Environment International 135 (2020) 105316

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Validation of self-reported perception of proximity to industrial facilities: MCC-Spain study

Adela Castelló[a,b,c,*], Beatriz Pérez-Gómez[b,c], David Lora-Pablos[c,d,e], Virginia Lope[b,c], Gemma Castaño-Vinyals[c,f,g,h], Facundo Vitelli-Storelli[i,j], Trinidad Dierssen-Sotos[c,k], Pilar Amiano[c,l], Marcela Guevara[c,m], Víctor Moreno[c,n,o], Macarena Lozano-Lorca[p,q], Adonina Tardón[c,r,s], Juan Alguacil[c,t], Marta Hernández-García[u], Rafael Marcos-Gragera[c,v,w], Maria Dolores Chirlaque López[c,x], Eva Ardanaz[c,m], Jesús Ibarluzea[c,l,y], Inés Gómez-Acebo[c,k], Antonio J. Molina[i,j], Cristina O'Callaghan-Gordo[c,f,h], Nuria Aragonés[c,z], Manolis Kogevinas[c,f,g,h], Marina Pollán[b,c], Javier García-Pérez[b,c]

[a] School of Medicine, University of Alcalá, Av. de Madrid, Km 33,600, 28871 Alcalá de Henares, Madrid, Spain
[b] Cancer and Environmental Epidemiology Unit, Department of Epidemiology of Chronic Diseases, National Center for Epidemiology, Carlos III Institute of Health, Calle de Melchor Fernández Almagro, 5, 28029 Madrid, Spain
[c] Consortium for Biomedical Research in Epidemiology & Public Health (CIBER Epidemiología y Salud Pública – CIBERESP), Av. de Monforte de Lemos, 3-5, 28029 Madrid, Spain
[d] Clinical Research Unit (i+12), Hospital Universitario, 12 de Octubre, Av. de Córdoba, s/n, 28041 Madrid, Spain
[e] Spanish Clinical Research Network (SCReN), C/ Profesor Martín Lagos S/N, 28040 Madrid, Spain
[f] ISGlobal, Barcelona, Carrer del Rosselló, 132, 08036 Barcelona, Spain
[g] IMIM (Hospital del Mar Medical Research Institute), Carrer del Dr. Aiguader, 88, 08003 Barcelona, Spain
[h] Universitat Pompeu Fabra (UPF), Campus del Mar, Carrer del Dr. Aiguader, 80, 08003 Barcelona, Spain
[i] The Research Group in Gene – Environment and Health Interactions (GIIGAS)/Institut of Biomedicine (IBIOMED), Universidad de León, Campus Universitario de Vegazana, 24071 León, Spain
[j] Faculty of Health Sciences, Department of Biomedical Sciences, Area of Preventive Medicine and Public Health, Universidad de León, Campus Universitario de Vegazana, 24071 León, Spain
[k] Universidad de Cantabria – IDIVAL, Avenida Cardenal Herrera Oria s/n, 39011 Santander, Spain
[l] Public Health Division of Gipuzkoa, Biodonostia Health Research Institute, Ministry of Health of the Basque Government, P° Dr. Beguiristain s/n, 20014 San Sebastian, Spain
[m] Instituto de Salud Pública de Navarra, IdiSNA, Calle Leyre, 15, 31003 Pamplona, Spain
[n] Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology (ICO) and Oncobell Program, Bellvitge Biomedical Research Institute (IDIBELL), Hospital Duran i Reynals, Avinguda de la Gran Via de l'Hospitalet, 199-203, 08908 Hospitalet de Llobregat, Barcelona, Spain
[o] Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Carrer de Casanova, 143, 08036 Barcelona, Spain
[p] Department of Preventive Medicine and Public Health, School of Medicine, University of Granada, Av. de la Investigación, 11, 18016 Granada, Spain
[q] Granada Health Research Institute (ibs.GRANADA), Doctor Azpitarte 4 4ª Planta, Edificio Licinio de la Fuente, 18012 Granada, Spain
[r] Instituto Universitario de Oncología, Universidad de Oviedo. Facultad de Medicina, Oviedo, Spain
[s] Spain Instituto de Investigación Sanitaria del Principado de Asturias, Oviedo, Spain
[t] Centro de Investigación en Salud y Medio Ambiente (CYSMA), Universidad de Huelva, Campus Universitario de El Carmen, 21071 Huelva, Spain
[u] Cancer and Public Health Area, FISABIO – Public Health, Avda. de Catalunya, 21, 46020 Valencia, Spain
[v] Epidemiology Unit and Girona Cancer Registry, Oncology Coordination Plan, Department of Health, Autonomous Government of Catalonia, Catalan Institute of Oncology, Carrer del Sol, 15, 17004 Girona, Spain
[w] Descriptive Epidemiology, Genetics and Cancer Prevention Group, Biomedical Research Institute (IDIBGI), Carrer de Santiago Ramón y Cajal, 30, 17190 Salt, Girona, Spain
[x] Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Campus de Ciencias de la Salud, Carretera Buenavista s/n, 30120 El Palmar, Murcia,

Abbreviations: MCC-Spain, Multicase-control study in Spain; AUC, Area under the curve; aOR, adjusted odds ratio; 95%CI, 95% confidence interval; GIS, Geographic Information Systems

* Corresponding author at: School of Medicine, University of Alcalá, Av. de Madrid, Km 33,600, 28871, Alcalá de Henares, Madrid, Spain.
E-mail addresses: adela.castello@uah.es (A. Castelló), bperez@isciii.es (B. Pérez-Gómez), david@h12o.es (D. Lora-Pablos), vicarvajal@isciii.es (V. Lope), gemma.castano@isglobal.org (G. Castaño-Vinyals), fvites00@estudiantes.unileon.es (F. Vitelli-Storelli), trinidad.dierssen@unican.es (T. Dierssen-Sotos), epicss-san@euskadi.eus (P. Amiano), mguevare@cfnavarra.es (M. Guevara), v.moreno@iconcologia.net (V. Moreno), macarenalozano@ugr.es (M. Lozano-Lorca), atardon@uniovi.es (A. Tardón), alguacil@dbasp.uhu.es (J. Alguacil), hernandez_margarb@gva.es (M. Hernández-García), rmarcos@iconcologia.net (R. Marcos-Gragera), mdolores.chirlaque@carm.es (M.D. Chirlaque López), me.ardanaz.aicua@cfnavarra.es (E. Ardanaz), mambien3-san@euskadi.eus (J. Ibarluzea), ines.gomez@unican.es (I. Gómez-Acebo), ajmolt@unileon.es (A.J. Molina), cristina.ocallaghan@isglobal.org (C. O'Callaghan-Gordo), nuria.aragones@salud.madrid.org (N. Aragonés), manolis.kogevinas@isglobal.org (M. Kogevinas), mpollan@isciii.es (M. Pollán), jgarcia@isciii.es (J. García-Pérez).

https://doi.org/10.1016/j.envint.2019.105316

Available online 06 January 2020
0160-4120/ © 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

Spain
[y] School of Psychology, University of the Basque Country (UPV/EHU), Tolosa Hiribidea, 70, 20018 San Sebastián, Spain
[z] Epidemiology Section, Public Health Division, Department of Health of Madrid, C/San Martín de Porres, 6, 28035 Madrid, Spain

ARTICLE INFO

Handling Editor: Adrian Covaci

Keywords:
Self-reported perception
Residential proximity
Case-control study
Sensitivity
Specificity
AUC
Industrial pollution
MCC-Spain

