non-synchronization in data collection at follow-up

To check if the duration of the time lag between the urine sample collection date and interview date caused heterogeneities in smoking outcomes, we divided the participants into the following four core groups (Figure 1):

Core 1: Participants whose late pregnancy interview date and urine sample collection were on the same day (49 in total).

Core 2: Participants whose interview date and urine sample date were within 2 weeks (620 in total).

Core 3: Participants whose interview date and urine sample date were more than 2 weeks, but less than 4 weeks apart (146 in total).

Core 4: Participants with greater than 4 weeks lag between interview and urine sample date (55 in total).

Applying Dukic and colleagues' calibration approach [2], these 870 participants were cross-tabulated according to their reporting behaviours (over-reporter, accurate reporter, under-reporter and extreme under-reporter) at baseline and follow-up (Supplementary materials S2).

In order to utilize the Breslow Day test for homogeneity assessment, we further collapsed these reporting behaviours into two simplified groups: accurate and over reporters forming the positive reporters; under and extreme under reporters forming negative reporters (Table 1). The Breslow-Day test on Table 1 showed that these four core groups present homogenous reporting behaviours (p=0.24). This implied that the impacts of non-synchronisation in data collection on reporting behaviour patterns were ignorable, supporting combining all 870 participants for analysis. Additionally, the Mantel-Haenszel chi-squared test showed that the participants' baseline and follow-up reporting behaviours were associated (p<0.01).

Table 1. Simplified reporting behaviours at baseline and follow-up for core 1 - core 4. In each core, participants were classified as positive or negative reporters at baseline and follow-up.

		Follow-up: N, %							
		Core 1		Core 2		Core 3		Core 4	
		PE*	NE*	PE*	NE*	PE*	NE*	PE*	NE*
Baseline: N, %	PE*	20 (40.8%)	8 (16.3%)	227 (36.6%)	99 (16.0%)	45 (30.8%)	29 (19.9%)	19 (34.5%)	7 (12.7%)
	NE*	5 (10.2%)	16 (32.7%)	71 (11.5%)	223 (36.0%)	23 (15.8%)	49 (33.6%)	8 (14.5%)	21 (38.2%)

*PE: positive reporter; NE: negative Reporter.

To investigate the homogeneity in reporting behaviour shifting between the four core groups, participants were regrouped into three categories: consistent reporter, increasingly positive reporter and increasingly negative reporters (Table 2). Consistent reporters were participants whose reporting behaviours at follow-up were the same as theirs at baseline (second row). Positive-shifting reporters were participants who became more positive in reporting at follow-up (first row), while negative-shifting reporters became more negative in reporting at follow-up (reporting less smoking in self-report; third row). The Chi-squared test of Table 2 showed the proportions of participants with different reporting behaviour shifting were homogeneous across the four core groups (p=0.69). This evidence also supported the validity of combining the 870 participants.

Table 2. Contingency table on reporting behaviour shifting for core 1 - core 4. In each core, participants were classified as positive-shifting, consistent or negative-shifting reporters.

N, % (95% CI)	Core 1	Core 2	Core 3	Core 4
PSR*	10, 20.4% (11.5%, 33.6%)	102, 16.5% (13.7%, 19.6%)	30, 20.5% (14.8%, 27.8%)	12, 21.8% (12.9%, 34.4%)
CR*	29, 59.2% (45.2%, 71.8%)	388, 62.6% (58.7%, 66.2%)	83, 56.8% (48.7 %, 64.6%)	35, 63.6% (50.4%, 75.1%)
NSR*	10, 20.4% (11.5%, 33.6%)	130, 21.0% (17.9%, 24.3%)	33, 22.6% (16.6%, 30.0%)	8, 14.5% (7.6%, 26.2%)

*PSR: positive-shifting reporter; CR: consistent reporter; INR: negative-shifting reporter.

In the Building Blocks trial, the impact of intervention on participants' reporting behaviours was also of scientific interest [1] and we carried out an additional analysis to investigate this. The results (Supplementary material S3) showed that 263 of 431 (61.0%) participants who received usual care were consistent reporters, compared to 272 of 439 (62.0%) of participants who received the FNP intervention. The Breslow-Day test (Supplementary material S4) indicated no significant treatment effects on participants' reporting behaviours (p=0.92).

Incompleteness in data collection at follow-up

To address incompleteness of data collection in partial data cases (222 participants with missing cotinine level at follow-ups), reporting behaviours of 870 full data cases were tabulated into Table 3. In terms of their reporting behaviour shifting, 535 (61.5%) participants were consistent reporters (diagonal entries), while 154 (17.7%) participants became positive-shifting reporters (lower sub-diagonal entries) and 181 (20.8%) became negative-shifting reporters (upper sub-diagonal entries). We then used this table as transition rate matrix to test the robustness of assuming the partial data cases were consistent reporters.

