

1 **“Footprinting” missing epidemiological data for cervical cancer: a case study in India**

2

3 Irene Man*, Damien Georges, Maxime Bonjour, Iacopo Baussano

4

5 Early Detection, Prevention and Infections Branch, International Agency for Research on Cancer

6 (IARC/WHO), Lyon, France

7

8 * Corresponding author: Irene Man, Early Detection, Prevention and Infections Branch,

9 International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08,

10 France; tel: +33 4 7273 8017; email: mani@iarc.fr (IM)

11

12 **Running title:** Footprinting missing cervical cancer data

13 **Main text word count:** 3704

14 **Abstract word count:** 274 (max. 300)

15 **Number of references:** 31

16 **Tables:** 3

17 **Figures:** 3

18

19 **Abstract**

20 Background

21 Context-specific cervical cancer epidemiological data are essential to derive local impact
22 projections of cervical cancer preventive measures. However, these are not always available, in
23 particular in low- and middle-income countries (LMICs), where impact projections are essential
24 to plan cervical cancer control programs.

25 Methods and Findings

26 We developed a framework, hereafter named Footprinting, to approximate the sexual behavior,
27 human papillomavirus (HPV) prevalence, and/or cervical cancer incidence data needed for impact
28 projections. The framework was applied to a case study in India, the country with the highest
29 expected cervical cancer burden but still limited access to cervical cancer prevention. With our
30 Footprinting framework, we 1) identified clusters of Indian states with similar cervical cancer
31 incidence patterns, 2) classified states without incidence data to the identified clusters based on
32 similarity in sexual behavior data, 3) approximated missing cervical cancer incidence and HPV
33 prevalence data based on available data within each cluster. Two main patterns of cervical cancer
34 incidence, characterized by high and low incidence, were identified for 6 and 8 Indian states,
35 respectively. States in the low-incidence cluster were characterized by less sexual activity with
36 non-regular partners in men and earlier sexual debut in women. Based on these patterns, all 11
37 Indian states with missing cervical cancer incidence data were classified to the low-incidence
38 cluster. Finally, missing data on cervical cancer incidence and HPV prevalence were approximated
39 based on the mean of the available data within each cluster.

40 Conclusions

41 With the Footprinting framework, we enabled approximation of missing cervical cancer
42 epidemiological data and derivation of context-specific impact projection of cervical cancer
43 prevention measures, assisting public health decisions on cervical cancer prevention in India and
44 other LMICs.

45

46 Keywords

47 Cervical cancer incidence, HPV prevalence, sexual behavior, cervical cancer prevention, public
48 health decisions, impact projection, clustering, classification.

49

50 **Introduction**

51 Cervical cancer is an important source of disease burden worldwide.[1] In 2020, the number of
52 new cases and deaths due to cervical cancer worldwide were estimated to be 604,000 and
53 342,000, respectively.[2] Vaccination against human papillomavirus (HPV), cervical cancer
54 screening, and treatment of pre-cancer and cancer can reduce the burden of cervical cancer,[3-5]
55 but access to these prevention measures is still limited in many places of the world, especially in
56 low- and middle-income countries (LMICs).[6, 7] To encourage the scale-up of cervical cancer
57 prevention worldwide, the World Health Organization has developed a global strategy to
58 eliminate cervical cancer as a public health problem.[8] The strategy proposes an elimination
59 target of 4 cases per 100,000 women-years (age-standardized) and three intervention targets:
60 90% of girls vaccinated against HPV by age 15; 70% of women receiving twice-lifetime
61 screening with high-performance testing; and 90% of women having access to cervical pre-
62 cancer and cancer treatment, and palliative care.

63 In order for the aspirational global targets to be perceived as realistic, achievable, and
64 equitable, they must be adapted to local context.[9] The local need for and the impact of cervical
65 cancer prevention measures depend on the burden of cervical cancer in a given population,
66 which is determined by context-specific sexual behavior, and related to the HPV prevalence.[10]
67 Local data on these aspects are therefore crucial for deriving projections of the health and
68 economic impact of possible interventions. When based on adequate data, impact projections of
69 cervical cancer prevention measure can help local health authorities set adequate public health
70 targets and allocate resources accordingly.[11]

71 However, local epidemiological data for cervical cancer needed to derived impact
72 projections are sometimes missing. High-quality type- and age-specific data on HPV prevalence

73 and cervical cancer incidence from local populations are often unavailable. The same holds for
74 adequate data on sexual behavior, e.g., data on sex outside marriage, which are also prone to
75 bias, e.g., social desirability and recall bias.[12, 13] When essential epidemiological data for
76 projections are missing, there are two main possible solutions: collection or approximation of the
77 necessary data. Collection of new data in a local context would be the ideal option. However,
78 this could be time- and resource-demanding and therefore not always feasible. Alternatively,
79 missing data on a given population can be approximated using available data from populations
80 sharing similar characteristics.

81 In this paper, we propose a framework, hereafter named Footprinting, for approximating
82 missing cervical cancer epidemiological data for a selected number of geographical units within
83 a larger geographical target area where we would like to derive impact projections of cervical
84 cancer prevention. The framework is presented through a case study in India, the country with
85 the world's highest expected burden of cervical cancer [5] but only very limited access to
86 cervical cancer prevention measures.[14] To assist local public health decision-making in India,
87 we applied Footprinting to approximate missing Indian state-specific cervical cancer incidence
88 and HPV prevalence data and so to enable impact projections of cervical cancer preventive
89 measures with state-specific granularity.

90 **Methods and Materials**

91 *Background to the Indian case study*

92 India, the entire target geographical area of this case study, consists of 25 geographical units of
93 states or groups of states (both referred to as states hereafter). Detailed data on sexual behavior
94 are available in all Indian states through a national survey.[15] However, we only identified 14
95 out of 25 Indian states with cervical cancer incidence data [16, 17], and two Indian states with
96 the high-quality type- and age-specific HPV prevalence data as well as the cervical cancer
97 incidence data needed for impact projections.[18-20] See **Table S1** for the definition of the 25
98 states and an overview of data availability by state.