ABSTRACT

Background: Self-reported data about environmental exposures can lead to measurement error.
Objectives: To validate the self-reported perception of proximity to industrial facilities.
Methods: MCC-Spain is a population-based multicase-control study of cancer in Spain that recruited incident cases of breast, colorectal, prostate, and stomach cancer. The participant's current residence and the location of the industries were geocoded, and the linear distance between them was calculated (gold standard). The epidemiological questionnaire included a question to determine whether the participants perceived the presence of any industry at ≤1 km from their residences. Sensitivity and specificity of individuals' perception of proximity to industries were estimated as measures of classification accuracy, and the area under the curve (AUC) and adjusted odds ratios (aORs) of misclassification were calculated as measures of discrimination. Analyses were performed for all cases and controls, and by tumor location, educational level, sex, industrial sector, and length of residence. Finally, aORs of cancer associated with real and self-reported distances were calculated to explore differences in the estimation of risk between these measures.
Results: Sensitivity of the questionnaire was limited (0.48) whereas specificity was excellent (0.89). AUC was sufficient (0.68). Participants with breast (aOR(95%CI) = 2.03 (1.67;2.46)), colorectal (aOR(95%CI) = 1.41 (1.20;1.64)) and stomach (aOR(95%CI) = 1.59 (1.20;2.10)) cancer showed higher risk of misclassification than controls. This risk was higher for lower educational levels ($aOR_{<\text{ primary } vs.\text{ university}}$ (95%CI) = 1.78 (1.44;2.20)), among younger participants ($aOR_{22\text{-}54\text{ years } vs.\ 73\text{-}85\text{ years}}$ (95%CI) = 1.32 (1.09;1.60)), and for some industrial sectors: pharmaceutical (aOR(95%CI) = 29.02 (19.52;43.14)), galvanization (aOR(95%CI) = 14.14 (6.78;29.47)), and ceramic (aOR(95%CI) = 12.73 (7.22;22.44)). Participants living ≤1 year in the study area showed a lower risk of misclassification (($aOR_{\leq 1\ vs.\ >15\text{ years}}$ (95%CI) = 0.56 (0.36;0.85)). The use of self-reported proximity *vs*. real distance to industrial facilities biased the effect on cancer risk towards the nullity.
Conclusions: Self-reported distance to industrial facilities can be a useful tool for hypothesis generation, but hypothesis-testing studies should use real distance to report valid conclusions. The sensitivity of the question might be improved with a more specific formulation.

1. Introduction

Epidemiological questionnaires usually include self-reported data about exposures of interest, such as diet, personal and familiar background information, or occupational history (Härmä et al., 2017; Kee et al., 2017; Naska et al., 2017; Sediq et al., 2018). The use of self-reported measurements can be very useful, since this information is inexpensive, feasible and relatively easy to obtain. However, if the collection method is not validated, the obtained information might be erroneous, leading to measurement error and potential biases when these data are used in studies about exposure and risk assessment (Hartge and Cahill, 2008).

In the case of studies that evaluate the effect of exposures to pollutant sources in health, it is often difficult to know the exact locations of individuals and/or hazardous sources (Cordioli et al., 2014; García-Pérez et al., 2008; Piro et al., 2008). In these situations, the use of a validated self-reported method to obtain the information is essential. Accordingly, some authors have explored the validity of self-reported residential exposure to environmental (Cordioli et al., 2014) and traffic (Cesaroni et al., 2008; Heinrich et al., 2005) pollution. Some of them showed a relatively high (Cesaroni et al., 2008; Cordioli et al., 2014) and others a weak (Heinrich et al., 2005) agreement between real and self-reported exposures. Similarly, Rull et al. (Rull et al., 2006) reported a weak agreement between observational and questionnaire measures of pesticide exposure and Daniau et al. (Daniau et al., 2013a, 2013b) claimed that the perception of pollution, especially sensory information such as odors, affects self-reported health. Also, some researchers, have found a positively biased risk of disease when using self-reported information on proximity to agricultural industries and crops (Handal et al., 2015) and others claimed a possible over reporting of pollution problems among individuals with respiratory problems (Piro et al., 2008). However, to our knowledge, no study assessed the concordance between perceived (self-reported) and observed (real distance) proximity to pollutant factories from different sectors.

In the present paper, we explored the validity of the self-reported perception of proximity to industrial facilities by sample characteristics and type of industrial sectors in the context of a population-based multicase-control study of incident cancers carried out in Spain (MCC-Spain).

2. Materials and methods

2.1. Study area and subjects

Details of the MCC-Spain study are described in detail elsewhere (Castaño-Vinyals et al., 2015). Briefly, between September 2008 and December 2013 histologically confirmed incident cases of breast (n = 1738), prostate (n = 1112), colorectal (n = 2140) and stomach (n = 459) cancers were recruited in 12 provinces from north, south, east, west, and central Spain to ensure geographical representation (map: http://www.mccspain.org/wp-content/uploads/2016/04/05_Mapa_nodos2.jpg). All participants were between the age of 20 and 85 and resided in the catchment areas of the hospitals for at least 6 months prior to recruitment. A common set of 3941 population-based controls were randomly selected from administrative records of selected primary care health centers located within the catchment areas of the hospitals, and were frequency matched to the overall distribution of cases by sex, age (in 5-year age groups), and region (province). The Ethics Committee of each participating center approved the study protocol and all participants signed an informed consent before the recruitment.

2.2. Data collection

Information on socio-demographic factors, lifestyle, and personal/family medical history was collected with a questionnaire administered by trained personnel in a face-to-face interview. Such questionnaire included the following question: "Is your current residence less than

1 km from a factory or industry?" (See Supplementary Material I for more detail). We classified as "yes", if the participant perceived there was some industry ≤1 km from her/his residence, or "no", if the participant did not perceive it. Missing responses were not included in the analyses.

2.3. Residential locations

Each participant's current residence was geocoded into Universal Transverse Mercator Zone 30 (ED50) coordinates using Google Earth Pro. All coordinates were individually checked using the "street-view" application included in Google Earth Pro, and the National Cadastre, which includes the names and numbers of streets and buildings.

2.4. Industrial facility locations

The information about industries governed by the Integrated Pollution Prevention and Control Directive and installations included in the European Pollutant Release and Transfer Register, corresponding to 2009, was used. Due to the identification of errors in the address of several industries, each one of them was thoroughly checked. Following the same methodology applied in previous studies (García-Pérez et al., 2019, 2008), we used the "street-view" application included in Google Earth Pro, the Google Maps server, the "Yellow pages" web page (which allows the search of addresses and companies), the Spanish Agricultural Plots Geographic Information System (which includes orthophotos and topographic maps showing the names of the industries) (Spanish Ministry of Agriculture and Fishing Food and Environment, 2019), and the web pages of the industries themselves to ensure that the location of the industrial facility was accurate. The final industrial database included information about the coordinates of the 2809 industrial facilities located in the study area.

2.5. Statistical analysis

The characteristics of the individuals and their proximity to industrial facilities were summarized using basic descriptive statistics (frequencies, medians, and interquartile range (IQR)).

For each participant, the shortest Euclidean distance between the coordinates of the individual's residence and the coordinates of any of the 2809 industrial facilities was calculated. This distance was used as the gold standard in the analyses.

As measures of classification accuracy, sensitivity (probability of the questionnaire to classify correctly those participants with residence at ≤1 km from any industry) and specificity (probability of the questionnaire to classify correctly those participants with residence at > 1 km from any industry), with their 95% confidence intervals (95%CIs) were calculated. The agreement between observed and perceived proximity to industrial facilities (discrimination of the question) was assessed by the area under the curve (AUC) and its 95%CI. These analyses were stratified by tumor location, educational level, age (quartiles in controls), sex, length in the current residence, and industrial sector.

Additionally, with the purpose of providing an adjusted measure of the discrimination, the risk of misclassification by sample characteristics and industrial sector was calculated. To obtain these estimates, three types of analyses were performed using binary mixed logistic regression models with the province as a random effects term. For all analyses, the outcome (misclassification) was defined as: 0 = participant correctly classified (if she/he lived and perceived ≤1 km or lived and perceived > 1 km), or 1 = participant misclassified (in the other cases):

(a) First analysis: the adjusted odds ratios (aORs) of misclassification (based on the distance to any type of industry) and 95%CIs by tumor location were obtained with 4 different models, one for each tumor, adjusted by educational level, age, length of residence and, in the case of stomach and colorectal cancer, also by sex. The aORs for educational level, age, sex, and length of residence were estimated including in a different model these three variables and the case/control status.
(b) Second analysis: we defined the observed proximity to each industrial sector as: 0 = No industries of any sector ≤1 km; 1 = at least one industry of the industrial sector in question ≤1 km. The

Table 1
Description of sample characteristics, observed and perceived proximity to an industrial facility, and distance to the nearest industry by type of tumor.

