"Virtual controls" and Geolocalization to search for Environmental correlates of Hypospadias

Adrien de La Vaissière, David Baker,
Pierre Bougnères et Alain-Jacques Valleron
Inserm U1169 – Hôpital Bicêtre

BACKGROUND

Incidence of hypospadias varies considerably across countries, ranging from 4 to 43 cases per 10,000 births . Environmental factors might explain these differences. The classical approach is to use case-control studies and questionnaires to identify these factors. However, this approach suffers from the unavoidable arbitrariness of the definition of controls and from recall bias.

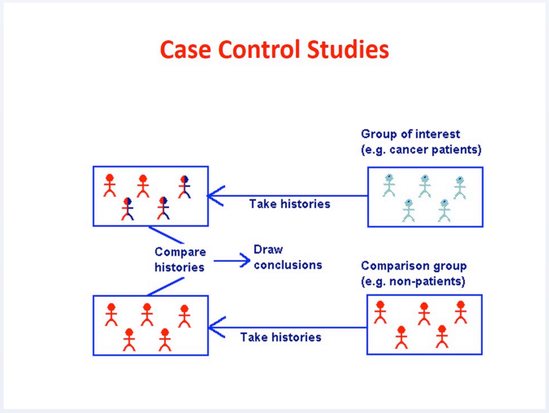

Figure 1: the past exposures to the studied environmental factor observed in a group of patients with the disease of interest ("cases") is compared with the past exposures of comparable subjects without the disease ("controls").

How define a good choice of controls?

Miettinen (1985) stated that a control should be someone who –if he had developed the disease- would have been recruited among the cases of the study. [1]

New opportunities for environmental epidemiology arise with the increasing availability of **public databases** informing on local environmental exposures. These public databases can be **mapped**, using geographic information systems (GIS) to the addresses **of geolocalized** patients of interest. This can be used to compare environmental exposures by mapping the addresses of the cases and of the controls to the environmental geographic databases of interest.

A more straightforward approach is to choose randomly a set of places on the map, then compute the exposure at these places, and compare the exposure of the cases of interest to this reference. In other words, the idea is to define "virtual controls" instead of "physical controls" to evaluate the reference exposure. This is appealing because:

1/ it avoids to get (costly) "physical controls" whose choice bears some arbitrariness

2/ the choice of virtual controls (defined by a given algorithm) is more flexible

Common sense tells however that these virtual controls cannot be just randomly sampled on the map, as they would in this cases most likely fall in uninhabited places, while cases would be in majority in high population density spots.

Instead of selecting arbitrarily a unique set of virtual controls, we sampled different algorithms for an optimal choice of "virtual controls".

OBJECTIVES

To describe our methodology as a proof-of-concept of a "virtual controls"approach to search for environmental markers (factors) of a disease, applied here to hypospadias.

METHODS

I- The starting point is to define the **population** in which the virtual controls will be sampled. This population must be identical to the population from which cases are extracted. The basic idea is that, if the controls would have got the disease, they would have had the same probability of being selected than cases.

When the cases are recruited by clinical centers, the base population is harder to define because it depends on location and characteristics of these clinical centers. Our cohort included 8766 cases of hypospadias coming from 16 specialized surgical centers.

We defined several different populations from which we draw the « virtual controls ». For exemple

- in squares around each case
- in circles containing k% of the cases around each clinical center.

We also considered in another sampling only the population living in rural locations. (thanks to INSEE definition [2])

II- The French Statistical body (INSEE) provides annual estimates of the population numbers by age-class in each square of a 200m x 200m grid covering France. The map can be viewed as a set of 2 278 213 pixels, each of them corresponding to an elementary square of the grid. Assume that the environmental exposures are known for 1000 subjects, and that 1000 virtual controls are needed, the problem is to appropriately choose at random 1000 pixels across the grid.

The random selection process chosen guarantees that the geographical density of the virtual controls be equal to the density of the population of reference. To achieve this goal, the probability for selecting each "pixel" was taken proportional to the population density in the "square" corresponding to this pixel. This was achieved by using the R package PPS developed by Jack G.[3] This algorithm provides virtual control subjects that mirror closely the general population.

III- We can now use the Corine Land Cover (CLC) database, which is an inventory of land cover in 44 classes. France is divided in 100m*100m squares, and in each of them, we have one of the 44 values (for example: Vineyards, Mixed forest, Discontinuous urban fabric …). In our study, we are going to focus on vineyards.

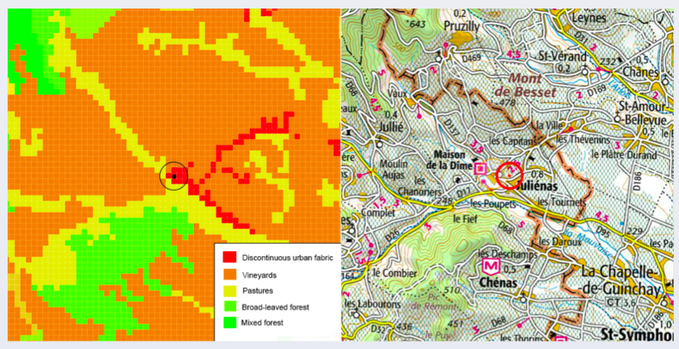

Fig. 2 – Example of a CLC map around the residence of a case

For each adress (case or "virtual controls"), we can draw circles and rings of different radius (250m, 500m, 1km, 2km, 5km…) around each point and check if there is any vineyard inside or not.

RESULTS

For every set of "virtuals controls", we compared the distance to the closest vineyards between cases and controls. Here are the results for one set of "virtuals controls" sampled in rural locations inside circles including 90% of rural cases around each clinical center.

	<250m	250/500m	500/1000m	1000/2000m	2000/5000m	>5000m (non exposed)
Hypospadias	126	169	201	252	397	1638
« virtual controls »	59	90	127	187	336	1698
Odds-ratio	2,21	1,95	1,64	1,40	1,22	
P-value	P < 0,0001	P < 0,0001	P < 0,0001	0,0011	0,0132	
Min(95%CI)	1,61	1,49	1,30	1,14	1,04	
Max(95%CI)	3,04	2,54	2,07	1,71	1,44	

Table 1 - odds-ratios comparing expositions to vineyards in the different rings for « virtuals controls » drawn in rural locations inside circles including 90% of rural cases around each clinical center.

CONCLUSION

The results are consistent with an association between hypospadias and vineyards. Nevertheless, we obtain similar results with other types of CLC, like forest. Further investigation is needed. We are also going to crosscheck those results with other databases, in particular the "parcel"[4] database.

Computational optimization is needed to reduce computation times.

REFERENCES

•[1] Miettinen, O. S. (1985). « The "case-control" study: valid selection of subjects. » J Chronic Dis 38(7): 543-548.
• A.J. Valleron / C.R. Acad. Sci. Paris, Sciences de la vie / Life Sciences 323 (2000) 617–628
• [2] http://www.insee.fr/fr/methodes/default.asp?page=zonages/unites_urbaines.htm
• [3] Gambino, J. G. (2005). « pps: Functions for PPS sampling. » http://CRAN.R-project.org/package=pps.
• [4] http://professionnels.ign.fr/bdparcellaire

55ESPE 346--P1
Gonads & DSD
Pierre Bougneres
DOI: 10.3252/pso.eu.55ESPE.2016
Poster presented at: ESPE 2016 PARIS