99

100 *Footprinting framework*

101 We propose a framework, Footprinting, to approximate missing data on the three key aspects of
102 cervical cancer epidemiology: sexual behavior, HPV prevalence, and cervical cancer incidence.
103 Missing data across geographical units are assumed to occur in a hierarchical manner, i.e., there
104 is an ordering of geographical units according to their levels of data availability (**Figure 1**). At
105 the highest level, there are a small number of geographic units for which data on all key aspects
106 are available. For the Indian case study, there were two such states. The remaining geographic
107 units are further divided into the levels of intermediate and low data availability. In the Indian
108 case study, there were 12 states with cervical cancer incidence and sexual behavior data, which
109 were assigned to the level of intermediate data availability. The remaining 11 states had only
110 sexual behavior data and were assigned to the low level of data availability. For reason that will
111 become clear shortly, the three data sources with increasing data availability are labelled as
112 “Bottleneck”, “Pattern” and “Footprint” data (**Figure 1**).

113 To address the hierarchical form of missing data, we propose a three-step approach, with the
114 successive steps labelled as “Clustering”, “Classification”, and “Projection”. In brief, the
115 approach identifies clusters of geographical units sharing similar patterns of cervical cancer
116 epidemiology and uses the available data within each cluster to approximate data and extrapolate
117 impact projections to geographical units with lower data availability. The details of the
118 respective steps are as follows.

119 *1. Clustering step*

120 In the Clustering step, clusters of geographical units sharing similar patterns of cancer
121 epidemiology are identified based on the Pattern data, which has large enough coverage
122 over the target geographical area. In the Indian case study, cervical cancer incidence data
123 were available in 14 out of 25 Indian states and hence suitable for this purpose. As a
124 result of the Clustering step, each Indian state of intermediate level of data availability
125 was matched with a state with high level of data availability that shared a similar pattern
126 of cancer incidence. In order for all states of intermediate data availability to be matched,
127 the number of clusters must be chosen so that each cluster contained at least one state
128 with high data availability.

129 *2. Classification step*

130 In the second step, geographical units of the lowest level of data availability, which were
131 not clustered in the previous step, are classified into the identified clusters. Classification
132 is based on the similarity between geographical units according to the Footprint data,
133 which should be available for the entire target geographical area, i.e., sexual behavior
134 data in the Indian case study. As in the Clustering step, the result of the Classification

135 step is that each Indian state with the lowest level of data availability was matched to
136 states with higher levels of data availability within the same cluster.

137 *3. Projection step*

138 In the last step, missing data are approximated based on the available data from other
139 geographical units within the same cluster, e.g., mean or median of the available data. If
140 the Classification step also provided the probability of belonging to each cluster,
141 approximation could even be based on weighted values of different clusters. With the
142 approximated data, it is then possible to construct projection models, i.e., HPV
143 transmission and cervical cancer progression models, and derive context-specific impact
144 projections for each geographical unit separately. Alternatively, a less computationally
145 demanding approach would be to construct projection models for the geographical units
146 of the highest level of data availability only and to, subsequently, scale the derived
147 projections to the other geographical units within the same clusters.

148 Data sources

149 In this section, we describe the data sources used in the India case study. The primary source of
150 cervical cancer incidence data, which was used as Pattern data in the Clustering step, was cancer
151 registry data from volume XI of Cancer Incidence in Five Continents (CI5).[16] It comprised
152 incidence data from 16 cancer registries in 10 of the 25 Indian states (some states had more than
153 one registry). In addition, cervical cancer incidence data were extracted from the 2012-2016
154 report by the Indian National Centre for Disease Informatics and Research (NCDIR) to provide
155 data from 17 additional cancer registries not included in CI5.[17] Combining the two sources
156 provided incidence data for 33 registries in 14 Indian states. Cervical cancer incidence was

157 reported in number of cases per 100,000 women-years, stratified by 5-year age groups from age
158 15 to 79 years. See **Figure S1** for the extracted incidence data by state.

159 Sexual behavior data, which were used as Footprint data in the Classification step, were from
160 the National Behavior Surveillance Survey by the National AIDS Control Organization of India
161 in 2006, which was the most recent edition at the moment of writing.[15] Data for all 25 Indian
162 states were available in the survey. Sexual behavior data by Indian state in the form of aggregate
163 statistics of survey respondents were available for the following 4 groups of 12 variables:

- 164 • *Median age of first sex* – stratified by residence (*urban/rural*) and sex (*male/female*),
165 resulting in 4 variables.
- 166 • *Proportion of respondents reporting sex with non-regular partners in the last 12 months*
167 – stratified by residence (*urban/rural*) and sex (*male/female*), resulting in 4 variables.
- 168 • *Proportion of male respondents reporting sex with commercial partners in the last 12*
169 *months* – stratified by residence (*urban/rural*), resulting in 2 variables.
- 170 • *Proportion of male respondents by number of commercial partners in the last 12 months*
171 – restricted to respondents with at least one commercial partner and divided into three
172 categories (*1/2-3/>3*). As the 3 proportions always sum up to one and are therefore
173 correlated, we omitted one category, resulting in 2 variables.

174 See **Figure S2** for the extracted sexual behavior data by state.

175

176 Statistical methods

177 The statistical method employed in the Clustering step to cluster registry-specific cervical cancer
178 incidence data was a Poisson-regression-based CEM clustering algorithm,[21, 22] described in
179 detail in **Appendix S1**. Briefly, clusters of age-specific cervical cancer incidence were obtained

180 by likelihood-based optimization under Poisson regression model. The Poisson regression model
181 for each cluster was characterized by three parameters: an intercept, one parameter for age, and
182 one for the square of age. This parametric form was chosen to match the general pattern of
183 incidence through age, namely, increasing from zero incidence from the youngest age group,
184 then decreasing after reaching a maximum (**Figure S1**). Application of the clustering method
185 required prefixing the number of clusters k . The goodness-of-fit of each k -clustering was
186 evaluated based on the Bayesian information criterion (BIC). To transform the obtained
187 clustering of registry-specific data to clustering of Indian states for states with multiple registries,
188 we assigned each state to the cluster that included the highest number of its registries, i.e.,
189 according to a majority rule.