	Breast Cancer		Prostate Cancer		Colorectal and stomach cancer		
	Controls	Cases	Controls	Cases	Controls	Colorectal Cases	Stomach Cases
Educational level n (%)							
University graduate	416 (21%)	312 (19%)	420 (22%)	169 (15%)	836 (22%)	206 (11%)	48 (11%)
Secondary school	607 (31%)	554 (33%)	537 (28%)	239 (22%)	1144 (29%)	380 (21%)	94 (21%)
Primary school completed	594 (30%)	552 (33%)	649 (34%)	434 (39%)	1243 (32%)	695 (38%)	177 (39%)
Less than primary school	336 (17%)	265 (16%)	329 (17%)	258 (23%)	665 (17%)	560 (30%)	132 (29%)
Age, n (%)							
73–85	369 (19%)	220 (13%)	556 (29%)	230 (21%)	925 (24%)	649 (35%)	172 (38%)
65–72	378 (19%)	236 (14%)	624 (32%)	411 (37%)	1002 (26%)	447 (24%)	106 (24%)
55–64	432 (22%)	445 (26%)	540 (28%)	388 (35%)	972 (25%)	476 (26%)	85 (19%)
22–54	774 (40%)	782 (46%)	215 (11%)	71 (6%)	989 (25%)	269 (15%)	88 (20%)
Sex, n (%)							
Male	0 (0%)	0 (0%)	1935 (100%)	1100 (100%)	1935 (50%)	1177 (64%)	304 (67%)
Female	1953 (100%)	1683 (100%)	0 (0%)	0 (0%)	1953 (50%)	664 (36%)	147 (33%)
Length of residence							
> 15 years	1251 (64%)	966 (57%)	1406 (73%)	828 (75%)	2657 (68%)	1362 (74%)	303 (67%)
11–15 years	202 (10%)	204 (12%)	162 (8%)	75 (7%)	364 (9%)	113 (6%)	43 (10%)
6–10 years	268 (14%)	234 (14%)	189 (10%)	98 (9%)	457 (12%)	185 (10%)	50 (11%)
1–5 years	184 (9%)	217 (13%)	132 (7%)	77 (7%)	316 (8%)	133 (7%)	48 (11%)
≤1 year	47 (2%)	61 (4%)	39 (2%)	19 (2%)	86 (2%)	44 (2%)	7 (2%)
Observed and perceived proximity to an industrial facility, n (%)							
Observed distance ≤1 km	239 (12.2%)	291 (17.3%)	184 (9.5%)	133 (12.1%)	423 (10.9%)	314 (17.1%)	70 (15.5%)
Perceived distance ≤1 km	257 (13.2%)	335 (19.9%)	273 (14.1%)	167 (15.2%)	530 (13.6%)	374 (20.3%)	74 (16.4%)
Distance to the nearest industry in km, median (IQR)							
All individuals	2.65 (2.30)	2.19 (2.15)	2.63 (2.13)	2.20 (1.92)	2.64 (2.20)	2.21 (2.32)	2.36 (2.31)
Among those who perceived ≤1 km	1.25 (2.10)	1.24 (1.46)	1.78 (1.94)	1.48 (1.51)	1.55 (2.05)	1.19 (1.45)	1.27 (1.40)
Among those who perceived > 1 km	2.77 (2.27)	2.42 (2.12)	2.75 (2.11)	2.34 (1.99)	2.76 (2.19)	2.48 (2.46)	2.57 (2.48)

aORs of misclassification (based on the distance to any sector-specific industry) and 95%CIs by industrial activity were estimated in 19 independent models including each sector-specific misclassification variable in one of them and adjusting the estimations by case control status, educational level, age, sex, and length of residence.

(c) Third analysis: additionally, the differences in the aOR of misclassification between cases and controls by educational level, age, sex, length of residence, and industrial sector were assessed, including an interaction term between these variables and the case-control status in each of the models described before.

(d) Fourth analysis: to explore possible differences in the effect of self-reported and real distance to industrial facilities on cancer risk, we estimated the OR of breast, prostate, colorectal, and stomach cancer associated with these two measurements. For this purpose, we fitted eight binary logistic regression models, one for each tumor *vs*. self-reported/real distance combination, adjusted by educational level, age length of residence and, in the case of stomach and colorectal cancer, also by sex.

Finally, the following sensitivity and supplementary analyses were carried out: in order to evaluate the effect of the number of industries close by in the global percentage of misclassification, the percentage of incorrect self-perception of proximity to an industrial facility among individuals living close to one, two and three or more industrial facilities was calculated. We also provided a comparison of the median distance and interquartile range (IQR) between individuals who perceived living ≤1 km from an industry and those who perceived living > 1 km by industrial sector as a complementary descriptive material (Supplementary Material II). Additionally, to explore the possible bias in sector-specific self-perception caused by the number of industries or the presence of industries from other sectors different from the one under study, the models from the third analysis were adjusted excluding individuals living close to more than one industry (Supplementary Material III).

All analyses were performed with Stata/MP 15.0.

3. Results

After accounting for the missing information in the perceived distance from industrial pollutant sources, 1683 breast cancer cases, 1100 prostate cancer cases, 1841 colorectal cancer cases and 451 stomach cancer cases were included in the study. For the analyses of stomach and colorectal cancer, we used the full sample of 3888 controls, for breast cancer, we selected only the 1953 female controls, and for the analyses of prostate cancer, we selected only the 1935 male controls.

Table 1 shows the description of the sample by case/control status. Cases were less educated than controls for all cancer locations explored. Regarding the age, women with breast cancer (mean ± sd = 59.02 ± 13.25, data not shown) and men with prostate cancer (mean ± sd = 66.05 ± 7.35) were younger than controls (mean ± sd = 56.42 ± 12.67 and mean ± sd = 66.23 ± 9.71 respectively),

Table 2
Measures of classification accuracy (sensitivity and specificity) and discrimination (area under the curve and adjusted odds ratio) of misclassification using observed ("gold standard") and perceived (questionnaire) proximity to industries (≤1 km), by control/tumor location, educational level, age, sex, and length of residence.

