Reviewer's report

Title: Estimation of HPV prevalence in young women in Scotland; monitoring of future vaccine impact

Version: 1 Date: 9 September 2013

Reviewer: Kate Soldan

Reviewer's report:

This is a nice study of HPV prevalence pre-immunisation. It could be improved by a clearer focus on the key findings and using the data and methods accordingly, rather than describing everything without clear indication of relevance. The potential bias in participants in the defaulters component is not acknowledged, but may not be relevant if these data are not intended to be combined with the LBC data for comparison with follow-up - in which case, it needs to be made clear why are they being reported here.

Major

1. Abstract methods and results: please add some mention of the age (20-21yrs) of women in the study (currently reads as “women” implying a wider age range). The mention in the background is insufficient. Also in line 87 of introduction (“..estimates of HPV prevalence in unvaccinated women aged 20-21 in Scotland.”).

2. The weighting of the prevalence estimates needs to be clearer (lines 169-173) and better explained/discussed. Please state in the methods and tables which results were weighted for which variables. If weighting for reminder status, please also report the (crude) prevalence by reminder status. Also in line 30, abstract: is it possible to state which variable(s) used for weighting e.g. “Age-weighted estimates…”? In this result is it “Estimates weighted for deprivation and reminder status…” (presumably not for specimen type here)? Also, the line 173 suggests the weighting was assumed to successfully “account for any influence of recruitment factors”, which seems optimistic and unlikely to be correct (see later re 13% participation of defaulters).

3. HPV type groupings: The results are given for any HPV type, and for various sub-groups including HR HPV types. This latter group is, in my view, far more interesting and relevant than “any HPV”. Please reconsider making “HR HPV” the primary outcome of analysis instead of “any HPV” throughout (including, for e.g. in line 272, discussion). Please also (which ever grouping used) be clear for each result, e.g. line 195 should read “The prevalence of any HPV was 32.2% for urine…”, not “..of HPV was..”. Line 200-2002 reports prevalence excluding HPV16/18 and excluding cross-protection types: this is a strange way to present the results after defining the groups for analysis by inclusion, not exclusion. If a particular point is to be made about the prevalence estimates reported here (i.e. for HPV types excluding...), then it would probably fall better entirely within the discussion. However, I do not currently find this result discussed there, so can
perhaps be omitted entirely. The section on multiple infections (starting line 233), in particular, would be more relevant if restricted to multiple HR types. The prevalence infected with more than one type (lines 234-238) is fairly meaningless, as comparisons with other studies depend entirely on which types are detected by the assay and so included in “any HPV” Lines 239-244 put emphasis on specific type co-infections: what is the point of these results in the current analysis (if only for comparison with post-imms data, to reserve for post-imms analysis?)? The risk factors for multiple infections (next section, starting line 245) are slightly more meaningful, in so far as exposure to any HPV is associated with exposure to HR HPV (the HPV of interest for health), but I would still prefer this to be for HR only.

4. Type distribution by specimen type: I am not clear what the objective of this analysis is (lines 203-216), in terms of establishing the baseline for comparison with subsequent post-immunisation data from LBCs. Lines 206-209 seem to take the comparisons between types too far, given the uncertainties. Lines 210-216 report an analysis of unstated power, and the meaning in terms of use of the data from self-taken specimens is not clear. If this is a method applied to check/permit inclusion of the self-taken specimen data in the “baseline” for subsequent use, then this belongs in the methods, not the results. Is “marginally” higher prevalence of some types either statistically of clinically significant?

5. Care should be taken when comparing odds ratios with the use of "less likely" or "more likely", or “1.55 times more likely”. See literature on interpretation of odds ratios vs relative risks for relatively common outcomes (such as HPV).

6. The difference in multiple infections by specimen type is reported as if this is an epidemiological observation. Is that intended? Presumably this is a reflection of different detection sensitivities for the different specimen types, and related to overall detected prevalence? Is the lower odds of multiple infection in urines to be expected given lower detected prevalence in urines? I am not convinced, by the analysis as presented, that the difference by non-white resident quintile is not due to confounding with specimen type.