190 In the Classification step, we assigned the remaining states without cervical incidence
191 data to the identified clusters based on Random Forest (RF) using the sexual behavior data as
192 Footprint data.[23] The RF classifier was constructed using sexual behavior data from states with
193 identified clusters. The predictive value of each variable was evaluated with the mean decrease
194 in accuracy, which expressed how much the accuracy of the model decreased if the variable were
195 excluded. The performance of the classification step was validated by both out-of-bag error
196 estimate and by applying the constructed classifier to the sexual behavior data from states with
197 identified clusters. Subsequently, the constructed classifier was applied to the sexual behavior
198 data from states without identified clusters, providing the probability to belong to each cluster.
199 Each state was classified to the cluster receiving the highest probability. Classification was
200 performed using the R package *party*. See **Appendix S2** for details of the RF-based classification
201 method.

202 Finally, in the projection step, missing data on cervical cancer incidence and HPV
203 prevalence were approximated based on the mean within each cluster. Derivation of impact
204 projections was reported elsewhere.[24, 25]

205

206 **Results**

207 *Clustering of cervical cancer incidence patterns*

208 Clusterings of registry-specific cervical cancer incidence were obtained for up to four prefixed
209 clusters (**Figure 2**). Model fit improved substantially when increasing the number of prefixed
210 clusters from two to three, with the BIC reducing from 6933 to 5700 (**Table 1**). Further increase
211 in the number of prefixed clusters to four only led to marginal improvement in model fit, with a
212 small reduction in BIC from 5700 to 5532, and a poorly defined cluster of only one registry.
213 With the number of prefixed clusters set at five, the clustering method no longer converged. We
214 concluded that two or three clusters were fitting to describe the patterns of cervical cancer
215 incidence in the available data.

216 The identified clusters of cervical cancer incidence differed in terms of magnitude of
217 incidence and location of maximum incidence (**Figure 2**). Cluster 1 of the 2-clustering had a low
218 maximum incidence of 47 cases per 100,000 women-years at age group 60-64 years, compared
219 to cluster 2 with its higher maximum incidence of 91 cases per 100,000 women-years at the
220 earlier age group of 55-59 years (**Figure 2, Table 1**). When allowing three clusters, an additional
221 cluster characterized by intermediate maximum incidence of 64 cases per 100,000 women-years
222 at age group 60-64 years was identified (**Figure 2, Table 1**). This additional cluster mainly
223 consisted of registries that had previously been assigned to the low-incidence cluster, i.e., cluster
224 1 of the 2-clustering, while having a relatively high incidence (**Figure 2, Table 2**). See
225 supplementary **Table S2** for additional details of the obtained clusterings.

226 The obtained clusterings of registries were then used to derive clusterings of Indian states
227 based on the majority rule (**Table 2**). When using the 2-clustering of registries, none of the states
228 were exclusively attributed to cluster 2, hence 2-clustering could not be used for the

229 classification step. When using 3-clustering, still none of the states were exclusively attributed to
230 cluster 2, however, 8 and 4 states were assigned to clusters 1 and 3, i.e., the clusters with low and
231 intermediate incidence, respectively. Hence, we combined cluster 2 and 3, i.e., the clusters with
232 high and intermediate incidence as the new “high-incidence” cluster, while keeping cluster 1 as
233 the “low-incidence” cluster. Note that, with these newly defined clusters, each cluster still
234 contained at least one state with the highest level of data availability, which was necessary for
235 the projection step. Note, however, that with the new definition of clusters, we could no longer
236 distinguish patterns of early or late peak of incidence.

237

238 *Classification to cervical cancer patterns based on sexual behavior data*

239 A RF classifier was constructed using the sexual behavior data corresponding to the states with
240 identified clusters. The variables with the highest, second, and third highest predictive values
241 were “*proportion of urban male respondents reporting sex with non-regular partners in the last*
242 *12 months*”, “*median age of first sex in rural males*”, and “*median age of first sex in urban*
243 *females*”, respectively (**Table S3**). In particular, there was a good separation between the high-
244 and low-incidence clusters in terms of “*proportion of urban male respondents reporting sex with*
245 *non-regular partners in the last 12 months*”, with high proportions associated with the high-
246 incidence cluster (**Figure 3**). High values of “*median age of first sex in females*” and low values
247 of “*median age of first age in males*” were also associated for the high-incidence cluster,
248 although the separation was less clear.

249 The estimated out-of-bag error of the constructed classifier was 29%. When applying the
250 constructed classifier to the Indian states with identified clusters, only Karnataka and Other
251 North Eastern States (2 of 14 states; 14%) were wrongly classified to the low-incidence cluster

252 (Table 3). Visualization shows that the sexual behavior data of these two states more resemble
253 the other states belonging to the low-incidence cluster, despite being clustered into the high-
254 incidence cluster (Figure S2, Figure 3).

255 Subsequently, the constructed classifier was applied to classify the remaining states
256 without cervical cancer incidence data and thus with unknown cluster. All 11 remaining Indian
257 states received a higher probability to belong to the low-incidence cluster (Table 3). Indeed,
258 Figure 3 shows that the sexual behavior data of the states with unknown cluster (indicated in
259 gray) were generally closer to the sexual behavior data of the states of the low-incidence cluster
260 (indicated in red) than those of the high-incidence cluster (indicated in blue). Hence, we
261 identified in total 19 and 6 states for the low- and high-incidence clusters, respectively.