	Observed/Perceived ≤1 km								
	No/No	Yes/Yes	No/Yes	Yes/No	Sens (95%CI)	Spec (95%CI)	AUC[a] (95%CI)	%Misc[a]	OR[b] (95%CI)
Total	6843	591	889	640	0.48 (0.45;0.51)	0.89 (0.88;0.89)	0.68 (0.67;0.70)	17.1%	
Breast Cancer									
Controls	1576	119	138	120	0.50 (0.43;0.56)	0.92 (0.91;0.93)	0.71 (0.68;0.74)	13.2%	1
Cases	1199	142	193	149	0.49 (0.43;0.55)	0.86 (0.84;0.88)	0.67 (0.64;0.70)	20.3%	2.03 (1.67;2.46)
Prostate cancer									
Controls	1563	85	188	99	0.46 (0.39;0.54)	0.89 (0.88;0.91)	0.68 (0.64;0.71)	14.8%	1
Cases	857	57	110	76	0.43 (0.34;0.52)	0.89 (0.86;0.91)	0.66 (0.61;0.70)	16.9%	0.95 (0.76;1.18)
Colorectal and stomach cancer									
Controls	3139	204	326	219	0.48 (0.43;0.53)	0.91 (0.90;0.92)	0.69 (0.67;0.72)	14.0%	1
Colorectal cases	1310	157	217	157	0.50 (0.44;0.56)	0.86 (0.84;0.88)	0.68 (0.65;0.71)	20.3%	1.41 (1.20;1.64)
Stomach cases	338	31	43	39	0.44 (0.32;0.57)	0.89 (0.85;0.92)	0.66 (0.60;0.73)	18.2%	1.59 (1.20;2.10)
Educational level									
University graduate	1342	49	115	65	0.43 (0.34;0.53)	0.92 (0.91;0.93)	0.68 (0.63;0.72)	11.5%	1
Secondary school	1862	151	245	153	0.50 (0.44;0.55)	0.88 (0.87;0.90)	0.69 (0.66;0.72)	16.5%	1.42 (1.17;1.72)
Primary school completed	2279	252	324	246	0.51 (0.46;0.55)	0.88 (0.86;0.89)	0.69 (0.67;0.71)	18.4%	1.68 (1.39;2.03)
Less than primary school	1360	139	205	176	0.44 (0.39;0.50)	0.87 (0.85;0.89)	0.66 (0.63;0.68)	20.3%	1.78 (1.44;2.20)
Age[c]									
73–85	1709	116	207	164	0.41 (0.36;0.47)	0.89 (0.88;0.91)	0.65 (0.62;0.68)	16.9%	1
65–72	1710	134	231	127	0.51 (0.45;0.58)	0.88 (0.87;0.90)	0.70 (0.67;0.73)	16.3%	0.95 (0.80;1.12)
55–64	1778	155	251	182	0.46 (0.41;0.51)	0.88 (0.86;0.89)	0.67 (0.64;0.70)	18.3%	1.16 (0.98;1.37)
22–54	1646	186	200	167	0.53 (0.47;0.58)	0.89 (0.88;0.91)	0.71 (0.68;0.74)	16.7%	1.32 (1.09;1.60)
Sex									
Male	3468	268	481	299	0.47 (0.43;0.51)	0.88 (0.87;0.89)	0.68 (0.65;0.70)	17.3%	1
Female	3375	323	408	341	0.49 (0.45;0.53)	0.89 (0.88;0.90)	0.69 (0.67;0.71)	16.8%	1.05 (0.93;1.19)
Length of residence									
> 15 years	4580	416	678	442	0.48 (0.45;0.52)	0.87 (0.86;0.88)	0.68 (0.66;0.70)	18.3%	1
11–15 years	633	50	61	55	0.48 (0.38;0.58)	0.91 (0.89;0.93)	0.69 (0.65;0.74)	14.5%	0.79 (0.64;0.98)
6–10 years	818	61	76	69	0.47 (0.38;0.56)	0.91 (0.89;0.93)	0.69 (0.65;0.74)	14.2%	0.74 (0.61;0.90)
1–5 years	622	50	56	63	0.44 (0.35;0.54)	0.92 (0.89;0.94)	0.68 (0.63;0.73)	15.0%	0.78 (0.63;0.97)
≤1 year	179	13	14	11	0.54 (0.33;0.74)	0.93 (0.88;0.96)	0.73 (0.63;0.84)	11.5%	0.56 (0.36;0.85)

[a] AUC: Area under the curve; %Misc: Percentage of misclassification.

[b] Adjusted odds ratio of misclassification, including classification (0 = Correct; 1 = Misclassified) as the outcome and: – case/control for breast and prostate cancer (in two different models), educational level, age, and length of residence as independent variables to estimate the coefficients for breast and prostate cancer. – case/control for colorectal and stomach cancer (in two different models), educational level, age, sex, and length of residence as independent variables to estimate the coefficients for colorectal and stomach cancer. – case/control status, educational level, age, sex, and length of residence as independent variables to estimate the coefficients for educational level, age, and sex.

[c] Quartiles of the distribution among controls.

while colorectal (mean ± sd = 66.77 ± 10.86) and stomach cancer cases (mean ± sd = 59.02 ± 13.25) were older than controls (mean ± sd = 66.37 ± 12.31). For these two last tumors, the percentage of males was bigger than the percentage of women. The percentage of individuals who have resided for less than 1 year was similar for all subgroups explored except for breast cancer cases, which showed a higher proportion of just arrived residents. The prevalence of individuals whose current residence was located at ≤1 km from any industry was lower for controls than for cases (10.9% for all controls, 17.3% for breast cancer cases, 12.1% for prostate cancer cases, 17.1% for colorectal cancer cases, and 15.5% for stomach cancer cases). The median distance was also smaller for cases than for controls (2.64 km for all controls, 2.19 km for breast cancer cases, 2.20 km for prostate cancer cases, 2.21 km for colorectal cancer cases, and 2.36 km for stomach cancer cases)). Similarly, this median distance was smaller among those who perceived living ≤1 km from an industry than among those who perceived living > 1 km for all subgroups explored (see Supplementary Material II for data on median distance by industrial sector).

Measures of classification accuracy (sensitivity and specificity) and discrimination (AUC) using observed ("gold standard") and perceived (questionnaire) proximity to industries (≤1 km) are shown in Table 2 (by control/tumor location, educational level, age, and sex) and Table 3 (by industrial sector). As can be seen in Table 2, the capability of the questionnaire to classify correctly those participants with residence at ≤1 km from an industrial facility (sensitivity) was limited for all subgroups explored (0.48 for the total population, and figures between 0.41 and 0.54 for the rest of the variables), while the validity of the questionnaire to classify correctly those participants with residence at > 1 km (specificity) was excellent (0.89 for the total population, and figures between 0.86 and 0.93 for the rest of the variables). The general capability of the questionnaire to classify the individuals into ≤1 km or > 1 km from an industrial facility was only sufficient (0.68 for the total, and values between 0.65 and 0.73 for the rest of the variables) according to the current cut-off points (Šimundić, 2009), mainly due to the low sensitivity of the questionnaire. Breast cancer cases showed a two-fold increase in the odds of misclassification with respect to controls. This increase was 41% and 59% for stomach and colorectal cancer cases, respectively. Similarly, the odds of misclassification was higher for the youngest participants, with those aged 22–54 showing an 32% higher odds than those aged 73–85. Additionally, the odds of misclassification was 78% higher among those with less than primary school, 68% higher among those with primary school and 42% higher among those with secondary school education than among university graduates. Finally, when we compared with participants who resided in the study area for > 15 years, those who reside for 11–15 years, 6–10 years, 1–5 years, and ≤ 1 year showed a 21%, 26%, 22%, and 44% lower odds of misclassification, respectively.

As for the self-reported perception of the presence of industrial facilities according to the industrial sector (Table 3), our results showed no important aORs of misclassification for industries of inorganic chemical sectors, mining, non-hazardous waste, combustion, and cement and lime in concordance with the higher sensitivity, specificity, and AUC values showed for these industrial facilities. For the rest of the industrial sectors, the aORs were substantial and especially striking for industries of pharmaceutical products (29 times higher odds of misclassification), galvanization (> 14 times) and ceramic (> 12 times), followed by installations belonging to the surface treatment using organic solvents, urban waste-water treatment, food and beverage, ship building, surface treatment of metals and plastic, hazardous waste, and paper and wood production sectors (with odds of misclassification between 5 and 8 times bigger). In all cases, the high misclassification risk was explained by the low awareness of the individuals in the sample about their proximity to industries (low sensitivity of the questionnaire). Specificity was the same for all industrial types, since the question on self-reported proximity to industries did not differentiate by sectors.

The stratified analyses by case-control status (Table 4) revealed that

Table 3
Measures of classification accuracy (sensitivity and specificity) and discrimination (area under the curve and adjusted odds ratio) of misclassification using observed ("gold standard") and perceived (questionnaire) proximity to industries (≤1km) by industrial sector.