Table 3. Cross-tabulated reporting behaviours (over, accurate, under or extreme under reporters) of 870 full data cases at their baseline and follow-up.

		Follow-up n, %			
		O*	A*	U*	E*
Baseline n, %	O*	8, 0.9%	15, 1.7%	4, 0.5%	12, 1.4%
	A*	3, 0.3%	286, 32.9%	115, 13.2%	12, 1.4%
	U*	3, 0.3%	60, 6.9%	36, 4.1%	24, 2.8%
	E*	9, 1.0%	34, 3.9%	44, 5.1%	205, 23.6%

*O: over-reporter; A: accurate reporter; U: under reporter; E: extremely under reporter.

The first step was to use these participants' baseline reporting behaviours to directly impute their follow-up reporting behaviours by the consistent reporter assumption. Then we calculated their calibrated self-reported tobacco use, which was an average of 4.5 (SD=6.0) cigarettes per day. In the second step, we used the Table 3 as the benchmark transition rates to carry out multiple imputation on follow-up reporting behaviours. For example, if *participant A* was an over-reporter at baseline, then following the first row of the table, her probability of becoming an over-reporter at follow-up would be 8/(8+15+4+12)=20.5%. Similarly we could work out the probability of her becoming an accurate reporter (15/(8+15+4+12)=38.5%), under reporter (4/(8+15+4+12)=10.3%) or extreme under reporter (12/(8+15+4+12)=30.8%). The reporting behaviour of *participant A* at follow-up was then allocated by Mont-Carlo simulation with these prior transition probabilities, i.e., we would

allocate *participant A* as an over, accurate, under or extreme under reporter with probabilities of 20.5%, 38.5%, 10.3% and 30.8% respectively. For the multiple imputation process, ten simulated data sets were generated with imputed reporting behaviours at follow-up. The pooled mean and standard deviation of calibrated cigarettes were calculated, which was averagely 4.5 cigarettes per day (SD=5.5). These figures were consistent to the results from step one, which confirming the robustness of consistent reporter assumption.

DISCUSSION

To our best knowledge, this is the first study to tackle the non-synchronisation issue in smoking data analysis with a formal testing procedure. For the Building Blocks study, we demonstrated no heterogeneity in reporting behaviours among participants with different extent of non-synchronisation in smoking data collection. This result provided reassurance to our decision to combine these participants in the main trial analysis.

We also formally tested the robustness of assuming consistent reporting behaviours among participants with partial data, by borrowing the strength of constructed transitional rate matrix from full data cases. The underline claim was that the partial data cases and those with full data were the same when it comes to their consistency of reporting over time. Again, results confirmed the validity of this assumption for our study.

There were 526 participants with neither urine samples nor self-reports collected at follow-up; these participants were not included in the main trial analysis. One theoretical analytic option for this participant group would be to impute both their follow-up cotinine levels and self-reports from their baseline values. This approach was clearly not appropriate in the Building Blocks trial, as the intervention was hypothesised to reduce participants' smoking activities. In observational studies of smoking, direct imputation may be a reasonable approach.

At baseline interview, about 50% participants were negative reporters (either under-reporters or extreme under-reporters). This high underestimation rate suggests a strong sense of social undesirability in reporting smoking in this population group (teenage women experiencing their first-time pregnancy, aged 16.9-18.8).

In the Building Blocks trial [1], we found no difference in the proportion of smokers and average number of cigarettes smoked per day between the treatment and control groups. Our additional analysis assessed the impact of the intervention on the participants' reporting behaviours at follow-up; that we found no significant difference in reporting behaviours across trial arms provides additional support for the main trial conclusions.

There are limitations to this study. The calibration approach we employed in this study [2] did not consider participants' demographic and metabolic heterogeneity. It has been shown that women differ in how quickly they metabolise nicotine and this may also change over time during pregnancy [20]. Therefore, the interpretation of the results should be restricted to the population level rather than the individual level. With 1618 participants, the Building Blocks study has strengthened our understanding of smoking behaviours in young woman with first-time pregnancies, however, we should be cautious in extrapolating the results to other populations.

Our pragmatic methodological approach can be extended to other studies with similar issues, where self-reported and biomarker outcomes can be calibrated but where there are non-synchronization and incompleteness issues.