7. Discussion: The representativness of the study population of the “general population” is over-stated. E.g. (line 264-5)“have obtained samples from both those attending screening and defaulters in order to accurately calculate pre-immunisation HPV prevalence ion this cross-section of the population.” And line 325. With only 13% participation of defaulters, this is not totally justified – as is stated in line 316-17.

8. Lines 281-284 introduce possible association between both the urban/rural variable and levels of deprivation and HPV prevalence. If so, is it surprising that adjustment for one did not alter the magnitude or significance very much of the other (table 3?)? Could you report the association (co-linearity) between these variables within the study data?

9. There is no mention of intention to conduct post-immunisation collections from defaulters, possibly due to the low participation rate. In which case, the “baseline” established here is for comparison to LBC data. I am not totally clear whether the plan is to use baseline estimates that include the self-taken swab data, and
defaults, for comparison with women attending for screening in the follow-up
(3000 stated in paper starting line 162), i.e. combined data. Could you clarify the
place of the self-taken swab data in useful “characterisation of a pre-vaccine
baseline”?

Minor

1. Line 30, abstract “either urine (n=378) and self-taken swab..” should be “either
urine (n=378) or self-taken swab..”

2. Line 46, abstract: remove “is”.

3. Line 59, introduction: 2 or the 4 references cited are for oral HPV prevalence,
which seems strange in background to a paper about monitoring genital HPV.

4. Line 66-68, introduction: It is not exactly true that “the best estimates of HPV in
the general population come from studies that consider samples from cervical
screening programmes”. Women participating in cervical screening programmes
(especially with no qualification as to coverage of these programmes) are a)
patients and b) somewhat self-selected in response to screening offer/invitation.
There have been studies of HPV conducted in general population surveys (e.g.
NATSAL in UK, NHANES in US) which would tend to be considered the best
(albeit not without their own biases) for estimating general population prevalence.
Screening-based studies may be preferred because of link to cytology/disease
status, and ease, rather than because of representative of general population.

5. Line 71, introduction: suggest “..feeds into models of vaccine impact and cost
effectiveness to inform..”.

6. Line 74, introduction: decrease in “prevalence” is the expected vaccine effect,
not necessarily (or measurably) of “circulation” of infection.

7. Line 93-94, introduction: modelling is one method to contribute to
understanding, not as an aim in itself. Suggest delete “modelling and”.

8. Line 133, methods: delete “the”. Also paragraph break seems to come 1
sentence too early.

9. Line 191, results: suggest include total self-taken samples from defaulters
(709?) as well as split by sample type.

10. Line 223-4, results: do you mean “"the linear effect of quintile of non-white
residents became...”

11. Line 228 and 229. It would help to include “(i.e. least deprived)” and “(i.e.
most deprived”) after SIMD5 and SIMD1, respectively.

12. Lines 230, 232, and others. Please indicate when reporting adjusted odds
ratios (“aOR”).

13. Tables: Could percentages be shown in more traditional percentage format,
rather than as proportions, throughout? E.g. 68% not 0.68.

14. Table 2 and 3. Would suggest that only 1 or 2.d.p are necessary.

15. Table 4 is labelled Table 5

16. I would favour the numbers (numerators and denominator) included in Figure
2, not only percentages.

17. Please could you state in table title or footnotes what the data are weighted or adjusted for, where relevant, and be consistent in all tables. Table 3 column headers include “unadjusted” and “adjusted” ORs, vs title of “Adjusted odds”. Table 4(labelled as 5) is titled “Adjusted odds..” but column heading is “OR” (not “adjusted OR”).

18. Line 259: delete “to”

19. Line 263: consider revising “population-based”

20. There is some duplication in the discussion, e.g. line 268-9 and line 286-7, line 276-7 and line 314-315.

21. Line 291: suggest delete “clinically”

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Needs some language corrections before being published

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

Declaration of competing interests:

none