	Observed/Perceived ≤1 km								
	No/No[a]	Yes/Yes	No/Yes[a]	Yes/No	Sensitivity (95%CI)	Specificity (95%CI)	AUC[b] (95%CI)	%Misc[b]	aOR[c] (95%CI)
Industrial sector									
Any industrial sector	6843	591	889	640	0.48 (0.45;0.51)	0.89 (0.88;0.89)	0.68 (0.67;0.70)	17.1%	7.49 (6.50;8.64)
Combustion installations	6843	36	889	11	0.77 (0.62;0.88)	0.89 (0.88;0.89)	0.83 (0.76;0.89)	11.6%	1.52 (0.75;3.08)
Production and processing of metals	6843	174	889	90	0.66 (0.60;0.72)	0.89 (0.88;0.89)	0.77 (0.74;0.80)	12.2%	3.83 (2.85;5.14)
Galvanization	6843	15	889	22	0.41 (0.25;0.58)	0.89 (0.88;0.89)	0.65 (0.56;0.73)	11.7%	14.14 (6.78;29.47)
Surface treatment of metals and plastic	6843	165	889	206	0.44 (0.39;0.50)	0.89 (0.88;0.89)	0.66 (0.64;0.69)	13.5%	6.67 (5.29;8.41)
Mining industry	6843	31	889	4	0.89 (0.73;0.97)	0.89 (0.88;0.89)	0.89 (0.83;0.94)	11.5%	1.32 (0.45;3.85)
Cement and lime	6843	27	889	6	0.82 (0.65;0.93)	0.89 (0.88;0.89)	0.85 (0.78;0.92)	11.5%	2.15 (0.87;5.36)
Glass and mineral fibers	6843	30	889	23	0.57 (0.42;0.70)	0.89 (0.88;0.89)	0.73 (0.66;0.79)	11.7%	2.38 (1.35;4.18)
Ceramic	6843	23	889	34	0.40 (0.28;0.54)	0.89 (0.88;0.89)	0.64 (0.58;0.71)	11.9%	12.73 (7.22;22.44)
Organic chemical industry	6843	56	889	44	0.56 (0.46;0.66)	0.89 (0.88;0.89)	0.72 (0.67;0.77)	11.9%	3.88 (2.54;5.94)
Inorganic chemical industry	6843	5	889	1	0.83 (0.36;1.00)	0.89 (0.88;0.89)	0.86 (0.70;1.00)	11.5%	0.75 (0.09;6.59)
Pharmaceutical products	6843	41	889	151	0.21 (0.16;0.28)	0.89 (0.88;0.89)	0.55 (0.52;0.58)	13.1%	29.02 (19.52;43.14)
Hazardous waste	6843	53	889	45	0.54 (0.44;0.64)	0.89 (0.88;0.89)	0.71 (0.66;0.76)	11.9%	5.35 (3.42;8.36)
Non-hazardous waste	6843	8	889	2	0.80 (0.44;0.97)	0.89 (0.88;0.89)	0.84 (0.71;0.97)	11.5%	1.49 (0.31;7.23)
Disposal or recycling of animal waste	6843	23	889	18	0.56 (0.40;0.72)	0.89 (0.88;0.89)	0.72 (0.65;0.80)	11.7%	2.54 (1.34;4.81)
Urban waste-water treatment plants	6843	9	889	18	0.33 (0.17;0.54)	0.89 (0.88;0.89)	0.61 (0.52;0.70)	11.7%	7.62 (3.33;17.45)
Paper and wood production	6843	42	889	34	0.55 (0.43;0.67)	0.89 (0.88;0.89)	0.72 (0.66;0.78)	11.8%	5.18 (3.14;8.55)
Food and beverage sector	6843	48	889	39	0.55 (0.44;0.66)	0.89 (0.88;0.89)	0.72 (0.67;0.77)	11.9%	7.59 (4.81;11.96)
Surface treatment using organic solvents	6843	66	889	56	0.54 (0.45;0.63)	0.89 (0.88;0.89)	0.71 (0.67;0.76)	12.0%	7.66 (5.09;11.52)
Ship building	6843	19	889	17	0.53 (0.35;0.70)	0.89 (0.88;0.89)	0.71 (0.62;0.79)	11.7%	7.53 (3.74;15.19)

[a] Since the self-perception question is not sector-specific, those who report not living close (n = 6843) and those who report living near (n = 889) to an industrial facility among the total number of individuals who do not reside at ≤1 km from an industrial facility (n = 7732) are the same for all industrial sectors.

[b] AUC: Area under the curve; %Misc: Percentage of misclassification.

[c] Adjusted odds ratio of misclassification, including classification (0 = Correct; 1 = Misclassified) as the outcome, proximity to each industrial group as the main exposure, and case/control, educational level, age, sex, and length of residence as potential confounders.

Table 4
Adjusted odds ratio of misclassification using observed ("gold standard") and perceived (questionnaire) proximity to industries (≤1 km) by educational level, age, sex, length of residence, and industrial sector, and separately for cases and controls.

	Observed/Perceived ≤1 km				CONTROLS		Observed/Perceived ≤1 km				CASES		
	No/No	Yes/Yes	No/Yes	Yes/No	%Misc[a]	OR[b] (95%CI)	No/No	Yes/Yes	No/Yes	Yes/No	%Misc[a]	OR[b] (95%CI)	p-int
Educational level													0.006
University graduate	745	14	60	17	9%	1	597	35	55	48	14.0%	1	
Secondary school	952	55	89	48	12%	1.33 (0.99;1.79)	910	96	156	105	20.6%	1.45 (1.13;1.87)	
Primary school completed	950	85	124	84	17%	2.11 (1.59;2.81)	1329	167	200	162	19.5%	1.43 (1.12;1.83)	
Less than primary school	492	50	53	70	18%	2.17 (1.57;2.98)	868	89	152	106	21.2%	1.55 (1.19;2.03)	
Age													0.971
≤54	745	40	75	65	15%	1	964	76	132	99	18.2%	1	
55–64	827	46	92	37	13%	0.84 (0.64;1.09)	883	88	139	90	19.1%	1.03 (0.83;1.27)	
65–72	783	45	90	54	15%	1.05 (0.81;1.36)	995	110	161	128	20.7%	1.23 (1.00;1.51)	
> 72	784	73	69	63	13%	1.18 (0.89;1.55)	862	113	131	104	19.4%	1.42 (1.13;1.79)	
Sex													0.049
Male	1563	85	188	99	15%	1	1905	183	293	200	19.1%	1	
Female	1576	119	138	120	13%	0.91 (0.75;1.10)	1799	204	270	221	19.7%	1.15 (0.99;1.33)	
Length of residence													
> 15 years	2125	138	250	144	0.148	1	2455	278	428	298	21.0%	1	0.971
11–15 years	294	24	25	21	0.126	0.83 (0.59;1.16)	339	26	36	34	16.1%	0.77 (0.58;1.02)	
6–10 years	378	23	28	28	0.123	0.79 (0.58;1.08)	440	38	48	41	15.7%	0.71 (0.55;0.91)	
1–5 years	263	13	20	20	0.127	0.83 (0.58;1.18)	359	37	36	43	16.6%	0.76 (0.58;0.99)	
≤1 year	71	6	3	6	0.105	0.59 (0.29;1.20)	108	7	11	5	12.2%	0.54 (0.31;0.92)	
Industrial sector	No/No[c]	Yes/Yes	No/Yes[c]	Yes/No			No/No[c]	Yes/Yes	No/Yes[c]	Yes/No			
Any industrial sector	3139	204	326	219	14%	9.92 (7.81;12.60)	3704	387	563	421	19.4%	6.48 (5.45;7.70)	0.004
Combustion installations	3139	11	326	0	9%	–	3704	25	563	11	13.3%	1.81 (0.87;3.77)	
Production and processing of metals	3139	95	326	35	10%	3.37 (2.18;5.20)	3704	79	563	55	14.0%	4.22 (2.89;6.15)	0.425
Galvanization	3139	7	326	12	10%	22.68 (8.27;62.16)	3704	8	563	10	13.4%	8.57 (3.16;23.23)	0.166
Surface treatment of metals and plastic	3139	30	326	34	10%	6.91 (4.09;11.67)	3704	135	563	172	16.1%	6.62 (5.12;8.56)	0.883
Mining industry	3139	23	326	1	9%	0.52 (0.07;3.90)	3704	8	563	3	13.2%	2.91 (0.75;11.32)	0.131
Cement and lime	3139	17	326	3	9%	2.21 (0.63;7.74)	3704	10	563	3	13.2%	2.10 (0.56;7.81)	0.954
Glass and mineral fibers	3139	2	326	0	9%	–	3704	28	563	23	13.6%	2.52 (1.42;4.48)	
Ceramic	3139	9	326	15	10%	17.68 (7.41;42.23)	3704	14	563	19	13.5%	10.00 (4.82;20.75)	0.318
Organic chemical industry	3139	16	326	18	10%	7.18 (3.46;14.92)	3704	40	563	26	13.6%	2.85 (1.69;4.81)	0.042
Inorganic chemical industry	3139	0	326	0	9%	–	3704	5	563	1	13.2%	0.75 (0.09;6.59)	
Pharmaceutical products	3139	8	326	87	12%	144.07 (65.20;318.34)	3704	33	563	64	14.4%	11.64 (7.40;18.31)	0.000
Hazardous waste	3139	23	326	17	10%	7.09 (3.58;14.03)	3704	30	563	28	13.7%	4.49 (2.58;7.79)	0.292
Non-hazardous waste	3139	1	326	0	9%	–	3704	7	563	2	13.2%	1.63 (0.33;8.07)	
Disposal or recycling of animal waste	3139	1	326	1	9%	5.60 (0.34;93.06)	3704	22	563	17	13.5%	2.44 (1.26;4.70)	0.576
Urban waste-water treatment plants	3139	1	326	0	9%	–	3704	8	563	18	13.5%	8.20 (3.48;19.32)	
Paper and wood production	3139	2	326	1	9%	6.93 (0.56;85.81)	3704	40	563	33	13.7%	5.12 (3.08;8.53)	0.820
Food and beverage sector	3139	9	326	7	10%	9.28 (3.29;26.16)	3704	39	563	32	13.7%	7.24 (4.38;11.98)	0.673
Surface treatment using organic solvents	3139	38	326	24	10%	6.07 (3.39;10.87)	3704	28	563	32	13.8%	9.39 (5.44;16.23)	0.267
Ship building	3139	7	326	9	10%	16.70 (5.77;48.38)	3704	12	563	8	13.3%	4.16 (1.64;10.55)	0.051