FUNDING

This paper is based on independent research commissioned and funded by the National Institute for Health Research Policy Research Programme (reference 006/0060). The views expressed are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health and Social Care or its arm's length bodies, and other Government Departments.

DECLARATION OF INTERESTS

None declared.

ACKNOWLEDGEMENT

The authors would like to thank Vanja Dukic for providing helpful comments to this research. We also thank the journal reviewers for insightful comments on earlier draft of this manuscript.

REFERENCES

1. Michael Robling, Marie-Jet Bekkers, Kerry Bell, et al. Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (Building Blocks): a pragmatic randomised controlled trial. The Lancet, 2016; Volume 387, Issue 10014 , 146 – 155.
2. V.M. Dukic, M. Niessner, N. Benowitz, S. Hans, L.S. Wakschlag. Modeling the relationship of cotinine and self-reported measures of maternal smoking during pregnancy: a deterministic approach. Nicotine Tob Res, 2007; 9 (4), pp. 453-465.
3. Pickett, K.E., Kasza, K., Biesecker, G., Wright, R.J, Wakschlag, L.S. Women who remember, women who don't: A methodological study of maternal recall of smoking in pregnancy. Nicotine Tob Res, 2009; 11(10): 1166–1174.
4. Lantini R, McGrath AC, Stein LA, Barnett NP, Monti PM, Colby SM. Misreporting in a randomized clinical trial for smoking cessation in adolescents. Addict Behav, 2015; 45: 57-62.
5. Piasecki TM, Solhan M, Trull TJ, Hufford MR. Assessing clients in their natural environments with electronic diaries: rationale, benefits, limitations, and barriers. Psychol Assessment, 2007; 19: 25-43.
6. Barker C. Research methods in clinical psychology : an introduction for students and practitioners. Third edition. ed: Chichester, West Sussex : Wiley Blackwell; 2016.
7. De Vaus DA. Surveys in social research. Sixth edition. ed: Abingdon, Oxon : Routledge; 2014.
8. Shipton D, Tappin DM, Vadiveloo T, Crossley JA, Aitken DA, Chalmers J. Reliability of self reported smoking status by pregnant women for estimating smoking prevalence: a retrospective, cross sectional study. BMJ, 2009; 339: b4347.
9. Connor Gorber S, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M.The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. Nicotine Tob Res, 2009; 11(1):12-24.
10. Newell SA, Girgis A, Sanson-Fisher RW, Savolainen NJ. The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population: a critical review. AM J PREV MED, 1999; 17: 211-29.

11. Bailey BA. Using expired air carbon monoxide to determine smoking status during pregnancy: preliminary identification of an appropriately sensitive and specific cut-point. Addict Behav, 2013; 38: 2547-50.
12. Burstyn I, Kapur N, Shalapay C, et al. Evaluation of the accuracy of self-reported smoking in pregnancy when the biomarker level in an active smoker is uncertain. Nicotine Tob Res, 2009; 11: 670-8.
13. Dempsey, D., Jacob, P. III, Benowitz, N.L. Accelerated metabolism of nicotine and cotinine in pregnant smokers. J. Pharmacol Exp Ther, 2002; 301, 594–598.
14. Patterson F, Benowitz N, Shields P, et al. Individual differences in nicotine intake per cigarette. Cancer Epidem Biomar, 2003; 12(5): 468-71.
15. Etzel RA. A review of the use of saliva cotinine as a marker of tobacco smoke exposure. Prev Med, 1990; 19(2): 190-7.
16. Owen, L. , McNeill, A. Saliva cotinine as indicator of cigarette smoking in pregnant women. Addiction, 2001; 96, 1001 – 1006 .
17. Smith JJ, Robinson RF, Khan BA, Sosnoff CS, Dillard DA. Estimating cotinine associations and a saliva cotinine level to identify active cigarette smoking in alaska native pregnant women. Matern Child Healt J. 2014; 18(1):120-8.
18. Pickett KE, Wakschlag LS, Leventhal BL. Fluctuations in maternal smoking during pregnancy. Obstet Gynecol 2003; 101:140-7.
19. Pickett KE, Rathouz PJ, Kasza K, Wakschlag LS, Wright RJ. Self-reported smoking, cotinine levels and patterns of smoking in pregnancy. Paedia and Perinat Ep 2005; 19:368-376.
20. Vanja M. Dukic, Marina Niessner, Kate E. Pickett, Neal L. Benowitz, and Lauren S. Wakschlag. Calibrating Self-Reported Measures of Maternal Smoking in Pregnancy via Bioassays Using a Monte Carlo Approach. Int J Env Res Pub He. 2009; 6(6): 1744–1759.