262 Finally, missing cervical cancer incidence data and HPV prevalence were approximated
263 based on the mean within each cluster (Figure S3). Approximation of HPV prevalence was
264 based on the only one prevalence survey we could identify per cluster.(20,21) We verified that
265 the HPV prevalence reported by the survey corresponding to the high-incidence cluster was
266 higher than the prevalence reported by the one corresponding to the low-incidence cluster: HPV
267 prevalence of 16.9% vs. 9.8% (in women in the age range of approximately 20-60 years). This
268 1.7-fold difference in HPV prevalence was in the same order of magnitude as the 1.9-fold
269 difference we found for the age-standardized cervical cancer incidence between the two clusters
270 (17.9 vs. 9.01 cases per 100,000 women-years). The final step of deriving impact projections of
271 cervical cancer preventive measures for the whole of India with state-specific granularity was
272 reported elsewhere.[24, 25]

273

274

275 **Discussion**

276 In this paper, we developed the Footprinting framework to approximate missing cervical cancer
277 epidemiological data in some geographical units when deriving impact projections of cervical
278 cancer prevention measures for a larger geographical area. In brief, the framework identified
279 clusters of geographical units sharing similar patterns of cervical cancer epidemiology and uses
280 the available data within each cluster to approximate data and extrapolate impact projections to
281 geographical units with lower data availability. The framework was demonstrated by a case
282 study of approximating missing cervical cancer incidence and HPV prevalence data for a
283 selection of Indian states. With the application, we have derived, for the first time, impact
284 projections of cervical cancer prevention measures for the whole of India with state-specific
285 granularity.[24, 25]

286 Moreover, this work has also generated a deeper understanding of cervical cancer
287 epidemiology throughout the country. We found that India can be divided into two main groups
288 of 19 and 6 Indian states or groups of states that are characterized by low or high cervical cancer
289 incidence, respectively. As expected, and in line with previously studies, individuals, in
290 particular men, in high-incidence states have more sexual activity with non-regular partners,
291 including commercial partners, than in low-incidence states.[26, 27] While early sexual debut in
292 women has also previously been suggested to be associated with high cervical cancer incidence
293 and HPV positivity,[26, 27] it was associated with lower cervical cancer incidence in the dataset
294 we considered. We hypothesize that, for the Indian context, early sexual debut is common in
295 states with a larger rural population, among whom less sexual activity occurs with non-regular
296 partners, which is the main determining factor for a lower risk of cervical cancer. With the
297 urbanization of rural areas, which often entails evolving socio-cultural norms, it is possible that

298 more Indian states may shift to a high cancer incidence pattern, with an accompanying early peak
299 in incidence.[28]

300 It should be noted that our Footprinting framework is similar to other extrapolation
301 approaches previously used in model-based projection studies targeting large geographical areas
302 with missing data.[29, 30] While the rationale behind different extrapolation approaches are
303 similar, which is to approximate missing data from other geographical units with similar
304 epidemiological indicators, there are also differences. A strength of our framework is that it
305 relies on the observed patterns of epidemiology in the data to select key geographical units from
306 which impact projections are extrapolated to other units instead of working with a predefined
307 selection of key units. This allows the selection of key units that maximizes the representation of
308 different epidemiological patterns in the data. Moreover, it also helps to pinpoint geographical
309 units that could be interesting for future data collection efforts. Secondly, we made use of a
310 newly developed clustering method [21, 22] that is able to assess the similarities between age-
311 specific patterns of cervical cancer incidence, which have not been considered by previous
312 studies.

313 Our application of Footprinting to the Indian case study also bears some resemblance
314 with the extrapolation of cervical cancer incidence by GLOBOCAN in the process to obtain
315 nationwide estimates of cervical cancer incidence in India.[31] Essentially, in GLOBOCAN,
316 missing incidence was extrapolated based on the identity being urban or rural as a footprint,
317 while we used sexual behavior for this purpose and considered each state separately, which is
318 necessary for state-specific impact projections. As a result, we neglected the variation between
319 rural and urban areas within Indian states, which is a limitation of our analysis. We expect that
320 Footprinting with further stratification of states by rural/urban identity could improve the

321 approximation. Furthermore, our nationwide estimate of cervical cancer incidence derived from
322 aggregating the state-specific estimates (reported in a separate manuscript [24]) was lower than
323 the estimate reported by GLOBOCAN. This could be explained by the use of different methods
324 of extrapolation and the fact that we included data from 17 additional cancer registries with
325 relatively low incidence not included in GLOBOCAN estimates.

326 Various possible adaptations of the proposed Footprinting framework are worth
327 mentioning. Firstly, in the less than ideal situation where none of the relevant cervical cancer
328 epidemiology data are available in some geographical units, data on indicators of human
329 development and geographical location, could also be used as Footprint data. Secondly, while we
330 focused on epidemiological data for cervical cancer in this work, Footprinting could be used to
331 approximate missing economic data (e.g., treatment or vaccine delivery costs) that are needed to
332 assess the health economic impact of cervical cancer prevention measures, given that relevant
333 footprint data can be defined and collected.

334 Through this work, we have provided a comprehensive framework to deal with the
335 important and ubiquitous challenge of missing data on cervical cancer epidemiology. By using
336 the proposed framework, it is possible to derive robust yet context-specific impact projections for
337 cervical cancer preventive measures for a wide range of geographical settings. Such projections
338 can assist local health authorities in planning and implementing cervical cancer prevention that is
339 adapted to the local needs and resources and so intensify the efforts to reduce the high burden of
340 cervical cancer still existing in many countries in low-resource settings.

341 **Acknowledgments**

342 This study was funded by the Bill & Melinda Gates Foundation (grant numbers: OPP48979; INV-
343 039876). The funder had no role in study design, data collection and analysis, decision to publish,
344 or preparation of the manuscript. For the authors identified as personnel of the International
345 Agency for Research on Cancer or World Health Organization, the authors alone are responsible
346 for the views expressed in this article and they do not necessarily represent the decisions, policy
347 or views of the International Agency for Research on Cancer or World Health Organization. The
348 designations used and the presentation of the material in this Article do not imply the expression
349 of any opinion whatsoever on the part of WHO and the IARC about the legal status of any country,
350 territory, city, or area, or of its authorities, or concerning the delimitation of its frontiers or
351 boundaries.