[a] %Misc: Percentage of misclassification.

[b] Adjusted odds ratio of misclassification, including classification (0 = Correct; 1 = Misclassified) as the outcome and: – educational level, age,sex, and length of residence as independent variables interacting with case/control status in 4 separate models to calculate the aORs for the categories of these variables. – proximity to each industrial group as the main exposure interacting with case/control status and tumor location, educational level, age, sex, and length of residence as potential confounders to calculate the aORs by industrial sector.

[c] Since the self-perception question is not sector-specific, those who report not living close (n = 6843) and those who report living near (n = 889) to an industrial facility among the total number of individuals who do not reside at ≤1 km from an industrial facility (n = 7732) are the same for all industrial sectors.

there are differences between cases and controls regarding the aORs of misclassification by educational level, sex and some industrial sectors. Even if the direction of the aOR of misclassification by educational level was the same for both groups it seemed to be stronger among controls. The same happened to the industrial sectors: those with the highest aORs of misclassification were the same for both cases and for controls, but the magnitude of the association was significantly smaller among cases. As for the case of sex, the risk of misclassification was higher in women than in men but only among the cases.

Finally, the results revealed that the ORs of misclassification for breast, prostate, colorectal, and stomach cancer were stronger when using the real distance to industrial facilities (aOR_{breast} (95%CI): 1.88 (1.53;2.33); $aOR_{prostate}$(95%CI): 1.13 (0.87;1.48); $aOR_{colorectal}$ (95%CI): 1.55 (1.30;1.84); $OR_{stomach}$ (95%CI): 2.23 (1.62;3.08)) than when using the self-reported distance (aOR_{breast} (95%CI): 1.64 (1.36;1.98); $aOR_{prostate}$ (95%CI): 0.86 (0.68;1.08); $aOR_{colorectal}$ (95%CI): 1.60 (1.36;1.87); $OR_{stomach}$ (95%CI): 1.30 (0.98;1.74)) (data only shown in the text).

4. Discussion

To our knowledge, this is the first study assessing the concordance between observational (real) and self-reported measures of proximity to industries from different sectors. In summary, sensitivity of the questionnaire (capability to classify correctly proximity to industries) was limited, whereas specificity (capability to classify correctly non-proximity to industries) was excellent. The risk of misclassification was higher among cases of breast, colorectal, and stomach cancer, among the youngest participants, those with lower education, individuals more rooted in the study area and for industries of pharmaceutical products, galvanization, ceramic, surface treatment using organic solvents, urban waste-water treatment, food and beverage, ship building, surface treatment of metals and plastic, hazardous waste, and paper and wood production sectors. The misclassification resulting from the self-perceived proximity to industrial facilities biased the associations with cancer risk towards the null hypothesis for all explored types of cancer.

Breast cancer cases and, to a lesser extent, colorectal, and stomach cancer cases, showed a higher risk of misclassification than controls after adjusting for education, age, sex, and length of residence. A priori, it would be expected a better accuracy in discriminating the proximity to pollution sources and a certain tendency to overestimate the exposure among cases since, after the diagnosis, cancer patients overthink about past exposures that might be related to the disease (Coughlin, 1990). This is true for colorectal cancer cases that showed a lower underestimation of the exposure (higher sensitivity) and a higher overestimation the exposure (lower specificity). Breast and stomach cancer cases showed lower sensitivity and lower specificity than controls, showing that in these groups, both exposure and non-exposure are partially wrongly classified. As for the case of prostate cancer, cases showed lower sensibility and same specificity than controls, suggesting that in this group only the exposure is underestimated. The stronger aOR of breast, prostate, colorectal, and stomach cancer risk observed when using the measured *vs.* self-reported distance to industrial facilities is concordant with sensitivity and specificity data. Sensitivity is bad for all groups. Therefore, many cases living near an industrial facility do not report proximity to pollutant sources, which biases the risk towards the null. Even though the lower specificity indicates an overestimation of the exposure that might positively bias the risk, the ability of the questionnaire to classify correctly the non-exposed, was very good. Therefore, the lower sensitivity has a bigger impact in the results than the moderate specificity, explaining why the observed effect of proximity to industrial facilities on cancer risk is lower when using self-reported data.

The overestimation of exposure among diseased people (Daniau et al., 2013a, 2013b; Piro et al., 2008) and the lower concordance between observed and self-reported neighborhood data among less educated and younger people (Bailey et al., 2014) found in this work, have been reported in previous studies. Our results, which indicate that more rooted people have a higher risk of misclassification (mainly due to an overestimation of the exposure) than those that just arrived in the study area, are in agreement with the results published by other authors (Shi and He, 2012). These authors claim that the length of residence is linked to a higher self-perception of air, water and noise pollution, which can overestimate the real exposure. On the other hand, Avruskin et al. (Avruskin et al., 2008) claimed that the number of years in residence do not influence the agreement between the measured and self-reported proximity estimations. In our case, we believe that a possible explanation for this is that people who moved recently has explored more thoroughly the area and are more conscious about their real distance to industries.

