Appendix 3 (as supplied by the authors): Methodological approach to missingness

The proportion of missingness that is considered acceptable for multiple imputation has been an area of active debate, yet we are unaware of any explicit cut-point, nor do we believe that one exists simply on the basis of proportion missing as a single criterion [1]. While the proportion of missingness is a critical factor to consider, it is also important to consider the mechanism of missingness (i.e., missing not at random versus missing at random), and the number and strength of the relationship between the variables included in the imputation [2]. First, we evaluated which variables could explain the mechanism of missingness (i.e., whether p16 was tested or not), and which variables had a strong correlation with the expected outcome (i.e., whether p16 was positive or negative). The variables, as it happens, were the same. This is a strength because it reduces the need to include too many auxiliary variables in the imputation. HPV status was most likely to be missing if patients had clinical and socio-demographic indicators of being HPV-, and this observation was consistent with clinical practice. We assumed that HPV status did not influence whether or not samples were tested in any other way apart from the variables included in the multiple imputation model. Under such a missingness pattern, complete-case analyses may be biased [3, 4]. The complete-case analysis in Figure 4a revealed no change over time in the prevalence of HPV+ OPC, an observation that is contrary to anecdotal experience and data from other jurisdictions. To address this limitation we calibrated the results using survival data, which was external to the multiple imputation (to choose a p16 positivity cut-point at which to consider a patient as p16+ or p16-). This was needed because the Pr(HPV+) values were, in general, too high. It was not surprising that we observed clear splitting of the survival curves by Pr(HPV+) quintiles, since the factors that were used to impute Pr(HPV+) are also strong prognostic factors. This cut-point was further validated using leave-one-out cross validation. Then, we performed wide sensitivity analyses making extreme assumptions about the multiple imputation cutpoints (Figure 4D), yet still showing the dominance of the conclusion that there has been a substantial rise in the rate of p16+ HPV related oropharyngeal cancer. The results are also consistent with clinical experience, and so we are confident that the biases inherent in multiple imputation of datasets with moderate to high missingness are not a major limitation in this report, yet one that merits careful interpretation of the rate of rise.

Appendix to: Habbous S, Chu KP, Lau H, et al. Human papillomavirus in oropharyngeal cancer in Canada: analysis of 5 comprehensive cancer centres using multiple imputation. *CMAJ* 2017. DOI:10.1503/cmaj.161379.

References

1. Dong Y, Peng CY. Principled missing data methods for researchers. *Springerplus* 2013;2:222.
2. Hardt J, Herke M, Leonhart R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Med Res Methodol* 2012;12:184.
3. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
4. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010;29:2920-31.