352

353 **Competing interests**

354 The authors have declared that no competing interests exist.

355

356 **Contributors**

357 IB was responsible for funding acquisition and supervision of this study. IM, DG, and IB co-
358 designed and co-led the data curation, formal analysis, investigation, and visualization of the study
359 finding. MB contributed to the development of statistical methodology and validation of study
360 results. All authors had full access to all the data in the study and had final responsibility for the
361 decision to submit for publication. IM and IB drafted the original draft. All authors contributed to
362 review and editing and have approved the final manuscript.

363 **Data Availability Statement**

364 All data used in the present study were openly available at:

365 [https://www.aidsdatahub.org/sites/default/files/resource/national-bss-general-population-india-](https://www.aidsdatahub.org/sites/default/files/resource/national-bss-general-population-india-2006.pdf)

366 [2006.pdf](https://www.aidsdatahub.org/sites/default/files/resource/national-bss-general-population-india-2006.pdf) for the sexual behavior data published by the National AIDS Control Organisation

367 Ministry of Health and Family Welfare Government of India, at <http://ci5.iarc.fr> for the cervical

368 cancer incidence data published by the International Agency for Research on Cancer, and at

369 https://www.ncdirindia.org/All_Reports/Report_2020/resources/NCRP_2020_2012_16.pdf for

370 the cervical cancer incidence data published by the National Centre for Disease Informatics and

371 Research of India. All data produced in the present study are available upon reasonable request to

372 the authors.

373

374 **References**

- 375 1. de Martel C, Plummer M, Vignat J, Franceschi S. Worldwide burden of cancer attributable to
376 HPV by site, country and HPV type. *International Journal of Cancer*. 2017;141(4):664-70. doi:
377 10.1002/ijc.30716.
- 378 2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer
379 Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185
380 Countries. *CA Cancer Journal for Clinicians*. 2021;71(3):209-49. doi: 10.3322/caac.21660.
- 381 3. Lei J, Ploner A, Elfström KM, Wang J, Roth A, Fang F, et al. HPV Vaccination and the Risk of
382 Invasive Cervical Cancer. *New England Journal of Medicine*. 2020;383(14):1340-8. doi:
383 10.1056/nejmoa1917338.
- 384 4. Bouvard V, Wentzensen N, Mackie A, Berkhof J, Brotherton J, Giorgi-Rossi P, et al. The IARC
385 Perspective on Cervical Cancer Screening. *The New England journal of medicine*. 2021;385(20):1908-18.
- 386 5. Bonjour M, Charvat H, Franco EL, Piñeros M, Clifford GM, Bray F, et al. Global estimates of
387 expected and preventable cervical cancers among girls born between 2005 and 2014: a birth cohort
388 analysis. *The Lancet Public Health*. 2021;6(7):e510-21. doi: 10.1016/s2468-2667(21)00046-3.
- 389 6. Bruni L, Saura-Lázaro A, Montoliu A, Brotons M, Alemany L, Diallo MS, et al. HPV vaccination
390 introduction worldwide and WHO and UNICEF estimates of national HPV immunization coverage 2010–
391 2019. *Preventive Medicine*. 2021;144:106399. doi: 10.1016/j.ypmed.2020.106399.
- 392 7. de Sanjose S, Tsu VD. Prevention of cervical and breast cancer mortality in low-and middle-
393 income countries: A window of opportunity. *International Journal of Women's Health*. 2019;11:381-6.
394 doi: 10.2147/IJWH.S197115.
- 395 8. WHO, Global strategy to accelerate the elimination of cervical cancer as a public health problem
396 [cited 2021 29 October]. Available from: <https://www.who.int/publications/i/item/9789240014107>.
- 397 9. Tsu VD. Cervical cancer elimination: are targets useful? 2020;395(10224):539-40. doi:
398 10.1016/S0140-6736(20)30219-1.

- 399 10. Guan P, Howell-Jones R, Li N, Bruni L, De Sanjosé S, Franceschi S, et al. Human papillomavirus
400 types in 115,789 HPV-positive women: A meta-analysis from cervical infection to cancer. *International*
401 *Journal of Cancer*. 2012;131(10):2349-59. doi: 10.1002/ijc.27485.
- 402 11. Goldie SJ, Goldhaber-Fiebert JD, Garnett GP. Chapter 18: Public health policy for cervical
403 cancer prevention: The role of decision science, economic evaluation, and mathematical modeling.
404 *Vaccine*. 2006;24(SUPPL. 3):S155-S63. doi: 10.1016/j.vaccine.2006.05.112.
- 405 12. Morris M, Vu L, Leslie-Cook A, Akom E, Stephen A, Sherard D. Comparing estimates of
406 multiple and concurrent partnerships across population based surveys: Implications for combination HIV
407 prevention. *AIDS and Behavior*. 2014;18(4):783-90. doi: 10.1007/s10461-013-0618-6.
- 408 13. Kelly CA, Soler-Hampejsek E, Mensch BS, Hewett PC. Social desirability bias in sexual
409 behavior reporting: Evidence from an interview mode experiment in rural Malawi. *International*
410 *Perspectives on Sexual and Reproductive Health*. 2013;39(1):14-21. doi: 10.1363/3901413.
- 411 14. Sankaranarayanan R, Basu P, Kaur P, Bhaskar R, Singh GB, Denzongpa P, et al. Current status of
412 human papillomavirus vaccination in India's cervical cancer prevention efforts. *The Lancet Oncology*.
413 2019;20(11):e637-44. doi: 10.1016/S1470-2045(19)30531-5.
- 414 15. National Behavioural Surveillance Survey: General Population 2006. National AIDS Control
415 Organisation Ministry of Health and Family Welfare Government of India [cited 2021 1 February].
416 Available from: [https://www.aidsdatahub.org/sites/default/files/resource/national-bss-general-population-](https://www.aidsdatahub.org/sites/default/files/resource/national-bss-general-population-india-2006.pdf)
417 [india-2006.pdf](https://www.aidsdatahub.org/sites/default/files/resource/national-bss-general-population-india-2006.pdf).
- 418 16. Bray F, Colombet M, Mery L, Piñeros M, Znaor A, Zanetti R, et al. Cancer Incidence in Five
419 Continents, Vol. XI (electronic version). Lyon: International Agency for Research on Cancer 2017 [cited
420 2021 1 February]. Available from: <http://ci5.iarc.fr>.
- 421 17. Report of National Cancer Registry Programme 2012-2016. National Centre for Disease
422 Informatics and Research 2020 [cited 2021 1 October]. Available from:
423 https://www.ncdirindia.org/All_Reports/Report_2020/resources/NCRP_2020_2012_16.pdf.