The results by industrial sector indicated a higher awareness of proximity to industries from the cement and lime, mining, non-hazardous waste, combustion, and inorganic chemical sectors, which can be explained by different factors. Some of these industrial sectors, such as mining or combustion installations, are located in northern areas of Spain, which have a long tradition of industrial production and group different type of industries in specific zones. In a sensitivity analysis (data not shown), we found that the percentage of misclassification decreased as the number of close industries increased (percentage of misclassification: 60%, 43%, and 36% for proximity to one, two, and three or more industrial facilities respectively). In relation to mining, some authors have found a greater perception of pollution in residents living in coal mine areas (Shi and He, 2012) possibly due to the economical and working dependence of the inhabitants of these areas of this industrial activity. Another explanation can be the great extension of these industrial areas and the high visibility of the equipment used for their exploitation. In the case of combustion or cement plants, they have tall smokestacks, which are visible from large distances by the nearest populations. Concerning non-hazardous waste industries, the malodorous emissions they release have a large radius of spread (Cheng et al., 2019), so populations residing close to these pollution sources could be more aware of their presence nearby. On the other hand, their poor visibility and their small smokestacks can explain the lack of awareness of residential proximity to industries of pharmaceutical products, galvanization, and ceramic.

Some authors have previously focused the problem of validation of self-reported exposure to traffic-related pollution exclusively (Cesaroni et al., 2008; Gunier et al., 2006; Heinrich et al., 2005), proximity to crops (Avruskin et al., 2008; Rull et al., 2006), or occupational exposures (Hu et al., 2002; Perry et al., 2006). Some of these studies report a good agreement between self-reported and real measurements (Cesaroni et al., 2008; Gunier et al., 2006; Heinrich et al., 2005; Hu et al., 2002) while others claim that the validity of self-reported data is weak (Avruskin et al., 2008; Perry et al., 2006; Rull et al., 2006). In relation to industrial pollution sources, there are few studies and the existing focus in a single type of activity (Shi and He, 2012) or in the general perception of industrial pollution not considering the type of activity (Cordioli et al., 2014). Shi et al. (Shi and He, 2012) evaluated the association between the self-perceived air, water and noise pollution (scale of 1–5) with the real proximity to coal industries (< 1 km; 1–2 km; 2–3 km; 3–4 km; 4–5 km; and > 5 km). They found no significant correlation of measured distance to a coal mining area with self-perceived air pollution and a slight negative relationship with self-reported water and noise pollution. Cordioli et al. (Cordioli et al., 2014) evaluated the association between self-reported type of zone of residence (rural/residential/industrial) and GIS-derived proxy for exposure to industrial pollutants (presence/absence of emissions of total suspended particles and volatile organic compounds into the atmosphere inside three buffers within a radius of 100, 500, and 1000 m from each residence). The results showed a mild agreement between the two methods for 100 m buffer (50%) and a strong agreement for 500 (94%) and 1000 m (100%) buffers. However, only 18 residences were

within the industrial area and the conclusions were based on descriptive analyses and unadjusted bivariate tests over categorical variables. In our study, the question included in the questionnaire was more detailed, including the distance between the residence and the industry: "Was your current residence less than 1 km from a factory or industry?" and the gold standard was the measured Euclidean distance between the last individual's residence and the industrial installation.

One of the main lessons or conclusions obtained from our study is that individuals do not perceive or identify the presence of nearby industries (low sensitivity). A similar result was presented in Marcon et al. (2015) with a secondary result of a study of health risk perception in environmental surveys. This study revealed that, despite the fact that 37.3% of the participants have their residence close (< 1 km) to a chipboard or wood factory, only 5.8% reported living in industrial areas. The differences in the measurement standard set for self-reported and real distance might be behind this low sensitivity and the high specificity of the question in our questionnaire. Measured distance is calculated as the shortest straight line (Euclidean distance) between the participant's residence and the industrial facility. However, it is highly probable that when participants answer the question about self-perception they think about the geographical distance they need to travel to get to the facility, which in most cases is longer than the Euclidian distance. As a result, in most cases when the Euclidian distance is > 1 km, geographical distance also is > 1 km making participants correctly classify themselves as further away from an industrial facility (high specificity). However, in some cases the Euclidian distance might be ≤1 km but geographical distance might be > 1 km, which makes participants to classify themselves as further when they are ≤1 km in a straight line from the facility (low sensitivity). Therefore, the reformulation of the question as "Is your current residence less than 1 km in a straight line from a factory or industry?" might improve its sensibility.

Regarding the limitations of the present study, it is important to note that, even if we explored the validity of self-reported proximity to industries by type of sector, the question did not differentiate by industrial activity. Therefore, sector-specific self-perception might be biased by the number of industries in a 1 km radius or by the presence of industries from other sectors different from the one under study. To explore this possibility, we carried out some additional analyses excluding those individuals that reside ≤1 km from more than one industrial facilities, and the conclusions reached with the new results are similar to those obtained with the full dataset (see Supplementary material III). On the other hand, some industrial sectors such as inorganic chemical industry or non-hazardous waste treatment plants presented an insufficient number of individuals living nearby to obtain reliable conclusions. As for the rest of the industrial sectors, numbers are also moderate, so the analyses by industrial activity need to be interpreted with caution and should be reproduced in future studies. Special attention to these limitations must be paid out for the analyses of cases and controls separately.

One of the main strengths of our study is the large sample size and the completeness of its analyses, which consisted of an in-depth survey study of validation stratified by tumor, educational level, sex, length of residence, and industrial sector. Another strength was the exhaustive process of address and industrial geocoding and validation carried out in the study, which has allowed us to have the exact coordinates of individuals and industrial installations.

5. Conclusions

The results showed a high specificity and a low sensitivity of the question to classify correctly proximity to industries. Self-reported perception to industrial facilities only partially captures the reality and results in an underestimation of the real association between this exposure and cancer risk. Therefore, the current question might be a useful tool for hypothesis generation or pilot studies, but not for studies testing scientific hypotheses. However, the sensitivity of the self-reported data might be improved with a more specific formulation: "Is your current residence less than 1 km in a straight line from a factory or industry?"

Declaration of Competing Interest

The authors declared that there is no conflict of interest.

Acknowledgments

The authors thank all those who took part in this study providing questionnaire data. This study was funded by: Scientific Foundation of the Spanish Association Against Cancer (*Fundación Científica de la Asociación Española Contra el Cáncer (AECC)* – EVP-1178/14), Spain's Health Research Fund (*Fondo de Investigación Sanitaria* – FIS 12/01416), the Spanish Ministry of Economy and Competitiveness (Carlos III Institute of Health; PI12/00488, PI12/00265, PI12/01270, PI12/00715, PI12/00150, PI08/1770, PI08/0533, PI08/1359, PS09/00773, PS09/01286, PS09/01903, PS09/02078, PS09/01662, PI11/01403, PI11/01889, PI11/00226, PI11/01810, PI11/02213, PI14/01219 and Río Hortega CM13/00232), the Catalan Government 2009SGR1489 & 2014SGR756-F, the Fundación Marqués de Valdecilla (API 10/09), the ICGC International Cancer Genome Consortium CLL, the Junta de Castilla y León (LE22A10-2), the Consejería de Salud of the Junta de Andalucía (PI-0571), the Conselleria de Sanitat of the Generalitat Valenciana (AP 061/10), the Recercaixa (2010ACUP 310), and the Regional Government of the Basque Country by European Commission grants FOOD-CT-2006-036224-HIWATE.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2019.105316.

References

Avruskin, G.A., Meliker, J.R., Jacquez, G.M., 2008. Using satellite derived land cover information for a multi-temporal study of self-reported recall of proximity to farmland. J. Expo. Sci. Environ. Epidemiol. 18, 381–391. https://doi.org/10.1038/sj.jes.7500610.

Bailey, E.J., Malecki, K.C., Engelman, C.D., Walsh, M.C., Bersch, A.J., Martinez-Donate, A.P., Peppard, P.E., Nieto, F.J., 2014. Predictors of discordance between perceived and objective neighborhood data. Ann. Epidemiol. 24, 214–221. https://doi.org/10.1016/j.annepidem.2013.12.007.

Castaño-Vinyals, G., Aragonés, N., Pérez-Gómez, B., Martín, V., Llorca, J., Moreno, V., Altzibar, J.M., Ardanaz, E., de Sanjosé, S., Jiménez-Moleón, J.J., Tardón, A., Alguacil, J., Peiró, R., Marcos-Gragera, R., Navarro, C., Pollán, M., Kogevinas, M., MCC-Spain Study Group, 2015. Population-based multicase-control study in common tumors in Spain (MCC-Spain): rationale and study design. Gac. Sanit. 29, 308–315. https://doi.org/10.1016/j.gaceta.2014.12.003.

Cesaroni, G., Badaloni, C., Porta, D., Forastiere, F., Perucci, C.A., 2008. Comparison between various indices of exposure to traffic-related air pollution and their impact on respiratory health in adults. Occup. Environ. Med. 65, 683–690. https://doi.org/10.1136/oem.2007.037846.

Cheng, Z., Sun, Z., Zhu, S., Lou, Z., Zhu, N., Feng, L., 2019. The identification and health risk assessment of odor emissions from waste landfilling and composting. Sci. Total Environ. 649, 1038–1044. https://doi.org/10.1016/j.scitotenv.2018.08.230.

Cordioli, M., Ranzi, A., Freni Sterrantino, A., Erspamer, L., Razzini, G., Ferrari, U., Gatti, M.G., Bonora, K., Artioli, F., Goldoni, C.A., Lauriola, P., 2014. A comparison between self-reported and GIS-based proxies of residential exposure to environmental pollution in a case-control study on lung cancer. Spat. Spatio-Temporal Epidemiol. 9, 37–45. https://doi.org/10.1016/j.sste.2014.04.004.

Coughlin, S.S., 1990. Recall bias in epidemiologic studies. J. Clin. Epidemiol. 43, 87–91. https://doi.org/10.1016/0895-4356(90)90060-3.

Daniau, C., Dor, F., Eilstein, D., Lefranc, A., Empereur-Bissonnet, P., Dab, W., 2013a. Study of self-reported health of people living near point sources of environmental pollution: a review. First part: health indicators. Rev. Epidemiol. Sante Publique 61, 375–387. https://doi.org/10.1016/j.respe.2013.02.014.

Daniau, C., Dor, F., Eilstein, D., Lefranc, A., Empereur-Bissonnet, P., Dab, W., 2013b. Study of self-reported health of people living near point sources of environmental pollution: a review. Second part: analysis of results and perspectives. Rev. Epidemiol. Sante Publique 61, 388–398. https://doi.org/10.1016/j.respe.2013.05.010.

García-Pérez, J., Boldo, E., Ramis, R., Vidal, E., Aragonés, N., Pérez-Gómez, B., Pollán, M.,

López-Abente, G., 2008. Validation of the geographic position of EPER-Spain industries. Int. J. Health Geogr. 7, 1. https://doi.org/10.1186/1476-072X-7-1.
García-Pérez, J., Gómez-Barroso, D., Tamayo-Uria, I., Ramis, R., 2019. Methodological approaches to the study of cancer risk in the vicinity of pollution sources: the experience of a population-based case-control study of childhood cancer. Int. J. Health Geogr. 18, 12. https://doi.org/10.1186/s12942-019-0176-x.
Gunier, R.B., Reynolds, P., Hurley, S.E., Yerabati, S., Hertz, A., Strickland, P., Horn-Ross, P.L., 2006. Estimating exposure to polycyclic aromatic hydrocarbons: a comparison of survey, biological monitoring, and geographic information system-based methods. Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. 15, 1376–1381. https://doi.org/10.1158/1055-9965.EPI-05-0799.
Handal, A.J., McGough-Maduena, A., Páez, M., Skipper, B., Rowland, A.S., Fenske, R.A., Harlow, S.D., 2015. A pilot study comparing observational and questionnaire surrogate measures of pesticide exposure among residents impacted by the ecuadorian flower industry. Arch. Environ. Occup. Health 70, 232–240. https://doi.org/10.1080/19338244.2013.879563.
Härmä, M., Koskinen, A., Ropponen, A., Puttonen, S., Karhula, K., Vahtera, J., Kivimäki, M., 2017. Validity of self-reported exposure to shift work. Occup. Environ. Med. 74, 228–230. https://doi.org/10.1136/oemed-2016-103902.
Hartge, P., Cahill, J., 2008. Field methods in epidemiology. In: Rothman, K.J., Greenland, S., Lash, T.L. (Eds.), Modern Epidemiology, third ed. Lippincott Williams & Wilkins, Philadelphi, PA, USA, pp. 492–510.
Heinrich, J., Gehring, U., Cyrys, J., Brauer, M., Hoek, G., Fischer, P., Bellander, T., Brunekreef, B., 2005. Exposure to traffic related air pollutants: self reported traffic intensity versus GIS modelled exposure. Occup. Environ. Med. 62, 517–523. https://doi.org/10.1136/oem.2004.016766.
Hu, Y.A., Smith, T.J., Xu, X., Wang, L., Watanabe, H., Christiani, D.C., 2002. Comparison of self-assessment of solvent exposure with measurement and professional assessment for female petrochemical workers in China. Am. J. Ind. Med. 41, 483–489. https://doi.org/10.1002/ajim.10069.
Kee, C.C., Lim, K.H., Sumarni, M.G., Teh, C.H., Chan, Y.Y., Nuur Hafizah, M.I., Cheah, Y.K., Tee, E.O., Ahmad Faudzi, Y., Amal Nasir, M., 2017. Validity of self-reported weight and height: a cross-sectional study among Malaysian adolescents. BMC Med. Res. Methodol. 17, 85. https://doi.org/10.1186/s12874-017-0362-0.
Marcon, A., Nguyen, G., Rava, M., Braggion, M., Grassi, M., Zanolin, M.E., 2015. A score for measuring health risk perception in environmental surveys. Sci. Total Environ. 527–528, 270–278. https://doi.org/10.1016/j.scitotenv.2015.04.110.
Naska, Androniki, Lagiou, Areti, Lagiou, Pagona, 2017. Dietary assessment methods in epidemiological research: current state of the art and future prospects. F1000Res 6, 926. https://doi.org/10.12688/f1000research.10703.1.
Perry, M.J., Marbella, A., Layde, P.M., 2006. Nonpersistent pesticide exposure self-report versus biomonitoring in farm pesticide applicators. Ann. Epidemiol. 16, 701–707. https://doi.org/10.1016/j.annepidem.2005.12.004.
Piro, F.N., Madsen, C., Naess, O., Nafstad, P., Claussen, B., 2008. A comparison of self reported air pollution problems and GIS-modeled levels of air pollution in people with and without chronic diseases. Environ. Health Glob. Access Sci. Source 7, 9. https://doi.org/10.1186/1476-069X-7-9.
Rull, R.P., Ritz, B., Shaw, G.M., 2006. Validation of self-reported proximity to agricultural crops in a case-control study of neural tube defects. J. Expo. Sci. Environ. Epidemiol. 16, 147–155. https://doi.org/10.1038/sj.jea.7500444.
Sediq, R., van der Schans, J., Dotinga, A., Alingh, R.A., Wilffert, B., Bos, J.H., Schuiling-Veninga, C.C., Hak, E., 2018. Concordance assessment of self-reported medication use in the Netherlands three-generation Lifelines Cohort study with the pharmacy database iaDB.nl: The PharmLines initiative. Clin. Epidemiol. 10, 981–989. https://doi.org/10.2147/CLEP.S163037.
Shi, X., He, F., 2012. The environmental pollution perception of residents in coal mining areas: a case study in the Hancheng mine area, Shaanxi Province, China. Environ. Manage. 50, 505–513. https://doi.org/10.1007/s00267-012-9920-8.
Šimundić, A.-M., 2009. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC 19, 203–211.
Spanish Ministry of Agriculture and Fishing Food and Environment, 2019. SIGPAC [WWW Document]. URL http://sigpac.mapa.gob.es/fega/visor/ (accessed 11.5.19).