- 424 18. Franceschi S, Rajkumar R, Snijders PJF, Arslan A, Mahé C, Plummer M, et al. Papillomavirus
425 infection in rural women in southern India. *British Journal of Cancer*. 2005;92(3):601-6. doi:
426 10.1038/sj.bjc.6602348.
- 427 19. Dutta S, Begum R, Mazumder D, Mandal SS, Mondal R, Biswas J, et al. Prevalence of Human
428 Papillomavirus in Women Without Cervical Cancer: A Population-based Study in Eastern India.
429 *International Journal of Gynecological Pathology*. 2012;31(2):178-83. doi:
430 10.1097/PGP.0b013e3182399391.
- 431 20. Kataria I, Bhandari P, Saraswathi LR, Siddiqui M, Sankaranarayanan R. Review of HPV
432 prevalence data in Indian by RTI Internation.
- 433 21. Subtil F, Boussari O, Bastard M, Etard JF, Ecochard R, Génolini C. An alternative classification
434 to mixture modeling for longitudinal counts or binary measures. *Statistical Methods in Medical Research*.
435 2017;26(1):453-70. doi: 10.1177/0962280214549040.
- 436 22. Klich A, Ecochard R, Subtil F. Trajectory clustering using mixed classification models. *Statistics*
437 *in Medicine*. 2021;40(15):3425-39.
- 438 23. Breiman L. Random Forests. *Machine Learning*. 2001;45(1).
- 439 24. Man I, Georges D, M de Carvalho T, Ray Saraswathi L, Bhandari P, Kataria I, et al. Evidence-
440 based impact projections of single-dose human papillomavirus (HPV) vaccination in India (Manuscript in
441 preparation). 2022.
- 442 25. de Carvalho TM, Man I, Georges D, Saraswati LR, Bhandari P, Kataria I, et al. Health economic
443 impact of the introduction of single-dose HPV vaccination in India (manuscript in preparation).
- 444 26. Vaccarella S, Franceschi S, Herrero R, Muñoz N, Snijders PJF, Clifford GM, et al. Sexual
445 behavior, condom use, and human papillomavirus: Pooled analysis of the IARC human papillomavirus
446 prevalence surveys. *Cancer Epidemiology Biomarkers and Prevention*. 2006;15(2):326-33. doi:
447 10.1158/1055-9965.EPI-05-0577.

- 448 27. Schulte-Frohlinde R, Georges D, Clifford GM, Baussano I. Predicting cohort-specific cervical
449 cancer incidence from population-based HPV prevalence surveys: a worldwide study. *American Journal*
450 *of Epidemiology*. 2021;191(3):402-12. doi: 10.1093/aje/kwab254.
- 451 28. Baussano I, Lazzarato F, Brisson M, Franceschi S. Human papillomavirus vaccination at a time
452 of changing sexual behavior. *Emerging Infectious Diseases*. 2016;22(1):18-23. doi:
453 10.3201/eid2201.150791.
- 454 29. Brisson M, Kim JJ, Canfell K, Drolet M, Gingras G, Burger EA, et al. Impact of HPV vaccination
455 and cervical screening on cervical cancer elimination: a comparative modelling analysis in 78 low-income
456 and lower-middle-income countries. *The Lancet*. 2020;395(10224):575-90. doi: 10.1016/S0140-
457 6736(20)30068-4.
- 458 30. Qendri V, Bogaards JA, Baussano I, Lazzarato F, Vänskä S, Berkhof J. The cost-effectiveness
459 profile of sex-neutral HPV immunisation in European tender-based settings: a model-based assessment.
460 *The Lancet Public Health*. 2020;5(11):e592-e603. doi: 10.1016/S2468-2667(20)30209-7.
- 461 31. Ferlay J, Ervik M, Lam F, Colombet M, Piñeros M, Znar A, et al. *Global Cancer Observatory:*
462 *Cancer Today*. Lyon, France: International Agency for Research on Cancer 2020 [cited 2022 15
463 February]. Available from: <https://gco.iarc.fr/today>.
- 464

Table 1. Estimated parameters of clusters of cervical cancer incidence patterns.

Number of prefixed clusters	BIC *	Cluster label <i>i</i>	Number (%) of registries in cluster	Maximum incidence §	Maximum incidence pattern	Maximum incidence age group †	Maximum incidence age group pattern
2	6933	1	27 (82%)	47 cases	Low	60-64 year	Late
		2	6 (18%)	91 cases	High	55-59 year	Early
3	5700	1	19 (58%)	38 cases	Low	60-64 year	Late
		2	5 (15%)	92 cases	High	55-59 year	Early
		3	9 (27%)	64 cases	Intermediate	60-64 year	Late
4	5532	1	18 (55%)	39 cases	Low	60-64 year	Late
		2	5 (15%)	92 cases	High	55-59 year	Early
		3	9 (27%)	64 cases	Intermediate	60-64 year	Late
		4	1 (3%)	20 cases	Very low	60-64 year	Early

* Bayesian information criterion for evaluating the goodness-of-fit of obtained clustering

§ Maximum incidence given in cases per 100,000 women-years

† Five-year age group in which the maximum incidence is located

Table 2. Clustering of cervical cancer incidence of Indian states based on clustering of registries.

State/group of states *	2-clustering		3-clustering			4-clustering			
	1 (low, late)	2 (high, early)	1 (low, late)	2 (high, early)	3 (interm., late)	1 (low, late)	2 (high, early)	3 (interm., late)	4 (very low, early)
Andhra Pradesh	•		•			•			
Assam	•••		•••			••			•
Delhi	•				•			•	
Gujarat + Dadra & Nagar Haveli	•		•			•			
Karnataka	•				•			•	
Kerala + Lakshadweep	••		••			••			
Madhya Pradesh	•				•			•	
Maharashtra	••••••	•	••••		•••	••••		•••	
Manipur	••		••			••			
Other North Eastern States §	•••••	••••	••••	••••	•	••••	••••	•	
Punjab + Chandigarh	•				•			•	
Sikkim	•		•			•			
Tamil Nadu + Puducherry	•	•		•	•		•	•	
West Bengal + Andaman & Nicobar Islands	•		•			•			

Each circle represents the count of one registry being assigned to the corresponding cluster. Gray shading represents the cluster including the highest number of registries, either exclusively or in a draw with another cluster.

Cluster labels and the corresponding patterns of maximum incidence and maximum incidence age group, given in the second row were defined in the third, sixth and eighth columns of Table 1, respectively.

* States/or groups of states were defined as reported in the 2006 National Behavior Surveillance Survey of the National AIDS Control Organization of India.[15]

§ Other North Eastern States included Arunachal Pradesh, Nagaland, Meghalaya, Mizoram, and Tripura.

Abbreviation: Interm, intermediate.

Table 3. Identified and classified cluster of cervical cancer incidence pattern by Indian state.

Cervical cancer incidence data ^o	State/group of states [*]	Identified cluster [†]	Classified cluster [‡]	Probability of belonging to the low-incidence cluster
Available	Andhra Pradesh	low	low	0.60
	Assam	low	low	0.69
	Delhi	high	high	0.42
	Gujarat + Dadra & Nagar Haveli	low	low	0.69
	Karnataka	high	low	0.63
	Kerala + Lakshadweep	low	low	0.60
	Madhya Pradesh	high	high	0.44
	Maharashtra	low	low	0.57
	Manipur	low	low	0.65
	Other North Eastern States§	high	low	0.53
	Punjab + Chandigarh	high	high	0.41
	Sikkim	low	low	0.63
	Tamil Nadu + Puducherry	high	high	0.38
	West Bengal + Andaman & Nicobar Islands	low	low	0.71
Unavailable	Bihar	-	low	0.67
	Chhattisgarh	-	low	0.66
	Goa + Daman & Diu	-	low	0.54
	Haryana	-	low	0.66
	Himachal Pradesh	-	low	0.58
	Jammu & Kashmir	-	low	0.63
	Jharkhand	-	low	0.71
	Orissa	-	low	0.68
	Rajasthan	-	low	0.66
	Uttar Pradesh	-	low	0.64
	Uttarakhand	-	low	0.69

^o Availability of cervical cancer incidence data was based on the incidence data from volume XI of Cancer Incidence in Five Continents (CI5) and the 2012-2016 report of the National Centre for Disease Informatics and Research (NCDIR).[16, 17]

^{*} States/groups of states were defined as reported in the 2006 National Behavior Surveillance Survey of the National AIDS Control Organization of India.[15]

[§] Other North Eastern States included Arunachal Pradesh, Nagaland, Meghalaya, Mizoram, and Tripura.

[†] Identified clusters derived in the Clustering step.

[‡] Classified clusters derived in the Classification step. A given state was classified to the low-incidence cluster if the probability of belonging to the low-incidence cluster (given in the next column) was above 0.50. For the Indian states with available cervical cancer incidence data and hence already in an identified cluster, classification was done for the purpose of validation.

Figure Legends

Figure 1. Hierarchical structure of availability of cervical cancer epidemiological data.

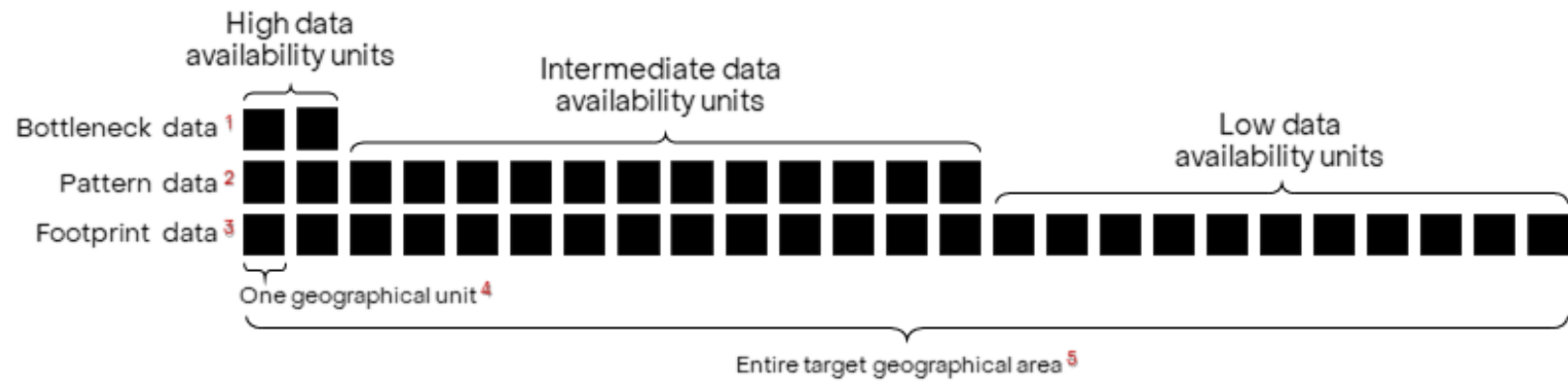
Figure 2. Identified clusters of registry-specific cervical cancer incidence.

Clusterings under (A) 2, (B) 3, and (C) 4 pre-fixed clusters. Each panel within a row corresponds to a cluster within a k -clustering, with the cluster label given on top of the panel. The cervical cancer incidence data were extracted from volume XI of Cancer Incidence in Five Continents (CI5) [16] and the 2012-2016 report by the Indian National Centre for Disease Informatics and Research (NCDIR) [17]. Black: cluster mean of cervical cancer incidence; Dark gray: registry incidence assigned to the cluster; Light gray: registry incidence assigned to other clusters.

Figure 3. Sexual behavior data from NACO by Indian state.

Indian state-specific data on (A) median age of first sex, (B) proportion of respondents reporting sex with non-regular partners in the last 12 months, (C) proportion of male respondents reporting sex with commercial partners in the last 12 months, and (D) proportion of male respondents by number of commercial partners in the last 12 months. Each violin plot and the associated cloud of circles correspond to a sexual behavior variable. Each circle corresponds to the data of a state (or group of states). The data were extracted from the 2006 National Behavior Surveillance Survey of the National AIDS Control Organization of India.[15] Blue and red: Indian states identified in the high and low cervical cancer incidence clusters. Gray: states without cervical cancer incidence data and therefore unknown cluster.

Figure 1. Hierarchical structure of availability of cervical cancer epidemiological data.



India case study

- 1 HPV prevalence data
- 2 Cervical cancer incidence data
- 3 Sexual behavior data
- 4 Indian state
- 5 India

Figure 2. Identified clusters of registry-specific cervical cancer incidence.

Clusterings under (A) 2, (B) 3, and (C) 4 pre-fixed clusters. Each panel within a row corresponds to a cluster within a k -clustering, with the cluster label given on top of the panel. The cervical cancer incidence data were extracted from volume XI of Cancer Incidence in Five Continents (CI5) [16] and the 2012-2016 report by the Indian National Centre for Disease Informatics and Research (NCDIR) [17]. Black: cluster mean of cervical cancer incidence; Dark gray: registry incidence assigned to the cluster; Light gray: registry incidence assigned to other clusters.

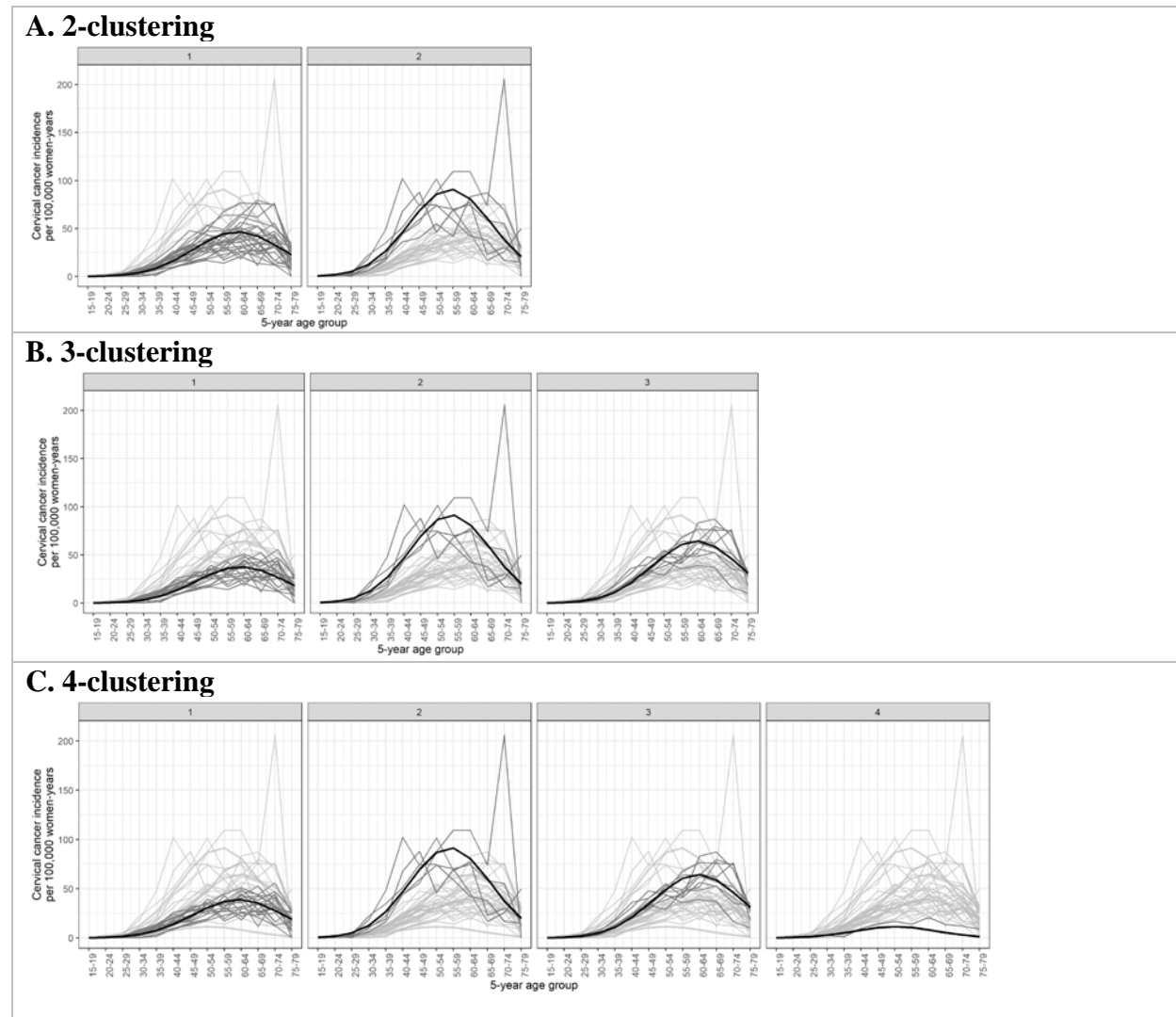


Figure 3. Sexual behavior data from NACO by Indian state.

Indian state-specific data on (A) median age of first sex, (B) proportion of respondents reporting sex with non-regular partners in the last 12 months, (C) proportion of male respondents reporting sex with commercial partners in the last 12 months, and (D) proportion of male respondents by number of commercial partners in the last 12 months. Each violin plot and the associated cloud of circles correspond to a sexual behavior variable. Each circle corresponds to the data of a state (or group of states). The data were extracted from the 2006 National Behavior Surveillance Survey of the National AIDS Control Organization of India.[15] Blue and red: Indian states identified in the high and low cervical cancer incidence clusters. Gray: states without cervical cancer incidence data and therefore unknown cluster.

